# Power and Sample Size Estimation Techniques for Fisheries Management: Assessment and a New Computational Tool

KENNETH G. GEROW

*Department of Statistics, University of Wyoming, Laramie,*

*Wyoming 82071-3332, USA*

## Abstract

Formulae for sample size calculations in the literature are often predicated on unrealistic assumptions (e.g. equal variances) or unrealistic, or simply undesirable, designs (two independent samples of the same size). In addition, sample size and power calculations usually involve repeated use of any given formula, as the researcher works through varying inputs (alpha, power choices, design choices, etcetera), making the process of considering sample size and power issues unpleasantly tedious and error-prone. I survey formulae currently in the fisheries literature, describing deficiencies. I then discuss sample size formulae that correct these deficiencies, and describe a freely available Excel tool that enables the calculations.

## Introduction

Fisheries biologists often need to determine if a management action or other event has resulted in a change in fish populations. Estimation of sample sizes (e.g. number of net-sets or reaches of a stream) needed at two points in time for purposes of detecting changes in indices of fish density such as catch per unit effort is of substantial interest among fisheries scientists. Often the point of such sample size calculations is to ensure adequate power for detecting some specified change in a mean. In that context, statistical power is defined as the probability of correctly rejecting the null hypothesis of no change when some specified alternate is correct. The estimation of power given a sampling context or sample size given a target value for power requires inputs such as (1) choice of alpha (chance of falsely rejecting the null); (2) declaration of whether the test will be one-or two-tailed; (3) a stated value for the alternate (usually some “biologically significant” value); (4) choice of design (e.g. independent samples, paired samples), and (5) some assumption about the behavior of variation in the data (e.g. assume equal variances, or variances proportional to means). Figure 1 illustrates the concepts for a realistic situation. In this paper, some physical entity (e.g., one net-set, or one reach of a stream) constitutes a sample, within which one would measure, say, the number of fish.

Sample size calculations are often computed following explicit formulae using the standard normal (*z*) distribution (e.g. Gryska et al. 1997, Allen et al. 1999), or the *t*-distribution (e.g., Krueger et al. 1998, Bryant et al. 2004). I shall refer to these papers collectively as G, A, K & B. Owen (1965) showed that if an alternate hypothesis is true, the correct distribution of the test statistic for the difference between two means (or the mean difference for paired data) is a skewed variation of the *t*-distribution called the noncentral *t*-distribution. The approximations (*z* or *t*) are sometimes adequate, as I shall illustrate below, but there are important other problems with the use of the formulae in G, A, K & B. In particular, they explicitly assume two independent samples of the same size, and that variances are equal (or follow a specific model (Krueger et al. 1998) in the two populations that generated the samples. Note that this structure precludes consideration of a paired design, such as sampling the same sites at the second time.

In this paper I illustrate that a direct approximation of the noncentral *t*-distribution, (Abramowitz and Stegun, 1970) is not difficult to compute, and show it to be superior to the *z* or *t* methods. I then consider sample design issues and variation/mean relationships, and discuss relevant problems associated with formulae proffered by G, A, K, & B. This will lead to a set of power and sample size formulations that reflect realistic fisheries sampling scenarios. I illustrate with variations on a realistic example for a scenario of samples taken at two points in time crafted after data collected in Wyoming (M.C. Quist, Iowa State University, unpublished notes). Finally, I describe an Excel tool (freely available at statsalive.com) to enable the calculations.

## Methods

Many well-known formulae, using either the *z*-distribution (Snedecor and Cochran 1989) or the *t*-distribution (Sokal and Rohlf 1995) are available for calculating sample size at different levels of statistical power when testing differences in means with two independent samples. The simplest formula to use is based on the *z*-distribution:

$$n = 2\left(z_\alpha + z_\beta\right)^2 s^2 / d^2 , \tag{1}$$

where *d* is the purported difference in two means from populations, whose (assumed common) estimated SD is *s*, with random samples of size *n*; $z_\alpha$ and $z_\beta$ are values from the standard normal distribution that account for the chance $(\alpha)$ of false significance and chance $(\beta)$ of falsely failing to reject when in fact the difference *d* is true (power = $1-\beta$). For sake of illustrating the thinking behind sample size formulae, a derivation of this formula is in Appendix A. The *t*-based formula (Zar, 1999; p. 132) is essentially the same as the *z*-based formula, except that *z*-distribution elements are replaced by *t*-distribution elements:

$$n = 2\left(t_{\alpha,df} + t_{\beta,df}\right)^2 s^2 / d^2 . \tag{2}$$

These formulae use *z* and *t* as approximations to the correct distribution. If the null is not true and some specified difference *d* in the means at two times exists, the test statistic $d/SE(d)$ has a noncentral *t*-distribution characterized by degrees of freedom and a so-called noncentrality parameter $\delta$ (delta), which is none other than a standardized version of the postulated difference: $\delta = d/SE(d)$ (Owen 1965). I compared three methods (*z*, *t*, and an approximation to the noncentral *t* (see Appendix B for details) to the exact calculations using Minitab 14 (2005) for a small, medium, and large effect, over a range of sample sizes.

The context for formulae (1) and (2) is a two-independent-samples design with equal sample sizes and assumed equal variances from each population. In reality, variation often changes with means in biological data, independent samples may not be the best design, and sample sizes certainly don't have to be equal. I critique sample size formulae in the fisheries literature (G, A, K, & B ), and argue that, for biologically realistic scenarios and scientifically better designs, that sample size formulae need to be able to account for a variety of variation/mean relationships and accommodate paired sampling.

