NOTES FOR SUMMER STATISTICS INSTITUTE COURSE

COMMON MISTAKES IN STATISTICS –
SPOTTING THEM AND AVOIDING THEM

Day 3: Type I and II Errors and Power

MAY 23 – 26, 2011

Instructor: Martha K. Smith

CONTENTS OF PART 3

I: TYPE I ERROR AND SIGNIFICANCE LEVEL

Type I Error:

Rejecting the *null* hypothesis when it is in fact true is called a ***Type I error***.

Significance level:

Many people decide, before doing a hypothesis test, on a maximum p-value for which they will reject the null hypothesis. This value is often denoted α (alpha) and is also called the ***significance level***.

When a hypothesis test results in a p-value that is less than the significance level, the result of the hypothesis test is called *statistically significant*.

Confusing statistical significance and practical significance is a **common mistake**.

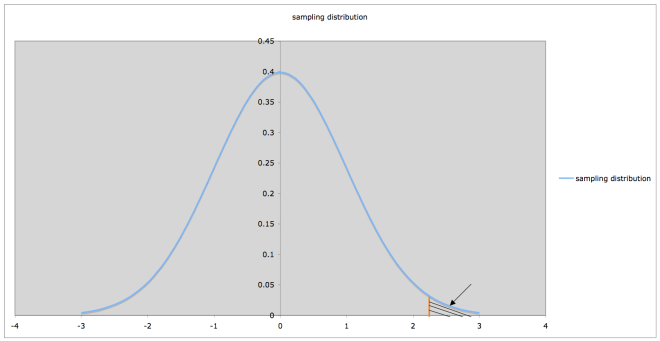
Example: A large clinical trial is carried out to compare a new medical treatment with a standard one. The statistical analysis shows a statistically significant difference in lifespan when using the new treatment compared to the old one.

- However, the increase in lifespan is at most three days, with average increase less than 24 hours, and with poor quality of life during the period of extended life.
- Most people would not consider the improvement practically significant.

Caution: The larger the sample size, the more likely a hypothesis test will detect a small difference. Thus *it is especially important to consider practical significance when sample size is large.*

Connection between Type I error and significance level:

A significance level α corresponds to a certain value of the test statistic, say t_α, represented by the orange line in the picture of a sampling distribution below (the picture illustrates a hypothesis test with alternate hypothesis "$\mu > 0$").



- Since the shaded area indicated by the arrow is the p-value corresponding to t_α, that p-value (shaded area) is α.
- To have p-value less than α, a t-value for this test must be to the right of t_α.
- So the probability of rejecting the null hypothesis when it is true is the probability that $t > t_\alpha$, which we have seen is α.
- In other words, *the probability of Type I error is α.*
- Rephrasing using the definition of Type I error:
  The significance level α is the probability of making the wrong decision when the null hypothesis is true.
- *Note*:
  - α is also called the *bound on Type I error*.
  - Choosing a value α is sometimes called *setting a bound on Type I error*.

Claiming that an alternate hypothesis has been "proved" because it has been rejected in a hypothesis test is a **common mistake** in using statistics. (This is one instance of the mistake of "expecting too much certainty" discussed in Part I.)

- There is always a possibility of a Type I error; the sample in the study might have been one of the small percentage of samples giving an unusually extreme test statistic.
- This is why *replicating experiments* (i.e., repeating the experiment with another sample) is important. The more experiments that give the same result, the stronger the evidence.
- There is also the possibility that the sample is biased or the method of analysis was inappropriate; either of these could lead to a misleading result.

II: PROS AND CONS OF

SETTING A SIGNIFICANCE LEVEL

- Setting a significance level (*before* doing inference) has the *advantage* that the analyst is not tempted to chose a cut-off on the basis of what he or she hopes is true.
- It has the *disadvantage* that it neglects that some p-values might best be considered borderline.
    - *This is one reason why it is important to report p-values when reporting results of hypothesis tests. It is also good practice to include confidence intervals corresponding to the hypothesis test.*
    - For example, if a hypothesis test for the difference of two means is performed, *also* give a confidence interval for the difference of those means.
    - If the significance level for the hypothesis test is .05, then use confidence level 95% for the confidence interval.

III.TYPE II ERROR

Not rejecting the *null* hypothesis when in fact the *alternate* hypothesis is true is called a ***Type II error***.

