NOTES FOR SUMMER STATISTICS INSTITUTE COURSE

COMMON MISTAKES IN STATISTICS – SPOTTING THEM AND AVOIDING THEM

Day 3: Type I and II Errors, Power, and Multiple Inference

MAY 26 - 29, 2015

Instructor: Martha K. Smith

CONTENTS OF DAY 3

(If time permits, we will start on some of the material listed for Day 4)

II. TYPE II ERROR

(Recall: Falsely rejecting a true null hypothesis is called a ***Type I 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.

New complication: "The alternate hypothesis" in the definition of Type II error may refer to the alternate hypothesis in a hypothesis test (a "general" alternate hypothesis), or it may refer to a "specific" alternate hypothesis.

Example/Elaboration: 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 <u>general</u> *alternate hypothesis "μ > 0".*
- Or we might talk about the Type II error relative to the <u>specific</u> *alternate hypothesis "μ = 1" (or " μ = 0.5", or ...).*
- 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 for two reasons:
  1. It's more feasible to calculate β than the probability of Type II error relative to the general alternative.
  2. What's usually important is the ability to detect a difference of practical importance, rather than any difference however minute.
- See the discussion of power below for related detail.

III: 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*

Analogy: Results of a criminal trial.

- The “null hypothesis” is "defendant is not guilty."
- The “alternate hypothesis” 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:

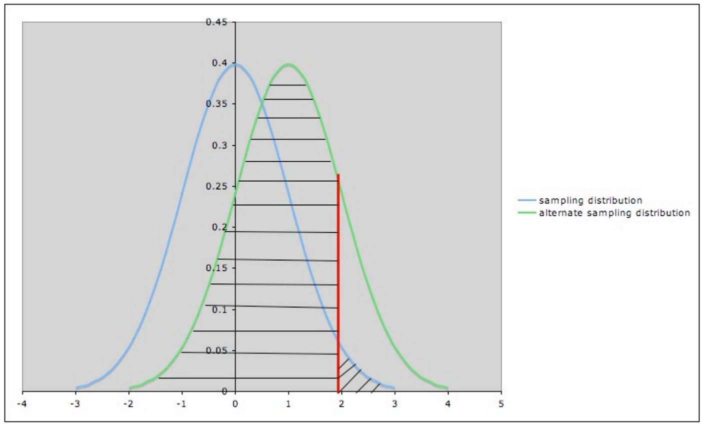
		Truth	
		Not Guilty	Guilty
Verdict	Guilty	*Type I Error --* Innocent person goes to jail (and maybe guilty person goes free)	*Correct Decision*
	Not Guilty	*Correct Decision*	*Type II Error --* Guilty person goes free

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 *both* the Type I error *and* the Type II error

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



In the diagram,

- The blue (leftmost) curve is the *sampling* distribution of the test statistic *assuming the null hypothesis* "μ = 0."
- The green (rightmost) curve is the *sampling* distribution of the test statistic *assuming the specific alternate hypothesis* "μ =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.

→ What happens to the Type II error probability (β) if we:

a. increase α?

b. decrease α?

IV. DECIDING WHAT SIGNIFICANCE LEVEL TO USE

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

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.
    - See discussion of Multiple Inference below.

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’re 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.

Example 3: Some vaccines are made from weakened strains of the pathogen causing the disease in question.

- In these cases, each batch of the vaccine needs to be tested for virulence (that is, the virus needs to be tested to be sure it is weakened enough that it does not cause the disease, or only causes a case that is minor but still results in immunity).
- The null hypothesis would be "the vaccine does not produce serious disease."
- The alternate hypothesis would be "the vaccine does produce serious disease"
- A type II error here would have serious consequences,.
- Thus it is important to have a high Type II error rate for such tests.
  - Indeed, in these cases, the Type II error rate is often set at 99%, whereas in much research, a Type II error rate of 80% is considered acceptable.

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 are serious consequences of each alternative, so 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's important to report p-values even if you set a significance level.*
  - It's *not* enough just to say, "significant at the .05 level," "significant at the .01 level," etc. Unfortunately, reporting p-values this way is a **very common mistake**.

- 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's 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.

