A New View of Statistics

Will Hopkins © 2014

CONTENTS

Items on the same line or separated by dots · are on the same page.

Confidence Limits

Generalizing via Confidence Limits

P Values and Statistical Significance

Getting It Wrong

Statistical Models

Models: Important Details

Complex Models

ABOUT THESE PAGES

Last updated 16 Dec 2005

I have written these pages for researchers and students in the sport and exercise sciences. I also hope to get hits from students and researchers struggling to understand stats in other disciplines.

If you're new to stats, most of what you read here will be a new view. But even if you have done some stats, there's plenty here that's new. For example, I've discarded most details of computation, in the hope that you will get a better understanding of the concepts. Let's leave the computations to the computers! You'll also find a new unified treatment of effect statistics and their magnitudes, a new emphasis and heaps of new stuff on validity and reliability, new valid methods to calculate reliability, a new exalted position for confidence intervals, a new attack on statistical significance and hypothesis testing, the first plain-language explanation of Bayesian analysis on the Web, a new way to understand all statistical models, a new simple treatment of non-parametric analyses, a new method of doing repeated measures with missing values (yes, it's true!), new simple ways to estimate sample sizes, and best of all, a highly ethical new way to reduce sample size. And as you may have noticed, I am blazing a trail with the use of plain language for a text of this sort.

To give the pages a bit of color, I have turned the "view" into a view of hills and mountains of statistical challenges. I add features to the picture, as I climb them. Please email me if there are any you would like included, or if I'm sending people up the wrong track.

I hope some of the ideas and methods here will end up being referenced. Check how your institution or the journal you are submitting to wants you to reference websites. Here is a one way to reference the book:

Hopkins WG (2000). A New View of Statistics, http://newstatsi.org. Accessed 31 April 2006.

The (2000) refers to the first date of publication.

References to specific pages or sections should include a subject heading, for example:

Hopkins WG (2003). Bayesian analysis. In: A New View of Statistics,
http://newstats.org/generalize.html#Bayes. Accessed 31 April 2006.

The (2003) refers to the update date at the bottom of the page. Note that newstats.org and sportsci.org/resource/stats are synonymous.

To bookmark a particular page, you will have to right-click or control-click to open it in a new frame first. When it loads, you won't have a navigation/index frame on the left. You also won't get the navigation frame if you type in the URL for an individual page to load it. Clicking on Home brings up the navigation/index frame.

Acknowledgments

The following people have given valuable advice on content, layout, or stats problems:
Frank Katch , Fred Lam and Peter Herbison (University of Otago), Mary Ann Wallace, Russell Wolfinger (SAS Institute), Greg Atkinson.

Summarizing Data:
SIMPLE STATISTICS & EFFECT STATISTICS

People hate numbers, and they can't understand them in bulk. That's why you have to **summarize data** when you present results of your research. You probably know most of the peaks on this part of the statistical map already: frequency distributions, simple statistics like the mean and standard deviation, effect statistics like correlation coefficients, and so on. You may not have attempted to master things like effect size, relative frequencies and risks, a scale of magnitudes for effects, dimension reduction, validity, reliability, and the finer points of how many digits to use, but they're all easy enough.

The only other big feature on the statistical map is generalizing to a population. That's where you use a bunch of numbers from a few subjects to make inferences about *everyone's* numbers. More about that later.

BASICS

Let's make sure you understand terms like **data**, **variable**, **frequency distribution**, **probability**, and **statistics**.

Data and Variables

Data are usually just a set of numbers. Often they are a set of numbers representing the same kind of thing, like body weight. That "thing" is called a **variable**, because the numbers vary from subject to subject. If the numbers are the same, the thing is called a **constant**.

I said *usually* a set of numbers, because some data are a set of labels, names, or levels. Again, when these labels represent the same kind of thing, that thing is a variable. For example, the labels *male* and *female* are values for the variable sex. Variables with numbers as values are called **numeric**; variables with names or labels as values are called **nominal**, for obvious reasons.

Numeric variables come in several varieties. Things like height and weight are the usual kind. These can have just about any value to as many decimal places as we like, so we call them **continuous**. An example of a variable that is not continuous is a **count**, such as the number of injuries a person has experienced.

One other kind of variable can't decide whether it's numeric or nominal. A good example is competitive level, with values of *novice, club, national, international.* There is an obvious order in the levels: novice is at the bottom, club is next, and so on, so we call the variable **ordinal**. It's usual to recode each level with an integer (1=novice, 2=club, 3=national, 4=international).

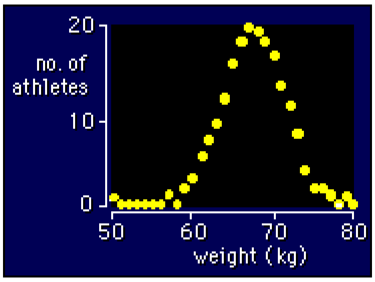
Here is an example of a data set with three nominal variables, two continuous numeric variables, one ordinal variable, and one counting variable:

subject	**weight**	**height**	**sex**	**sport**	**level**	**injuries**
AJH	63	170	female	swimming	3	0
CJD	78	185	male	basketball	2	0
NIH	68		female	basketball	5	2
ERF	69	177	male		1	1
MBA	etc.	etc.	etc.	etc.	etc.	etc.

Each row in the data set represents the values of all the variables for one subject. It's called an **observation**. You can have **missing values** in the data set, too, as shown.

Data for more than a few subjects and variables have to be summarized to make them palatable. You can't trot out the whole data set every time you want to talk about it.

Frequency Distributions

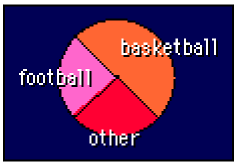


For numeric variables, one important way to summarize the values is to graph them as a **frequency distribution**. Here's what the weights of 200 athletes might look like in a frequency distribution done as a **scatter plot**, which shows a point for the number of times each weight occurs.

You can also show the frequencies as vertical bars rather than points, in which case the figure is called a **histogram**. Most stats programs also have a clever way to show the values as a kind of histogram called a **stem-and-leaf plot**. When you see one it will be obvious what is going on.

It's normal for data to have a symmetrical bell-shaped frequency distribution like the one shown. Hence the name: the **normal distribution**. Exactly why most things are normally distributed is a bit of a mystery.

When you have lots of values for a variable, it's a good idea to get a stats program to do a frequency distribution or stem-and-leaf plot, so you can see if there are any obviously wrong **outliers**. Outliers are often just errors in data entry. It might be worth checking out the original data for the person on 50 kg in the above figure. You certainly would if the value was 40 kg. Even if the value is correct, you might have a good reason to exclude that observation.



Summarizing the values of a nominal variable like sex is a simple matter. All you need is the frequency of each level. For example, a group of athletes might consist of 101 basketballers, 49 footballers, and 51 others. One occurrence of the new sport football would be an example of an outlier in need of correction. You can display the frequencies graphically as proportions in a **pie chart**, as shown, or as a bar divided up in the right proportions. Pie charts seem to be frowned on in scientific publications, but you see them in magazines.

Probability

When you keep having a shot at something, like rolling a six-sided die and hoping for a "four", what proportion of your shots end up being successful? If it's a symmetrical die, the answer is obviously 1 in 6. That **proportion** is known as **probability**. We usually write the proportion or probability as p. In this example, p = 1/6 = 0.1666 = 0.17 (to two decimal places).

Probability is obviously a number between 0 and 1. When it's 0, there's no way you'll be successful, and when it's 1 you'll win every time. You can't have negative probability.

We can represent probability in several other ways. In the above example, we can talk about 1 chance in 6, 17 times in 100, 17%, a likelihood of 0.17, or 17% likely. You'll also meet **odds** of 1 to 5, which means 1 success for every 5 failures. Odds of 1 to 1 means a 50% chance of something happening (as in tossing a coin and getting a head), and odds of 99 to 1 means it will happen 99 times out of 100 (as in bad weather on a public holiday).

A **probability distribution** is just a frequency distribution with each frequency divided by the total number of observations. It follows (although it's not obvious) that the area under a probability distribution has something to do with the probability of getting certain numbers. In the above example of the distribution of people's weights, if you draw someone at random from the population; the chance that they will have a weight between 60 and 70 kg is the area under the curve between 60 and 70 kg. What's more, the total area under a probability distribution is 1. Hmm... Too academic. Too technical. Not essential. But you *will* need to feel comfortable with probability when we deal with p value, confidence limits, relative risk and odds ratio.

Statistics

A statistic is a number summarizing some aspect of the data. There are three kinds of statistic: **simple** statistics, **effect** statistics, and **test** statistics. Simple statistics are also known as **univariate** statistics, because they summarize the values of one variable. Effect statistics summarize the relationship between the values of two or more variables. Simple and effect statistics are **descriptive** statistics, as opposed to test statistics, which can wait until later!

Let's start with simple statistics. The simplest of all is a **count** of the number of numbers or levels, also known as **sample size**. That's as far as it goes for a nominal variable. For numeric variables, we usually use two more simple statistics to give people an idea of what the original numbers are like: a statistic to represent the **middle** values of the data (on the next page), and a statistic to show how the data are **spread** out (on following pages).

Simple Statistics: THE MIDDLE

A statistic that represents the middle of the data is called a **measure of centrality**. The best is the **mean** or **average**. Just add up all the numbers and divide by the sample size. The mean is the best measure, partly because it uses more information in the data than any other measure of centrality.

The **median**, or "middle" number, can be useful for data with a non-normal distribution. To work it out, arrange the numbers in rank order (smallest to largest), then count in from one end until you find the middle. (If the sample size is an even number, take the average of the two middle numbers.) The median is not affected by outliers, which is a big point in its favor. But if you're interested in getting an estimate of the center of a population or of a subgroup of a population--and you usually are--the median is a coarse or "noisy" measure.

The **mode**, or most frequent number, is the only other measure of centrality you'll ever encounter. I've never used it.

Simple Statistics: THE SPREAD

Some statistics give an idea of **spread, variation**, or **dispersion** of the numbers. The simplest measure of spread is the **range**, expressed either as the biggest and smallest number in the data (e.g. 61-74), or as the difference between the biggest and smallest (e.g. 13).

The range is a bad measure of spread, for two reasons. First, it's dictated by outliers, whether they're errors in data entry or genuine values. Secondly, the range is dependent on the size of your sample: the more numbers, the bigger the range is likely to be. Two measures of spread that avoid these problems are the **standard deviation** (SD) and **percentile ranges**. I'll deal with these separately, and with these other measures of variation: the **root mean square error** (RMSE) and the **standard error of the estimate** (SEE). I explain on a separate page why the **standard error of the mean** is a measure of spread you should *not* use.

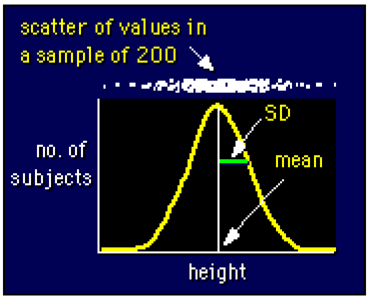
The statistics most people use to describe a set of numbers are sample size, mean, and standard deviation. All you need to define the shape of the normal distribution is the mean and the standard deviation. The mean and standard deviation are often written as mean ± SD: 67.8 ± 3.6 kg, for example.

In dealing with the spread in a bunch of numbers, we often think about the numbers as representing values of some characteristic, such as weight, for different subjects. But the bunch of numbers could represent the weight of a single subject measured many times. We talk about **between-subject variation** and **within-subject variation** to distinguish between these two types of spread. Within-subject variation comes up soon as a useful measure of reliability.

Standard Deviation

The standard deviation is usually the best measure of spread. It has a complicated definition: take the distance of each number from the mean, square it, average the result, then take the square root. In short, it's the root mean square of the distances (or differences) from mean. It's usually abbreviated as SD in scientific journals and as s in stats books and stats journals.

Actually, when you take the mean or average of the squares, you have to divide by n - 1 (one less than the sample size). Dividing by n gives you a biased estimate. Obviously for large n it doesn't matter whether you use n or n-1, but for n<20 it starts to make a difference. When using a calculator to work out a standard deviation, press the s_{n-1} button, not s_n.



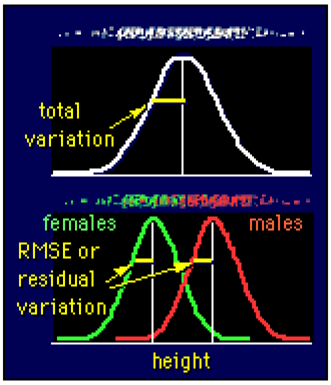
The figure shows how big one SD looks on a frequency distribution for a normally distributed variable like height. I've shown the frequencies (number of subjects) as a continuous curve rather than as discrete points for each value of height. The best way to think about the SD is that about two-thirds of the values of a variable are found within one SD each side of the mean.

The standard deviation is sometimes expressed as a percent of the mean, in which case it's known as a **coefficient of variation**. When the SD and mean come from repeated measurements of a single subject, the resulting coefficient of variation is an important measure of reliability. This form of within-subject variation is particularly valuable for sport scientists interested in the variability an individual athlete's performance from competition to competition or from field test to field test. The coefficient of variation of an individual athlete's performance is typically a few percent.

A measure of spread closely related to the SD is the **variance**, which is simply the square of the SD. I can't show you variance on a diagram. Statisticians prefer it to the SD, but it's not much use for researchers.

Root Mean-Square Error (RMSE)

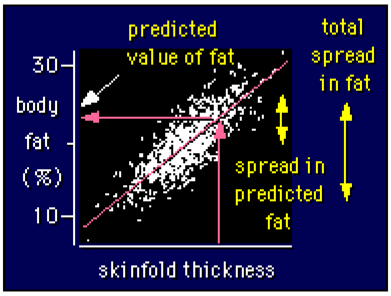
The RMSE is a kind of generalized standard deviation. It pops up whenever you look for differences between subgroups or for other effects or relationships between variables. It's the spread left over when you have accounted for any such relationships in your data, or (same thing) when you have fitted a statistical model to the data. Hence its other name, **residual variation**. I'll say more about residuals for models, about fitting models in general, and about fitting them to data like these much later.



Here's an example. Suppose you have heights for a group of females and males. If you analyze the data without regard to the sex of the subjects, the measure of spread you get will be the **total variation**. But stats programs can take into account the sex of each subject, work out the means for the boys and the girls, then derive a *single* SD that will do for the boys and the girls. That single SD is the RMSE. Yes, you can also work out the SDs for the boys and girls separately, but you may need a single one to calculate effect sizes. You can't simply average the SDs.

Standard Error of the Estimate (SEE)

The SEE is another example of a root mean square error. This time we're fitting a line to the data, to make predictions. The SEE tells us something about the accuracy of the predictions.

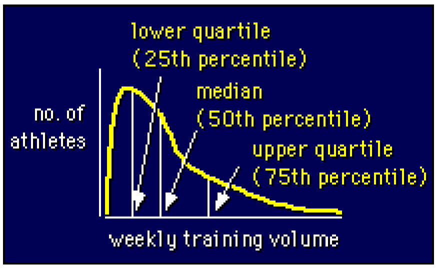


The figure shows an important example: how to predict body fat from skinfold thickness. You measure the skinfold thickness and body fat of several hundred subjects, then draw the best straight line through the points. The SEE represents the scatter of points about the line for any given value of skinfold thickness, which means it's the "error"--actually a standard deviation--in predicting body fat from a given value of skinfold thickness. As drawn for these imaginary data, it's about 3%. Whenever you measure the skinfold thickness on subjects in future and use the straight line to predict their body fat, you will know that you could be wrong by typically 3%.

Incidentally, the SEE--the scatter of body fat about the line for a given skinfold thickness--is assumed to be the same for every value of skinfold thickness. In other words, it doesn't matter where you are on the line, it's the same scatter in the vertical direction. I know it looks like there is less scatter at the ends of the line, but that's only because there are less points there. A hard one for newbies to understand!

Here's another important "incidentally". You can use a prediction line only for subjects similar to (drawn from the same population as) the subjects you used to make the prediction line in the first place. A line based on active young female athletes is no good for predicting body fat in sedentary middle-aged males. The SEE would also be wrong.

Percentile Ranges



The most common of these is the **interquartile range**, although even this is a seldom-visited feature on the statistical map. It is used with the median to give an idea of centrality and spread of **skewed** or otherwise grossly non-normally-distributed variables. Measures of training are often skewed enough to merit use of percentiles instead of the mean and standard deviation. For example, weekly training in a group of novice athletes might have a median of 5 and an interquartile range of 3-12 hours/week.

Here's something challenging for the real lovers of numbers. The mean ± SD encloses 68% of the data on average for a normally distributed variable. So if you want to use a percentile range that corresponds to the mean ± SD, what should it be? Answer: 16th-84th. If I had my way, this measure would replace the interquartile range. We could call it the standard percentile range…

We're right out on the horizon now. Let's get back to familiar territory.

EFFECT STATISTICS

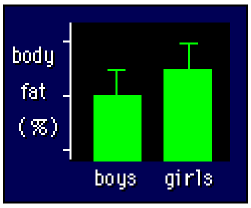
All the statistics we've met so far summarize a set of values of a single variable. But it's possible to have statistics describing the relationship between *two or more* sets of numbers. These are the statistics that really matter in research.

There is no agreed generic name for these statistics. I've seen *measures of effect* in the literature, so let's call them **effect statistics**. The *effect* refers to the idea that one variable has an effect on another. The main effect statistics are the **difference in means** (this page), the**correlation coefficient** (next page) and **relative frequency** (following page). Each of these effect statistics comes in several varieties, I close this section with a page on **a scale of magnitudes** for effect statistics

Difference in Means

Sometimes you can express the finding of a study simply as a difference between the means of two groups, in the original raw units of measurement. For example, a group of elite male runners has a mean body mass of 66.4 kg, whereas a subelite group weighs in with an average of 68.5. The difference of 2.1 kg is the effect.

Depending on the variable, you often want to talk about the difference between means as a **percent difference.** In the above example, you could say that the subelite runners are 3.2% heavier than the elites (2.1/66.4 = 0.032 = 3.2%). Percent differences are a natural way to express differences in the mean of variables that need log transformation. Percent effects are particularly appropriate for measures of athletic performance.



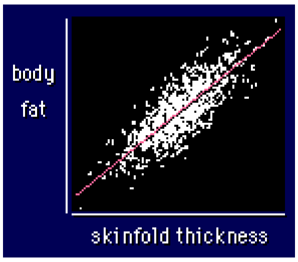
Converting the difference to a percent is one way to make the difference **dimensionless**, and therefore more generic. Another important way is to express the difference as a fraction or multiple of a standard deviation. Work out the difference between the means, then divide it by the average standard deviation for the two groups. What you end up with is the **standardized difference in the means,** a number that represents "how many standard deviations" the two groups differ by. Look closely at the imaginary example in the figure and work out the effect size for the difference in body fat between boys and girls. Answer: one unit, or 1.0, or one standard deviation. Note that the unit of measurement for body fat is irrelevant. I've shown it as the usual % of body mass, but it could be kg or pounds--the effect-size statistic has the same value. The effect-size statistic is appropriate for studies of population health, where differences or changes in means that impact the average person are paramount.

The standardized difference in the means is sometimes known simply as the effect-size statistic, although this term confuses the concept with the magnitude of other kinds of effect. A page on this topic comes up shortly. Meanwhile, to get you thinking about it, how big is the effect shown in the above figure? This difference of one standard deviation has been regarded as *large*, although I now think it's only a *moderate* effect. Anything less than 0.2 standard deviations isn't worth worrying about.

The example above is for two groups of subjects, but you should also use the concept of effect size when looking at changes in the mean as a result of an experiment. For example, the above two bars could represent muscle mass before and after treatment with anabolic steroids. In this case you use the SD of the pre scores only to standardize the effect. (I'd figured this out years ago, but until Oct 06 I missed a mistaken assertion on this page that you average the pre and post SDs. Sorry about that.) Some people think mistakenly that you should use the SD of the change scores to standardize effects in experiments. If you have a control group as well, you use the SD of all the pre scores, and you subtract the change in the control group from the change in the experimental group to get the magnitude of the experimental effect.

When you understand effect sizes, you'll know why you should always show standard deviations rather than standard errors of the mean with means. See the page devoted this important issue.

Correlation Coefficient



Let's return to our example of skinfolds and body fat. The correlation coefficient (r) indicates the extent to which the pairs of numbers for these two variables lie on a straight line. The correlation for this example is 0.9. If the trend went downward rather than upwards, the correlation would be -0.9. For perfect linearity, r = ±1. If there is no linear trend at all--for example, if there is a random scatter of points--the value of r is close to zero. Points distributed evenly around a circle would also give a correlation of near zero, because there would be no overall linear trend.

Which brings us to the question of how big a correlation has to be before it means anything. Correlations of less than 0.1 are as good as garbage. The correlation shown, 0.9, is very strong. Correlations have to be this good before you can talk about accurately predicting the Y value from the X value, especially when you want to use the result of the prediction to rank people. You can understand that by looking at the scatter of body fat about the line for a given value of skinfold thickness (the standard error of the estimate): it's still quite large, even for this correlation of 0.9. More on magnitudes of correlations shortly.

The details of calculation of correlations needn't concern us, because the stats packages do all that for us. But you should learn that the correlation between two variables X and Y is defined as the **covariance** of X with Y (covarXY) divided by the product of the standard deviation of X (stdevX) and the standard deviation of Y (stdevY):

r = covarXY/(stdevX·stdevY).

We've already met the variance: it's the mean value of all the differences from the mean multiplied by themselves (=squared). The covariance is similar: it's the mean value of all the pairs of differences from the mean for X multiplied by the differences from the mean for Y. If X and Y aren't closely related to each other, they don't *co-vary,* so the covariance is small, so the correlation is small. If X and Y are closely related, covarXY turns out to be almost the same as stdevX·stdevY, so the correlation is almost 1.

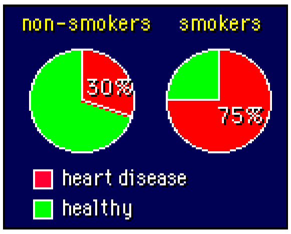
There are several important kinds of correlation, differing in the details of calculation. The most common is known as the **Pearson** (after a famous statistician). An older name is the **product-moment** correlation, which refers to the way it's calculated. The Pearson is what you get when you fit the best straight line to a set of points, such that the points are closest to the line when measured in the Y direction--the usual least-squares line, in other words. The topic of fitting lines and curves comes up in more detail later.

By the way, if the X and Y variables have the same standard deviation, the slope of the line is the correlation coefficient. Or to put it another way, if you **normalize** the X and Y variables by dividing them by their standard deviations, the slope of the line is the correlation coefficient.

Two more important kinds of correlation are the **Spearman** and **intraclass correlation coefficient (ICC)**. The Spearman comes up later in connection with non-parametric tests. The ICC is used as a measure of the reliability of a variable, whereas the Pearson is used for thevalidity of the variable. The values of the Pearson, Spearman, and intraclass correlation coefficients are usually similar for the same set of data.

The strength of the relationship between X and Y is sometimes expressed by squaring the correlation coefficient and multiplying by 100. The resulting statistic is known as **variance explained** (or **R^2**). Example: a correlation of 0.5 means 0.5^2x100 = 25% of the variance in Y is "explained" or predicted by the X variable. The reason why squaring a correlation results in a proportion of variance is a consequence of the way correlation is defined. You don't need to know the details right now. See later.

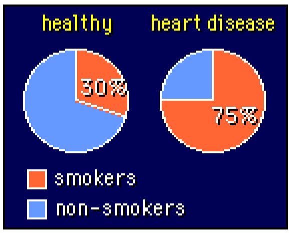
Difference in Frequency



We seldom use the raw counts of something when we compare frequencies. In the example shown on the right, there may have been 600 non-smokers who had heart disease, out of a sample of 2000 non-smokers altogether, so we usually talk about 30% of the non-smokers having heart disease. A percent frequency makes it easier to compare the rate of heart disease in other groups, for example smokers. But how do we now actually compare the frequencies? The simplest way is to subtract them: the difference in rate of heart disease is 45%. I think that's the best way, but it is not the usual way. Instead, researchers usually divide one frequency by the other. In the example, smokers would be 75/30, or 2.5 times as likely to develop heart disease as non-smokers. Or to put it another way, the **relative risk** of developing heart disease for smokers is 2.5. If the frequency of heart disease was the same in both groups, the relative risk would be 1.0, and if the frequency was less in smokers, the relative risk would be less than 1.0.

It's hard to put a figure on what are considered small, medium and large differences between the frequencies of something in two groups, because it depends on the frequencies. If one group has about 50% with a characteristic, a frequency of 60% or 40% in the other group can be considered small. That difference corresponds to a relative risk of about 1.2 (or 0.8, depending which way around the frequencies are). Once the frequencies get low (e.g. 1% in one group), relative risks have to be 2 or more before people get excited.

Notice that the two groups differ in **exposure** to something that might cause the disease. A somewhat different statistic, the **odds ratio**, is used when the basis of the grouping is whether subjects already have the disease: in other words, when the groups are **cases** and **controls**. In the example shown, the odds of being a smoker in the heart-disease group are 75/25 = 3. Similarly, the odds of being a smoker in the healthy group are 30/70 = 0.43. The odds ratio is therefore 3/0.43 = 7. Interpret this statistic as "seven people with heart disease smoke for every healthy person who smokes". Or, if you had two people in front of you, a healthy person who smokes and a person with heart disease, you would break even in the long run by betting at odds of 7:1 that the person with heart disease is a smoker. Fine, but I still have trouble getting my brain around this statistic. Are those odds good or bad, in terms of the effect of smoking on heart disease? I don't know. I guess I don't work with this statistic enough to have a feel for it. (I used to have here "seven smokers have heart disease for every one smoker who doesn't" or "if you are a smoker, odds are 7 to 1 that you have heart disease", but these interpretations are wrong. Thanks, Chris Rhoads!).



A Scale of Magnitudes for Effect Statistics

Suppose you get a correlation of 0.47 between two variables. Is that big or small, in the scheme of things? If you haven't a clue, you're not alone. Most people don't know how to interpret the magnitude of a correlation, or the magnitude of any other effect statistic. But people can understand *trivial, small, moderate,* and *large,* so qualitative terms like these need to be used when you discuss results. One day, stats programs will include these terms in their output. In the meantime, we have to do the job manually using a scale of magnitudes. I'll now explain a scale of magnitudes for linear trends (using the correlation coefficient), differences in means (using the standardized difference), and relative frequencies (using relative risks, odds ratios, and differences in frequencies).

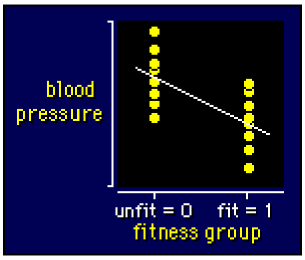
Correlations

Jacob Cohen has written the most on this topic. In his well-known book he suggested, a little ambiguously, that a correlation of 0.5 is large, 0.3 is moderate, and 0.1 is small (Cohen, 1988). The usual interpretation of this statement is that anything greater than 0.5 is large, 0.5-0.3 is moderate, 0.3-0.1 is small, and anything smaller than 0.1 is insubstantial, trivial, or otherwise not worth worrying about. His corresponding thresholds for standardized differences in means are 0.8, 0.5 and 0.2. He did not provide thresholds for the relative risk and odds ratio.

For me, the main justification for this scale of correlations comes from the interpretation of the correlation coefficient as the slope of the line between two variables when their standard deviations are the same. For example, if the correlation between height (X variable) and weight (Y variable) is 0.7, then individuals who differ in height by one standard deviation will on average differ in weight by only 0.7 of a standard deviation. So, for a correlation of 0.1, the change in Y is only one-tenth of the change in X. That seems a reasonable justification for calling 0.1 the smallest worthwhile correlation. I guess it's also reasonable to accept that a change in Y of one half that in X (corresponding to r = 0.5) is also the threshold for a large effect, and r = 0.3 seems a logical way to draw the line between small and moderate correlations.

Threshold values for standardized differences or changes in means and for relative frequency can be derived by converting these statistics to correlations. The procedure is a little artificial, so the resulting values need to be scrutinized to ensure they make sense. Here's how it's done.

Differences in Means



To work out a scale of magnitudes for differences or changes in means, you need a dimensionless measure comparable to the correlation coefficient. The best and possibly only such measure is the standardized difference. Cohen used the letter *d* to represent the standardized difference, and it is often known as *Cohen's d*. To see how to get thresholds for d from those for correlations, let's introduce a new predictor variable with the value of 0 for one group and 1 for the other, as shown in this example for the effect of fitness on blood pressure. (We can assign any number at all to each group, not just 0 and 1.) We then calculate the correlation between this variable and the dependent variable. If the standardized difference between the means is d (the difference in the means divided by the standard deviation in either group, here assumed to be the same), it's possible to show from the definition of a correlation that r = d/√(d^2+4), or rearranging, d = 2r/√(1-r^2). It follows that correlations of 0.1, 0.3, and 0.5 correspond to standardized differences in means of 0.20, 0.63, and 1.15.

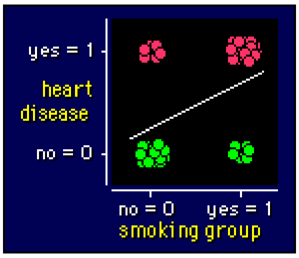
Problem! Cohen's thresholds for small, moderate and large are 0.20, 0.50 and 0.80. The lowest of these two sets of values agree (0.20), but the others don't. Cohen derived his thresholds from a consideration of non-overlap of the distributions of values in the two groups. He chose certain arbitrary amounts of non-overlap as defining his thresholds. The thresholds for *small* obviously correspond, but the others don't.

Something like Cohen's thresholds for standardized differences can be got by making the independent variable normally distributed, then "dichotomizing" it by splitting its values down the middle to make the two fitness groups. Correlations of 0.1, 0.3, and 0.5 then turn into standardized differences of 0.17, 0.50, and 0.87: yet another set of thresholds! Which set is correct? I think that this dichotomizing operation throws away information, and that therefore the values of 0.17, 0.50 and 0.87 underestimate the thresholds.

I'm happy to agree with Cohen that 0.20 is the threshold for smallest standardized differences in a mean. If we also assume that the thresholds of 0.1, 0.3 and 0.5 for correlations are acceptable, there is another approach to demonstrating that the other thresholds for standardized differences in the mean should be 0.63 and 1.15. Assume further that the X and Y variables are normally distributed. Consider first a correlation of 0.1. Imagine you are comparing two individuals with X values that differ by an amount a. They will, of course, have different Y values. From one of the meanings of the correlation coefficient, the difference in the Y values is a.r.SDy/SDx, where SDy and SDx are the standard deviations of the Y and X variables. To standardize this difference, we have to divide it by the appropriate standard deviation, which in this case is the standard error of the estimate, given by SDy√(1-r^2). The standardized difference in the Y values is therefore a.r.SDy/SDx/(SDy√(1-r^2)) = (a/SDx)(r/√(1-r^2)). So, if we want a smallest correlation of 0.1 to be equivalent to a smallest standardized difference of 0.20 between two individuals, the individuals have to differ on average by 2 standard deviations of the X values: (2SDx/SDx)(0.1/√(1-0.1^2) = 0.20. It follows that the standardized difference corresponding to any correlation r should be the difference corresponding to 2 standard deviations of the X values, and the formula to convert a correlation to an equivalent standardized difference in the means is therefore 2r/√(1-r^2). Note that this formula is the same as in the first paragraph of this section, so the thresholds for moderate and large are 0.63 and 1.15.

One reality check on these thresholds comes from considering the average separation between individuals in a normally distributed population. It turns out to be 1.13 standard deviations, which is a standardized difference of 1.13. So we have to ask: is it reasonable that the average difference between individuals in a population should be on the threshold between moderate and large? I think so, and I therefore think that Cohen's 0.5 and 0.8 are too low to define the thresholds for moderate and large effects.

Relative Frequencies



To work out a scale for comparing frequencies, we have to code not only the grouping variable, but also the dependent variable. See the example on the right, in which a cluster of points represents the frequencies for each level of the independent and dependent variables. Once again the values of 0 and 1 for the variables don't matter, but if we represent the frequencies as percents in each group, we get something really nice. For the example shown, heart disease was 75% in the smoking group and 30% in the non-smoking group. The difference in frequencies (75 - 30 = 45%) divided by 100 is 0.45, which turns out to be the correlation between our two newly coded variables. This result--the correlation times 100 equals the difference in percent frequencies--is true for all frequencies. The threshold correlations of 0.1, 0.3, and 0.5 therefore convert to thresholds of 10, 30 and 50 for differences in percent frequencies between the occurrence of something in two groups.

Now, are you happy with the notion that a difference of 10% in the frequency of something between two groups is indeed *small?* For example, if you made sedentary people active and thereby reduced the incidence of heart disease from 55% to 45% in some age group, would that be a small gain? At first glance you'd think this gain might be better described as *moderate.* Perhaps the way to view it is that the 10% in question is only one tenth of the entire group. On an absolute population basis, we may be talking about a lot of people, but it's still only one in 10. The threshold between *moderate* and *large* represents something that affects half the group, which seems OK. The boundary between *small* and *moderate* (three people in 10) is also acceptable.

Frequency differences do not convert simply into relative risks, because the values of this statistic depend on the frequencies in each group. For example, the threshold frequency difference of 10% for the smallest worthwhile effect represents a relative risk of 55/45 or 1.22 if the frequencies are 55% and 45%, but the relative risk is 11 if the frequencies are 11% and 1%. The odds ratio is even more sensitive to the absolute frequencies in each group. The smallest values for the relative risk and odds ratio occur when the frequencies in the two groups are symmetrically disposed about 50% (55-45, 60-40, 65-35 and so on).

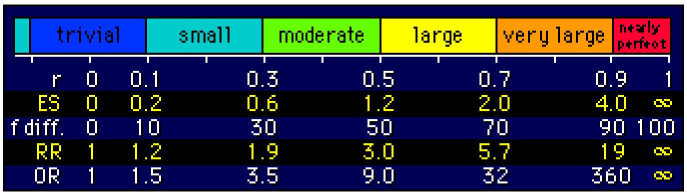
The Complete Scale

It seems to me that the vista of large effects is left unexplored by Cohen's scale. Surely more than just *large* can be applied to the correlations that lie between 0.5 and 1? What's missing from the picture is a rationale for breaking up this big half of the scale with a couple more levels. Here's the way I do it:

	trivial	small	moderate	large	very large	nearly perfect	perfect
Correlation	0.0	0.1	0.3	0.5	0.7	0.9	1
Diff. in means	0.0	0.2	0.6	1.2	2.0	4.0	infinite
Freq. diff.	0	10	30	50	70	90	100
Rel. risk	1.0	1.2	1.9	3.0	5.7	19	infinite
Odds ratio	1.0	1.5	3.5	9.0	32	360	infinite

I've adopted a Likert-scale approach by using *very* for the level above large, and I've assigned it to a correlation of 0.7 to keep the scale linear for correlations and frequency differences. A level of magnitude above *very large* is warranted for correlations, because a value of 0.9 is a kind of threshold for validity when the associated straight line is used to rank individuals, and reliability needs to be greater than 0.9 to be most useful for reducing sample sizes in longitudinal studies. I've opted for *nearly perfect* to describe these correlations. Values for the other effect statistics were calculated as before, and the values for the relative risk and odds ratio are the minimum values for these statistics.

To finish, here is a graphical representation of the scale...



	trivial		small	moderate	large	very large	nearly perfect
r	0	0.1	0.3	0.5	0.7	0.9	1
ES	0	0.2	0.6	1.2	2.0	4.0	∞
f diff.	0	10	30	50	70	90	100
RR	1	1.2	1.9	3.0	5.7	19	∞
OR	1	1.5	3.5	9.0	32	360	∞

...and a table of synonyms for the descriptors (for simplicity, only for the correlation coefficient). Use of these synonyms shouldn't lead to any confusion about the magnitude of the effect:

Correlation Coefficient	Descriptor
0.0-0.1	trivial, very small, insubstantial, tiny, practically zero
0.1-0.3	small, low, minor
0.3-0.5	moderate, medium
0.5-0.7	large, high, major
0.7-0.9	very large, very high, huge
0.9-1	nearly, practically, or almost: perfect, distinct, infinite

SAS programs that generated the results on this page are attached.

Other Effect Statistics

Cohen devised several other effect statistics and discussed their magnitudes, but I have not seen these statistics in publications. He also considered whether, for example, variance explained (the correlation squared) might be a more suitable scale to represent magnitude of linearity, especially when you take into account the useful additive property of variance explained in such things as stepwise regression. He rejected it, though, because a correlation of 0.1 corresponds to a variance explained of only 1%, which he thought did not convey adequately the magnitude of such a correlation. I agree.

The so-called **common-language effect statistic** (McGraw & Wong, 1992) or **probability of superiority** represents a more recent attempt on the summit of a universal scale of magnitudes. This statistic is easiest to understand when you compare two groups whose means differ. The probability of superiority is the probability that someone drawn at random from one group will have a higher value than someone drawn from the other group. The problem here is that no difference between the means implies a value of 50% or 0.5 (equal chance that the person will have a higher or lower value). A value of 50% for no difference doesn't feel right.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New Jersey: Lawrence Erlbaum.

McGraw, K. O., & Wong, S. P. (1992). A common language effect-size statistic. Psychological Bulletin, 111, 361-365.

Summarizing Data:
DIMENSION REDUCTION

Dimension reduction is a way of devising one or two variables to summarize the information contained in a whole lot of other variables. The three methods of dimension reduction are **principal components analysis**, **factor analysis**, and **cluster analysis**

PRINCIPAL COMPONENTS ANALYSIS

When you take lots of different measurements in a study, you sometimes want to combine them in some way to derive just one or two measures that summarize some aspect of the data. For example, you might have five different measures of body size, but would like to have simply one or two summary measure that combine the five. The summary measures are provided by principal components analysis.

All you do is tell the stats program what variables you want it to analyze. It comes up with a linear combination of the variables that somehow captures the biggest amount of common variation in all of them. It then goes on to produce another linear combination that captures the biggest amount of variation in what's left, and so. If you start with three variables, you'll get three principal components. The nice thing about them is, they are not correlated with each other, so they represent three totally independent measures. Exactly what they represent in reality has to be decided by looking at the **weighting factors** that the stats program derives to make the principal components. Sometimes it's not obvious that they represent anything meaningful, and you might have to abandon this approach.

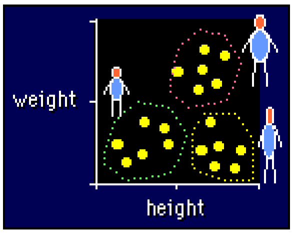
FACTOR ANALYSIS

Here you want combinations of variables with equal weighting, and you're generally not concerned if the resulting composites are correlated. This method is used by psychologists (or their statisticians) to derive distinct dimensions of the psyche from subsets of items in multi-item questions. Factor analysis divides the items into subsets such that items correlate well within each subset but not so well between subsets. Each subject then gets a mean score for the items in each subset . The researcher has to decide what to call the mean scores by looking at the wording of the items.

It's a few years since I did factor analysis, which is why this section is so short! If there is a demand for it, I will include the detail on such things as promax rotation and deciding where to draw the line for inclusion of variables in a factor.

CLUSTER ANALYSIS

This is a particularly severe form of dimension reduction that reduces all variables and data down to one variable with only a few values (e.g. group A, group B, and group C). It's easiest to understand from the example in the figure, which shows heights and weights for a bunch of people who obviously fall into three groups or "clusters":



You can let the stats program decide on the number of clusters, or you can force it to find as many as you like. The program decides which observations belong to which cluster by minimizing the distances between points in each cluster. You are not restricted to two variables, of course. It's impossible to imagine clusters for more than three variables (unless you are an Einstein), but the stats program handles it without any problem.

Cluster analysis is used in market research, where you want to identify a few major target groups in a population. And it's a cool way of identifying groups in the population with particular lifestyles. Variables used in the cluster analysis might be age, sex, socio-economic status, level of physical activity, measures of diet, and so on.

Summarizing Data:
PRECISION OF MEASUREMENT

How precise are your measurements? An important question, because the lower the precision, the more subjects you'll need in your study to make up for the "noise" in your measurements. Even with a larger sample, noisy data can be hard to interpret. And if you are an applied scientist in the business of testing and assessing clients, you need special care when interpreting results of noisy tests.

The two most important aspects of precision are **reliability** and **validity**. Reliability refers to the reproducibility of a measurement. You quantify reliability simply by taking several measurements on the same subjects. Poor reliability degrades the precision of a single measurement and reduces your ability to track changes in measurements in the clinic or in experimental studies. Validity refers to the agreement between the value of a measurement and its true value. You quantify validity by comparing your measurements with values that are as close to the true values as possible. Poor validity also degrades the precision of a single measurement, and it reduces your ability to characterize relationships between variables in descriptive studies.

The concepts of reliability and validity are related. For example, a little thought will satisfy you that measurements can be reliable but not valid, and that a valid measurement must be reliable. But we usually deal with these two concepts separately, either because most researchers study them separately, or because bringing the two concepts together is mathematically difficult. I've had a shot at combining them, but there's much more work to do.

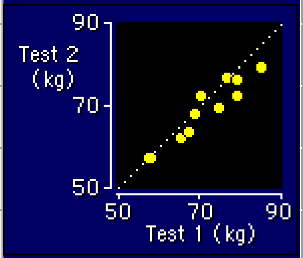
Here's the route map for this excursion. We begin with measures of reliability, then there are separate pages for applications of reliability and calculations for reliability. We'll deal with measures of validity and calculations for validity on one page, followed by applications of validity. Along the way there are three spreadsheets for various calculations: the precision of a subject's true value using reliability or validity, calculating reliability between pairs of trials, and calculating validity. Then there's a quick and easy page on precision in reporting measurements, and finally a page devoted to the all-important question of mean ± SD vs mean ± SEM. Some of the material on these pages is in Hopkins (2000).

Update Oct 2011: view this slideshow on validity and reliability for an overview of the important principles.

For a **Powerpoint presentation** (slide show) on the essentials of reliability and some of its uses (assessing individuals, estimating sample sizes, estimating individual responses), click here. This presentation was part of a mini-symposium entitled "Reliability, a Crucial Issue for Clinicians and Researchers" at the 2001 annual meeting of the American College of Sports Medicine in Baltimore.

MEASURES OF RELIABILITY

The most common form of reliability is **retest reliability**, which refers to the reproducibility of values of a variable when you measure the same subjects twice or more. Let's get down to the detail of how we quantify it. The data below, and the figure, show an example of high reliability for measurement of weight, for 10 people weighed twice with a gap of two weeks between tests. I'll use this example to explain the three important components of retest reliability: change in the mean, typical error, and retest correlation. I'll finish this page with two other measures of reliability: kappa coefficient and alpha reliability.

Test 1	Test 2	
57.5	57.4	
65.6	63.2	
67.0	66.5	
68.5	69.9	
70.8	72.8	
72.2	70.1	
74.9	75.6	
76.0	75.2	
76.1	72.8	
83.1	79.0	

Change in the Mean

The dotted line in the figure is the line representing identical weights on retest. Notice that most of the subjects are below the line: they were a bit lighter in the second test. To put a number on the change in weight, you subtract the mean of all the subjects for Test 1 (71.2 kg) from that for Test 2 (70.3 kg). The result (-0.9 kg) is the **change in the mean**: the difference between the means for two tests. The change consists of two components: a **random change** and a **systematic change**.

Random change in the mean is due to so-called **sampling error**. This kind of change arises purely from the typical error, which is like a randomly selected number added to or subtracted from the true value every time you take a measurement. The random change is smaller with larger sample sizes, because the random errors from all the measurements contributing to the mean tend to cancel out more.

Systematic change in the mean is a non-random change in the value between two trials. If the drop in weight in our example is a systematic change, it could be due to changes in the the subjects' behavior between trials. In tests of human performance that depend on effort or motivation, subjects might also perform the second trial better because they want to improve. Performance can be worse in a second trial if fatigue from the first trial is present at the time of the second trial. Performance can also decline in a series of trials, owing to loss of motivation.

Systematic change in the mean is an important issue when subjects perform a series of trials as part of a monitoring program. The subjects are usually monitored to determine the effects of an intervention (e.g., a change in diet or training), so it is important to perform enough trials to make learning effects or other systematic changes negligible before applying the intervention.

Systematic change is less of a worry for researchers performing a controlled study, because only the relative change in means for both groups provides evidence of an effect. Even so, the magnitude of the systematic change is likely to differ between individuals, and these individual differences make the test less reliable by increasing the typical error. You should therefore choose or design tests or equipment with small learning effects, or you should get subjects to perform practice (familiarization) trials to reduce learning effects.

How do you tell whether an observed change in the mean is a reproducible systematic effect? You work out and interpret the confidence limits for the mean, which represent the likely range of the true (systematic) change.

Typical Error of Measurement

Notice that our subjects didn't have exactly the same weight in the first and second tests. Sure, part of the problem is that everyone got a bit lighter, but even when you take the shift in the mean out of the picture, the weights on retest aren't exactly the same. To see what I mean, imagine that you reweighed one subject many times, with two weeks between each weighing. You might get something like:

72.2, 70.1, 68.5, 69.9, 67.9, 69.6...

The first few weights show a slight trend downwards--our subjects decided to lose a bit of weight, remember--then the weights level off, apart from a random variation of about a kilogram. That random variation is the **typical error**. We quantify it as the standard deviation in each subject's measurements between tests, after any shifts in the mean have been taken into account. The official name is the **within-subject standard deviation**, or the **standard error of measurement**. From now on I will refer to it as the **typical error of measurement**, or simply typical error, because its value is indeed the typical error or variation in a subject's value from measurement to measurement.

We talk about variation in measurements as *error,* but it's important to realize that only part of the variation is due to error in the sense of **technological error** arising from the apparatus. In fact, in the above example the variation is due almost entirely to **biological variation**in the weight of the subject. If we were to reweigh the subject with two minutes between weighings rather than two weeks, we'd get pure technological error: the noise in the scales. (We might have to take into account the fact that the subject would be getting slightly lighter all the time, through evaporation or trips to the bathroom.) *Measurement error* is a statistical term that covers variation from whatever source. It would be better to talk about measurement *variation* or typical *variation,* rather than *error,* but I might have trouble convincing my colleagues...

I've explained the notion of typical error as variation for one subject, but in practice you calculate the average typical error for all the subjects. You can calculate it even when there are only two tests, and even when there is a shift in the mean between those tests. See the page on calculations for reliability and the reliability spreadsheet for details. For the weight data shown in the figure, the typical error is 1.4 kg.

You can derive a closely related measure of error simply by calculating each subject's standard deviation, then averaging them. The result is the **total error of measurement**, which is a form of typical error contaminated by change in the mean. On its own the total error is not a good measure of reliability, because you don't know how much of the total error is due to change in the mean and how much is due to typical error. Some researchers and anthropometrists have used this measure, nevertheless.

An important form of the typical error is the **coefficient of variation**: the typical error expressed as a percent of the subject's mean score. For the above data, the coefficient of variation is 2.0%. The coefficient of variation is particularly useful for representing the reliability of athletic events or performance tests. For most events and tests, the coefficient of variation is between 1% and 5%, depending on things like the nature of the event or test, the time between tests, and the experience of the athlete. For example, if the coefficient of variation for a runner performing a 10,000-m time trial is 2.0%, a runner who does the test in 30 minutes has a typical variation from test to test of 0.6 minutes.

If you use the coefficient of variation rather than the raw typical error, it makes sense to represent any changes in the mean between tests as **percent changes**. In our example of body weights, the shift in the mean of -0.9 kg is -1.2%. The percent shifts, and the coefficient of variation, can be derived by analysis of the log-transformed variable. See the page on calculations for reliability for details.

All standard methods for calculating the typical error are based on the assumption that the typical error has the same average magnitude for every subject. If the typical error varies between subjects, statisticians say the data display heteroscedasticity, or **non-uniform error**. In this situation the analysis provides you with some kind of average typical error that will be too high for some subjects and too low for others. To get rid of heteroscedasticity, you have to either do separate analyses for subgroups of subjects with similar typical errors (e.g., males and females), or find a way to transform the variable to make the typical error for the transformed variable uniform. Log transformation often makes the error uniform when larger values of the original variable have more error. You should check for non-uniform error whenever you calculate reliability statistics. I explain how on the calculations page.

