What is Power?

- When conducting a research study using a statistical hypothesis test, power is the probability of getting statistical significance when the null hypothesis is false.

  - Key Points
    - Hypothesis testing
    - Power is a probability
    - Statistical significance
    - Null hypothesis is false
Hypothesis Testing Procedure

- Set up a null (or no effect) hypothesis (this is the hypothesis that the researcher would like to disprove).
- Conduct the study -- Draw one or more random samples (depending on your design) from the population, possibly apply a treatment, and measure each individual on the construct(s) of interest, compute any statistics necessary.
- Compare the results from the sample(s) to the null hypothesis. The sample result is based on a statistic calculated using the data from the sample(s). This statistic is often based on the sample mean, but other statistics are also used.
  - If the sample result is highly unlikely assuming the null hypothesis is true, reject the null hypothesis.
  - If the sample result is likely or just somewhat unlikely assuming the null hypothesis is true, retain the null hypothesis as a possibility.
Statistical Inference

- When we conduct a hypothesis test, the purpose is to draw a conclusion about a population based on data from a sample.

- **Population** – all of the individual units that are of interest (they might be people, schools, mice, or anything that can be counted).

- **Sample** – a subset of the population. We generally assume that we have a random sample which means each individual in the population has an equal chance of being selected for the sample.
How do we decide if the sample result is highly unlikely?

- Define highly unlikely in terms of probability. This is called alpha (α). It is also called the Type I error rate – the probability of rejecting a true null hypothesis. If the likelihood of the sample result is less than alpha then we feel confident saying that the null hypothesis must be false.
## Table 10.4-1  Decision Outcomes Categorized

**True Situation**

<table>
<thead>
<tr>
<th>Decision</th>
<th>$H_0$ true</th>
<th>$H_0$ false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fail to reject $H_0$</td>
<td>Correct acceptance Probability = $1 - \alpha$</td>
<td>Type II error Probability = $\beta$</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>Type I error Probability = $\alpha$</td>
<td>Correct rejection Probability = $1 - \beta$</td>
</tr>
</tbody>
</table>

Table 10.4-1 (p. 277) from Kirk (5 ed.)
Type I Error

- Incorrectly rejecting the null hypothesis. In other words, claiming that you found something in your study when in fact there was no effect.
- Researchers long ago decided that it was advantageous to minimize the probability of making a Type I error. That way you wouldn’t have lots of people running around making claims that they had discovered something when in fact the result was just due to chance.
Type I Error (cont.)

- In order to allow the researcher to have strict control over Type I errors, the hypothesis testing procedure is set up so that the researcher picks the value for the Type I error rate that s/he is willing to tolerate.
- The probability of a Type I error is called alpha ($\alpha$).
- By convention alpha is often set at .05. This means that the researcher is willing to accept a 5% chance of falsely rejecting the null hypothesis.
Type II Error

- There is another kind of error that can occur in hypothesis testing:
- Incorrectly retaining the null hypothesis. In other words, failing to find an effect when there actually is one.
- The probability of the Type II error is called beta (β).
- This type of error is not fixed in a typical study. Although if the researcher can pick the N to use, power analysis will allow you to plan ahead so that you can keep this type of error low.
Power

• Power is the probability of correctly rejecting the null hypothesis.
• In other words, power is the probability of correctly detecting an effect.
• Power = $1 - \beta$
• Power can be calculated if you know the size of the effect.
• Power can be estimated if you are willing to hypothesize a value for the effect.
Sampling Distributions

- Knowing the value you want to use for alpha is not enough. You need to translate that into a decision rule that identifies which sample results would qualify as sufficiently unlikely so as to allow you to reject the null hypothesis.
- To do this you need to know the probability distribution for the sample result. This is called the sampling distribution.
A Simple Example

- Suppose I roll a 6-sided die twice, record the number that is facing up each time, and calculate the average of the two values.
  - So if I got a 1 on the first roll and a 4 on the second roll, the mean would be 2.5.
  - This is a simple little experiment with a sample size (N) of 2.
What is the probability that the mean is 1?

