Chapter 12

Statistical Power

12.1 The concept

The power of an experiment that you are about to carry out quantifies the chance that you will correctly reject the null hypothesis if some alternative hypothesis is really true.

Consider analysis of a k-level one-factor experiment using ANOVA. We arbitrarily choose $\alpha = 0.05$ (or some other value) as our significance level. We reject the null hypothesis, $\mu_1 = \cdots = \mu_k$, if the F statistic is so large as to occur less than 5% of the time when the null hypothesis is true (and the assumptions are met).

This approach requires computation of the distribution of F values that we would get if the model assumptions were true, the null hypothesis were true, and we would repeat the experiment many times, calculating a new F-value each time. This is called the null sampling distribution of the F-statistic (see Section 6.2.5).

For any sample size (n per group) and significance level (α) we can use the null sampling distribution to find a critical F-value "cutoff" *before* running the experiment, and know that we will reject H_0 if $F_{\text{experiment}} \geq F_{\text{critical}}$. If the assumptions are met (I won't keep repeating this) then 5% of the time when experiments are run on equivalent treatments, (i.e. $\mu_1 = \cdots = \mu_k$), we will falsely reject H_0 because our experiment's F-value happens to fall above F-critical. This is the so-called Type 1 error (see Section 8.4). We could lower α to reduce the chance that we will make such an error, but this will adversely affect the power of the experiment as explained next.

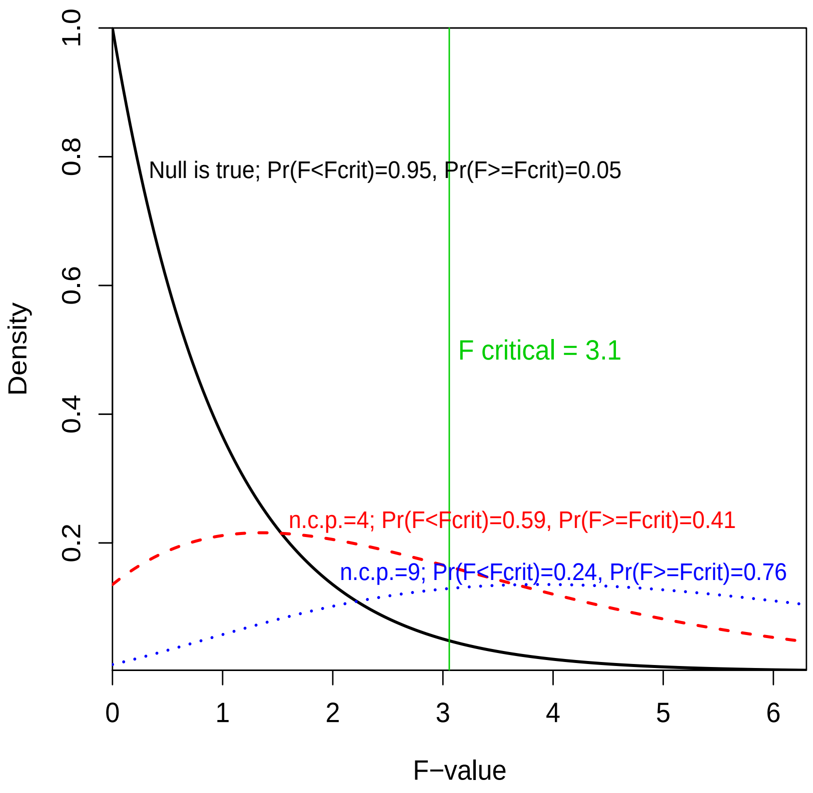


Figure 12.1: Null and alternative F sampling distributions.

Under each combination of n, underlying variance (σ^2) and some particular non-zero difference in population means (non-zero effect size) there is an *alternative* sampling distribution of F. An alternative sampling distribution represents how likely different values of a statistic such as F would be if we repeat an experiment many times when a particular alternative hypothesis is true. You can think of this as the histogram that results from running the experiment many times when the particular alternative is true and the F-statistic is calculated for each experiment.

As an example, figure 12.1 shows the null sampling distribution of the F-statistic for $k = 3$ treatments and $n = 50$ subjects per treatment (black, solid curve) plus the alternative sampling distribution of the F-statistic for two specific "alternative hypothesis scenarios" (red and green curves) labeled "n.c.p.=4" and "n.c.p.=9". For the moment, just recognize that n.c.p. stands for something called the “non-centrality parameter”, that the n.c.p. for the null hypothesis is 0, and that larger n.c.p. values correspond to less “null-like” alternatives.

Regarding this specific example, we note that the numerator of the F-statistic ($\text{MS}_{\text{between}}$) will have $k-1 = 2$ df, and the denominator($\text{MS}_{\text{within}}$) will have $k(n-1) = 147$ df. Therefore the null sampling distribution for the F-statistic that the computer has drawn for us is the (central) F-distribution (see Section 3.9.7) with 2 and 147 df. This is equivalent to the F-distribution with 2 and 147 df and with n.c.p.=0. The two alternative null sampling distributions (curves) that the computer has drawn correspond to two specific alternative scenarios. The two alternative distributions are called non-central F-distributions. They also have 2 and 147 df, but in addition have “non-centrality parameter” values equal to 4 and 9 respectively.

The whole concept of power is explained in this figure. First focus on the black curve labeled “null is true”. This curve is the null sampling distribution of F for *any* experiment with 1) three (categorical) levels of treatment; 2) a quantitative outcome for which the assumptions of Normality (at each level of treatment), equal variance and independent errors apply; 3) no difference in the three population means; and 4) a total of 150 subjects. The curve shows the values of the F-statistic that we are likely (high regions) or unlikely (low regions) to see if we repeat the experiment many times. The value of $\text{F}_{\text{critical}}$ of 3.1 separates (for k=3, n=50) the area under the null sampling distribution corresponding to the highest 5% of F-statistic values from the lowest 95% of F-statistic values. *Regardless of whether or not the null hypothesis is in fact true*, we will reject $H_0 : \mu_1 = \mu_2 = \mu_3$, i.e., we will *claim* that the null hypothesis is false, if our single observed F-statistic is greater than 3.1. Therefore it is built into our approach to statistical inference that among those experiments in which we study treatments that all have the same effect on the outcome, we will falsely reject the null hypothesis for about 5% of those experiments.

Now consider what happens if the null hypothesis is not true (but the error model assumptions hold). There are many ways that the null hypothesis can be false, so for any experiment, although there is only one null sampling distribution of F, there are (infinitely) many alternative sampling distributions of F. Two are shown in the figure. The information that needs to be specified to characterize a specific alternative sampling distribution is the spacing of the population means, the underlying variance at each fixed combination of explanatory variables (σ^2), and the number of subjects given each treatment (n). The number of treatments is also implicitly included on this list. I call all of this information an "alternative scenario". The alternative scenario information can be reduced through a simple formula to a single number called the non-centrality parameter (n.c.p.), and this additional parameter value is all that the computer needs to draw the alternative sampling distribution for an ANOVA F-statistic. Note that n.c.p.=0 represents the null scenario.