Another form of within-subject variation promoted by some statisticians is **reliability limits of agreement**, which represent the 95% likely range for the difference between a subject's scores in two tests. For example, if the limits of agreement for a measurement of weight are ±2.5 kg, there's a 95% chance that the difference between a subject's scores for two weighings will be within -2.5 kg and +2.5 kg (after any learning effect or other systematic change in the mean on retest has been taken out of the picture). Equivalently, if you reweighed a large number of subjects, 95% of them would have difference scores within -2.5 kg and +2.5 kg. The range defined by the limits of agreement is regarded as a kind of **reference range** for changes between pairs of measurements: in our example, any change between -2.5 and +2.5 kg is deemed to be normal variation; anything else is unusual enough to be indicative that a real change has occurred.

For a normally distributed variable, the limits of agreement are ±2.77 times the typical error. The 2.77 comes from the standard deviation of the difference score (which is root2 times the typical error) multiplied by 1.96 (which includes 95% of observations of the difference score). So even though they are very different in definition, the fact that the typical error and limits of agreement are proportional makes their properties similar. Which is the better measure of reliability? I prefer typical error, because limits of agreement are harder to understand, they are harder to apply to the error of a single measurement, they are too large as a reference range for making a decision about a change in a subject's measurements (more about this issue on the next page), and they have to be converted into a typical error for most statistical calculations.

Retest Correlation

When you plot test and retest values, it's obvious that the closer the values are to a straight line, the higher the reliability. A **retest correlation** is therefore one way to quantify reliability: a correlation of 1.00 represents perfect agreement between tests, whereas 0.00 represents no agreement whatever. In our example the correlation is 0.95, which represents very high reliability.

OK, do we need the correlation coefficient? Why can't we just use the typical error? Hmmm... Well, the two are certainly related, because a small typical error usually means a high correlation. But they also measure different things. The typical error is a pure measure of variation within each subject, whereas the correlation coefficient tells us something about the reproducibility of the rank order of subjects on retest. A high correlation means the subjects will mostly keep their same places between tests, whereas a low correlation means they will be all mixed up. Even a correlation as high as 0.95 implies some loss of order, as you can see in our example in the columns of weights. I've rank-ordered the weights in the first column (Test 1) to show you that the ordering is degraded somewhat in the second column (Test 2). It might help you understand if you think about the possibility of *negative* correlations for reliability. Such things exist and are even worse than zero, because they imply that the rank order of subjects in the first test tends to be *reversed* in the second test.

There is another important difference between typical error and retest correlation. Typical error can be estimated from a sample of subjects that is not particularly representative of the population you want to study. For example, the sample can be homogeneous relative to the population, or you can do multiple retests on just a few subjects. Either way, you can usually assume the resulting typical error applies to any subject in the population. But the retest correlation is sensitive to the nature of the sample used to estimate it. For example, if the sample is homogeneous, the correlation will be low. Or if multiple tests are performed on only a few subjects, the resulting estimate of correlation will be "noisy" (take my word for it). So whenever you interpret a correlation, remember to take into consideration the sample that was used to calculate it.

How do you calculate the retest correlation? The usual Pearson correlation coefficient is acceptable for two tests, but it overestimates the true correlation for small sample sizes (less than ~15). A better measure of the retest correlation is the **intraclass correlation coefficient**or ICC. It does not have this bias with small samples, and it also has the advantage that it can be calculated as a single correlation when you have more than two tests. In fact, the intraclass correlation is equivalent to the appropriate average of the Pearson correlations between all pairs of tests. You use analysis of variance or repeated measures to do the calculation, as detailed in reliability calculations.

Pearson and intraclass correlations are unaffected by any shift in mean on retest. So, in our example, the fact that the weights are down a bit in the second test has no effect on the correlation coefficient. And that's the way it should be. The question of any change in the mean value on retest should be kept separate.

By the way, I don't know what *intraclass* means. I presume the *intra* refers to the way typical error enters into the calculation of the correlation.

Kappa Coefficient: Reliability of Nominal Variables

Reliability can also be defined for nominal variables, to represent the consistency with which something is classified on several occasions. For example, how consistent are subjects in their choice of favorite sport, or in agreeing or disagreeing with a statement? The best measure is something called the **kappa coefficient**. It is analogous to a correlation coefficient and has the same range of values (-1 to +1). As far as I know, there is nothing analogous to typical error or change in the mean for nominal variables.

Alpha Reliability

Sport psychologists often produce a variable by effectively averaging the scores of two or more items from a multi-item questionnaire or inventory. The **alpha reliability** of the variable is derived by assuming each item represents a retest of a single item. For example, if there are five items, it's as if the five scores are the retest scores for one item. But the reliability is calculated in such a way that it represents the reliability of the *mean* of the items, not the reliability of any single item. So, for example, the alpha reliability of 10 items would be higher than that of 5 similar items.

Alpha reliability should be regarded as a measure of internal consistency of the mean of the items at the time of administration of the questionnaire. It is not test-retest reliability. For that, the questionnaire has to be administered on two or more occasions.

APPLICATIONS OF RELIABILITY

The applications are: estimating sample size for an experiment, estimating the extent of individual responses to a treatment in an experiment, assessing an individual with a single measurement or repeated measurements (with a spreadsheet for doing the calculations), andcomparing precision of measures provided by tests, items of equipment, or operators of the equipment. These applications impact on the **design of reliability studies** that give you estimates of reliability you want to use, so information about design is scattered through this page. There's also a section on sample size for reliability studies at the end. Follow this link for a Powerpoint presentation on use of reliability to assess an individual, to estimate sample size in experiments, and to estimate individual responses to a treatment.

Sample Size for an Experiment

In an experiment, you measure something on your subjects (e.g., performance), do something to them, then measure them again to see the effect of what you've done. The effect shows up as a change in mean performance between the two measurements. The more reliable your measure of performance, the more precision there will be in the change in mean performance, so the less subjects you will need. To get an estimate of the number of subjects, you have to include a value for the **smallest worthwhile change** that could result from your treatment. After all, if you see such a change in your subjects, you should be able to conclude that a change of something like that magnitude really does happen with the treatment. I've devised various formulae to work out the sample size that gives adequate precision the smallest change, using either retest correlation or typical error as the measure of reliability. When I first got into this game, I favored the retest correlation. These days I'm all for typical error. Here's a summary emphasizing the crucial role of typical error, followed by an explanation of each point:

1. Find the noise in your measure--the value of the typical error from a reliability study with individuals and a time frame similar to those of your intended study.
2. Decide on the smallest signal--the smallest clinically or practically worthwhile change in the measure for your study group.
3. If the noise is less than the smallest signal, you can use the measure to make precise estimates of any experimental effects with a single test and retest and a sample of modest size.
4. If the noise is greater than the smallest signal, the measure will provide acceptable precision for effects smaller than the noise only with more testing (more subjects, or more pre and post tests).

The important point in Point 1 is to make sure the **conditions** and **subjects** in the reliability study are similar to those in the intended experiment. In particular, the **time between consecutive pairs of trials** in the reliability study should be similar to the time between the pre and post tests in the experiment. For example, if you intend to look at the effects of a two-month nutritional intervention on body fat, the reliability of body fat measurements with two months between measurements will give you a more realistic idea of sample size than the higher reliability you are likely to see if only two hours separate the measurements. With two hours between measurements, the typical error is likely to arise only from **technological error**: error arising from the apparatus or the operator of the apparatus. With two months between measurements, some subjects will get fatter and some will get thinner, so the typical error will include also **biological "error"**: true variation within the subjects. It's against this background of biological variation and technological error that you measure changes in body fat resulting from your intervention.

Researchers don't devote enough attention to Point 2. I go into this point in detail on several pages: a scale of magnitudes, and formulae for sample size. In summary, for most studies of health, injury, and fitness of normal folks, the smallest effect is 0.2 of the between-subject standard deviation. For studies of athletic performance, the smallest effect is 0.3-0.5 of the typical variation (standard deviation) that a top athlete displays from competition to competition.

If only all our measures were as good as those in Point 3! The "modest" sample size based on adequate confidence limits is ~$8s^2/d^2$ for a crossover, or 4x as many when there is a control group, where s is the noise (typical error or within-subject standard deviation) and d is the signal (smallest worthwhile change). So, when the noise in the measure is negligible compared with the smallest effect (s<<d), you can in theory do the experiment with one subject in a crossover and two in a controlled trial (one each in the treatment and control groups). But you should still use ~10 subjects, to be confident that the subjects in your study are representative of a wider population.

Most often the noise is greater than the smallest signal, as in Point 4. The noise comes either from technological error, or from random real changes in the subjects over the time frame of the study, or from individual responses to the treatment. Whatever the source of the noise, acceptable precision for the smallest effect demands either a large sample size (>>8 in a crossover; >>32 in a controlled trial) or several pre and post tests on each subject. Extra pre tests and post tests effectively reduce the noise in the measure, because you analyze the change between the average of the pre tests and the average of the post tests. It's the only option when your pool of subjects is limited.

It's certainly a good idea to do a reliability study before an experiment, either to estimate sample size or to make sure you've got your techniques right. But if you are reasonably confident about the techniques, I advocate getting stuck straight into the experiment. As I explain insample size on the fly, if your treatment turns out to have a big effect, you needn't have done all the extra testing to get adequate precision.

Individual Responses to a Treatment

When the response to an experimental treatment differs between subjects, we say that there are individual responses to the treatment, or that there are individual differences in the response. For example, a treatment might increase the power output of athletes by a mean of 3%, but the variation in the true enhancement between individual athletes might be a standard deviation of 2.5%. In this example, most athletes would show positive responses to the treatment, some athletes would show little or no response, and some would even respond negatively. Note that this figure of 2.5% is not simply the standard deviation of the difference scores, which would include variation due to typical error. When I refer to individual responses, I mean variation in the true effect free of typical error. Although the primary aim in an experiment is to estimate the mean enhancement, it is obviously important to know whether the individual responses are substantial. Analysis of reliability offers one approach to this problem.

When individual responses are present, subjects show a greater variability in the post-pre difference score. Analysis of the experimental group as a reliability study therefore yields an estimate of the typical error inflated by individual responses. Comparison of this inflated typical error with the typical error of the control group or with the typical error from a reliability study allows you to estimate the magnitude of the individual responses as a standard deviation (2.5% in the above example). If the experiment consists of a pre-test, an intervention, and a post-test, the estimate is readily derived from basic statistical principles as root($2s^2_{expt}$ - $2s^2$), where s_{expt} is the inflated typical error in the experimental group, and s is the typical error in the control group or in a reliability study. For example, if the typical error in the experimental group is 2%, and the typical error in the control group or in a reliability study is 1%, the standard deviation of the individual responses is 2.5% (= root6).

If you use the typical error from a reliability study to estimate the individual responses in your experiment, make sure the reliability study has a time frame and subjects similar to those in your experiment. And if your experiment is a crossover, there is no control group, so you*have* to use the typical error from a reliability study. Alternatively, use a complex crossover in which your subjects do several tests for each of the treatments.

You can also used mixed modeling to estimate individual responses. It's awfully complicated, but the extra effort is worth it, because you also get confidence limits for the estimate. When individual responses are present, the obvious next step is to identify the subject characteristics that predict the individual responses. The appropriate analysis is repeated-measures analysis of covariance, with the likely subject characteristics (e.g., age, sex, fitness, genotype) as covariates. Follow this link for more.

Assessing an Individual

Whenever you take a measurement to assess someone's fitness, fatness, or other characteristics, your measurement is contaminated by "noise"--the typical error. Sometimes the noise is small enough to neglect, as in measurement of body mass with any reasonable set of scales. But if the noise is not negligible, you should be up front about your uncertainty when you report the measurement to your patient or client. There are a couple of ways to express this uncertainty. I'll explain in detail shortly. First, here are the main points. Typical error and the smallest clinically/practically worthwhile/important change have the same key roles here as they do in sample-size estimation for experimental studies (see above):

- **The typical error** of measurement is the key to making sense of a single measurement or a change in a measurement.
  - The typical error needs to come from a short-term reliability study of individuals similar to the one you are assessing.

- **For a single measurement**, the typical error really is the typical amount by which any single observed value is different from the true value.
- **For a change between two measurements**, take into account not only the typical error (the "noise") but also the smallest clinically important change (the smallest "signal").
    - If the noise is much less than the smallest signal, your measure is precise. Trust any change you see between a single test and retest.
    - If the noise is much greater than the smallest signal, your measure is too noisy to be useful. Find a less noisy test.
    - If the noise is about the same as the smallest signal, your measure is useful, but take into account the uncertainty in your measurements by using likelihoods or likely limits. You should also try to take multiple measurements and either average them to reduce the noise or look for a trend over time between the tests.

Typical Error for Assessing Individuals

When you wanted the sample size for an experiment, it was important to use an estimate of reliability from a reliability study with the same time between trials as in the experiment. But for a single measurement or a change in a measurement on an individual, you need an estimate of reliability with the minimum of biological variation. The period between measurements in the reliability study therefore needs to be brief. By *brief* I mean a period over which you wouldn't expect any real change in the variable you are assessing. For retests of skinfolds, *brief* could be an hour--anything longer and changes in the subject's posture or state of hydration might affect the measurements. For retests of physical performance, leave just enough time for all your subjects to recover from the fatigue of the previous test.

If there is a systematic change in the mean in the reliability study, do you take that into account in your subsequent assessments? In general, no, because changes in the mean in the reliability study will usually be due to changes within the subjects.

In what follows, I often refer to the true value of a subject's measurement. By *true* I mean the value free of typical error, which is the value you would get if you took hundreds of measurements on the subject and averaged them. There might still be a systematic error in this "true" measurement, but you would need to do a validity study to sort that out. That kind of systematic error is less likely to be a problem when you are interested in a change or difference between measurements, because the error will tend to disappear when you subtract one measurement from another.

A Single Measurement

You measure a gymnast and find a sum of seven skinfolds of 45.2 mm. The true value won't be *exactly* 45.2 mm, so one way to take measurement error into account is to specify a **likely range or limits for the true value**: a range within which the true value is likely to fall (for example, 42.2 to 48.2 mm). *Likely* can be anything we like. In research projects we usually opt for 95% likely, and later on I devote a whole page to the concept of confidence limits for generalizing from a sample to a population. The meaning is much the same here; the only difference is that we're talking about an individual rather than an average effect in the population. The 95% confidence or likely limits for an individual's true value have a 95% chance of enclosing that individual's true value. Or you can say the odds are 19 to 1 that the subject's true value will be within the range. You get 95% limits by multiplying the typical error by about ±2.0. Let's say your typical error is 1.5 mm for the sum of seven skinfolds on a sample of female gymnasts similar to your subject. The true value of the skinfold sum is therefore 95% likely to be within 45.2 ± 2.0x1.5, or 42.2 to 48.2 mm.

Do you tell the gymnast the 95% likely range? No, probably not. A certainty of 95% may be OK for research, but it's too much for assessing an individual. The range represented by ±1.0x the typical error--a 68% range, or 2 to 1 odds of enclosing the true value--is probably the best default way to convey your uncertainty about the true value. It's certainly the easiest to use! You just say to the gymnast, "the odds are 2 to 1 that your real skinfold thickness is between 45.2 ± 1.5, or 43.7 to 46.7 mm". If you are feeling more cautious, say instead "the odds are 9 to 1 that your real skinfold thickness is between 45.2 ± 2.5, or 42.7 to

47.7 mm." The table below summarizes the likely ranges, the odds, and the factors to multiply by your typical error. You can also use a spreadsheet for precision of a subject's true value.

Factors for generating likely (confidence) limits for the true value of a single measurement or of a difference or change in a measurement. "Likely" is defined by several values of probability or odds.

Likelihood that the limits will contain the true value		Multiply typical error by ± this factor to get the limits for…	
Probability	Odds	a single measurement	a change in a measurement
52%	1 to 1	0.71	1.00
68%	2 to 1	1.00	1.41
80%	4 to 1	1.28	1.81
90%	9 to 1	1.65	2.33
95%	19 to 1	1.96	2.77[a]

[a]This factor generates the 95% limits of agreement.

When the typical error is given as a percent, an approach similar to the above is usually accurate enough. For example, if the typical percent error is 3.0%, the 68% likely range of the true value of a single measurement is ±1.0x3.0 = ±3.0% of the observed value. If you get percent limits of 10% or more, this method become less accurate, so you have to use log transformation. But don't worry, it's all taken care of in the spreadsheet.

The factors shown in the table are values of the t statistic for the given probability. The factors get a bit larger for typical errors based on smaller sample sizes, reflecting more uncertainty about the magnitude of the typical error from smaller samples. For 20 subjects measured twice, the factors are accurate enough. If you assess subjects frequently, you should estimate the typical error of your measurement from a larger amount of retesting--otherwise you're likely to mislead *all* your subjects about the accuracy of their assessments through using an estimate of typical error that is much higher or much lower than the true typical error. See below for more on this issue.

The other way to take error into account when you assess a subject is to specify the **likelihood** (probability or odds) that the subject's true value is greater than (or less than) a reference value. This method is better for changes in a measurement between tests, but I'll illustrate it here with a simple example. If a skinfold thickness of 42 mm or more had some special significance, you could say to the gymnast "there's a 98% chance that your skinfolds are thicker than 42 mm", or "odds are 50 to 1 that your skinfolds are thicker than 42 mm". The probability and odds come straight from the first example shown on the spreadsheet.

Monitoring for a Change between Measurements

The uncertainty in a change between measurements is more than in a single measurement, because a change involves two measurements, each of which has error. But you double the variance, not the typical error, so the typical error in a change score is root2 times the typical error. The likely limits for a change in a measurement are therefore root2 times the limits for a single measurement. See the table above for the factors corresponding to the different likelihoods. I have incorporated these factors into the spreadsheet. For an example, let's measure our gymnast again, one month later. Her skinfolds were 45.2, but now they're 48.5 mm. The coach wants to know if she is really putting on fat. What do you tell the coach?

First, let's try likely limits. As before, let's assume the typical error is 1.5 mm. The easiest likely limits to calculate for a change score are the 50% limits: simply plus or minus the typical error. The observed change is 3.3 mm, so you'd say there's a 50% chance, or odds of 1:1, that the true change is between 3.3-1.5 and 3.3+1.5, or 1.8 and 4.8 mm. If we opt for a range that has odds of 4:1 of including the true change (an 80% likely range), the limits are 3.3 ±1.81x1.5, or 0.6 and 6.0 mm. And so on. Fine, but what percent limits should you use in these practical situations, and how do you use them to decide whether a real change has occurred? Rather than try to answer these hard questions, I will take you through a better method of assessing change.

The better method is based on calculating the likelihood that the true change is bigger than a reference value. For the reference value, you choose the **smallest clinically important or worthwhile change**. In the above example (observed increase of 3.3 mm, typical error of 1.5 mm), let's say that an increase in skinfolds of 2.0 mm is the smallest change worth worrying about. Obviously, the gymnast's observed change of 3.3 mm is already more than 2 mm, but how likely is it that the *true* change is more than 2 mm? From the spreadsheet, the likelihood is 73%, or odds of 3 to 1. We should also work out the likelihood that the gymnast's skinfolds have actually decreased (even though we observed an increase). The smallest worthwhile decrease would be 2.0. From the spreadsheet, the chance that the decrease has been greater than 2.0 (< -2.0) is only 1%, or odds of 1:136. Your advice to the coach? "Odds are 3 to 1 there's been a substantial increase in skinfold thickness, and there's a negligible chance that her skinfolds have decreased. You can assume she's fatter."

This example is reasonably clear cut, mainly because the typical error or noise (1.5 mm) is somewhat less than the smallest important change (2 mm). Basically, our measure is precise relative to any changes that matter, so any changes we observe with such a measure are trustworthy. But what if the noise is about equal to the smallest signal? The Powerpoint presentation has a couple of examples for an arbitrary variable with a typical error of 1.0 and a smallest important effect of 0.9. If the observed effect is 1.5, chances are 66% the true effect is clinically positive, 29% the true effect is clinically trivial, and 5% the true effect is clinically negative. It's reasonable to conclude the true effect is (probably) clinically positive. If the observed effect is a clinically trivial 0.5, the likelihood that the effect really is trivial is only 45%, whereas there's a 55% chance something really worthwhile has happened (39% positive, 16% negative). You can conclude that maybe nothing has happened, but acting on it would depend on the relative costs and benefits of taking action or doing nothing.

When the typical error is much greater than the smallest worthwhile change, we will often observe clinically worthwhile changes that are due to error of measurement rather than to any real change. The measure is therefore too noisy to be useful. The chances that real positive or negative changes have occurred (using the spreadsheet) confirm this state of affairs. For example, if the typical error is three times the smallest clinically worthwhile change, and we observe the smallest worthwhile change, the chance of a real positive change having occurred is 50%, or odds of 1:1, but the chance of a real negative change having occurred is 32%, or odds of 1:2..

Noisy measures can still be useful for characterizing worthwhile changes smaller than the noise, but we have to reduce the noise by performing multiple pre and post tests; we then either compare means of the pre and post tests or look for a trend across all the tests. On the other hand, observed changes *greater* than the typical error may still be trustworthy, if you expected them. In the present example, even a change equal to the typical error (three times the smallest worthwhile change) has likelihoods of a true positive value (68% or 2:1) or a true negative value (17% or 1:5) that would satisfy a practitioner who was expecting such a large change in the subject. But if true changes of such large magnitude are unlikely, we should be prepared to discount large observed changes as measurement error.

By basing our assessment partly on the change we think we're likely to see, we are assessing the individual in a **Bayesian** fashion. Bayesian analysis is a quantitative method for taking into account our prior belief about something, in this case the subject's true value or change in the true value. Experienced clinicians and practitioners adopt this approach qualitatively when they reject unlikely test results. Bayesian analysis ostensibly allows this kind of decision-making to be quantitative. But how can we quantify strength of a belief? For example, if we believe a change couldn't be outside ±3, where does the ±3 come from, and what likely limits define *couldn't?* 80%, 90%, 95%, 99%... ? At the moment I can't see a satisfactory answer to these questions, but whatever, I have included Bayesian adjustment for the likelihoods and likely limits in the spreadsheet. It took me so long to do, I'd hate to think the time was wasted!

Putting all these examples together with lots of deep thought, I came up with the bullet points at the start of this section on assessing an individual. Go back there now, read them again, and make sure you understand and learn them.

Some researchers have tried to use limits of agreement to make decisions about change in an individual. According to these researchers, you can trust an observed change only if it's greater than the limits of agreement. But limits of agreement are so big (2.8 typical errors) that clinically important trustworthy changes often fall within them. You end up having to ignore changes in your subjects that in some settings might be life-threatening! No, we must abandon limits of agreement as a clinical tool.

Comparing Individuals

All the above calculations for the change in a single subject's measurements also apply to making decisions about the difference between two subjects. In the above example, the second measurement of skinfold thickness (48.5 mm) could have been a measurement of skinfold thickness of another subject. Your conclusion would be that the second subject has skinfolds 3.3 mm thicker than the first, with odds of 4 to 1 that the real difference in skinfold thickness is between 0.6 and 6.0 mm. Better still, you could say that the odds of a real difference in skinfold thickness (more than 2 mm) are 3 to 1.

Spreadsheet for Assessing an Individual

In this spreadsheet I use the typical error of measurement and a subject's observed value to estimate likely limits for the subject's true value and to estimate the likelihood that the subject's true value is greater than a reference value. I do the same for the change between two observed values. I also include likelihoods and likely limits for the estimate of a true criterion value derived from a validity study. Finally, I've gone to a lot of probably pointless trouble to add Bayesian adjustments in a second spreadsheet (part of the same file).

Precision of the estimate of a subject's value: Excel latest	Help!

Comparing Reliability of Measures

Choosing between two items of laboratory equipment, choosing a good performance test or test protocol, deciding whether an anthropometrist has a reached a certain level of skill... these are all applications where you need to compare reliability of the measures produced by the equipment, the performance tests, or the anthropometrist. Recall that we have three main measures of reliability: change in the mean, typical error, and retest correlation. Which of these should you use when comparing the reliability of items of equipment, tests, anthropometrists, and so on?

Systematic changes in the mean can be an issue when comparing measures: in general, the bigger the changes between trials, the less desirable the measure. But comparing the typical errors is much more important, because the equipment, protocol, or anthropometrist that produces a measure with less typical error is providing a more accurate measure. Retest correlation contains the typical error, but the fact that it also contains the between-subject standard deviation makes the comparison of correlations either noisy (when there are different subjects in the two reliability studies) or computationally difficult (when the same subjects are in both studies). Besides, there is no point in comparing retest correlations, if you have already compared typical errors. I therefore will not deal with comparison of retest correlations.

When setting up a study to compare typical errors, keep in mind that the typical error always consists of biological variation arising from the subjects and technological variation arising from the items. The aim is usually to compare the technological variation, so try to make the biological variation as small as possible. For example, when comparing the reliability of two anthropometrists, you would get them to measure the same subjects within an hour, to avoid any substantial biological variation. Similarly, when comparing the reliability of power provided by two ergometers, use athletes as subjects, because they are usually more reliable than non-athletes.

Comparing the reliability of two items (protocols, equipment, or operators) is straightforward when *different* subjects are used to get the reliability for each item. Confidence limits for the ratio of the typical errors between corresponding trials in the two groups can be derived from an F ratio. Use Item 4 in the spreadsheet for confidence limits for this purpose. To compare changes in the mean between corresponding pairs of trials for the two measures, you will need to use an unpaired t test of the change scores. Using the *same* subjects has more power but requires analysis by an expert. (The analysis needs a mixed model, in which the equipment is a fixed effect, trial number is a fixed effect, subjects is a random effect, and a dummy random variable is introduced to account for the extra within-subject variance associated with measures on one of the items. Confidence limits for the extra variance tell you how different the typical errors could be. The model also provides an estimate of the difference in changes in the mean between the items, or you can use a paired t test.)

In the previous section I said that the 95% likely range is too conservative for assessing individuals, and I also said that it's difficult to decide on what percent range to use. The same argument and difficulty applies for comparison of typical errors in a clinical or field setting. It won't hurt to calculate, say, the 80% likely range, but I think clinicians and practitioners (and you!) will have a better chance of understanding what I'm getting at if you use likelihood that one typical error is substantially smaller or larger than the other. You compare typical errors by dividing one by the other, to get a ratio. A ratio of 1.1 or maybe 1.2 is my best guess at the minimum worthwhile difference in reliabilities, so you calculate the likelihood (as a probability or odds ratio) that one measure has a typical error at least 1.1x (or 1.2x) bigger than the other. It's all on the spreadsheet for confidence limits.

Sample Size for Reliability Studies

How many subjects and retests do you need in a reliability study? That depends on how precise you need to be with your estimate of reliability of the measure. That, in turn, depends on how reliable the measure itself turns out to be: the higher the reliability, the less precise the estimate of reliability needs to be, so the fewer the number of subjects or retests you will need. Let me explain this principle with an example of assessment of individuals. Suppose the variable in question is some measure of human performance. Suppose the smallest change in performance that matters to subjects is 2.0 units (seconds, cm, kg, %, or anything you like). If your measure of performance turns out to have a typical error of 0.2 units in a reliability study, a 50% uncertainty in this estimate (that is, a factor of 1.5, or 0.2/1.5 to 0.2x1.5, or 0.1 to 0.3 units) makes little difference to the precision of estimates of small changes in performance (~2.0 units). I mean, it doesn't matter too much if an estimate of a change in performance of 2.0 units is accurate to ±0.1 or ±0.3 units. But if the typical error turns out to be 3.0 units (that is, similar to the smallest change in performance that matters to subjects), a 50% uncertainty in the typical error (2.0 to 4.5 units) makes a big difference to the precision of estimates: 2.0 ± 2.0 units isn't very good, but it's a lot better than 2.0 ± 4.5 units. In other words, when the typical error is similar in magnitude to what matters for your subjects, your uncertainty in the typical error needs to be a lot smaller than 50%, and that means more subjects in the reliability study.

Phew! Let's see what sample size we'll need for the estimate of reliability for each application of reliability. The applications are: estimating sample size for an experiment, comparing reliability of different measures, estimating individual responses in an experiment, and assessing an individual. We'll assume modest reliability: a typical error of the same order of magnitude as the smallest change that matters to subjects. We'll find that samples of 50 subjects tested three times gives reasonable precision for the estimate of the typical error. That's assuming you can combine the data for all three trials to estimate the typical error. If there is a substantial learning or practice effect on the typical error between the first and second trials, you will need another trial--four in total--so you can combine the last three.

Sample Size for Reliability Studies...
...for Estimation of Sample Size for an Experiment

When you use a value of the typical error to estimate the sample size for an experiment, uncertainty in the typical error translates into uncertainty in the sample size you will need for the experiment. Sample size for an experiment is inversely proportional to the square of the typical error, so uncertainty in the typical error balloons into much bigger uncertainty in sample size for an experiment. You can check the effect of number of subjects and retests on precision of the typical error by plugging numbers into the appropriate cells of Item 3 on thespreadsheet for confidence limits. Give the typical error a value of 1.0 then pretend you got this value from a reliability study of either 10 subjects tested twice (= 9 degrees of freedom). You will find that the 95% confidence limits for the true typical error are 0.69 to 1.83; square these and you get the uncertainty in sample size as factors of 0.47 to 3.33. In other words, if you predicted a sample size of, say, 40 subjects in the experiment on the basis of a typical error of 1.0, what you might really need is anything from 19 to 133. Well, that's far too wide a range! Let's try a reliability study with 50 subjects tested three times. The range in sample size becomes 31 to 54, which is still quite a lot of uncertainty, but I guess it's OK.

This calculation is based on 95% limits of uncertainty for the typical error, which may be a bit too conservative for the likely limits of the sample size in the experiment. If instead we use 67% likely limits, we end up with something more like the typical variation in the estimate of sample size based on the reliability study. For a reliability study of 10 subjects tested twice, the typical variation in our estimate of sample size would be, for example, 28 to 72. Still too wide. Test them three times and you get 30 to 59. That's better, but the required sample size could easily be outside these 67% limits.

So what's my advice? If you have the time, money, and subjects for a large reliability study, go for it. Otherwise you're better off devoting your resources to the experiment by using sample size on the fly: stop testing subjects when you have adequate precision for the effect.

Sample Size for Reliability Studies...
...for Comparing Reliability of Measures

When you want to compare the reliability of two measures, the worst-case scenario is that you observe similar reliabilities for the two measures. (You might see why this is worst-case in a minute.) In this scenario, you want to conclude that there are no substantial differences between the measures. The easiest way to compare typical errors is compute their ratio and its confidence limits. Therefore, you will be able to conclude there is no substantial difference if the upper limit of the ratio is only a little greater than 1.00 and the lower limit is only a little less than 1.00. Let's generate some confidence limits for the ratio using Item 4 (ratio of standard deviations) in the spreadsheet for confidence limits. Make the two typical errors the same (e.g. 3.0), and pretend each has come from a study with 100 degrees of freedom (51 subjects, 3 trials). You'll see that the 95% confidence limits for the ratio of the typical errors are 0.82 to 1.22. In other words, the true values of the typical errors could be about 20% different from each other. That amount of uncertainty is marginal, in my view, but again, 95% confidence limits are probably too stringent in a real-life situation where you are choosing between two items of equipment. The 80% confidence limits for the ratio are 0.88 to 1.14, which make me feel more comfortable about concluding that there is no real difference in the reliability of the two measures. I feel even more comfortable looking at the likelihood the the true value of the ratio is greater than 1.2: it's only 3%, or odds of 1 in 28. There is no substantial difference in the reliability of these two measures, if by "substantial" we mean different by a factor of 1.2 or more.

Things aren't so bad when you observe a big difference between the typical errors of the measures, because you will need less subjects to conclude that one really is substantially worse (larger) than the other. Try it for yourself with the spreadsheet: make the observed typical errors 2.0 and 3.0, give them both only 20 degrees of freedom, make the reference ratio 1.15, say, then look at the likelihood that one typical error is substantially greater than the other: 88%, or odds of 7:1. Not much doubt about it--they're different!

Finally, if you can use the same subjects for both reliability studies, you're bound to get better precision for the ratio and therefore a reduction in sample size required to make firm conclusions about the relative magnitudes of the typical errors. Sorry, I haven't worked out how big the reduction is yet. You can't do it with the spreadsheet--you have to use mixed modeling or bootstrapping.

Sample Size for Reliability Studies...
...for Estimating Individual Responses

Estimation of individual responses to a treatment boils down to a comparison of the typical errors of two groups (the treatment and control groups), so the sample size must be the same as for a comparison of the reliability of two measures..

Sample Size for Reliability Studies...

...for Assessing an Individual

At first glance it appears you can use as few as 20 subjects and two trials to estimate a typical error without substantially degrading the precision of an individual assessment. Check the spreadsheet for precision of a subject's true value to see what I mean. In Item 1, put in an observed value of 50, a typical error of 2.0 from two trials, and compare the likely limits for the subject's true value when the typical error is based on 20 subjects vs 2000 subjects. With 20 subjects the 80% likely limits for the subject's true value are 47.3 to 52.7, or 50 ± 2.7; for 2000 subjects the limits are 47.4 to 52.6, or 50 ± 2.6. In other words, there's a negligible increase in the likely limits (= loss of precision) for the smaller sample size. But wait a minute... the typical error based on a sample of 20 subjects and two trials is really noisy. Check the spreadsheet for confidence limits and you'll see, for example, that a typical error of 2.0 has 95% likely limits of 1.5 to 2.9. That's a big range in precision. What gives?

Well, 20 subjects and two tests definitely give you almost as much accuracy as a zillion subjects and tests, and that's fair enough if you are assessing only one individual. If another clinician tested 20 subjects twice, then assessed another individual, it would be the same story. But there's likely to be a big difference between your typical errors; for example, yours might be 2.5, and the other clinician's might be 1.7. Your assessments of, say, 80% likely limits based on a typical error of 2.5 would really be ~90% likely limits, while the other person's 80% likely limits based on a typical error of 1.7 would be ~70% likely limits. You're both giving misleading assessments, and so would many other clinicians who tested only 20 subjects twice. Yet averaged over all clinicians and all subjects, the true values of 80% of subjects would be within the likely limits that each clinician tells each subject. The trouble is that your assessments will be *consistently* misleading, if you are unlucky enough to get a typical error of 2.5 or 1.7 with your batch of 20 subjects. A typical error based on 50 subjects and three tests would usually be in the range of 1.8 to 2.2, and if you used 2.2 in your assessments, your 80% likely limits would be less than 85% limits in reality, which seems OK to me. But I'm still thinking about it...

CALCULATIONS FOR RELIABILITY

Make sure you understand the page on reliability before tackling this page. I explain here how to analyze data for two trials using simple but effective methods. To combine three or more trials you need more sophisticated procedures, such as analysis of variance or modeling variances. I go into heaps of detail about checking for non-uniform error in your data, and I have a few words on biased estimates of reliability. Finally, you can download a spreadsheet for calculating reliability between consecutive pairs of trials, complete with raw and percent estimates and confidence limits for typical error, change in mean, and retest correlation. The spreadsheet has data adapted from real measurements of skinfold thickness of athletes.

Two Trials

Analyzing two trials is straightforward. All the necessary calculations are included in the spreadsheet for reliability. When you have three or more trials, I strongly recommend that you first do separate analyses for consecutive pairs of trials (Trial1 with Trial2, Trial2 with Trial3, Trial3 with Trial4, etc.). That way you will see if there are any substantial differences in the typical error or change in the mean between pairs of trials. Such differences are indicative of learning or practice effects. If there is no substantial change in the typical error between three or more consecutive trials, analyze those trials all together to get greater precision for your estimates of reliability.

Typical Error

The values of the change score or difference score for each subject yield the typical error. Simply divide the standard deviation of the difference score by root2. For example, if the difference scores are 5, -2, 6, 0, and

-3, the standard deviation of these scores is 4.1, so the typical error is 4.1/root2 = 2.9. This method for calculating the typical error follows from the fact that the variance of the difference score (s^2_{diff}) is equal to the sum of the variances representing the typical error (s) in each trial: $s^2_{diff} = s^2 + s^2$, so $s = s_{diff}$/root2.

To derive this within-subject variation as a coefficient of variation (CV), log-transform your variable, then do the same calculations as above. The CV is derived from the typical error (s) of the log-transformed variable via the following formula:

$$CV = 100(e^s - 1),$$

which simplifies to 100s for s<0.05 (that is, CVs of less than 5%). You will also meet this formula on the page about log-transformation, where I describe how to represent the standard deviation of a variable that need log transformation to make it normally distributed. As I describe on that page, I find it easier to interpret the standard deviation and shifts in the mean if I make the log transformation 100x the log of the variable. That way the typical error and shifts in the mean are already approximately percents. To convert them to exact percents, the formula becomes $100(e^{s/100} - 1)$.

We sometimes show the typical error with a ± sign in front of it, to indicate that a subject's observed value varies by typically ± the typical error whenever we measure it. For example, the typical error in a monthly measurement of body mass might be ±1.5 kg. When we express the typical error as a CV, we can also think of it as ±2.1% (if the subject weighed 70 kg), but strictly speaking it's more appropriate to show the variation as ×/÷1.021. In other words, from month to month the body mass is typically high by a factor of 1.021 or low by a factor of 1/1.021. These factors come from the assumption that the log-transformed weight rather than the weight itself is normally distributed. Now, ×1.021 is the same as 1 + 0.021, and 1/1.021 is almost exactly 1 - 0.021, so it's OK to show the CV as ±2.1%. But when the CV is bigger than 5% or so, the use of the minus sign gets more inaccurate. For example, if the CV is 35%, the value of the variable varies typically by a factor of 1.35 to 1/1.35, or 1.35 to 0.74, or 1 + 0.35 to 1 - 0.26, which is certainly **not** the same as 1 + 35% to 1 - 35%. You can still write ±35%, but be aware that the implied typical variation in the observed value is ×/÷1.35.

Changes in the Mean

A simple way to get these is to do paired t tests between the pairs of trials. Do it on the log-transformed variable and you'll get approximate percent changes in the mean between trials. Use the same formulae as for the CV to turn these into exact percent changes.

Retest Correlation

A simple Pearson correlation is near enough. If the variable is closer to normally distributed after log transformation, you should use the correlation derived from the log-transformed variable. Alternatively calculate the intraclass correlation coefficient from the formula ICC = $(SD^2 - sd^2)/SD^2$, where SD is the between-subject standard deviation and sd is the within-subject standard deviation (the typical or standard error of measurement). These standard devations can come from different subjects, if you want to estimate the retest correlation by combining the error in one study applied to a different group. The spreadsheet for the ICC has this formula and confidence limits for the ICC.

Note that the above relationship allows you to calculate the typical error from a retest correlation, when you also know the between-subject standard deviation: sd = SD·root(1 - r). Strictly speaking the r should be the intraclass correlation, but there is so little difference between the Pearson and the ICC, even for as few as 10 subjects, that it doesn't matter.

Three or More Trials

I deal here with the procedures for getting the average reliability across three or more trials. The simplest and possibly the most practical or realistic procedure is simply to average the reliability for the consecutive pairs of trials. Well, it's not that simple to average the standard deviations representing the typical error, because you have to weight their squares by the degrees of freedom, then take the square root. I've done it for you in the reliability spreadsheet. The resulting average is the typical error you would expect for the average time between consecutive pairs of trials, and you usually make that the same (e.g., 1 week) when you design the reliability study.

There are more complicated procedures for getting the average reliability, using ANOVA or repeated-measures analyses. There is no spreadsheet for these procedures. I'll describe the usual approach, which is based on the assumption that there is a single random error of measurement that is the same for every subject for every trial. That is, whenever you take a measurement, a random number comes out of a hat and gets added to the true value. The numbers in the hat have a mean of zero, and their standard deviation is the error of measurement that you want to estimate. Or to put it another way, no matter which pairs of trials you select for analysis, either consecutive (e.g., 2+3) or otherwise (e.g., 1+4), you would expect to get the same error of measurement. This assumption may not be particularly realistic, if, for example, you did 5 trials each one week apart: the error of measurement between the first and last trial is likely to be greater than between trials closer together. If you estimate the error assuming it is the same, you will get something that is too large for trials close together and too small for trials further apart.

To understand this section properly, read the pages on statistical modeling. In a reliability study or analysis, you are asking this question: how well does the identity of a subject predict the value of the dependent variable, when you take into account any shift in the mean between tests? (If the variable is reliable, the value of the variable is predicted well from subject to subject. If the variable is unreliable, it isn't much help to know who the subject is.) So the model is simply:

dependent variable <= subject test

In other words, it's a two-way analysis of variance (ANOVA) of your variable with *subject* and *test* as the two effects. Do NOT include the interaction term in the model! The analysis is not done as a repeated-measures ANOVA, because the subject term is included in the model explicitly. Experts with the Statistical Analysis System *can* use a repeated-measures approach with mixed modeling, as described below in modeling variances.

Typical Error
The root mean-square error (RMSE) in the ANOVA is a standard deviation that represents the within-subject variation from test to test, averaged over all subjects. If your stats package doesn't provide confidence limits for it, use the spreadsheet for confidence limits.

If you use a one-way ANOVA in which the only effect is *subject,* the RMSE will be contaminated by any change in the mean between trials. (In a two-way ANOVA, the *test* effect takes out any change in the mean.) The resulting RMSE represents the total error of measurement. You can also derive the total error by calculating each subject's standard deviation, squaring them, averaging them over all subjects, then taking the square root. This procedure works for two trials, too. I don't recommend total error as a measure of reliability, because you don't know how much of the total error is due to change in the mean and how much is due to typical error.

Changes in the Mean
Your stats program should be able to give you confidence limits or p values for each consecutive pairwise comparison of means. If it gives you only the p values, convert these to confidence limits using the spreadsheet for confidence limits.

Shifts in the mean and typical error as percents are derived from analysis of the log-transformed variable. See the previous section for the formula.

Retest Correlation
Scrutinize the output from the ANOVA and find something called the F value for the subject term. The retest correlation, calculated as an intraclass correlation coefficient (ICC), is derived from this F value:

ICC = (F - 1)/(F + k - 1),

where k = (number of observations - number of tests)/(number of subjects - 1). In the case of no missing values, number of observations = (number of tests)·(number of subjects), so k is simply the number of tests. For example, a reliability study of gymnastic skill consisted of 3 tests on 10 subjects. There were 28 observations instead of 30, because two athletes missed a test each, so k = (28-3)/(10-1) = 2.78. The F ratio for subjects was 56. Reliability was therefore (56-1)/(56+2.78-1) = 0.95.

I used to have this formula in the spreadsheet for confidence limits, then I removed it for many years, thinking that people don't need it. Recently (2009) I've started expressing predictability of competitive athletic performance as an ICC, and I found I do need it and related formulae. So they're back, in their own spreadsheet for the ICC.

The ICC formula came from Bartko (1966), although he used sums of squares rather than F values. His formula for k when there are missing values is complex and appears not to be the same as the one I have given above. The *random* statement in Proc Glm of the Statistical Analysis System generates k, and I have found by trial and error that my formula gives the exact value.

Your stats program will give you p value for the *subject* term and the *test* term. The p value for *subject* is not much use. It tells you whether the ICC is statistically significantly different from zero, but that's usually irrelevant. The ICC is usually at 0.7-0.9 or more, so there's no way it could be zero. More important are the confidence limits for the ICC and for the typical error. The p value for *test* addresses the issue of *overall* differences between the means of the tests, but with more than two tests you should pay more attention to the significance of consecutive pairwise differences (to see where any learning effects fade out). I'd prefer you to show the confidence intervals for the differences, rather than the p values. If your stats program doesn't give confidence intervals, use the spreadsheet for confidence limits for the typical error, and the spreadsheet for the ICC for confidence limits for the ICC. By the way, stats programs don't provide a p value for the typical error, because there's no way it can be zero.

The typical error or root mean square error (RMSE) from one group of subjects can be combined with the between-subject standard deviation (SD) of a second group to give the reliability correlation for the second group. This approach is handy if you do repeated testing on only a few subjects to get the within-subject variation, but you want to see how that translates into a reliability correlation when you combine it with the SD from single tests on a lot more subjects. You simply assume that the within-subject variation is the same for both groups, then apply the formula that defines the reliability correlation:

ICC = (SD^2 - typical error^2)/SD^2.

(This formula can be derived simply enough from the definition of correlation as the covariance of two variables divided by the product of their standard deviations.) The spreadsheet for the ICC deals with this scenario, too.

For non-normal variables, your analyses in the main study are likely to be non-parametric. So it makes sense to derive a non-parametric reliability. Just do the ANOVA on the rank-transformed variable. The within-subject variation is hard to interpret, though.

Attention sport psychologists: if the repeated "tests" are simply the items of an inventory, the alpha reliability of the items (i.e., the consistency of the mean of the items) is (F - 1)/F.

For nominal variables (variables with categories as values rather than numbers), the equivalent of the ICC is the kappa coefficient. Your stats program should offer this option in the output for the procedure that does chi-squared tests or contingency tables.

Modeling Variances for Reliability

A reliability studiy is just an experiment without an intervention, so any method for analyzing an experiment will work for a reliability study. Modeling variances is one such method. In SAS, you model variances with Proc Mixed, using the model for simple repeated measures. The procedure produces the within variance and its confidence limits. It also produces the retest correlation as an intraclass correlation, but to get its confidence limits you'll have to use the spreadsheet for confidence limits. I don't know whether the other major stats programs have procedures like Proc Mixed for modeling variances.

Non-Uniform Error of Measurement

I've already introduced the concept of non-uniform error (heteroscedasticity) to describe the situation when some subjects are more reliable than others. You should always check whether your typical error is non-uniform, but you will need plenty of subjects to make any definite conclusions. One good way to check is to calculate the typical error for different subgroups. Often the typical error varies with the magnitude of the variable, so try splitting your subjects into a top half and a bottom half and analyzing them separately. For the data on skinfold thickness in the spreadsheet for reliability, the typical errors of the bottom and top halves are 0.48 and 1.03 mm (not shown on the spreadsheet--you'll have to do it yourself). It certainly looks like subjects with a bigger sum of skinfolds have more variability, but with only 10 subjects in each half, there's a lot of uncertainty about just how big the difference really is.

Depending on the sample and the variable, you should also analyze the typical errors for subgroups differing in sex, athletic status, age group, and so on. You sometimes find that any differences in reliability between such groups arise mainly from differences in the magnitude of the variable; for example, if log transformation removes any non-uniformity of error related to the magnitude of the variable, you will probably find that the subgroups for sex, age or whatever now have the same percent typical errors.

A more statistical approach to checking for differences in the typical error between subjects is to look at the scatter of points in the plot of the two trials. The scatter at right angles to the line of identity should be the same wherever you are on the line (and for whatever subgroups). If there is more scatter at one end, the subjects at that end have a bigger typical error. It's often difficult to tell whether the scatter is uniform on the plot, especially when reliability is high, because the points are all too close to the line. An easier way is to plot the change score against the average of the two trials for each subject. I have provided such a plot on the spreadsheet. (It's not obvious even on this plot that the subjects with bigger skinfolds have more variability. Again, more subjects are needed.) I've also provided a complete analysis for the log-transformed variable. A uniform scatter of the change scores after log-transformation implies that the coefficient of variation (CV, or percent typical error) is the same for all subjects, and the analysis of the log-transformed variable provides the best estimate. Look at the plots of the difference scores and you will see that the scatter is perhaps a little more uniform after log transformation. When I analyzed the bottom and top halves of the log-transformed variable, I got CVs of 1.1% and 2.0%. These CVs are a little closer together than their corresponding raw typical errors, so it would be better to represent the mean typical error for the full sample as 1.7% rather than 0.83 mm. But really, you need more subjects...

