Uni- and Bivariate Power

Note that the relationship between risk and power is unidirectional. Power depends on risk, but risk is completely independent of power. The reason for this is as follows: α (which completely defines and determines risk) is always directly involved in the decision-making process, because the *p*-value from the inferential test is compared to α when you decide whether to reject the null hypothesis (i.e., "significant" vs "not significant"). Therefore, the value of α has an effect on the decision regardless of whether the null hypothesis is really true or really false. In contrast, β (which defines power, albeit via 1– β) does not explicitly appear in the decision-making process. Therefore, while β is used to describe the accuracy of the decision-making process when the null hypothesis is false, it does not have anything to do with what occurs when the null hypothesis is true.

Factors that Influence Risk

None. α is whatever you set it to be. We use 5%. Deal.

Factors that Influence Power

The magnitude of the violation of the null hypothesis.
The spread and shape of the hypothetical sampling distribution, which depends on
    the assumed or inferred shape of sampling distribution,
    the size of the sample, and
    the variability within the sample, which depends on
        the variability within the population.
The level of risk (i.e., α).

Recall (from the chapter on null-hypo statistical testing in general) that power is the probability of rejecting the null hypothesis on the assumption that it is false. It is denoted by 1–β. In psychology, there is a movement afoot to do all that we can to set power to .80 when-ever we conduct an inferential test. In other words, it is suggested that you not run an experiment or conduct a study unless you have an 80% chance of rejecting a false null hypothesis.

Why do we set risk to 5%, but only aim for 80% power?

This has more to do with history and the social rules of science than with statistics. In psychology, we seem to have a particular distaste for false alarms and/or we don't seem to worry so much about misses. Please do not read that as derogatory. The fact that we set risk so low and don't worry so much about power has a very useful function in a world where anyone is allowed to collect data and submit a manuscript. It allows us to protect the field from the unwanted influence of lousy experimenters. (Feel free to read that one as derogatory.)

To understand this, note the following: (1) journal space is limited and we have decided to give priority to papers that report significant results, (2) α is under the direct control of the data analyst, (3) β is not under direct control, and (4) most people don't know how to calculate β and/or would be shocked to learn how high it usually is.

First, note again that α is under direct control of the data analyst. Early on, psychology (as a field) opted to set this value at .05, because this seemed a good trade-off between errors and how many subjects would typically have to be run in experiments. So consider α to be fixed for all time.

Next, the reason for giving priority to papers that report significant results is this: anybody -- and I really mean anybody -- can run an experiment that fails to reject the null hypothesis. All you need to do is employ such lousy techniques as to guarantee a large amount of "noise" (or error variance) in the sample; this will widen the hypothetical sampling distribution to such an extent that no null hypothesis -- regardless of how wrong it might be -- could ever be rejected using $\alpha = .05$.

☺ Note: If you are one of those people for whom the label "lousy experimenter" is appropriate, take heart. You can still have a successful career in advertizing or work for a drug company (which would pay about 10 times as much as psychology -- not that I'm bitter). Think about it this way: what does it mean when an ad says that "nothing has been shown to work better" or "no side effects above those with placebo"? It means that a study failed to reject the null hypothesis (that other products work just as well or that side-effects with the drug are the same as with a placebo). So if "sample variance" are two of your middle names, off you go! A generic drug company has a corner office with a window just waiting for you.

Finally, because experiments that fail to reject the null hypothesis are rarely published, we need not worry too much about power, or so goes the logic. Low power would only be a problem (to the field) if non-significant results were often published and this started to make people believe that the effects being tested did not exist. Put another way, the field self-corrects for low power by having lots of practitioners. If an effect really does exist (i.e., the null hypothesis really should be rejected), then somebody, someday, will find the effect and publish a paper.

The danger in all this (and the reason that some statisticians have begun to argue that we really need to pay more attention to power) is that low power can have serious effects on the thought processes of individual researchers. If you conduct your work on the assumption that a failure to reject the null hypothesis should be taken as evidence that the effect of interest does not exist, but you never bother to check your level of power, then you might close off very promising lines of research on the basis of an experiment that had very little chance of detecting the effect.