The figure shows alternative sampling distributions for two alternative scenarios in red (dashed) and blue (dotted). The red curve represents the scenario where $\sigma = 10$ and the true means are 10.0, 12.0, and 14.0, which can be shown to correspond to n.c.p.=4. The blue curve represents the scenario where $\sigma = 10$ and the true means are 10.0, 13.0, and 16.0, which can be shown to correspond to n.c.p.=9. Obviously when the mean parameters are spaced 3 apart (blue) the scenario is more un-null-like than when they are spaced 2 apart (red).

The alternative sampling distributions of F show how likely different F-statistic values are if the given alternative scenario is true. Looking at the red curve, we see that if you run many experiments when $\sigma^2 = 100$ and $\mu_1 = 10.0, \mu_2 = 12.0$, and $\mu_3 = 14.0$, then about 59% of the time you will get $F < 3.1$ and $p > 0.05$, while the remaining 41% of the time you will get $F \geq 3.1$ and $p \leq 0.05$. This indicates that for the *one* experiment that you can really afford to do, you have a 59% chance of arriving at the incorrect conclusion that the population means are equal, and a 41% chance of arriving at the correct conclusion that the population means are not all the same. This is not a very good situation to be in, because there is a large chance of missing the interesting finding that the treatments have a real effect on the outcome.

We call the chance of incorrectly retaining the null hypothesis the Type 2 error rate, and we call the chance of correctly rejecting the null hypothesis for any given alternative the power. Power is always equal to 1 (or 100%) minus the Type 2 error rate. High power is good, and typically power greater than 80% is arbitrarily considered "good enough".

In the figure, the alternative scenario with population mean spacing of 3.0 has fairly good power, 76%. If the true mean outcomes are 3.0 apart, and $\sigma = 10$ and there are 50 subjects in each of the three treatment groups, and the Normality, equal variance, and independent error assumptions are met, then any given experiment has a 76% chance of producing a p-value less than or equal to 0.05, which will result in the experimenter correctly concluding that the population means differ. But even if the experimenter does a terrific job of running this experiment, there is still a 24% chance of getting $p > 0.05$ and falsely concluding that the population means do *not* differ, thus making a Type 2 error. (Note that if this alternative scenario is correct, it is impossible to make a Type 1 error; such an error can only be made when the truth is that the population means do *not* differ.)

Of course, describing power in terms of the F-statistic in ANOVA is only one example of a general concept. The same concept applies with minor modifications for the t-statistic that we learned about for both the independent samples t-test and the t-tests of the coefficients in regression and ANCOVA, as well as other statistics we haven't yet discussed. In the cases of the t-statistic, the modification relates to the fact that "un-null-like" corresponds to t-statistic values far from zero on either side, rather than just larger values as for the F-statistic. Although the F-statistic will be used for the remainder of the power discussion, remember that the concepts apply to hypothesis testing in general.

You are probably not surprised to learn that for any given experiment and inference method (statistical test), the power to correctly reject a given alternative hypothesis lies somewhere between 5% and (almost) 100%. The next section discusses ways to improve power.

For one-way ANOVA, the null sampling distribution of the F-statistic shows that when the null hypothesis is true, an experimenter has a 95% chance of obtaining a p-value greater than 0.05, in which case she will make the correct conclusion, but 5% of the time she will obtain $p \leq 0.05$ and make a Type 1 error. The various alternative sampling distributions of the F-statistic show that the chance of making a Type 2 error can range from 95% down to near zero. The corresponding chance of obtaining $p \leq 0.05$ when a particular alternative scenario is true, called the power of the experiment, ranges from as low as 5% to near 100%.

12.2 Improving power

For this section we will focus on the two-group continuous outcome case because it is easier to demonstrate the effects of various factors on power in this simple setup. To make things concrete, assume that the experimental units are a random selection of news websites, the outcome is number of clicks (C) between 7 PM and 8 PM Eastern Standard Time for an associated online ad, and the two treatments are two fonts for the ads, say Palatino (P) vs. Verdana (V). We can equivalently analyze data from an experiment like this using either the independent samples t-test or one-way ANOVA.

One way to think about this problem is in terms of the two confidence intervals for the population means. Anything that reduces the overlap of these confidence intervals will increase the power. The overlap is reduced by reducing the common variance (σ^2), increasing the number of subjects in each group (n), or by increasing the distance between the population means, $|\mu_V - \mu_P|$.

This is demonstrated in figure 12.2. This figure shows an intuitive (rather than mathematically rigorous) view of the process of testing the equivalence of the population means of ad clicks for treatment P vs. treatment V. The top row represents population distributions of clicks for the two treatments. Each curve can be thought of as the histogram of the actual click outcomes for one font for all news websites on the World Wide Web. There is a lot of overlap between the two curves, so obviously it would not be very accurate to use, say, one website per font to try to determine if the population means differ.

The bottom row represents the sampling distributions of the sample means for the two treatments based on the given sample size (n) for each treatment. The key idea here is that, although the two curves always overlap, a smaller overlap corresponds to a greater chance that we will get a significant p-value for our one experiment.

Start with the second column of the figure. The upper panel shows that the truth is that σ^2 is 100, and $\mu_V = 13$, while $\mu_P = 17$. The arrow indicates that our sample has $n = 30$ websites with each font. The bottom panel of the second column shows the sampling distributions of sample means for the two treatments. The moderate degree of overlap, best seen by looking at the lower middle portion of the panel, is suggestive of less than ideal power.

The leftmost column shows the situation where the true common variance is now 25 instead of 100 (i.e., the s.d. is now 5 clicks instead of 10 clicks). This

