

Power and Sample Size Concepts and Calculations

For a variety of statistical settings, one commonly approaches power and sample size issues from one of three starting points:

What sample size is required to achieve a confidence interval of specified width?


What sample size is necessary to achieve a specified power?

What power is achieved given a specified sample size?

In order to do any of these calculations, an estimate of the standard deviation in your sample (or possibly in each of several samples) is required. In many situations, power and sample size calculations are done before collecting data (in fact, it is usually a good idea to do so). Fortunately, it is possible to provide an estimate of the SD *without* having seen any data. If you can estimate the range (largest minus smallest) of values you will likely see, that range divided by three or four is a decent guess for the SD.

Sample Size Determined by Precision Goals

The width, which we will characterize by “the margin of error,” of confidence intervals is linked directly to sample size. A smaller desired margin of error requires greater precision (i.e. a smaller standard error), which in turn requires a larger sample size. In these notes, we assume that the sampling distribution of the relevant statistic is normal.

We'll address the concepts using as example one of the simplest of settings, a confidence interval (CI) for a population mean, based on a single sample mean, assuming Normality for the sampling distribution of the mean. If you aren't familiar with basic confidence interval construction and terminology,  .

There are three essentially equivalent ways you can impose some chosen precision on the CI. The one we'll use here is the margin of error, *m.e.*:

$$m.e. = t_{n-1, .95} \frac{s}{\sqrt{n}} \int_{.95}^{.975}$$

The margin of error is called by some the confidence interval “half-width.” Another common approach (technically equivalent) is to state the desired precision in terms of the CI width *w*. ($w = 2(m.e.)$).

A third way is as a percent, as in, “We want to be within plus or minus 10%.” Here the desired precision (we’ll call it *relative margin of error*, and label it with *rme*) is relative to the size of the mean; the connection to *m.e.* is

$$rme = 100\% \times \frac{m.e.}{\bar{y}}.$$

The take-home point: confidence interval precision in terms of CI margin of error, width, or relative margin of error are all equivalent. It is easy to translate from any one style to either of the other two. Given a choice of margin of error, the connection to sample size is straightforward conceptually, and, at first glance, looks like it leads to an easy computation:

$$m.e. = t_{n-1, .975} \frac{s}{\sqrt{n}}$$

↔ (square both sides; swap *n* and *m.e.*)

$$n = \left(t_{n-1, .975} \frac{s}{m.e.} \right)^2.$$

The problem with this formula is that one needs to know *n* in order to know *t* (since the correct *t*-distribution depends on the degrees of freedom) in order to calculate *n* in order to compute *t* in order... headache material.

A way out of the problem: it is a fact that as *n* increases, the multiplier $t_{n-1, (1-\alpha/2)}$ gets closer and closer to $z_{(1-\alpha/2)}$ (the corresponding


value from a standard Normal distribution). In fact, $t_{n-1,(1-\alpha/2)}$ is always going to be larger than $z_{(1-\alpha/2)}$ for any sample size. If we do an initial sample size estimate using $z_{(1-\alpha/2)}$ in the formula (which we *can* do easily), we will have a sample size that is smaller (likely not by much) than actually required. We then iteratively increase our estimated sample size until we reach the desired precision. An example may make this clearer.

Example.

A biologist wishes to make a 95% CI for a population mean, and wants the interval to be plus or minus 10% of the target mean. She estimates, after conferring with colleagues, that the mean will be about 50. These together mean that her desired CI margin of error is 5. She and her colleagues also estimate that the standard deviation will be about 10.

Steps	Formula	Calculation
Use basic formula to get a (slight under-) estimate	$n = \left(\frac{z_{.975} \times s}{m.e.} \right)^2$	$n = \left(\frac{1.96 \times 10}{5} \right)^2 = 15.37$
Use 16 (rounded up from 15.37) to see what m.e. is attained.	$m.e. = \frac{t_{16-1,.975} \times s}{\sqrt{n}}$	$m.e. = \frac{2.13 \times 10}{\sqrt{16}} = 5.325.$
That's bigger than our target. Add 1 to the sample size; try again	$m.e. = \frac{t_{n-1,.975} \times s}{\sqrt{n}}$	$m.e. = \frac{2.12 \times 10}{\sqrt{17}} = 5.142.$

<p>That's still bigger than our target. Add 1 to the sample size; try again. This time it works.</p>	$m.e. = \frac{t_{n-1, .975} \times s}{\sqrt{n}}$	$m.e. = \frac{2.11 \times 10}{\sqrt{18}} = 4.973.$
--	--	--

Clearly this would be pretty tedious to do routinely. To do these calculations easily and effortlessly,  .


Sample Size to Achieve Specified Power

The power of a statistical test is the probability of correctly detecting an existing effect (we'll define effect shortly). Power is a function of

1. the effect size (power increases as effect size increases);
 2. the chosen alpha level (power increases with larger alpha levels);
- and
3. the sample size (power increases as sample size increases).