- Example 2 below provides a situation where the concept of Type II error is important.

Note: "The alternate hypothesis" in the definition of Type II error may refer to the alternate hypothesis in a hypothesis test, or it may refer to a "specific" alternate hypothesis.

Example: In a t-test for a sample mean μ, with null hypothesis "μ = 0" and alternate hypothesis "μ > 0":

- We might talk about the Type II error relative to the *general alternate hypothesis "μ > 0"*.
- Or we might talk about the Type II error relative to the *specific alternate hypothesis "μ = 1"*.
- Note that *the specific alternate hypothesis is a special case of the general alternate hypothesis*.

In practice, people often work with Type II error relative to a *specific* alternate hypothesis.

- In this situation, the probability of Type II error relative to the specific alternate hypothesis is often called β.
- In other words, β is the probability of making the *wrong* decision when the *specific alternate* hypothesis is *true*.
- The specific alternative is considered since it is more feasible to calculate β than the probability of Type II error relative to the general alternative.
- See the discussion of power below for related detail.

IV: CONSIDERING BOTH TYPES OF ERROR TOGETHER

The following table summarizes Type I and Type II errors:

		Truth (for population studied)	
		Null Hypothesis True	Null Hypothesis False
Decision (based on sample)	Reject Null Hypothesis	*Type I Error*	*Correct Decision*
	Don't reject Null Hypothesis	*Correct Decision*	*Type II Error*

An analogy that can be helpful in understanding the two types of error is to consider a defendant in a trial.

- The null hypothesis is "defendant is not guilty."
- The alternate is "defendant is guilty."
- A Type I error would correspond to convicting an innocent person.
- Type II error would correspond to setting a guilty person free.
- The analogous table would be:

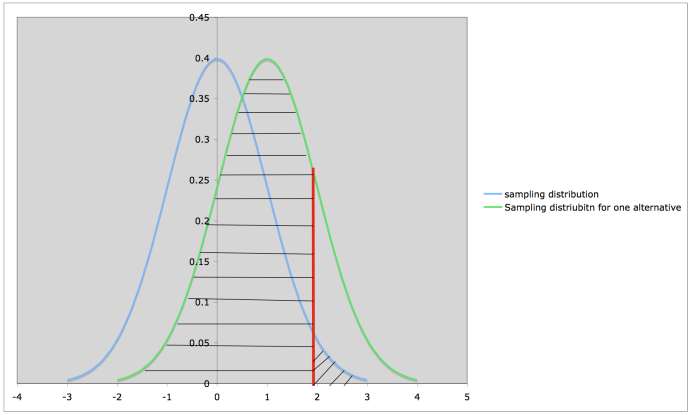
		Truth	
		Not Guilty	Guilty
Verdict	Guilty	*Type I Error* -- Innocent person goes to jail (and maybe guilty person goes free)	*Correct Decision*
	Not Guilty	*Correct Decision*	*Type II Error* -- Guilty person goes free

Note:

- This could be more than just an analogy if the verdict hinges on statistical evidence (e.g., a DNA test), and where rejecting the null hypothesis would result in a verdict of guilty, and not rejecting the null hypothesis would result in a verdict of not guilty.
- This analogy/example shows that *sometimes* a Type I error can be more serious than a Type II error. (However, this is *not always* the case).

The following diagram illustrates the Type I error and the Type II error

- against the specific alternate hypothesis "$\mu = 1$"
- in a hypothesis test for a population mean μ,
- with null hypothesis "$\mu = 0$,"
- alternate hypothesis "$\mu > 0$",
- and significance level $\alpha = 0.05$.



In the diagram,

- The blue (leftmost) curve is the *sampling* distribution of the test statistic assuming the null hypothesis "$\mu = 0$."
- The green (rightmost) curve is the *sampling* distribution of the test statistic assuming the specific alternate hypothesis "$\mu = 1$".
- The vertical red line shows the cut-off for rejection of the null hypothesis:
  - The null hypothesis is rejected for values of the test statistic to the *right* of the red line (and *not* rejected for values to the *left* of the red line).
- The area of the diagonally hatched region to the *right* of the red line and under the *blue* curve is the probability of type I error (α).
- The area of the horizontally hatched region to the *left* of the red line and under the *green* curve is the probability, β, of Type II error against the specific alternative.