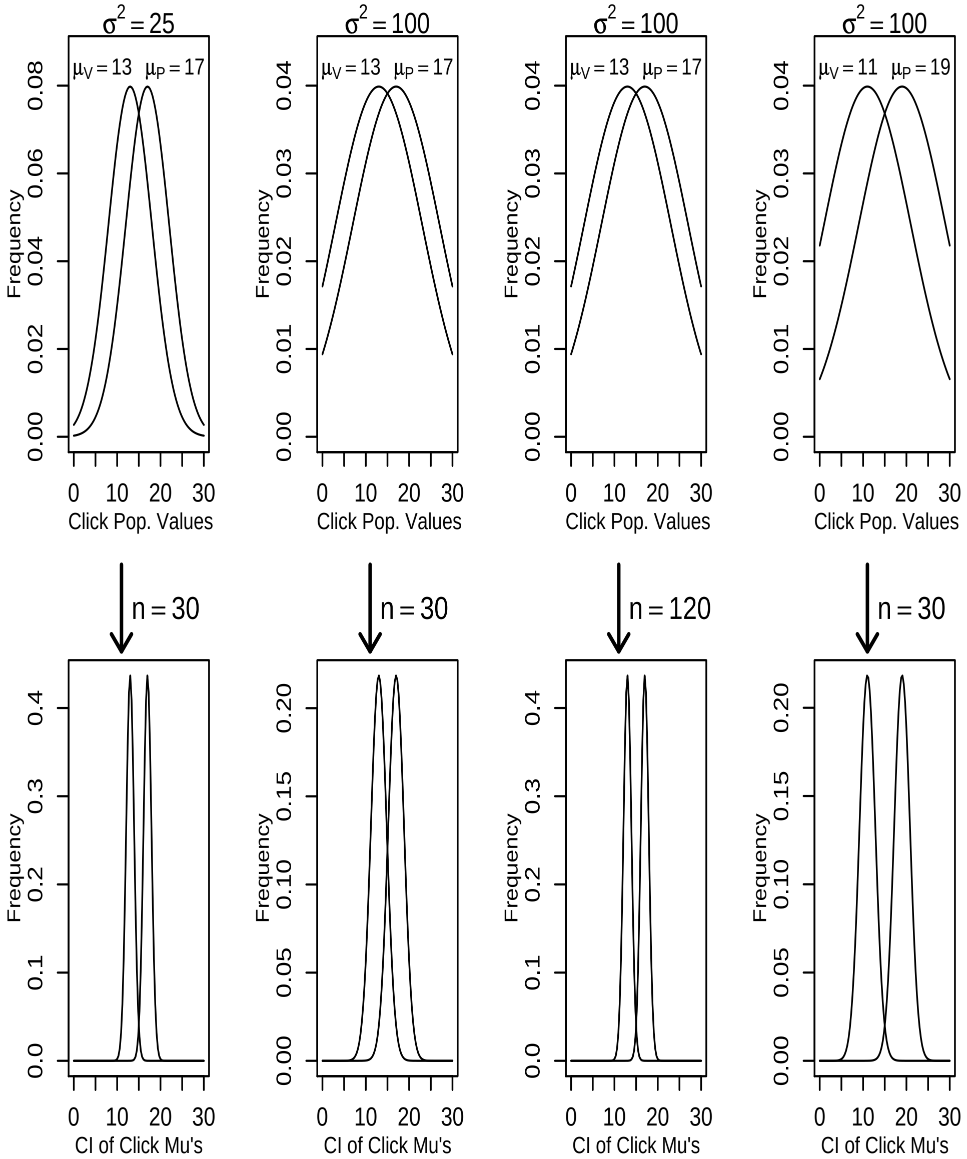


Figure 12.2: Effects of changing variance, sample size, and mean difference on power. Top row: population distributions of the outcome. Bottom row: sampling distributions of the sample mean for the given sample size.

markedly reduces the overlap, so the power is improved. How did we reduce the common variance? Either by reducing some of the four sources of variation or by using a within-subjects design, or by using a blocking variable or quantitative control variable. Specific examples for reducing the sources of variation include using only television-related websites, controlling the position of the ad on the website, and using only one font size for the ad. (Presumably for this experiment there is no measurement error.) A within-subjects design would, e.g., randomly present one font from 7:00 to 7:30 and the other font from 7:30 to 8:00 for each website (which is considered the "subject" here), but would need a different analysis than the independent-samples t-test. Blocking would involve, e.g., using some important (categorical) aspect of the news websites, such as television-related vs. non-television related as a second factor whose p-value is not of primary interest (in a 2-way ANOVA). We would guess that for each level of this second variable the variance of the outcome for either treatment would be smaller than if we had ignored the television-relatedness factor. Finally using a quantitative variable like site volume (hit count) as an additional explanatory variable in an ANCOVA setting would similarly reduce variability (i.e., σ^2) at each hit count value.

The third column shows what happens if the sample size is increased. Increasing the sample size four-fold turns out to have the same effect on the confidence curves, and therefore the power, as reducing the variance four-fold. Of course, increasing sample size increases cost and duration of the study.

The fourth column shows what happens if the population mean difference, sometimes called (unadjusted) effect size, is increased. Although the sampling distributions are not narrowed, they are more distantly separated, thus reducing overlap and increasing the power. In this example, it is hard to see how the difference between the two fonts can be made larger, but in other experiments it is possible to make the treatments more different (i.e., make the active treatment, but not the control, "stronger") to increase power.

Here is a description of another experiment with examples of how to improve the power. We want to test the effect of three kinds of fertilizer on plant growth (in grams). First we consider reducing the common variability of final plant weight for each fertilizer type. We can reduce measurement error by using a high quality laboratory balance instead of a cheap hardware store scale. And we can have a detailed, careful procedure for washing off the dirt from the roots and removing excess water before weighing. Subject-to-subject variation can be reduced by using only one variety of plant and doing whatever is possible to ensure that the plants are of similar size at the start of the experiment. Environmental variation can be reduced by assuring equal sunlight and water during the experiment. And treatment application variation can be reduced by carefully measuring and applying the fertilizer to the plants. As mentioned in section 8.5 reduction in all sources of variation except measurement variability tends to also reduce generalizability.

As usual, having more plants per fertilizer improves power, but at the expense of extra cost. We can also increase population mean differences by using a larger amount of fertilizer and/or running the experiment for a longer period of time. (Both of the latter ideas are based on the assumption that the plants grow at a constant rate proportional to the amount of fertilizer, but with different rates per unit time for the same amount of different fertilizers.)

A within-subjects design is not possible here, because a single plant cannot be tested on more than one fertilizer type.

Blocking could be done based on different fields if the plants are grown outside in several different fields, or based on a subjective measure of initial "healthiness" of the plants (determined before randomizing plants to the different fertilizers). If the fertilizer is a source of, say, magnesium in different chemical forms, and if the plants are grown outside in natural soil, a possible control variable is the amount of nitrogen in the soil near each plant. Each of these blocking/control variables are expected to affect the outcome, but are not of primary interest. By including them in the means model, we are creating finer, more homogeneous divisions of "the set of experimental units with all explanatory variables set to the same values". The inherent variability of each of these sets of units, which we call σ^2 for any model, is smaller than for the larger, less homogeneous sets that we get when we don't include these variables in our model.

Reducing σ^2, increasing n, and increasing the spacing between population means will all reduce the overlap of the sampling distributions of the means, thus increasing power.

12.3 Specific researchers' lifetime experiences

People often confuse the probability of a Type 1 error and/or the probability of a Type 2 error with the probability that a given research result is false. This section attempts to clarify the situation by looking at several specific (fake) researchers' experiences over the course of their careers.

Remember that a given null hypothesis, H_0, is either true or false, but we can never know this truth for sure. Also, for a given experiment, the standard decision rule tells us that when $p \leq \alpha$ we should reject the null hypothesis, and when $p > \alpha$ we should retain it. But again, we can never know for sure whether our inference is actually correct or incorrect.