The other danger is more subtle, but has been getting more attention of late. If you adopt the attitude that low power isn't a problem because, with enough people running experiments, someone will find the effect, then you are walking straight into a form of "*p*-hacking." Assume, for the moment, that a certain effect does not exist, but would be really cool if it did. So lots of people -- by which I mean more than 20 -- run experiments looking for the effect. Well, with that many chances for a false-alarm error to occur, one of those experiments will probably produce a significant result. And then it will be published. And everyone else (who got non-significant results because, as I said, the effect does not exist) will slap their foreheads and say "doh!" to conceal their jealousy of the one person or group that, in truth, just published a false-alarm error. In any event, the belief in the effect will be stronger than before and maybe even stronger than if the effect were uninteresting and nobody else went looking for it.

Also, it can be argued that conducting experiments with low power is unethical. If the experiment

had no chance of finding the effect, then the subjects were exposed to the risks of the study with no hope of any benefits to society.

For these reasons -- plus the very practical fact that granting agencies often demand that you conduct a power analysis for every experiment -- we will be paying close attention to power in all of the analyses that are covered in this course.

Power Defined as the Number of Subjects Required, N*

With all of the above said, it must now be admitted that the precise value of $1-\beta$ is not very useful, because it can only be calculated *after* the experiment or study has been run and analyzed (which "seals" the data). What we want is a method of getting power to .80 (or more) ***in advance***; we want to run experiments and studies that have .80 power. However, as you'll see in a moment, this will require some guess-work.

Recall from above that power depends on many factors, several of which are not under the control of the researcher. In particular, power depends on the size of the effect (i.e., by how much the null hypothesis is actually violated), the amount of unexplained variability (i.e., the standard error), the size of the sample (N), and the level of risk (i.e., α). We always set α at .05, so consider that to be both known and fixed. The size of the effect and the amount of error are not known, but we treat them as fixed and get estimates of their values from previous research. That leaves one thing as the determinant of power: the size of the sample. The way that modern, before-the-fact (*a priori*) power analyses are done is to calculate the number of subjects that are required to have at least .80 power (on the assumption that the effect has a certain size and error). This number of required subjects is referred to as ***N**** ("en star").

Calculating N*

Step 1: Estimate the Standardized Effect Size

*Note: this section used to be very, very long. I started with a variety of different measures of effect size (incl. Cohen's **d**, which used to be popular) and worked each different measure to a common end point. That's all gone. It's not how things are done these days.*

The first step to finding N* is to combine the two estimated values that influence power (i.e., the effect's size and error) into a single number. This number is the **standardized effect size**. Not only does combining the size of the effect and the error into a single number make it easier to calculate N*, but because we use the standard error as our measure of the latter, the units cancel, such that the standardized effect size has no units and can be compared across any set of dependent measures.

The modern way of quantifying the standardized effect is in terms of the proportion of variance explained (PoVe or just PoV). The PoVe combines both of the two determinants of power into a single value and it works for any type of data: within-subject experiments, between-subject experiments, pseudo-experiments, and correlations. Therefore, to complete this step of a power analysis, you need to come up with a best-guess about $p\eta^2$ (if you are planning to run a

paired-samples *t*-test), η^2 (if you are planning to run an independent-samples *t*-test), or r^2 (if you are planning to run a bivariate correlation of some sort). You can get this best guess from a pilot experiment or from previous, existing experiments that are similar to what you are planning to run. Note, here, how useful it is to have the formula for converting any value of *t* (with a certain df) to a value of $p\eta^2$ or η^2; even if the authors of the previous, similar experiment didn't report a measure of effect size or association, you can calculate it yourself from what they did provide.

The magical equation that converts the results from previous experiments to a PoVe is this:

$$\text{PoVe (i.e., } p\eta^2 \text{ or } \eta^2 \text{ or } r^2) \quad = \quad t^2 / (t^2 + df)$$

Step 2: Calculate f^2

In the second step, you convert the value of PoVe (i.e., $p\eta^2$ or η^2 or r^2) to something that more directly matches what matters to power, which is f^2 ("little-f-squared"). It's a very simple transformation, but it is critical. The relationship between PoVe and power is not linear; as PoVe approaches 1.00, power shoots up, such that very few subjects are required for significance. The actual relationship is this:

$$f^2 = \frac{p\eta^2}{1 - p\eta^2} \quad \text{or} \quad \frac{\eta^2}{1 - \eta^2} \quad \text{or} \quad \frac{r^2}{1 - r^2}$$

Note the parallel between f^2 and an *F*-statistic: both are the ratio of explained variance to unexplained variance. That isn't an accident. The value of f^2 is close to being a best guess of what *F* will be if you replicate the experiment.