When you analyze three or more trials using ANOVA or repeated measures, the equivalent of the difference scores is the residuals in the analysis, and the equivalent of the average of the two trials is the predicted values. The standard deviation of the residuals is the typical error, so if the residuals are bigger for some subjects (some predicteds), the typical error is bigger for those subjects. Try to coax your stats program into producing a plot of the residuals vs the predicteds. Click for more information about residuals and predicteds, and aboutbad residuals (heteroscedasticity).

Biased Estimates of Reliability

Some statisticians think mistakenly that reliability should be calculated with a one-way ANOVA, in which you leave out the term for the identity of the tests. The trouble is, a one-way ANOVA produces an estimate of retest correlation that is biased low for small samples, and it is even lower if the means differ between trials. The within-subject variation from the analysis is the same as the total error, which will be larger than the typical error when there is any systematic change in the mean between trials. Neither of these estimates of reliability should be used to estimate sample sizes for longitudinal studies.

The Pearson correlation coefficient is also a biased estimate of retest correlation: it is biased high for small sample sizes. For example, with only two subjects you always get a correlation of 1! For samples of 15 or more subjects, the ICC and the Pearson do not usually differ in the first two decimal places.

I used to think that limits of agreement were biased high for small samples, because I thought they were defined as the 95% confidence limits for a subject's change between trials. (The formula for confidence limits includes the t statistic, which is affected by sample size in such a way that the limits defined in this way would be biased high for small samples.) But apparently Bland and Altman, the progenitors of limits of agreement, did not define limits of agreement as 95% confidence limits; instead they defined them as a "reference range", generated by multiplying the typical error by 2.77, regardless of the size of the sample that is used to estimate the typical error. In other words, the limits of agreement represent 95% confidence limits for a subject's true change only if the typical error is derived from a large sample. With this definition, the limits of agreement are only as biased as the typical error.

Surprisingly, even the typical error is biased! Yes, the square of the typical error (a variance) is unbiased, so the square root of the variance must be biased low for small samples. In practical terms, typical errors derived from samples of, say, 10 subjects tested twice will look a bit smaller on average than typical errors derived from hundreds of subjects or many retests. This bias in the typical error does not affect any statistical computations involving the typical error.

Spreadsheet for Calculating Reliability

The spreadsheet computes the following measures of reliability between consecutive pairs of trials: change in the mean, typical error, retest correlation (Pearson and intraclass), total error, and limits of agreement. Data in the spreadsheet are from a study of the reliability of the sum of seven skinfolds for a group of athletes.

The spreadsheet now includes averages for the consecutive pairwise estimates of error, with confidence limits. This approach to combining more than two trials is probably more appropriate than the usual analysis of variance or repeated-measures analysis that I describeabove (and which, in any case, I can't set up easily on a spreadsheet). I have also included averages of trial means and standard deviations, in case you want to report these as characteristics of your subjects.

Pairwise reliability analyses: Excel spreadsheet

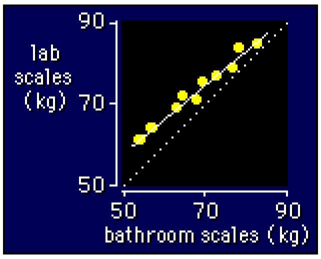
See also the spreadsheet for the ICC, when you have between- and within-subject standard deviations and you want the ICC and its confidence limits, or you have the ICC and you want its confidence limits, or you have an F ratio from an ANOVA and you want the ICC and its confidence limits.

MEASURES OF VALIDITY

Update Oct 2011: view this slideshow on validity and reliability for an overview of the important principles. If you haven't read the general introduction to precision of measurement, do so now. A variable or measure is valid if its values are close to the true values of the thing that the variable or measure represents. In plain language, it's valid if it measures what it's supposed to. This concept of validity is known as **concurrent validity**, and it's the only one I will deal with here.

Measures of validity are similar to measures of reliability. With reliability, you compare one measurement of a variable on a group of subjects with another measurement of the same variable on the same subjects. With validity, you also compare two measurements on the same subjects. The first measurement is for the variable you are interested in, which is usually some **practical variable** or measure. The second measurement is for a variable that gives values as close as you can get to the true values of whatever you are trying to measure. We call this variable the **criterion variable** or measure. The three main measures of reliability--change in the mean, within-subject variation, and retest correlation--are adapted to represent validity. I call them the estimation equation, typical error of the estimate, and validity correlation. There is also a measure of limits of agreement. I have a little to say on validity of nominal variables (kappa coefficient, sensitivity, and specificity), and I finish this page with a spreadsheet for calculating validity.

You will find that correlation has a more prominent role in validity than in reliability. Most applications of validity involve differences between subjects, so the between-subject standard deviation stays in the analysis and can be expressed as part of a correlation. In contrast, most applications of reliability involve changes within subjects; when you compute changes, the between-subject variation disappears, and with it goes correlation.



Let's explore these concepts with an example similar to the one I used for reliability. Imagine you are a roving applied sport scientist, and you want to measure the weight of athletes quickly and easily with portable bathroom scales. You check out the validity of the bathroom scales by measuring a sample of athletes with the scales and with certified laboratory scales, as shown in the figure. I've shown only 10 points, but in practice you'd use probably 20 or so, depending on how good the scales were.

Note that I have assigned the observed value of the variable to the X axis, and the true value to the Y axis. That's because you want to use the observed value to predict the true value, so you must make the observed value the independent variable and the true value the dependent variable. It's wrong to put them the other way around, even though you might think that the observed value is dependent on the true value.

Estimation Equation

The dotted line in the figure represents perfect validity: identical weights on the bathroom and lab scales. The solid line is the best straight line through the observed weights. Notice how the lighter weights are further away from the true value. That trend away from the true value is represented by the **estimation** or **calibration equation.** Any deviation away from the dotted line represents a **systematic offset**.

Notice also that a straight line is a pretty good way to relate the observed value to the true value. You'd be justified in fitting a straight line to these data and using it to predict the true weight (lab scales) from the observed weight (bathroom scales) for any athletes on your travels.

By the way, you won't always get a straight line when you plot true values against observed values. When you get a curve, fit a curve! You can fit polynomials or more general non-linear models. You know you have the right curve when your points are scattered fairly evenly around it. Use the equation of the curve to predict the true values from the observed values.

You can also use a practical measure that looks nothing like the criterion measure. For example, if you are interested in predicting body fat from skinfold thickness, the practical measure would be skinfold thickness (in mm) measured with calipers, and the criterion measure could be body fat (as percent of body mass) measured with dual-emission X-ray absorptiometry (DEXA). You then find the best equation to link these two measures for your subjects.

There are sometimes substantial differences in the estimation equation for different groups of subjects. For example, you'd probably find substantially different equations linking skinfold thickness to body fat for subjects differing in such characteristics as sex, age, race, and fitness. Sure, you can derive separate equations for separate subgroups, but it's usually better to account for the effect of subject characteristics by including them as independent variables in the estimation equation. For that you need the technique of multiple linear regression, or you could even go to exotic multiple non-linear models. A stepwise or similar approach will allow you to select only those characteristics that produce substantial improvements in the estimation of the criterion.

Typical Error of the Estimate

Notice how the points are scattered about the line. This scatter means that any time you use the line to estimate an athlete's true weight from the bathroom scales, there will be an error. The magnitude of the error, expressed as a standard deviation, is the **typical error of the estimate**: it's the typical error in your estimate of an athlete's true weight. We've met this term already as the standard error of the estimate. I used to call it the **standard deviation of the estimate**. Now I prefer typical error, because it *is* the typical amount by which the estimate is wrong for any given subject. In the above example, the typical error of the estimate is 0.5 kg.

The typical error of the estimate is usually in the output of standard statistical analyses when you fit a straight line to data. If you fit a curve, the output of your stats program might call it the root mean-square error or the residual variation. Some stats programs provide it in the squared form, in which case you will have to take the square root. Your program almost certainly won't give you confidence limits for the typical error, but you should nevertheless calculate them and publish them. See the spreadsheet for confidence limits.

In the sections on reliability, I explained that the within-subject variation can be calculated as a percent variation--the coefficient of variation--by analyzing the log of the variable. The same applies here: take logs of your true and observed values, fit a straight line or curve, then convert the typical error of the estimate to a percent variation using the same formula as for reliability. See reliability calculations for the formula. Analysis of the logs is included in the validity spreadsheet. In the above example the standard deviation is 0.7%. Expressing the standard deviation as a percent is particularly appropriate when the scatter about the line or curve gets bigger for bigger untransformed values of the estimate. Taking logs usually makes the scatter uniform. See log transformation for more.

New-Prediction Error

If a validity study has a small sample size (<50 subjects), the typical error of the estimate is accurate only for the subjects *in the validity study.* When you use the equation to predict a *new* subject's criterion value, the error in the new estimate--let's call it the **new-prediction error**--is larger than the original typical error of the estimate. Why? Because the calibration equation (intercept and slope) varies from sample to sample, and the variation is negligible only for large samples. The variation in the calibration equation for small samples therefore introduces some extra uncertainty into any prediction for a new subject, so up goes the error. The uncertainty in the intercept contributes a constant additional amount of error for any predicted value, but the error in the slope produces a bigger error as you move away from the middle of the data. Your stats program automatically includes these extra errors when you request confidence limits for predicted values. You will find that the confidence limits get further away from the line as you move out from the middle of the data. The effect is noticeable only for small samples or only for predicted values that are far beyond the data.

So, exactly how big is the error in a predicted value based on a validity or calibration study with a small sample size? If you have enough information from the study, you can work out the error accurately for any predicted value. Obviously you need the slope, intercept, and the typical error of the estimate. You also need the mean and standard deviation of the practical variable (the X values). I've factored all these into the formulae for the upper and lower confidence limits of a predicted value in the spreadsheet for analysis of a validity study. I've also included them in the validity part of the spreadsheet for assessing an individual. (In that spreadsheet I've used the mean and standard deviation of the criterion or Y variable, because it's convenient to do so, and the difference is negligible.)

When you don't have access to the means or standard deviations from the validity study, you can work out an average value for the new-prediction error, on the assumption that your new subject is drawn randomly from the same population as the subjects in the validity study. One approach to calculating this error is via the **PRESS statistic**. (PRESS = Predicted REsidual Sums of Squares.) I won't explain the approach, partly because it's complicated, partly because the PRESS-derived estimate is biased high, and partly because I have better estimates. For one predictor variable, the exact formula for the new-prediction error appears to be the typical error multiplied by root(1+1/n+1/(n-3)), where n is the sample size in the validity study. I checked by simulation that this formula works. I haven't yet worked out the exact formula for more than one predictor variable, but my simulations show that the typical error multiplied by root[(n-1)/(n-m-2)] is pretty good, where m is the number of predictor variables.

Researchers in the past got quite confused about the concept of error in the prediction of new values. They used to split their validity sample into two groups, derive the estimation equation for one group, then apply it to the second group to check whether the error of the estimate was inflated substantially. That approach missed the point somehow, because the error was bound to be inflated, although they apparently didn't realize that the inflation was usually negligible. And whether or not they found substantial inflation, they should still have analyzed all the data to get the most precise estimates of validity and the calibration equation. The PRESS approach has a vestige of that data-splitting philosophy. Not that it all matters much, because most validity studies have more than 50 subjects, so the new-prediction error from these studies is practically identical to the typical error of the estimate.

A final point about the new-prediction error: don't use it to compare the validity of one measure with that of another, even when the sample sizes are small and different. Use the typical error, which is an unbiased and unbeatable measure of validity, no matter what the sample size. (Actually, it's the square of the typical error that is unbiased, but don't worry about that subtlety.)

Non-Uniform Error of the Estimate

You will recall that calculations for reliability are based on the assumption that every subject has the same typical error, and we used the term heteroscedasticity to describes any non-uniform typical error. The same assumption and terminology underlies calculations for the validity, and the approach to checking for and dealing with any non-uniformity is similar.

Start by looking at the scatter of points on the plot of the estimation equation. If every subject has the same typical error of the estimate, the scatter of points, measured in the vertical direction on the graph (parallel to the Y axis), should be the same wherever you are on the line or curve. It's difficult to tell when the points lie close to the line or curve, so you get a much better idea by examining the difference between the observed and the predicted values of the criterion for each subject. These differences are known as the residuals, and it's usual to plot the residuals against predicteds values. I have provided such a plot on the spreadsheet, or click here to see a plot from a later section of this text. If subjects in one part of the plot have a bigger scatter, they have a bigger typical error (because the standard deviation of the residuals is the typical error). The calculated typical error of the estimate then represents some kind of average variation for all the subjects, but it will be too large for some subjects and too small for others.

To get an estimate of the typical error that applies accurately to all subjects, you have to find some way to transform the criterion and practical measures to make the scatter of residuals for the transformed measures uniform. Once again, logarithmic transformation often reduces non-uniformity of the scatter in situations where there is clearly more variation about the line for larger values of the criterion. A uniform scatter of the residuals after log transformation implies that the typical error, when expressed as a percent of the criterion value, is the same for all subjects; the typical error from analysis of the log-transformed measures then gives the best estimate of its magnitude. I have included an analysis of the log-transformed measures in the spreadsheet, although for the data therein it is clear that the scatter of residuals is more uniform for the raw measures than for the log-transformed measures.

If you fit a curve rather than a straight line to your data, the standard deviation of the residuals (the root mean square error) still represents the typical error in the estimate of the criterion value for a given practical value. To estimate the typical error from the spreadsheet might be too difficult, though, because you will have to modify the predicted values according to the type of curve you used. It may be easier to use a stats program. The typical error in the output from the stats program will be labeled either as the SEE, the root mean-square error, or the residual error. Some stats programs provide the typical error as a variance, in which case you will have to take the square root.

When you have subgroups of subjects with different characteristics (e.g., males and females), don't forget to check whether the subgroups have similar typical errors. To do so, you should label the points for each subgroup in the plot of residuals vs predicteds, because what looks like a uniform scatter might conceal a big difference between the subgroups. If there is a big difference, you shouldn't use a composite estimation equation for the two groups; instead, you should derive separate equations and separate typical errors for each subgroup.

Validity Limits of Agreement

By analogy with reliability limits of agreement, we can define **validity limits of agreement** as the 95% likely range or reference range for the difference between a subject's values for the criterion and practical measures. Let's try to understand this concept using the data in the validity spreadsheet.

The data are from a validity study in which the practical measure was body fat estimated using a Bod Pod, and the criterion measure was body fat measured with a DEXA scan. The units of body fat are percent of body mass (%BM). The limits of agreement (not shown in the spreadsheet) are -2.9 to 7.9 %BM, or 2.5 ± 5.4 %BM. You can interpret these numbers in two ways: there's a 95% chance that a subject's "true" (DEXA) body fat is within 2.5 ± 5.4 %BF of his or her Bod Pod value; or, if you measured a large number of subjects in the Bod Pod, 95% of them would have a DEXA body fat within 2.5 ± 5.4 %BF of their Bod Pod values. The value *2.5* in this example is the mean of the criterion-practical difference (or the difference between the means of the criterion and practical measures); it is sometimes known as the **bias** in the practical measure, but don't confuse this concept with the small-sample bias I described in connection with measures of reliability. The value ±*5.4* on its own is usually referred to as the limits of agreement; it is ±2.0x the standard deviation of the criterion-practical difference (= 2.7). The standard deviation of the criterion-practical difference is itself known as the **pure error** or **total error**.

Limits of agreement are related to the typical error of the estimate. When the slope of the estimation equation is exactly 1, the pure error is the same as the typical error, so in this special case the limits of agreement are twice the typical error. If the slope differs from 1, the limits of agreement are greater than twice the typical error. If the calibration equation is a curve rather than a straight line, the limits of agreement will also be greater than twice the typical error.

Advocates of limits of agreement encourage authors to plot the criterion-practical differences against the mean of these measures (or against the criterion). The resulting plot is similar to a plot of the residuals against the predicteds from the analysis of the estimation equation: if the estimation equation is a straight line of slope close to 1, the criterion-practical differences are the same as the residuals, and the mean of the criterion and practical is near enough to the predicted value. The plot will therefore allow you to check for heteroscedasticity. If the calibration equation is a straight line with slope different from 1, or if it is a curve, the scatter of points in the plot of the criterion-practical differences will show a trend towards a straight line or a curve, so it will be harder to tell if heteroscedasticity is present.

Validity limits of agreement suffer from problems similar to those of reliability limits of agreement: they are harder to understand than the typical error, and they are too large as a reference range for making a decision about a subject's true (criterion) measurement. The fact that the nature of the estimation equation affects the magnitude of the limits is also a serious problem. Unfortunately some authors have published limits of agreement without an estimation equation or the typical error, so readers cannot properly assess the practical measure and the published data cannot be used to recalibrate the practical measure.

Validity Correlation

The properties of the validity correlation are similar to those of the retest correlation. In particular...

- The correlation is a measure that combines within- and between-subject variation. *Within* here refers to the typical error of the estimate.
- The correlation gives you a good idea of how well the observed value of a variable (weight on bathroom scales in our example) retains the true rank order of subjects. Correlations >0.90 are needed to retain reasonable order in the ranking. Don't use those bathroom scales to assign athletes to competitive classes based on weight unless the validity correlation is well above 0.90!
- The correlation is unaffected by any systematic offset.
- The correlation is sensitive to the nature of the sample used to estimate it. For example, if the sample is homogeneous, the correlation will be low. So whenever you interpret a correlation, remember to take the sample into consideration.

- In contrast, the typical error of the estimate can be estimated from a sample of subjects that is not particularly representative of the population you want to study. You can usually assume the estimate applies to any subject in the population.

When it comes to calculating the validity correlation, you don't have much choice: if you fit a straight line to the data, the correlation is a Pearson correlation coefficient--there is no equivalent intraclass correlation coefficient. If you fit a curve, the stats program should provide you with a goodness-of-fit statistic called the variance explained or the R-squared. Just take the square root of this statistic and you have the equivalent of the Pearson correlation coefficient.

An estimate of validity correlation can also be obtained by taking the **square root of the concurrent reliability correlation**. By *concurrent reliability* I mean the immediate retest reliability, rather than the retest reliability over the time frame of any experiment you may be planning. This relationship between validity and reliability comes about because reliability is the correlation of something with itself (and there is error in both measurements), whereas validity is something correlated with the real thing (so there is error in only one measurement). The relationship can be derived from the definition of correlation (covariance divided by product of standard deviations) applied to the validity and reliability correlations.

The square root of concurrent reliability represents the maximum possible value for validity. The actual validity correlation could be less, because a measure can have high reliability and low validity. To put it another way, a measure can produce nonsense consistently!

Validity can be difficult to measure, because the true value of something can be difficult to assay. Measures other than the true value are called **surrogates**. These measures usually result in underestimates of validity when they are correlated with observed values, for obvious (I hope) reasons. Here's an example. Body density obtained by underwater weighing is often referred to as the gold standard for estimating percent body fat, but it is only a surrogate for true percent body fat. So if you are validating a skinfold estimate of body fat against the value obtained by underwater weighing, the validity correlation will be lower than if you validated the skinfold estimate against a more accurate method than underwater weighing, for example, a DEXA scan. Similarly the typical error of the estimate will be smaller when you validate skinfolds against DEXA rather than underwater weighing.

Validity of Nominal Variables

Validity of nominal variables can be expressed as a **kappa coefficient**, a statistic analogous to the Pearson correlation coefficient. Validity of nominal variables doesn't come up much in sport or exercise science--there's usually no question that you've got someone's sex or sport right--but it's a big issue in clinical medicine, where yes/no decisions have to be made about the presence of a disease or about whether to apply an expensive treatment. In cases where the variable has only two levels, clinicians have come up with other measures of validity that are easier to interpret than correlations. For example, **sensitivity** is the proportion or percent of true cases (people with a disease) correctly categorized as having the disease by the instrument/test/variable, and **specificity** is the proportion of true non-cases (healthy people) correctly categorized as being healthy. I have been unable to find or devise a simple relationship between the kappa coefficient and these two measures. One of these days...

Spreadsheet for Calculating Validity

This spreadsheet shows an example of a simple linear relationship between a practical measure (body fat derived from body density, estimated with the Bod Pod) and a criterion measure (body fat derived from dual energy X-ray absorptiometry, or DEXA). To use the spreadsheet, replace these data with your own data.

The spreadsheet estimates the calibration equation and the following measures of validity: typical error of the estimate, new-prediction error, correlation coefficient, and limits of agreement (but don't use them!). Analysis of log-transformed data is included for estimation of errors as percents of the mean.

Analysis of validity studies: Excel latest	Help!

APPLICATIONS OF VALIDITY

The applications are: tweaking up the sample size for a cross-sectional study, assessing an individual to predict her/his criterion value, comparing the validity of measures (to select a good one), and deciding whether a measure is valid enough for monitoring for changes in an individual's criterion value. I also consider sample size for validity studies on this page.

Sample Size for a Cross-Sectional Study

Just as reliability affected sample size in experimental or longitudinal studies, validity impacts sample size in descriptive or cross-sectional studies. In such studies, you measure each variable only once, and your outcomes are relationships between the variables. The lower the validity, the more the relationships are degraded, so the bigger the sample size you need to characterize them. For this application it's easier to discuss the effects of validity by considering the validity correlation rather than the typical error of the estimate.

The effect on the magnitude of the relationship between variables is proportional to the validity correlations of each variable. For example, suppose you are interested in the relationship between physical activity and health, and suppose that the true underlying relationship corresponds to a correlation of 0.50. If your measure of physical activity has a validity correlation of 0.7, then in your study of health and physical activity you will observe a correlation of only 0.5x0.7, or 0.35 (plus or minus sampling error, of course). The sample size required to detect a degraded relationship is inversely proportional the square of the validity correlation coefficient of each variable in the relationship. In our example, $1/0.70^2 = 2.0$, so you have to double the number of subjects. That's bad news, because most psychometric and subjective behavioral measures appear to have validities of 0.7 at best. Objective measures taken on lab instruments or in the field usually have validities of 0.8-0.9 or better, so you can often ignore the effect of validity of such variables on the magnitude of effects and the required sample size. Go to the section on sample size for cross-sectional studies for more information about the actual sample sizes you need.

Assessing an Individual

When you use a prediction equation to estimate a criterion value from a practical value (e.g., body fat from a sum of skinfolds), you should take into account the typical error of the estimate in much the same way as you do the typical error of measurement for a single measurement. You use the same factors to generate the likely range of the predicted value of the criterion (the factors for a *single* measurement in the table), but you multiply them by the typical error of the estimate. If the typical error is based on a study of less than 50 subjects, you will need to use a new-prediction error instead of the typical error, as explained earlier. The calculations are in the appropriate section of the spreadsheet for assessing an individual.

Example: You measure a client's skinfolds. You dig around in the literature and find an estimation equation that was developed for predicting body fat as a percent of body mass (%BM) in a large number of subjects similar to your client. The client's predicted body fat is 26.4 %BM, and the typical error of the estimate for the equation based on a large sample of similar subjects is 2.1 %BM. From the table in the section on reliability, the factor to multiply by the typical error for an 80% likely range is 1.28, which makes the limits 26.4 ± 1.28x2.1, or 23.7 to 29.1. You say to the client: "Your predicted body fat is 26.4 %BM, but the odds are 4 to 1 that your true (DEXA) body fat is somewhere between 24 and 29 %BM." Use the spreadsheet to generate these limits, and also the likelihood that the client's true value is greater than some reference value. For example, the likelihood that her true body fat is greater than 25 %BM is 74%, or odds of 3 to 1.

Comparing Validity of Measures

Just as the typical error of measurement was the best measure for comparing reliability of instruments, operators, or protocols, the typical error of the estimate is the best measure for comparing their validity, Do not compare the new-prediction errors, however derived: these are appropriate only for assessing individuals. As I explained with comparing measures of reliability, use the spreadsheet for confidence limits to calculate 80% or 90% likely ranges for the ratio of typical errors determined with different subjects, and to get likelihoods for the true ratio being greater that a reference ratio. Get an expert to use mixed modeling to estimate likely ranges when the same subjects are used to determine the typical errors.

Validity for Monitoring Changes

Our discussion of validity thus far has been concerned with the validity of a single measurement on an individual. But we often use a practical measure to **monitor for changes in a criterion measure**. For example, we use changes in skinfolds to infer that there have been changes in a subject's body fat. You might think that changes in skinfolds would have to reflect changes in body fat, but what if the amount of non-fat tissue in a skinfold is affected substantially by the subject's state of hydration or the menstrual cycle? In this situation a change in skinfold thickness may or may not represent a change in body fat, so skinfold thickness would no longer be a trustworthy measure for tracking body fat.

How do we decide whether skinfolds or some other practical measure is trustworthy? There are three approaches: correlation of spontaneous changes, correlation of induced changes, and correlation of original variables. The reliability of the practical and criterion measures usually has to be taken into account, so the statistics get quite complex. That might explain why no-one has yet published an adequate account of any of these approaches. I will therefore restrict this section to a qualitative overview.

Correlation of Spontaneous Changes

The obvious way to see how well changes in a practical measure track changes in a criterion measure is to measure some subjects, wait long enough for spontaneous changes to occur in some of them, measure them again, then plot the changes in the criterion measure against changes in the practical measure. If you get a very strong correlation (>0.95) you know the practical measure is trustworthy. The trouble is, you usually get a low correlation. Why? Because the real changes between measurements are usually of the same order of magnitude as the noise (the typical errors) in each measurement. The change scores for each measure therefore have a big contribution from the typical errors, which are random and uncorrelated, so the correlated true changes get lost in the noise in your plot of the change scores. You can estimate what the true correlation would be with the typical errors out of the picture, but if the observed correlation is poor, you will need hundreds of subjects to get enough precision for the estimate of the true correlation to decide whether the practical measure is any good.

Correlation of Induced Changes

Another approach is to make large changes happen by giving some kind of treatment to half your subjects. You then see how well the practical measure tracks the criterion measure in that half relative to the other half by correlating the change scores of all the subjects together. Even if you are successful in finding an effective treatment and subjects willing to undergo the treatment, you will have validated the practical measure only for changes induced by that particular treatment. In other words, you still won't know whether the practical measure is good for tracking spontaneous changes or changes brought about by other treatments.

Correlation of Original Variables

The third approach is to analyze data from a standard validity or calibration study. If the correlation between the practical measure and the criterion measure is near enough to perfect (>0.95), the two measures are effectively identical, so changes in the practical measure must track changes in the criterion. All the previous remarks about the correlation between change scores apply to the correlation between raw scores: the observed correlation will usually be a lot less than 0.95, because the correlation between the true values of the practical and criterion measures is degraded by the typical errors; you can estimate the true validity correlation by taking the concurrent retest reliability correlations into account; the true correlation needs to be greater than 0.95; and if the typical errors have a large degrading effect on the correlation, you will need hundreds of subjects in the validity and reliability studies to make a firm conclusion. You also need a reasonably good validity correlation to start with, which you won't get if your subjects are a homogeneous subgroup. Another problem is that even the true correlation between the measures may turn out to be less than 0.95, yet the practical measure will still track changes well. For example, the amount of non-fat tissue in skinfolds might vary between individuals with the same body fat (resulting in a relatively poor correlation between skinfolds and body fat), but the amount of non-fat tissue might not change with hydration status (so changes in skinfolds will still mirror changes in fat). This problem does not arise with the first two approaches, because the constant amount of non-fat tissue in each subject's skinfolds disappears from the change in skinfolds.

Each of these three approaches has its strengths and weaknesses. The third approach is best for a heterogeneous group of subjects, but only if it produces a very high and precise estimate of the true correlation. If the group is homogeneous, or if the true correlation is poor, you will have to use one of the two change-score approaches. Inducing changes with an appropriate treatment may give you a good estimate of the correlation between the change scores, but you end up validating the practical measure only for the treatment you used. The greatest strength of the first approach is that it validates the practical measure for tracking the changes that occur in the normal course of events, but the validation won't be clear cut if the changes are too small.

Sample Size for Validity Studies

As with reliability, sample size for estimation of validity is dictated by the need for precision. In this case precision of the typical error of the estimate or the new prediction error is the main consideration. You don't have the option of performing more than two tests; instead, you have to get adequate precision by increasing the number of subjects. For a reliability study involving a noisy measure, I recommended a minimum of 50 subjects tested three times. In terms of degrees of freedom (which dictate the precision of estimates of typical error), that is equivalent to about 100 subjects tested twice, so that is the preferred minimum sample size for a validity study of a noisy practical measure.

When there are several independent variables (regressors) in the prediction equation, an important consideration is ensuring that the typical error is uniform across the range of the regressor (or between subgroups represented by the regressor). Extrapolating from what I said about sample size for comparison of typical errors of measurement, I suggest adding 100 subjects for each extra regressor. (After all, if there are substantial differences in the typical error of the estimate between subgroups, and if the differences are resistant to transformation, you will have to perform separate analyses for each subgroup, each of which will require 100 subjects.) Many published validity studies with multiple regressors have involved several hundred subjects, but I don't think the choice of sample size in those studies was driven by consideration of uniformity of error. Another important consideration is keeping the new-prediction error from increasing substantially. It's easy to show (using Item 3 of the spreadsheet for a subject's true value) that increasing the number of subjects by 50 for each regressor after the first will ensure the new-prediction error is no more than 1% larger than the typical error. No worry there, if you use 100 subjects per regressor.

HOW MANY DIGITS?

Stats programs routinely crank out 8-figure accuracy for computed statistics. Your data are hardly ever good enough to justify that sort of precision. In any case, too many digits make data hard to comprehend, and most people hate numbers! So when you present your statistics in print or on a slide, it's important to show as few digits as possible.

Most statistics need either two significant digits (the first two digits), or two decimal places when the number is less than 1.0:

Percentages:	73%, 7.3%, 0.73%
Correlations:	0.97, 0.23, 0.05
Relative risks or odds ratios:	12, 2.4, 0.64
Effect sizes:	2.6, 0.51, 0.07

SDs usually need two significant digits. The mean must match the precision of the SD:

23500 ± 1300 (*not* 23538 ± 1341 etc.)
2350 ± 130
235 ± 13
2.35 ± 0.13
0.235 ± 0.013

The SD in descriptive statistics for height, weight, and age can often be shown with just one significant digit. After all, it doesn't really matter whether your subjects were 67 ± 5 or 67.3 ± 5.4 kg in weight:

height: 178 ± 7 cm
weight: 67 ± 5 kg
age: 23 ± 4 y

Naturally, if weight was an outcome variable, you would need to show two significant figures.

Avoid p values, but if you have to give in to the demands of a journal reviewer or editor who hasn't seen the Light, show no more than two significant digits: p = 0.007, 0.04, 0.35. See later for more about p values and statistical significance.

MEAN ± SD or MEAN ± SEM?

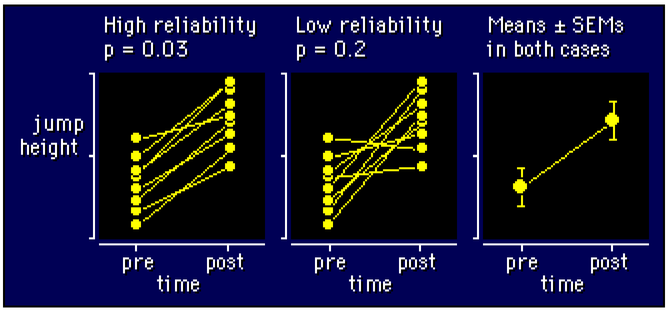
The standard deviation (SD) represents variation in the values of a variable, whereas the standard error of the mean (SEM) represents the spread that the mean of a sample of the values would have if you kept taking samples. So the SEM gives you an idea of the accuracy of the mean, and the SD gives you an idea of the variability of single observations. The two are related: SEM = SD/(square root of sample size).

Some people think you should show SEMs with means, because they think it's important to indicate how accurate the estimate of the mean is. And when you compare two means, they argue that showing the SEMs gives you an idea of whether there is a statistically significantdifference between the means. All very well, but here's why they're heading down the wrong track:

- For descriptive statistics of your subjects, you need the SD to give the reader an idea of the spread between subjects. Showing an SEM with the mean is silly.
- When you compare group means, showing SDs conveys an idea of the magnitude of the difference between the means, because you can see how big the difference is relative to the SDs. In other words, you can see how big the effect size is.
- It's important to visualize the SDs when there are several groups, because if the SDs differ too much, you may have to use log transformation or rank transformation before you compute confidence limits or p values. If the number of subjects differs between groups, the SEMs won't give you a direct visual impression of whether the SDs differ.
- If you think it's important to indicate statistical significance, show p values or confidence limits of the outcome statistic That's more accurate than showing SEMs. Besides, does anyone know how much SEMs have to overlap or not overlap before you can say the difference is significant? And does anyone know that the amount of overlap or non-overlap depends on the *relative* sample sizes?

- Most importantly, when you have means for pre and post scores in a repeated-measures experiment, the SEMs of these means do NOT give an impression of statistical significance of the change--a subtle point that challenges many statisticians. So if the SEMs don't show statistical significance in experiments, what's the point of having them anywhere else?

  Here's a figure to illustrate why SEMs don't convey statistical significance. It's for imaginary data in an experiment to increase jump height. The change in height is significant (p=0.03) when the measurement of jump height has high reliability, but not significant (p=0.2) when the reliability is low. But the SEMs are the same in both cases:



- The SEMs of the post-pre change scores in a treatment and control group *would* indicate statistical significance. But if you show the change scores, you should show the confidence interval for the change, not the SEM. You should also show the SD of the change scores for the treatment and control groups, because a substantial increase in the SD of the change scores in a treatment group relative to a control group indicates individual responses to the treatment. SEMs of the change scores would alert you to the possibility of individual responses only if the sample size was the same in both groups.

So when you see SEMs in a publication, smile, then mentally convert them into SDs to see how big the differences are between the groups. For example, if there are 25 subjects in a group, increase the size of the SEM by a factor of 5 (= square root of 25) to turn it into an SD.

The bottom line: never show SEMs. Never. Trust me.

Here endeth precision of measurement and summarizing data. On the next page we start generalizing to a population.

GENERALIZING TO A POPULATION

You have a bunch of numbers for a **sample** of subjects. But people don't really want to know about your sample, which was a one-off set of observations that will never be taken again. People are much more interested in what you can say about the **population** from which your sample was drawn. Why? Because things that are true for the population are true for a lot more people than just for your sample. Hence the second major purpose of statistics: **generalizing** from a sample to a population. It's also known as **making inferences** about a population on the basis of a sample. By the way, the term *population* doesn't mean the entire population of a country. It just means everyone in a well-defined group; for example, young adult male trained distance runners.

I deal first with **confidence limits**, which are the simplest and best way to understand generalization. **Bootstrapping**, **meta-analysis**, and **Bayesian analysis** are applications of confidence limits that I include on this page. On the next page are the related concepts of **p values and statistical significance**, followed by **type I and II errors** and a mention of **bias**. You can also download a slideshow that deals with all the material on these three pages, and more.

The second section is devoted to how we use **statistical models** or **tests** to generalize the relationships between variables. To generalize properly you need a sample of adequate size, so I deal with methods for **estimating sample size** in the final section.

Generalizing to a Population: CONFIDENCE LIMITS

GENERALIZING VIA CONFIDENCE LIMITS

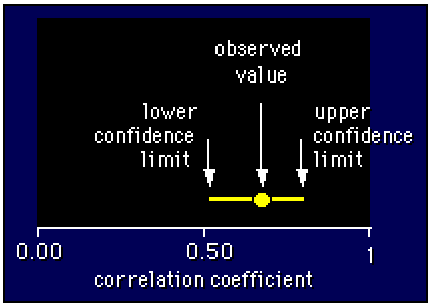
What can you say about the population when all you've got is a sample? Well, to start with, the value of a statistic (e.g. a correlation coefficient) derived from a sample is obviously one estimate of the value in the population. But the sample is only an approximation for the population, so the statistic is also only an approximation. If you drew a different sample, you'd get a different value.

The only way you can really get the population value is to measure everyone in the population. Even if that was possible, it would be a waste of resources. But it *is* possible to use your sample to calculate a range within which the population value is likely to fall. "Likely" is usually taken to be "95% of the time," and the range is called the **95% confidence interval**. The values at each end of the interval are called the **confidence limits**. All the values between the confidence limits make up the confidence interval. You can use *interval* and *limits*almost interchangeably.

Learn this plain-language definition: **the confidence interval is the likely range of the true value**. Note that there is *only one* true value, and that the confidence interval defines the range where it's most likely to be. The confidence interval is NOT the variability of the true value or of any other value between subjects. It is nothing like a standard deviation. If there are individual differences in the outcome, then there is more than one true value, but we'll deal with that later.

Another important concept embodied in confidence limits is **precision of estimation**. The wider the confidence interval, the less the precision. Research is all about getting **adequate precision** for things like a correlation coefficient, a difference in the mean between groups, the change in a mean following a treatment, and so on.

An Example



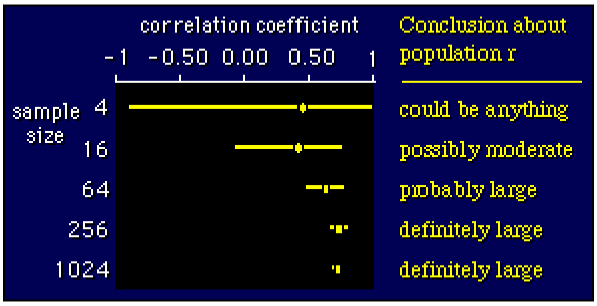
Suppose you observed a correlation of 0.68 between height and weight of 64 healthy undergraduate females. The 95% confidence limits are 0.52 and 0.79, which means that there's a 95% chance that the correlation between more-or-less *all* healthy undergraduate females is between 0.52 and 0.79. The figure shows it graphically. The confidence interval is the length of the line between the limits. You would report this result formally in a research paper as follows: the correlation between height and weight was 0.68; the 95% confidence interval was 0.52 to 0.79. I prefer the following less formal rendition: the correlation... was 0.68, and the likely range was 0.52 to 0.79.

Notice that the confidence limits in the above example are not spaced equally on each side of the observed value. That happens with non-normally distributed statistics like the correlation coefficient. Most other statistics are normally distributed, so the observed value falls in the middle of the confidence interval. For example, an observed enhancement in performance of 2.3% could have confidence limits of 1.3 to 3.3%. In such cases, you can use a ± sign to express the outcome in the following way: the enhancement was 2.3%, and the likely range (or confidence interval or limits) was ±1.0%. Of course, you mean by this that the limits are 2.3-1.0 and 2.3+1.0.

The lower and upper confidence limits need to be interpreted separately. The *lower* (or numerically smaller) limit shows how *small* the effect might be in the population; the *upper* limit shows how large the effect might be. Of course, you'll never know whether it really is that small or big unless you go out and measure the whole population. Or more subjects, anyway. Which brings us to the next important point: the more subjects, the narrower the confidence interval.

Effect of Sample Size on the Confidence Interval

Here's a figure showing how the width of the confidence interval depends on the number of subjects, for a correlation coefficient. It's the sort of thing you would get if you took bigger and bigger samples from a population.



Notice that you can't say anything useful about the population correlation when the sample has only 4 subjects. Already with 16 subjects you get the idea that it could be moderately positive. With 64 subjects the correlation is definitely positive and probably large, although it could also be moderate. The sample of 256 nails it as a large effect, and 1024 subjects give too much precision. The conclusions I have shown in the above figure are only approximate. Since drawing this figure, I have come up with an exact approach to making conclusions like *probably large.* See below.

The Confidence Interval and Statistical Significance

If the confidence interval does not overlap zero, the effect is said to be **statistically significant**. In the above figure, the results for the sample sizes of 64, 256, and 1024 are all statistically significant, whereas the other results are not statistically significant. We can also define statistical significance using something called a p value, but I'll deal with that on the next page.

We have a couple of plain-language ways of talking about something that is statistically significant: we say that *the true value is unlikely to be zero,* or that *there is a real effect.* These aren't bad ways to think about statistical significance, and you can sort of understand them by looking at the above figure, but they're not strictly correct. After all, the true value of something is never exactly zero anyway. I'll pick this issue up on the next page, under hypothesis testing.

The value for a statistic corresponding to no effect in the population is called the **null value**. For correlations and changes in the mean, the null value is zero. If the outcome statistic is a relative risk or odds ratio, the null value is 1 (equal risk or odds). So for these statistics, the result is statistically significant if the confidence interval does not overlap 1.

A Spreadsheet for Confidence Limits

To calculate confidence limits for a statistic, a stats program works out the variation between subjects, then estimates how that variation would translate into variation in your statistic, if you kept taking samples and measuring the statistic. (You don't have to take extra samples to get the variation from sample to sample.) When you tack that variation onto the value of your sample statistic, you end up with the confidence interval. The calculation requires some important simplifying assumptions, which I will deal with later.

Unfortunately, some stats programs don't provide confidence limits, but they all provide p values. I've therefore made a spreadsheet to calculate confidence limits from a p value, as explained on the next page. The calculation works for any normally distributed outcome statistic, such as the difference between means of two groups or two treatments. I've included calculations for confidence limits of relative risks and odds ratios, correlations, standard deviations, and comparison (ratio) of standard deviations.

I've also added columns to give chances of clinically or practically important effects. Make sure you come to terms with this stuff. It is more important than p values.

Update Oct 2007: the spreadsheet now generates customizable clinical and mechanistic inferences, consistent with an article on inferences in Sportscience in 2005. The inferences are also consistent with an article on sample-size estimation in Sportscience in 2006.

Spreadsheet for confidence limits and inferences: Download

Bootstrapping (Resampling)

Another way of getting confidence limits, when you have a reasonable sample size, is by the wonderful new technique of **bootstrapping**. It's a way of calculating confidence intervals for virtually any outcome statistic. It's tricky to set up, so you use it only for difficult statistics like the difference between two correlation coefficients for the same subjects. And you'll need an expert with a high-powered stats program to help you do it.

For example, you might want to use a fitness test in a large study, so you do a pilot first to see which of two tests is better. The tests might be submaximal exercise tests to determine maximum oxygen uptake. "Better" would mean the test with higher validity, in other words the test with the higher correlation with true maximum oxygen uptake. So you might get a sample of 20 subjects to do the two tests and a third maximal test for true maximum oxygen uptake. The validity correlations turn out to be 0.71 and 0.77. Sure, use the test with the higher correlation, but what if it's more difficult to administer? Now you begin to wonder if the tests are really *that* different. The difference is 0.06. That's actually a trivial difference, and if it was the real difference, it wouldn't matter which test you used. But the observed difference is never the real difference, and that's why we need confidence intervals. If the confidence interval was 0.03 to 0.09, you'd be satisfied that one test is a bit better than another, but that it still doesn't really matter, and you would choose the easier test. If the confidence interval was -0.11 to 0.23, you couldn't be confident about which test is better. The best decision then would be to test more subjects to narrow down the confidence interval.

Anyway, bootstrapping is how you can get the confidence interval. The term *bootstrapping* refers to the old story about people lifting themselves off the ground by pulling on the backs of their own boots. A similar seemingly impossible thing occurs when you *resample* (to describe it more formally) to get confidence intervals. Here's how it works.

For a reasonably representative sample of maybe 20 or more subjects, you can recreate (bootstrap) the population by duplicating the sample endlessly. Sounds immoral, if not impossible, but simulations have shown that it works! Next step is to draw, say, 1000 samples from this population, each of the same size as your original sample. In any given sample, some subjects will appear twice or more, while others won't be there at all. No matter. Next you calculate the values of the outcome statistic for each of these samples. In our example above, that would be the difference between the correlations. Finally, you find the middle 95% of the values (i.e. the 2.5th percentile and the 97.5th percentile). That's the 95% confidence interval for your outcome! Cool, eh?

The median value from your 1000 samples should be virtually the same as the value from the original sample. If it's not, something is wrong. Sometimes the variables have to be **transformed** in some way to get over this problem. For example, to get the confidence interval for the difference between correlation coefficients, you first have to convert the correlations using something called the **Fisher z transformation**: $z = 0.5\log[(1 + r)/(1 - r)]$. This equation looks horribly complicated, but all it does is make the correlations extend out beyond the value 1.0. It makes them behave like normally distributed variables.

How do you "duplicate endlessly" to recreate the population? Actually you don't duplicate the data set. If your original sample had 20 observations, you use a random number generator in the stats program to select a sample of 20 from these 20. Then you do it again, and again, and again...

At the moment I don't know of a good rule to decide when a sample is big enough to use bootstrapping. Twenty observations seems to be OK. Note, though, that if you have subgroups in your data set that are part of the outcome statistic, you need at least 20 in each subgroup. For example, if you wanted to compare a correlation in boys and girls, you would need at least 20 boys and 20 girls.

And now for a test of your understanding. If you can recreate the population by duplicating the sample endlessly, why bother with all that resampling stuff? Why not just work out the value of the statistic you want from say a one-off sample of a million observations taken from this population? With a million observations, it will be really accurate! Answer: Well, ummm... the value you calculate from a million observations will be almost exactly the same as the value from your original sample of 20. You're no better off. OK, it was a silly question.

Meta-Analysis

I deal with meta-analysis here, because it is an application of confidence intervals. Meta-analysis is literally an analysis of analyses, which is near enough to what it is really: a synthesis of all published research on a particular effect (e.g. the effect of exercise on depression). The aim is to reach a conclusion about the magnitude of the effect in the population.

The finding in a meta-analytic study is the mean effect of all the studies, with an overall confidence interval. In deriving the mean, more weight is given to studies with better designs: more subjects, proper random selection from the population, proper randomization to any experimental and control groups, double blinding, and low dropout rate. Studies that don't meet enough criteria are sometimes excluded outright from the meta-analysis.

Whenever you read a meta-analysis involving longitudinal (experimental) studies, check to make sure the statistician used the correct standard deviation to calculate the effect size. It should always be the average standard deviation of the before and/or after scores. Some statisticians have used the standard deviation of the before-after difference score, which can make the effects look much bigger than they really are.

Bayesian Analysis

Bayesian analysis is a kind of meta-analysis in which you combine **observed data** with your **prior belief** about something and end up with a **posterior belief**. In short, it's a way to **update your belief**. Clinicians use this approach informally when they try to diagnose a patient's problem. They have a belief about possible causes of the problem, and they probe for symptoms, test for signs of possible diseases, and order blood tests or scans or whatever to get data that will make their belief in one cause much greater than other possible causes. Fine, and no-one disputes the utility of this approach in the clinical setting with an individual patient or client. The disputes arise when statisticians try to apply it to the analysis of research data from a sample of a population. Let's start with the usual approach (also known as the **frequentist** approach) to such data, then see how a Bayesian would handle it.

Suppose you're interested in the effect of a certain drug on performance. You study this problem by conducting a randomized controlled trial on a sample of a population. You end up with confidence limits for the true effect of the drug in the population. If you're a frequentist you publish the confidence limits. But if you're a Bayesian, you also factor in your prior belief about the efficacy of the drug, and you publish **credibility limits** representing your posterior (updated) belief. For example, you might have believed the drug had no effect (0.0%), and you were really skeptical, so you gave this effect confidence limits of -0.5% to +0.5%. You then did the study and found a positive effect of 3.0%, with confidence limits of 1.0% to 5.0%. Combine those with your prior belief and you end up with a posterior belief that the effect of the drug is 0.6%, with confidence limits of -1.0% to 3.2%. Let's assume a marginal effect is 1%, a small effect is 3%, and a moderate effect is 5%. A Bayesian concludes (from the credibility limits of -1.0% to 3.2%) that the drug has anything from a marginal negative effect to a small positive effect. A frequentist concludes (from the confidence limits of 1.0% to 5.0%) that the drug has anything from a marginal positive to a moderate positive effect.

There are formal procedures for combining your prior belief with your data to get your posterior belief. In fact, the procedure works just like a meta-analysis of two studies: the first study is the one you've just done to get an observed effect with real data; the other "study" is your prior belief about what the effect was. The observed effect and your belief are combined with weighting factors inversely proportional to the square of the widths of their confidence intervals. For example, if you have a very strong prior belief, your confidence (= credibility) interval for your belief will be narrow, so only a markedly different observed effect with a narrow confidence interval will change your belief. On the other hand, if you are not at all sure about the effect, your confidence interval for your prior belief will be wide, so the confidence limits for your posterior belief won't be much different from those provided by the data. To take this example to an extreme, if you have no prior belief, the posterior confidence limits are identical to those provided by the data.