Next we need to clarify the definitions of some common terms. A "positive" result for an experiment means finding $p \leq \alpha$, which is the situation for which we reject H_0 and claim an interesting finding. "Negative" means finding $p > \alpha$, which is the situation for which we retain H_0 and therefore don't have enough evidence to claim an interesting finding. "True" means correct (i.e. reject H_0 when H_0 is false or retain H_0 when H_0 is true), and "false" mean incorrect. These terms are commonly put together, e.g., a false positive refers to the case where $p \leq 0.05$, but the null hypothesis is actually true.

Here are some examples in which we pretend that we have omniscience, although the researcher in question does not. Let $\alpha = 0.05$ unless otherwise specified.

1. Neetika Null studies the effects of various chants on blood sugar level. Every week she studies 15 controls and 15 people who chant a particular word from the dictionary for 5 minutes. After 1000 weeks (and 1000 words) what is her Type 1 error rate (positives among null experiments), Type 2 error rate (negatives among non-null experiments) and power (positives among non-null experiments)? What percent of her positives are true? What percent of her negatives are true?

   This description suggests that the null hypothesis is always true, i.e. I assume that chants don't change blood sugar level, and certainly not within five minutes. Her Type 1 error rate is $\alpha = 0.05$. Her Type 2 error rate (sometimes called β) and power are not applicable because no alternative hypothesis is ever true. Out of 1000 experiments, 1000 are null in the sense that the null hypothesis is true. Because the probability of getting $p \leq 0.05$ in an experiment where the null hypothesis is true is 5%, she will see about 50 positive and 950 negative experiments. For Neetika, although she does not know it, every time she sees $p \leq 0.05$ she will mistakenly reject the null hypothesis, for a 100% error rate. But every time she sees $p > 0.05$ she will correctly retain the null hypothesis for an error rate of 0%.

2. Stacy Safety studies the effects on glucose levels of injecting cats with subcutaneous insulin at different body locations. She divides the surface of a cat into 1000 zones and each week studies injection of 10 cats with water and 10 cats with insulin in a different zone.

   This description suggests that the null hypothesis is always false. Because Stacy is studying a powerful treatment and will have a small measurement error, her power will be large; let's use 80%=0.80 as an example. Her Type 2 error rate will be β=1-power=0.2, or 20%. Out of 1000 experiments, all 1000 are non-null, so Type 1 error is not applicable. With a power of 80% we know that each experiment has an 80% chance of giving $p \leq 0.05$ and a 20% chance of given $p > 0.05$. So we expect around 800 positives and 200 negatives. Although Stacy doesn't know it, every time she sees $p \leq 0.05$ she will correctly reject the null hypothesis, for a 0% error rate. But every time she sees $p > 0.05$ she will mistakenly retain the null hypothesis for an error rate of 100%.

3. Rima Regular works for a large pharmaceutical firm performing initial screening of potential new oral hypoglycemic drugs. Each week for 1000 weeks she gives 100 rats a placebo and 100 rats a new drug, then tests blood sugar. To increase power (at the expense of more false positives) she chooses $\alpha = 0.10$.

   For concreteness let's assume that the null hypothesis is true 90% of the time. Let's consider the situation where among the 10% of candidate drugs that work, half have a strength that corresponds to power equal to 50% (for the given n and σ^2) and the other half correspond to power equal to 70%.

   Out of 1000 experiments, 900 are null with around 0.10*900=90 positive and 810 negative experiments. Of the 50 non-null experiments with 50% power, we expect around 0.50*50=25 positive and 25 negative experiments. Of the 50 non-null experiments with 70% power, we expect around 0.70*50=35 positive and 15 negative experiments. So among the 100 non-null experiments (i.e., when Rima is studying drugs that really work) 25+35=60 out of 100 will correctly give $p \leq 0.05$. Therefore Rima's average power is 60/100 or 60%.

> Although Rima doesn't know it, when she sees $p \leq 0.05$ and rejects the null hypothesis, around 60/(90+60)=0.40=40% of the time she is correctly rejecting the null hypothesis, and therefore 60% of the time when she rejects the null hypothesis she is making a mistake. Of the 810+40=850 experiments for which she finds $p > 0.05$ and retains the null hypothesis, she is correct 810/(810+40)=0.953=95.3% of time and she makes an error 4.7% of the time. (Note that this value of approximately 95% is only a coincidence, and not related to $\alpha = 0.05$; in fact $\alpha = 0.10$ for this problem.)
>
> These error rates are not too bad given Rima's goals, but they are not very intuitively related to $\alpha = 0.10$ and power equal to 50 or 70%. The 60% error rate among drugs that are flagged for further study (i.e., have $p \leq 0.05$) just indicates that some time and money will be spent to find out which of these drugs are not really useful. This is better than not investigating a drug that really works. In fact, Rima might make even more money for her company if she raises α to 0.20, causing more money to be wasted investigating truly useless drugs, but preventing some possible money-making drugs from slipping through as useless. By the way, the overall error rate is (90+40)/1000=13%.

Conclusion: For *your* career, you cannot know the chance that a negative result is an error or the chance that a positive result is an error. And these are what you would really like to know! But you do know that when you study "ineffective" treatments (and perform an appropriate statistical analysis) you have only a 5% chance of incorrectly claiming they are "effective". And you know that the more you increase the power of an experiment, the better your chances are of detecting a truly effective treatment.

It is worth knowing something about the relationship of power to confidence intervals. Roughly, wide confidence intervals correspond to experiments with low power, and narrow confidence intervals correspond to experiments with good power.

The error rates that experimenters are really interested in, i.e., the probability that I am making an error for my current experiment, are not knowable. These error rates differ from both α and β=1-power.

12.4 Expected Mean Square

Although a full treatment of "expected mean squares" is quite technical, a superficial understanding is not difficult and greatly aids understanding of several other topics. EMS tells us what values we will get for any given mean square (MS) statistic under either the null or an alternative distribution, on average over repeated experiments.

If we have k population treatment means, we can define $\bar{\mu} = \frac{\sum_{i=1}^{k} \mu_i}{k}$ as the mean of the population treatment means, and $\lambda_i = \mu_i - \bar{\mu}$ (where λ is read "lambda"), and $\sigma_A^2 = \frac{\sum_{i=1}^{k} \lambda_i^2}{k-1}$. The quantity σ_A^2 is not a variance, because it is calculated from fixed parameters rather than from random quantities, but it obviously is a "variance-like" quantity. Notice that we can express our usual null hypothesis as $H_0 : \sigma_A^2 = 0$ because if all of the μ's are equal, then all of the λ's equal zero. We can similarly define σ_B^2 and σ_{A*B}^2 for a 2 way design.

Let σ_e^2 be the true error variance (including subject-to-subject, treatment application, environmental, and measurement variability). We haven't been using the subscript "e" up to this point, but here we will use it to be sure we can distinguish various symbols that all include σ^2. As usual, n is the number of subjects per group. For 2-way ANOVA, a (instead of k) is the number of levels of factor A and b is the number of levels of factor B.

The EMS tables for one-way and two-way designs are shown in table 12.1 and 12.2.