V. DECIDING WHAT SIGNIFICANCE LEVEL TO USE

This should be done *before analyzing* the data -- *preferably before gathering the data*. There are (at least) two reasons why this is important:

1) The significance level desired is one criterion in deciding on an appropriate sample size.
   - See discussion of Power below.

2) If more than one hypothesis test is planned, additional considerations need to be taken into account.
   - More tomorrow

The choice of significance level should be based on the consequences of Type I and Type II errors:

1. If the *consequences of a Type I error are serious or expensive*, a very *small* significance level is appropriate.

   Example 1: Two drugs are being compared for effectiveness in treating the same condition.
   - Drug 1 is very affordable, but Drug 2 is extremely expensive.
   - The null hypothesis is “both drugs are equally effective.”
   - The alternate is “Drug 2 is more effective than Drug 1.”
   - In this situation, a Type I error would be deciding that Drug 2 is more effective, when in fact it is no better than Drug 1, but would cost the patient much more money.
   - That would be undesirable from the patient’s perspective, so a *small* significance level is warranted.

2. If the consequences of a Type I error are not very serious (and especially *if a Type II error has serious consequences*), then a *larger* significance level is appropriate.

Example 2: Two drugs are known to be equally effective for a certain condition.

- They are also each equally affordable.
- However, there is some suspicion that Drug 2 causes a serious side effect in some patients, whereas Drug 1 has been used for decades with no reports of serious side effects.
- The null hypothesis is "the incidence of serious side effects in both drugs is the same".
- The alternate is "the incidence of serious side effects in Drug 2 is greater than that in Drug 1."
- Falsely rejecting the null hypothesis when it is in fact true (Type I error) would have no great consequences for the consumer.
- But a Type II error (i.e., failing to reject the null hypothesis when in fact the alternate is true, which would result in deciding that Drug 2 is no more harmful than Drug 1 when it is in fact more harmful) could have serious consequences from a consumer and public health standpoint.
- So setting a large significance level is appropriate.

Comments:

- Neglecting to think adequately about possible consequences of Type I and Type II errors (and deciding acceptable levels of Type I and II errors based on these consequences) *before* conducting a study and analyzing data is a **common mistake** in using statistics.
- Sometimes there may be serious consequences of each alternative, so some compromises or weighing priorities may be necessary.
  - The trial analogy illustrates this well: Which is better or worse, imprisoning an innocent person or letting a guilty person go free?
  - *This is a value judgment; value judgments are often involved in deciding on significance levels.*
  - *Trying to avoid the issue by always choosing the same significance level is itself a value judgment.*
- Different people may decide on different standards of evidence.
  - This is another reason why *it is important to report p-values even if you set a significance level.*
  - It is *not* enough just to say, “significant at the .05 level,” “significant at the .01 level,” etc.
- Sometimes different stakeholders have different interests that compete (e.g., in the second example above, the developers of Drug 2 might prefer to have a smaller significance level.)
  - This is another reason why it is important to report p-values in publications.
- See Wuensch (1994) for more discussion of considerations involved in deciding on reasonable levels for Type I and Type II errors.
- See also the discussion of Power below.
- Similar considerations hold for setting confidence levels for confidence intervals

VI: POWER OF A STATISTICAL PROCEDURE

Overview

The ***power*** of a statistical procedure can be thought of as *the probability that the procedure will detect a true difference of a specified type.*

- As in talking about p-values and confidence levels, the reference category for "probability" is the sample.
- Thus, power is the probability that a randomly chosen sample
    - satisfying the model assumptions
    - will give evidence of a difference of the specified type when the procedure is applied,
    - if the specified difference does indeed occur in the population being studied.
- Note also that power is a conditional probability: the probability of detecting a difference, *if* indeed the difference does exist.

In many real-life situations, there are reasonable conditions that we would be interested in being able to detect, and others that would not make a practical difference.

Examples:

- If you can only measure the response to within 0.1 units, it doesn't really make sense to worry about falsely rejecting a null hypothesis for a mean when the actual value of the mean is within less than 0.1 units of the value specified in the null hypothesis.
- Some differences are of no practical importance -- for example, a medical treatment that extends life by 10 minutes is probably not worth it.

In cases such as these, neglecting power could result in one or more of the following:

- Doing much more work or going to more expense than necessary
- Obtaining results which are meaningless
- Obtaining results that don't answer the question of interest.