A positive aspect of the Bayesian approach is that it encapsulates the manner in which we assimilate research findings. New evidence that agrees with our preconceived notions reinforces our beliefs, whereas we tend to disregard evidence that flies in the face of our cherished prejudices or has no apparent mechanism. Sure, but even as a frequentist you can tackle these issues *qualitatively* in the Discussion section of your paper. If you try to *quantify* your prior belief, you run into two problems. First, your belief and the real data are combined with weighting factors, but they are otherwise on an equal footing. That's acceptable to a frequentist only if it's quite clear that the outcome of the Bayesian analysis is still only a belief, not a real effect. Secondly, exactly how do you convert a belief into a quantitative effect, and how do you give it confidence limits? (Bayesians give their belief a complete probability distribution, but the principle is the same.) You could--and probably do--base the belief on the results of other studies, but you might just as well meta-analyze these other studies to get your prior "belief". In that case, though, your posterior "belief" will be identical to a meta-analysis of all the studies, including the one you've just done. In other words, it's not a Bayesian analysis any more.

Bayesian analysis may be justified where a decision has to be made with limited real data. The prior belief could be the average belief of several experts. When I hear of a specific example, I will update this page. Meanwhile, click here for a response to this section from Mike Evans, a Bayesian.

P VALUES AND STATISTICAL SIGNIFICANCE

The traditional approach to reporting a result requires you to say whether it is statistically significant. You are supposed to do it by generating a **p value** from a **test statistic.** You then indicate a significant result with "p<0.05". So let's find out what this p is, what's special about **0.05**, and when to use p. I'll also deal with the related topics of **one-tailed vs two-tailed tests**, and **hypothesis testing**.

What is a P Value?

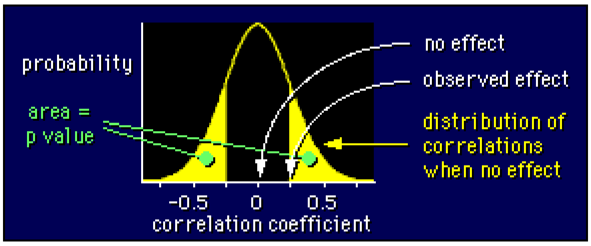
It's difficult, this one. P is short for probability: the probability of getting something more extreme than your result, when there is no effect in the population. Bizarre! And what's this got to do with statistical significance? Let's see.

I've already defined statistical significance in terms of confidence intervals. The other approach to statistical significance--the one that involves p values--is a bit convoluted. First you assume there is no effect in the population. Then you see if the value you get for the effect in your sample is the sort of value you would expect for no effect in the population. If the value you get is unlikely for no effect, you conclude there *is* an effect, and you say the result is "statistically significant".

Let's take an example. You are interested in the correlation between two things, say height and weight, and you have a sample of 20 subjects. OK, assume there is no correlation in the population. Now, what are some unlikely values for a correlation with a sample of 20? It depends on what we mean by "unlikely". Let's make it mean "extreme values, 5% of the time". In that case, with 20 subjects, all correlations more positive than 0.44 or more negative than -0.44 will occur only 5% of the time. What did you get in your sample? 0.25? OK, that's not an unlikely value, so the result is not statistically significant. Or if you got -0.63, the result would be statistically significant. Easy!

But wait a minute. What about the p value? Yes, umm, well... The problem is that stats programs don't give you the threshold values, ±0.44 in our example. That's the way it *used* to be done before computers. You looked up a table of threshold values for correlations or for some other statistic to see whether your value was more or less than the threshold value, for your sample size. Stats programs *could* do it that way, but they don't. You want the correlation corresponding to a probability of 5%, but the stats program gives you the probability corresponding to your observed correlation--in other words, the probability of something more extreme than your correlation, either positive or negative. That's the p value. A bit of thought will satisfy you that if the p value is less than 0.05 (5%), your correlation must be greater than the threshold value, so the result is statistically significant. For an observed correlation of 0.25 with 20 subjects, a stats package would return a p value of 0.30. The correlation is therefore not statistically significant.

Phew! Here's our example summarized in a diagram:



The curve shows the probability of getting a particular value of the correlation in a sample of 20, when the correlation in the population is zero. For a particular observed value, say 0.25 as shown, the p value is the probability of getting anything more positive than 0.25 *and*anything more negative than -0.25. That probability is the sum of the shaded areas under the probability curve. It's about 30% of the area, or a p value of 0.3. (The total area under a probability curve is 1, which means absolute certainty, because you have to get a value of some kind.)

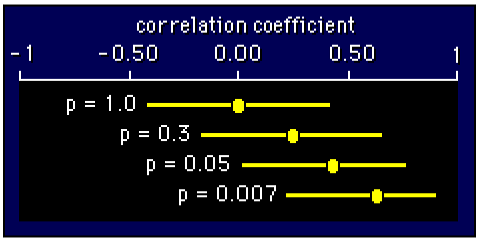
Results falling in that shaded area are not really unlikely, are they? No, we need a smaller area before we get excited about the result. Usually it's an area of 5%, or a p value of 0.05. In the example, that would happen for correlations greater than 0.44 or less than -0.44. So an observed correlation of 0.44 (or -0.44) would have a p value of 0.05. Bigger correlations would have even smaller p values and would be statistically significant.

Test Statistics

The stats program works out the p value either directly for the statistic you're interested in (e.g. a correlation), or for a test statistic that has a 1:1 relationship with the effect statistic. A test statistic is just another kind of effect statistic, one that is easier for statisticians and computers to handle. Common test statistics are t, F, and chi-squared. You don't ever need to know how these statistics are defined, or what their values are. All you need is the p value, or better still, the confidence limits or interval for your effect statistic.

P Values and Confidence Intervals

Speaking of confidence intervals, let's bring them back into the picture. It's possible to show that the two definitions of statistical significance are compatible--that getting a p value of less than 0.05 is the same as having a 95% confidence interval that doesn't overlap zero. I won't try to explain it, other than to say that you have to slide the confidence interval sideways to prove it. But make sure you are happy with this figure, which shows some examples of the relationship between p values and 95% confidence intervals for observed correlations in our example of a sample of 20 subjects.



The relationship between p values and confidence intervals also provides us with a more sensible way to think about what the "p" in "p value" stands for. I've already said that it's the probability of a more extreme (positive or negative) result than what you observed, when the population value is null. But hey, what does that really mean? I get lost every time I try to wrap my brain around it. Here's something much better: if your observed effect is positive, then half of the p value is the probability that the true effect is negative. For example, you observed a correlation of 0.25, and the p value was 0.30. OK, the chance that the true value of the correlation is *negative* (less than zero) is 0.15 or 15%; or you can say that the odds of a negative correlation are 0.15**:**0.85, or about 1 to 6 (1 to 0.85/0.15). Maybe it's better to it turn around and talk about a probability of 0.85 (= 1 - p/2), or odds of 6 to 1, that the true effect is *positive.*Here's another example: you observed an increase in performance of 2.6%, and the p value was 0.04, so the probability that performance really did increase is 0.98, or 49 to 1. Check your understanding by working out how to interpret a p value of exactly 1.

So, if you want to include p values in your next paper, here is a new way to describe them in the Methods section: "Each p value represents twice the probability that the true value of the effect has any value with sign opposite to that of the observed value." I wonder if reviewers will accept it. In plain language, if you *observe* a positive effect, 1 - p/2 is the probability that the *true* effect is positive. But even with this interpretation, p values are not a great way to generalize an outcome from a sample to a population, because what matters is clinical significance, not statistical significance.

Clinical vs Statistical Significance

As we've just seen, the p value gives you a way to talk about the probability that the effect has *any* positive (or negative) value. To recap, if you observe a positive effect, and it's statistically significant, then the true value of the effect is likely to be positive. But if you're going to all the trouble of using probabilities to describe magnitudes of effects, it's better to talk about the probability that the effect is *substantially* positive (or negative). Why? Because we want to know the probability that the true value is big enough to count for something in the world. In other words, we want to know the probability of **clinical or practical significance**. To work out that probability, you will have to think about and take into account the **smallest clinically important positive and negative values of the effect**; that is, the smallest values that matter to your subjects. (For more on that topic, see the page about a scale of magnitudes.) Then it's a relatively simple matter to calculate the probability that the true value of the effect is greater than the positive value, and the probability that the true value is less than the negative value.

I have now included the calculations in the spreadsheet for confidence limits and likelihoods. I've called the smallest clinically important value a "threshold value for chances [of a clinically important effect]". You have to choose a threshold value on the basis of experience or understanding. You also have to include the observed value of the statistic and the p value provided by your stats program. For changes or differences between means you also have to provide the number of degrees of freedom for the effect, but the exact value isn't crucial. The spreadsheet then gives you the chances (expressed as probabilities and odds) that the true value is **clinically positive** (greater than the smallest positive clinically important value), **clinically negative** (less than the negative of the smallest important value), and **clinically trivial** (between the positive and negative smallest important values). The spreadsheet also works out confidence limits, as explained in the next section below.

Use the spreadsheet to play around with some p values, observed values of a statistic, and smallest clinically important values to see what the chances are like. I've got an example there showing that a p value of 0.20 can give chances of 80%, 15% and 5% for clinically positive, trivial, and negative values. Wow! It's clear from data like these that editors who stick to a policy of "publishable if and only if p<0.05" are preventing clinically useful findings from seeing the light of day.

I have written two short articles on this topic at the Sportscience site. The first article introduces the topic, pretty much as above. The second article summarizes a **Powerpoint slide show** I have been using for a seminar with the title *Statistical vs Clinical or Practical Significance,* in which I explain hypothesis testing, P values, statistical significance, confidence limits, probabilities of clinical significance, a qualitative scale for interpreting clinical probabilities, and some examples of how to use the probabilities in practice. Download the presentation (91 KB) by (right-)clicking on this link. View it as a full slide show so you see each slide build.

Confidence Limits from a P Value

Stats programs often don't give you confidence limits, but they always give you the p value. So here's a clever way to derive the confidence limits from the p value. It works for differences between means in descriptive or experimental studies, and for any normally distributed statistic from a sample. Best of all, it's on a spreadsheet! I explain how it works in the next paragraph, but it's a bit tricky and you don't have to understand it to use the spreadsheet. Link back to the previous page to download the spreadsheet.

I'll explain with an example. Suppose you've done a controlled experiment on the effect of a drug on time to run 10,000 m. Suppose the overall difference between the means you're interested in is 46 seconds, with a p value of 0.26. From the definition of the p value (see top figure on this page), we can draw a normal probability distribution centered on a difference of 0 seconds, such that there is an area of 0.26/2 = 0.13 to the right of 46 and a similar area to the left of -46. Or to put it another way, the area between -46 and 46 is 1-0.26 = 0.74. If we now shift that distribution until it's centered over 46, it represents the probability distribution for the true value. We know that the chance of the true value being between 0 and 92 is 0.74, so now all we need is the range that will make the chance 0.95, and that will be our 95% confidence interval. To work it out, we use the fact that the distribution is normal. That allows us to calculate how many standard deviations (also known as the z score) we have to go on each side of the mean to enclose 0.74 of the area under the normal curve. We get that from tables of the cumulative normal distribution, or the function NORMSINV in an Excel spreadsheet. Answer: 1.13 standard deviations. Ah, but we know that 1.96 standard deviations encloses 95% of the area, and because the 1.13 standard deviations represents 46 seconds, our confidence interval must be -46(1.96/1.13) to +46(1.96/1.13), i.e. -34 to +126.

Fine, except that it's not really a normal distribution. With a finite number of subjects, it's actually a t distribution, so we have to use TINV in Excel. What's more, the 95% confidence limits are really a titch more than 1.96 standard deviations each side of the mean. Exactly how much more depends on the number of subjects, or more precisely, the number of degrees of freedom. With your own data, search around in the output from the analysis until you find the degrees of freedom for the error term or the residuals. Put it into the spreadsheet, along with the observed value of the effect statistic, and its p value (not the p value for the model or for an effect in the model, unless it is the statistic). If you can't find the number of degrees of freedom on the output, the spreadsheet tells you how to calculate it. And if you don't get it exactly right, don't worry: the confidence limits hardly change for more than 20 degrees of freedom.

One Tail or Two?

Notice in the first figure on this page that the p value is calculated for *both* tails of the distribution of the statistic. That follows naturally from the meaning of statistical significance, and it's why tests of significance are sometimes called **two tailed**. In principle you could eliminate one tail, double the area of other tail, then declare statistical significance if the observed value fell within the one-tailed area. The result would be a **one-tailed** test. Your Type I error rate would still be 5%, but a smaller effect would turn out to be statistically significant. In other words, you would have more power to detect the effect.

So how come we don't do all tests as one-tailed tests? Hmm... The people who support the idea of such tests--and they are a vanishing breed--argue that you can use it to test for, say, a positive result only if you have a good reason for believing *beforehand* that the outcome will be positive. I hope I am characterizing their position correctly, because I don't understand it. What is a "good reason"? It seems to me that you would have to be *absolutely certain* that the outcome would be positive, but in that case running the test for statistical significance is pointless! I therefore don't buy into one-tailed tests. If you have any doubts, revert to the confidence-interval view of significance: one-sided confidence intervals just don't make sense, but confidence limits equally placed on each side of the observed value is unquestionably a correct view.

Except that... there is a justification for one-tailed tests after all. You just interpret the p value differently. P values for one-tailed tests are half those for two-tailed tests. It follows that the p value from a one-tailed test is the exact probability that the true value of the effect has opposite sign to what you have observed, and 1 - p is the probability that the true value of the effect has the same sign, as I explained above. Hey, we don't have to muck around with p/2. So here's what you could write in the Methods section of your paper: "All tests of significance are one-tailed in the direction of the observed effect. The resulting p values represent the probability that the true value of the effect is of sign opposite to the observed value." Give it a go and see what happens. Such a statement would be anathema to reviewers or statisticians who assert that an observed positive result is not a justification for doing a one-tailed test for a positive result. They would argue that you are downgrading the criterion for deciding what is "statistically significant", because you are effectively performing tests with a Type I error of 10%. Fair enough, so don't mention statistical significance at all. Just show 95% confidence limits, and simply say in the Methods: "Our p values, derived from one-tailed tests, represent the probability that the true value of the effect is of sign opposite to the observed value."

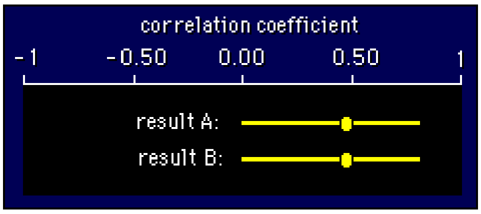
But as I discussed above, the probability that an effect has a *substantially* positive (or negative) value is more useful than the probability that the effect has *any* positive (or negative) value. Confidence limits are better than one-tailed p values from that point of view, which is why you should always include confidence limits.

Why 0.05?

What's so special about a p value of 0.05, or a confidence interval of 95%? Nothing really. Someone decided that it was reasonable, so we're now stuck with it. P < 0.01 has also become a bit of a tradition for declaring significance. Both are hangovers from the days before computers, when it was difficult to calculate exact p values for the value of a test statistic. Instead, people used tables of values for the test statistic corresponding to a few arbitrarily chosen p values, namely 0.05, 0.01, and sometimes 0.001.

These values have now become enshrined as the threshold values for declaring statistical significance. Journals usually want you to state which one you're using. For example, if you state that your **level of significance** is 5% (also called an **alpha level**), then you're allowed to call any result with a p value of less than 0.05 significant. In many journals results in figures are marked with one asterisk (*) if p<0.05 and two (**) if p<0.01.

Some researchers and statisticians claim that a decision has to be made about whether a result is statistically significant. According to this logic, if p is less than 0.05 you have a publishable result, and if p is greater than 0.05, you don't.Here's a diagram showing the folly of this view of the world. One of these results is statistically significant (p<0.05), and the other isn't (p>0.05). Which is publishable? Answer: both are, although you'd have to say in both cases that more subjects should have been tested to narrow down the likely range of values for the correlation. And in case you missed the point, the exact p values are 0.049 and 0.051. Don't ask me which is which!



Some journals persist with the old-fashioned practice of allowing authors to show statistically significant results with p<0.05 or p<0.01, and non-significant results with p>0.05. Exact p values convey more information, but confidence intervals give a much better idea of what could be going on in the population. And with confidence intervals you don't get hung up on p values of 0.06.

Hypothesis Testing

The philosophy of making a decision about statistical significance also spawned the practice of **hypothesis testing**, which has grown to the extent that some departments make their research students list the hypotheses to be tested in their projects. The idea is that you state a**null hypothesis** (i.e. that there is no effect), then see if the data you get allow you to reject it. Which means there is no effect until proved otherwise--like being innocent until proved guilty. This philosophy comes through clearly in such statements as "let's see if there is an effect".

What's wrong here? Well, people may be truly innocent, but in nature effects are seldom truly zero. You probably wouldn't investigate something if you really believed there was nothing going on. So what really matters is **estimating** the magnitude of effects, not **testing**whether they are zero. But that's only a philosophical issue. There are more important practical issues. Getting students to test hypotheses diverts their attention from the magnitude of the result to the magnitude of the p value. Read that previous sentence again, please, it's *that*important. So when a student researcher gets p>0.05 and therefore "accepts the null hypothesis", s/he usually concludes erroneously that there is no effect. And if s/he gets p<0.05 and therefore "rejects the null hypothesis", s/he still has little idea of how big or how small the effect could be in the population. In fact, most research students don't even know they are supposed to be making inferences about population values of a statistic, even after they have done statistics courses. That's how hopelessly confusing hypothesis testing and p values are.

"Let's see if there is an effect" isn't too bad, if what you mean is "let's see if there is a *non-trivial* effect". That's what people really intend. But a test for statistical significance does not address the question of whether the effect is non-trivial; instead, it's a test of whether the effect is greater than zero (for an observed positive effect). And it's easy to get a statistically significant effect that could be trivial, so hypothesis testing doesn't do a proper job. With confidence limits you can see immediately whether the effect could be trivial

Research *questions* are more important than research hypotheses. The right question is "how big is the effect?" And I don't just mean the effect you observe in your sample. I mean the effect in the population, so you will have to show confidence limits to delimit the population effect.

Using P Values

When I first published this book, I was prepared to concede that p values have a use when you report lots of effects. For example, with 20 correlations in a table, the ones marked with asterisks stand out from the rest. Now I'm not so sure about the utility of those asterisks. The non-significant results might be just as interesting. For example, if the sample size is large enough, a non-significant result means the effect can only be trivial, which is just as important as the effect being substantial. And if the sample size isn't large enough, a non-significant result with the lower confidence limit in the trivial region (e.g. r = 0.34, 95%CL = -0.03 to 0.62) is arguably only a tad less interesting than a statistically significant result with the lower confidence limit still in the trivial region (e.g. r = 0.38, 95%CL = 0.02 to 0.65). So I think I'll harden my attitude. No more p values.

By the way, if you *do* report p values with your outcome statistics, there is no point in reporting the value of the test statistic as well. It's superfluous information, and few people know how to interpret the magnitude of the test statistic anyway. But you must make sure you give confidence limits or *exact* p values, and describe the statistical modeling procedure in the Methods section.

GETTING IT WRONG

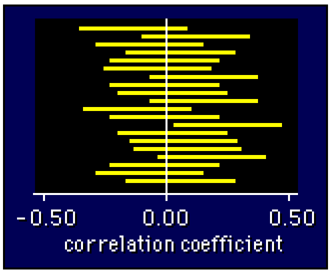
The words *probability* and *confidence* seem to come up a lot. You should be getting the message that few things are definite in our discipline, or in any empirical science. Sometimes we get it wrong.

From the point of view of confidence intervals, getting it wrong is simply a matter of the population value being outside the confidence interval. I call it a **Type O error**. You can think of the "O" as standing either for "outside (the confidence interval)" or for "zero" (as opposed to errors of Type I and II, which it supersedes). For 95% confidence limits the Type O error rate is 5%, by definition. From the point of view of hypothesis testing, getting it wrong is much more complicated. You can be responsible for a false alarm or Type I error, and a failed alarm or Type II error. An entirely different way to get things wrong is to have bias in your estimate of an effect. This page ends with a link to download a PowerPoint slide presentation, in which I summarize and in some instances extend important points from these pages.

Type I Error

A level of significance of 5% is the rate you'll declare results to be significant when there are no relationships in the population. In other words, it's the rate of false alarms or false positives. Such things happen, because some samples show a relationship just by chance.

For example, here are typical 95% confidence intervals for 20 samples of the same size for a population in which the correlation is 0.00. (The sample size is irrelevant.) Notice that one of the correlations is statistically significant. If that happened to be your study, you would rush into print saying that there is a correlation, when in reality there isn't. You would be the victim of a Type I error. Of course, you wouldn't know until others--or you--had tested more subjects and found a narrower confidence interval overlapping zero.



Cumulative Type I and Type O Error Rates

The only time you need to worry about setting the Type I error rate is when you look for a lot of effects in your data. The more effects you look for, the more likely it is that you will turn up an effect that seems bigger than it really is. This phenomenon is usually called the **inflation of the overall Type I error rate**, or the **cumulative Type I error rate**. So if you're going fishing for relationships amongst a lot of variables, and you want your readers to believe every "catch" (significant effect), you're supposed to reduce the Type I error rate by adjusting the p value downwards for declaring statistical significance.

The simplest adjustment is called the **Bonferroni**. For example, if you do three tests, you should reduce the p value to 0.05/3, or about 0.02. This adjustment follows quite simply from the meaning of probability, on the assumption that the three tests are independent. If the tests are not independent, the adjustment is too severe.

Those of us who use confidence intervals rather than p values have to be aware that **inflation of the Type O error** also happens when we report more than one effect. For example, if there are two independent effects, the probability that at least one will be outside its confidence interval is about 10%. We could increase the width of our confidence intervals to bring the overall probability back to 5%. For example, Bonferroni-adjusted 95% confidence intervals for three effects would each be 98% confidence intervals. Adjusting the confidence intervals in this or some other way will keep the purists happy, but I'm not sure it's such a good idea. I prefer to see the raw 95% confidence intervals, and I prefer to make my own mental adjustment when there are lots of effects. I just look at the results and think to myself, OK, the population value might be outside the interval for one or two of those effects (depending on how many results are reported). The fact that the effects are reported in one publication is no justification for widening the confidence intervals, in my view. You might just as well argue that all the confidence intervals in the entire issue of the journal should be widened, to keep the cumulative error rate for the issue in check! And why stop with one issue... So I don't think confidence intervals or p values should be adjusted, but I know many will disagree.

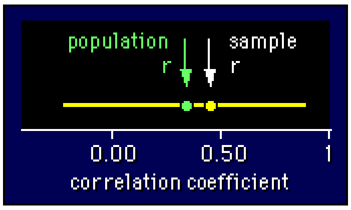
Why not use a lower p value all the time, for example a p value of 0.01, to declare significance? Surely that way only one in every 100 effects you test for is likely to be bogus? Yes, but it is harder to get significant results, unless you use a bigger sample to narrow down that confidence interval. In any case, you are entitled to stay with a 5% level for one or two tests, if they are **pre-planned**--in other words, if you set up the whole study just to do these tests. It's only when you tack on a lot of other tests afterwards (so-called **post-hoc** tests) that you need to be wary of false alarms.

Controlling the Type I error comes up a lot in analysis of variance, when you do comparisons between several groups or levels. For more insights see estimates and contrasts in one-way ANOVA and estimates and contrasts in repeated-measures ANOVA.

Type II Error

The other sort of error is the chance you'll *miss* the effect (i.e. declare that there is no significant effect) when it really is there. In other words, it's the rate of failed alarms or false negatives. Once again, the alarm will fail sometimes purely by chance: the effect is present in the population, but the sample you drew doesn't show it.

The smaller the sample, the more likely you are to commit a Type II error, because the confidence interval is wider and is therefore more likely to overlap zero. Here's an example in which a Type II error has occurred for a correlation. Imagine you got this result:



I've indicated where the population correlation is for this example, but of course, in reality you wouldn't know where it was. I've made the true correlation about 0.40, which is well worth detecting. But it hasn't been detected, because the confidence interval overlaps zero. A big-enough sample size would have produced a confidence interval that didn't overlap zero, in which case you would have detected a correlation, so no Type II error would have occur red. Now, a test of your understanding: where would the population r have to be on the figure for a Type II error NOT to have been made? Answer: on or close to 0.00.

The Type II error needs to be considered explicitly at the time you design your study. That's when you're supposed to work out the sample size needed to make sure your study has the **power** to detect anything useful. For this purpose the usual Type II error rate is set to 20%, or 10% for really classy studies. The power of the study is sometimes referred to as 80% (or 90% for a Type II error rate of 10%). In other words, the study has enough power to detect the smallest worthwhile effects 80% (or 90%) of the time.

Here's something interesting that no-one seems to mention: **cumulative Type II error rate**--in other words, the chance that you will miss at least one effect when you test for more than one. Is your head starting to spin? Mine is! Don't worry, just go back to confidence limits and the notion of cumulative Type O error. When you are looking at lots of effects, the near equivalent of inflated Type II error is the increased chance that any one of the effects will be bigger than you think it could be (bigger than its upper confidence limit). Come to think of it, the near equivalent of inflated Type I error is the increased chance that any one of the effects will be smaller than you think.

Bias

People use the term **bias** to describe deviation from the truth. That's the way we use the term in statistics, too: we say that a statistic is biased if the average value of the statistic from many samples is different from the value in the population. To put it simply, the value from a sample tends to be wrong.

The easiest way to get bias is to use a sample that is in some way a non-random sample of the population: if the average subject in the sample tends to be different from the average person in the population, the effect you are looking at could well be different in the sample compared with the population.

Some statistics are biased, if we calculate them in the wrong way. Using n instead of n-1 to work out a standard deviation is a good example. There is also bias in some reliability statistics. Building up a sample size in stages can also result in bias, as I describe in sample size on the fly.

SLIDES ON CONFIDENCE LIMITS

Click here to download a PowerPoint 97/98 set of 30 slides on the topic "Planning, Performing, and Publishing Research with Confidence Limits", which I presented on this topic at the annual meeting of the American College of Sports Medicine in Seattle, June 4 1999. If you have trouble downloading or opening the file, click here

Generalizing to a Population:
STATISTICAL MODELS

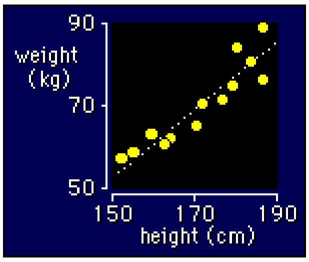
WHAT IS A MODEL?

Can you see that women are usually different from men in certain characteristics? Can you see that taller people are heavier, in general? Can you see that participation rates differ between sports? Fine, you're already an expert in the use of models! All we're going to do now is formalize your intuitive understanding, and put numbers on everything. Let's hope we don't destroy your intuition in the process!

What do these three examples have in common? Something affected by or related to something else? Yes, a model is **a relationship between variables**. The relationships we deal with are usually simple: women are shorter than men, by a fixed amount; body mass is proportional to height or maybe height^2; the chance that any given person will participate in a particular sport is a simple function of age, sex, socio-economic status, or whatever.

Inasmuch as models are relationships between variables, I could have dealt with them under the general heading of Summarizing Data, and in particular in the pages on effect statistics. Certainly, if our only aim was to characterize the relationship in a sample, then that's where these pages should have been. But we fit a model to data from a sample almost always to make a statement about the model in the population. That is, we want to make a statement about the precision of the estimate of the effect statistic(s) describing the model, using things likeconfidence limits and/or chances of clinical benefit (or P values and/or statistical significance, if you are stuck in the 20th Century). So I deal with models here, under the heading of Generalizing to a Population. Let's be clear, though: a model is another way of summarizing data using effect statistics.

On the next pages I'll get more technical about how different kinds of variable produce different models. Meanwhile, let's take a sneak preview of a simple model.



Here are some imaginary heights and weights of a sample of adults. As soon as you plot data like these, you want to draw a straight line through them. The straight line is the model. You decide you want to draw one, and the stats program does the rest. It finds the equation of the straight line that fits the data best. It also produces a correlation coefficient, which is a measure of how well the line fits (or, same thing, how close the relationship between height and weight comes to being a straight line). And, inasmuch as the data are a sample, the program even produces confidence limits for the line, or a p value for a test of whether there is a line in the population at all. In fact, statistical modeling and statistical testing mean the same thing.

Is this all too easy, or what? It gets a bit more complicated for things like analysis of covariance, repeated measures categorical modeling, and so on, but the principle is the same.

. SIMPLE MODELS AND TESTS

I was confused by the wide variety of models until I found a simple way of categorizing them. The trick is to think about the variables in the model as either **numeric** or **nominal**, and as either **dependent** or **independent**.

You already know about numeric and nominal variables: numeric variables have numbers as values, and nominals have names or levels. Either type can be dependent or independent. The variable you're most interested in is known as the dependent variable, because it might be dependent on, or affected by, something else that you've measured, which is therefore an independent variable. For example people's weight (dependent variable) might depend on their height (independent variable). *Independent* is not a very good term, because you can have several independent variables, and they may not be independent of each other. So, a better term for independent variables is **effects**, because they have an effect on the dependent variable. They're also known as **predictor** or **explanatory** variables, for obvious reasons. A nominal predictor variable is also known as a **grouping** variable, because it divides the data up into groups.

Now let's talk about the relationships between variables. I'm going to use a short-hand method to represent the relationship between a dependent and independent variable. For example, if I want to show that height affects weight, I will write:

weight <= height

The "<=" is a backwards-pointing arrow, by the way! Read the expression as "weight is affected by height".

Sure, it would be more sensible to write height => weight, and read it as "height affects weight", but statisticians are used to seeing the dependent variable on the left. It goes back to writing things like Y = X + 1. We don't write X + 1 = Y (although we could). So in general, let's write

dependent <= independent

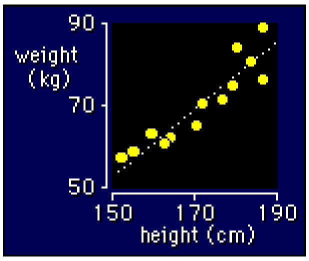
Now, if we just substitute nominal and numeric variables for the dependent and independent variables, we'll end up with four different simple models. Here they are, with their names:

numeric <= numeric	Linear Regression
numeric <= nominal	T Test and One-Way ANOVA
nominal <= nominal	Contingency Table
nominal <= numeric	Categorical Modeling

I detail each model on the next four pages.

Linear Regression

Let's use the same example that I used to introduce the concept of statistical models. As you can see, data for two variables like weight and height scream out to have a straight line drawn through them. The straight line will allow us to predict any person's weight from a knowledge of that person's height. Obviously, the prediction won't be perfect, so we will also be able to say how strong the linear relationship is between weight and height, or how well the straight line fits the data (the goodness of fit).



Here's how we represent the model:
  model: numeric <= numeric
  example: weight <= height

You normally think about a straight line as Y = mX + c, where m is the slope and c is the intercept. The way I would write this relationship, using the above notation, is simply Y <= X. We don't have to worry about showing the constants, but the stats program worries about them. They're the **parameters** in the model.

The Slope
The most interesting parameter in a linear model is usually the slope. If the slope is zero, the line is flat, so there's no relationship between the variables. In the example, the slope is about 0.75 kg per cm (an increase in weight of 0.75 kg for each cm increase in height). We can also calculate the slope in two ways that don't have those ugly units (kg per cm).

One way is to calculate the percent change in weight per percent change in height. It's unusual, but sometimes it's the best way, especially for variables that need log transformation. The slope expressed as % per % comes directly out of the analysis of log-transformed variables.

The other way to remove the units is to **normalize** the two variables by dividing their values by their standard deviations, then fit the straight line. The resulting slope is known as a **standardized regression coefficient**. It represents the change in weight, expressed as a fraction of the standard deviation, per standard deviation change in height. You can also generate it by multiplying the slope (in kg per cm) by the ratio of the standard deviations for height over the standard deviation for weight. In a simple linear regression, the value of the standardized regression coefficient is exactly the same as the correlation coefficient, and you can interpret its magnitude in the same way. In the example, the value is about 0.9, or a difference of 0.9 standard deviations in weight per change of one standard deviation in height. That's a really strong relationship!

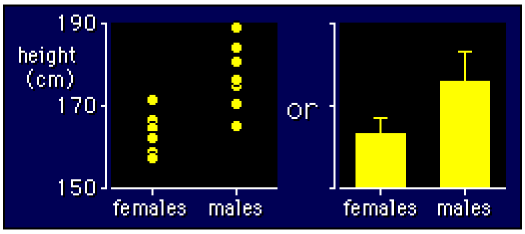
Goodness of Fit

The stats program works out values for the slope and intercept (the parameters) that give the best fit. I'll explain how after I've dealt with all four simple models. Meanwhile, we want a measure of how good the fit is. The correlation coefficient is one such measure. Another way to represent the fit is to square the correlation coefficient, multiply it by 100, then call the result the **percent of variance explained**, or **percent R^2**. For example, the R^2 represents the proportion of variation in weight that can be attributed to height, when there is a linear relationship between weight and height. A correlation of 0.9 is equivalent to an R^2 of 0.81 or 81%. I'll explain more about goodness of fit in a few pages' time.

The p value or the confidence interval for the correlation coefficient tell us how good the fit is likely to be in the population. The program can also give confidence intervals or p values for the slope and intercept. The correlation coefficient can be considered as a test statistic for whether the line fits the data at all. But stats programs can also produce another statistic for this purpose, called the F ratio. The values for F are quite different from those for r, but there is a one-to-one relationship between them, and the r and the F have the same p value for a given sample.

T Test and One-Way ANOVA

model: numeric <= nominal
example: height <= sex



In other words, if you know someone's sex, what does that tell you about their height? Or, how well do the height data fall into two groups when you label the values by sex? The test statistic for the test of whether sex has an effect on height is called Student's t, or just t. Hence the name of this model, the t test.

When there are three or more levels for the nominal variable, a simple approach is to run a series of t tests between all the pairs of levels. For example, we might be interested in the heights of athletes in three sports, so we could run t test for each pair of sports. (Note that this approach is not the same as a paired t test. That comes later.) A more powerful approach is to analyze all the data in one go. The model is the same, but it is now called a one-way analysis of variance (ANOVA), and the test statistic is the F ratio. So t tests are just a special case of ANOVA: if you analyze the means of two groups by ANOVA, you get the same results as doing it with a t test.

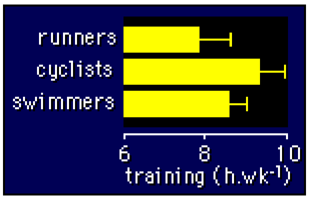
The term *analysis of variance* is a source of confusion for newbies. In spite of its name, ANOVA is concerned with differences between *means* of groups, not differences between *variances.* The name analysis of *variance* comes from the way the procedure uses variances to decide whether the means are different. A better acronym for this model would be ANOVASMAD (analysis of variance to see if means are different)! The way it works is simple: the program looks to see what the variation (variance) is *within* the groups, then works out how that variation would translate into variation (i.e. differences) *between* the groups, taking into account how many subjects there are in the groups. If the observed differences are a lot bigger than what you'd expect by chance, you have statistical significance. In our example, there are only two groups, so variation between groups is just the difference between the means.

I won't bother with trying to represent this model as an equation like Y = mX + c. Suffice to say that it can be done, simply by making an X variable representing sex that has the value 0 for females and 1 for males, say (or vice versa). So it is also a "linear" model, even though we don't normally think about it as a straight line. The parameters in the model are simply the mean for the females and the mean for the males.

The spreadsheet for analysis of controlled trials includes a comparison of the means (and standard deviations) of two groups at baseline. You can use it for any tests of two independent groups, as in the above example.. Ignore all the stuff related to comparisons of changes in the mean in the two groups.

Comparisons of Means

With a t test, the thing we're most interested in is, of course, a comparison of the two means. You should think about the best way to express the difference in the means for your data: raw units, percent difference, or effect size. And don't forget to look at and discuss the magnitude of the difference and the magnitude of its confidence limits.



With three or more levels for the nominal variable, we can start asking interesting questions about the differences between pairs or combinations of means. Such comparisons of means are known as **estimates** or **contrasts**. For example, suppose we are exploring the relationship between training hours per week (the dependent variable) and sport (the nominal independent variable). Suppose sport has three levels: runners, cyclists, and swimmers, as shown. We can ask the question, are there differences overall between the sports? The answer would be given by the p value for sport in the model. And what about the difference between cycling and running? Yes, we can dial up the difference and look at its p value or confidence interval. We do that by subtracting the value for the parameter (the mean) for cycling from that for running, using the appropriate syntax in the stats program. We could even ask how different swimming was from the average of running and cycling, and so on. There's also a special kind of contrast (polynomials) you can apply if the levels are a numbered sequence and you want to describe a curve drawn through the values for each level.

If you're expressing a difference between means as an effect size, the standard deviation to use in the calculation is the root mean square error (RMSE) in the ANOVA. An ANOVA is based on the assumption that the standard deviation in the same in all the groups, and the RMSE represents the estimate of that standard deviation. You can think of the RMSE as the average standard deviation for all of the groups.

With lots of contrasts, the chance of any one of them being spuriously statistically significant--in other words, the overall chance of a Type I error--goes up. So stats programs usually have built-in ways of controlling the overall Type I error rate in an ANOVA. Basically they adjust the p value down for declaring statistical significance, although you don't see it like that on the printout. These methods have statisticians' names: Tukey, Duncan, Bonferroni... They're also known as post-hoc tests or simply post hocs. I don't use them, because I now use confidence limits and clinical significance rather than statistical significance, so I don't test anything.

One approach to controlling the Type I error rate with multiple contrasts is simply not to perform the contrasts unless the overall effect is significant. In other words, you don't ask *where* the differences are between groups unless there *is* an overall difference between groups. Sounds reasonable, but wait a moment! If there is no overall statistically significant difference between groups, surely none of the contrasts will turn up significant? Yes, it can happen! There's jitter in the p values, and there's nothing to say that the p value for the overall effect is any more valid than the p value for individual contrasts. So if you've set up your study with a particular contrast in mind--a pre-planned contrast--go ahead and do that contrast, regardless of the p value for the overall effect. Performing the pre-planned contrast does not have to be contingent upon obtaining significance for the overall effect. Those of us who prefer confidence intervals to p values can understand why: the estimate of the difference between groups has a confidence interval that may or may not overlap zero, and the confidence interval for the overall effect (expressed in some measure of goodness of fit) may or may not overlap zero. There is no need to reconcile the two.

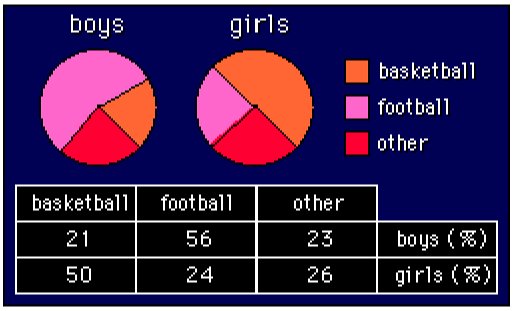
Goodness of Fit

What statistic do we use to talk about how well the ANOVA model fits the data? It's not used that frequently, but you can extract an R^2 just like you do for a straight line. The R^2 represents how well all the levels of the grouping (nominal) variable fit the data. More aboutgoodness of fit soon.

Contingency Table (Chi-Squared Test)

---

model: nominal <= nominal
example: sport <= sex



basketball	football	other	
21	56	23	boys (%)
50	24	26	girls (%)

What effect does a kid's sex have on the kind of sport s/he likes? That's the sort of question we address with this model, as shown in the example of the sport preferences of a sample of boys and girls.

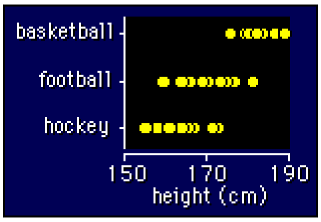
The word *contingency* in the name of the model refers, I guess, to the relationship between the two variables. *Table* speaks for itself. The test for whether there is any relationship at all is known as the chi-squared test, from the test statistic, chi squared (c^2: this will come up as c^2 if your browser doesn't show symbols). It's pretty obvious that there's a strong relationship in the example. Whether the relationship is significant would depend on the number of boys and girls.

We don't normally think about parameters for this model, but they would be the probabilities of opting for each sport, for each sex. Goodness of fit is also not usually calculated, but various analogs of the correlation coefficient (e.g. the kappa coefficient) make their appearance occasionally. Those outcome measures that we have already met, the relative risk and odds ratio, make sense only for 2 x 2 tables or for comparing 2 x 2 cells in a bigger table. Most stats programs can calculate the confidence intervals for these outcome measures

When you have more than two rows or columns in the table (e.g. the three sports above), the chi-squared test tells you whether there is any relationship, but it doesn't tell you where the differences are. Now, just as we can do pairwise tests for the different levels of a grouping variable in an ANOVA, we can in principle test for differences between frequencies of males and females in pairs of sports, or between one sport and the rest, or whatever. In the above example, it's clear that the "other" category does not differ between sexes (which is actually a comparison of "other" with basketball and football combined, if you think about it), whereas every other pairwise comparison looks like it could be different. The funny thing is, there is no tradition for doing such pairwise tests in a contingency table, or for controlling the type I error, not that I know of anyway. All that people do is state whether there is an effect overall or not, then eyeball the frequencies in the table and comment on where the biggest differences are. Strange...

Categorical Modeling

model: nominal <= numeric
example: sport <= height



Another name for this model is **discriminant function analysis**, because, for example, you end up with a function of height that allows you to predict which sport a person belongs in. Whether the person would do *better* in that sport is another question requiring different variables, of course!

This model is the least common of the four. It's much easier to turn the model around to make it height <= sport, and apply... what? Yes, an ANOVA. Strictly speaking, though, if the research calls for height to be the independent variable, then you should apply categorical modeling, and express your outcomes as an effect of height on the probability of being in the different sports. You end up with horrible outcome measures like an odds ratio per unit of height, which blows away everyone except card-carrying statisticians! By the way, the test statistic is chi-squared.

Another approach is to treat each sport as a separate variable, then code the value as 1 if the person belongs to the sport and 0 if not. You can also group the sports in some sensible way and again code a variable as 0 or 1 if the person belongs to that group. You then treat these variables as numeric and analyze them in the usual way. You have to assume the sample size is big enough to ensure the sampling distribution of the outcome statistic is normal. I explain what all this means shortly.

A special case of categorical modeling is **logistic regression**. You have to use this model when the dependent variable is ordinal. A page devoted to this problem also comes up shortly. You could also turn simple models like these around and analyze them as ANOVAs, but you shouldn't.

MODELS: IMPORTANT DETAILS

On this page are details of how a stats program fits a model, which you will need to understand before you tackle the other major topic on this page, calculating confidence limits and p values. You'll find that your data sometimes violate assumptions the stats programs make when they perform the calculations. One fix--the t test for unequal variances--is at the bottom of this page. Other fixes are on the following pages: log transformation, rank transformation, non-parametric models and tests, models for ordinal dependent variables, and non-linear models.

How a Stats Program Fits a Model

Key terms you will meet here are **parameters**, **predicted values**, **residuals**, and **goodness of fit**.

Parameters
Recall that, to fit a straight line to data, you need a slope and an intercept for the line. The slope and intercept are called **parameters** of the model.

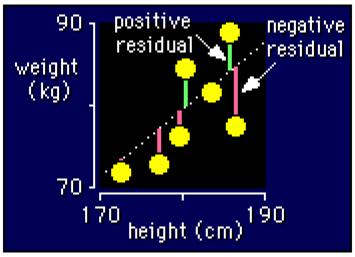
If the model is a t test (for example, heights of girls vs boys) or simple ANOVA (heights of three or more subgroups), the parameters are single values of height for each subgroup that best fit the data. The values are, of course, the means of each subgroup.

To complete the picture, what are the parameters if we're modeling the frequency of something in one or more groups (for example, the prevalence of injury in different sports)? Too easy: it's just the frequency in each group, or more exactly, the probability that a person in a each group will have an injury or whatever.

Solution and Residuals
As we've seen, a relationship or model is represented by parameters like the slope and intercept of a line,

or the means of groups. To fit a model, the stats program finds values of the parameters that fit the data best. These values are called the **solution**. But HOW is it done?

The standard method is to find the values of the parameters that produce the minimum difference between the **observed** values of the dependent variable and the values of the dependent variable that would be **predicted** by the model. The difference between an observed and a predicted value is a **residual**. It's easiest to understand when the model you are fitting is a straight line. See the diagram, which is the top corner of the height-weight graph blown up so you can see what's what. The observed values are just the weights. The predicted values are the weights on the line corresponding to each observed weight. You should be able to see that if you drew a straight line further away from the points, the residuals overall would get bigger.



The program actually minimizes the sum of the *squares* of the residuals. Why not just minimize the sum of the raw residuals? Let's just accept that it works best to square the residuals first. So when you fit a straight line, it's known as the **least-squares line**. This line doesn't always look quite like the best line--the slope sometimes looks a bit too shallow--but that's because of the way the distances are measured and minimized only in the Y direction. Trust me, it IS the best line! The same applies when you fit curves rather than straight lines. By the way, the root-mean-square error derived from any model is the standard deviation of the residuals, and the mean of the residuals is always zero.

If the model we're fitting is means for different groups (in an ANOVA), the predicted values are just the means for each group, and the residuals are the differences between, for example, each girl's values and the girls' mean, and ditto the boys.. Easy stuff. And when we're looking at different frequencies of something in different groups (contingency table), the predicted values are just the observed frequencies. There aren't any residuals as such in that case, but you start to get them when you have categorical modeling. No need to understand this subtlety, though.

Goodness of Fit

I've already introduced the concept of goodness of fit for simple linear regression. I stated that the correlation is a good way to describe it, and that 100x the square of the correlation--the percent of variance explained--is also used. Now that you know about residuals, I can explain goodness of fit a bit more.

Obviously, the smaller the residuals, the better the fit. One measure of the magnitude of the residuals is their standard deviation, alias the root mean square error. But what can we compare the error with to get a generic measure of goodness of fit? Answer: the standard deviation of the dependent variable itself, before we try to fit any model. This standard deviation represents the amount of variation in the dependent variable, and the error represents the variation that's left over after we fit the model. But statisticians like to make things complicated, right? So they square the standard deviation to get **total variance,** and they square the error to get **error variance.** The total variance minus the error variance is... wait for it... the **variance explained** by the model. Divide the variance explained by the total variance and you have something equivalent to the square of a correlation coefficient--we call it the **goodness-of-fit R^2** for the model. Multiply it by 100, and you have... the percent of total variance explained by the model, or just the **percent of variance explained.** Cool!

Stats books have lots of formulae involving sums of squares, which are what we used to use to calculate statistics in the days before computers. Sums of squares are directly related to variances. The total sum of squares is the sum of the squares of each observed value after the mean has been subtracted from it. The residual sum of squares is exactly what it says. Subtract the residual SS from the total SS, divide by the total SS, and you have another formula for R^2.

The R^2 is also identical to the square of the correlation between the observed values and the values predicted by the model--quite a nice way of thinking about goodness of fit for a complex model. And of course, for a simple linear regression, the R^2 for the model is the same as r^2, the square of the correlation coefficient.

The stats program should give you a p value for the R^2, which will help you make decisions about the linear relationship between the dependent variable and independent variable. What programs currently won't do is give you confidence limits for the R^2. Maybe we don't really want it anyway. It's easier to interpret R rather than R^2, as discussed in the page on scale of magnitudes. So take the square root of the R^2, then work out the confidence interval of this correlation using the Fisher z transformation. (The "n" in the Fisher formula in this case is the number of degrees of freedom for the error term in the linear model, minus 1.) I've set it all up on the spreadsheet for confidence limits.

Goodness of fit for models in which the dependent variable is nominal is a bit trickier. As I mentioned earlier, goodness of fit is not usually calculated for these models, but various analogs of the correlation coefficient (e.g. the **kappa coefficient**) can be used. The clinical measures of sensitivity and specificity can also be regarded as measures of goodness of fit.