*Step 3: Calculate df**

The next step is where we take α into account. Whether a given value of *t* or *F* constitutes sufficient evidence to reject the null hypothesis depends on both α and the degrees of freedom. The value of degrees of freedom depends on the number of subjects and the design that's employed; we'll deal with those later. For now, we'll focus on α.

In one of the most important advances in the history of power analysis, Jacob Cohen (see, esp., the 1988 version of *Statistical Power*) came up with a set of multipliers for converting f^2 to the number of subjects required to have .80 power for any of a variety of values of α and $1-\beta$. The one drawback to Cohen's approach was that he presented it in a manner that only works for between-subject design. In what follows, I'll give you a version that works for all types of design. The one "trick" to getting Cohen's method to work was to add an additional step: first calculate the required value of df (for a given f^2 and α and $1-\beta$); then convert this value of df* (i.e., degrees of freedom required for .80 power) to N* later.

The general formula for calculating df* (based on Cohen's work) is this:

$$df^* \quad = \quad L / f^2$$

where L is a semi-magical constant, worked out by Cohen, that depends on the number of independent variables or predictors in the design, as well as both α and 1– β. For all forms of *t*-test and bivariate correlation (i.e., situations with exactly one predictor), the value of L for α=.05 and 1–β=.80 is 7.85. (For folks who do power analyses, 7.85 is right up there with 1.96 as a constant to memorize.) Thus, the df* equation for *t*-tests and bivariate correlation is:

$$df^* = 7.85 / f^2$$

Note: in this case, round anything up to get a whole number. And I really mean *anything*. Round up from X.000001 to X+1. *Why?* Because we promised <u>*at least*</u> .80 power. To guarantee the "at least" part, we round anything up.

*Step 4: Calculate N**

This is where the type of design comes back into play, because different designs produce different values of df for the same number of actual subjects.

A univariate or paired-samples *t*-test has N–1 df, where *N* is the number of pairs in the latter case. Therefore, to get N*, just add one to the value of df* from Step 3.

An independent-samples *t*-test has N–2 df. Therefore, add 2 to df* to get the total number of subjects and then divide the resulting N* by 2 (rounding anything up) to get the number of subjects per group.

A bivariate correlation of any sort also has N–2 df, so add 2 to df* to get N*.

Complications

Caveat #1: In the case of a point-biserial correlation (i.e., a dichotomous variable, such as sex, crossed with a quantitative variable, such as anxiety), the value of N* will only ensure .80 power if approximately half of the subjects fall into each of the two "pseudo-groups" (i.e., when the dichotomous variable splits 50/50 in the sample). As the dichotomous variable deviates from a 50/50 split, you need to run some extra subjects. How many extra depends on this formula:

$$\text{new } N^* = N^* / (4\, p_0\, p_1)$$

where p_0 is the expected proportion of subjects in the pseudo-group coded by zero and p_1 is the expected proportion in the pseudo-group coded by one. (By definition: $p_0 + p_1$ must = 1.00) Note that when $p_0 = p_1 = 0.50$ (i.e., when the two pseudo-groups are equal in size), the above formula does not alter N* at all.

Caveat #2: Similar to the above, for independent-samples *t*-tests, N* will only ensure .80 power if the two groups are equal in size (i.e., when both have N*/2 members). If you plan to use group sizes that are unequal (for what-ever reason -- although I've yet to hear a good one), then the value of N* must be increased using the method above, where p_0 and p_1 are now the proportions that you plan to put in each group. Apply this correction to the original, total N* -- i.e., before you split N*

into the sizes of each of the two groups. Then multiply the new N* by p_0 or p_1 to get the final sizes for the two groups. As always, round anything up.

Practice

$p\eta^2$ for a within-subjects effect is .65. How many subjects are needed for .80 power?

$f^2 = 1.8571$ df* = 4.2270 → 5 N* = 6

η^2 for a between-subjects effect is .35. How many subjects are needed *per group* for .80 power?

$f^2 = 0.5385$ df* = 14.5775 → 15 N* = 17, thus 9 per group

Same as above, but you plan, for some reason, to put twice as many subjects in one of the groups.

original N* = 17 new N* = 19.2221 → 20 thus 14 and 7

r^2 for a certain relationship is .40. How many subjects are needed for .80 power?

$f^2 = 0.6667$ df* = 11.7744 → 12 N* = 14