Elaboration

For many *confidence interval procedures*, power can be defined as:

> The probability (again, the reference category is "samples") that the procedure will produce an interval with a half-width of at least a specified amount.

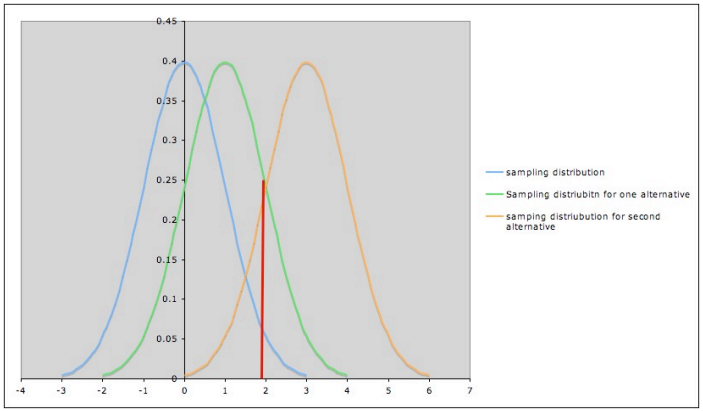
For a *hypothesis test*, power can be defined as:

> The probability (again, the reference category is "samples") of rejecting the null hypothesis under a specified condition.
>
> *Example*: For a one-sample t-test for the mean of a population, with null hypothesis Ho: $\mu = 100$, you might be interested in the probability of rejecting Ho when $\mu \geq 105$, or when $|\mu - 100| > 5$, etc.
>
> As with Type II Error, we may think of power for a hypothesis test in terms of *power against a specific alternative* rather than against a general alternative.

Example: If we are performing a hypothesis test for the mean of a population, with null hypothesis H_0: $\mu = 0$ and alternate hypothesis $\mu > 0$, we might calculate the power of the test *against the specific alternative* H_1: $\mu = 1$, or *against the specific alternative* H_3: $\mu = 3$, etc.
The picture below shows three sampling distributions:

- The sampling distribution assuming H_0 (*blue; leftmost* curve)
- The sampling distribution assuming H_1 (*green; middle* curve)
- The sampling distribution assuming H_3 (*yellow; rightmost* curve)

The red line marks the cut-off corresponding to a significance level $\alpha = 0.05$.



- Thus the area under the *blue* curve to the *right of the red line* is 0.05.
- The area under the *green* curve the to *right of the red line* is the probability of rejecting the null hypothesis ($\mu = 0$) if the specific alternative H_1: $\mu = 1$ is true.
  - In other words, this area is *the power of the test against the specific alternative H_1: $\mu = 1$*.
  - We can see in the picture that in this case, this power is greater than 0.05, but noticeably less than 0.50.
- Similarly, the area under the *yellow* curve the to *right of the red line* is *the power of the test against the specific alternative H_3: $\mu = 3$*.
  - Notice that the power in this case is much larger than 0.5.

This illustrates the general phenomenon that *the farther an alternative is from the null hypothesis, the higher the power of the test to detect it.* (See Claremont Graduate University WISE Project Statistical Power Demo for an interactive illustration.)

Note:

- For most tests, it *is* possible to calculate the power against a specific alternative, at least to a reasonable approximation. (More below and in Appendix)
- It is *not* usually possible to calculate the power against a general alternative, since the general alternative is made up of infinitely many possible specific alternatives.

Power and Type II Error

Recall: The Type II Error rate β of a test against a specific alternate hypothesis test is represented in the diagram above as the area under the sampling distribution curve for that alternate hypothesis and to the *left* of the cut-off line for the test. Thus

β + (Power of a test against a specific alternate hypothesis)

= total area under sampling distribution curve

= 1,

so

$$Power = 1 - \beta$$

Factors that Affect the Power of a Statistical Procedure

Power depends on several factors in addition to the difference to be detected.