- If the die is a fair die, the probability would be $1/36$. 
  $(1/6*1/6)$ which equals .028 or 2.8%.
- This tells us that if we were to repeat this experiment 100 times, we would only expect to get a sample mean of 1 about three times.
## Sampling Distribution

### N=2

<table>
<thead>
<tr>
<th>Sample Mean</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.8%</td>
</tr>
<tr>
<td>1.5</td>
<td>5.6%</td>
</tr>
<tr>
<td>2</td>
<td>8.3%</td>
</tr>
<tr>
<td>2.5</td>
<td>11.1%</td>
</tr>
<tr>
<td>3</td>
<td>13.9%</td>
</tr>
<tr>
<td>3.5</td>
<td>16.7%</td>
</tr>
<tr>
<td>4</td>
<td>13.9%</td>
</tr>
<tr>
<td>4.5</td>
<td>11.1%</td>
</tr>
<tr>
<td>5</td>
<td>8.3%</td>
</tr>
<tr>
<td>5.5</td>
<td>5.6%</td>
</tr>
<tr>
<td>6</td>
<td>2.8%</td>
</tr>
</tbody>
</table>
What if \( N=3? \)

- Suppose I roll a 6-sided die three times, record the number that is facing up each time, and calculate the average of the three values.
  - So if I got a 2 on the first roll, a 6 on the second roll, and a 5 on the third role, the mean would be 4.33.
  - Now what is the probability that the mean is 1?
    - If the die is a fair die, the probability would be \( \frac{1}{216} \).
      \( \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \) which equals 0.005 or 0.5%.
    - This tells us that if we were to repeat this experiment 100 times, we would expect to get a sample mean of 1 either not at all or maybe just once.
## Sampling Distribution

### N=3

<table>
<thead>
<tr>
<th>Sample Mean</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>1.33</td>
<td>1.4%</td>
</tr>
<tr>
<td>1.67</td>
<td>2.8%</td>
</tr>
<tr>
<td>2</td>
<td>4.6%</td>
</tr>
<tr>
<td>2.33</td>
<td>6.9%</td>
</tr>
<tr>
<td>2.67</td>
<td>9.7%</td>
</tr>
<tr>
<td>3</td>
<td>11.6%</td>
</tr>
<tr>
<td>3.33</td>
<td>12.5%</td>
</tr>
<tr>
<td>3.67</td>
<td>12.5%</td>
</tr>
<tr>
<td>4</td>
<td>11.6%</td>
</tr>
<tr>
<td>4.33</td>
<td>9.7%</td>
</tr>
<tr>
<td>4.67</td>
<td>6.9%</td>
</tr>
<tr>
<td>5</td>
<td>4.6%</td>
</tr>
<tr>
<td>5.33</td>
<td>2.8%</td>
</tr>
<tr>
<td>5.67</td>
<td>1.4%</td>
</tr>
<tr>
<td>6</td>
<td>0.5%</td>
</tr>
</tbody>
</table>
What do you notice?

<table>
<thead>
<tr>
<th></th>
<th>N=2</th>
<th>N=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(mean = 1)</td>
<td>2.8%</td>
<td>.5%</td>
</tr>
<tr>
<td>p(mean &gt;= 5)</td>
<td>16.7%</td>
<td>9.3%</td>
</tr>
</tbody>
</table>

- The probability of an extreme result is smaller when the result is based on more information (in this case more rolls of the die)
Experiment

- Say we’re interested in testing the null hypothesis, “This is a fair die” against an alternative hypothesis that “High numbers are more likely on this die than low numbers.”
- We decide to use the mean to test this hypothesis (there are other ways to test the same hypothesis, but for our purpose, this is the way we want to test it).
- Let’s also suppose we go with the conventional Type I error rate of .05 ($\alpha=.05$).
Experiment (cont.)