V: POWER OF A STATISTICAL PROCEDURE

Overview

The ***power*** of a hypothesis test can be thought of as *the probability that the test 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 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'd like to be 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 testing for vaccine virulence, it is very important to be able to detect virulence, so high power is especially important.

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

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

Elaboration

The *power* of a hypothesis test is 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 H_0: $\mu = 100$, you might be interested in the probability of rejecting H_0 when $\mu \geq 105$, or when $|\mu - 100| > 5$, etc.

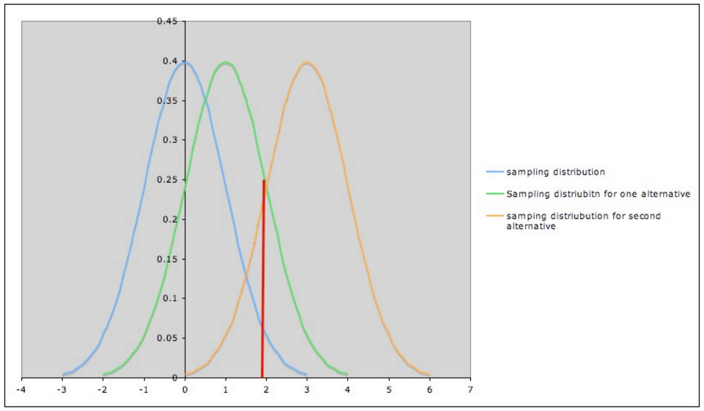
As with Type II Error, we may think of power for a hypothesis test in terms of *power against a specific alternative* rather than against a general alternative.

Example: If we're 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 for our test statistic:

- 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 $\alpha = 0.05$.



→Where would we reject/not reject the null hypothesis?

From the above, we conclude (*how?*) that:

- 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, the 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's *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 (cf p. 7). Thus

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

= total area under sampling distribution curve

= 1,

so

Power = 1 - β

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*

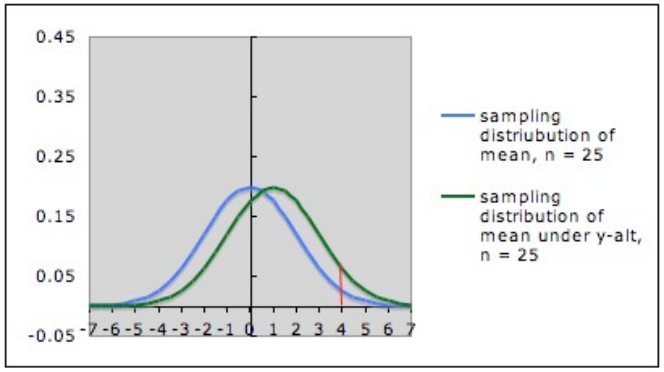
This can be seen in the diagram illustrating power:

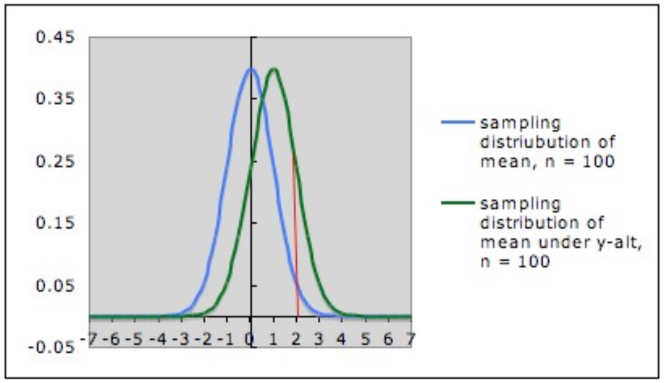
- *Increasing* the significance level α will move the red line to the *left*, and hence will *increase power*.
- Similarly, decreasing significance level decreases power.

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.
  - Why are the curves in the second graph skinnier?
- Note that both graphs are in the same scale. In both pictures, the blue curve is centered at 0 (corresponding to the null hypothesis) and the green curve is centered at 1 (corresponding to the alternate hypothesis).
- In each picture, the vertical red/orange 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 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.