## Results and Discussion

*Using z- and t- approximations*. –The *z*-approximation tends to over-estimate true power, quite dramatically for a large effect (e.g. difference in means) and small sample size combination (Figure 2); less so otherwise. The corollary of the *z*-tool overestimating power is that it underestimates the required sample size, but has the benefit of being a closed-form formula.

The *t*-distribution, on the other hand, tends to under-estimate power with a smaller bias than the upward bias of the *z*-approximation (Figure 2), but introduces a small paradox. The formula for a *t*-distribution requires the degrees of freedom (*df*), which depends on sample size [in the case of equal sample sizes and assumed equal SD, $df = 2(n-1)$ ]. Thus knowledge of *n* is required, which is what we are trying to use the formula to compute.

This is commonly resolved by using the *z*-distribution formula to get an initial sample size. Estimated power is then obtained for that sample size: $t_{\beta^*,df} = \frac{d}{s}\sqrt{\frac{n}{2}} - t_{\alpha,df}$ . A tabular query of this *t*-value will reveal that $\beta^*$ is larger than desired (i.e., the power will be smaller); the protocol is then to increase the sample size by one, and check again, and to repeat until the desired power is obtained.  This process usually requires only a few iterations to reach the estimated sample size.

The direct approximation to the noncentral *t*-distribution (Abramowitz and Stegun, 1970) is slightly more cumbersome to use than the *t*-distribution approximation, but matches the exact answers best (Figure 2). Its direct use is analogous to using the *t*-distribution: generate an initial *z*-based estimate, check the power obtained, and increase the sample size iteratively until desired power is achieved. Electronic calculators can bypass this need to iterate because of their blinding speed: as quickly as you can choose a sample size, power is instantly calculated, and it is effortless (and fast!) to choose different sample sizes until a satisfactory power is obtained.

*Practical problems* -- The formulae in G, A, K, & B, while relatively convenient to use, suffer in practice from several important features, not related to the *z* or *t* approximations. Variation in biological data is often in some way proportional to means, so the assumption of equal standard deviations is incorrect, and can result in quite misleading results.  Further, the formulae are designed for a study with two independent samples, each of the same size.  In monitoring programs, it is often much better to used paired sampling, revisiting the same sites a second time. The failure to accommodate paired samples may be viewed as the greatest weakness of these formulae. Second, (in cases where the sampling occurs from one season to the next) sample size calculations often ensue after a season of field work (see Gryska et al. 1997, Krueger et al. 1998), by which time sample size one is fixed. A better question in that event is, “Given my first sample size, how many samples do I need the next time?”

Gryska et al. (1997) based their calculations on log-transformations, an approach that often stabilizes variances. It also, as they noted, induces distributions that are often more normally distributed. As an author on that study, I would no longer suggest that method routinely. Most importantly, it does not address the independent versus paired samples issue. Second, the normality of the data is not critical (except in the case of very small sample sizes) because the difference in means will have an approximately normal distribution almost without regard to the original distribution, given sufficient sampling. Third, equal variances are not a requirement for *t*-tests.

One could use a calculation that is based on the Poisson distribution, as was done in Krueger et al. (1998). In a Poisson distribution, commonly used as a model for count data, the mean is equal to the variance. Patchiness of habitat and consequent changes in abundance of fishes, for instance, would lead to increases in variance beyond what a Poisson would predict. In that case, sample size calculations using the Poisson would be biased because estimates would be based on an assumed variance that is too low. Their approach also employs the assumption of two independent samples (with equal sample sizes), which is not often going to be the optimal design.

I note that the formula in Krueger et al. (1998) is incorrect. Their Table 6 shows that required sample sizes are larger for detection of an increase and smaller for detection of a decrease, to be expected when using the Poisson distribution as a basis for the calculations. Unfortunately, examination shows their formula to be symmetric, yielding the same answer whether one seeks to test for increases or decreases in abundance.

*Effective and efficient sample size calculation*. -- A useful tool for sample size calculations would allow a researcher to easily compare the sampling requirements of various designs (two independent samples (not restricted to equal sample sizes), paired samples, a mix of paired and independent data), and incorporate different variation patterns (equal SD, arbitrary unequal SD, variation proportional to mean). I will now discuss some of these issues in more detail.

*Choosing effect sizes for power and sample size calculations*: -- Bryant et al. (2004) followed Cohen (1988) in choosing a small effect size to be 20% of the estimated population standard deviation (SD), medium to be 50% of the SD, and a large effect to be 80%. Effect sizes should always be chosen to be ones of particular interest to the situation at hand, not based on interests outside the study. When advising scientists on power and sample size matters, I suggest they think of small, medium, and large, as follows. For simplicity, I take the difference between two means as the effect being considered. A small effect is the smallest difference that elicits your interest. A large effect is the smallest difference that you would definitely not want to fail to detect. Medium is, say, the average of the small and large. For any given sample size, the power will be largest for the large effect, and smallest for the small effect. Ideally, you will arrive at a sample size that is logistically and financially feasible, yet yields sufficient power where it counts i.e., larger effects.

*Choosing a targeted power* -- Obtaining statistical power is a situation where the law of diminishing returns applies. In a situation where, say, a sample of 25 yields 70% power, you'd need 30 to achieve 80%, 40 to achieve 90%, and almost 50 to obtain 95% power. Thus, increasing power has real costs. I suggest that the choice of power to aim for depends on the consequences. While, for instance, 80% power may be reasonable in some circumstances, it may be too low or too costly for others.