Calculating Confidence Limits

Calculating the value of an effect statistic like the difference between two means is usually easy. Calculating the confidence interval or confidence limits and/or the p value for the true value of the statistic is another matter. In the usual models (t tests, ANOVA, linear or curvilinear regression), the calculations are based on three simplifying assumptions: **independence of observations**, **normality of sampling distribution**, and **uniformity of residuals**. Let's see what happens and what you have to do if your data violate these assumptions.

Independence of Observations

Independence of observations refers to the notion that the value of one datum is unrelated to any other datum. In other words, knowing the value of one observation gives you no information about the value of any other. To see what happens if this assumption is violated, let's take an extreme case. Imagine you are doing calculations on what you think is a large data set, but unbeknownst to you, someone has inflated the sample size simply by duplicating every observation. The observations in such a data set are definitely not independent! The correct confidence interval or p value for a given effect in the data would be given by an analysis based on the original sample size, obviously. But the confidence interval you get with the spuriously inflated sample will be narrower (by a factor of about 1/root2, or 0.7), and the corresponding p value will be smaller too. In general, then, lack of independence of observations results in incorrectly narrow confidence limits and incorrectly small p values, because the effective sample size is less than what you think it is.

Observations that are not independent are also said to be **correlated** or **interdependent** . There are some clever tests for independence in some specific situations, but in general you have to decide yourself--without recourse to statistical tests--whether there is substantial interdependence among the observations in your data set.

An obvious example of interdependence occurs in any intervention: the subjects each provide two or more observations before and after the intervention, and all the observations belonging to a given subject usually have similar values compared with values from other subjects. The usual approach to such data is a repeated-measures analysis or mixed modeling.

A statistic summarizing the amount of independence in a set of observations is the **degrees of freedom**. Well, actually, the degrees of freedom summarizes the amount of independence in the *residuals* in your model--and that's as it should be, because the residuals are what the stats program uses to calculate the confidence interval. The degrees of freedom is simply the total number of independent bits of error in the residuals. Here's an example: fit a straight line to 10 points and you will have 10 residuals but only 8 degrees of freedom, because the model estimates two parameters--the slope and intercept of the line. Some stats procedures account for interdependence of residuals in some complex models by estimating a reduced number of degrees of freedom for the residuals.

You don't have to worry about the details of degrees of freedom, but you should be aware that the more parameters you estimate in your model, the more degrees of freedom you lose. That's not a problem if your sample size is large, but with a small sample size the uncertainty in the magnitude of the error will translate into substantially wider confidence intervals, because the width of the confidence interval is proportional to the value of a t statistic for the number of degrees of freedom of the residuals. The effect starts to bite when your model reduces your number of degrees of freedom to 10 or less. So if you have small sample sizes, you'll need to keep your models simple.

Normality of Sampling Distribution

The sampling distribution of any outcome statistic is the distribution you would expect to get for the values of the statistic, if you repeated your study many times. To calculate the confidence limits for the true value of most statistics, a stats program has to assume that this distribution is normal. If your raw data have a normal-looking distribution, the sampling distribution of all the usual outcome statistics based on the data will definitely be normal, so there's no problem. But even if your raw data are not normally distributed, the sampling distribution of a given statistic is often so close to normal that you can trust the confidence limits and p value.

Question: When *can't* you trust the confidence limits and p value?
Answer: Depends on how non-normal your residuals are, and how small your sample is.

Question: How non-normal, how small, and *how come!?*
Answer: Let's address *how come* first. The residuals sort-of add together to give you the sampling distribution of your statistic. And when you add enough randomly varying things like residuals together, even though each of them is not normally distributed, they smooth out into a normal distribution. You can actually prove it mathematically, and the proof is called the **central limit theorem**. Of course, the more non-normal the residuals, the bigger the sample size you will need to get a normal sampling distribution. But there are apparently no rules about how non-normal the residuals and how small the sample size need to be before an analysis falls over. I could find nothing on the Web and I got no joy when I inquired on a statistics mailing list, so I did some simulations to find out about *how non-normal* and *how small.*

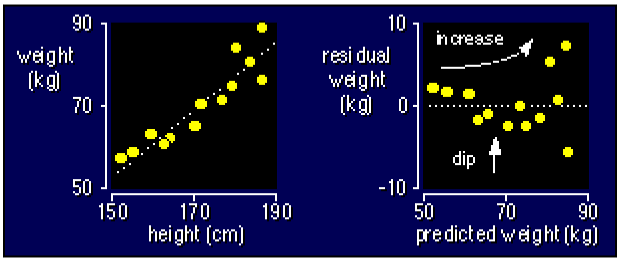
I used a variable that has grossly non-normal residuals: an ordinal variable having only two values (0 and 1). This variable is what researchers use to code no/yes responses or 2-point **Likert scales** in questionnaires. I restricted the analyses to unpaired t tests of two groups with various sample sizes (for example, a comparison of the responses of 10 boys vs 30 girls). I found that the confidence limits started to go wrong for sample sizes of 10 or less if the average response in one or both groups was <0.3 or >0.7 (corresponding to more than 70% of the responses in each group being on one or other level of the 2-point Likert scale). I also tried ordinal variables with 3, 4 and 5 values, corresponding to Likert scales with 3, 4 and 5 levels. For any kind of reasonably realistic spread of responses on these scales, the confidence limits were accurate for samples of 10 or more in each group. The confidence limits went awry only when responses were stacked up on the bottom or top level of the scale, in the same manner as for the 2-point scale. Even then they came right for samples of 50 or so.

My conclusion is that people (me included) have worried needlessly about non-normality of residuals. The time to get worried is when the residuals look really awful and you have a sample of only 10 or so subjects. When that happens, you'll have to try other approaches:logistic regression in the case of Likert-type responses stacked up on one or other extreme value, and some kind of **transformation** for everything else. I explain transformations on the next few pages, starting with log transformation.

By the way, do not test for non-normality of the residuals. Residuals that look only remotely normal will work fine in your analyses, even though the test tells you they aren't normal. And with large sample sizes, residuals that look indistinguishable from normal sometimes return a positive test for non-normality, but your analyses will definitely be OK here.

Uniformity of Residuals

Your residuals are uniform if their mean is zero and their scatter (standard deviation) is the same for any subsample or subgroup of observations. So what does all *that* mean? I'll try to make things clear with an example. Here are the data for weight vs height for the linear regression you saw earlier. I've also plotted the residuals against the predicteds, which is the best way to check for non-uniformity. (Go back up this page if you need to remind yourself what residuals and predicteds are.)

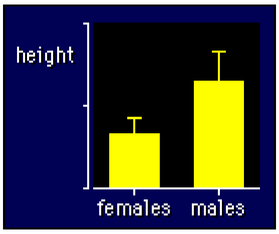


Notice that the residuals are more scattered for larger predicted values: that means the standard deviation is larger for the larger values of height. Notice also that there is a dip in the middle of the plot of the residuals: that means the mean of the residuals is positive for the lowest heights, negative for middle heights, and possibly positive again for the highest heights. The dip and the scatter are a bit hard to see, I admit, but you get used to spotting such things. They're easier to see with larger sample sizes. Actually, you can spot the dip and the scatter on the original plot on the left--look closely and you will see that the scatter about the line gets bigger for bigger heights, and that the data tend to curve about the line of best fit. Another way to spot non-uniformity is to group the predicted values into quantiles of equal numbers of observations, for example five groups (quintiles), then plot the means and standard deviations of the residuals against the mean of each quantile of the predicteds. In the above example, the mean of the residuals would be positive for the lowest group of predicteds, then the means would go negative, then the highest subgroup would be positive again. The standard deviation would be small in the lowest group and largest in the highest group.

The jargon to describe this behavior of residuals is **heteroscedasticity** (hetero- = uneven; -scedasticity = scatter). I prefer **non-uniform residuals** or just **bad residuals**. So what? Well, the dip in the residuals tells you that the straight line doesn't fit the data properly, as you can see from the raw data, too. So you should either fit a **non-linear model** (a curve) instead of a straight line, or you should **transform the data** to get a straight line. We'll deal with non-linear models shortly. Transforming the data means changing the values of a variable in some systematic way. The most common ways are **log transformation** and **rank transformation**. I go into the details of transformations starting on the next page.

So much for the dip in the residuals, but what about the increasing scatter? If the scatter is not the same everywhere, the confidence limits won't be right, because stats programs work out the confidence limits (and p values) on the assumption that the scatter is the same. For regression-type models, as in the above example, you have no choice: you have to find a transformation that makes the scatter uniform, as we will see on the next page. For t tests, there is a simpler solution.

The unpaired t test often gives rise to non-uniform residuals, but it's really easy to spot them and to do something about them. The residuals in an unpaired t test are simply the differences between each observation and the mean within each of the two groups. The mean of the residuals in each group is therefore automatically zero, so you don't have to worry about that. The scatter of the residuals is simply the standard deviation of the observations within each group.



For example, if you are comparing females and males, check to see how different the standard deviation is for the males vs the females (see the figure). Then, if the sample size is the same in each group, forget about it! Yes, it doesn't matter how different the standard deviations are, you get the right confidence limits when the groups are the same size. But if the groups differ in size (e.g., by a factor of 1.1 or more) and the standard deviations differ too (also by a factor of 1.1 or more), then you gotta do something. And the answer is... use the **t test with unequal variances**! It would be better to call it the t test with unequal standard deviations, but statisticians prefer to use the term *variance* (the square of the standard deviation). Most stats programs offer this option along with the usual t test. Your stats program may even do an extra test of whether the variances within the two groups are equal, but don't take any notice of the p value for that test. Look instead at the size of the standard deviations and the sample sizes, then make your decision about which form of the t test to use. Actually, when the variances are the same and the sample sizes are the same, the confidence limits provided by the two tests are practically identical, so you might as well always use the t test with unequal variances

Uniformity of Residuals in Complex Models

I have been dealing with uniformity of residuals in simple models, which consist of one predictor variable (height or sex in our examples). In more complex models--those with two or more predictor variables--you should check the uniformity of the residuals not just across the range of predicted values but also across the range of values of each of the predictors. You do that by getting your stats package to output all the residuals with corresponding values of each predictor variable. For each predictor you then do a plot of the residuals (Y axis) against the values of the predictor (X axis) and look for non-uniformity.

As in the first example above, you get a better idea of any non-uniformity by plotting the mean and standard deviation of the residuals. For each nominal predictor variable, you show the means and standard deviations on a single plot that includes each level of the variable. For each numeric predictor, you group the values of the predictor into quantiles of equal numbers of observations, just like I explained above for the predicted values. You then plot the means and standard deviations of the residuals against the mean of each quantile of the predictor. Any consistent pattern in the mean of the residuals indicates that the mathematical form of the model for that predictor is inadequate. For example, you might need to introduce a quadratic or a non-linear term for that variable into the model . Any substantial difference in the standard deviation of the residuals for different levels or values of the predictor indicates that you need to transform the dependent variable.

Log Transformation for Better Fits

In log transformation you use natural logs of the values of the variable in your analyses, rather than the original raw values. Log transformation works for data where you can see that the residuals get bigger for bigger values of the dependent variable. Such trends in the residuals occur often, because the error or change in the value of an outcome variable is often a **percent** of the value rather than an **absolute** value. For the same percent error, a bigger value of the variable means a bigger absolute error, so residuals are bigger too. Taking logs "pulls in" the residuals for the bigger values. Here's how.

A percent error in a variable is actually a multiplicative factor. For example, an error of 5% means the error is typically 5/100 times the value of the variable. When you take logs, the multiplicative factor becomes an additive factor, because that's how logs work: log(Y*error) = log(Y) + log(error). The percent error therefore becomes the same additive error, regardless of the value of Y. So your analyses work, because your non-uniform residuals become uniform. This feature of log transformation is useful for analysis of most types of athletic performance and many other measurements on humans.

Percent Effects from Log-Transformed Variables

If the percent error in a variable is similar from subject to subject, it's likely that treatment effects or differences between groups expressed as percents are also similar from subject to subject. It therefore makes sense to express a change or difference as a percent rather than as a raw number. For example, it's better to report the effect of a drug treatment on high-jump performance as 4% rather than 8 cm, because the drug affects every athlete by 4%, but only those athletes who jump 2 m will experience a change of 8 cm. In such situations, the analysis of the log-transformed variable provides the most accurate estimate of the percent change or difference. Make sure you use natural logs, not base-10 logs, then analyze the log-transformed variable in the usual way.

Suppose you end up with a difference of 0.037 (you'll often get small numbers like this).

Now multiply it by 100, and hey presto, the difference in your mean is 3.7%. Actually, multiplying by 100 is an approximation, and it's near enough only for differences <0.05 (5%). The exact percent difference is given by $100(e^{diff} - 1)$, where e is exponential e and diff is the difference provided by the analysis of the log-transformed variable (see explanation box). This formula simplifies to 100diff only for diff <0.05.

I find it easier to interpret the diffs (differences or changes) in a log-transformed variable if I use 100x the log of the variable as the log transformation. That way the diffs are already approximately percents. For example, instead of getting a change of 0.037, you will get 3.7, which means approximately 3.7%. To convert this diff to an exact percent, the formula is $100(e^{diff/100} - 1)$, obviously! A diff of 3.7 is really $100(e^{3.7/100} - 1) = 3.8\%$.

Explanation of $100(e^{diff} - 1)$ and 100diff

If Z = log(Y) and Z' = log(Y'),
then diff = Z' - Z = log(Y') - log(Y) = log(Y'/Y).
But Y'/Y = 1+(Y'-Y)/Y = 1+(percent change in Y)/100.
Therefore e^{diff} = Y'/Y = 1+(percent change in Y)/100.
Therefore percent change in Y = $100(e^{diff} - 1)$.
For small diff, $e^{diff} = 1 + diff$,
so percent change in Y is approximately 100diff.

It's easy to get confused when the percent change is large. For example, a change of 90% means that the final value is (1 + 90/100) or 1.90 times the initial value. A change of 100% therefore means that the final value is (1 + 100/100) or 2.0 times the initial value. A 200% increase means that the value has increased by a factor of 3, and so on. A negative percent change can also be confusing. (In a previous version of this paragraph, my interpretation of large negative changes was wrong!) A change of -43% means that the final value is (1 - 43/100) or 0.57 times the initial value. An 80% fall means that the final value is only 0.20 times the initial value, and so on.

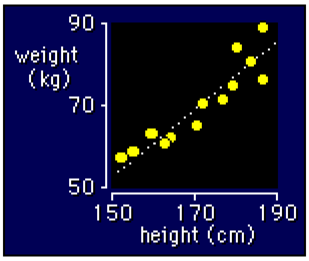
When variables need log transformation to make them normal, how do you represent their means and standard deviations? I think a hybrid approach is best. Convert the mean of the log-transformed variable back to raw units using the back-transformation $Y = e^{mean}$ (if your transformation was Z = logY) or $Y = e^{mean/100}$ (if you used Z = 100logY). Keep the standard deviation as a percent variation or coefficient of variation (CV). Calculate this CV in the same way as for differences or changes in the variable: if SD is the standard deviation of the log-transformed variable, the approximate CV is simply 100SD, and the exact CV is $100(e^{SD} - 1)$. If you used 100log for your transformation, the approximate CV is simply the SD itself, and the exact CV is $100(e^{SD/100} - 1)$.

You can convert the CV into a raw standard deviation, but it's messy and I don't recommend it. Back-transforming the SD as e^{SD} is incorrect. Instead, you have to show the upper and lower values of the mean ± standard deviation as $e^{mean + SD}$ and $e^{mean - SD}$. With a bit of algebra, you can show that $e^{mean + SD}$ is equal to the back-transformed mean times 1 + CV, and $e^{mean - SD}$ is the back-transformed mean times 1/(1 + CV). Hence a CV of, say, 23% represents a typical variation in the mean of ×1.23 through ×1/1.23. As I explain on the page about calculating reliability as a CV, it's OK to write ±CV, provide you realize what it really means.

CAUTION. With log and other non-linear transformations, the back-transformed mean of the transformed variable will never be the same as the mean of the original raw variable. Log transformation yields the so-called geometric mean of the variable, which isn't easily interpreted. Rank transformation yields the median, or the middle value, which at least means something you can understand. The square-root and arcsine-root transformations for counts and proportions yield goodness-knows-what. Usually it's the effects you are interested in, not the mean values for groups, so you don't need to worry. But if the means are important, for example if you want the true mean counts of injuries to come out of your analysis, you will have to use a cutting-edge modeling approach that does not require transformation, such as binomial regression.

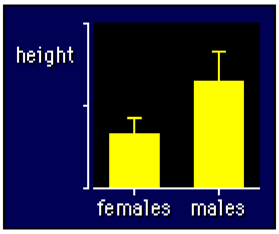
If you're graphing means and standard deviations of a variable that needed log transformation, use a log scale on the axis. Here's how. Plot the values you get from the log-transformed data without back-transformation, but delete the tick marks and put new ticks corresponding to values of the original raw variable that you would normally show on a plot. (You will struggle to understand what I am getting at here. Persevere. And if you use Excel to do your graphs, paste the graph into Powerpoint and do the editing there.) The error bar or bars go onto the plot without and fiddling. In fact, you can put the error bar anywhere on the axis.

More Examples of Log Transformation



Let's get back to the issue of goodness-of-fit with log transformations. In a previous example with weights and heights (see the figure at right), it's clear that people's weights get more variable for heavier people--quite reasonable when you think about it--so taking logs of the weight would be a good thing to try. When you fit a straight line, log transformation of the independent variable may also remove the "dip" in the residuals that we saw with this example on the previous page. So taking logs of the heights and the weights in the above example would make the model much fitter!

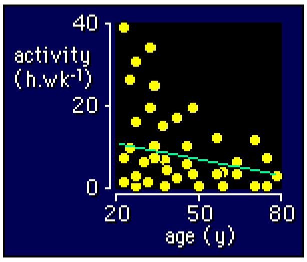
Many relationships that have a curve in them respond well to log-log transformation. To get technical, all models of the form Y = aX^n convert to simple linear models when you take logs: logy = loga + nlogX. The relationship between weight (Y) and height (X) is a particularly good example. The value of the parameter n is given by the slope of the log-log plot, and it is about 1.7, or nearly 2, which is why we normalize body weights by dividing by the height squared to get the so-called *body mass index.* It would be better to divide by height to the power of 1.7, but that's another story.



Now check out the figure at right, from our example of the effect of sex on height. Do you think there's a need for log transformation here? You bet! Just look at the differences in the standard deviations on the bar graph: the males have a bigger standard deviation and a bigger mean, so log transformation is indicated (provided the means and standard deviations are pretty-much in proportion). Analysis of log-transformed height will give the difference between the females and males as a percent. You can also analyze these data without transformation by using the t test with unequal variances. What you will get then is the absolute difference in height between the average female and the average male. There's nothing wrong with that, if it's what you want.

Another case for some sort of transformation is where the standard deviation is about the same size as, or even bigger than, the mean. This sort of thing sometimes happens when variables have very skewed distributions. Example: the level of deliberate physical activity in adults, where you have most people hovering around zero hours per week, and the rest doing up to 10 hours a week or even more. So the mean ± SD might be 0.7 ± 1.8 hours per week. It doesn't mean that some people are doing negative hours per week! For such awful data we could use rank transformation: see the next page.

Rank Transformation: Non-Parametric Models



Take a look at the awful data on the right. It's clear that activity is greater at younger ages, so you want an outcome statistic to summarize that important finding. Fitting a line or a curve would do the trick Let's keep it simple and fit a line, in the usual least-squares way. The slope of the line is what you want, and its value is something like 1.0 hours reduction in activity per decade of age. You also want confidence limits or a p value for the slope. So how do you go about it?

The least-squares approach gives you confidence limits and a p value for the slope, but you can't believe them, because the residuals are grossly non-uniform. You don't have to plot residuals vs predicteds to see that--just look how much bigger the spread is about the line for younger subjects. The bigger spread occurs for bigger values of activity, so that's a strong indication for log transformation. Unfortunately you can't take logs here, because some subjects have zero activity, and you can't take the log of zero. (You get minus infinity!) I've seen people attempt to solve this problem by adding a small number, such as 1 hour, to everyone's activity, then taking logs. I don't really agree with this approach, because it means changing some of the data.

What to do? The best approach is to use bootstrapping, but that's a big ask for most researchers. The next-best approach is to "take ranks" rather than to take logs. In other words, **rank transform** the dependent variable. What does *that* mean, exactly? Simply that you arrange the values of activity for every subject in rank order, then assign the smallest a value of 1, the next smallest a value of 2 etc., etc. Now do your modeling in the normal way, but use the variable representing the rank as the dependent variable. You have just performed a**non-parametric analysis**--more about that below. Rank transformation usually results in uniform residuals (same scatter for any age) for the rank-transformed variable. You should check that they are indeed uniform. If they aren't, you are no better off.

Confidence Limits via the P Value for the Rank-transformed Variable

But wait! You originally wanted confidence limits for the slope of the line of activity vs age. The analysis of rank-transformed activity will give you confidence limits for a slope, but it will be the slope of the ranks of the activity, not the slope of activity itself. The slope and its confidence limits in rank units are just about impossible to interpret, and that's true of all analyses involving rank transformation. So now what? Well, I've agonized over this one for some years, and I now have the solution. You probably won't find this one anywhere else, but I think it's the way to go.

The analysis of the rank-transformed variable gives you a p value for the outcome statistic, in this case the slope of the line. You now assume that the p value applies to the slope of the line you got by analyzing the untransformed data. Next assume you have a sample of sufficient size that the central limit theorem comes into action to give you a normal sampling distribution for your slope. Therefore combine the p value and the slope to calculate the confidence limits for the slope, using the spreadsheet for confidence limits. Done!

Confidence Limits via Cohen's Effect-size Statistic for the Rank-transformed Variable

Another approach to getting confidence limits for the outcome statistic with a rank-transformed variable is to calculate a Cohen-type effect size (change in the mean divided by a standard deviation). Again, you won't see this approach anywhere else, but again, it works well. The only drawback is that most folks still aren't used to Cohen effect sizes. Let me remind you that this outcome statistic is ideal for studies of average subjects in a population, but it's no good for studies of performance of competitive athletes.

I'll start with the simple case of the difference in the mean of two groups: for example, the mean heights of females vs males. Rank-transform height without regard to sex, then do an unpaired t test (the unequal variances version) on the rank-transformed height. Calculate the effect size for the difference between the means of two groups by taking the difference between the means of the ranked height, then dividing by the average standard deviation of the ranked variable within the two groups. (You might have to generate the average standard deviation yourself, if the t test doesn't give it to you. Average the variances, not the standard deviations, then take the square root. If you've done an ANOVA rather than a t test, the root-measn square error is the average standard deviation you want.) Divide the upper and lower confidence limits of the difference in the mean by the average standard deviation to get approximate confidence limits for the effect size.

I have checked by simulation that this Cohen-type estimate is unbiased for normally distributed variables. In other words, on average it gives the same effect size as the analysis of an untransformed normally distributed variable. Cool! Strictly speaking, the confidence limits for the effect size should be derived using something called the non-central t statistic, to take into account uncertainty in the standard deviation. With a reasonable sample size you don't have to worry about this detail. One day soon I will provide a spreadsheet to do the calculation.

You can take a similar approach to express the slope of a straight line in effect-size units, when the straight line comes from the rank-transformed variable. In this case you divide the slope and its confidence limits by the standard error of the estimate (or the root-mean square error) from the regression analysis of the rank-transformed variable. In the above example, you might get something like 0.7 Cohen units per decade, and whatever confidence limits. If you are interested in the difference over a decade, 0.7 would be a moderate effect on thescale of magnitudes. Over two decades, the difference would be 1.4 units, which would be large.

It's also possible to avoid dealing directly with the slope to express the magnitude of the effect of X on Y (here age on activity). Just rank-transform the Y, then calculate the correlation coefficient and its confidence limits. Interpret the magnitudes using the scale of magnitudes. This is the simplest and possibly the best method of all, provided you aren't particularly interested in the magnitude of the effect for different differences (sic) in X.

Non-parametric Analyses

Your stats program will probably convert values of a variable to ranks with the click of a mouse. Or if you select **non-parametric analysis** in the stats program, it will do the transformation without you realizing it, because *a non-parametric analysis is a parametric analysis on a rank-transformed variable.* The term *non-parametric* refers to the fact that you are no longer modeling, for example, the means of your groups, because that information is lost when you take ranks. But you are still performing a parametric analysis, so the term is a misnomer.

The names statisticians use for non-parametric analyses are misnomers too, in my opinion: Kruskal-Wallis tests and Kolmogorov-Smirnov statistics, for example. Good grief! These analyses are simple applications of parametric modeling that belie their intimidating exotic names. But there is one name you need to know: a non-parametric correlation coefficient is called a **Spearman correlation coefficient**. Most stats programs will calculate this at the click of a mouse, but note that it is derived by ranking *both* variables. Most of the time you need to rank only the dependent variable, not the independent variable too.

Actually, some non-parametric analyses come close to being truly non-parametric--things like the signed rank-sum test. But even here you are modeling probabilities, so it's still debatable whether they should be called non-parametric. The simplest example is the **sign test**. It's worth a paragraph, because it tests your understanding of p values. Here's the problem: what's the minimum number of *all positives* or *all negatives* that need to come up for you to decide whether there's a significant difference? For example, if you have a group of seven athletes, and they all get better after you've done something to them, is that statistically significant? (Let's leave aside the question of a control group.) Look at it from the point of view of tossing a coin. If you toss a coin several times and get all heads or all tails, how many tosses does it take before you decide the coin is fishy? Let's start with three tosses. The chance of getting three heads *or* three tails is 0.5*0.5*0.5 + 0.5*0.5*0.5, i.e. 0.25, so three isn't enough. Four heads or four tails in a row occurs with a probability of 0.125, and so on until we get to six in a row (p = 0.03) and eight in a row (p = 0.008). So you need six positives or negatives in a row to declare significance at the 5% level, and eight at the 1% level.

Here's a good final question. Why not play it safe with non-uniform residuals by doing all analyses after rank transformation? Hmmm... Well, rank transformation throws away some information, so it can't be as good as using the original variable. But the loss of information only starts to bite when you have small sample sizes. In other words, with small sample sizes, non-parametric analyses are less likely to detect effects, or the power is reduced, or the confidence intervals are wider. So use parametric analyses wherever possible. Besides, it's easier to interpret the outcomes from a parametric model.

Ordinal Dependent Variables

Outcome variables with only a few possible values, such as 1, 2 or 3, need special treatment. Variables like this are called **ordinal**, because they indicate an ordering of responses. They crop up often in questionnaires, where people have to tick one response from a choice like*less, the same,* or *more.* The choices make up a so-called **Likert scale**. We use integers to number and record the responses, but the responses aren't integers. All the integers do is indicate order in the levels of the response. That's why ordinal variables are neither numeric nor nominal.

If we treat the ordinal variable as nominal, we lose the information about the ordering. But if we try to treat it as a numeric variable, we might violate one or more of the assumptions we make when we calculate confidence limits or p values. I used to think such violations were a frequent problem, but it turns out that they are rare. Most of the time you can use t tests for comparisons of groups, and if you are fitting lines or curves, you can use rank transformation to get rid of non-uniform residuals.

As I pointed out earlier, the rare situations occur when the responses of a Likert-type variable are almost all stacked up on the top or bottom level of the scale. Rank transformation doesn't work in these situations either, because the rank-transformed variable has the same problems as the raw variable. In such situations, the only way forward is to model the **probabilities** of responses being at each level. It's called **logistic regression**. You've met this before as categorical modeling. The only difference is that logistic regression takes into account the fact that the different levels of the outcome are ordered, whereas plain old categorical modeling treats the outcome as a nominal variable, without any implied order in the levels of the variable.

Logistic regression gives you a way out when you have a variable, like habitual intense physical activity, with almost everyone on zero and a smear of values for the rest. Split the values of your outcome variable into a number of ordered levels (the first being zero, of course), then do a logistic regression on those levels. You are actually transforming the ugly continuous variable into a more manageable Likert-scale (ordinal) variable.

Counts and Proportions as Dependent Variables

If your dependent variable represents a **count** (e.g., the number of injuries in different sports) or a **proportion** (e.g., the percent of Type I muscle fibers in a muscle biopsy from different athletes), analysis can be a challenge. Once again, the problem with the usual analyses is the possibility of violation of one or more of the assumptions we have to make when calculating confidence limits or the p value.

What I said on the last few pages about t tests of ordinal variables and t tests of Likert-scale variables applies also to counts: t tests are usually OK, and they will fall over only when you have a small sample size and more than 70% of your subjects score zero counts (because then the sampling distribution of the difference between the means won't be close enough to normal).

When you are fitting lines or curves, you also have to worry about non-uniformity of residuals. With counts, this worry is very real, because the variation in a given count from sample to sample depends on how big the count is. For example, the typical variation (standard deviation) in a count is usually simply the square root of the count, so a count of about 400 injuries varies typically by ±20, whereas a count of about 40 injuries varies typically by ±6. I hope it's obvious that the residuals for injury counts of 400 will therefore be much larger than those for counts of about 40. Rank transformation would fix these non-uniform residuals, but better approaches are available: **binomial regression**, **Poisson regression**, **square-root transformation** and **arcsine-root transformation**.

Binomial and Poisson Regression

When counts have a smallish upper bound (e.g., the number of injured players in a squad of 24 is at most 24), the counts from sample to sample vary according to what is known as a **binomial distribution**. When the upper bound is very large compared with the observed values of the count (e.g., the number of spinal injuries in American football each year), the counts have a **Poisson distribution**. With a good stats program, you can dial up an analysis that uses either of these distributions. The result is a **binomial regression** or a **Poisson regression**. In the Statistical Analysis System, you can do these analyses with Proc Genmod. *Genmod* stands for **generalized linear modeling**, which is an advanced form of general linear modeling that allows for the properties of non-normally distributed variables such as counts and proportions based on counts.

Don't feel intimidated by *binomial* and *Poisson.* Are you happy with the notion that the values of most variables have the bell-shaped normal distribution? OK, counts or proportions of something don't have the normal shape when the counts are small, so we need different mathematics to describe their shapes, and different names for them. As counts get larger, the shapes of the binomial and Poisson distributions tend towards the normal shape. You still have the problem of non-uniform residuals, though, because the variability from observation to observation for larger counts is more (in absolute values) or less (in percentage terms) than for smaller counts. Binomial and Poisson regressions and other forms of generalized linear modeling take care of the non-uniformity. For more on generalized linear modeling, in particular the specification and use of distributions and link functions, read this message I sent to the Sportscience email list in July 2004. .

Square-root and Arcsine-root Transformation

One way to deal with non-uniform residuals is to transform the variable. We've seen that log transformation works for some variables, and rank transformation works for most variables as a last resort. Is there a transformation for **counts** that will allow us to use normal analyses instead of binomial or Poisson regression? Yes, provided you aren't close to some upper bound in the counts, just use the **square root** of the counts in the usual analyses. When you've derived the outcome statistic and its confidence limits, assess their magnitudes with Cohen's or my scale of effect sizes, as I explained for rank transformation. You can't back-transform an effect (such as a difference between means) into a count by squaring it, but you can get a feel for the magnitude as a count relative to the mean by adding the value of the effect appropriately to the mean of the square-rooted counts, then squaring it. Square the mean for comparison. Add each of the confidence limits of the effect to the square-rooted mean and square it to get a feel for the precision of the magnitude.

Read the cautionary note about how the value of a back-transformed mean is not the same as the mean of the raw variable. For a simple example, imagine you have a team with only one injury this season and another team with nine injuries. The mean of the raw number of injuries is (1+9)/2 = 5. But the mean of the root-transformed injury counts is (1+3)/2 = 2, and when you square 2 to back-transform it you get 4!

Proportions require an exotic transformation called **arcsine-root**. To use this transformation, express the proportion as a number between 0 and 1 (e.g., 210 Type I muscle fibers in a biopsy of 542 total fibers represents a proportion of 210/542 = 0.387). Now take the square root and find the inverse sine (arcsine) of the resulting number; in other words, find the angle whose sine is the square root of the proportion. (The angle can be in degrees or radians, where 360 degrees is 2 pi radians.) Use that weird variable in your analysis, but*weight each observation by the number in the denominator of the proportion,* to ensure that the residuals in the analysis are uniform. You'll have to read the documentation for your stats program to see how to apply a weighting factor. To gauge magnitude of effects with an arcsine-root transformed variable, apply the Cohen or Hopkins scale, as explained for rank transformation. The appropriate standard deviation is the root-mean square error from the analysis of the transformed variable, because this error should take into account the weighting factors. As is the case for counts, back-transformation of the observed effect works only if you add the effect appropriately to the mean before taking its sine and squaring it. Multiply the result by 100 if you want it as a percent. Do the same with the confidence limits.

The square root and arcsine-root transformations work well even for low counts or zero proportions. As with ordinal variables, you'll get into trouble only with small sample sizes when more than 70% of your subjects have a score of zero or a proportion of zero. Then you*have* to use binomial or Poisson regression.

Phew! The square-root and arcsine-root approaches are complex. I recommend that you come to terms with a stats package that offers binomial and Poisson regression or generalized linear modeling.

Linear and Non-Linear Models

A linear model is one in which the independent variable is added or multiplied together with the parameters. A non-linear model has exponents, logarithms, or other complicated functions of the independent variable and parameters.

Some non-linear models can be reduced to linear models to make it easier to do the fitting. For example, if your Y values curve upwards like a simple quadratic in relation to your X values, then it might be appropriate to fit $Y = aX^2$. You could reduce this model to a linear one simply by introducing a new variable called S (say), which has the same values as X^2. You then fit the linear model Y = aS. Some stats programs generate these new variables automatically when you fit quadratics, cubics, or other higher order **polynomials**. More on theseshortly.

Most non-linear models can't be reduced to a simple linear model in this way. But a good stats program can fit non-linear models as complex as you like. All you do is choose the mathematical form of the model; the stats program then calculates the values of the parameters that give the best fit to your data, as explained earlier. The usual method is to minimize the sum of the squares of the residuals.

Note: Whatever model you fit, you should check visually that it really does fit the trend in the data. In other words, plot the curve and see if your points are fairly evenly scattered about it. Or get the stats program to plot residuals against predicteds from the model, then eyeball the plot to make sure you haven't got bad residuals.

COMPLEX MODELS

I've split complex models up into two main groups: **more than one predictor (independent) variable**, starting on this page, **and repeated-measures models** later on. In between there's a short page on **more than one dependent variable**, and **variables of uncertain status**. All details about model fitting on the previous few pages apply to all these models.

More Than One Predictor (Independent) Variable

In other words, models like :

weight <= height sex

This model is called an **analysis of covariance** (ANCOVA) when one predictor variable is numeric (height) and the other is nominal (sex). *Covariance* refers to the fact that height "co-varies" with the dependent variable, so height is also known as a **covariate**. Other names for models with two or more predictor variables include **multiple linear regression** when all variables are numeric and **two-way analysis of variance** (or three-way ANOVA etc) when all are nominal. In essence they are all the same. Before we go into each model in detail, let's understand what it means to have more than one predictor variable. Let's stay with the above example.

What the Model Means
It's easiest to think about the model as a tool for predicting weight when you know a person's height and sex. If there IS a relationship between weight and height, then knowing a person's height will tell you something about his or her weight. Similarly, if there IS a relationship between weight and sex, then knowing a person's sex will also allow you to say something about her or his weight. And if you know both height and sex, you'll be able to be even more specific about weight. So that's the question that the overall model poses: **what do the predictor variables taken together tell you about the dependent variable?**

Stats programs can calculate the usual goodness-of-fit R^2 for the model, which you can interpret as a measure of *how much* the independent variables tell you about the dependent variable. In formal terms the R^2 is the percentage of the variance in the dependent variable explained or predicted by the independent variables. You can also get a test statistic for the full model and its associated p value. You could use the p value to work out confidence limits for the overall R, but otherwise these statistics aren't worth worrying about. Much more important are effects derived from the predictor variables, as I will describe now.

"Controlling" for Something
The overall relationship is seldom the main focus when you have more than one predictor variable in the statistical model. Instead, these models are used to address a much more important question: what is the effect of something when we **take into account** something else? It's such an important concept, statisticians have some jargon for it: they talk about **controlling** for something, **adjusting** for something, or **partialing** something out. For example, what is the effect of sex on weight, when we take height into account? Think about it. Boys are heavier than girls, but boys are taller than girls, and taller people are heavier, so if we take into account the difference in height between boys and girls, is there any "real" difference in weight between them? A trivial question here maybe, but not if your outcome variable is an athlete's performance or health, and you control for time spent training before you look at the effect of sex or sport or region or whatever. And of course, it's also important to know about the effect of training on performance or injuries when you take into account differences between sexes or sports or regions.

What do we really mean when we control for height or take height into account in the comparison of the weights of boys and girls? Simply this: if boys and girls had the same height, what would be the difference in weight? And that's exactly what the statistical analysis tells us: it gives us **the effect of a predictor with all other predictors held constant**. When you do your usual estimates or contrasts for the effects you're interested in, or inspect the solution of the model, the answers you get are automatically adjusted for the presence of all the other predictor variables, as if they are all set to come constant value. For example, you get the difference in the mean weight of boys and girls who have the same (mean) height. Note that the analysis automatically controls for every predictor variable, so you can also address the question: what's the effect of height on weight when you take sex into account? What you get from the analysis for this question is the average slope of the lines for the boys and the girls, as if there was an equal number of boys and girls in the study. I'll delve into these issues more on the next page.

Why do the estimates for a given predictor represent the effect of the predictor with all other predictors in the model held constant? I'm not sure of the best way to answer this question. I've satisfied myself by considering that a linear model with two numeric predictor variables represents a plane in 3-D space. The stats program finds the least-squares plane of best fit. With a bit of thought and 3-D doodling I was able to see how the value of the coefficient of each variable is the "slope" for that variable with the other predictor variable held constant.

Mechanism Variables and Confounders

In the above example, suppose we adjust or control for height and find no substantial difference in the mean weight of boys compared with that of girls. Is it therefore reasonable to say that differences in height are responsible for the differences in weight between boys and girls? Yes! In fact, I call height a **mechanism variable** for the effect of sex on weight: sex affects height, and height affects weight. Any variable (here, height) on the causal path between the predictor (sex) and the dependent (weight) will reduce the effect of the predictor on the dependent when that variable is included as a covariate in a multiple linear regression. So if you see such a reduction, the covariate could be a mechanism variable. If you don't see a reduction, the covariate can't be a mechanism variable. A reduction in the effect is necessary but not sufficient for the covariate to be a mechanism.

Some researchers also call height a **confounding variable** or a **confounder** in the relationship between sex and weight. When you use the word *confounder* to describe height, you are implying that it sort-of makes the difference between boys and girls seem bigger than it really is. Boys are heavier than girls, of course, but height is confounding (or even compounding) the difference. There might be no difference when you take account of height. In fact, girls might even be heavier than boys. Fair enough, but the word *confounder* should be reserved for a different kind of covariate, one that has or could have a causal effect jointly on the predictor and the dependent. Let's consider another example to make the point clear. Consider the effect of physical activity on health in a cross-section of the population. Do the analysis without regard to the age of the subjects and you will find a really strong relationship. Cool, jobs for exercise professionals! Now control for age and you will find the relationship gets a lot weaker. Curses! It's likely that age is the real cause of most of the relationship between activity and health: age reduces physical activity and age reduces health. We say that the effect of physical activity on health is confounded by age. It's only when we control for age that we see the effect of differences in activity on the health of people of the same age.

What happens in the above example if we make age the predictor variable and physical activity the covariate? Age on its own will have a strong effect on health, but control for physical activity and you will find the relationship gets a lot weaker. So, you would be justified in regarding physical activity as a possible mechanism for the effect of age on health. Wow, that's cool again! Whether the effect of physical activity on health is really causal or just coincidental cannot be resolved with cross-sectional data. You have to do interventions and a repeated-measures analysis to sort that out. I explain how to include a mechanism variable as a covariate in such analyses later on.

Interactions
I now have to introduce you to another fearful challenge: **interactions**. Let's just have a look first, then we'll climb it in several ways on the next few pages.

Height has an overall effect on weight, and sex has an overall effect on weight. But maybe the effect of height on weight is a bit different for boys than for girls: maybe being taller has a bigger effect on weight for boys than for girls. We show that in the model with the so-called interaction term, which is represented by multiplying height and sex together:

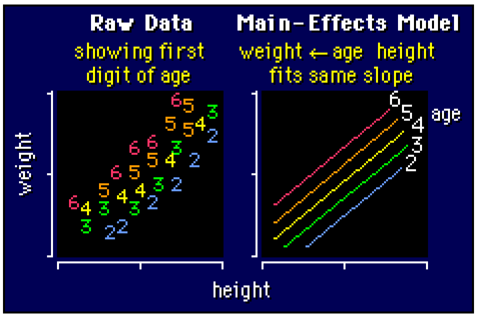
weight <= height sex height*sex

This will all make sense when we deal with the specific models. Meanwhile one more bit of jargon. Height and sex are called **main effects**, to distinguish them from the interaction term. When you have more than two main effects, you can have more than one interaction. When you have all the different combinations of the effects, including the interactions, you have what's called a **full model**.

A Warning!
There are several traps for the unwary when you have more than one predictor variable. **Read the following pages carefully** or you might jump to wrong conclusions with your data.

Multiple Linear Regression

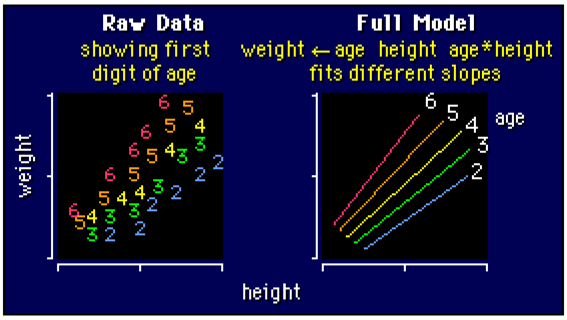
model: numeric <= numeric1 numeric2... + interactions
example: weight <= height age height*age



The example shows weights and heights of a sample of people aged between 20 and 60. Each person is represented by a number, which is the person's age rounded to the nearest decade (2 = 15-24 years, 3 = 25-34 years, etc.). Look closely at the way the numbers are distributed. What would you conclude about the effect of age on weight, for any given height? Right! People get heavier as they get older.

Multiple linear regression is the model to use when you want to look at data like these, consisting of two or more numeric independent variables (height, age) and a numeric dependent variable (weight). In this first example, the only effect of age is to produce a uniform increase in weight, irrespective of height. It's just as correct to say there is a uniform increase in weight with height, irrespective of age. These interpretations come straight from the model. Or you can look at the graphical interpretation and think about the effect of age as altering the intercept of the weight-height line in a uniform way. But what about when there's an interaction?

Interpreting the Interaction Term



As you can see, the effect of an interaction is to make different slopes for different ages. The slopes change in a nice linear way with increasing age, just as the intercepts did (and still do). In the example, I've given older people a greater weight for a given height than younger people, which is not necessarily realistic. Real data would certainly not show such clear-cut effects of either height or weight, anyway.

It's one thing for me to show you a clear-cut example with colors for the different ages. It's quite another matter for you to interpret real data, without a colored graph. If you get a substantial interaction with your data, I suggest you look at the values of the parameters in the solution. Use them to work out how your outcome variable is affected by a range of values of the independent variables. That's the only way you will sort out what's going on.

By the way, for publication you would not plot them as I have shown here. In fact, generally you don't plot the data for linear regressions, be they simple or multiple, unless the data show interesting non-linear effects.

Paradoxically Insubstantial Effects

On the previous page I pointed out how one independent variable can make another seem insubstantial in an ANCOVA. The same is true here. It's important, so let's take an example.

Suppose you want to predict running-shoe size (dependent variable) from an athlete's height and weight. These two variables are well correlated, but let's assume the correlation is almost perfect. When two variables have an almost perfect correlation, it means they effectively measure the same thing, even if they are in different units. Now let's put them both into the model. Will weight tell you anything extra about shoe size, when height is already in the model? No, because weight isn't measuring anything extra, so it won't be substantial in the model. But hey, height won't be substantial with weight in the model, for the same reason. So you have the bizarre situation where neither effect is substantial, and yet both are obviously substantial! If you didn't know about this phenomenon, you might look at the p values for each effect in the model, see that they are both greater than 0.05, and conclude that there is no significant effect of either height or weight on shoe size.

The trick is to look at the p value for the whole model as well. None of the effects might be significant, but the whole model will be very significant. And you should always look at the main effects individually, as simple linear regressions or correlations, before you go to the multiple model. You'd find they were both substantial/significant.

So in this example, would you use both independent variables to predict shoe size? Not an easy question to answer. I'd look to see just how much bigger the R^2 gets with the second independent variable in the model, regardless of its statistical significance. More on this, next.

Now for two important applications of multiple linear regression: **stepwise regression**, and on the next page, **polynomial regression**.

Stepwise Regression

model: numeric <= numeric1 numeric2 numeric3...
example: competitive speed <= a set of fitness-test variables
No figure is needed for this one. No interactions either, thank goodness! Numeric1, numeric2, and so on are independent variables, and you try to find the best ones for predicting your dependent variable.

An obvious example is where your dependent variable is some measure of competitive performance, like running speed over 1500 m, and your independent variables are the results of all sorts of fitness tests for aerobic power, anaerobic power, and body composition. What's the best way to combine the tests to predict performance? An interesting and possibly useful question, because you can use the answer for talent identification or team selection. (Why not use the 1500-m times for that purpose? Hmmm...) Anyway, in stepwise regression the computer program finds the lab test with the highest correlation (R^2) with performance; it then tries each of the remaining variables (fitness tests) in a multiple linear regression until it finds the *two* variables with the highest R^2; then it tries all of them again until it finds the *three*variables with the highest R^2, and so on. The overall R^2 gets bigger as you add in more variables. Ideally of course, you hope to explain 100% of the variance.

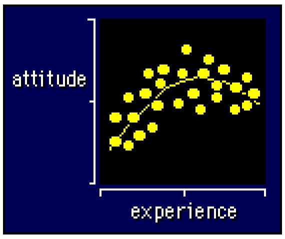
Now, even random numbers will explain *some* of the variance, because you never get exactly zero for a correlation with real numbers. So you need an arbitrary point at which to cut off any further variables from entering the analysis. It's done with the p value, and the default value is 0.15. When a variable enters the model with a p value >0.15, the stepwise procedure halts. You'd hardly call a p value of 0.15 significant, but it's OK if you're using stepwise regression as an exploratory tool to identify the potentially important predictors.

The question of what variables you finally include for your prediction equation is not just a matter of the p values, though. You should be looking at the R^2 and deciding whether the last few variables in the stepwise analysis add anything worthwhile, regardless of their significance. If the sample size isn't as big as it ought to be, there's a good chance that the last few variables will contribute substantially to the R^2, and yet not be statistically significant. You should still use them, but knowing that their real contributions could be quite a bit different.

OK, what is a worthwhile increase in the R^2 as each variable enters the model? Take the square root of the total R^2 after each variable has entered, then interpret the resulting correlations using the scale of magnitudes. If the correlations are in the moderate-large range, an increase of 0.1 or more is worthwhile. If the correlation is in the very large to almost perfect range, then smaller increases (0.05 or even less) are worthwhile, as I explain later.

Finally, a warning! If two independent variables are highly correlated, only one will end up in the model with a stepwise analysis, even though either can be regarded as predictors. Go back up this page for the reason. And as discussed in the previous paragraph, the decision to keep both in the model depends on the R.

Polynomial Regression



The figure shows data that lend themselves to fitting a polynomial. As you can see, there is a so-called **curvilinear** trend in an outcome measure when it is plotted against an independent variable. In the example the dependent variable is some sort of attitude in athletes, but it could be performance or just about anything. The independent variable is often some measure of time--here it's years of competitive experience. Maybe the athletes start to go stale in this sport after a certain time, and you'd like to be able to say so, quantitatively, with confidence limits. Polynomial regression is the answer for these data and for most curvilinear data that either show a maximum or a minimum in the curve, or that could show a max or min if you extrapolated the curve beyond your data. Log transformation is more likely to fit a curve that shows an ever-increasing or ever-decreasing trend, although often it makes sense to fit a polynomial to log-transformed data.