Remember that all of the between-subjects ANOVA F-statistics are ratios of mean squares with various means squares in the numerator and with the error mean square in the denominator. From the EMS tables, you can see why, for either design, under the null hypothesis, the F ratios that we have been using are appropriate and have "central F" sampling distributions (mean near 1). You can also see why, under any alternative, these F ratios tend to get bigger. You can also see that power can be increased by increasing the spacing between population means ("treatment strength") via increased values of $|\lambda|$, by increasing n, or by decreasing σ_e^2. This formula also demonstrates that the value of σ_e^2 is irrelevant to the sampling distributing of the F-statistic (cancels out) when the null hypothesis is true, i.e., $\sigma_A^2 = 0$.

Source of Variation	MS	EMS
Factor A	MS_A	$\sigma_e^2 + n\sigma_A^2$
Error (residual)	MS_{error}	σ_e^2

Table 12.1: Expected mean squares for a one-way ANOVA.

Source of Variation	MS	EMS
Factor A	MS_A	$\sigma_e^2 + bn\sigma_A^2$
Factor B	MS_B	$\sigma_e^2 + an\sigma_B^2$
A*B interaction	MS_{A*B}	$\sigma_e^2 + n\sigma_{AB}^2$
Error (residual)	MS_{error}	σ_e^2

Table 12.2: Expected mean squares for a two-way ANOVA.

For the mathematically inclined, the EMS formulas give a good idea of what aspects of an experiment affect the F ratio.

12.5 Power Calculations

In case it is not yet obvious, I want to reiterate why it is imperative to calculate power for your experiment *before* running it. It is possible and common for experiments to have low power, e.g., in the range of 20 to 70%. If you are studying a treatment which is effective in changing the population mean of your outcome, and your experiment has, e.g., 40% power for detecting the true mean difference, and you conduct the experiment perfectly and analyze it appropriately, you have a 60% chance of getting a p-value of greater than 0.05, in which case you will erroneously conclude that the treatment is ineffective. To prevent wasted experiments, you should calculate power and only perform the experiment if there is a reasonably high power.

It is worth noting that you will not be able to calculate the “true” power of your experiment. Rather you will use a combination of mathematics and judgement to make a useful estimation of the power.

There are an infinite number of alternative hypothesis. For any of them we can increase power by 1) increasing n (sample size) or 2) decreasing experimental error (σ_e^2). Also, among the alternatives, those with larger effect sizes (population mean differences) will have more power. These statements derive directly from the EMS interpretive form of the F equation (shown here for 1-way ANOVA):

$$\text{Expected Value of F} = \text{Expected value of } \frac{MS_A}{MS_{\text{error}}} \approx \frac{\sigma_e^2 + n\sigma_A^2}{\sigma_e^2}$$

Obviously increasing n or σ_A^2 increases the average value of F. Regarding the effect of changing σ_e^2, a small example will make this more clear. Consider the case where $n\sigma_A^2 = 10$ and $\sigma_e^2 = 10$. In this case the average F value is 20/10=2. Now reduce σ_e^2 to 1. In this case the average F value is 11/1=11, which is much bigger, resulting in more power.

In practice, we try to calculate the power of an experiment for one or a few reasonable alternative hypotheses. We try not to get carried away by considering alternatives with huge effects that are unlikely to occur. Instead we try to devise alternatives that are fairly conservative and reflect what might really happen (see the next section).

What we need to know to calculate power? Beyond k and alpha (α), we need to know sample size (which we may be able to increase if we have enough resources), an estimate of experimental error (variance or σ_e^2, which we may be able to reduce, possibly in a trade-off with generalizability), and reasonable estimates of true effect sizes.

For any set of these three things, which we will call an "alternative hypothesis scenario", we can find the sampling distribution of F under that alternative hypothesis. Then it is easy to find the power.

We often estimate σ_e^2 with residual MS, or error MS (MSE), or within-group MS from previous similar experiments. Or we can use the square of the actual or guessed standard deviation of the outcome measurement for a number of subjects exposed to the same (any) treatment. Or, assuming Normality, we can use expert knowledge to guesstimate the 95% range of a homogenous group of subjects, then estimate σ_e as that range divided by 4. (This works because 95% of a normal distribution is encompassed by mean plus or minus 2 s.d.) A similar trick is to estimate σ_e as 3/4 of the IQR (see Section 4.2.4), then square that quantity.

Be careful! If you use too large (pessimistic) of a value for σ_e^2 your computed power will be smaller than your true power. If you use too small (optimistic) of a value for σ_e^2 your computed power will be larger than your true power.

12.6 Choosing effect sizes

As mentioned above, you want to calculate power for "reasonable" effect sizes that you consider achievable. A similar goal is to choose effects sizes such that smaller effects would not be scientifically interesting. In either case, it is obvious that choosing effect sizes is not a statistical exercise, but rather one requiring subject matter or possibly policy level expertise.

I will give a few simple examples here, choosing subject matter that is known to most people or easily explainable. The first example is for a categorical outcome, even though we haven't yet discussed statistical analyses for such experiments. Consider an experiment to see if a certain change in a TV commercial for a political advisor's candidate will make a difference in an election. Here is the kind of thinking that goes into defining the effect sizes for which we will calculate the power. From prior subject matter knowledge, he estimate that about one fourth of the voting public will see the commercial. He also estimates that a change of 1% in the total vote will be enough to get him excited that redoing this commercial is a worthwhile expense. So therefore an effect size of 4% difference in a favorable response towards his candidate is the effect size that is reasonable to test for.

Now consider an example of a farmer who wants to know if it's worth it to move her tomato crop in the future to a farther, but more sunny slope. She estimates that the cost of initially preparing the field is \$2000, the yearly extra cost of transportation to the new field is \$200, and she would like any payoff to happen within 4 years. The effect size is the difference in crop yield in pounds of tomatoes per plant. She can put 1000 plants in either field, and a pound of tomatoes sells for \$1 wholesale. So for each 1 pound of effect size, she gains \$1000 per year. Over 4 years she needs to pay off \$2000+4(\$200)=\$2800. She concludes that she needs to have good power, say 80%, to detect an effect size of 2.8/4=0.7 additional pounds of tomatoes per plant (i.e., a gain of \$700 per year).

Finally consider a psychologist who wants to test the effects of a drug on memory. She knows that people typically remember 40 out of 50 items on this test. She really wouldn't get too excited if the drug raised the score to 41, but she certainly wouldn't want to miss it if the drug raised the score to 45. She decides to "power her study" for $\mu_1 = 40$ vs. $\mu_2 = 42.5$. If she adjusts n to get 80% power for these population test score means, then she has an 80% chance of getting $p \leq 0.05$ when the true effect is a difference of 2.5, and some larger (calculable) power for a difference of 5.0, and some smaller (calculable) non-zero, but less than ideal, power for a difference of 1.0.

In general, you should consider the smallest effect size that you consider interesting and try to achieve reasonable power for that effect size, while also realizing that there is more power for larger effects and less power for smaller effects. Sometimes it is worth calculating power for a range of different effect sizes.