*Estimating standard deviations*: -- Experienced biologists often find it relatively easy to provide an *a priori* estimate of the range of values they will likely observe; an *a priori* estimate of the SD can then be easily constructed. For many distributions, the SD in a sample of modest sample size (I have 20 in mind) is approximately 1/3 to 1/4 of the range (maximum minus minimum) of values in the sample. Using 1/3 will provide a slightly more conservative estimate of SD than will 1/4.

*Impact of variation:mean relationships* -- In situations where variation is larger in populations with larger means, it is easier to detect a decrease of a given size than an increase of the same size. The reason is that the standard error of a difference in means will be smaller in the former case (see Figure 1). Suppose that we understand standard deviations to be approximately equal to means and that an initial data set of 25 observations with mean $\bar{Y}_1 = 10,$ and sample SD the same. Let us consider testing for a 50% increase over time (mean of second sample is expected to be 15) with a second independent sample of the same size. We expect the SD in that sample to be $s_2 = 15$. Then $SE\left(\bar{Y}_1 - \bar{Y}_2\right) = \sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}} = \sqrt{\frac{10^2}{25} + \frac{15^2}{25}} = 3.61$. On the other hand, for a 50% decrease, we might expect $s_2 = 5.0,$ whence $SE\left(\bar{Y}_1 - \bar{Y}_2\right) = \sqrt{\frac{10^2}{25} + \frac{5^2}{25}} = 2.24.$ The smaller *SE* in the second case will make detection of a decrease more likely (power is 59%) than an increase (power is 28%). In particular, let us use $\alpha = 0.05$ in a two-tailed test. In order to have 80% power to detect a 50% increase the required sample size at each time is approximately 102; for a 50% decrease, it is only 40.

Two simple relationships that are relatively easy to consider are that the standard deviation or the variance is proportional to the mean. The latter is the correct relationship for Poisson-distributed data (specifically, they are equal); the former expresses variation greater than that expected in a Poisson (which is likely realistic for many situations). Lacking any evidence or theoretical justification for a particular variation/mean relationship, I recommend doing calculations using both. Often, they will yield sample size/power calculations that are not much different. If the results are quite different, biologists might consider using the more conservative result until calculations can be refined with more data.

If you have data representing a range of means, it is possible to establish whether standard deviations or variances are more closely in a proportional relationship to means. For example, Figure 3 illustrates SD:mean and variance:mean patterns for counts of six species of fish (each counted in 49 reaches) (unpublished data from M. Quist). In that example, the SD:mean ratios showed less relative variability (as measured by the coefficient of variation), which suggest consideration of SD:mean proportionality. I use that ast the basis for illustrations in this paper. It is more often the case that such data do not exist, especially if one is conducting *a priori* power calculations.

*Design* -- The simplest realistic modification to the two independent, equal sized, samples design is the case where the one sample size $\left(n_1\right)$ is fixed, and you want to know how many samples to take at the second time to achieve your power goals. Continuing with the foregoing examples, suppose, $n_1 = 25$. Power for a 50% decrease will not budge above 75%

(attained for $n_2 = 200$) no matter how many samples we take at the second time, and appears to top out around 70% for a 50% increase. Initially, this seems paradoxical: how can it be that power fails to increase as we increase our sampling effort? The cause, it turns out, is the fixed initial sample size. In the case of a 50% decrease, $SE\left(\overline{Y}_1 - \overline{Y}_2\right) = \sqrt{\frac{10^2}{25} + \frac{5^2}{n_2}} \geq 2$; it will approach 2 (but go no lower) as the second sample size $\left(n_2\right)$, increases. This implies, as a matter of practicality, that one ought to do as much as possible in the first year of a monitoring effort to minimize the chance of this occurring.

A paired design is often the most powerful design for detecting changes across time, provided there is sufficient correlation within sites across time. Using the sample size calculations requires an estimate of the correlation between sites. In most instances this requires either an educated (based on experience with similar data) or a conservative guess. A conservative estimate of the correlation could be computed by doing a regression analysis with the second sampling time as the response, and the first as the predictor, (it is not important which roles you assign). As an aid in estimating correlation in the absence of data, the excel tool (described later) includes an interactive, dynamic application with which you can study correlation between two hypothetical samples.

For example, if this regression computes an $r^2$ of 16% and that is considered to be on the low side for such a relationship, this means that that the correlation (*r* is 0.4) is a conservative estimate. In the event of paired sampling, the estimated standard error of the mean of the differences is $SE\left(\overline{diff}\right) = \sqrt{\left(s_1^2 + s_2^2 - 2\rho s_1 s_2\right)/n}$, where *n* is the sample size, $s_1$ and $s_2$ are the standard deviations of the two samples, and $\rho$ is the within-site correlation from one sampling period to the next. As the correlation goes up, the standard error goes down. Continuing with the foregoing example, the standard error for the difference in means given there has been a 50% increase is 3.6 with two independent samples of size 25. If the samples were paired (with correlation was 0.4), the SE is $\sqrt{\left(s_1^2+s_2^2-2\rho s_1 s_2\right)/n}=\sqrt{\left(10^2+15^2-2\times 0.4\times 10\times 15\right)/25}=2.86$, a 20% reduction. The consequence of pairing is that you would need 66 samples to have 80% power to detect a 50% increase, and 28 for a 50% decrease; 66 may still frighten, but it is a lot smaller than 102!