Often the different points come from the same subjects, especially when time is the independent variable. You can still fit polynomials to such data, but you have to use repeated-measures models. I deal with repeated-measures polynomials later, but the interpretation of the numbers describing the shape of the curve is the same, and I deal with that here.

A Simple Polynomial

model: numeric <= numeric numeric^2 numeric^3...

example: attitude <= experience experience^2

Notice the subtle difference from the model for multiple linear regression on the previous page. Here the numbers 2, 3... represent **powers** of the **same** variable. It might be easier to see if I write: $Y <= X \;\; X^2 \;\; X^3$... The stats program fits the polynomial $Y = a + bX + cX^2 + dX^3$... to the data. Polynomials are a special case of the more general non-linear models. Check that page out again right now!

For data that are shaped like a parabola, you probably won't need more than a quadratic model ($Y <= X \;\; X^2$). If the curve is trends up again at one end, you'll need a cubic model. Curves with multiple kinks need even higher-order terms. It's rare to go past a quadratic, though.

When you fit a model like $Y <= X \;\; X^2$, the stats program finds the best quadratic curve to fit the data. In other words, it will find the best values for the **coefficients** (or parameters) a, b and c in the equation $Y = a + bX + cX^2$. The value of a represents the overall position of the curve up and down the Y axis; for example, an increase of 1 unit in a shifts the whole curve up the Y axis by 1 unit. The value of b represents the amount of overall upward or downward linear (straight-line) trend in the values of Y as you move along the X axis; in other words, if you draw a straight line to fit all the points well, b is the slope of the line, which is the same thing as the increase (or decrease, if b is negative) in Y for each 1-unit increase in X. For the data in the figure, b would represent the change in attitude per year of experience. The value of c represents the amount of curvature in the data; in the present example, c would be negative, because the parabola is upside down. I find it easier to interpret c visually if I transform the X values so they range from -1 to +1. If I then fit a curve with this new independent variable, the value of c that I get is about the amount that the values of Y sit above (or fall below, if c is negative) a straight line at either end of the X range.

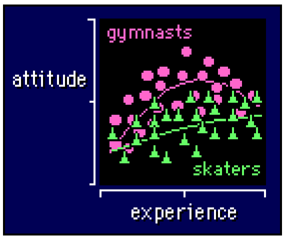
Remember that you can derive these coefficients or parameters as raw values, as percents, and as normalized regression coefficients, just like the slope in a simple linear regression. Make sure you interpret their magnitudes and their confidence limits!

Caution! The linear term in a quadratic polynomial represents the overall effect as you go from low to high values of the independent variable. The quadratic term doesn't impact this overall effect--in fact, including the quadratic when there is curvature in the trend will make the estimate of the linear term more precise. But if you include a *cubic* term in the polynomial, the cubic also contributes to the overall effect of going from low to high values of the independent variable. **This extra contribution of the cubic makes it impossible to interpret the linear term as representing the difference between low and high values of the independent variable.** This problem is particularly important when you are using polynomial contrasts in a repeated-measures analysis, where the independent variable is time or trial number. The easiest way to avoid the problem is to avoid including a cubic or quintic in the polynomial. If you do include these higher order terms, and you want an estimate of the difference between the effect of low and high values of the independent variable (e.g., first test vs last test), you will have to derive an estimate for the high minus the low values.

Don't forget that you can assess the contribution of each term of the polynomial to the variance explained (R^2) by the model. If your stats program doesn't give you the R^2 for each term, find the total sum of squares and the sums of squares for each effect in the output, then calculate the R^2 for the quadratic term by dividing its sum of squares by the total sum of squares, multiplied by 100 to convert it to a percent. Phew! Interpret the R^2 by taking its square root and working out the confidence limits of the resulting correlation, as described earlier.

A Polynomial With a Nominal Effect

The next figure shows an extension of the above model to test for differences between two sports. Let's build up the model term by term. We'll need sport as a main effect, to see how much overall difference there is in the mean attitude for the two sports:



attitude <= sport

The main trend with experience is linear, and we want to know about the differences in the slopes, so we need a full ANCOVA model:

attitude <= sport experience sport*experience

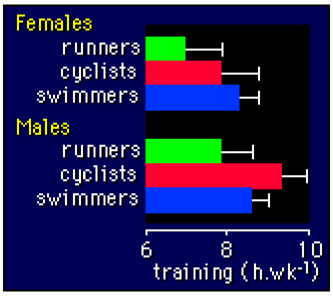
And finally, there is curvature for at least one sport, so we need to fit a quadratic term overall, and a quadratic term that might differ between the two sports. The way to do that is to include the quadratic term as a main effect and as an interaction with sport. So here's the full model:

attitude <= sport experience sport*experience experience^2 sport*experience^2

The p value for sport*experience^2 tells you whether any difference in the curvature for the two sports is statistically significant. Once again you express this difference as a contribution to the overall R^2 for the model, as described for the simpler example above.

Two-Way ANOVA

model: numeric <= nominal1 nominal2 nominal1*nominal2
example: training <= sex sport sex*sport



This model is like a one-way ANOVA with an extra grouping variable. I've put the groups up the Y axis to make it easier to show the names of the sports. The outcome variable (hours of training per week) is still the dependent variable, even though it is shown on the X axis. By the way, the *two* in *two-way ANOVA* refers to the two main effects in the model, not to the two levels for sex in this example. So a three-way ANOVA would have three main effects, and so on. Now let's see what each effect in the model tells us about the effects of sex and sport on training.

The main effects are easy to understand. Sex tells us how different the training is between sexes overall. In these data it's obvious that the males are doing more training than the females. The numeric value for the difference would be given by the values of the two levels for sex in the solution for the model. You would get your program to perform an estimate or a contrast between the two levels of sex to get confidence intervals or p values. Note that the resulting value for the difference between sexes is equivalent to the difference you would get between the means for the males and the females with equal numbers of males and females in each of the three sports. If you have different numbers in each sport, the difference between the raw means for all the males and all the females will be different from the stats program's estimate of the difference. The estimate from the stats program is usually the one you're interested in.

In a similar fashion, the sport effect gives overall differences between the sports. You would do pairwise estimates or contrasts to see how different the sports are from each other. In the example, runners are clearly different from cyclists and swimmers, but cyclists are about the same as swimmers, because the difference goes one way for females and the other way for males. Again, the estimates for the differences between sports are the same as what you would get with equal numbers of males and females in each sport.

Which leaves us with the interaction. It tells us about the overall trend for differences in the sports within females compared with the trend within males. Well, it looks like something interesting is going on with the trends, because the means for the cyclists and swimmers swap over between females and males. To find out how big the differences are, you would do estimates or contrasts for the different levels of the interaction term. There are six levels in sex*sport: two for sex, and three for sport. In alphabetical order, the levels are female·cyclists, female·runners, female·swimmers, male·cyclists, male·runners, male·swimmers. The estimate of the magnitude of the female-male swap over for the cyclists and swimmers would be given by the value for (male·cyclists - male·swimmers) - (female·cyclists - female·swimmers). This difference in the differences is about one hour, so you'd say "males did relatively more cycling than swimming in comparison with females; the combined difference was about an hour..." And as before, think about the difference as what you would get with equal numbers of males and females in each sport.

With luck your stats program will calculate confidence limits for all the differences between groups. It will certainly calculate p values anyway, and you can convert these to confidence limits by downloading a spreadsheet.

Don't forget to keep one eye on the standard deviations when you compare means of groups. In the above example, many of the differences between groups are equivalent to at least one standard deviation (an effect size of at least 1.0), so they are moderate to large. Differences between standard deviations are also important. Notice that the standard deviations for the swimmers are about half those in the other sports. Maybe the swimmers all train in squads with similar training programs. That could be a problem, in more ways than one. First, if the subjects *are* drawn from just a few squads, their values of training will not be independent of each other, so we'll have to use repeated-measures modeling. And secondly, different standard deviations violate an assumption our usual analyses are based on. We can get around this problem by transforming the training times. Log transform? No, the standard deviations would need to be bigger for more training. Rank transform? Yes, non-parametric analysis is called for here. Just rank the entire column of data for training, then do the analysis as usual.

More Than One Dependent Variable

These are called **multivariate** models. In other words, things like:

jump sprint <= sex

You read this as: what is the effect of sex on a person's ability to jump *and* sprint? Sure, sex might have an effect on each of these separately, but let's put them together and look at the overall effect on both. This would be an example of multivariate analysis of variance (MANOVA). The test statistics are unusual (e.g. Hotelling t^2, Wilks' lambda).

Multivariate models are easy in principle, but in practice it's hard to interpret the outcome statistics. I advise you to analyze your dependent variables separately, or do a dimension reduction first, then analyze each dimension separately.

Multivariate models have been adapted for analysis of repeated measures. For example, replace sprint in the above model with a second measurement of jump, and you could write:

jump1 jump2 <= sex,

where jump1 and jump2 are jump heights on the first and second occasion. I deal with this approach to repeated measures later.

Models With Variables of Uncertain Status

There's only one kind of model here, but it goes by various names: path analysis, structural equation modeling (SEM, not to be confused with standard error of the mean), and linear structural relationships (LISREL).

I've never used this kind of modeling, so this section will be brief and untrustworthy. As far as I can see, the technique can be applied only in cross-sectional studies with hundreds of subjects. It represents an attempt to establish a chain of cause and effect between variables. The stats program does it by looking at all the correlations between all the variables, then creating the best chain, like so:

numeric <= numeric <= numeric <= numeric...

The program produces correlations for each link, and a correlation between the variable on the far right (the cause) and the far left (the effect). Validity of measurement for each variable, where known, can be taken into account.

There is now a huge literature on this topic, which in my opinion goes way beyond what is justified for cross-sectional data. Let's face it, cross-sectional studies can only ever provide suggestive evidence; in the end you need longitudinal studies to nail cause and effect. That's where repeated-measures models come to the fore. Read on.

REPEATED MEASURES MODELS

So far, all the models we have looked have been for data from **cross-sectional** or **descriptive studies**. These are studies in which each person is observed only once, so for each variable you have only one value per person. To put it another way, each row in the data set is for a different subject.

Now, what about **longitudinal studies**, in which people are observed more than once? In particular, what about **interventions** or **experiments**, where you compare values of a dependent variable before and after you try something like a training program or a potentially active drug? You can analyze data from these studies with the procedures used for cross-sectional data *only* if you can assume that the residuals are uniform--have the same standard deviation--for each of the repeated measurements. But in general, you *can't* assume such uniformity: subjects will show more variation on some repeated measurements than on others, usually because of differences between measurements in the effects of time or the treatment. So you have to use repeated-measures models.

We'll start on this page with the simple case of only two trials for only one group of subjects (no between-subject effect). On the next page I'll extend it to several groups (a between-subjects effect, e.g. an experimental and control group). Then I'll deal with more than two trials, first without a between-subjects effect, then with a between-subjects effect, before I deal with other repeated-measures models including the simple, robust approach of within-subject modeling. Then there is a page on how to use the mixed procedure in the Statistical Analysis System, with links to . Finally, I devote a page to a problem that can arise in repeated-measures analyses, regression to the mean. But first, some other resources I have created since writing these pages: a slideshow, a stand-alone article, and some spreadsheets.

Slideshow on Repeated Measures

For a Powerpoint slideshow (340 kB) dealing with most aspects of repeated-measures analyses, click here. I presented this talk at the 2003 annual meeting of the American College of Sports Medicine. The sections are Basics (analysis by ANOVA, within-subject modeling, and mixed modeling; fixed and random effects; individual responses and asphericity), Accounting for Individual Responses, Analyzing for Patterns of Responses, and Analyzing for Mechanisms. The information in the slide show complements the information on these pages. Read both.

Articles and Spreadsheets for Straightforward Repeated Measures

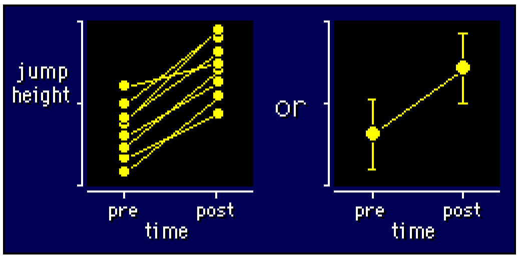
I have created spreadsheets for analysis of repeated-measures data from controlled trials and crossovers. You add the raw observations, the spreadsheet does the rest. I have also written articles at the Sportscience site explaining important issues in such analyses, and how the spreadsheets address them. (Links to the Sportscience articles will not work if you are using a copy of these pages off-line.)

Click to view the 2006 article, which explains the use of a covariate and has links to earlier articles. See also an article on the different kinds of controlled trials in the 2005 issue, which explains the names I have used below for the spreadsheets.

Click to download the spreadsheet for pre-post parallel-groups trials, the spreadsheet for post-only crossovers (which also works for the paired t-test model on this page), and the spreadsheet for pre-post crossovers. The following links will download earlier version that do not include the covariate and other enhancements: spreadsheet for controlled trials, spreadsheet for crossovers, and fully controlled crossovers.

Paired T Test or Repeated-Measures ANOVA with two trials and no between-subjects effect

model: numeric <= (subject) trial
example: jumphgt <= (athlete) time



Don't try to understand the model yet. Just look at the example in the figure, which shows individual values on the left and means and standard deviations on the right. There is one measurement on each of eight athletes before (pre) and after (post) a training program aimed at increasing jump height, with no control group. This sort of design is sometimes described as one in which the subjects "act as their own controls", although this description fits any longitudinal study, whether or not there is a control group.

The results can be displayed as shown in the left-hand panel, with pre and post heights linked for each subject. The right-hand panel shows the more usual way of connecting the means by a line. By the way, it's wrong to use a bar graph, because the pre and post data are from the same subjects.

It doesn't look anything like it, but this model is actually a two-way ANOVA. If I'd drawn bars instead of points for the pre and post heights, you might have seen that it is at least a one-way ANOVA, time being the nominal effect (with two levels, pre and post), and height the dependent numeric variable. So let's get started with jumphgt <= time.

The other effect is hidden in the right-hand figure, but it's clear in the left-hand side: the identity of the subjects. We introduce this variable as a way to link each subject's measurement of height at the pre and post times. Hence the full model: jumphgt <= (athlete) time. In the general model, one term in the ANOVA is the identity of the subjects, and the other term is the identity of the time points or trials.

Hang on. Why (athlete) rather than athlete? Well, the variable representing the identity of the subjects is a bit different from all the other variables we've met so far. The subjects are usually a random sample of a population, so this variable is known as a **random effect**. If we repeated the study, we could have a different sample of subjects, each with different values drawn randomly from the population. In contrast, the identity of the time points is a **fixed effect**, because this variable would have the same values and levels (pre and post) in any repeat of the study. Look back at the nominal variables in the other models we've dealt with and you'll see that they are all fixed effects. For example, sex always has values male or female in every sample, and we assume the effect of maleness or femaleness is the same for every male or female. For more information on fixed and random effects, see the slideshow on repeated measures. If you want to work with mixed models, make sure you get familiar with my "hats" metaphor for random effects, as explained in the slideshow.

So, I've put parentheses around the subject term to indicate that it's a random effect, and to let you know that stats programs don't normally include the subject term in the model in the way that I have here. If I left the parentheses out, I would imply that the subject term is a fixed effect. It *is* possible to analyze your data as a straightforward non-repeated-measures ANOVA with the subject term as a fixed effect, but the results you get are appropriate only for repeated-measures data that have uniformity of residuals. I deal with that later under the heading sphericity or covariance structure.

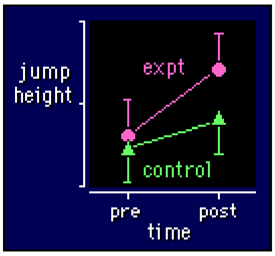
We don't have the interaction term athlete*time in the model, partly because athlete is a random effect, and partly because we would need multiple measurements for subjects at the pre and post time points for the interaction term to make any sense. Let's leave aside this complexity.

It all sounds awfully complicated, but in practice it's straightforward. You have two lots of measurements performed on the same subjects, and all you want to know is how the means have changed. Most stats programs can do that for you without you having to worry about models like the above. All you do is click up a **paired t test**, which produces a p value for the difference in the means, and hopefully a confidence interval. The paired t test has the same internal workings as the unpaired t test, which is why they share the same name.

On the next page we'll add a control group. After all, the athletes might jump higher in the post test simply because they have learned how to do the test, not because they responded to your training program. A group that does everything the same as the experimental group, other than the training program, "controls" for this and other problems. But the main reason I'm talking about a control group now is to explain the terminology in the heading for this page. Having a control group in a repeated-measures design is an example of a **between-subjects effect**, because there are different subjects in the control and experimental groups. Hence *no between-subjects effect* in the title of this section. Time or trial is a **within-subjects effect**, because the same subjects experience the different levels of that effect.

Repeated-Measures ANOVA with two trials plus a between-subjects effect

model: numeric <= (subject) group trial group*trial
example: jumphgt <= (athlete) group time group*time



Let's take the experiment on the previous page, where we attempted to increase jump height with some sort of experimental treatment. As before, we measure jump height pre and post the treatment in a group of subjects, the experimental group (*expt*in the figure). But now we also have a second group who get a different treatment, and the aim of the experiment is to compare the change in jump height in the two groups. If that different treatment is nothing at all, or a sham treatment (a **placebo**), the second group is called a *control* group--hence the name for this sort of experiment, a **controlled trial**.

Let's analyze it the easy way first. For each subject, subtract the pre height from the post height to get a change score. Now compare the change scores in the two groups using an unpaired t test. Use the unequal-variances version of the t test, because the standard deviation (square root of the variance) of the change scores in the experimental group is likely to be larger than that in the control group, owing to individual responses to the treatment. The spreadsheet for controlled trials can do it all for you. If you have three groups (e.g. two experimental groups and one control group), use a new spreadsheet for each pairwise comparison of groups. You can also use a one-way ANOVA on the change scores, but beware: ANOVA assumes equal variances (standard deviations) of the change scores in all the groups. See the slide show on repeated measures for an explanation of these subtleties.

Now for the model, which is the hard way. We have to do it, though, because you need to understand the model for later complexities with repeated measures. Let's start with the simple model from the previous page:

jumphgt <= (athlete) time

This model represents the obvious fact that jump height is affected by time (depends whether it's the pre-test of the post-test) and the identity of the athlete (depends how good a jumper s/he is). But we now have two groups of subjects (control and expt), so we have to add a term to show that athletes in one group could jump differently from those in the other:

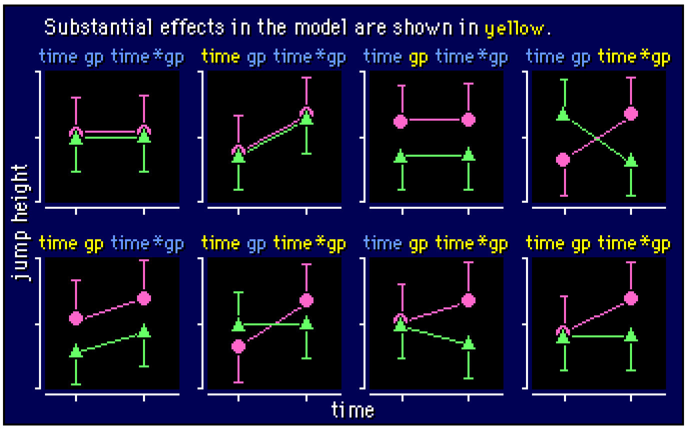
jumphgt <= (athlete) time group

Technically the model is now a three-way ANOVA, but no-one ever calls it that. OK, what tells us whether the experimental group did better in the post test, relative to the control group? The group effect? No, this term represents the *overall* difference between the groups, counting pre and post tests. We're missing a term, of course: the interaction time*group. This term is the first thing you look at to see how your treatment worked.. So the full model is:

jumphgt <= (athlete) time group time*group

By the way, the order of time and group in the model is irrelevant, and time*group is the same as group*time.

The data for this model seem simple enough (pre and post means and SDs for two groups), but interpreting the substantiveness/significance of each term in the model can be confusing. So here are examples illustrating the eight possible combinations of insubstantial and substantial effects for the different terms in the model. Don't go past this section until you understand all eight parts of this diagram:



The last two examples on the lower right are the ones we usually want in a study: no difference between the control and experimental groups in the pretest, and a nice big divergence on post-test. The fact that main effects are substantial in these two examples is irrelevant. The other two examples with a substantial interaction also illustrate treatments that worked, but the outcomes are not ideal, because in both cases the groups are different in the pretest. A worry, because it means that one or both of the groups can't be representative of the population, at least as far as jump height is concerned. And non-representative samples mean non-generalizable findings!

Finally how do we calculate the magnitude of the experimental effect? Easy. The post score minus the pre score for the experimental group is the main thing, but we have to subtract off any change in the control group. To do it as an estimate or contrast in the repeated-measures ANOVA, combine the four levels of time*group in the following way: (post·expt - pre·expt) - (post·cont - pre·cont).

Special Case: Simple Crossovers

In a simple crossover design, half the subjects get a control treatment followed by an experimental treatment, while the other half get the treatments the other way around. People usually analyze the data as a simple paired t test, which means they effectively subtract the control response from the experimental response for each subject, without regard for the order of treatment. In a minute I'll show you a better way, using the above ANOVA model, and I'll generalize it to multiple crossovers. First, more about simple crossovers.

Why split the subjects into two groups and cross the treatments over? Because if all subjects get the control and experimental treatments in the same order, you won't know whether any change you see is truly an effect of the treatment, or just an effect of being tested a second time--a *practice* or *learning effect.* When you split the subjects, the group that gets the control first has the practice effect added to the experimental treatment, whereas the group that gets the experimental first has the practice effect added to the control treatment. So when you average the difference scores, the practice effect disappears and you are left with the treatment effect, provided the two groups have the same number of subjects.

Fine, but there's a problem. When there *is* a practice effect, you get two clusters of difference scores. For example, if the practice effect is about the same size as the treatment effect, one set of difference scores will be around zero, and the other will tend to be twice as large as the treatment effect. The average is still equal to the treatment effect, but the effect appears to be more variable between subjects. The result is a bigger (worse) confidence interval for the treatment effect, or a bigger (worse) p value, or less power to detect the treatment effect.

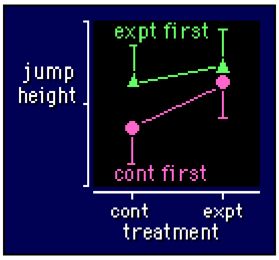
Another potential problem is *carry over.* For the group that gets the experimental treatment first, it's important that any effect of the treatment disappears by the time that group gets the control treatment--otherwise the difference between control and experimental treatments for that group will be reduced. The result will be an apparently smaller treatment effect overall, and an apparent practice effect. For example, if the treatment effect carries over completely, the analysis will produce a treatment effect that is half its true value, and an apparent practice effect of the same magnitude. So **you can't do a training study as a crossover**, unless you are confident that the adaptations produced by the experimental training program decay away before subjects get the control program.

You might be able to get over the problem of carry over by increasing the time between the two treatments. But the longer the time, the less reliable the dependent variable is likely to be, which means a wider confidence interval for the difference between the treatments.

One way around the problem of practice and carry-over effects is to throw out the crossover altogether. Replace it with a properly controlled study, in which you split the subjects into two groups, give both groups a pre-test, then administer the control treatment to one group and the experimental treatment to the other, and finally do a post-test on both groups. Any practice effect should be the same for both groups, so it disappears when you calculate the change in the experimental group minus the change in the control group.

So why bother with a crossover at all? For a very good reason: you get the same confidence interval for the treatment effect with *one quarter* the number of subjects as in a fully controlled design, provided there are no practice and carry-over effects. For such a big saving in time and expense, always consider a crossover before a fully controlled study. Minimize any carry-over effect by allowing adequate time between the treatments. And don't worry about the practice effect, because ANOVA takes care of it. Here's how:

model: numeric <= (subject) treat group treat*group
example: jumphgt <= (athlete) treat group treat*group
The figure shows data for an example of a simple crossover, in which an experimental treatment increased jump height relative to a control treatment. I've separated the data for the two groups (control treatment first, experimental treatment first) to illustrate a practice effect, which adds to the difference between experimental and control treatments for the group that had the control treatment first, but reduces the difference for the other group. The data also illustrate that randomization of athletes to the two groups resulted in one group (expt first) being somewhat better jumpers overall.



The model has the same form as the model at the top of this page, but the time effect is now replaced with treat, which has two levels (cont and expt). The other main effect, group, now represents which group each subject was assigned to (contfirst, exptfirst). The interaction term treat*group has four levels (cont·contfirst, cont·exptfirst, expt·contfirst, and expt·exptfirst).

The difference between the two levels of the treatment effect (expt - cont) tells you the thing you're most interested in: how well the treatment worked relative to control. The difference between the two levels of the group effect (exptfirst - contfirst) tells you how different your two groups of subjects were, so it's a measure of how well you randomized your subjects to the two groups. The interaction gives you the size of the practice effect, and I'll leave you to figure out that the appropriate contrast is 0.5*(expt·contfirst - expt·exptfirst - cont·contfirst + cont·exptfirst). If that's too challenging, here's another way to get the practice effect. First, make another repeated-measures variable called trial in your data set. Trial is almost the same as treat, but trial has values of the dependent variable corresponding to the first and second trial, whereas treat has values corresponding to control and experimental treatments. Now do the ANOVA with group, trial, and group*trial in the model. The practice effect comes straight from trial in this model.

Get your stats program to give you confidence intervals for all these contrasts, please, not just the p values! And if you plot your data for publication, show the two groups as I have done in the above example.

A bonus for this method of analyzing crossovers is no absolute requirement for an equal number of subjects in each group. It's still best to have equal numbers, but if you get dropouts in one group, the resulting treatment effect is not biased by any practice effect. It would be biased if you used a paired t test to analyze the data.

Users of the Statistical Analysis System have the option of modeling the data in a slightly more intuitive way. Instead of having a group effect in the model, use a variable called trial, which has values *first* and *second* (or *1* and *2).* This variable indicates whether the given observation represents each subject's first or second trial or test. Here's the model:

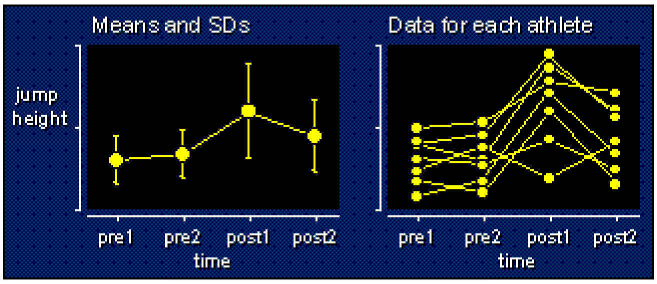
model: numeric <= (subject) treat trial treat*trial

It looks similar to the previous model, but trial is actually a second within-subject factor, which we haven't dealt with yet. It turns out that traditional methods of repeated-measures ANOVA can't handle this model, because each subject has values for only two of the four combinations of treat and trial. But the new mixed procedure in SAS handles it brilliantly. Just use the treat term to get the estimate of the difference between the experimental and control treatments, and use the trial term to get the practice effect. An appropriate combination of the levels of treat*trial gives the difference between the means of the two groups of subjects with treatment and practice effects partialed out, if you want to check how evenly the subjects were randomized to the two treatment sequences.

The above models can be generalized to multiple crossovers: crossovers with several treatments. More about those after the next page, which deals with more than two trials.

Repeated-Measures ANOVA with three or more trials and no between-subjects effect

model: numeric <= (subject) trial
example: jumphgt <= (athlete) time



Check back and you'll see it's the same model as for two trials with no between-subjects effect: adding extra trials doesn't usually mean a different model. This kind of design--multiple repeated measurements without a control group--is sometimes called a **time series**. In the above example, there are two trials (pre1 and pre2) to establish a baseline of performance before some kind of treatment, then two trials (post1 and post2) to see the effect of the treatment. There's a big effect at post1, but it's wearing off by post2.

One way to analyze these data is to do a series of paired t tests. Post1 vs pre2 is the first comparison you would want to do. You'd also be interested in post2 vs post1, and possibly pre2 vs pre1, post 1 vs the mean of pre1 and pre2, and so on. An analysis that takes into account all the tests is more elegant and more powerful. The trouble is, generally we can't analyze such data using conventional ANOVA. The example shows several reasons why. See if you can spot them before reading on.

You should have noticed that the standard deviation is bigger for the post1 and post2 trials. Different SDs are a problem for conventional ANOVA, but if that was the only problem, we could fix it by doing a non-parametric analysis via rank transformation of the dependent variable. No, the real problems are apparent only when you look at the data for the individual athletes. One of them appears to be a **negative**

responder to the training program, and another is possibly a **non-responder**. What's more, the ordering of the subjects between pre2 and post1 or between post1 and post2 was not nearly as good as the ordering between the baseline tests. It's this change in consistency of ordering, or to give it its statistical term, the change in reliability between tests, that stymies the normal ANOVA. **Individual differences** in the response to the treatment between subjects is the reason for the loss of ordering here.

A change in reliability shows up as different correlations between pre1 and pre2, pre2 and post1, etc. When these correlations get too different and/or the standard deviations get too different, it's called **loss of sphericity** or **asphericity**. Statisticians examine sphericity in something called a **covariance matrix**, which neatly summarizes correlations and standard deviations for all the levels of the within-subject effect (time or trial). I will provide more information about covariances soon on the page devoted to the use of Proc Mixed in the Statistical Analysis System. Meanwhile, let's look at the three fixes for this problem.

Fix #1: Multivariate ANOVA

Someone worked out that you can treat the values of the dependent variable for each trial as separate dependent variables. In our example, jumphgt becomes jumphgt1 (values at pre1), jumphgt2 (values at pre2), etc. The data set would look like this:

jumphgt1	**jumphgt2**	**jumphgt3**	**jumphgt4**
163	165	171	168
167	166	170	167
etc.	etc.	etc.	etc.

Notice that time as a variable has disappeared: it's been absorbed into the four new variables for jump height, but it reappears as a within-subjects factor when you run the analysis. The variable subject has also disappeared: it's not needed, because there is only one row per subject and no ambiguity is possible.

It's difficult to write these four new variables into a model. Obviously they go on the left-hand side, like so:

jumphgt1 jumphgt2 jumphgt3 jumphgt4 <=

but what goes on the right-hand side? Nothing! Looks silly, but SAS makes you show it like this when you analyze a data set like the above.

I don't recommend the multivariate ANOVA approach. For starters, all it provides is a p value for the overall effect of time. It doesn't provide estimates or p values for the individual contrasts of interest (post1 minus pre2 etc.). What's more, I've shown by doing simulations that the p value it does produce is too big with some kinds of data and too small with others. Another big problem is **missing values**: if one of your subjects missed one of the tests, that subject is omitted from the analysis entirely.

Fix #2: Adjusted Univariate ANOVA

This method has been the most widely used. The analysis is done as a conventional two-way ANOVA with one dependent variable (hence *univariate)* and effects for subject and trial (time in our example). The program then uses the covariance matrix to come up with a correction factor that leads to a different p value for the effect of trial. You choose from two factors: Greenhouse-Geisser epsilon or Huynh-Feldt epsilon.

Fix #3: Within-subject Modeling

In this approach, you avoid the problems of repeated measures by not doing them! Instead, you convert each subject's repeated measurements into a single number, then do paired or unpaired t tests or simple ANOVAs on those numbers. I explain this approach later and in the slideshow.

Fix #4: Modeling Covariances (Mixed Models)

Suppose you have data like the previous example, where the standard deviations and correlations for the repeated measures are all over the place. Don't adjust for them: make them part of the model! Yes you can, with Proc Mixed in the Statistical Analysis System (SAS). It's a major breakthrough. The procedure is known as modeling covariances, because standard deviations and correlations can be expressed more generally as covariances (nothing to do with analysis of covariance, by the way). Unfortunately the instructions for the procedure that does it in SAS are incomprehensible to all but highly trained statisticians. But if you can find one of those, you will be delighted, for the following reasons:

- This method of modeling can handle missing values! No longer do you lose the entire data for a subject who missed one or two tests.
- You can dial up just about any structure of standard deviations and correlations.
- The model starts to look like a proper univariate ANOVA again. Unfortunately you still don't specify the subject term in a natural way (as an overt effect in the model), not in SAS anyway. The model statement is reserved for fixed effects. The model for our example would be jumphgt <= time. You specify the identity of subjects as a random effect.
- The data set is submitted in the more natural univariate format. For example:

athlete	**time**	**jumphgt**
Jo	pre1	163
Jo	pre2	165
Jo	post1	171
Jo	post2	168
Kim	pre1	167
etc.	etc.	etc.

By the way, the term *mixed* refers either to the fact that you are modeling a mixture of means and covariances, or (same thing) to the fact the model consists of a mixture of random and fixed effects. The subject term in a repeated-measures model is a random effect. Random effects produce variance that has to be accounted for in the model.

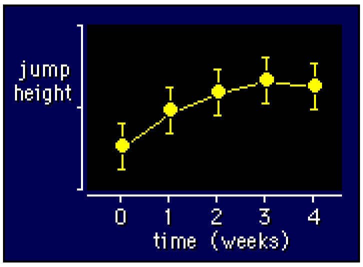
I have now added SAS programs for analyzing repeated-measures data with the mixed procedure in SAS. Link to them from the page devoted to Proc Mixed.

Estimates or Contrasts

OK, let's assume we've got a method that accounts for lack of sphericity. Now for the question of estimates or contrasts between the mean jump heights at the different times. You can dial up any contrast you like, if you and the stats program are good enough! For example, was the jump height straight after the intervention higher than the mean of the baseline values (and what's the confidence interval on the difference)? Some stats programs offer standard contrasts. Examples: **One level with every other**, would be the obvious contrast to apply to post1 in the above example. **Each level with the one immediately preceding** is good for determining where a change takes place in a time course, although you can easily get the situation where no successive contrasts are significant, and yet there is obviously a significant trend upwards or downwards. That's where **polynomial contrasts** come to the rescue: the ANOVA procedure fits a straight line, and/or a quadratic, and/or a cubic, etc. to the means.

Polynomial Contrasts

Here's an example of data that would be ideally suited to fitting a straight line and a quadratic. It's jump height and time again, but I've added an extra time point and made a curvilinear trend:



The magnitude and significance of the linear component would tell you about the general upward trend in performance, while the quadratic component would tell you how it is leveling off with time. If your stats program's a good one, it will offer polynomial contrasts as an option. Otherwise you will need a high-powered helper to combine the levels of the time effect in a way that generates the coefficients of the polynomials. You can adjust for unequal intervals between the time points, too, if your stats program or helper are really good. (SAS users can fit a polynomial directly in the model with Proc Mixed.)

By the way, what if the data in the above figure were not repeated measures? In other words, what if there were different subjects at each time point? For example, the data could represent level of physical activity in samples drawn from a population at monthly intervals. Could you still do polynomial contrasts? Of course. You do it within a normal ANOVA.

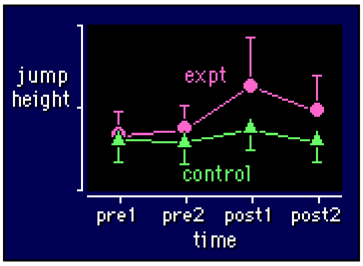
Controlling Type I Error with Repeated Measures

Keeping the overall chance of a type I error in check efficiently for multiple contrasts between levels of a repeated-measures factor seems to be theoretically difficult. The SAS program simply doesn't offer the option. I don't worry about it anyway, because I don't believe in testing hypotheses. If you are a p-value traditionalist, use the Bonferroni correction. And as I explained earlier, do specific estimates/contrasts regardless of the p value for the overall effect.

Repeated-Measures ANOVA with three or more trials plus a between-subjects effect

---

model: numeric <= (subject) trial group trial*group
example: jumphgt <= (athlete) time group time*group



You should be able to see that this model is the previous two merged together. The interpretations of the main effects and interaction term from the first of the previous models (two time points, two groups) still apply. And all the problems with sphericity from the model on the previous page (three or more trials in one group of subjects) still have to be addressed.

To summarize:

- Use univariate ANOVA with adjustment for non-sphericity, or model the covariance matrix with proc mixed and/or a statistician.
- Trial*group tells you how the experiment worked.
- Contrasts between different levels of trial and group for the trial*group term tell you where the experimental group differs from the control or other groups.
- Get the program to give you estimates and confidence intervals, not just contrasts and p values.

If you didn't know any better, you might try to analyze these data by doing a series of unpaired t tests for each time point. That would be foolish, for three reasons: the power to detect differences would be lousy, because you would not be making use of changes in each subject's values; you would not be taking into account any differences between the two groups at baseline; and finally you would not impress the reviewers and editor of the journal you submit the research to. You could fix the first two criticisms by subtracting each subject's mean baseline value from the post1 and post2 values, then doing unpaired t tests on these difference scores.

Special Case: Multiple Crossovers and Latin Squares

Recall that a simple crossover is a design in which all subjects receive two treatments. We analyzed the data with a treatment effect and a group effect that indicated which treatment each subject got first:

model: numeric <= (subject) treat group treat*group
example: jumphgt <= (athlete) treat group treat*group

OK, but what about more than two treatments? For example, can we have a crossover in which every subject gets a control treatment and two experimental treatments aimed at increasing jump height? Sure, just use the above model. Treat is the repeated-measures or within-subject effect, with three or more levels. The group effect represents the sequence of treatments that each subjects is assigned to, and it also has three or more levels, as we'll see. And in the same manner as for a simple crossover, with SAS you can use a trial effect instead of a group effect to indicate whether each treatment was first, second, third... for each subject (see earlier):

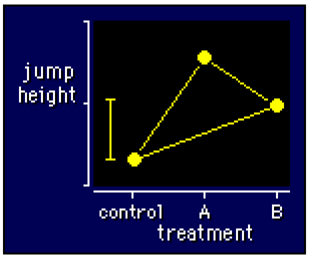
model: numeric <= (subject) treat trial treat*trial

Whichever model you use, the good news is that you still need only about a quarter of the subjects of a fully controlled study! Here's how to set up and analyze these multiple crossovers.

First, randomize your subjects to the various sequences of treatments. For example, if you have two experimental treatments (A and B), and a control treatment (C), there are six possible sequences: A-B-C, A-C-B, B-C-A, B-A-C, C-A-B, and C-B-A. It's best to use all these sequences, because if one of the treatments has a carry-over effect, it will affect all the other treatments equally.

Next, decide on sample size. By running a simulation for this design, I've found that about 12 subjects give acceptable confidence limits for the pairwise comparisons of the treatment effects, for very high reliability (r=0.95). So you could start with two subjects doing each of the six sequences of treatments, then do more subjects if necessary, as described in sample size on the fly. If one or two subjects miss a treatment, or if you lose one or two subjects completely, no great problem: the data don't have to be "balanced" to give unbiased estimates, provided none of the treatments carry over.

Now do the work, get the data, and analyze them. If you can use proc mixed in SAS, I recommend the crossover model with trial: it gives slightly better confidence intervals, and it's easy to estimate the learning effects from the levels of the trial effect (e.g. for three treatments there's one learning effect between first and second trials and another between second and third trials). See the simulation for the program statements. Non-SAS users will have to use the usual crossover model with a group effect, but it's fiendishly difficult to work out the appropriate combination of levels of treat*group to get the learning effects.



How best to plot data for a multiple crossover? For a simple crossover I suggested showing the means for the two groups of subjects for control and experimental treatments. That approach is no good for multiple treatments, because there'll be too many groups and not enough subjects in each group. One solution is simply to plot the means for each treatment. Connect all the points together, as I have done in the figure, to show they all have the same repeated-measures relationship to each other. Or if the treatments can be put into a sensible order, such as an increasing dose of something, plot the treatments in order along the X axis and connect the points sequentially. Either way, you shouldn't use a bar graph.

Standard deviations are a problem with multiple crossovers. You could plot the standard deviations for each treatment, but they will be inflated by any learning effects. Stats wizards using proc mixed in SAS can extract the composite between-subject SD from the ANOVA. This SD includes within-subject retest error, but it is not affected by the treatment and learning effects. It's therefore the best measure of variation by which to assess visually the magnitude of the treatment effects shown in your plot.

In the text of the Results section, give the raw differences between the means for the treatments and the confidence intervals for these differences. Or when appropriate (e.g. for most athletic performance measures), show percent differences and their confidence intervals, as provided by analysis of log-transformed performance measures.

When there are four treatments altogether (e.g. a control and three experimental treatments), there are 24 possible sequences of treatments. You could randomize one subject to each sequence of treatments, but you might not need 24 subjects to get acceptable confidence limits for your comparisons. But with random assignment of less than 24 subjects, you might not end up with balance in the way treatments follow each other. A problem if one of the treatments has a carry-over effect, because it will have the greatest effect on the treatment that follows it most times. So it's better to randomize to a subset of sequences that ensures every treatment follows every other treatment the same number of times. Any carry-over effect will then affect every other treatment equally. Such a balanced subset of sequences is called a **Latin square**. Here's the Latin-square set for four treatments, A, B, C, and D:

Sequence 1: A B C D
Sequence 2: B D A C
Sequence 3: D C B A
Sequence 4: C A D B

Check and you'll see that each treatment follows every other treatment only once. Your sample size obviously has to be a multiple of 4 to keep the balance. For example, for 12 subjects, assign three at random to each of the four sequences.

Here's a balanced set of sequences for five treatments. In this case you need 10 sequences to keep the balance (and each treatment is followed by every other treatment twice), so you'll need multiples of 10 subjects in your study:

1: A B C D E
2: B D A E C
3: D E B C A
4: E C D A B
5: C A E B D
6: E D C B A
7: C E A D B
8: A C B E D
9: B A D C E
10: D B E A C

We're into seldom-trodden territory now, but I must record the trick of generating these Latin squares, in case you want to do more than five treatments. Instead of labeling the treatments with letters (A, B, C...), let's label them with numbers (1, 2, 3...). Assume n treatments. Here is the Latin square for n even. As above, the sequences of treatments are given by the rows, not the columns:

1	2	n	3	n-1	4	n-2	.	.
2	3	1	4	n	5	n-1	.	.
3	4	2	5	1	6	n	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

For n odd, use the above set of sequences, plus its mirror image. Check out the sets for five treatments given above, to see what I mean. Thanks, Dennis Loiselle!

Other Repeated-Measures Models

I deal with several here: an extra between-subject effect (example: male and female subjects), two or more within-subject factors (example: the same subjects get several treatments at several time points), a general type of within-subject model (you fit data to each subject separately, then combine the fits), inclusion of covariates in the model to analyze for individual responses, trends in repeated sets of trials, and mechanisms, and finally troublesome variables that might require transformation or more advanced approaches.

An Extra Between-Subject Effect

model: numeric <= (subject) trial group trial*group
sex sex*trial sex*group sex*trial*group
or simply: numeric <= (subject) trial | group | sex
example: jumphgt <= (athlete) time | group | sex

I've used the first example from the previous page. The only difference is that we now know our subjects are a mixture of males and females. The analysis gets complicated, because sex could affect everything--that's why there are so many interaction terms in the model. But we're usually interested only in the extent to which the sexes differ for the effect of the treatment, so that means comparing the appropriate levels of sex*trial*group. In short, find the levels for trial*group that tell you what you want to know, then find the difference between those for the females and those for the males in the sex*trial*group term.

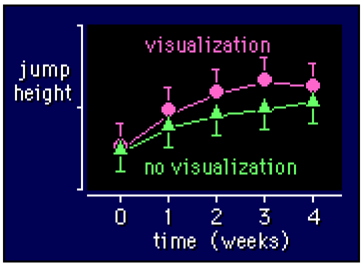
A word of warning! The term trial*group gives you the overall effect of the treatment between the experimental and control groups, but with sex in the model it's the average of the effect on the females and males. In other words, it's the expected effect of the treatment with equal numbers of males and females, even though your sample may have had unequal numbers.

The simple notation of a vertical bar (|) between effects is what the Statistical Analysis System uses to indicate that you want to include all possible main effects and interactions in the analysis. I usually leave them all in, because I subscribe to the idea that all independent variables have some effect on the outcome variable, however small. The only harm that can come from including all the interaction terms is loss of degrees of freedom, and therefore widening of the confidence intervals. On the other hand, if the effects of sex on the treatment are substantial, then inclusion of sex*trial*group will actually make the confidence intervals smaller, because it will eliminate the variability in the effect of the treatment that was due to sex.

In any case, if the effect of sex is substantial, you will want to know about it! For that reason, if the outcome of your study is as important for males as for females, you should try to have equal numbers of male and female subjects. Your confidence interval for the overall effect on females and males combined will be only a bit wider than if you had subjects of one sex, but of course the effect will be the average of the effect on (equal numbers of) females and males. The sex*trial*group interaction will tell you how different the males are from the females, although the confidence interval for the comparison of the effect on females and males will be about twice as wide as that for the overall effect. So, to properly delimit the difference between females and males, you will need four times as many subjects as for a single-sex study. That's the bad news. The good news is that you will end up with a wonderfully narrow confidence interval for the overall effect.

Two or More Within-Subject Effects

model: numeric <= (subject) trial condition trial*condition
example: jumphgt <= (athlete) time test time*test



Imagine that the intervention is a program aimed at training athletes to use visualization before a jump (and thereby jump higher). At weekly intervals the athletes perform a jump with and without visualization. The figure shows a possible outcome, in which visualization starts to work after a couple of weeks of training:

The model has the same form as in the previous example, but I've replaced group (representing separate measurements on different subjects) with condition (representing separate measurements on the same subjects). The interpretation of trial and condition in the model is the same as that for trial and group. Coaxing your stats program to deal with two or more within-subject effects will be a challenge!

In the example, I've renamed trial and condition to time and test to make things a bit clearer. The interpretation of time is obvious. Test represents the test of jump height, and it has two levels: visualization and no visualization. The interaction effect time*test tells us about the difference between the two time courses, and contrasts between the different levels of time and test for this effect tell us when the effect of visualization differs from no visualization. A polynomial contrast would show a substantial difference in linear and quadratic components, indicating that jumping with visualization shows a more rapid improvement in jump height initially (the linear effect), but that the gap is closing by Week 4 (the quadratic effect). You'll have to think really hard about this one.

An unsophisticated approach to these data would be to perform a series of paired t tests for each time point. For example, you might find that the difference between visualization and no visualization is significant at Weeks 2 and 3, but not at the other times. This approach does not take into account any difference between the tests at Week 0, so it's not valid. An acceptable fix is to subtract the jump height at Week 0 from that at each of the other weeks (for the two tests separately), then do a paired t test to compare visualization with no visualization at Weeks 1 through 4.

Within-Subject Modeling

This is a name I've devised for an approach that reduces or avoids the complexity of repeated-measures analyses. Basically, you derive a single measurement from the repeated measurements on each subject, then apply an appropriate simple analysis to the single measurement. The post-pre change score is the simplest example of a single measurement, and you would analyze the change scores with the unequal-variances version of the unpaired t statistic.

The approach works well for more complex derived measurements, too. For example, imagine you're looking at the effect of overtraining on recovery of heart rate following a standard bout of exercise. Let's say you record heart rate at half-minute intervals for three minutes. OK, that makes seven repeated measurements. Do you use repeated-measures ANOVA on these heart rates? Well, you could, I suppose. But what about if you want to fit an exponential decay curve on the heart rates, and extract the time constant. I defy anyone to deal with that within a repeated-measures model. It's only possible if you fit an exponential curve to each subject's data separately, extract a time constant for each subject, then use the time constant in your subsequent analyses. Hence the name *within-subject modeling*: you fit the same model to each subject and extract one or more parameters, which you then use for further analysis. r.

In fitting the model to each subject, you don't have to worry about distributions of residuals. Subsequent modeling with the parameters does need to be done properly, though. That modeling could be cross-sectional or longitudinal (repeated measures). For example, if the seven measurements of heart rate are taken on only one occasion for each subject, subsequent analysis of the parameter(s) describing the change in heart rate will be cross-sectional. But if the seven measurements are taken on several occasions, each subject provides several estimates of the parameter(s), so repeated-measures modeling will be necessary.