12.7 Using n.c.p. to calculate power

The material in this section is optional.

Here we will focus on the simple case of power in a one-way between-subjects design. The "manual" calculation steps are shown here. Understanding these may aid your understanding of power calculation in general, but ordinarily you will use a computer (perhaps a web applet) to calculate power.

Under any particular alternative distribution the numerator of F is inflated, and F follows the non-central F distribution with $k-1$ and $k(n-1)$ degrees of freedom and with "non-centrality parameter" equal to:

$$\text{n.c.p.} = \frac{n \cdot \sum_{i=1}^{k} \lambda_i^2}{\sigma_e^2}$$

where n is the proposed number of subjects in *each* of the groups we are comparing. The bigger the n.c.p., the more the alternative sampling distribution moves to the right and the more power we have.

Manual calculation example: Let $\alpha = 0.10$ and $n = 11$ per cell. In a similar experiment MSE=36. What is the power for the alternative hypothesis $H_A : \mu_1 = 10$, $\mu_2 = 12$, $\mu_3 = 14$, $\mu_4 = 16$?

1. Under the null hypothesis the F-statistic will follow the central F distribution (i.e., n.c.p.=0) with $k-1=3$ and $k(n-1)=40$ df. Using a computer or F table we find $F_{\text{critical}} = 2.23$.

2. Since $\bar{\mu}$=(10+12+14+16)/4=13, the λ's are -3,-1,1,3, so the non-centrality parameter is

$$\frac{11(9+1+1+9)}{36} = 6.11.$$

3. The power is the area under the non-central F curve with 3,40 df and n.c.p.=6.11 that is to the right of 2.23. Using a computer or non-central F table, we find that the area is 0.62. This means that we have a 62% chance of rejecting the null hypothesis if the given alternate hypothesis is true.

4. An interesting question is what is the power if we double the sample size to 22 per cell. df_{error} is now 21*4=84 and F_{critical} is now 2.15. The n.c.p.=12.22. From the appropriate non-central F distribution we find that the power increases to 90%.

In practice we will use a Java applet to calculate power.

In R, the commands that give the values in the above example are:

```
qf(1-0.10, 3, 40) # result is 2.226092 for alpha=0.10
1-pf(2.23, 3, 40, 6.11) # result is 0.6168411
qf(1-0.10, 3, 84) # result is 2.150162
1-pf(2.15,3, 84, 12.22) # result is 0.8994447
```

In SPSS, put the value of 1-α (here, 1-0.10=0.90) in a spreadsheet cell, e.g., in a column named "P". The use Transform/Compute to create a variable called, say, "Fcrit", using the formula "IDF.F(P,3,40)". This will give 2.23. The use Transform/Compute to create a variable called, say, "power", using the formula "1-NCDF.F(Fcrit,3,40,6.11)". This will give 0.62.

12.8 A power applet

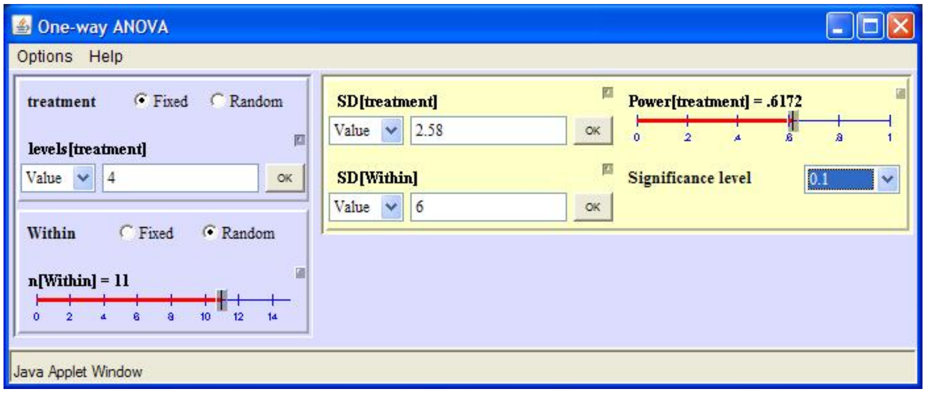
The Russ Lenth power applet is very nice way to calculate power. It is available on the web at http://www.cs.uiowa.edu/~rlenth/Power. If you are using it more that occasionally you should copy the applet to your website. Here I will cover ANOVA and regression. Additional topic are in future chapters.

12.8.1 Overview

To get started with the Lenth Power Applet, select a method such as Linear Regression or Balanced ANOVA, then click the “Run Selection” button. A new window will open with the applet for the statistical method you have chosen. Every time you see sliders for entering numeric values, you may also click the small square at upper right to change to a text box form for entering the value. The Help menu item explains what each input slider or box is for.

12.8.2 One-way ANOVA

This part of the applet works for one-way and two-way balanced ANOVA. Remember that balanced indicates equal numbers of subjects per group. For one-way ANOVA, leave the “Built-in models” drop-down box at the default value of “One-way ANOVA”.



One-way ANOVA

treatment ● Fixed ○ Random

levels[treatment] Value 4

Within ○ Fixed ● Random

n[Within] = 11

SD[treatment] Value 2.58

SD[Within] Value 6

Power[treatment] = .6172

Significance level 0.1

Figure 12.3: One-way ANOVA with Lenth power applet.

Enter “n” under “Observations per factor combination”, and click to study the power of “F tests”. A window opens that looks like figure 12.3.

On the left, enter “k” under “levels[treatment] (Fixed)”. Under “n[Within] (Random)” you can change n.

On the right enter σ_e (σ) under “SD[Within]” (on the standard deviation, not variance scale) and α under “Significance level”. Finally you need to enter the "effect size" in the form of "SD[treatment]". For this applet the formula is

$$\text{SD[treatment]} = \sqrt{\frac{\sum_{i=1}^{k} \lambda_i^2}{k-1}}$$

where λ_i is $\mu_i - \bar{\mu}$ as in section 12.4.