For the mixed design, we have a set of *n* paired samples at times one and two (with values labeled as $Y_{1j}$ and $Y_{2j}$; $j=1,2,\ldots,n$) and a third, independent sample at time two, of size $n_3$. We have available two estimators of the difference in means across time: the mean of differences for the paired samples: $\frac{1}{n}\sum\left(Y_{1j}-Y_{2j}\right)$, and the difference in means: $\left(\bar{Y}_1-\bar{Y}_3\right)$. I combined them in a weighted average (with weights chosen to reflect their precision):

$\widehat{diff}=\frac{w_1}{n}\sum\left(Y_{1j}-Y_{2j}\right)+w_2\left(\bar{Y}_1-\bar{Y}_3\right)$ with

$$SE\left(\widehat{diff}\right)=\sqrt{\frac{s_1^2}{n}+\frac{\left(w_1 s_2\right)^2}{n}-\frac{2w_1\rho s_1 s_2}{n}+\frac{\left(w_2 s_2\right)^2}{n_2}}. \qquad (3)$$

See Appendix B for a derivation and notation. In our ongoing example, we've already established that a paired sample of size 28 is enough to have 80% power for a 50% decrease, assuming a correlation of 0.4. For a 50% increase, with 30 pairs, we will need more than 300 additional points to obtain 80% power. The point made above that fixing the initial sample size can cause power to "top out" at somewhere less than 100% holds for this mixed design also.

*A tool for calculations*. -- Consider the several choices and estimates that are required for such a calculation: (1) choice of $\alpha$; (2) estimates of standard deviation (which may be complicated by variation/mean structure); (3) choice of one- or two-tailed test; (4) choice of size of effect; (5) choice of whether to do the calculation for an increase or a decrease in the mean, which will matter if there is some variation:mean relationship in the populations; and (6) choice of design: paired (for which one needs to estimate the correlation), two independent samples, or a combination of these designs.

There are many (several excellent) easy-to-use power/sample size calculators available on the Web. None that I found accommodate easily incorporation of variation/mean relationships, paired designs (they do implicitly: one can use a single-sample *t*-calculator, since a paired design reduces to a single sample analysis, but estimating the relevant inputs is not easy), or the possibility of a mixed design (paired plus some extra that are independent).

The need to re-do sample size calculations potentially many times in a given study will inhibit full consideration of all the possibilities.  In response to that, I developed an Excel tool (Figure 4) that allows exploration of sample size calculations flexibly and dynamically.  The tool has a main sheet in which to enter inputs (Figure 4), a sheet with graphs to illustrate the power situation, given those inputs (Figure 1), and two additional sheets that interactively allow the user to explore variation:mean relationships and correlation between samples. I used that tool to perform calculations for the examples illustrating this paper.  The tool is freely available at www.statsalive.com.

*Robustness of calculations* – Necessarily, sample size calculations will often be built around guesses for standard deviations, correlations, and/or variation:mean relationships. It is useful, then, to study the robustness of your calculations to those guesses. With electronic calculators, it is usually easy to re-do the calculations with different inputs (for instance, try a slightly larger

SD or smaller correlation). In the examples in this paper, I assumed correlation was equal to 0.4. To study the impact of mis-specification, I re-did the calculations using correlation of 0.3 (corresponding to an $R^2$ of 9%: very low). Recall: I found that a sample of size 28 (paired samples) was adequate to obtain 80% power for a 50% decrease (using correlation 0.4). With 0.3, I need 30. I rest easy that my guess of 0.4 (if incorrect and too large) is likely not leading to grossly wrong sample size calculations.

## Management Implications

Sample size calculations (given power goals) or power calculations (given sample size) can be a critical tool in effective resource allocation by fisheries managers. Sample size formulae commonly cited in the literature for comparisons at two points in time are predicated on equal sample sizes for two independent samples; this design is often not realistic (the first sample may already have been taken by the time of doing the calculations) or not the best design (paired samples are more powerful). Some of the formulae assume equal variances (others a restricted relationship to means), which is unrealistic in many biological situations, wherein variation changes with means. In addition, the tedium of cranking through many calculations (varying parameter estimates, and choices such as $\alpha$, and power goals) inhibits full exploration of the possibilities and trade-offs in a given situation. The Excel tool I created eliminates the tedium, making it quick and painless to sort through many scenarios as part of a power and sample size analysis.

# Acknowledgements

This work was inspired by a fisheries biologist (M. Quist) who had to conduct power calculations in dozens of situations, and who swore it would drive him mad. I thank Jeff Beck, Wayne Hubert, Mike Quist, and Frank Rahel for their comments on an earlier draft. Conversations with M. Quist and Roy Irwin sharpened much of the thinking that wen into this project. Anonymous reviewers (one in particular whose detailed comments and perspective were most helpful) improved this work.

## References

Abramowitz, M., and I.E. Stegun, Eds. 1970. Handbook of Mathematical Functions. Dover Publications, Inc., New York.

Allen, M.S., Hale, M.M., and W.E. Pine III. 1999. Comparison of Trap Nets and Otter Trawls for Sampling Black Crappie in Two Florida Lakes. North American Journal of Fisheries Management 19: 977-983

Bryant, M.D., Caouette, J.P., and B.E. Wright. 2004. Evaluating Stream Habitat Survey Data and Statistical Power from Southeast Alaska. North American Journal of Fisheries Management 24:1353-1362.

Cohen, J. 1988. Statistical Power Analyses for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale New Jersey.

Gryska, A.D., Hubert, W.A., and K.G. Gerow. 1997. Use of Power Analysis in Developing Monitoring Protocols for the Endangered Kendal Warms Springs Dace. North American Journal of Fisheries Management 17:1005-1009.

Krueger, K.L., Hubert, W.A., and R.M. Price. 1998. Tandem-Set Fyke Nets for Sampling Benthic Fishes in Lakes. North American Journal of Fisheries Management 18:184-160.