A real advantage of within-subject modeling is that most research students can do it without complex statistical analyses. Another advantage is that the analyses require fewer assumptions about the repeated-measures structure of the data, so the p values and confidence limits are more trustworthy. But mixed modeling, properly applied, is more powerful, especially when you want to include covariates in the analysis. See the slideshow on repeated measures for more examples and explanations of within-subject modeling.

Covariates in Repeated-Measures Analyses

By adding terms called **covariates** to the usual (fixed-effects) model, you can analyze for the following: the extent to which a subject characteristic accounts for **individual responses** to a treatment, the effect of the treatment on **trends in repeated sets of trials**, and the extent to which the effect of the treatment was due to changes in a putative **mechanism variable**. All is explained in the slideshow on repeated measures that I referred to on the first page on repeated measures. Click here to download the slideshow, which is an updated and extended version of an earlier slideshow on covariates in repeated measures.

Repeated-Measures Analysis of Troublesome Variables

In earlier pages on non-repeated-measures models, I showed how to deal with dependent variables that don't produce uniform normally distributed residuals. The same approaches apply to repeated-measures models. Thus, you will often need to log-transform or rank-transform a variable before analyzing it. When you rank-transform, make sure you do it to all the observations in one shot, not to each repeated measurement separately. The data will need to be in the form of one row per trial (as for mixed modeling), not one row per subject (as for ANOVA), for you to do the rank transformation correctly within an Excel spreadsheet.

An exact analysis of ordinal variables, such as those derived from Likert scales, requires repeated-measures logistic regression, but the analyses are difficult for newbies and the outcome statistics (odds ratios) are hard to interpret. For most variables, including even those with only two levels (yes or no, injured or not...), you can code each level of the variable as consecutive integers (0 and 1; 1, 2, 3, 4, and 5; and so on) and analyze it as if it was a well-behaved continuous normally distributed variable. Sure, the residuals are anything but normal, but as before, you can count on the central limit theorem to make the sampling distribution of the effect statistic normal, so the confidence limits or p values will be trustworthy. If responses for one or more groups are severely stacked up at one end of the scale, you will need a large sample size (possibly >20) for the central limit theorem to do its thing. I can't say exactly how many, but I hope to do some simulations to get an idea. The unequal-variances t test came through with flying colors for modest sample sizes (10-20) in my simulations with non-repeated-measures ordinal variables, and it will probably do equally well when applied to change scores derived from ordinal variables. We can't assume that mixed modeling will perform as well, because its method of estimation is different from that in the t test. Simulation will reveal all.

Nominal dependent variables can be analyzed by repeated-measures categorical modeling, if you want outcomes expressed as odds ratios. Otherwise treat each level of the nominal variable as a separate variable coded 0 or 1 (as I suggested under categorical modeling), then analyze each variable with conventional repeated-measures approaches. For example, you get schoolkids to tick one of four boxes representing the most important reason for playing sport. You collect the questionnaire, then show half of them a video aimed at convincing them that winning is (or isn't) everything. Finally you give them a fresh copy of the questionnaire to fill in. To what extent did the video change their attitude? You might even administer the questionnaire again a month later to see how the changes lasted. To do the analysis, treat each reason as a separate variable, code it 1 if the kid ticked it, or 0 if not, then use the unequal-variances t statistic to investigate differences in the changes between the group who saw the video and those who didn't. The magnitude of the outcome is the proportion of kids who changed their choice of the given reason.

Variables representing proportions or counts require root or arcsine-root transformation before you give them the usual repeated-measures analysis. The more exact approach is to use binomial or Poisson regression. Proc Genmod in SAS does it for repeated measures.

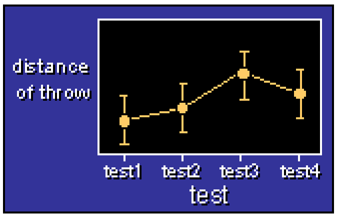
The next page deals specifically with the use of mixed modeling in the Statistical Analysis System.

Proc Mixed for Repeated Measures

On this page I introduce several examples of repeated-measures data, and I provide programs to analyze them using Proc Mixed in the Statistical Analysis System (SAS). Proc Mixed uses mixed modeling, a concept I have already introduced and which I will explain here in more detail soon. I will also explain covariance matrices. Meanwhile, here are some general remarks about the examples and the programs, followed by specific remarks for each example and links to the programs.

It's more than five years since I wrote these pages and the SAS programs! My approach has become more sophisticated during that time, but I haven't had a chance to update things here yet. If you are running SAS and would like advice and/or copies of my recent programs, contact me.

The data in each example are for athletes tested on several occasions to determine the distance they can throw an object, such as a javelin. (The figure shows the simplest example.) In each example, the program creates a sample of athletes drawn randomly from a population with a normal distribution of throwing ability. Next, the program generates normally-distributed within-subject random variation, which is simply the variation in performance that each subject experiences between tests. It then adds a change or changes in performance between some tests, for example changes resulting from a training program. Finally it uses Proc Mixed to analyze the data.



If you re-run any of these programs, the random variation between and within subjects will produce a slightly different outcome, so the data may not look exactly like what's in the figure accompanying each example. It's the same as repeating the study with a different sample of subjects. Try it and see, then play with the sample size, the between- and within-subject variation, and the magnitude of the change or changes in performance.

The main aim of the analysis is to calculate the changes in performance between tests, and the confidence limits or p values for the changes. Calculating the changes is usually easy: you just subtract the mean of one or more tests from the mean of one or more other tests. Calculating confidence intervals or p values is the hard part. That's when you need a procedure like Proc Mixed or analysis of variance. The procedure can also output the changes in performance, to save you doing it on a spreadsheet.

Here are the examples:

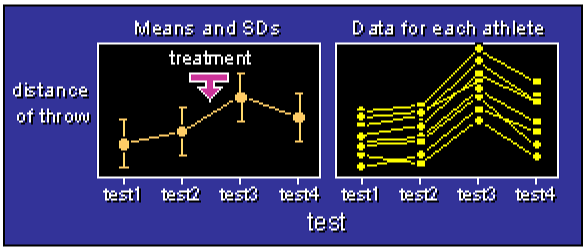
- Simple repeated measures
- Adding a control group
- Fitting polynomials
- Individual differences, and covariates

More to come soon!

To analyze your own data, you will need to get help from someone who knows how to set up a link from the SAS program to the data file on your computer. That means adding a *filename* statement that links to a data step containing an *infile* statement. I might provide examples of that soon, too.

Simple Repeated-Measures

The figure shows data for a single group of subjects who were tested four times. A treatment between test2 and test3 (for example, supplementing for a week with a potentially ergogenic nutritional like creatine) seems to have produced an increase in the distance of the throw.

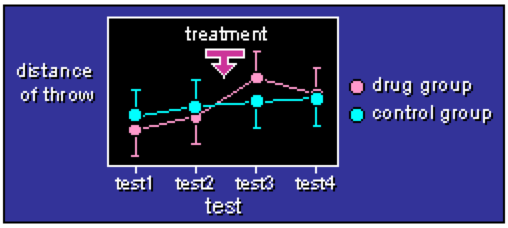


As noted above, the aim of the analysis is to calculate the mean increase in the distance of the throw between the tests, and its confidence limits or p value.

The data for each athlete show the kind of consistency of performance you expect when there are no problems with sphericity, as I discussed on the page devoted to three or more tests and no between-subjects effect. In the accompanying program, the analysis with an unstructured covariance matrix serves as a check for such problems. If there aren't any (and I didn't deliberately generate any), you use a covariance matrix with compound symmetry, as shown in the program. It won't make much sense until I provide a full explanation. Soon.

Adding a Control Group

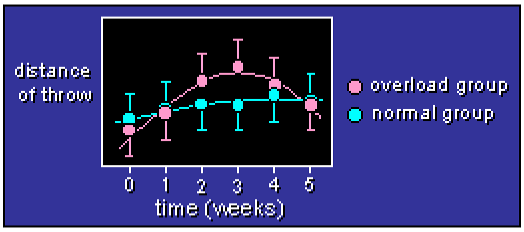
The data are the same as above, but this time there is a control group who don't get any special treatment between test2 and test3. The treatment could be something like a week of supplementing with creatine (the drug group) or an inactive substance (the control group).



Again, the aim of the analysis is to determine the confidence limits for the increase, but this time it's the increase in the drug group relative to (minus) the increase in the control group. Make sure you understand the concepts on the pages devoted to two trials plus a between-subjects effect and three or more trials plus a between-subjects effect before you try the SAS program.

Fitting Polynomials

The figure shows data for two groups of athletes. After a baseline test at time=0, one group did overload training, while the other group continued with normal training.

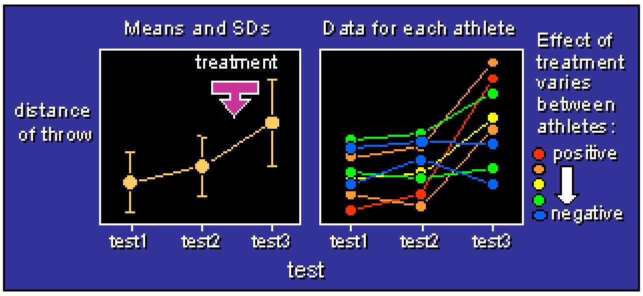


The aim of the analysis is to compare the enhancement in the overload group with that of the normal group at various times. You can also compare the linear and quadratic components of the trends between the groups. I'll add lots more on this topic soon.

See the earlier section on polynomial contrasts before you try the SAS program.

Individual Differences, and Covariates

When different subjects have a different response to a treatment, we say that there are **individual differences** in the response. I first touched on this possibility when I dealt with repeated measures with three or more trials and no control group. Here I've limited it to three trials only, but I have included a control group (not shown in the figure). In this example, a treatment between test2 and test3 has produced an overall increase in distance of a throw, but individual athletes differ widely in their response to the treatment.



The aim of the analysis is...

- To estimate the overall effect of the treatment. That's usually the change in the mean between test2 and test3. In the example, the distance of the throw is increased by 3 m following the treatment.
- To estimate the variability in the change in the mean between test2 and test3. That represents the individual differences, and it's best expressed as a standard deviation. In the example, the standard deviation is 2 m. The increase in distance is therefore 3 ± 2 m, which means that typical enhancements for individual athletes (to the nearest meter) might be 2, 4, 3, 0, -1, 3, 6...
    - Researchers sometimes calculate the standard deviation for the difference between test3 and test2. This standard deviation includes within-subject variation, so it is always larger than the true measure of individual differences.
- To account for the individual differences with a subject characteristic, such as age or percent of type 1 muscle fibers. In the example, the ± 2 m is attributed entirely to another variable, such that a change in that variable of one unit produces a change in the throw of 2 m.
- To calculate the confidence limits for all of the above. For example, the confidence limits for the 3 m might be 1 to 5 m, and the confidence limits for the 2 m might be 0 to 5 m.

The SAS program generates data for the treatment group, first without individual differences between test2 and test3, then with them. It also generates the data for a control group who have no shifts in the mean and no individual differences between test2 and test3. The analysis includes test1-test2 comparisons too.

This analysis is difficult, so I have included the output of the program and annotated it a little. Here's a summary of the output:

When there are no individual differences in the treatment group, the change in performance between test2 and test3 is 2.7 ± 0.3 units (mean ± SD); the 95% confidence limits for the mean are 1.6 and 3.8; the 95% confidence limits for the SD are -1.1 and 1.2.

With individual differences present, the change in performance between test2 and test3 is 2.5 ± 2.0 units (mean ± SD); the 95% confidence limits for the mean are 0.9 and 4.1; those for the SD are 1.3 and 4.4.

With individual differences present, and with a covariate that explains them included in the analysis, the change in performance between test2 and test3 is 2.7 ± 0.3 units (mean ± SD); the 95% confidence limits for the mean are 1.5 and 3.9; those for the SD are -1.1 and 1.2. The value of the covariate is 1.8 units of throwing performance per unit of covariate; its 95% confidence limits are 0.5 and 3.2.

REGRESSION TO THE MEAN

Do a fitness test on a bunch of subjects. Rank the subjects by their score and select the bottom half of the bunch. Retest the bottom half. The average score of the bottom half will probably improve somewhat on retest. Similarly, the average score of the top half will probably drop somewhat on retest. These changes in performance are called *regression to the mean.* The name refers to a tendency for subjects who score below average on a test to do better next time, and for those who score above average to do worse.

The group you select doesn't have to be the bottom or top half, and the test doesn't have to be the first one. Any group or even any subject you choose with an average score below or above the mean of all the subjects in a given test will probably move (regress) noticeably closer to the mean in another test. In general the scores don't move completely to the mean–they just get closer to it. It is therefore more accurate to call the phenomenon regression *towards* the mean.

OK, so low scorers tend to get better on retest, and high scorers tend to get worse? Well, no, actually. Depending on the nature of your data, the change in the scores towards the mean may be partly or even entirely a statistical artifact. If it's entirely an artifact, the true scores of the subjects don't really change on retest–it just looks that way. When that happens in, for example, a training study, your analysis might lead you to conclude that the least fit subjects got a big benefit from the training, whereas the fittest subjects got a smaller benefit or may even have got worse. In reality, all subjects may have increased in fitness by a similar amount, regardless of initial fitness. Your conclusion about the effect of initial fitness could be artifactual garbage.

Regression to the mean can lead to similar mistakes with repeated observation or testing of the health or performance of an individual. Consider a patient with a chronic health problem. Depending on the problem, symptoms can fluctuate in severity over a period of weeks or months, for no apparent reason. When the symptoms get really bad, the patient may try a new alternative therapy. The symptoms then improve, because they were bound to improve from their atypical severe level. The patient can be forgiven for thinking that the new therapy worked. Later on, the patient stops taking the new therapy, the symptoms get bad again, the patient takes the therapy again, the symptoms improve... Get the picture? You can imagine a similar scenario with an athlete who turns in a particularly bad performance, then does something about it. Whatever the athlete does, it's likely to work–artifactually. Now you can understand why there is so much snake oil on the shelves of drug stores.

I'll now deal with the nature of artifact when you analyze data from a group of individuals. The subsections are: the cause of the artifact, the magnitude of the artifact, and how to avoid the artifact.

Cause of the Artifact

Regression to the mean occurs because of noise (error) in the test score. Noise refers to the random fluctuations in a subject's score between tests–the typical or standard error of measurement. When you select subjects who scored low in one test, their scores were low partly because the noise just happened to make the scores low in that test. In other words, their true scores aren't really as low as the scores you selected. When you retest these low scorers, their scores in the retest will on average be their true scores (plus or minus the noise of the test, of course), so the scores are likley to rise. For the same reason, high scorers selected by you in one test are likely to fall on retest. Average scorers, on the other hand, are equally likely to rise or fall, so on average they don't change. The overall pattern is therefore for scores different from the mean in one test to regress towards the mean in another test.

The noise responsible for regression to the mean can come from two sources: the measuring instrument (technical or technological noise) and the subjects themselves (within-subject variation from test to test). I use the word *instrument* in its most generic sense: it could be a questionnaire, a device for measuring oxygen consumption, or whatever. If the noise comes solely from the instrument, regression to the mean is unquestionably an artifact. But if the noise is due to within-subject variation, there is a sense in which the regression to the mean is real. I'll explain with an example.

Suppose you administer two fitness tests several months apart. Several months is long enough for many subjects to change their fitness substantially: some will be fitter, some less fit. The "noise" in the test could be due almost entirely to these random but real within-subject changes in fitness. So when you select a subgroup with low fitness scores in the first test, the increase in their fitness in the second test is a real increase. If the increase is real, is there still a problem? Yes, because you could easily attribute the increase in fitness to something you had done between the tests, such as a training or nutritional intervention. The increase in fitness is real, but some of it was going to happen anyway, regardless of whatever you did. There are many papers in the literature in which the authors did not take account of regression to the mean when they claimed that their treatment produced a bigger increase in fitness on subjects with lower initial fitness.

Magnitude of the Artifact

There is a simple formula for estimating the magnitude of regression to the mean: on retest, scores will move towards the mean by a fraction given by $1 – r$, where r is the reliability correlation between test and retest scores. So, if $r = 0.9$, and you select a group of subjects whose average score is, say, 20 units above the mean, you can expect the average scores of those subjects to drop on retest by an average of $(1 – 0.9)x20$, or 2 units. Obviously, the smaller the r, the bigger the fractional move towards the mean. In the extreme case of $r = 0$, scores on retest regress on average all the way back to the mean. The $1 – r$ formula comes from the page Regression to the Mean at Bill Trochim's stats site. There is no proof or reference for the formula at his site, but it checks out with my simulations.

The retest correlation is involved in regression to the mean, because the correlation is a measure of the magnitude of the noise in the measurement. The formula for r is $(SD^2 – sd^2)/SD^2$, where sd is the within-subject standard deviation (the typical or standard error of measurement, or the noise) and SD is the usual between-subject standard deviation in either test. Rearranging, $1 – r$ = the fractional shift towards the mean $= sd^2/SD^2$. If sd is small relative to SD, there is little regression to the mean. At the other extreme, when SD = sd, subjects are effectively identical (the only difference between subjects is noise), so all pre-selected scores that differ from the mean will, on average, regress completely to the mean on retest.

The above formulae will allow you to estimate how much of a change in the mean is artifactual, but you should also be concerned about precision of the estimate, that is, the confidence limits for the true value. Bill Trochim does not have a formula for the confidence limits for the adjusted change in the mean. In the next section I will explain how to use the formula and get confidence limits.

How to Avoid the Artifact

Regression to the mean is a problem only when there is substantial noise in your dependent variable and you subdivide your subjects into groups that differ in their mean scores in one of the tests. Using the best test available is one way to reduce the effect of noise, but that won't reduce noise represented by real random changes in the subjects over the period between the tests. Of course, you can avoid the problem by not subdividing your subjects on the basis of their initial scores, but it is nice to know how a subject's initial score affects the outcome of a treatment. For example, you should find out if people with high initial scores get little benefit, because it's a waste of time using the treatment on such people. There are two approaches: correct the change scores using a formula, or use a control group. I once had an additional approach on this page, based on using the mean of each subject's pre- and post-test scores to subdivide the subjects. This approach eliminates regression to the mean, but it works properly only when the effect of the subject's pre-test score on the effect of the treatment is small (relative to the between-subject standard deviation in the pre-test). In general, you won't know how big the effect of the pre-test score is, so I have had to shelve this approach for the time being.

Correct the Change Scores

To use this approach, you will need to know either the retest correlation coefficient (r) or the within-subject variation (standard deviation, sd) for the dependent variable. Both must come from a reliability study with subjects and time between tests similar to those in your study. In my experience, an appropriate reliability study is often not available in the literature, so you will have to guestimate the reliability from less applicable reliability studies. Guestimate an sd rather than an r, because r is sensitive to the between-subject standard deviation of the subjects in the reliability study.

Armed with the reliability sd or r, proceed as follows. Subtract the pre-test mean of all subjects from each subject's pre-test score. Multiply that difference either by sd^2/SD^2 or by (1 - r), where SD is the usual between-subject standard deviation of your subjects in the pre-test. Now add the result (or subtract it when it is negative) to the post-pre change score for that subject. This corrected change score is free of the artifact. Use it in your analyses just as you would any change score. For example, do an unpaired t test to compare subjects with low vs high pre-test scores. Better still, plot the corrected change scores on the Y axis against the pre-test scores on the X axis. If the points form something like a line, derive the slope of the line as an estimate of the effect of pre-test score on the effect of the treatment.

Be aware that the confidence interval (or p value) for any effects involving the adjusted change score will be too small if the reliability study had a small sample size, owing to uncertainty in the estimate of sd or r. The effects, such as the difference between high and low scorers or the slope of the line in the examples above, will also be biased if the r or sd from the reliability study are substantially different from what your subjects would show in a reliability study with the same time between tests as in your study.

Use a Control Group

Using a control group is a better approach than correcting the change score. Actually, the approaches are fundamentally the same, because the control group is effectively the most appropriate reliability study for correcting the change scores. But don't use the control group to correct each subject's change score. Instead, analyze the effect of the pre-test score on the change score in both groups in the same manner, then compare the effect in the treatment group with that in the control group. The analysis will require a two-way analysis of variance (ANOVA) or covariance (ANCOVA). For example, suppose Ychng is the dependent variable representing each subject's post-pre change score, suppose Group has levels *control* and *intervention*, and suppose Prescore represents the pre-test score. The model is:

Ychng <= Group Prescore Group*Prescore.

If Prescore has the numeric values of the pre-test score, the model represents an ANCOVA. If instead you have coded the pre-test scores into two levels, such as *low* and *high*, the model is a 2-way ANOVA. Not that it matters what you call it--either way, you are interested only in the interaction term Group*Prescore, which yields the difference between the groups in the effect of the pre-test score on the change score (that is, on the effect of the treatment).

Watch out for non-uniform error! The standard deviation of the change scores in the treatment group may be larger than that in the control group, and there may be differences in the standard deviation for different values of Prescore, when there is a substantial true effect of pre-test score on the change score. The only way to take such non-uniform error into account properly is to use mixed modeling to specify different error terms for the different groups. Sorry, that's the way it is, guys. It's time you upskilled to the mixed model.

Generalizing to a Population:
ESTIMATING SAMPLE SIZE

Update Oct 2007: The following pages are now largely superseded by an extensive article on sample-size estimation published in Sportscience in 2006 with an accompanying slideshow and spreadsheet. I suggest you read the article first. There are a few formulae on the following pages that are not in the article.

I get more requests for information about sample sizes than about any other aspect of stats. I've come up with approaches and formulae that you won't find anywhere else, and that's not because they're wrong, either!

First, I'll deal with the need for the right number of subjects in a study: the main considerations are publishability of your findings, and the ethics of wasting resources. Then I spend a page on a new look at the traditional approach to what determines sample size, which leads to the formulae. I then present a new approach, sample-size estimation based on confidence intervals, with the good news that you need half the usual number of subjects. You'll almost certainly get away with an even smaller sample, if you use sample size "on the fly". Finally I encourage you to use simulation to work out sample size for complex designs or unusual outcome statistics.

THE RIGHT NUMBER OF SUBJECTS

With too few subjects, the confidence interval on your outcome is too wide to allow any useful conclusion. For example, you could get a big positive effect, but that's not very exciting or publishable if the wide confidence interval shows that the effect could actually be negative--in other words, if it's not statistically significant. Even if you observe a trivial effect, a small sample means a wide confidence interval, so the effect could still be large and positive or large and negative. Such results are hard for journals to accept.

With the right number of subjects, you have a narrow confidence interval on your outcome. It's sufficiently narrow that any worthwhile effects are statistically significant, which means you won't have missed anything. And even statistically non-significant results are publishable, because you can say that the effect is trivial. In my view, being able to say that an effect is too small to worry about is just as important as saying that it is large.

Too many subjects gives you a nice narrow confidence interval, but it's more narrow than you need. For example, it would be silly to have so many subjects that you could say a correlation lies between 0.725 and 0.729. That's far too much precision. Most of the time you'd be happy to say that it's 0.7, but not 0.8 or 0.6.

The ethical committees that grant approval for research projects are becoming more aware of the need to have the right number of subjects in a study. They require you to document your estimation of the required sample size, and they will not grant approval for research projects with too few or too many subjects. Small samples are unethical, because you can't be specific enough about the size of the effect in the population. Large samples are also unethical, because they represent a waste of resources.

You can sometimes justify a suboptimal sample size by arguing it's for a pilot study to determine reliability or validity, which in turn will allow you to estimate the sample size for a larger-scale study. A suboptimal sample size is also the starting point for sample size on the fly. But let's continue with the traditional approach and some formulae on the next page.

WHAT DETERMINES SAMPLE SIZE?

The traditional approach to estimation of sample size is based on statistical significance of your outcome measure. You have to specify the **smallest effect** you want to detect, the **Type I and Type II error rates**, and the **design** of the study. I present here new formulae for the resulting estimates of sample size. I also include new ways to adjust for **validity and reliability**, and I finish with sample sizes required for several complex cross-sectional designs.

I also advocate a new approach to sample-size estimation based on width of the confidence interval of your outcome measure. In this new approach, your concern is with the precision of your estimate of the effect, not with the statistical significance of the effect. The formulae on these pages still apply, but you **halve the sample sizes**.

The Smallest Effect Worth Detecting

I've already spent a whole page on magnitudes of effects. You should go back and make sure you understand it before proceeding. Or take a risk and read on!

Let's look at a simple example of the smallest effect worth detecting. Your research project includes the question of differences in height of adults in two regions. This sounds like a trivial project, but hey, the difference might be caused by a nutritional deficit, environmental toxin, level of physical activity, or whatever. OK, what difference in height would you consider to be the smallest difference worth noticing or commenting on? Almost everyone reading this paragraph will automatically start thinking either in inches or centimeters. So what's your choice? An inch, or 2.5 cm? Sounds like a nice round figure! Let's go with it for now.

To use my approach to sample-size estimation, you convert this difference into a value for the effect-size statistic. To do that, you divide it by the standard deviation, expressed in the same units. The standard deviation here is just the usual measure of spread, except that we have two groups. So let's assume we have an average of the standard deviation in both groups. Let's say it is 2 inches, or 5 cm. So, if you want to detect 2.5 cm, and the standard deviation is 5.0 cm, the smallest effect worth detecting is 2.5/5.0, or 0.5.

I'll talk about what I mean by *detecting* in a minute. First, more about the smallest effect. You'll discover shortly that the required number of subjects is quite sensitive to the magnitude of the smallest worthwhile effect. In fact, halving the magnitude quadruples the number of subjects required to detect it. So the way you decide on the smallest effect is important. How did we arrive at that minimum difference of 2.5 cm? In my experience, most researchers dream up a number that sounds plausible, just like we did here. Well, sorry, but you just can't do it like that. In fact, you don't have the freedom to choose the minimum effect. In all but a few special cases, it's the threshold for small effects on the scale of magnitudes: 0.2 for the Cohen effect-size statistic, 10% for a frequency difference, and 0.1 for a correlation. You need the same sample size to detect each of these effects, and as we'll see, it's 800 subjects for a simple cross-sectional study in the old-fashioned way of doing the figuring. It's even more than 800 when you factor in the validity of your variables. But don't panic. We'll also see that there are ways of reducing this number, sometimes drastically.

Type I and II Error Rates

Now, what do I mean by *detecting?* Simply that if the real difference between the two groups in the population is 2.5 cm (an effect size of 0.5), you want to be sure that it will turn up as statistically significant in the sample that you draw for your study. If it doesn't turn up as statistically significant, you have failed to detect something that you were interested in. Make sense? So our definition of *statistical* *significance,* and our idea of what it means *to be sure that it will turn up,* both impact on the required sample size.

First, statistical significance. The difference is statistically significant, by definition, if the 95% confidence interval does not overlap zero, or if the p value for the effect is less than 0.05. Values of 95% or 0.05 are also equivalent to a Type I error rate of 5%: in other words, the rate of false alarms in the absence of any population effect will be 5%. We don't have any choice here. It has to be 5%, or less preferably, but most researchers opt for 5%. If you want a lower rate of false alarms, say 1%, you will need more subjects.

Now, what about being sure that the effect will turn up? In other words, if the effect really is 2.5 cm in the populations, how sure do we want to be that the difference observed in our sample will be statistically significant? We don't have any choice here, either. We have to be at least 80% sure of detecting the smallest effect. To put it another way, the **power of the study** to detect the smallest effect has to be at least 80%. Or to put it yet one more way, the Type II error rate--the rate of failed alarms for the smallest effect--is set at 20% or less. That's one chance in five of missing the thing you're looking for!?! Sounds a bit high, but keep in mind that it is the rate for the *smallest* worthwhile effect. The chance of missing larger effects is smaller. Once again, if you want to make the error rate lower, say 10%, you will need more subjects.

Research Design

We're stuck with having to detect 0.2 for the effect-size statistic, 10% for a frequency difference, or 0.1 for a correlation. And we're stuck with false and failed alarms of 5% and 20%. All that's left now is how we're going to go about it: the research design. When it comes to sample sizes, there are only two sorts of research design: **cross-sectional** and **longitudinal**.

Cross-sectional designs include correlational, case-control, and any other design with single observations for each subject. Some so-called prospective designs, where subjects are followed up over time, are cross-sectional if there is only one value for each variable for each subject. Cross-sectional studies need heaps of subjects, and the number is affected by the validity of the variables.

Longitudinal designs include time series, experiments, controlled trials, crossovers, and anything else where the dependent variable is measured twice or more. The data have to be subjected to repeated-measures analysis. The usual thing with these designs is a measurement before and after you do something, to see if what you do has any effect. Whether or not you have a control group, it is always the case that subjects "act as their own controls", because there are always pre and post measurements on the subjects. Longitudinal designs generally need far fewer subjects than cross-sectional designs, depending on the reliability of dependent variable.

Sample Size for Cross-Sectional Studies

For variables with perfect validity, you can now look up tables or run special software to see how many subjects you need. (G*power is a great little free program for the purpose.) Or use the following simple formula I have worked out:

For Type I and II errors of 5% and 20%, the total number of subjects N is given by:

$N = 32/ES^2$, where ES is the smallest effect size worth detecting.

Example: for ES = 0.2, the total N is 800, which means 400 in each group for a case-control study or a study comparing males and females. So for our study of differences in height, we'd need 400 in each group.

What about if the outcome is a difference in the frequency of something in the two groups, for example the frequency of clinical obesity. The minimum worthwhile difference is 10% (e.g. 25% in one group and 35% in the other). You just think about that difference as being equivalent to an effect size of 0.2, and plug it into the formula: 400 in each group again.

And finally what about sample size to detect a correlation, for example the correlation between physical activity and body fat? Same story: 800 subjects to detect the minimum worthwhile correlation of 0.1, because a correlation of 0.1 is equivalent to an effect size of 0.2. For larger correlations use the scale of magnitudes to convert the correlation to an equivalent effect size, then plug it into the formula.

For the rare cases where you have the luxury of Type I and II errors of 1% and 10% respectively, the number is nearly double: **$N = 60/ES^2$**.

Validity of the variables can have a major impact on sample size in cross-sectional studies. The lower the validity, the more the "noise in the signal", so the more subjects you need to detect the signal. If the validity correlation of the dependent variable is v (Pearson, intraclass, or kappa), the number of subjects increases to **N/v^2**.

To detect a correlation between variables with validities v and w, the number is **$N/(v^2w^2)$**. Sample sizes may therefore have to be doubled or quadrupled when effects are represented by psychometric or other variables that have modest (~0.7) validity.

Sample Size for Longitudinal Studies

In our first example on this page, we had a cross-sectional design in which we were interested in the difference in height between people in two regions. Now, in a longitudinal design, we might want to know whether a stretching exercise makes people taller. Can you see that the same concept of minimum effect size still holds here? If we thought one inch was the smallest difference worth detecting between groups, then it has to be the smallest difference we would like to see as a result of our stretching exercise. (It might need a medieval rack to make people a whole inch taller!)

Once again we don't have a choice about that minimum effect: it's still an effect size of 0.2 standard deviations, and the standard deviation is still the usual standard deviation of the subjects. At the moment we have only one group of subjects, and the standard deviation before we put people on the rack is usually about the same as after the rack. So you can think about the minimum effect size as a fraction of either standard deviation. But note well: do **not** use the standard deviation of the before-after difference score.

Reliability of the dependent variable is the final piece of the jigsaw. The higher the reliability, the more reproducible are the values for each subject when you retest them, which makes it more likely you will detect a change in their values. So the higher the reliability, the less subjects you need to detect the minimum effect. Read the earlier section on sample size for an experiment for an overview of the role of typical error in sample-size estimation, and for an important detail about the conditions in a reliability study aimed at estimating sample size.

The rest of this section contains details of formulae that you may not need to worry about. You can use two forms of reliability in the formulae: retest correlation and within-subject variation.

Using the Retest Correlation

First, a couple of cautions. The retest correlation is for retests with the same time between the tests as you intend to have in your experiment. For example, if you are doing an intervention that lasts 2 months, you need a 2-month retest correlation. Don't use a 1-day retest correlation unless you have good grounds for believing that it will be the same as a 2-month retest correlation. Also, the spread between the subjects in your study has to be similar to the spread between the subjects in the reliability study. If the spread is different, the value of the retest correlation coefficient will be inappropriate. In that case you will need to calculate the appropriate value by combining the within (s) and between (S) standard deviations for your subjects using this formula:
retest correlation $r = (S^2-s^2)/S^2$.

Right, here's the strategy for working out the required sample size when you know the retest correlation:

- Work out the sample size of an equivalent cross-sectional study, **N**, as shown above. It's 800 in the traditional approach using statistical significance, or 400 using my new approach of adequate precision of estimation for trivial effects.
- Determine the reliability r of the outcome measure by consulting the literature or doing a separate study.
- For a simple design consisting of a single pre and post measurement on each subject, and no control group, the number of subjects is:
  n = (1 - r)N/2
  This formula applies also to simple crossover designs, in which subjects receive an experimental treatment and a control treatment. (One half get the experimental treatment first; the other half get the control treatment first.)
- If there is a control group, the total number of subjects required is:
  n = 2(1 - r)N
  Yes, you need *four* times the number of subjects when there is a control group, not *twice* the number. Hard to accept, I know.
- To take into account the validity of the outcome measure, multiply the above formulae by **$1/v^2$**, where v is the concurrent validity correlation (the correlation between the observed value and the true value of the variable). The simplest estimate of the concurrent validity is the square root of the concurrent reliability correlation for the outcome measure, so you simply divide the above formulae by the concurrent reliability correlation. In general, the concurrent reliability will be greater than the retest reliability

Using the Within-Subject Variation

You can also think about the difference between the post and pre means in terms of the within-subject variation (standard deviation). For example, if the performance of an individual athlete varies by 1% (the within-subject standard deviation expressed as a coefficient of variation), how many athletes should you test to detect a 1% change in performance, or a 2% change, or a 0.5% change? Here is the formula:

- To detect a fraction (f) of a within-subject standard deviation with 5% false alarms and 20% failed alarms:
  $n = 64/f^2$ with a full control group
  $n = 16/f^2$ for crossovers or experiments without a control group.
- Another way to represent the same formulae is to replace f with d/s, where d is the smallest worthwhile post-pre difference you want to detect, and s is the within-subject standard deviation:
  $n = 64s^2/d^2$ with a full control group
  $n = 16s^2/d^2$ for crossovers or experiments without a control group.
- **Remember to halve these numbers** when you justify sample size using the new approach based on acceptable precision of the outcome.

Example: You want to detect (p=0.05, 80% power) a 2% change in performance when the coefficient of variation is 2%. The corresponding value of f is 1.0, which means you'd need to test 16 athletes in a crossover design, or 32 in each of a control and experimental group. Or it's 8 or 16+16, if you justify sample size using precision of estimation.

What's the smallest value of f worth detecting? Is it 1.0? Not an easy question! To answer it, you usually have to bring in the between-subject variation one way or another. Why? Because you can't get away from the fact that the magnitude of a change in the value of a variable usually has to be thought about in terms of the variation in the values of that variable between subjects. That's what minimum worthwhile effect sizes are all about. For example, if the between-subject variation is 5%, the smallest difference worth detecting is

0.2*5% or 1%. So, if your within-subject variation of 2%, you have to chase an f of 0.5. But if the between-subject variation is 10%, the smallest worthwhile effect is 0.2*10% or 2%, so you chase an f of 1.0.

Once you bring the between-subject variation back into the picture, you have all the ingredients for expressing the reliability as a retest correlation, so you can use the formulae with the retest correlation. For example, a within of 2% and a between of 5% implies a retest correlation of $(5^2-2^2)/5^2$ or (25-4)/25 or 0.84. A within of 2% and a between of 10% implies a correlation of (100-4)/100, or 0.96. Use these correlations in the formulae for sample size and you'll get the same answers as in the formulae using f. But if you have a reasonable notion of the smallest worthwhile change in a variable without explicitly knowing the between-subject standard deviation or the correlation, use the formula with d and s (or f).

There is certainly one situation where it's better to use the within-subject variation: estimation of sample size in studies of athletic performance. When athletes are subjects and competitive performance is the outcome, the smallest worthwhile effect is an enhancement that increases the medal prospects of a *top* athlete, not the *average* athlete. For sports like track and field, this minimum effect is about 0.5 of the typical variation in a top athlete's performance between events. For example, if the typical variation between events is 1.0%, then you're interested in enhancements of about 0.5%. So if you use a lab test with the same typical error as the competitive event, f in the above formulae is simply 0.5, so you would need $64/0.5^2$, or 256 subjects for a fully controlled study. That's bad enough, but if your lab test has a typical variation of 2.0%, f is 0.5/2.0, which means 1024 subjects! Oh no! Clearly you need very reliable lab tests if you want to detect the smallest effects that matter to top athletes. See this Sportscience article for more information:

Hopkins WG, Hawley JA, Burke LM (1999). Researching worthwhile performance enhancements. Sportscience 3, sportsci.org/jour/9901/wghnews.html

Sample Size for Complex Cross-Sectional Studies

I'll deal with two groups of unequal size, more than two groups, and more than one independent variable. Anything else requires simulation.

Two Groups of Unequal Size

Up to this point I have assumed equal numbers in each group, because that gives the most power to detect a difference between the groups. But sometimes unequal numbers are justified.

The simplest case is where you have far more in one group than another. For example, you already have the heights for thousands of control subjects from all over the country, and you want to compare these with the heights of people from a particular region you are interested in. So, how many subjects do you need in that particular group? And the answer is... as few as one-quarter the usual number! But you will need to test, or have the data for, an "infinite" number of subjects in the other group for the number to be that low. How big is *infinite*? For the purposes of statistical power, about 5 times as many as in the special-interest group is close enough.

I have a formula, but to understand how to apply it will need a lot of thought. If you have samples of size n_1 and n_2, then your study will have the power *equivalent* to a study with a sample size of N equally divided between two groups, where:

$$\mathbf{N = 4\, n_1 n_2/(\, n_1 + n_2)}$$

For example, if you have data for 1000 controls (= n_1), and 800 (= N) is the number you would normally require for equal-sized groups, then the above formula shows that you need to test only 250 cases (= n_2). If you make n_1 very large, the formula simplifies to $N = 4\, n_2$, or $n_2 = N/4$, which is one-quarter the usual total number.

More Than Two Groups

Suppose we wanted to compare the heights of people in more than two regions. What should we do about the sample size? Do we need more than 400 in each region, less than 400, or just 400? And the answer is... it depends on what estimates or contrasts you want to perform.

If you are interested in comparing one particular region with another particular region, you will still need 400 in each of those regions to keep the same power to detect a difference. The fact that you have all those other regions in the analysis matters not a jot, I'm afraid. They don't increase the power of the design unless the number in each region is about 10 or less, which it never should be!

If you are interested in comparing one particular region with the mean of every other, you've got the usual two-group design, but with 400 subjects in the region of interest and 400 divided up equally into the other regions.

If you want to do every possible comparison between pairs of regions, or between pairs of groups of regions, things start to get complicated. As far as I can see, with six regions, say, only five completely independent comparisons are possible. So if you are concerned aboutinflation of the Type I error, you will need to apply Bonferroni's correction by reducing the p value to 0.05/5, or 0.01. Alas, a smaller p value means a bigger sample size. It's difficult to work out exactly what it should go up to, because somehow or other the inflated Type II error should also be taken into account. Certainly, nearly doubling the group size from the usual 400 would be a good start in this example, because as we've already seen on this page, that would be equivalent to a p value of 0.01 and a Type II error of 10%, instead of the usual 0.05 and 20%.

More Than One Independent Variable

Suppose you intend to measure half a dozen things like age, sex, body fat, whatever, and you want to know the effect of each of them on severity of injury in a particular sport. How many subjects do you need?

Before we get clever with complex models for this question, let's take in the big view. If we treat each variable as a separate issue, it should be obvious that there will be a problem with inflation of the Type I error: none of the variables you've measured might predict severity of injury in the population, but if you have enough variables, there's a good chance one will predict injury in your sample. So you'll need to reduce your p value using Bonferroni's 0.05/n, where n is the number of independent variables. This correction will be too severe if the independent variables are correlated, but I don't know how to adjust for that.

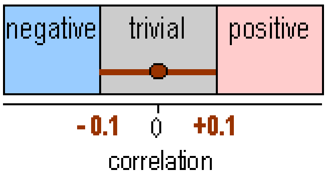
When you analyze the data, you should look at the effect of the independent variables separately to start with, but you will also end up using multiple linear regression, analysis of covariance, or some other complex model, with all the independent variables on the right-hand side of the model. As I explained on the first page devoted to complex models, you are now asking a question about how much each variable contributes to the severity of injury in the presence of (when you control for) the others. How many subjects do you need to answer this question? Theoretically the extra independent variables shouldn't make much difference, but I've checked by simulation to make sure. You need one extra subject for each extra independent variable. With five extra variables, that makes five extra subjects. Forget it. With a thousand or so subjects, five won't make any difference.

Here's a different problem involving more than one independent variable, where you don't have to worry about increasing the sample size to reduce the Type I error. Suppose you are currently predicting competitive performance from four lab and field tests, and you want to know whether it's worth adding an expensive fifth test to the test battery. For this sort of problem, you would model the data by doing a multiple linear regression, with the expensive test as the last independent variable in the model. So, how many subjects? It's a *specific* extra variable in this case, so there is no inflation of the Type I error, so the sample size is still about 800. But if all the field tests were in there on an equal footing, and you wanted to know which ones to drop out of the test battery, then it's back to the bigger sample size of the previous example. In this case you'd use stepwise regression with a reduced p value for entry of variables into the model.

SAMPLE SIZE BASED ON CONFIDENCE LIMITS

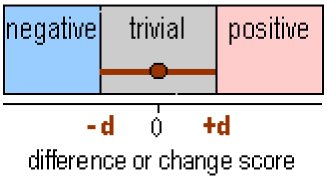
The traditional approach to sample size estimation requires the smallest worthwhile effects to be statistically significant. In other words, the approach is based on the relationship between the confidence interval and the null value of the outcome statistic. Why such a key role for the exact null value in the scheme of things? I believe it should be de-emphasized. If an effect is trivial, it doesn't matter whether it is zero, slightly positive, or slightly negative. And anyway, no real effects in nature are truly null.

So, I think it is more logical to use a sample size that ensures the true value of the outcome could not be substantially positive and substantially negative. In other words, the confidence interval for the outcome statistic should not overlap into values that are substantially positive and substantial negative. If it *does* overlap positive and negative values, you have to conclude that the true value could be positive or negative. To avoid this unsatisfactory conclusion, you need a small-enough confidence interval, which means a big-enough sample size.



You need the biggest sample size in this new approach when the observed value of the outcome statistic is zero or null. (You'll see why, eventually.) The figure shows an example for an observed correlation coefficient of zero and for ±0.10 as the smallest worthwhile effects. With a sample size of 400, the confidence interval for an observed correlation of 0.00 is -0.098 to +0.098, or just within ±0.10. A sample of 380 gives an exact fit to ±0.10. Thus with 95% confidence, a population correlation coefficient cannot be substantially positive and negative if the sample size is 380, which is half the value you're supposed to use with the traditional approach to sample-size estimation. The same argument and sample size apply to a descriptive study when the outcome is the difference between the mean of two groups or the relative frequency of something in two groups. The formulae on the previous page are still applicable, including those for longitudinal designs (experiments or interventions), but in all cases the sample sizes are halved. When the effects are large, you need even smaller samples. On the next page I show you how to get these sample sizes "on the fly".

The fact that the sample sizes using this new approach are half those of the old approach worries some statisticians. They say "your sample sizes give power of 50% rather than 80% for detecting the smallest effect". That's true, I admit, but we shouldn't be concerned with statistical significance any more. If you accept my rationale for basing sample size on precision of estimation, then you need half the sample size that you used to use. Or, to put it another way, people have been using samples that are twice as big as they needed. Sure, in one sense bigger samples are always better, because they give you more precision for the outcome. But too much precision represents an unethical waste of resources, so we've been getting an unethical amount of precision with our old sample sizes. Actually, the argument is more complex, because you really need several studies and even a meta-analysis to confirm a finding beyond reasonable doubt. No problem.

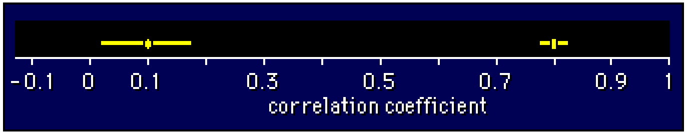


Here's another example, this time for an experiment. The figure shows an observed outcome of zero change and the more general case of the smallest worthwhile pre to post difference or change of ±d. If this is a crossover or a simple experiment without a control group, the confidence limits are ± root(2) x s/root(n) x $t_{0.975,\ df}$, where s is the within-subject standard deviation or typical error, n is the sample size, and t is the value of the t statistic for cumulative probability of 0.975 and df degrees of freedom (= n-1). Rearranging, $n = 2t^2s^2/d^2$. The value of t is approximately 2, so n is about $8s^2/d^2$. When n is small, t is a bit bigger than 2.0; for example, if d=s, the sample size is about 10 rather than 8. With a control group, the sample size is 4x as big.

SAMPLE SIZE "ON THE FLY"

CAUTION: Most of the material in this section is original and has not been subjected to formal peer review.

In the traditional approach to research design, you use a sample big enough to detect the smallest worthwhile effect. But hang on. You'll have wasted resources if the effect turns out to be large, because you need a smaller sample for a larger effect. For example, here is the confidence interval for a correlation of 0.1 with a sample of 800, which is what you're traditionally supposed to use to detect such correlations. Look what happens if the correlation turns out to be 0.8:



Far too much precision for a large correlation! So wouldn't it be better to use a smaller sample size to start with, see what you get, then decide if you need more? You bet! I call it sample size *on the fly,* because you start without knowing how many subjects you will end up with. The official name is **group-sequential design**, because you sample a *group* of subjects, then another group, then another group... in *sequence,* until you decide you've done enough.

I'll start this page with a potential drawback of group-sequential designs, **bias**. Then I'll describe a new method based on confidence intervals that is virtually free of bias. I'll detail the method on separate pages for correlations, differences between means, and differences between frequencies. On the last page I show how to use it for any design and outcome, I suggest what to say when you seek ethical approval to use this new method, and I give justification for a strong warning: **Do NOT use statistical significance to reach a final sample size on the fly**. I finish that page with a link for license holders to download a spreadsheet that will make calculations easier and more accurate.

Big Bias Bad

How come this method isn't in all the stats books? How come every ethical committee doesn't insist on it? Surely the less testing, the more ethical the method? Yes, but statisticians are wary of group-sequential designs, because the final value of the outcome statistic is **biased**. For example, if you are finding out how well two variables are correlated, and you adopt a group-sequential approach, the value of the correlation you end up with after two or three rounds of sampling will tend to be higher than it really is in the population. That's what bias means: samples *on average* yield a value for a statistic different from the population value. In this case the bias is high.

Where does this bias come from in a group sequential design? It's easy to see. You stop if you get a big effect, but you keep going if you get a small effect. You do the same thing again at Round #2, and Round #3, and so on: stop on a big effect, keep going on a small effect. Well, it's inevitable you'll end up with something higher than it ought to be, on average. But how high? That depends on how you start sampling and how you decide to stop. I have done simulations to show that the bias is substantial if you use statistical significance as your stopping rule, even for quite large initial sample sizes (see later). But the bias is trivial for the method I have devised using width of confidence intervals.

On the Fly with Confidence Intervals

My method for getting sample size on the fly came out of the conviction that confidence intervals are what make results interesting, not statistical significance. An effect with a narrow confidence interval tells you a lot about what is going on in a population; an effect with a wide confidence interval tells you little. And effects with narrow confidence intervals are publishable, regardless of whether they are statistically significant. So all we have to do is decide on the width of the confidence interval, then keep sampling until we get that width. That's it, in a nutshell. The rest is detail.