If the die is fair and all numbers are equally likely then the mean of the distribution would be 3.5 and we would expect the sample mean to be close to 3.5.

What would be our critical value be for rejecting the null hypothesis?

- If N= 3, the sample mean would have to be 5.33 or higher for us to be able to say that the probability of falsely rejecting the null hypothesis is less than .05.
What would the Type II error rate be?

- This depends on the characteristics of the alternative distribution. If high numbers are more likely then the mean will be something higher than 3.5. How much higher?
- Let’s just suppose that if the die is loaded, the true mean is 4.0 (This is based on assuming that 4, 5, & 6 are twice as likely to occur as 1, 2, and 3).
- Assuming the alternative hypothesis is true and the mean really is 4, the probability of getting a sample mean lower than 5.33 when in fact the alternative is true would be .889. In other words 89% of the time when this study was done with a loaded die (specifically with a die that has a mean of 4), the result would be incorrectly retaining the null hypothesis.
Thus what would the power be?

1 minus the probability of a Type II error

1 - .89 = .11

Thus only 11% of the time would we expect to correctly find that the die is loaded.
Experiment (cont.)

- Obviously this is a terrible study, but what can you expect with an N of 3. Luckily we didn’t have much invested in it, it just takes a few seconds to roll a die 3 times.
- But what if the study had been more involved, it cost money, or took a lot of the researcher’s time. Would you really want to invest those resources if there was only an 11 percent chance of actually detecting the effect?
- What if there was a 50% chance? 80%?
- At what point would you consider the probability high enough that you would want to devote your time and resources to the project?
What is Power Analysis?

- Power analysis is a process where a researcher specifies the details of a study and uses those details to either:

  - Determine the power of the study (the probability of finding a statistically significant result assuming that the null hypothesis is false).

  - Determine the sample size that would be needed to obtain a specified level of power.
Statistical Power

In order to determine statistical power, the research must know four things about the study.

- Design
- Alpha (including one- or two-tailed)
- Sample size
- Effect size
What is an effect size?

- There are many different measures of effect size, depending on the design of the study.
- The one that is easiest to understand is probably the correlation coefficient. The correlation coefficient is a measure of the degree of relationship between two variables.
- Another simple one is the mean difference. If your study involves comparing two groups, the difference in the means for the two groups is a measure of the effect size.
  - An even more useful measure of the effect size is the difference between the means divided by the pooled standard deviation (This was first introduced by Cohen in 1969 and is typically referred to as Cohen’s \(d\)).
Cohen’s d

\[ d = \frac{|m_1 - m_2|}{s_{\text{pooled}}} \]

- Thus \( d \) reports the mean difference in standard deviation units. \( d = 1.0 \) means that one group performed a standard deviation above the other group. \( d = 0.2 \) means that one group performed two-tenths of an SD above the other group.
The Big Problem with Power Analysis

- In order for a researcher to do a power analysis, the researcher must pick an effect size on which to base the calculation.
- However, typically the reason a researcher conducts a study is because s/he wants to find out if there is an effect and how big it is.
- So how can you do a power analysis prior to conducting a study, when the power analysis relies on a parameter that you aren’t going to have an estimate of until after you have conducted the study.
What to do?

• Some researchers will base the effect size on a rule of thumb. For Cohen’s \( d \) the rule of thumb goes this way:
  - Small effect \( d = .2 \)
  - Medium effect \( d = .5 \)
  - Large effect \( d = .8 \)

• Cohen advised (as paraphrased by Bausell & Li), when in doubt, hypothesize an effect size of \( .5 \).

• I strongly urge you not to rely too heavily on the rules of thumb. The average effect size over a great many studies may in fact be close to \( .5 \), but only someone with in-depth understanding of what is being studied can decide what a reasonable effect size is for a particular study.
How to pick an effect size?