For $H_A: \mu_1 = 10$, $\mu_2 = 12$, $\mu_3 = 14$, $\mu_4 = 16$, $\bar{\mu} = 13$ and $\lambda_1 = -3$, $\lambda_2 = -1$, $\lambda_3 = +1$, $\lambda_4 = +3$.

$$\begin{aligned} \text{SD[treatment]} &= \sqrt{\frac{\sum_{i=1}^{k} \lambda_i^2}{k-1}} \\ &= \sqrt{\frac{(-3)^2 + (-1)^2 + (+1)^2 + (+3)^2}{3}} \\ &= \sqrt{20/3} \\ &= 2.58 \end{aligned}$$

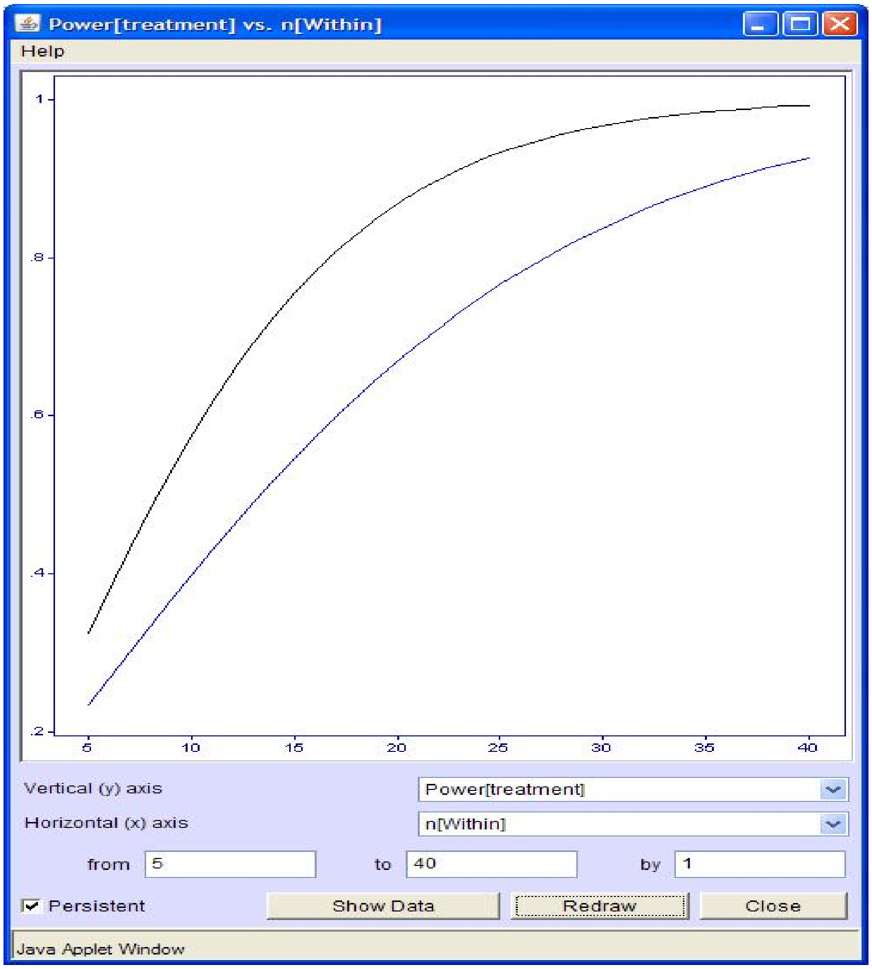
You can also use the menu item "SD Helper" under Options to graphically set the means and have the applet calculate SD[treatment].

Following the example of section 12.7 we can plug in SD[treatment]=2.58, $n = 11$, and $\sigma_e = 6$ to get power=0.6172, which matches the manual calculation of section 12.7

At this point it is often useful to make a power plot. Choose Graph under the Options menu item. The most useful graph has "Power[treatment]" on the y-axis and "n[Within]" on the x-axis. Continuing with the above example I would choose to plot power "from" 5 "to" 40 "by" 1. When I click "Draw", I see the power for this experiment for different possible sample sizes. An interesting addition can be obtained by clicking "Persistent", then changing "SD[treatment]" in the main window to another reasonable value, e.g., 2 (for $H_A: \mu_1 = 10$, $\mu_2 = 10$, $\mu_3 = 10$, $\mu_4 = 14$), and clicking OK. Now the plot shows power as a function of n for two (or more) effect sizes. In Windows you can use the Alt-PrintScreen key combination to copy the plot to the clipboard, then paste it into another application. The result is shown in figure 12.4. The lower curve is for the smaller value of SD[treatment].

12.8.3 Two-way ANOVA without interaction

Select "Two-way ANOVA (additive model)". Click "F tests". In the new window, on the left enter the number of levels for each of the two factors under "levels[row]



Vertical (y) axis: Power[treatment]

Horizontal (x) axis: n[Within]

from 5 to 40 by 1

☑ Persistent

Figure 12.4: One-way ANOVA power plot from Lenth power applet.

(Fixed)" and "levels[col] (Fixed)". Enter the number of subjects for each cell under "Replications (Random)".

Enter the estimate of σ under "SD[Residual]" and the enter the "Significance level".

Calculate "SD[row]" and "SD[col]" as in the one-way ANOVA calculation for "SD[treatment]", but the means for either factor are now averaged over all levels of the other factor.

Here is an example. The table shows cell population means for each combination of levels of the two treatment factors for which additivity holds (e.g., a profile plot would show parallel lines).

Row factor / Column Factor	Level 1	Level 2	Level 3	Row Mean
Level 1	10	20	15	15
Level 2	13	23	18	18
Col. Mean	11.5	21.5	16.5	16.5

Averaging over the other factor we see that for the column means, using some fairly obvious invented notation we get $H_{ColAlt} : \mu_{C1} = 11.5, \mu_{C2} = 21.5, \mu_{C3} = 16.5$. The row means are $H_{RowAlt} : \mu_{R1} = 15, \mu_{R2} = 18$.

Therefore SD[row] is the square root of $((-1.5)^2 + (+1.5)^2)/1$ which is 2.12. The value of SD[col] is the square root of $((-5)^2 + (+5)^2 + (0)^2)/2$ which equals 5. If we choose $\alpha = 0.05$, $n = 8$ per cell, and estimate σ at 8, then the power is a not-so-good 24.6% for H_{RowAlt}, but a very good 87.4% for H_{ColAlt}.

12.8.4 Two-way ANOVA with interaction

You may someday find it useful to calculate the power for a two-way ANOVA interaction. It's fairly complicated!

Select "Two-way ANOVA". Click "F tests". In the new window, on the left enter the number of levels for each of the two factors under "levels[row] (Fixed)" and "levels[col] (Fixed)". Enter the number of subjects for each cell under "Replications (Random)".

Enter the estimate of σ under "SD[Residual]" and the enter the "Significance level".

The treatment effects are a bit more complicated here. Consider a table of cell means in which additivity does not hold.

Row factor / Column Factor	Level 1	Level 2	Level 3	Row Mean
Level 1	10	20	15	15
Level 2	13	20	18	17
Col. Mean	11.5	20.0	16.5	16

For the row effects, which come from the row means of 15 and 17, we subtract 16 from each to get the λ values of -1 and 1, then find SD[row]=$\sqrt{\frac{(-1)^2+(1)^2}{1}} = 1.41$.

For the column effects, which come from the column means of 11.5, 20.0, and 16.5, we subtract their common mean of 16 to get λ values of -4.5, 4.0, and 0.5, and then find that SD[col]=$\sqrt{\frac{(-4.5)^2+(4.0)^2+(0.5)^2}{2}} = 4.27$.