What is the appropriate width for the confidence interval? On the previous page I argued that, for very small effects, a narrow-enough 95% confidence interval is one that makes sure the population effect can't be substantially positive *and* substantially negative. In the case of the correlation coefficient, the width of the resulting interval is 0.20 units. It turns out that we can make this width the required width of our confidence interval for all except the highest values of correlation coefficient. Here's why.

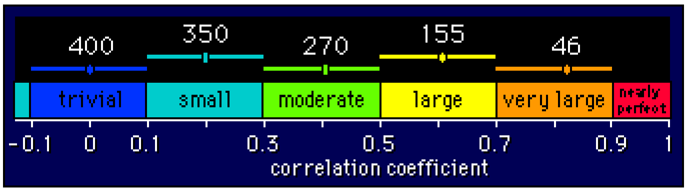
The threshold values of correlation coefficients for the different levels of the magnitude scale are separated by 0.20 units. This separation of 0.20 units must therefore represent what we consider to be a noticeable or worthwhile difference between correlations. It follows that the confidence interval should be equal to this difference: any wider would imply an uncertainty worth worrying about; any narrower would imply more certainty than we need. It's that simple!

Acceptable widths of confidence intervals for the other effect statistics are obtained by reading them off the magnitude scale. The interval for the effect-size statistic gets wider for bigger values of the statistic. The same is true of the relative risk and odds ratio, but confidence intervals for a difference in frequencies have the same width regardless of the difference.

A bonus of having a confidence interval equal to the width of each step on the magnitude scale is that the interval can never straddle more than two steps. So when we talk about a result in qualitative terms, we can say, for example, that it is *large,* or *moderate-large,* or*large-very large.* But fortunately we cannot say that it is *small-large* or similar, which seems to be a self-contradiction.

Actually, there are occasions when you need a narrower confidence interval. Remember that a correlation difference of 0.20 corresponds to a change of 20% in the frequency of something in a population group, so in matters relating to life and death an uncertainty of less than ±10% would be desirable. Correlations in the range 0.9-1.0 also need greater precision.

Right, let's get back on the main track. How come we need smaller samples for bigger effects? That's just the way it is with correlations. For the same width of confidence interval, you need less observations as the correlation gets bigger. Here's a figure showing the necessary sample size to give our magic confidence interval of 0.20 for various correlations:



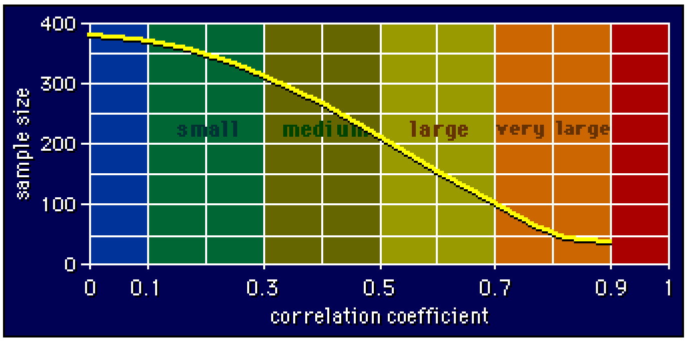
Notice that for very large correlations you need a sample size of only 50 or so, but to nail a correlation as being *small* to *very small,* you need more like 400. I'll now describe the strategy for correlations.

ON THE FLY FOR CORRELATIONS

The research question here is simply this: how linear is the relationship between two numeric variables, like weight and height? The extent of the linearity is captured beautifully by the correlation coefficient, so that's the outcome statistic we focus on.

As I've explained already on the previous page, to do the research on the fly, you keep sampling until the confidence interval for the correlation falls below 0.20. Here's how to go about it.

1. What's the maximum the correlation could ever be in the population you are studying? Start with a sample size that would give a confidence interval of 0.20 for that correlation. Use the graph below to read off this sample size. (The graph is just an adaptation of thefigure on the previous page, to allow you to get the sample size corresponding to any correlation.)



Curve is fitted to empirically-derived sample sizes that give confidence intervals of 0.20 for correlations in the "middle" of the steps of the magnitude scale. More information and simulation program.

2. The smallest sample size is about 45, which corresponds to correlations of 0.82 or higher. Correlations of 0.90 or more are a special case I'll deal with separately.
3. Do the practical work and calculate the correlation for the initial sample.
4. If the observed correlation is higher than the correlation corresponding to the initial sample size, the confidence interval must be less than 0.20, so the study is finished. If not, go to the next step.
5. Use the graph to read off the sample size that would give your correlation a confidence interval of 0.20.
6. Subtract the current total sample size from that sample size on the graph. The result is the number of subjects for the next lot of practical work.
7. Do the practical work, add the new observations to all the previous ones, then calculate the correlation for the whole lot.
8. If the correlations is higher than the previous correlation, the confidence interval must be less than 0.20. The study is finished. Otherwise go to Step 4.

Here's an example. You want to find the correlation between height and weight in a population. You think it will be very large, so you start with 45 subjects. You get a correlation of 0.71. The graph shows the corresponding sample size is about 95. So sample another 50 subjects (= 95 - 45), then calculate the correlation for all 95. You get 0.67, which means about 120 subjects. Off you go, test another 25. This time the correlation for all 120 subjects is 0.69. Stop. Publish.

The chance that you will finish on each round after the initial one is 50% or less, so the chance of having to go more than three extra rounds is about 10% or less. By then, my simulations show that typically you're adding only 5% to the total number of subjects, so you'll converge rapidly on the final correlation.

Confidence Limits for the Correlation

Naturally, you're expected to give the confidence limits of the correlation coefficient you end up with. If your stats program doesn't generate them, you'll have to use the Fisher z transformation: $z = 0.5\log[(1 + r)/(1 - r)]$. The transformed correlation (z) is normally distributed with variance $1/(n - 3)$, so the 95% confidence limits are given by $z \pm 1.96/\text{sqrt}(n - 3)$. You then have to back-transform these limits to correlation coefficients using the equation $r = [(e^{2z} - 1)/(e^{2z} + 1)]$. This is standard stuff for statisticians, but as a mere mortal you'll be struggling. I've set it up on the spreadsheet for confidence limits.

More on the Initial and Final Sample Sizes

You will be tempted to start with 45 every time, hoping that you won't have to do any more. But funnily enough, starting with this small sample, you could end up testing more subjects than necessary! For example, if the correlation in the populations is moderate (~0.4), a sample of 45 will sometimes produce a small correlation (~0.2), and when that happens you're supposed to test about 300 subjects on the next round. But if you had opted for, say, 200 to start with, you'd be unlikely to have to test another 150 on the next round.

But there's an acceptable cheat's way around this problem that allows you to start with 45 every time. All you do is set an upper limit on the number of subjects you will test, then take the limit off. For example, start with 45 subjects, but if the next round requires 250 more, you test only 100. Then you work out how many more you need from the total of 145, and test them.

However you do it, you'll get there in the end. And the answer will be trustworthy: I've found that the greatest bias occurs for correlations around 0.7-0.8, but it is only 0.01. This amount of bias--5% of the confidence interval--is negligible. What's more, the bias is insensitive to the initial sample size, and there is no noticeable extra bias when you set reasonable limits to the sample size on each extra round of sampling (e.g. 100 on the first round, 200 on the second and/or higher rounds). So even if you haven't got the resources to go to the full 400 subjects, you can still get a practically unbiased estimate of the correlation, albeit with a less-than-ideal confidence interval for the smallest correlations.

Adjusting for Imperfect Validity

Imperfect validity of one or both variables in the correlation degrades the apparent relationship between them. If the correlation you're chasing has a true value of r, and the validities are v and w, then the correlation you will observe, say r', is r·v·w, which is smaller than r. But when you write up the study, you will say that the correlation in the population is r'/(v·w). In other words, you inflate the observed correlation by a factor 1/(v·w), which is, or course, greater than 1. Uh huh! So that means the confidence interval is also inflated by the same factor. Curses, that means we'll need more subjects to make sure the larger correlation still has a confidence interval of 0.20. In fact, the final number of subjects is inflated by a factor $1/(v^2w^2)$. This factor popped up in the estimation of sample size using the traditional approach. You can use it on the fly, but it's a bit tricky. You have to inflate all sample sizes by the same factor on the way to detecting the correlation.

Here's an example. Suppose the validity correlations are 0.90 and 0.80. Overall that's 0.72, and 0.72^2 is 0.52. So start with 45/0.52 or 87 subjects. Suppose you get a correlation of 0.35. For perfect validity that would be a correlation of 0.35/0.72 or 0.49. On the graph that's equivalent to 220 subjects, but that's for perfect validity, so you need 220/0.52 or 423 subjects. So test 423 - 87 = 336 subjects. And so on. Mind-boggling, I'm afraid. It's all much simpler if you use the spreadsheet!

Nearly Perfect Correlations

You'll notice I've omitted correlations in the *nearly perfect* range on the graph for estimating sample sizes. If a correlation is this high, the relationship it represents is probably a reliability or a validity, or it may be a linear relationship used for predicting something. Confidence intervals less than 0.20 are needed for these correlations. Exactly how much less is a difficult question that I'm still working on.

Meanwhile, start with a sample of about 15 and see what you get for the correlation and for its confidence limits. You'll almost certainly find that the lower confidence limit is too low, unless you're lucky enough to get a correlation of 0.98 or 0.99. So you'll need more subjects. Estimate the sample size for the next round using the rule that the width of the interval is approximately inversely proportional to the square root of the sample size. Then test the extra subjects, recalculate the correlation and its confidence limits, and go to another round if necessary.

For example, let's suppose you get a correlation of 0.91 with 15 subjects. The 95% confidence limits are 0.97 and 0.75. Well, if the correlation is really 0.97, that's great for every possible purpose. But 0.75 is hopeless for applications requiring an almost perfect correlation! Obviously you need to narrow down the confidence interval. Halving the interval would help, which means a total of 4x as many subjects, or another 45. Test them, add them to the original 15, then recalculate. Suppose you get 0.93. The 95% confidence limits are now 0.96 and 0.89. Whether you stop at this point or go to another round of testing depends on whether 0.89 makes a big difference compared with 0.96, for the application you have in mind. I'd stop there if I was defining the validity of a variable for the purpose of seeing how many extra subjects I might need in a big cross-sectional study. I'd want to narrow down the interval a bit more if I wanted to use the underlying linear relationship to predict things like body fat from skinfold thickness. And I'd probably want to narrow it down more if the correlation was a reliability I was using to predict a sample size in a longitudinal study, using the old-fashioned approach.

For another example, imagine that you got a correlation of 0.98 with your initial sample of 15. The confidence limits are 0.96 and 0.99. No need to test any more subjects!

ON THE FLY FOR DIFFERENCES BETWEEN MEANS

How many subjects do you need to see how females and males differ in strength? For cross-sectional studies like this, where you're looking at the difference between means of two groups, you use the same method as for correlation coefficients,. The main difference is that you use the effect-size statistic rather than the correlation coefficient. You have to calculate its value each time yourself, because current stats programs don't.

A variant of the method also works for longitudinal studies--for example, where you want to compare the strength of females before and after they take a hormone that makes them like males. We'll come to those in a minute.

Cross-Sectional Studies

As before, you keep sampling until you get a sample size that gives an acceptable confidence interval for the outcome statistic, the effect size. But calculating the effect size causes a bit of a problem.

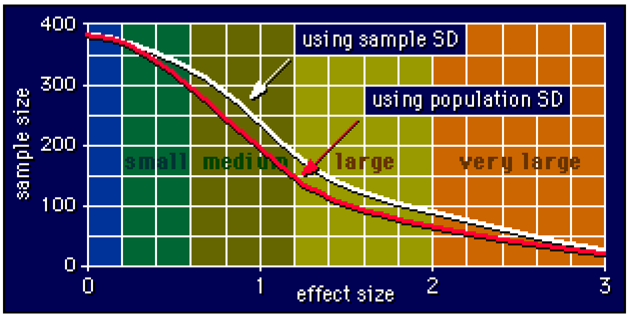
Recall that the effect size is the difference between the means divided by the average standard deviation of the two groups. Well, the standard deviation calculated from your sample introduces some error of its own, which contributes to error in the effect size. So if you have a more accurate estimate of the population standard deviation from elsewhere, use it instead of the value from your sample. It can mean 40 less subjects, depending on how big the effect is. It also makes calculating the confidence limits of the effect size a lot easier.

Here's the method, for either standard deviation.

1. Start with about 40 or more subjects (20 or more in each group), knowing that you might have to go to nearly 400 if the effect turns out to be trivial.
2. If validity is less than perfect, inflate the starting number and make further adjustments as described for correlations. Not an easy task!
3. Do the practical work, then calculate the difference between the means.
4. Convert the difference between the means into an effect size by dividing it by the standard deviation. Use the population standard deviation if available, or calculate the average standard

deviation of your two groups if not. Make sure you average the variances of the two groups, then take the square root to get the average standard deviation. Don't forget to log or rank transform the dependent variable if necessary (which may complicate the use of any available population standard deviation).

5. Use the appropriate curve on this graph to read off the sample size needed to give an acceptable confidence interval to your effect size. Or license holders can use the spreadsheet, which also adjusts for validity.



Each curve was drawn through the point in the middle of each step of the scale that gives a confidence interval just spanning the step. See the simulation program for more information.

6. If the sample size from the graph is less than the initial sample size, the confidence interval is already narrower than the acceptable confidence interval, so the study is finished. Otherwise go to the next step.
7. Subtract the current total sample size from that sample size on the graph. The result is the number of subjects for the next lot of practical work. You can "cheat" by doing the practical work on less than this number, if it's a big leap to nearly 400 from the previous number. This trick will help make sure you don't test too many subjects, as I described for correlations. If the effect turns out to be trivial, you will still eventually end up with nearly 400, of course!
8. Divide the extra subjects equally into the two groups, do the practical work, add the new observations to all the previous ones, then calculate the effect size for the whole lot.
9. If the effect size is greater than the previous value, the confidence interval must be narrower than the acceptable confidence interval, so the study is finished. Otherwise go to the next step.
10. Use the graph to read off the sample size needed to give an acceptable confidence interval to your effect size. Now go to Step 7, and continue in this fashion until you reach a sample size that gives an acceptable confidence interval.

Cool! You've got a value for the effect size, and you've done it with the minimum number of subjects, and it's practically unbiased by doing it on the fly, and you know that its confidence interval is narrow enough that it can't overlap more than two steps (colors) on the qualitative magnitude scale. But what exactly *is* the value of the confidence interval? If I end up refereeing your paper, I'll insist you put it in! Here's how to get it.

Confidence Limits for Effect Size (Cross-sectional Studies)

If you used the *population* standard deviation for sample sizing on the fly, get your stats program to produce the confidence interval of the raw difference between the means for the final sample. Divide this confidence interval by the population standard deviation and you have the exact confidence interval for the effect size. The observed effect size sits symmetrically in the middle of this confidence interval. If you can't get your stats program to produce the confidence interval of the difference score, the confidence interval of the effect size is given exactly by 2t·sqrt(4/N), where N is the total sample size, and t is the value of the t statistic for N - 2 degrees of freedom and cumulative probability 0.975. The value of t is near enough to 2.0.

If you used the *sample* standard deviation on the fly, the resulting effect size is biased a bit high for small total sample sizes (N). Adjust out the bias using this formula:
unbiased ES = (observed ES)(1 - 3/(4N - 1)).
Now use the following fairly accurate formula to calculate the 95% confidence interval for the unbiased effect size:
95% confidence interval = 4sqrt(4/N + ES^2/(N - 2)).
The confidence limits are therefore given fairly accurately by:
ES ± 2sqrt(4/N + ES^2/(N - 2)),
but that's only for ES<1.0. For larger values of ES, the limits start to sit asymmetrically about the observed value of ES. Then the going gets really tough. The exact values of the confidence limits are given by t·sqrt(4/N), where t is the value of the non-central t statistic with degrees of freedom = N - 2, non-central parameter = ES·sqrt(N/4), and cumulative probabilities of 0.025 and 0.975 for the lower and upper limits respectively. Only advanced stats programs can produce values for the non-central t statistic.

All the above formulae are available on the spreadsheet, with the exception of the non-central t statistic. I will add it when Excel does.

Reference for formulae:
Becker, B. J. (1988). Synthesizing standardized mean-change measures. British Journal of Mathematical and Statistical Psychology, 41, 257-278.

Longitudinal Studies

In longitudinal studies we are interested in seeing how much a mean changes as a result of an intervention, for example the change in swimming speed resulting from a new training technique. We compute the mean of the post minus pre scores to get the change. Now, the confidence interval of that post-pre difference is extremely sensitive to the reliability of the outcome measure. For almost perfect reliability, the confidence interval is very narrow compared with what it would be in a cross-sectional study, so we can get away with using a far smaller sample size than in a cross-sectional study.

But if we use the sample standard deviation to calculate the effect size, there is a major hitch. With the small sample sizes that are possible, the error in the standard deviation is proportionally larger, so the confidence interval of the effect size ends up large after all, so we lose the benefit of the high reliability and end up with larger sample sizes again. The calculations are difficult, too.

On the other hand, if we know or can guess the population standard deviation, all is saved. So I'll concentrate on a method that uses the population standard deviation, then deal briefly with the use of the sample standard deviation.

Using *Population* SD to Calculate Effect Size and its Confidence Limits

This method works for the effect size in cross-sectional or longitudinal designs of any kind, and for any estimates/contrasts between levels of within and between factors. Wow! The only challenge for you is to coax your stats program to produce a confidence interval for the raw difference between the means, or for whatever estimate/contrast you are interested in. You then simply convert that to a confidence interval for the effect size by dividing it by the population standard deviation, see if the confidence interval is narrow enough, and if it's not, work out how many more subjects you'll need.

This paragraph may confuse you. Skip to the method in the next paragraph if it does. To get an idea of the kind of sample sizes you can end up with, you can apply the formulae I presented earlier for the effects of reliability on sample size. The only difference is, the "N" in the formulae is now the sample size you would need for a cross-sectional study, as shown by the curve in the above graph for population SD. So, the sample size for a longitudinal study with a single pre and post measurement and no control group is N(1 - r)/2, where r is the reliability correlation coefficient. If there is a control group, you need twice as many in both groups, or 2N(1 - r) altogether. Let's check out an example on the graph above. If your effect size turns out to be in the middle of the medium range, you'd end up needing about 200 subjects for a cross-sectional study. But if your reliability is 0.9, that'll come down to 10 subjects for a study without a control group! Fantastic! If your reliability is 0.95--not out of the question for some outcome measures--you'd need only 10 subjects in each group of a properly controlled study. It will be even less for larger effects. But check the graph: you might still have to go to nearly double that number if the effect size turns out to be zero.

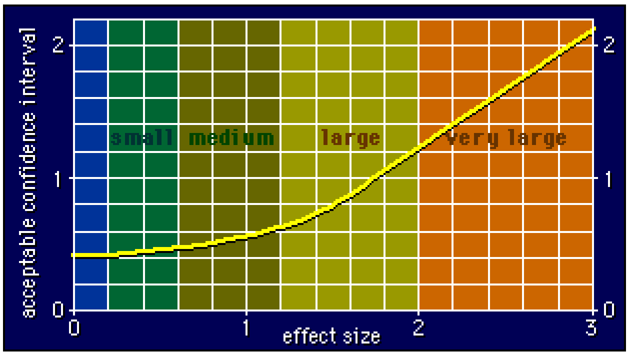
OK, here's how the method works. It's the usual iterative process, but this time it relies on the fact that the width of the confidence interval is inversely proportional to the square root of the sample size.

1. If you have high reliability and the effect is very large, ridiculously small sample sizes are possible. But you have to be careful when you're down to five or so subjects, because you might end up with a sample that is not typical of the population. Papers do get published with six subjects in each group, but I'd feel safer with a minimum of eight. If your reliability is unlikely to be better than 0.9, or your effects are probably small-medium, start with 10-15. That means 10-15 in a single group if it's a study without a control group, or 10-15 in each group if there's a control group or several experimental groups.

2. Do the practical work, then crunch the numbers to get the difference between the means of interest, or do whatever other estimate/contrast you like. By the way, when you have a control group, the difference you want is the post-pre difference score for the experimental group minus the post-pre difference score for the control group.

3. Get your stats program to produce the confidence interval for the difference. Convert it into effect-size units by dividing it by the population standard deviation. Convert the difference itself into an effect size in the same way.

4. Use this figure to read off the acceptable confidence interval for your effect size, or use the spreadsheet, which also performs subsequent calculations and takes account of less-than-perfect validity.



The way I derived this curve and validated the on-the-fly method is described on separate pages for longitudinal studies without a control group and with a control group.

5. If your observed confidence interval is less than the acceptable confidence interval, the study is obviously finished. If not, go to the next step.

6. Divide your observed confidence interval by the acceptable confidence interval, square the result, then multiply it by the total number of subjects you have tested. That's your next target total number of subjects.

7. Subtract the current total sample size from that target total. The result is the extra subjects for the next lot of practical work. Divide them equally into the groups, if there is more than one group.

8. Do the practical work, add the data to the previous data, then go to Step 3.

The confidence interval of the final effect size is no problem, this time. You've been calculating it all along.

Using *Sample* SD to Calculate Effect Size and its Confidence Limits

You go through the same steps as for use of the population SD, but you have to calculate the confidence interval for the effect size using the sample SD. You then use this calculated confidence interval in Step 3. Here's how to calculate the confidence interval. If you have a control group, I will assume it has the same number of subjects as the experimental group.

- Calculate the effect size using the average variance, as described in Step 4 for cross-sectional studies. If you've got a control group too, average all four variances before you take the square root.
- Correct out the bias in the effect size, using this formula:
  unbiased ES = (observed ES)(1 - 3/(4N - 1)), where N is the total sample size (experimental plus any control).
- Calculate the reliability (r) of the dependent variable, preferably as an intraclass correlation, but otherwise as a Pearson correlation. Do it using the experimental data: a shift in the mean due to the intervention does not affect the reliability. If you have a control group, use the average reliability of the control and experimental group. A proper average should be done via the Fisher z transform, but if the correlations are fairly similar it won't matter if you just take the usual mean.
- Calculate an approximate confidence interval for the ES using this formula:
  4sqrt(2(1 - r)/N + ES^2/(2(N - 1)) if there is no control group, or
  4sqrt(8(1 - r)/N + ES^2/(2(N - 4)) if there is a control group.

When you've done your sampling on the fly, the confidence limits of the effect size, for effect sizes <1, are given by the final effect size ± half the confidence interval. For effect sizes>1 there is that problem of the confidence interval not sitting symmetrically around the effect size...

For studies without a control group, the exact values of the confidence limits are given by t·sqrt(4(1 - r)/N), where t is the value of the non-central t statistic with degrees of freedom = N - 2, non-central parameter = ES·sqrt(N/(4(1 - r)), and cumulative probabilities of 0.025 and 0.975 for the lower and upper limits respectively.

For studies with a control group, the exact values of the confidence limits are given by t·sqrt(8(1 - r)/N), where t is the value of the non-central t statistic with degrees of freedom = N - 2, non-central parameter = ES·sqrt(N/(8(1 - r)), and cumulative probabilities of 0.025 and 0.975 for the lower and upper limits respectively.

If only the stats programs would do these calculations...! I've put most of them on the spreadsheet, but I can't do anything about non-central t statistics until Excel does.

If you've got this far, you will no doubt be interested in a simulation that validates the on-the-fly method for the case of no control group. It includes an empirical check on the formulae when there is a control group.

Now for something a little easier: on the fly for differences in frequencies.

ON THE FLY FOR DIFFERENCES BETWEEN FREQUENCIES

Now you're interested in things like the difference in the frequency of injury in two groups. For example, if you found that 47% of runners and 15% of cyclists have an injury each year, how many runners and cyclists would you have needed in the study for the result to be publishable? Publishability depends on the confidence interval for the difference between the frequencies, of course. Obviously 10 runners and 10 cyclists would give a hopelessly unpublishably wide confidence interval, and equally obviously 10,000 of each has got to be overkill!

You can use sample size on the fly to get the minimum number of subjects, but you don't get quite the same saving as for correlations or means. I've used simulation to see how many subjects you need to give acceptable confidence limits for a wide range of frequency differences. I've found that it's at least 100 subjects, even for very large effects, so that will have to be our starting number.

The other thing we need for sample size on the fly is an acceptably narrow confidence interval for the outcome statistic. It's straightforward if we use the difference in frequencies as the outcome, but it gets really complicated if we use relative risk or the odds ratio. Let me explain with the example of injury in runners and cyclists.

The difference in rates of injury can be expressed either as a difference in the percentage rates (47 - 15 = 32%), or as a relative risk of injury (runners have 47/15 = 3.1 times the risk of cyclists). The acceptable width of the interval for a difference in the percentage rates is a fixed 20%, as I explained earlier. In our example the difference is 32%, so the required publishable confidence limits are 22% to 42%. Expressed as a relative risk, these frequencies correspond to 3.1, with confidence limits 2.1 to 5.1. But suppose the original frequencies were 67% and 52%. The difference in frequencies is still 32%, and the acceptable confidence limits on this difference are still 22% to 42%. But now the corresponding relative risk is 1.9, with confidence limits 1.5 to 2.5 What a mess! The odds ratio misbehaves in the same way for case-control data.

So here's the method, based on the confidence interval for the differences in frequencies between the groups, expressed as percents.

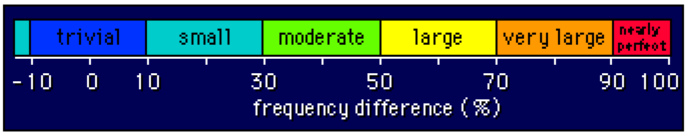
1. Start with a sample size of 100 (50 in each group).
2. Do the practical work. That often means interviewing subjects, or waiting for them to get sick or injured!
3. For each group, count up the number of subjects with the thing you're interested in (e.g. an injury). Express it as a percent for each group, then subtract one from the other. That's the frequency difference.
4. You are aiming for a confidence interval of 20% for that frequency difference. What's the current confidence interval? Once again, stats programs don't produce it, but it can be derived from something called the normal approximation to the binomial distribution. Here it is, in the right percent units:
   $392\cdot\text{sqrt}((n_1(n - n_1) + n_2(n - n_2))/n^3)$,
   where n_1 and n_2 are the numbers (not %) of subjects with the thing of interest in groups 1 and 2, and n is the number of subjects in EACH group (50 to start with). My simulations show that this formula is surprisingly accurate, even for very low n_1 and n_2 (~1%, with only 50 in each group!).
5. If this confidence interval is less than 20, the study is finished. Otherwise go to the next step.
6. To estimate the number of subjects required to bring the confidence interval down to 20, we make use of the fact that the width of the confidence interval is inversely proportional to the square root of the sample size. So, divide the current confidence interval by 20, square the result, and multiply it by the current number of subjects in each group. The result is the predicted total number of subjects needed.
7. Subtract the current number of subjects in each group from the predicted number. The result is the number of subjects needed in each group for the next round of practical work. You can "cheat" by doing the practical work on less than this number, if it's a big leap to nearly 400 from the previous number. This trick will help make sure you don't test too many subjects, as I described for correlations and effect sizes. If the difference in frequencies turns out to be trivial, you may still end up with a final sample size of up to 200 in each group.

8. Do the practical work on the extra subjects, add them to all the previous subjects, then go to Step 3.

All computations in the above procedure are available on the spreadsheet, which includes the case of unequal numbers of subjects in the groups.

How do you present the final outcome? Obviously you need to show the frequency of the injury or whatever as a percent in the two groups. You should also show the confidence limits for the difference in frequencies (confidence limits = the difference in frequencies ± half the confidence interval, which you will have calculated in the last iteration of the sampling process). That's it, as far as I am concerned, but for a clinical journal you may have to show a relative risk or an odds ratio. If the editor of the journal insists on one or other of these effect statistics, put it in, and get your stats program to calculate its confidence limits.

To describe the outcome of your research in qualitative terms, check where the confidence limits of the frequency difference fall on the scale of magnitudes. Here's a version of it for frequency differences:



For example, if the limits are 22% and 42%, the effect is small-moderate; if they are -5% and 15%, the effect is trivial-small, and so on.

ON THE FLY: MISCELLANEOUS

On this last page devoted to sample size on the fly, I explain how to use it for any design and any outcome statistic. I then suggest what to say to the ethical committee when you apply for approval. I also warn you not to use statistical significance for sampling on the fly.

ON THE FLY FOR OTHER DESIGNS

Whatever the design and whatever the outcome statistic, if your stats program can produce a confidence interval for the outcome statistic, you can sample on the fly. Here is the procedure. First I explain how to do it for outcome statistics whose confidence interval has a width proportional to the square root of the sample size.

1. Decide on an acceptable width for the confidence interval of your outcome statistic. If the outcome statistic is a correlation coefficient or a frequency difference, there's no problem: the acceptable widths are 0.20 for a correlation and 20% for a frequency difference. Or you can choose a narrower confidence interval for the frequency difference, if it's a matter of life and death.

2. If the outcome is an effect size, the width depends on the value of the effect size, as shown in the figure on the page devoted to differences between means.

3. For other outcome statistics, work out what seems like a reasonable acceptable width for its confidence interval. It may depend on the magnitude of the statistic. For example, the relative risk and odds ratio clearly need wider confidence intervals for larger values of the statistic.

4. Start with a reasonable sample size. If it's a cross-sectional design, it will probably be around 50 subjects. If it's a longitudinal design and the outcome is derived from the repeated measure, then 10 or so will probably do the trick, provided the reliability isn't too bad.

5. The rest will sound familiar! I've copied it from the method for means in longitudinal studies.

6. Do the practical work.

7. Calculate the value of the outcome statistic and its confidence interval.
8. If your observed confidence interval is less than the acceptable confidence interval, the study is finished. If not, go to the next step.
9. Divide your observed confidence interval by the acceptable confidence interval, square the result, then multiply it by the total number of subjects you have tested. That's your next target total number of subjects.
10. Subtract the current total sample size from that target total. The result is the extra subjects for the next lot of practical work.
11. Do the practical work, add the data to the previous data, then go to Step 7.

If the confidence interval of your outcome statistic is not inversely proportional to the square root of the sample size, replace Step 9 with the following elegant procedure (which allows you to work out the relationship between sample size and the width of the confidence interval):

1. Make a double-sized sample by simply duplicating the sample and adding it back in with itself.
2. Analyze the double-sized sample with the stats program to get the confidence interval.
3. Add the new sample to itself to get a sample four times as big, then analyze it for the confidence interval.
4. Repeat to analyze a sample eight times as big, and 16 times as big.
5. Now plot sample size vs confidence interval, connect the points with a smooth curve, and read off the sample size corresponding to an acceptable confidence interval for the value of the outcome statistic from Step 7. Now go to Step 10.

ON THE FLY FOR THE ETHICAL COMMITTEE

You need to convince the ethical committee that you have the resources to go to the usual large number of subjects, if the effect turns out to be small. So you will have to provide an estimate of the worst-case sample size. You'll have to justify it using my approach with confidence intervals (which requires half the usual number), because you can't let statistical significance get anywhere near sample size on the fly. The two do not mix, as we'll see shortly.

To do a cross-sectional study properly, you must have the resources to test hundreds of subjects, if necessary. Don't forget to take into account known or guessed validities, which could push the number up by a factor of two or three.

For a longitudinal study, reliability is crucial for calculating how many subjects you might need. If you don't know or can't guess the reliability, you have to tell the committee that you simply don't know how many subjects you might end up with. So tell them that testing 10 or so subjects per group will be enough to detect large effects if the reliability is almost perfect, and it will give you enough data to estimate roughly the final sample size otherwise. Indicate the total number you will be able to test, and admit that this number may not be enough if the reliability turns out to be low. You will end up with a confidence interval that is wider than optimum, but the result may still be publishable. There's nothing you can do about it, and there's no ethical justification for your application to be refused, if you've got everything else right. After all, if no-one knows the reliability, someone has to start testing to find out how many subjects are needed. And it makes sense to do it during the experiment itself rather than to waste resources on a reliability study. But if you already have data from a reliability study, point out that uncertainty in the reliability makes a big difference to the estimate of the worst-case final sample size, so you might still be wrong with your estimate.

DO NOT FLY WITH STATISTICAL SIGNIFICANCE

It's important to understand that you sample until you get a narrow confidence interval. You do NOT sample until you get statistical significance. Let's see why.

If statistical significance is your goal, you would presumably start with a sample big enough to give statistical significance for large effects. For example, you might start searching for a correlation of 0.6, which you would want to find statistically significant ($p<0.05$) 80% of the time. From the formulae, the number of subjects is 13, so let's say you start with this number. If you get statistical significance, you stop. If not, you test more subjects.

Seems OK, but there are two things wrong. If the correlation does turn out to be statistically significant on the first go, it has such a wide confidence interval that the correlation in the population is likely to be anything from practically perfect down to trivial. In other words, there's an effect, yes, but you end up with little idea of how big it is.

The other problem is more serious: bias! With a true correlation of 0.6, a starting sample size of 13, and up to three rounds of extra sampling, the sample correlation ends up at 0.65 on average. For a true correlation of 0.40, the sample correlation averages 0.50. This amount of bias is unacceptable. Starting with a bigger sample helps, but as long as you make stopping contingent upon statistical significance, you will have substantial bias for most values of correlation. For example, a true correlation of 0.20 and a starting sample of 45 produce a correlation of 0.25 on average in the final sample. You could start with hundreds of subjects, I suppose, but by then you'd have defeated the purpose of sample sizing on the fly!

I wonder if sampling on the fly using statistical significance is a widespread practice, without people realizing it. By *people* I mean everyone, including the experimenters themselves. It's all too easy to start a study with a small sample, stop if you get statistical significance, or do a few more subjects to bring a promising p value below the 0.05 threshold!

A FINAL WARNING. Opting for sample size on the fly, then sky diving as soon as you get statistical significance, is forbidden. If your paper comes to me for review, I will reject it on the grounds that the result is biased and that the confidence interval is too wide.

Spreadsheet

My apologies, folks. I have yet to do the spreadsheet, because it requires values of a statistic (non-central F) that is not yet available in Excel. I have a link to a plug-in that does the trick. One of these days I will do it. Meanwhile, use the graphs, or the generic method I outlined above.

SIMULATION FOR SAMPLE SIZE

Simulation is where you make up data and analyze them. It's valuable at the planning stage of a complex study, if you're not sure how many subjects you need. It's also a great way to work out how to use a stats program.

Steps in the Process

Simulating to get sample size consists of the following steps:

1. Generate values for the variable(s) for a sample of subjects, usually by drawing them at random from a normal distribution.

2. Introduce an effect, such as a difference between means or a correlation. Make the effect big to start with, but eventually you use the smallest worthwhile effect.
3. Adjust data to take into account likely validity and/or reliability, if necessary.
4. Calculate the effect statistic and either its confidence interval or its p value.
5. If a narrow confidence interval is your goal, go back to Steps 1 or 2 and repeat with different effect magnitudes or sample sizes until you get the sample size that will give the acceptable confidence interval for the smallest worthwhile effect. Stop.
6. If statistical significance is your goal, repeat Steps 1-4 hundreds of times, with new subjects each time, then work out how many times (as a percent) the smallest effect is statistically significant. Voilà, that is the power of the design, for the given sample size.
7. Repeat Step 6 with a bigger or smaller sample until you find the sample size that gives acceptable power (usually 80%) for the smallest effect. Stop.

Even if you can't make your program do the above steps automatically, it's worth making up some values and entering them by hand into a data set, then analyzing them. Make the effect you're interested in big to start with, so you can see which part of the output of the program corresponds to the thing you're looking for. Then try it with a small effect, and see if you still get significance. Remember that in the traditional approach, the smallest worthwhile effect is supposed to turn out significant 80% of the time.

Regard the rest of this page as an appendix. I describe how I generate subjects and variables in SAS. The SAS language is a kind of BASIC, so you should be able to follow it and adapt it to other programs. I show a simulation for a cross-sectional study (where validity can be an issue), and for a longitudinal study (where reliability is crucial), and only for a numeric dependent variable.

Cross-Sectional Study

Let's make two groups of 100 subjects differing by an effect size of 0.2 for a variable with validity 0.9.

In SAS the function rannor(0) generates one randomly chosen value for a normally distributed variable with population mean of 0 and SD of 1. (The "0" has nothing to do with a mean of 0, by the way. It is just a starting "seed" number.) Your stats program should have something like rannor(0). Here I have assigned it to a variable called true (standing for a subject's true value)

```
true=rannor(0)
```

I usually stick with means of 0 and SDs of 1, but if you wanted to make it, say, 70 ± 6, you'd write true=70+6*rannor(0)

These few lines of code generate 100 subjects:

```
do subject=1 to 100;
  true=rannor(0);
  output;
  end;
```

Now let's generate a variable called depvar (standing for dependent variable) with a validity of 0.9 (its correlation with true). I like to do it in such a way that depvar still has a mean of 0 and an SD of 1. I use rannor(0) again to generate a normally distributed source of error, then add a bit of it in with most of true. In the following, sqrt stands for square root:

```
do subject=1 to 100;
  true=rannor(0);
  depvar=0.9*true+sqrt(1-0.9**2)*rannor(0);
  output;
  end;
```

The fact that the population correlation of depvar with true is 0.9 follows from the definition of the correlation coefficient (for the geeks, the correlation coefficient = the covariance of the two variables, divided by their SDs). Here the covariance is 0.9, and the SDs are 1.

It's now dead easy to make another set of 100 subjects with a true effect size of 0.2 relative to the first 100. Study this closely, because it shows how a true effect of 0.2 is degraded to an observed effect of 0.9*0.2 when the validity is 0.9:

```
do subject=101 to 200;
  true=rannor(0)+0.2;
  var1=0.9*true+sqrt(1-0.9**2)*rannor(0);
  output;
  end;
```

Now do a t test and see if you get statistical significance. Write a program to do it 1000 times and see what percentage of the tests gives statistical significance, and hey presto, that's your power. It would be lousy with only 100 subjects in each group!

I'll leave it to you to work out how to generate a simulation for a correlation between two variables, each with its own less-than-perfect validity.

Longitudinal Study

The trick here is to generate two or more correlated repeated measures. We'll do two and call them repvar1 and repvar2. The correlation between the measures is the reliability correlation, of course. Once again you generate true values for your subjects, then add error, this time in a slightly different way . Let's generate repvar1 and repvar2 with a reliability correlation of 0.95 for 20 subjects:

```
do subject=1 to 20;
  true=rannor(0);
  repvar1=sqrt(0.95)*true+sqrt(1-0.95)*rannor(0);
  repvar2=sqrt(0.95)*true+sqrt(1-0.95)*rannor(0);
  output;
  end;
```

There are two ways to add in an effect, let's say 0.2 for repvar2. The normal way is to add it to the true value, just as we did for the cross-sectional design:

```
repvar2=sqrt(0.95)*(true+0.2)+sqrt(1-0.95)*rannor(0);
```

But sometimes repvar is the **criterion** outcome measure, so it may not be appropriate to consider that the effect is degraded by the less-than-perfect reliability. For example, if repvar represents competitive performance, we may be interested in detecting an effect of 0.2 for repvar, not for true. I'm still thinking about this one. In such cases, this is how you add in the effect:

```
repvar2=sqrt(0.95)*true+sqrt(1-0.95)*rannor(0)+0.2;
```

It's possible to add finite validity along with reliability for variables in a longitudinal simulation. If the reliability is r, the validity is v, and the effect size is es, then the following generates two variables (repvar1 and repvar2) that have a correlation of r with each other and that have a correlation of v with the true value:

```
do subject=1 to 20;
  true=rannor(0);
  errorv=rannor(0);
  repvar1=v*true+sqrt(r-v**2)*errorv+sqrt(1-r)*rannor(0);
  repvar1=v*(true+es)+sqrt(r-v**2)*errorv+sqrt(1-r)*rannor(0);
  output;
  end;
```

I have used this simulation to check that the formulae for longitudinal designs are correct.

SUMMARY: The Most Important Points

- Think about differences between group means in terms of standard deviations, not standard errors of the mean. Mean ± SD or Mean ± SEM?
- Learn exactly what a trivial, small, moderate, large, very large, and almost perfect effect is, for a correlation, a frequency difference, and the effect-size statistic. Magnitudes for Effect Statistics
- Present as few numbers as possible: no more than two significant digits for effect statistics and standard deviations. How Many Digits?
- Understand how validity impacts on cross-sectional studies and reliability impacts on longitudinal studies. Cross·Sectional Designs, Longitudinal Designs
- Test enough subjects to allow you to publish any result. Sample·Size Estimation, Based on Confidence Limits
- Try sample size "on the fly" in your next project, but base it on width of the confidence interval, not statistical significance. On The Fly
- Explore a stats program by making up data with an effect, then analyzing them. Simulation
- Before you do any modeling (statistical tests), look at your data to see what's going on. Effect Statistics
- Keep your eye on standard deviations or scatter of points, to decide whether log or rank transformation is needed before you fit a model. Residuals: Bad, Log Transformation, Rank Transformation
- If you have repeated measures with missing data, get a statistician to help you model covariances. Modeling Covariances
- Know the difference between statistically significant and substantial. Confidence Limits, Magnitudes for Effect Statistics
- Show confidence intervals instead or, or as well as, p values. P Values
- Stop asking "is there an effect?" Start asking "how big is the effect?" Hypothesis Testing
- Stop thinking about testing. Start thinking about estimating. Hypothesis Testing

QUIZ

Each question has either only one correct answer or one incorrect answer. The answers appear in the lower frame when you click on answer. Links to the appropriate sections of the text are also included.

1. A frequency distribution can be shown as

- a statistic
- a histogram
- a scatter plot
- a stem and leaf plot
answer · Basics

2. Simple statistics are

- for simpletons
- presented in stem and leaf plots
- things like correlations
- things like standard deviations
answer · Simple Statistics

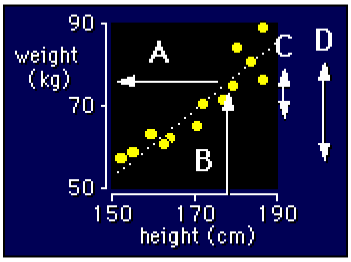
3. What would you do with a median?

- Use it do show spread.
- Use it for normally distributed data.
- Cross it against oncoming traffic
- Indicate the middle of some data.
answer · Simple Statistics

4. The following are measures of spread:

- standard deviation
- root mean square errors
- percentile ranges
- polyunsaturated margarine
answer · the Spread

5. Which arrow indicates the standard error of the estimate?



- A
- B

- C
- D
  answer · SEE

6. A relative risk is

- a risk of matrimony
- an outcome statistic
- a relative of the odds ratio
- a relative frequency
  answer · Relative Frequency

7. Differences between means are best thought about in terms of

- p values
- standard errors of the mean
- percentages of the mean
- standard deviations
  answer · Effect Size · Mean ± SD or Mean ± SEM?

8. Dimension reduction

- describes loss of precision.
- describes factor analysis.
- is an example of ANOVA.
- is a weight-loss program.
  answer · Dimension Reduction

9. Concerning reliability:

- It impacts most on descriptive studies.
- It can be expressed as an ICC.
- It can be expressed as a CV.
- It is quantified by 2-way ANOVA.
  answer · Reliability

10. Concerning validity:

- It impacts most on descriptive studies.
- It is the correlation between true and observed values.
- A valid measure must be reliable.
- A reliable measure must be valid.
  answer · Validity

11. Correct or incorrect expressions?

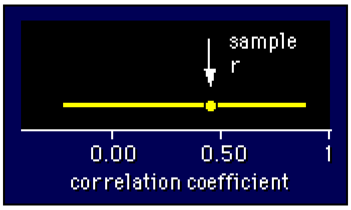
- height = 175 ± 6 cm
- VO2peak = 67 ± 5.1 ml/min/kg

- ICC = 0.87
- CV = 1.4%
  answer · How Many Digits

12. Confidence intervals...

- are a new form of sprint training.
- are calculated routinely by most stats packages.
- define the likely range of a population value.
- are inferior to p values for indicating magnitude of outcomes.
  answer · Confidence Intervals · What is a P Value?

13. A correlation coefficient and its confidence interval are shown in the figure.



We can conclude that:

- The correlation is significant.
- The true value of the correlation is likely to be 0.45.
- More subjects should be tested.
- A type II error has occurred.
  answer · More on the Lower and Upper Limits · Type II Errors

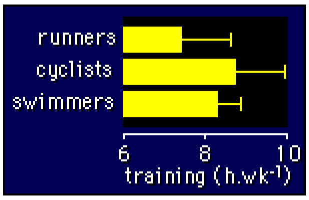
14. Concerning tests and test statistics:

- One-tailed tests are sometimes justified.
- Test statistics should always be shown.
- Chi-squared is a common test statistic.
- P = 0.06 means there is no effect.
  answer · What is a P Value? · Using P Values

15. Many samples, each of 100 observations, are drawn from a population in which there is a correlation of 0.70 between two variables. How often would you expect to find a statistically significant correlation?

- hardly ever
- about one time in 100
- about one time in 20
- almost always.
  answer · Type I Errors

16. What are appropriate comments about these data, which show mean weekly training durations for three groups of athletes? (Bars are SDs.)



- Differences between all groups are substantial. (See Effect Size.)
- The data should be analyzed by repeated-measures ANOVA.
- Log transformation appears to be necessary before analysis.
- Runners are lazier than cyclists.
answer · One-Way ANOVA

17. An outcome measured on a five-point scale (*not at all* to *always*)...

- is an example of an ordinal variable.
- has a behavior problem when it comes to residuals.
- should be analyzed by logistic regression.
- can be analyzed by ANOVA.
answer ·Ordinal Dependent Variables

18. Log transform a variable...

- if the values are too big.
- if the residuals (error) get bigger for bigger values of the variable.
- if you don't get statistical significance.
- if non-parametric tests are inappropriate.
answer · Log Transformation · Non-Parametric Models

19. Non-parametric tests usually...

- are parametric tests in disguise..
- involve rank transformation of the dependent variable.
- work for grossly non-normal data..
- should be attempted if parametric tests give p > 0.05.
answer · Non-Parametric Models

20. If we studied the effect of gender and body mass on sprint performance time, we would use the following model:

- unpaired t test
- ANCOVA
- ANOVA
- MANOVA
answer

21. Concerning multiple linear regression:

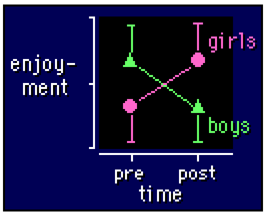
- Use it to fit curves as well as straight lines.
- Use it to control for the effect of numeric variables.

- It gives misleading results for highly correlated independent variables.
- Use it to fit multiple straight lines with several groups.
  answer · Multiple Linear Regression

22. Repeated-measures models...

- are used in descriptive studies.
- can be analyzed by modeling variances.
- are used when you have to repeat a failed test.
- are straightforward to analyze with stats programs.
  answer · Repeated-Measure
  ANOVA · RM·ANOVA: Three or more trials and no between·subjects effect

23. In a longitudinal study aimed at enhancing sport enjoyment, the following results were obtained



We can conclude that:

- Initial randomization to the two groups was poor.
- There is one between- and one within-subject factor.
- The time effect in the model is substantial.
- The time effect in the model is significant.
  answer · Two trials plus...

24. Concerning sample sizes for a controlled longitudinal study:

- Sample size is proportional to (1 - r), where r = reliability correlation.
- Controlled studies need 4x as many subjects as uncontrolled studies.
- Get sample size "on the fly" by testing until you get an acceptable confidence interval.
- None of the above.
  answer · What Determines Sample Size · Sample Size "On the Fly"

25. The size of a sample needed for a cross-sectional study...

- depends on the size of your research grant.
- is inversely proportional to the square of the validities of your measures.
- is a function of the largest effect you want to detect.
- depends on how many student researchers you have on the project.
  answer · The Right Number of Subjects · What Determines Sample Size

26. When you come home from climbing in the statistical mountains, you will tell the folks, amongst other things, that...

- from now on you will show as few numbers as possible.
- statistical modeling is no substitute for knowing your data.

- it's important to play with stats programs.
- from now on you will test rather that estimate.
  answer · Summary