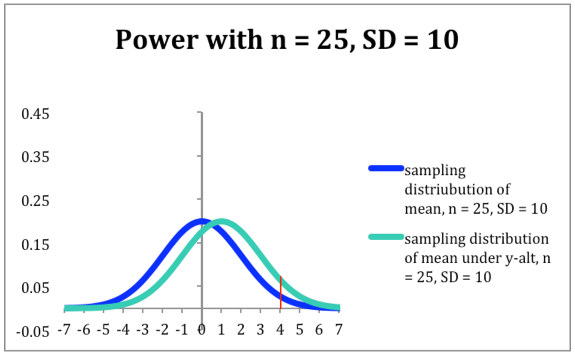
Note: Sample size needed to give desired power typically increases at an increasing rate as power increases. (e.g., in the above example, increasing the sample size by a factor of 4 increases the power by a factor of about 2; the graphics aren't accurate enough to show this well.)

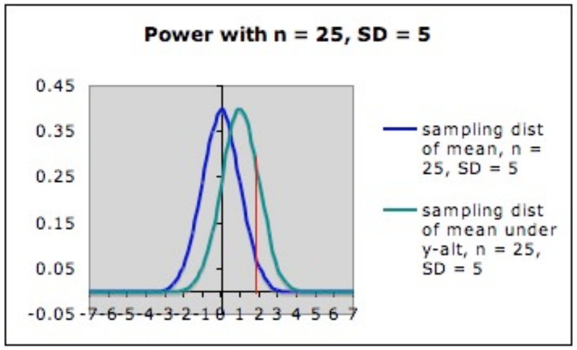
3. *Variance*

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/orange 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 line is 0.05.
- In each picture, the area under the *green* curve to the right of the 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. [Try 1) mean 0, st dev. 1; and 2) mean 1, st deviations 1 and 5]

Note: Variance can sometimes be reduced by using a better measuring instrument, by 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 *in principle* lets us 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.

- *In particular:* Power calculations need to take into account the specifics of the statistical procedure.

  - For example, there are many F-tests; they involve different calculations of the F-statistic, and thus require different power and sample size calculations.

  - In particular, there are many types of ANOVA; the test statistic depends on the experimental design, so power calculation depends on the experimental design.

Detrimental Effects of Underpowered or Overpowered Studies

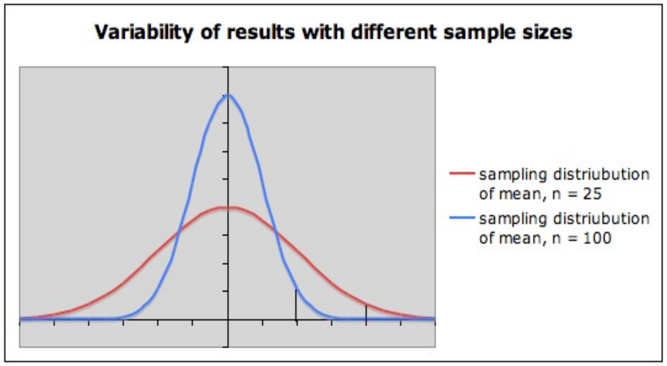
The most straightforward consequence of ***underpowered*** studies (i.e., those with low probability of detecting an effect of practical importance) is that *effects of practical importance are not detected*.

But there's another, more subtle, but important consequence:

> ***Underpowered*** studies result in *a larger variance of the estimates of the parameter being estimated.*
> - For example, in estimating a population mean, the sampling distributions of sample means in 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:

- It shows the sampling distributions of the mean for a variable with zero mean when sample size n = 25 (red/orange) 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 µ = 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 exhibiting a large effect than studies with high power*.

- *This may suggest an exaggerated effect, or even one that is not there*.