Owen, D. 1965. The Power of Student’t *t*-test. Journal of the American Statistical Association 60: 320-333.

Snedecor G.W. and W.G. Cochran. 1989. Statistical Methods, Eighth Ed.. Iowa State College Press, Ames, Iowa.

Sokal, R.R., and F.J. Rohlf. 1995. Biometry: The Prinicple and Practice of Statistics in Biological Research, Third Ed.. W.H. Freeman and Company, New York.

Zar, Jerrold. 1999. Biostatistical Analysis. 4th ed. Prentice Hall, Upper Saddle River, New Jersey.

## Appendix A: Derivation of *z*-based sample size formula

Assuming a common sample size *n* and common standard deviation *s* for both samples, let $z_\alpha$ denote the value from a standard normal distribution such that $100 \times {}^{\alpha}\!/_{2}\%$ (two-tailed test) or $100 \times \alpha\%$ (one-tailed test) of the distribution lies further from the mean. For example, with $\alpha = 0.05$, $z_\alpha = 1.96$ for a two-tailed test, and $z_\alpha = 1.645$ for a one-tailed test. The units are standard errors (of the difference in the two means). So, for instance, in a two-tailed test (using this criterion), an observed difference would have to be larger than 1.96 SEs for it be declared significant.

Similarly let $z_\beta$ be that value from the normal distribution with $100 \times \beta\%$ lying further from the mean. This value is chosen so that $1-\beta$ is the desired power of the test. For instance, if the desired power is 0.80, $z_\beta = 0.842$. By construction of the testing procedure (that is, with $\alpha$ and $\beta$ chosen), the difference in means (denoted by *d*) must be $z_\alpha + z_\beta$ standard errors from the null difference of zero. Given the assumption of a single SD and equal sample sizes, the SE of the difference in means is $\sqrt{2s^2/n}$, so the difference is equal to $d = \left(z_\alpha + z_\beta\right)\sqrt{2s^2/n}$. Now turn it all around. Choose *d* (and $z_\alpha$ and $z_\beta$), and estimate *s*. A little algebraic manipulation yields the sample size formula: $n = 2\left(z_\alpha + z_\beta\right)^2 s^2/d^2$.

## Appendix B: Combining paired and independent differences

Suppose we have a paired sample at times one and two (call them samples 1 and 2, with $n$ at each time), and a third, independent sample at time two, of size $n_3$. What is the standard error of the estimated difference in means from time one to time two and what df?

One choice of estimator is a weighted sum of the two individual estimators (one is the mean of differences; the other the difference in means:

$$\begin{aligned}\widehat{diff} &= \frac{w_1}{n}\sum\left(Y_{1j}-Y_{2j}\right)+w_2\left(\bar{Y}_1-\bar{Y}_3\right)=\frac{1}{n}\sum\left(\left(w_1+w_2\right)Y_{1j}-w_1Y_{2j}\right)-w_2\bar{Y}_3\\ &=\frac{1}{n}\sum\left(Y_{1j}-w_1Y_{2j}\right)-w_2\bar{Y}_3 \text{ (since } w_1+w_2=1)\end{aligned}$$

where $w_1$ and $w_2$ are weights assigned to two individual estimators $(w_1+w_2=1)$; $w_1=w_2=1/2$ represents the simple average of the two. The standard error is

$$SE\left(\widehat{diff}\right)=\sqrt{\frac{s_1^2}{n}+\frac{\left(w_1s_2\right)^2}{n}-\frac{2w_1\rho s_1s_2}{n}+\frac{\left(w_2s_2\right)^2}{n_2}}\ .$$

We will use weights proportional to the inverse of the variance for each individual estimator: let $V_1=\left(s_1^2+s_2^2-2\rho s_1s_2\right)\big/n,$ and $V_2=s_1^2\big/n_1+s_2^2\big/n_2.$ Then $w_1=\dfrac{V_1^{-1}}{V_1^{-1}+V_2^{-1}}$ and $w_2=1-w_1=\dfrac{V_2^{-1}}{V_1^{-1}+V_2^{-1}}$ (the denominator is just a scaling trick to make sure the two weights sum to one).

**Figures**

Figure 1. Screen shot of graphical output from Excel power and sample size calculator (available at www.statsalive.com). Illustrated are null and alternate distributions for a 50% change from a mean of 10. Sample sizes are 25 in each of two independent samples, and it has been assumed that standard deviations are proportional to means. In this scenario, it is easier to detect a decrease (power is 59%; left panel: area to the left of the yellow line under the orange curve) than an increase (power is 25%; right panel: area to the tight of the yellow line under the orange curve). Alpha has been set to 0.05, and the test is two-tailed.