To calculate "SD[row*col]" we need to calculate for each of the 6 cells, the value of $\mu_{ij} - (\bar{\mu} + \lambda_{i.} + \lambda_{.j})$ where μ_{ij} indicates the i^{th} row and j^{th} column, and $\lambda_{i.}$ is the λ value for the i^{th} row mean, and $\lambda_{.j}$ is the λ value for the j^{th} column mean. For example, for the top left cell we get 10-(16-4.5-1.0)=-0.5. The complete table is

Row factor / Column Factor	Level 1	Level 2	Level 3	Row Mean
Level 1	-0.5	1.0	-0.5	0.0
Level 2	+0.5	-1.0	0.5	0.0
Col. Mean	0.0	0.0	0.0	0.0

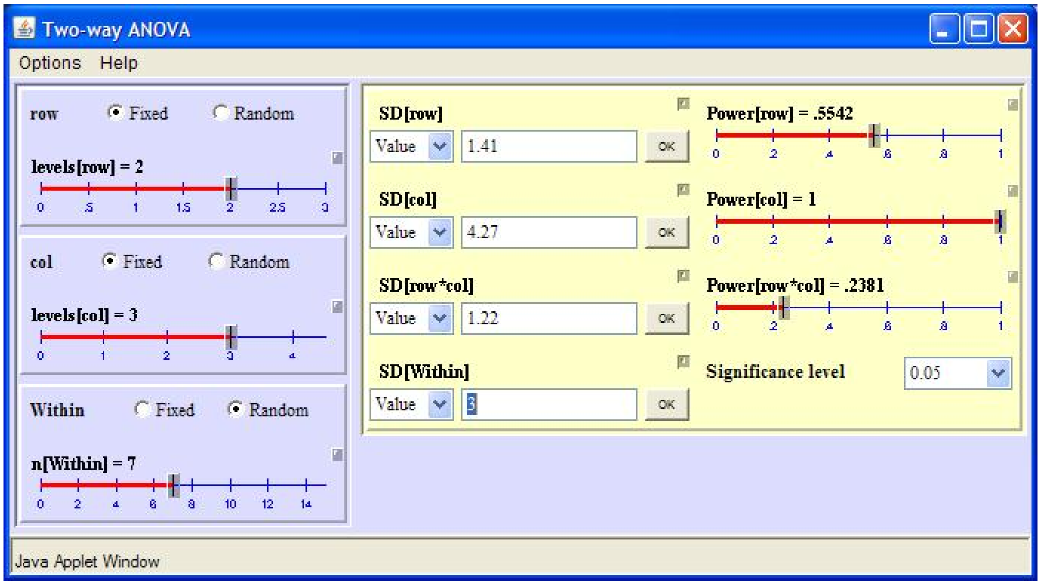
You will know you constructed the table correctly if all of the margins are zero. To find SD[row*col], sum the squares of all of the (non-marginal) cells, then divide by $(r-1)$ and $(c-1)$ where r and c are the number of levels in the row and column factors, then take the square root. Here we get SD[row*col]=$\sqrt{\frac{0.25+1.0+0.25+0.25+1.0+0.25}{1 \cdot 2}} = 1.22$.

If we choose $\alpha = 0.05$, $n = 7$ per cell, and estimate σ at 3, then the power is a not-so-good 23.8% for detecting the interaction (gettin an interaction p-value less than 0.05). This is shown in figure 12.5.

12.8.5 Linear Regression

We will just look at simple linear regression (one explanatory variable). In addition to the α, n, and σ, and the effect size for the slope, we need to characterize the spacing of the *explanatory* variable.

Choose "Linear regression" in the applet and the Linear Regression dialog shown in figure 12.6 appears. Leave "No. of predictors" (number of explanatory variables) at 1, and set "Alpha", "Error SD" (estimate of σ), and "(Total) Sample



row ● Fixed ○ Random

levels[row] = 2

col ● Fixed ○ Random

levels[col] = 3

Within ○ Fixed ● Random

n[Within] = 7

SD[row] Value 1.41

SD[col] Value 4.27

SD[row*col] Value 1.22

SD[Within] Value 3

Power[row] = .5542

Power[col] = 1

Power[row*col] = .2381

Significance level 0.05

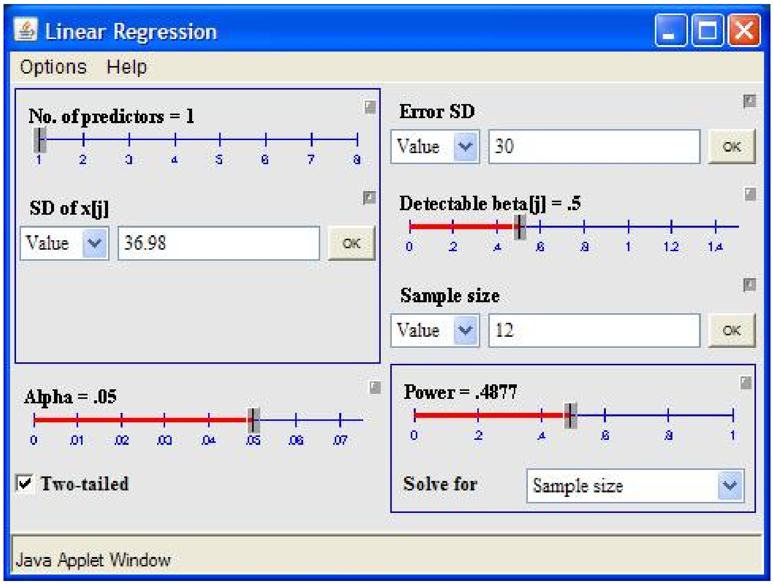
Figure 12.5: Two-way ANOVA with Lenth power applet.

size".

Under "SD of x[j]" enter the standard deviation of the x values you will use. Here we use the fact that the spread of any number of repetitions of a set of values is the same as just one set of those values. Also, because the x values are fixed, we use n instead of $n-1$ in the denominator of the standard deviation formula. E.g., if we plan to use 5 subjects each at doses, 0, 25, 50, and 100 (which have a mean of 43.75), then SD of x[j] $= \sqrt{\frac{(0-43.75)^2+(25-43.75)^2+(50-43.75)^2+(100-43.75)^2}{4}} = 36.98$.

Plugging in this value and $\sigma = 30$, and a sample size of 3*4=12, and an effect size of beta[j] (slope) equal to 0.5, we get power = 48.8%, which is not good enough.

In a nutshell: Just like the most commonly used value for alpha is 0.05, you will find that (arbitrarily) the most common approach people take is to find the value of n that achieves a power of 80% for some specific, carefully chosen alternative hypothesis. Although there is a bit of educated guesswork in calculating (estimating) power, it is strongly advised to make some power calculations before running an experiment to find out if you have enough power to make running the experiment worthwhile.



No. of predictors = 1

SD of x[j]: Value 36.98

Alpha = .05

☑ Two-tailed

Error SD: Value 30

Detectable beta[j] = .5

Sample size: Value 12

Power = .4877

Solve for: Sample size

Figure 12.6: Linear regression with Lenth power applet.