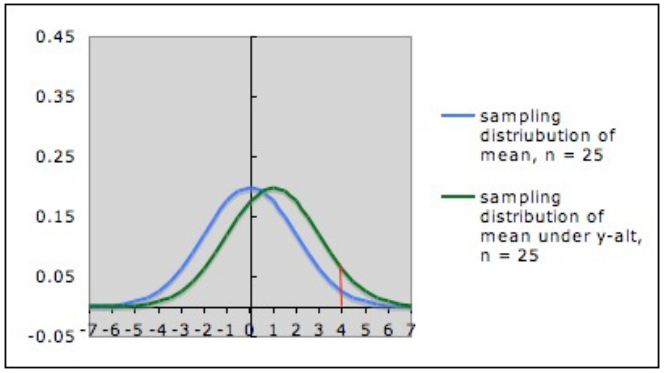
1. *Significance Level*

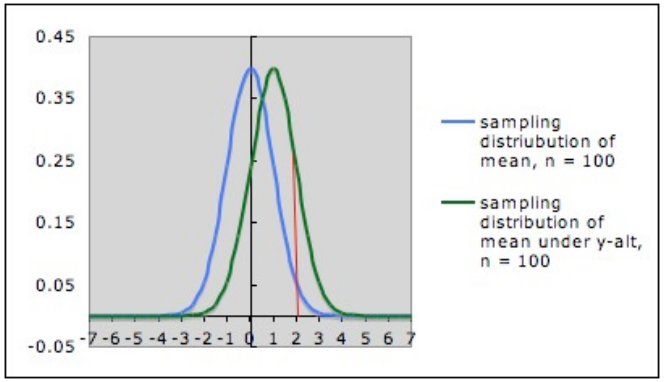
This can be seen in the diagram illustrating power: *Increasing* the significance level α will move the red line to the *left*, and hence will *increase power*. Similarly, decreasing significance level decreases power.

2. Sample Size

Example: The pictures below each show the sampling distribution for the mean under the null hypothesis $\mu = 0$ (blue -- on the left in each picture) together with the sampling distribution under the alternate hypothesis $\mu = 1$ (green -- on the right in each picture), but *for different sample sizes*.

- The first picture is for sample size n = 25; the second picture is for sample size n = 100.
- Note that both graphs are in the same scale. In both pictures, the blue curve is centered at 0 (corresponding to the the null hypothesis) and the green curve is centered at 1 (corresponding to the alternate hypothesis).
- In each picture, the red line is the cut-off for rejection with alpha = 0.05 (for a one-tailed test) -- that is, in each picture, the area under the *blue* curve to the right of the red line is 0.05.
- In each picture, the area under the *green* curve to the right of the red line is the power of the test against the alternate depicted. Note that this area is *larger* in the second picture (the one with larger sample size) than in the first picture.





This illustrates the general situation:

Larger sample size gives larger power.

The reason is essentially the same as in the example: Larger sample size gives a narrower sampling distribution, which means there is less overlap in the two sampling distributions (for null and alternate hypotheses).

See Claremont University's Wise Project's Statistical Power Applet (http://wise.cgu.edu/powermod/power_applet.asp) for an interactive demonstration of the interplay between sample size and power for a one-sample z-test.

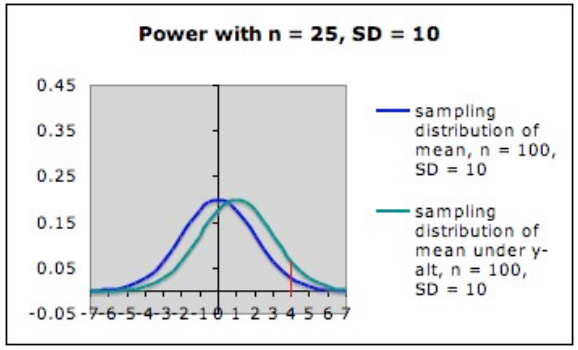
Note: Sample size needed typically increases at an increasing rate as power increases. (e.g., in the above example, increasing the sample size by a factor of 4 increases the power by a factor of about 2; the graphics aren't accurate enough to show this well.)

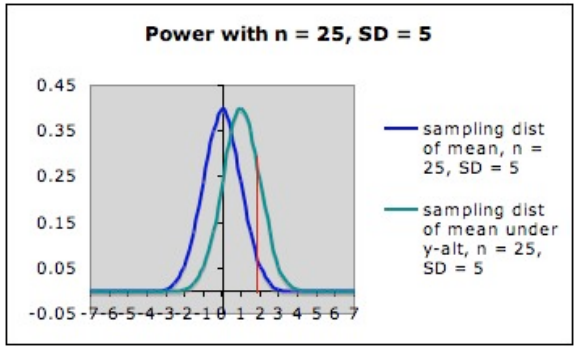
3. *Variance*

Power also depends on variance: *smaller variance yields higher power*.