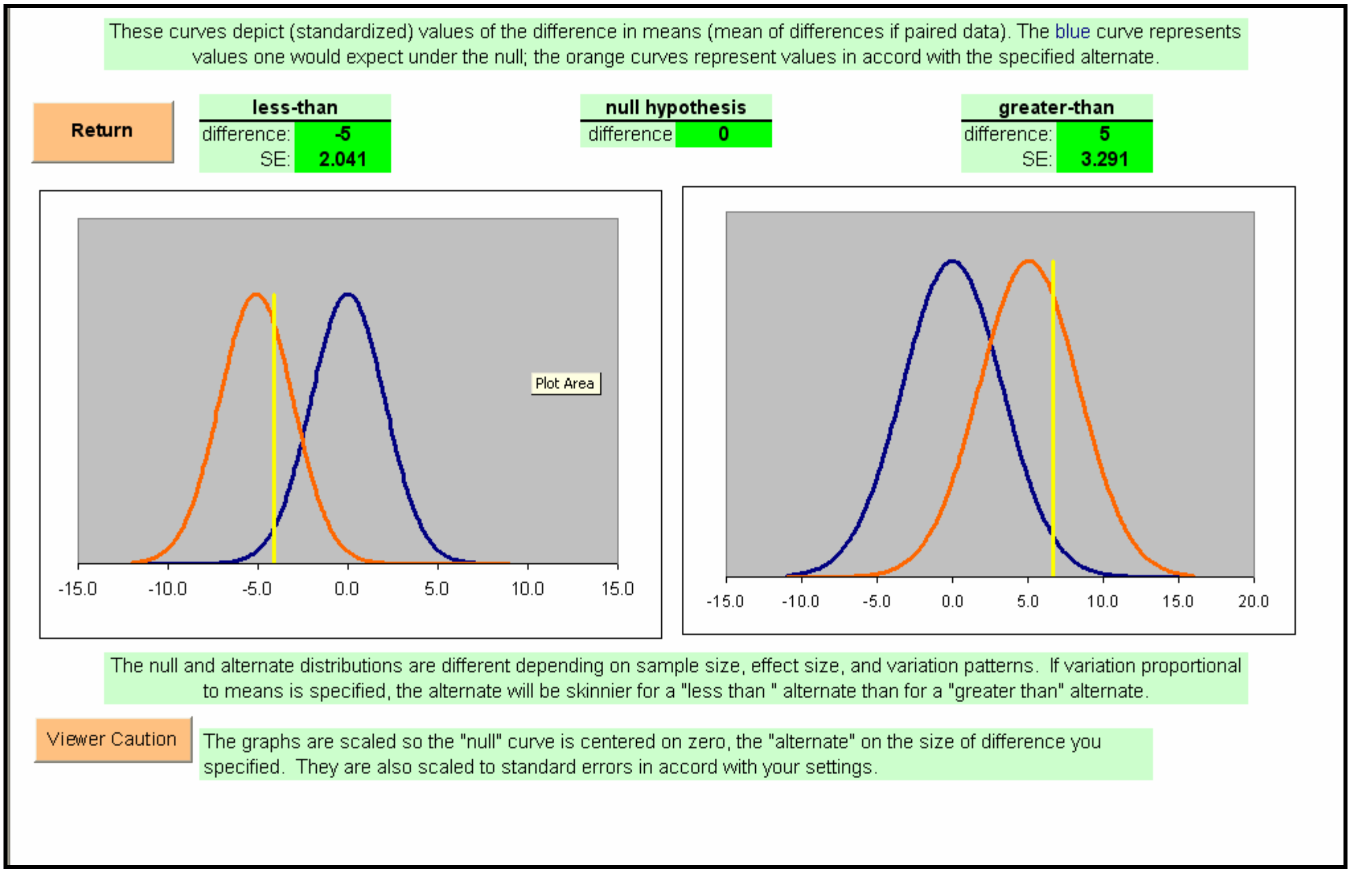
Figure 2. Estimated power functions for a statistically small (Panel A), medium (Panel B), and large (Panel C) effect. For each effect level, power functions are estimated using four approaches: exact (using the non-central *t*-distribution), *z*-based formula, *t*-based formula, and the direct approximation to the noncentral *t* (A&S). The *z*-based method overestimates true power; the *t*- underestimates it, especially for large effects and small sample sizes.

Figure 3. Screen shot of variation:means relationships worksheet on the Excel tool. In this example, the 6 observed SD/mean ratios display less relative variation (CV = 25%) than the 6 variance/mean ratios (CV = 50%), suggesting the proportionality relationship is more stable for SDs. Both relationships are reasonably linear (correlations around 90%).

Figure 4. Screen shot of main worksheet on the Excel tool. Tan (routine) and red (cautions) buttons display messages (dark grey in this black & white depiction). The user enters inputs using yellow buttons and sliders (light grey here); the buttons in the upper middle of the sheet will display the graphs sheet (see Figure 2), power curves, or appendices

containing documentation. In the live version, cells used directly by the user for inputs are colored yellow; cells containing other values are colored green.

Figure 1



These curves depict (standardized) values of the difference in means (mean of differences if paired data). The blue curve represents values one would expect under the null; the orange curves represent values in accord with the specified alternate.

**less-than**
difference: **-5**
SE: **2.041**

**null hypothesis**
difference **0**

**greater-than**
difference: **5**
SE: **3.291**

The null and alternate distributions are different depending on sample size, effect size, and variation patterns. If variation proportional to means is specified, the alternate will be skinnier for a "less than " alternate than for a "greater than" alternate.

The graphs are scaled so the "null" curve is centered on zero, the "alternate" on the size of difference you specified. They are also scaled to standard errors in accord with your settings.