**Way #1** Conduct a pilot study and estimate the effect size from the pilot data.

- This is an excellent idea – it allows you to work out the kinks in the study, possibly improve the instruments.
- I help a lot of people with their research and I can say that a pilot study is one of the best ways to improve your chances of conducting a successful study. And it gives you a little data on which to calculate an estimate of the effect size.
How to pick an effect size? (cont.)

- **Way #2** Find the effect size in studies that involved similar or related populations and variables to the planned study.
  - This is another thing that should always be done. Good studies are almost always based on a thorough review of the literature.
  - It is getting more and more common for studies to report effect sizes, but even when they don’t, the information needed for you to calculate an estimate of the effect size is generally reported.
How to pick an effect size? (cont.)

**Way #3** Determine what the effect size would need to be for anyone to care about your results.

- Imagine there is a drug to treat a childhood disorder. You suspect that this drug is causing children to not grow as tall as they otherwise would. In this case, you could decide on the effect size to use in the power analysis based on the minimum decrease in height that you think anyone would care about.

- Would people care if kids taking the medication were on average .1 inch shorter than those not taking it? What about if they were a half inch shorter? A full inch? At what point would the finding be something of interest?
How to go about doing a power analysis?

- First of all you need to decide what you want to calculate.
- For any design, there are four quantities that go into the power analysis:
  - Alpha
  - Effect size
  - N
  - Power
- If you specify three of these, you can calculate the fourth.
How To Estimate Power

- You can calculate power by hand for simple designs, and I have students in Intermediate statistics class do that for a very simple test.

- But it is much easier to have a computer do it and there are now a number of programs that will calculate power for a variety of designs.

- I am going to focus on a free program that you can download from the Internet called G*Power.
G*Power

http://www.psycho.uniduesseldorf.de/abteilungen/aap/gpower3

- Covers a variety of designs.
- It’s free
- There is a user’s guide, but it is a little patchy in terms of details.
- The easiest way to find it is to search on G*Power.
Let’s start by using G*Power to calculate the power of my little experiment to decide if the die was fair or loaded.

First we need to calculate the effect size.
- \( d = |m_{H_1} - m_{H_0}|/s = |4 - 3.5|/1.71 = .29 \)

Now, the underlying distributions here are not normal – it is in fact uniform under the null hypothesis – but we will see how close we get if normality is assumed.
Notice, power equals .10 which is remarkably close to the value we calculated. The difference is due to using a t-test rather than an exact test.
If I were to roll the die 10 times, rather than 3, power would increase to .22. Still not very good.

G*Power can also be used to find out what sample size I would need to use to have the power I want to have.
Power = .80 is typically considered a lower bound on what is acceptable. The power analysis indicates that I would need an N of 74.

If I conduct the study with the recommended N of 74, I would have an 80% chance of getting a statistically significant result (assuming of course that the effect size of .29 is accurate).
A More Interesting Example

- We are interested in whether or not a new drug is better for managing chronic back pain than the drug that is traditionally used.
- Suppose there was a special pain rating scale developed for doctors to use to rate the level of pain a patient is experiencing. In this rating scale higher scores mean the patient has more pain.
- Say we know that after a month of drug treatment, people who have the traditional drug have a mean score on the rating scale on 6.2.
Example (cont.)

- Suppose we plan to recruit a random sample of 20 people with chronic back pain and gave them a month’s worth of the new drug. After 1 month on the new drug, doctors will rate each patient on their level of pain.
- We are going to submit a proposal to the National Institutes of Health (NIH) for the money to conduct this study. But they aren’t likely to fund the study unless we can convince them that the study is likely to be successful. Success is defined as correctly determining that the new drug works.
- Thus we need a power analysis.
- Based on a small pilot study, we estimate that the mean on the new drug is 5.2 and the standard deviation is 1.8.
- Putting these values in G*Power, we find that our power is .77.
- Do we want to invest the resources needed for a 77% chance of a significant result? And, even if we are, is NIH going to be willing to give us the money to conduct the study if the chance of success is .77.
Study Design

One sample tests are rarely optimal for many reasons so let’s redesign the study but keep the fundamental scenario.

