# Planning for Power

Here I shall elaborate on Type I and Type II errors, and try to explain why power is a Good Thing in statistics, and why it is so elusive. Statistics is used to manage the uncertainty that arises from not knowing the true state of nature.

### Statistical power and conclusion validity

In the lectures on t-tests power was mentioned in relation to how the two-sample test is affected when there are serious problems of heterogeneity: It was neccessary to use much reduced degrees of freedom in the significance test in order to counteract the increased risk of a Type I error, which is a threat to conclusion validity.

On the other hand, reduced degrees of freedom and the consequent loss of power pose a new threat to conclusion validity in the guise of increased risk of a Type II error. To understand why, a close look at the notion of power is required.

# Statistical Power

Howell defines power as **the probability of correctly rejecting a false null hypothesis**. His treatment of power is mainly in terms of the formal statistics, and may be more difficult than Trochim's definition: **the odds that you observe an effect when it occurs**. Both definitions can be understood better if you refer back to Howell's Table 8.1 (3rd & 4th editions, or Table 11-1 in 2nd edition), or to the similar table in Trochim showing Type I and Type II errors in relation to hypothesis testing. I have reproduced a modified version of this table below:

A imagem não pode ser exibida. Talvez o computador não tenha memória suficiente para abrir a imagem ou talvez ela esteja corrompida. Reinicie o computador e abra o arquivo novamente. Se ainda assim aparecer o x vermelho, poderá ser necessário excluir a imagem e inseri-la novamente.

In the diagram the sampling distribution of the mean under the null hypothesis is shown in red, and the sampling distribution of the mean when the alternate hypothesis is true is shown in blue. For simplicity, a one-tailed test is illustrated.

Consider the situation on the left. If your sample mean (in green) falls in the rejection region, then by the rules you must reject the null hypothesis. This is correct but would in reality be a Type I error. Random sampling error has produced a result that is pure noise but looks indistinguishable from the situation on the right in which the alternate hypothesis is true and there is no error. Fortunately, we ourselves formulate the null hypothesis and also set the level of alpha according to our research aims, and therefore can manage that risk effectively:

Exploratory research

> We have little or no idea of the true state of nature. A relatively large risk of Type I error (e.g. .10) is acceptable because we primarily seek information.

Confirmatory research

> We have a reasonably good idea of the true state of nature but seek confirmation. Only a small risk of Type I error (e.g. .01) is acceptable.

Power refers to the odds that the outcome of your research will be in the top right-hand corner of the figure, where you correctly reject $H_0$. Thus power counteracts the threat of a Type I error: The inevitable (albeit manageable) risk that a decision to reject the null hypothesis could be wrong is countered by the power to ensure that it will be right. Unfortunately, this power depends partly on the true state of nature and is thus not entirely within your control. That fact is indicated in the table above by the probabilities which add to 1.0 downwards but not across. In plain words this means that the odds of a Type I error (alpha) are quite independent from the odds of a Type II error (beta). So, for example, you could have a high alpha (risk of a Type I error) and at the same time have very high power: Ideal for exploratory research where you might need to discern weak signals amidst much noise. On the other hand you could have high alpha and low power: The worst scenario in which there is little control over threats to conclusion validity.

### The true state of nature

From the above there are two important points to note:

- Correctly following the rules of statistical inference can nevertheless lead to an error.
- The type of error depends both on the research and on the ***true state of nature***.

This is why the table has two independent dimensions. Whatever decision you make is not going to affect reality (except perhaps in the long term), and whatever the state of nature is will not directly affect your decision.

### Inconclusive results

Consider next the threat of a Type II error (*ß or beta*). Again, this is partly beyond our control: Our null hypothesis might really be false but the effect we actually seek too subtle to be detected and the null hypothesis is mistakenly retained. The result would be written off as inconclusive. However, with sufficient power even very subtle effects can be detected. Therefore, if you can show that despite having to retain the null hypothesis your research had good power to detect an effect if there was one, then you can further argue that your result was not inconclusive. So power is the converse of *ß*, in other words 1 - *ß*.

## Planning for power

It is desirable to plan research so you have a reasonable chance of success. Unfortunately, this is seldom done. One reason is that research always involves an element of uncertainty. If the outcome can be predicted, why bother to do the research? Although it is not possible to predict the outcome, a reasonable estimate can often be made before you begin. The following table (based on Trochim) summarizes the required ingredients.

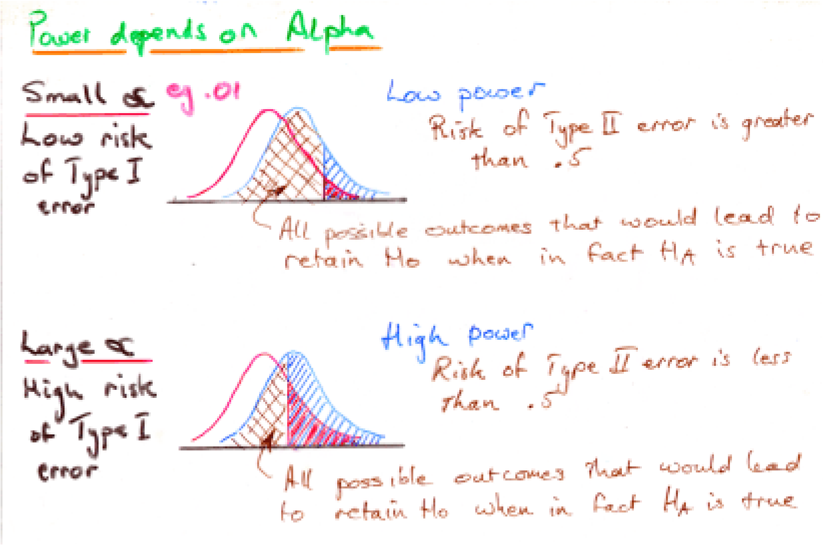
**Four components determine the outcome of research**

- Power: The odds that you will observe a treatment effect when it occurs.
- Alpha level: Acceptable risk, or odds that an observed result is due to chance.
- Effect size: Salience of the treatment relative to noise in measurement.
- Sample size: How many units of information accessible.