Figure 2

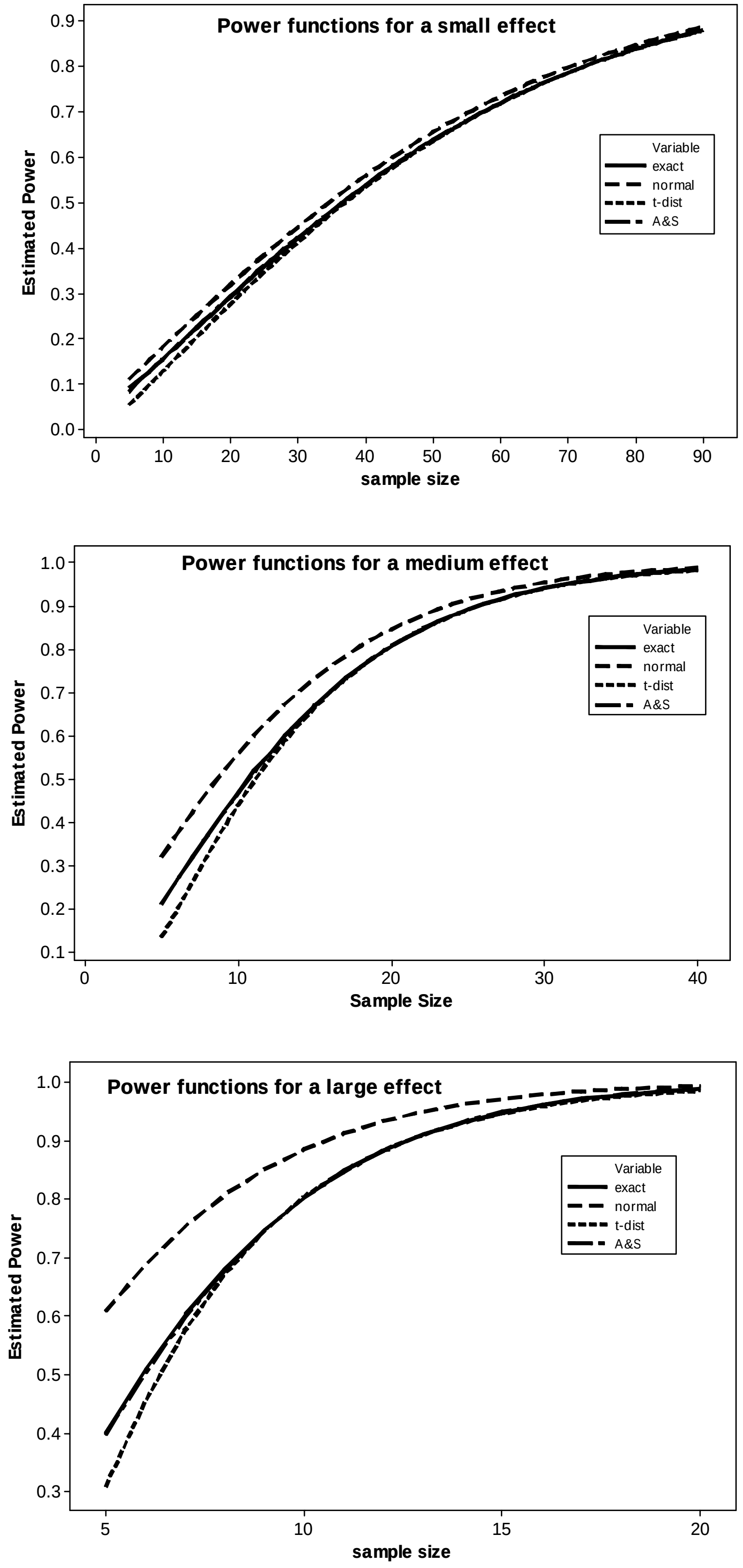
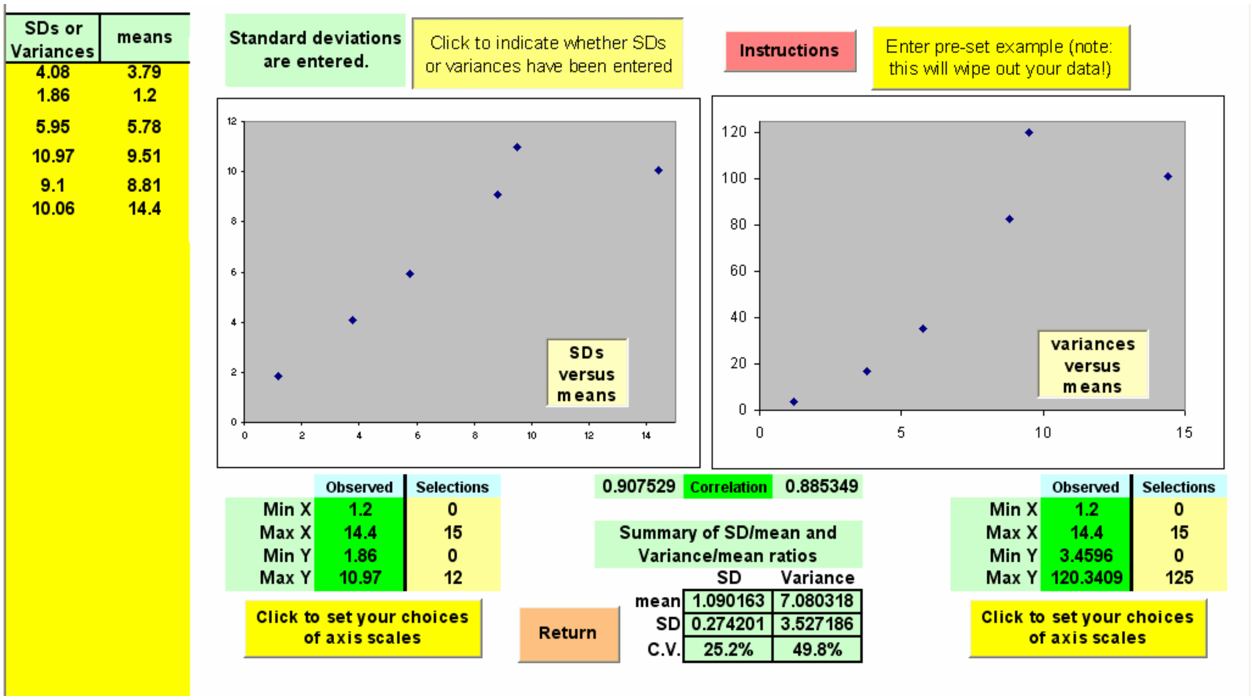