Two type of errors can occur when doing a hypothesis test. One, a so-called false significance (or false rejections of the null hypothesis), has a specified chance, alpha, of occurring. The other, a false retention of the

null, occurs when one fails to reject the null hypothesis, but in fact some alternate is true. This probability goes by the name beta. Power is 1 -

beta.  to visually study the interplay between these two probabilities.

Power of a Statistical Test

In this discussion, we will assume that the [sampling distribution](#) of the relevant statistic is normal. The power of a test is the probability of correctly rejecting the null hypothesis when a specified alternate is true.

Swift Fox Example: (adapted from a study by **Olson** (1999)).

One question of interest in a study of swift foxes was whether home range size increased from summer to fall. A sample of 10 summer home ranges yielded a mean of 1000ha, with a SD of 320ha. The sampling distribution of the mean is assumed to be normal, with standard deviation estimated from the standard error of the summer

sample to be $se(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{320}{\sqrt{10}} \approx 100$.

This is a one-tailed hypothesis test (they are only interested in increases), for which they chose to use the conventional $\alpha = 0.05$.

Effect Size

In order to do a power analysis, one needs to establish effect sizes for which to do the calculations. Effect size is the difference between the value of a parameter under the null hypothesis and its value under an alternate. Effect size is a commonly used generic term, the context-dependent meaning of which will hopefully be made clear through the following examples.

Example 1. In a fisheries study, 15 randomly chosen net sites were chosen. A one-sided, one-sample t-test was used to test whether the mean number of trout per net was 6 (the historical average), or if, in fact, the mean number appears to be declining. The biologist wanted to be sure to have high power to detect a decline of 2 fish per net, if indeed the decline was that large. Here, the “effect size” of interest was a change of 2 in the mean number of trout caught per net.

Example 2. A study was done to compare male and female salaries among biologists of similar rank and seniority in a federal agency. They performed a two-sided, two-sample t-test for the null hypothesis of no difference. In discussion, it was decided that if the true difference was more than \$2,000 per year, they wanted to be sure to detect it. Here, the “effect size,” was a difference of \$2000 in the mean salaries.

In a regression study, the “effect size” might be a difference in slopes (that is, some chosen difference from the value being tested in the null hypothesis). “Effect size,” is a generic term that truly gains its meaning from the particular context.

One way to approach the question of effect size is to choose a range of effect sizes. Pick an effect that is sufficiently small so that any effects (should they exist) that are any smaller are not of much interest. Pick an effect that is sufficiently large so that any effects that are any larger are quite important to detect. Analyze power/sample sizes for that range of effects.

Swift Fox Example. For this study, an increase of 10% (100ha) was chosen for the smaller value, and 50% (500ha) for the larger.

Statistical Decision Rule

For the swift fox example, the researchers used $\alpha = 0.05$, which dictates that p -values less than 0.05 would lead to rejection of the null hypothesis, while for p -values larger than 0.05, we would fail to do so (at least, this is the formal, mechanistic way to use p -values and alpha). In a Normal distribution, 5% of the values are larger than 1.65 standard deviations above the mean. Given our SE of 100, we would, expect

means larger than 1165 about 5% of the time if the null hypothesis (no increase in home range size) were true.

The key point to keep in mind is that a mean home range size larger than 1165 will have a p-value less than $\alpha = 0.05$, and will lead to rejection of the null hypothesis. Let's consider the consequences of this decision rule under the assumption that, in fact, the average home range in the fall is 1100ha (for an effect size of 100); and let's examine the consequences for 1500ha (effect size is 500). The other pieces we need for the calculations are: $SD = 320$; $n = 10$; $\alpha = .05$; and the fact that


the test is one-tailed. Write those figures on a scrap of paper, and  .

What power is achieved given a specified sample size?

The power of a statistical test is a function of the effect size (power increases as effect size increases); the alpha level (power increases with larger alpha levels); and the sample size (power increases as sample size increases). This question is the complement of calculating the sample size required to achieve a given power. Usually, this question is asked for existing studies, to seek a better understanding of their limitations and for planning for future studies. The swift fox example had home range size estimates from $n = 10$ fox pairs. What size of change (if any) in home range size *can* they detect with that sample size?

Calculations for Paired data

A paired t -analysis proceeds as an analysis of a single sample, after having subtracted the values in one sample from those in the other. If you have existing data to work with, then use the sample of differences to directly estimate the standard deviation of the differences SD_d and proceed following the single sample instructions. The difficulty in working through power and sample size calculations without existing data is that the standard deviation of the differences is a function of the standard deviation in each sample *and* the correlation between the samples, which is really hard to estimate without data. That said, if you are blessed with enough insight into your data, you can use as an estimate $SD_d = S\sqrt{2-2r}$, where S is the (assumed to be the same for both) standard deviation within each sample, and r is the estimated correlation between the samples.

The following is a tool that incorporates several designs (two independent samples, paired, and a mixture of the two) along with flexible variance structures (assumed equal variances, arbitrary (your choice) variances, SD proportional to means, etcetera):  .