In particular, *when there is a Type I error (falsely rejecting the null hypothesis), the effect will appear to be stronger with low power than with a high power.*

- This phenomenon is sometimes called *"The winner's curse,"* or *"The Proteus phenomenon,"* or "*The statistical significance filter."*
- 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" (discussed below).
- Recall that low power may result from:
  - Small sample size
  - Small alpha level
  - Large variance

In response to the winner's curse and other concerns, the terms *Type M error* and *Type S error* have recently been introduced as refinements of (or better alternatives to) the notion of Type I error:

- A *Type M error* occurs when the effect size estimate differs in size (magnitude) from the true value of the effect being estimated (as shown in the above diagram, assuming the vertical axis shown is at effect = 0)
- A *Type S error* occurs when the effect size estimate has a different sign than the true effect size.
  - This could be illustrated by a figure similar to the one above, but with vertical axis between the two short vertical lines.
- Example: Gelman and Weakliem (2009) responded to a claim by S. Kanazawa that "Beautiful parents have more daughters," by locating several more data sets appropriate for "testing" this claim.
  - In most of these, the proportion of girls born to beautiful people was less than 50%, suggesting that Kanazawa had a Type S error.
- For an alternative to power based on Type S and Type M errors, see Gelman and Carlin (2014).
  - This perspective has the advantage that it can be used either prospectively (to design a study) or retrospectively (to analyze an existing study).

- Recall from Day 2: Replicating studies is important because of the possibility of Type I error.
  - *The possibility of Type S and Type M errors makes this even more important*.
  - See Lehrer (2010) for a popular press article on this.
  - See Ioannidis (2014) for ideas on how to encourage replication and other practices that will improve the overall quality of research results.
  - For discussion of some recent efforts to promote replication studies, see Baker (2015) and the links and references therein.

Overpowered studies waste resources.

- When human or animal subjects are involved, an overpowered study can be considered unethical.
    - For more on ethical considerations in animal studies, see Festing (2010), Kilkenny et al (2010), or *Nature* Editors (2015)
- More generally, an overpowered study may be considered unethical if it wastes resources.

A common practice is to 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.*
- For example, in animal experiments, the *percentage* of animals wasted decreases as sample size increases, so performing many underpowered studies may waste more animals than carrying out one higher-powered study. (Currie, undated)

VI: 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 aren't practically significant.
- 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>.***

- Power decreases with decreasing sample size.
- Thus *a small sample size may not be able to detect an difference that is important*.
- *If* there's strong evidence that the power of a procedure will indeed detect a difference of practical importance, then accepting the null hypothesis *might* be appropriate.
    - However, it may be better to use a *test for equivalence*; see 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*.

- If the author of a study hasn't mentioned power, be skeptical.

- If the study has mentioned power, look carefully to see whether the power was calculated appropriately. (See items 4 - 7 below.)

- Remember the following quotes from Andrew Gelman's blog on the winner's curse (http://andrewgelman.com/2010/10/02/the_winners_cur/):
    - "If an estimate is statistically significant, it's probably an overestimate of the magnitude of your effect." (Andrew Gelman)
    - "Large estimates often do not mean 'Wow, I've found something big!' but, rather, 'Wow, this study is underpowered!' (Jerzy Wieczorek)

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

- If you use a sample size that's *too small* to detect a difference of practical significance, you may get a result that's not statistically significant even though there is a difference of practical significance, or you may obtain a result that misleadingly suggests significance.
    - Thus *you've expended considerable effort to obtain a result that doesn't really answer the question of interest*.

- If you use a sample size that's *larger than needed* to detect a relevant difference, you've also wasted resources.

- In addition to (or instead of) standard power calculations, do a "design analysis" as described by Gelman and Carlin (2014) to take into account Type M and Type S errors.
    - Even with a standard power analysis, it may be wise to base sample size calculations on a hypothesized effect size that is determined as discussed in Gelman and Carlin.

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. (*More on this below*.)

- 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. ***Calculating power using "standardized effect sizes" rather than considering the particulars of the question being studied.***

"Standardized effect sizes" (examples below) 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.

- Standardized effect sizes are important in meta-analysis, when considering studies that may use different measures that are on different scales.
- However, *in calculating power or sample size for a particular study, you're losing information if you use standardized effect sizes rather than entering their components into the calculation individually*.

Examples:

i. Cohen's effect size d is the ratio of the raw effect size (e.g., difference in means when comparing two groups) and a suitable 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)

ii. 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 *<u>prospectiv</u>*<u>e</u> or *a priori* power.

  It's a conditional probability, P(reject H_0 | H_a), *<u>calculated without using the data</u> to be analyzed*.

  In fact, *it's best calculated before even gathering the data, and taken into account in the data-gathering plan*.