Example: The pictures below each show the sampling distribution for the mean under the null hypothesis $\mu = 0$ (blue -- on the left in each picture) together with the sampling distribution under the alternate hypothesis $\mu = 1$ (green -- on the right in each picture), both with sample size 25, but *for different standard deviations of the underlying distributions*. (Different standard deviations might arise from using two different measuring instruments, or from considering two different populations.)

- In the first picture, the standard deviation is 10; in the second picture, it is 5.
- Note that both graphs are in the same scale. In both pictures, the blue curve is centered at 0 (corresponding to the the null hypothesis) and the green curve is centered at 1 (corresponding to the alternate hypothesis).
- In each picture, the red line is the cut-off for rejection with alpha = 0.05 (for a one-tailed test) -- that is, in each picture, the area under the *blue* curve to the right of the red line is 0.05.
- In each picture, the area under the *green* curve to the right of the red line is the power of the test against the alternate depicted. Note that this area is *larger* in the second picture (the one with smaller standard deviation) than in the first picture.





(See Claremont University's Wise Project's Statistical Power Applet at http://wise.cgu.edu/powermod/power_applet.asp or the Rice Virtual Lab in Statistics' Robustness Simulation at http://onlinestatbook.com/stat_sim/robustness/index.html for an interactive demonstration.)

Note: Variance can sometimes be reduced by using a better measuring instrument, restricting to a subpopulation, or by choosing a better experimental design (see below).

4. *Experimental Design*

Power can sometimes be increased by adopting a different experimental design that has lower error variance. For example, stratified sampling or blocking can often reduce error variance and hence increase power. However,

- The power calculation will depend on the experimental design.
- The statistical analysis will depend on the experimental design. (To be discussed tomorrow.)
- For more on designs that may increase power, see Lipsey (1990) or McClelland (2000)

***Calculating Sample Size to Give Desired Power*:** The dependence of power on sample size allows us, *in principle*, to figure out beforehand what sample size is needed to detect a specified difference, with a specified power, at a given significance level, if that difference is really there.

In practice, details on figuring out sample size will vary from procedure to procedure. See the Appendix for discussion of some of the considerations involved.

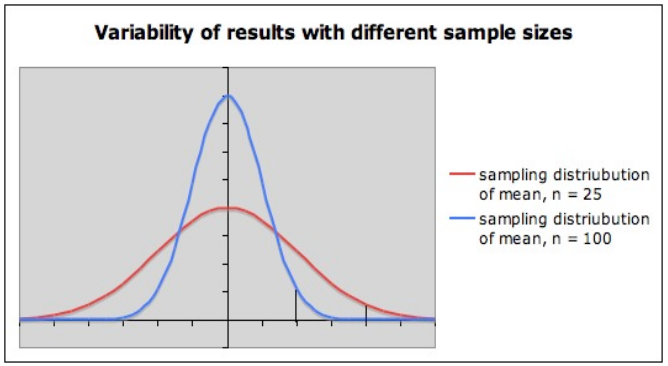
Detrimental Effects of Underpowered or Overpowered Studies

The most straightforward consequence of ***underpowered*** studies (i.e., those with low probability of detecting an effect of practical importance) is that effects of practical importance are not detected.

But there is a second, more subtle consequence: ***underpowered*** studies result in a larger variance of the estimates of the parameter being estimated. For example, in estimating a population mean, the sample means of studies with low power have high variance; in other words, *the sampling distribution of sample means is wide.*

This is illustrated in the following picture, which shows the sampling distributions for a variable with zero mean when sample size n = 25 (red) and when n = 100 (blue). The vertical lines toward the right of each sampling distribution show the cut-off for a one-sided hypothesis test with null hypothesis $\mu = 0$ and significance level alpha = .05. Notice that

- The sampling distribution for the *smaller* sample size (n = 25) is *wider* than the sampling distribution for the larger sample size (n = 100).
- Thus, *when the null hypothesis is rejected with the smaller sample size n = 25, the sample mean tends to be noticeably larger than when the null hypothesis is rejected with the larger sample size n = 100.*



This reflects the general phenomenon that *studies with low power have a larger chance of having a large sample mean than studies with high power.*

In particular, *when there is a Type I error (falsely rejecting the null hypothesis), the effect will appear to be stronger with a large sample size (higher power) than with a small sample size (higher power). This may suggest an exaggerated effect, or even one that is not there*. Thus, *when studies are underpowered, the literature is likely to be inconsistent and often misleading.*

- This problem is increased because of the "File Drawer Problem" (to be discussed tomorrow).