Power functions for a small effect
Estimated Power
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
sample size
0
10
20
30
40
50
60
70
80
90
Variable
exact
normal
t-dist
A&S
Power functions for a medium effect
Estimated Power
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Sample Size
0
10
20
30
40
Variable
exact
normal
t-dist
A&S
Power functions for a large effect
Estimated Power
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
sample size
5
10
15
20
Variable
exact
normal
t-dist
A&S

Figure 3



| SDs or Variances | means |
|---|---|
| 4.08 | 3.79 |
| 1.86 | 1.2 |
| 5.95 | 5.78 |
| 10.97 | 9.51 |
| 9.1 | 8.81 |
| 10.06 | 14.4 |

Standard deviations are entered.

Click to indicate whether SDs or variances have been entered

Instructions

Enter pre-set example (note: this will wipe out your data!)

| | Observed | Selections |
|---|---|---|
| Min X | 1.2 | 0 |
| Max X | 14.4 | 15 |
| Min Y | 1.86 | 0 |
| Max Y | 10.97 | 12 |

Click to set your choices of axis scales

0.907529 Correlation 0.885349

Summary of SD/mean and Variance/mean ratios

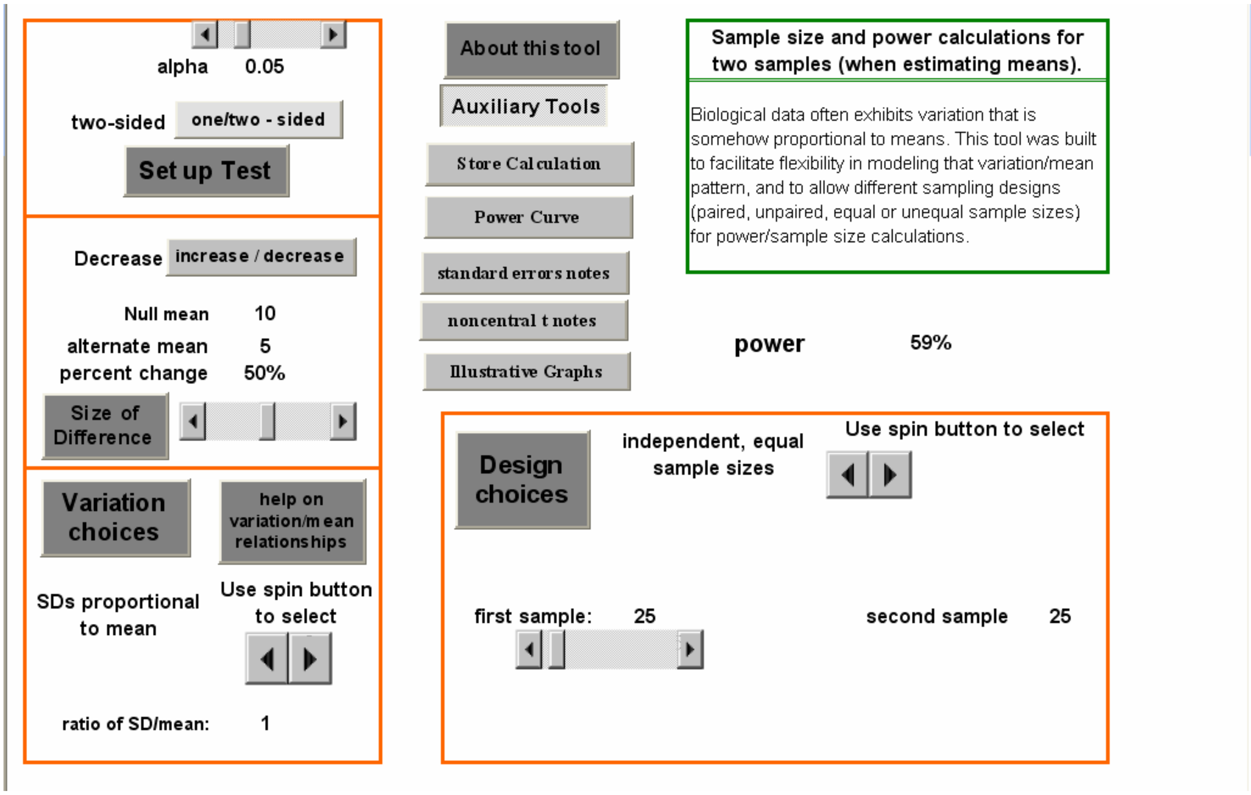
| | SD | Variance |
|---|---|---|
| mean | 1.090163 | 7.080318 |
| SD | 0.274201 | 3.527186 |
| C.V. | 25.2% | 49.8% |

Return

| | Observed | Selections |
|---|---|---|
| Min X | 1.2 | 0 |
| Max X | 14.4 | 15 |
| Min Y | 3.4596 | 0 |
| Max Y | 120.3409 | 125 |

Click to set your choices of axis scales

Figure 4



Sample size and power calculations for two samples (when estimating means).

Biological data often exhibits variation that is somehow proportional to means. This tool was built to facilitate flexibility in modeling that variation/mean pattern, and to allow different sampling designs (paired, unpaired, equal or unequal sample sizes) for power/sample size calculations.

alpha 0.05

two-sided

Decrease

Null mean 10

alternate mean 5

percent change 50%

SDs proportional to mean

ratio of SD/mean: 1

power 59%

independent, equal sample sizes

first sample: 25

second sample 25

1