- *<u>Retrospective</u>* power is calculated *after* the data have been collected, *using the data*.

  Depending on how retrospective power is calculated, it might (or might not) 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.

  Moreover, 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.*

- However, the "design calculations" recommended by Gelman and Carlin (2014) considering Type M and Type S errors *can* be done retrospectively.

8. ***Using the same sample size as a previous study to calculate power for a replication.***

Assume we do a new study with the same sample size as the previous study. What can we say about the power of the new study?

The picture below shows:

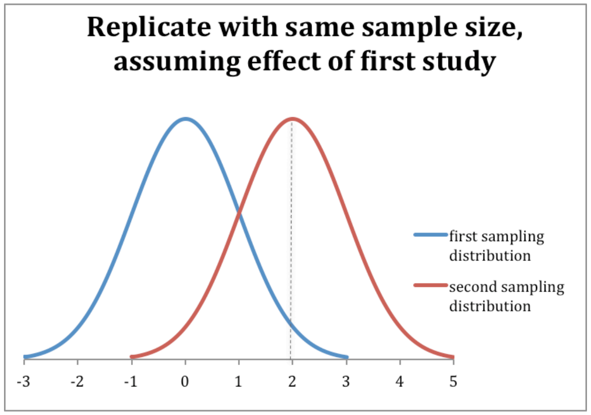
1. The sampling distribution for the previous study (blue)

2. The test statistic obtained from the sample used in the previous study (dashed line)

   Note: Since the second study has the same size as the first, then the *blue* curve is *also* the sampling distribution for the *second* study (assuming the null hypothesis of no effect)

3. The sampling distribution for the second study, assuming the specific alternate "effect is the estimate from the first study" (red)

Note: The red sampling distribution is centered at the test statistic from the previous study, because that is the test statistic resulting from that effect.



The graph shows that the power of the second study against the specific alternate hypothesis that the effect the estimate found in the previous study is about _______,

because ____.

In fact, an even smaller sample size could give a non-significant result, if the first hypothesis test happened to be a Type M error.

Real Example:

Psychologists Brian Nosek and Matt Motyl obtained a statistically significant (p = 0.01) result with sample size N = 1,979.

- However, before publishing their findings, they decided to do a replication study.
- They did a power analysis and determined that a sample size of 1300 would give power .995 to detect the effect size found in the original study at significance level .05.
- The replication study gave p = .59.
- See Nosek et al (2012) for details.

VII. THE FILE DRAWER PROBLEM

("PUBLICATION BIAS")

Publication bias refers to the influence of the results of a study on whether or not the study is published.

Examples of how results might influence the publication decision:

- Whether or not the results are statistically significant.
- Whether or not the results are practically significant.
- Whether or not the results agree with the hopes or expectations of the researcher or sponsor.

Publication bias is also called the ***file drawer problem***, especially when the nature of the bias is that studies which fail to reject the null hypothesis (i.e., that do not produce a statistically significant result) are less likely to be published than those that do produce a statistically significant result.

Older studies (see Sterling et al 1995, Song et al 2009, and Hopewell et al 2009) have reported indirect and/or anecdotal evidence of publication bias in the research literature.

The institution of the registry ClinicalTrials.gov in 2000 has now made it possible to do more direct studies of the file drawer problem.

A recent paper (Riveros et al, 2013) examined 600 clinical trials that had been registered on ClinicalTrials.gov and also had results posted there. Of these,

- Half did *not* have publications reported on PubMed
- There was also evidence of "selective" or "partial" publication bias:
    - Of the studies that had publications giving results for the primary outcomes, 73% listed adverse events on ClinicalTrials.gov, but only 45% listed adverse events in the published report.
    - Of these studies, 99% listed serious adverse events on ClinicalTrials.gov, but only 63% listed them in the published report.

Consequences of the File Drawer Problem:

1. Investigators may spend unnecessary effort conducting research on topics that have already been well researched but not reported because results were negative.

- Thus, *it is important to report negative results*.
    - But it's also important not to "spin" them. (See Couzin-Frankel, 2014)
- It's also important when planning research to search thoroughly for possible previous publications that have studied the same question.
    - If you can find negative results, this can help you plan appropriate sample size – or abandon the study altogether if results of the negative results were from a study with high power.