Overpowered studies waste resources.

- When human or animal subjects are involved, having an overpowered study can be considered unethical.
    - For more on ethical considerations in animal studies, see Festing (2010) or Kilkenny et al (2010)
- More generally, an overpowered study may be considered unethical if it wastes resources.

A common compromise between over-power and under-power is to try for power around .80. *However, power needs to be considered case by case, balancing the risks of Type I and Type II errors.*

IV: COMMON MISTAKES INVOLVING POWER

1. ***Rejecting a null hypothesis without considering practical significance.***

A study with large enough sample size will have high enough power to detect minuscule differences that are not of practical significance. Since power typically increases with increasing sample size, practical significance is important to consider.

2. ***Accepting a null hypothesis when a result is not statistically significant, <u>without taking power into account</u>.***

- Since power typically increases with increasing sample size, practical significance is important to consider.
- Looking at this from the other direction: Power decreases with decreasing sample size.
- Thus *a small sample size may not be able to detect an important difference*.
- If there is strong evidence that the power of a procedure will indeed detect a difference of practical importance, then accepting the null hypothesis is appropriate.
    - However, it may be better to use a *test for equivalence*; see the Appendix for references.
- Otherwise “accepting the null hypothesis” is *not appropriate* -- all we can legitimately say then is that we fail to reject the null hypothesis.

3. ***Being convinced by a research study with low power.***

As discussed above, *underpowered studies are likely to be inconsistent and are often misleading*.

4. ***Neglecting to do a power analysis/sample size calculation <u>before</u> collecting data***

- Without a power analysis, you may end up with a result that does not really answer the question of interest.
- You might obtain a result that is not statistically significant, but is not able to detect a difference of practical significance.
- You might also waste resources by using a sample size that is larger than is needed to detect a relevant difference.

5. ***Neglecting to take multiple inference into account when calculating power.***

If more than one inference procedure is used for a data set, then power calculations need to take that into account. Doing a power calculation for just one inference will result in an underpowered study. (*More on this tomorrow*)

- For more detail, see Maxwell and Kelley (2011) and Maxwell (2004)

6. ***Using standardized effect sizes rather than considering the particulars of the question being studied.***

"Standardized effect sizes" (sometimes called "canned" effect sizes) are expressions involving more than one of the factors that needs to be taken into consideration in considering appropriate levels of Type I and Type II error in deciding on power and sample size. *Examples*:

- Cohen's effect size d is the ratio of the raw effect size (e.g., difference in means when comparing two groups) and the error standard deviation. But each of these typically needs to be considered individually in designing a study and determining power; it's not necessarily the ratio that's important. (See Appendix)
- The correlation (or squared correlation) in regression. The correlation in simple linear regression involves three quantities: the slope, the y standard deviation, and the x standard deviation. Each of these three typically needs to be considered individually in designing the study and determining power and sample size. In multiple regression, the situation may be even more complex.

For specific examples illustrating these points, see Lenth, (2000) and (2001)

7. ***Confusing <u>retrospective</u> power and <u>prospective</u> power***.

- Power as defined above for a hypothesis test is also called *prospectiv*e or *a priori* power.
  - It is a conditional probability, P(reject H_0 | H_a), *calculated without using the data to be analyzed*.
  - In fact, it is best calculated before even gathering the data, and taken into account in the data-gathering plan.
- *Retrospective* power is calculated *after* the data have been collected, *using the data*.
- Depending on how retrospective power is calculated, it might be legitimate to use to estimate the power and sample size for a *future* study, but <u>cannot</u> legitimately be used as describing the power of the study from which it is calculated.
- However, some methods of calculating retrospective power calculate the power to detect the effect observed in the data -- which misses the whole point of considering practical significance. These methods typically yield simply a transformation of p-value. See Lenth (2000 for more detail.
- See Hoenig and Heisley (2001) and Wuensch et al (2003) for more discussion and further references.