With respect to planning, the key is that given any three of these, the fourth can be calculated. Hence you can plan for a suitable level of power. As defined, power is the chance of success. You want this to be as high as possible: Howell gives details for precise calculations, but what matters here is only: How can you maximize it?

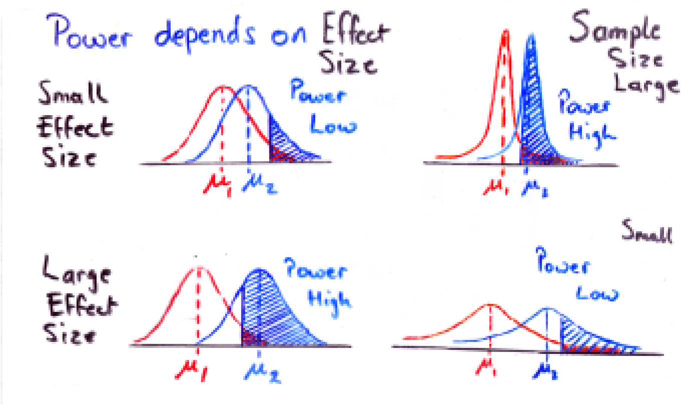
The figure below shows how the choice of *alpha* changes power in a one-tailed test. The red and blue curves in these diagrams are sampling distributions of the mean. They show all possible outcomes in the situation where either $H_0$ is true, or where $H_A$ is true.

It is only meaningful to speak of power and Type II errors in terms of the alternate hypothesis. Power is an area beneath the blue curve. The risk *ß* of a Type II error (in brown) when the null is retained but in fact is false, is also under the blue curve. Power is the remainder: `1 - ß`. It comprises all possible samples with a mean **greater** than what is required for rejection of $H_0$ and therefore occupies the region to the right (in this particular one-tailed test).


Power depends on Alpha
Small α eg .01
Low risk of Type I error
Low power
Risk of Type II error is greater than .5
All possible outcomes that would lead to retain Ho when in fact HA is true
Large α
High risk of Type I error
High power
Risk of Type II error is less than .5
All possible outcomes that would lead to retain Ho when in fact HA is true

Increasing alpha (e.g. .10) increases power, but also the risk of a Type I error. You are free to choose any significance level, but convention often fixes *alpha* at .05 or less.

The power increases with effect size:



Effect size is formally defined in terms of population parameters: We have little control over population parameters, they are fixed by the Almighty! How can we improve power? Trochim correctly points out that effect size really is the ratio of signal to noise (where noise is *sigma*): Poor methods can weaken the signal and increase the noise. Reliable methods will increase the observed effect size relative to the noise, and increase power.

What remains, is the sample size: Increasing sample size directly increases power.

This leads to the pragmatic side of the matter: Research is often costly. To obtain funding you might justify the need for your expenses in terms of the research goals. But do you have a reasonable chance of achieving these goals? In statistical terms, power can answer this question. It balances the pragmatic and statistical issues in a neat equation.

Suppose you are planning some research and work out that with the proposed methodology you have a 20% chance of making a real contribution to knowledge. Fund managers would look askance. How about a similar research proposal that can claim an 80% chance? The 20% versus 80% comparison is the question of power in pragmatic terms. It is the same as saying that one research proposal claims a .2 probability of correctly rejecting a false $H_0$, versus a claim of .8 probability to be able to do the same.

On the other hand, large samples can be costly. If it can be shown that reasonable power is available with a small sample rather than a big one then the costs and benefits can be evaluated. Such a comparison can be made, but only if it is possible to estimate power before the research is carried out.

## Risky Tradeoffs

It should now be clear that the possibility of error is always present, and that the risk of Type I errors is not independent of the risk of Type II errors. The following table is an attempt to summarize what was said above.

| **Exploratory** | **Confirmatory** |
|---|---|
| **Much noise** | **Little noise** |
| **Type I OK** | **Type II OK** |
| High alpha | Low alpha |
|  | **Good design** |

| **Minimize Noise** | **Randomize Noise** |
|---|---|
| Repeated measures | Randomized groups |
| Matched groups | |

**Reliable**

Instruments

Procedures

Refer to the comparison in Howell's Table 15.2 (3rd and 4th editions) for an example of the difference between using repeated measures (one sample tests) versus randomized groups (two-sample tests).

## Estimating power

What finally determines power is effect size, and that refers to the shadowy world of population parameters which lurk behind what we see indistinctly in our samples. What we see in our samples depends on the amount of information they contain.

Howell shows how to calculate the sample size needed to get any level of power. Unfortunately, the calculation is different for each kind of statistical test, and some statistical tests require computations that are only available in specialized computer packages. For these reasons power calculations are beyond the scope of this course. You need only deal with it on a conceptual level as outlined above, especially in terms of Trochim's discussion of threats and remedies.