2. Effects that are not real may appear to be supported by research.

- Recall: If a significance level of 0.05 is used, then in repeated studies, about 5% of studies of a situation where the null hypothesis is true will falsely reject the null hypothesis
- Thus*, if just (or even predominantly) the statistically significant studies are published, the published record misrepresents the true situation*. (More on this tomorrow)

3. Furthermore, papers that are published because of Type I errors, if underpowered, may show an exaggerated effect size (See above), increasing the misrepresentation.

Some Methods Proposed to Detect Publication Bias:

1. Rosenthal (1979) proposed a method, based on probability calculations, for deciding whether or not a finding is "resistant to the file drawer threat."

- This method has become known as the *fail-safe file drawer (or FSFD) analysis.*
    - Scargle (2000) has criticized Rosenthal's method on the grounds that it fails to take into account the bias in the "file drawer" of unpublished studies, and thus can give misleading results.
    - More recently, Simonsohn et al (2013) have pointed out that the prevalence of "p-hacking" (to be discussed tomorrow) invalidates Rosenthal's method.

2. Various types of plots have been used to try to detect publication bias. These plot some measure of precision against effect size, or vice-versa.

- Some such plots are called "funnel plots" because they typically have a funnel shape.
    - However, Lau et al (2006) point out some problems in using these plots.
    - See also Sterne et al (2011) for recommendations in using funnel plots.
- Recently, Simonsohn et al (2013) have proposed a method called "p-curving" to detect possible publication bias and/or p-hacking (to be discussed tomorrow).

3. Research registries have been instituted in some areas.

- For example, certain clinical trials are now required by law to be registered at the NIH database ClinicalTrials.gov.
- These are beginning to point to possible systemic problems, such as:
    - The "partial publication bias" mentioned above.
    - "We are finding that in some cases, investigators cannot explain their trial, cannot explain their data. Many of them rely on the biostatistician, but some biostatisticians can't explain the trial design.

      So there is a disturbing sense of some trials being done with no clear intellectual leader."

      Deborah Zarin, Director, ClinicalTrials.gov, quoted in interview in Marshall (2011)
- Registration does not solve other problems (including those discussed in this course) that can make the literature misleading.
    - See, for example, blog posts during June, July, and August at http://www.ma.utexas.edu/blogs/mks discussing problems with registered reports.

4. Additionally, full data may reveal a different story from what appears in published papers, conference proceedings and registries.

- Although such data is increasingly becoming more available, obtaining it can often still be difficult or impossible.
- See Doshi et al (2012) for an example.
    - The editorial preface to this article says: "After publication of a Cochrane review into the effectiveness of oseltamivir [Tamiflu] in 2009, the reviewers got access to thousands of pages of previously unavailable data. [The authors] describe how it shook their faith in published reports and changed their approach to systematic reviews."
    - The authors obtained over 3000 pages of study reports from one drug company, and over 25,000 pages from the European Medicines Agency.
    - The new review based on the additional data took the equivalent of two full-time researchers for 14 months.
    - They also point out how calculations based on electronic data bases may be questionable (e.g., because of lack of standardized definitions for complications).

- More recently, Jefferson et al (2014) studied risk of bias in reports on 14 clinical trials of oseltamivir
    - They compared risk estimates for three different levels of reporting. (In increasing order of information: journal publications, core reports, and full clinical trial reports.)
    - They found that risk of bias increased as documents provided more information.
- An accurate "history" of computational methods used is also an important source of data on research methods.
    - One method for facilitating this is sweave, http://www.stat.uni-muenchen.de/~leisch/Sweave/.

See the Appendix for suggestions for helping to deal with the File Drawer Problem.

VIII. MULTIPLE INFERENCE

"Recognize that any frequentist statistical test has a random chance of indicating significance when it is not really present. Running multiple tests on the same data set at the same stage of an analysis increases the chance of obtaining at least one invalid result. Selecting the one "significant" result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading."

Professionalism Guideline 8, *Ethical Guidelines for Statistical Practice*, American Statistical Association, 1997

Performing more than one statistical inference procedure on the same data set is called ***multiple inference***, or ***joint inference***, or ***simultaneous inference***, or ***multiple testing***, or ***multiple comparisons***, or ***the problem of multiplicity***.