We are interested in whether or not a new drug is better for managing chronic back pain than the drug that is traditionally used. We have 2 random samples, in one the subjects take the standard medication, in the other the subjects take the new drug. We want to know how many people we should include in the study in order to have a reasonable amount of power.
Two dependent Groups T-test

- Suppose that we match the patients prior to assigning them to the standard or new drug. The two patients with the most severe pain are a pair, the next two in terms of severity of pain are a pair, and so on. For each pair, one subject is randomly assigned the new drug, and the other gets the standard drug.

- Now we are doing a dependent groups t-test. This is similar to an independent groups t-test but we need to take into account the correlation between the scores for the two groups.

- A very important thing to keep in mind is that the effect size is now measured using a different formula. $d_z$ is still a standardized group difference, but now the SD that we are using is the standard deviation of the difference scores. This means that the values that you may have in your head for what is a small, medium, or large effect size, have to be thrown out.
Dependent Groups Effect Size

- The effect size index \( d_z \) is defined as:

\[
d_z = \frac{|\mu_z|}{\sigma_z} = \frac{|\mu_x - \mu_y|}{\sqrt{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy} \cdot \sigma_x \cdot \sigma_y}}
\]

- I haven’t found any rules of thumb for what is a small, medium, or large effect size using \( d_z \), but the values tend to be much larger than effect sizes using Cohen’s \( d \). So, don’t use the Cohen’s \( d \) rules of thumb.
One-way ANOVA

- What if we had 3 independent groups? Standard Drug, New Drug (Dose 1), New Drug (Dose 2).
- For an ANOVA, the effect size used by G*Power is Cohen’s $f$ which is a ratio of the standard deviation among the population means and the standard deviation of the scores within the groups.
Cohen’s $f$

- The effect size $f$ is defined as: $f = \sigma_m / \sigma$. In this equation $\sigma_m$ is the standard deviation of the group means $\mu_i$ and $\sigma$ the common standard deviation within each of the $k$ groups. The total variance is then $\sigma^2_t = \sigma^2_m + \sigma^2$.
- Cohen (1969, p.348) defined the following effect size conventions:
  - small $f = 0.10$
  - medium $f = 0.25$
  - large $f = 0.40$
Effect Sizes

- From [http://www.psycho.uni-duesselddorf.de/aap/projects/gpower/user_manual/user_manual_02.html#effect_size](http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/user_manual/user_manual_02.html#effect_size)

<table>
<thead>
<tr>
<th>Test</th>
<th>Index</th>
<th>small</th>
<th>medium</th>
<th>large</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-test on Means</td>
<td>d</td>
<td>0.20</td>
<td>0.50</td>
<td>0.80</td>
</tr>
<tr>
<td>T-test on Correlations</td>
<td>r</td>
<td>0.10</td>
<td>0.30</td>
<td>0.50</td>
</tr>
<tr>
<td>F-test (ANOVA)</td>
<td>f</td>
<td>0.10</td>
<td>0.25</td>
<td>0.40</td>
</tr>
<tr>
<td>F-test (MCR)</td>
<td>f²</td>
<td>0.02</td>
<td>0.15</td>
<td>0.35</td>
</tr>
<tr>
<td>Chi-square Test</td>
<td>w</td>
<td>0.10</td>
<td>0.30</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Last Notes

- Whenever possible, base your effect size estimate on in-depth knowledge of the construct being studied.
- Rules of thumb are really rough guidelines, don’t rely solely on them.
- Different measures of effect size (e.g., $d$, $f$, $r$) use different metrics. What is a small effect using one measure of effect size might be a large effect using another measure.