Performing multiple inference using frequentist methods *without considering the implications for Type I error* is a **common error** in research using statistics.

- For example, A. M. Strasak et al (2007) examined all papers from 2004 issues of the *New England Journal of Medicine* and *Nature Medicine* and found that 32.3% of those from *NEJM* and 27.3% from *Nature Medicine* were "Missing discussion of the problem of multiple significance testing if occurred."
- These two journals are considered the top journals (according to impact figure) in clinical science and in research and experimental medicine, respectively.

The Problem

Recall: If you perform a hypothesis test using a certain significance level (we'll use 0.05 for illustration), and if you obtain a p-value less than 0.05, then there are *three possibilities*:

1. The model assumptions for the hypothesis test are not satisfied in the context of your data.
2. The null hypothesis is false.
3. Your sample happens to be one of the 5% of samples satisfying the appropriate model conditions for which the hypothesis test gives you a Type I error – i.e., you falsely reject the null hypothesis.

Now suppose you're performing *two* hypothesis tests, *using the same data* for both.

- Suppose that in fact all model assumptions are satisfied and *both* null hypotheses are true.
- *There is in general no reason to believe that the samples giving a Type I error for one test will also give a Type I error for the other test.*
- See Jerry Dallal's Simulation
- This motivates considering the *joint Type I error rate*

Joint Type I error rate: This is the probability that a randomly chosen sample (of the given size, satisfying the appropriate model assumptions) will give a Type I error for *at least one* of the hypothesis tests performed.

The joint Type I error rate is also known as the ***overall Type I error rate***, or ***joint significance level***, or the ***simultaneous Type I error rate***, or the ***family-wise error rate*** (***FWER***), or the ***experiment-wise error rate***, etc.

- The acronym FWER is becoming more and more common, so will be used in the sequel, often along with another name for the concept as well.

Examples of common mistakes involving multiple inference:

1. An *especially serious* form of neglect of the problem of multiple inference is the one alluded to in the quote from the ASA ethics page:

- Trying several tests and reporting just one significant test, without disclosing how many tests were performed or correcting the significance level to take into account the multiple inference.
- *Don't do it!*
- To help you remember: Think Jelly Beans, http://xkcd.com/882/
- To help drive home the message, see more of Jerry Dallal's simulations:
  - http://www.jerrydallal.com/LHSP/jellybean.htm
  - http://www.jerrydallal.com/LHSP/cellphone.htm
  - http://www.jerrydallal.com/LHSP/coffee.htm

2. Some textbooks and software packages advise using a hypothesis test for equal variance before using a hypothesis test that has equal variance as a model assumption (e.g., equal variance two-sample t-test; standard ANOVA test).

- This can produce misleading results two ways
  - First, either test could produce Type I errors.
  - But the sequential use of the tests may lead to more misleading results than just the use of two tests.
- Zimmerman (2004) discusses this in more detail.

Multiple inference with confidence intervals

The problem of multiple inference also occurs for confidence intervals.

- In this case, we need to focus on the *confidence level*.
- *Recall*: A 95% confidence interval is an interval obtained by using a procedure that, for 95% of all suitably random samples, of the given size, from the random variable and population of interest, produces an interval containing the parameter we are estimating (assuming the model assumptions are satisfied).
- In other words, the procedure does what we want (i.e. gives an interval containing the true value of the parameter) for 95% of suitable samples.
- *If we're using confidence intervals to estimate two parameters, there's no reason to believe that the 95% of samples for which the procedure "works" for one parameter* (i.e. gives an interval containing the true value of the parameter) *will be the same as the 95% of samples for which the procedure "works" for the other parameter*.
- If we're calculating confidence intervals for more than one parameter, we can talk about the ***joint (***or ***overall*** *or* ***simultaneous*** *or* ***family-wise or experiment-wise) confidence level***.
- For example, a group of confidence intervals (for different parameters) has an ***overall 95% confidence level*** (or ***95% family-wise confidence level***, etc.) if the intervals are calculated using a procedure which, for 95% of all suitably random samples, of the given size from the population of interest, produces for *each* parameter in the group an interval containing that parameter (assuming the model assumptions are satisfied).

What to do about multiple inference

Unfortunately, *there is not (and can't be) a simple formula to cover all cases:*

- Depending on the context, the samples giving Type I errors for two tests might be the same, they might have no overlap, or they could be somewhere in between – and we can't know which might be the case.

- Various techniques for bounding the FWER (joint Type I error rate) have been devised for various special circumstances.
    - Some will be discussed below.

- There are also alternatives to considering FWER.
    - Some of these will be discussed below.

- For more information on other methods for specialized situatio, see, e.g., Hochberg and Tamhane (1987) and Miller (1981)

- See Efron (2010) for both an account of the history (Chapter 3) of the subject and discussion of some somewhat more recent developments in dealing with multiple inference, especially in large data sets.

Bonferroni method:

Fairly basic probability calculations show that *if the sum of the individual Type I error rates for different tests is ≤ α, then the overall ("family-wise") Type I error rate (FWER) for the combined tests will be* ≤ α.

- For example, if you're performing five hypothesis tests and would like an FWER (overall significance level) of at most 0.05, then using significance level 0.01 for *each* test will give an FWER (overall significance level) of at most 0.05.
- Similar calculations will show that if you're finding confidence intervals for five parameters and want an overall confidence level of 95%, using the 99% confidence level for each confidence interval will give you overall confidence level at least 95%. (Think of confidence level as 1 - *α.*)

The Bonferroni method can be a used as a fallback method when no other method is known to apply.

- However, if a method that applies to the specific situation is available, it will often be better (less conservative; have higher power) than the Bonferroni method, so calculate by both methods and compare.

- Holm's procedure (which depends on the Bonferroni idea, but in a more sophisticated way) is a relatively easy (e.g., on a spreadsheet) method that gives higher power than the basic Bonferroni method. It's described various places on the web – e.g., http://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method

The basic Bonferroni method is also useful for dividing up the overall Type I error between different types of inference.

- *Example*: If three confidence intervals and two hypothesis tests are planned, and an overall Type I error rate of .05 is desired, then using 99% confidence intervals and individual significance rates .01 for the hypothesis tests will achieve this.
- This method can also be used to apportion Type I error rate between *pre-planned inference* and *post-hoc inference*
  - *pre-planned inference*: the inferences planned as part of the design of the study
  - *post-hoc inference:* the inferences based on looking at the data and noticing other things of interest.
    - These are also called "data-snooping" – more on this tomorrow.
  - Example: If you plan 3 hypothesis tests, but might decide later to do more, you could plan to do the three "preplanned" hypothesis tests each at significance level .01, leaving .02 to divide between the data-snooping hypothesis tests
- However, *this apportioning should be done* <u>*before*</u> *analyzing the data.*

Whichever method is used, *it's important to make the calculations based on the number of tests that have been done,* <u>*not*</u> *just the number that are reported.*

- Remember Jelly Beans!

False discovery rate:

An alternative to bounding Type I error was introduced by Benjamini and Hochberg (1995): bounding the *False Discovery Rate*.

> The ***False Discovery Rate*** (FDR) of a group of tests is the *expected value of the ratio of falsely rejected hypotheses to all rejected hypotheses*.

("Expected value" refers to the mean of a distribution. Here, the distribution is the sampling distribution of the ratio of falsely rejected hypotheses to all rejected hypotheses tested.)

Note:

- The family-wise error rate (FWER) focuses on the possibility of making *any* Type I error among all the inferences performed.
- The false discovery rate (FDR) tells you what *proportion* of the *rejected* null hypotheses are, *on average*, really false.
- Bounding the FDR rather than the FWER may be a more reasonable choice when many inferences are performed, especially if there is little expectation of harm from falsely rejecting a null hypothesis.
- Thus it's increasingly being adopted in areas such as micro-array gene expression experiments or neuro-imaging.
- However, these may involve variations rather than the original definition given above; see Efron (2010) for more details.

As with the FWER, there are various methods of actually bounding the false discovery rate.

- For the original false discovery rate, see Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), and Benjamini and Yekutieli (2005)
- For variations of false discovery rate, see Efron (2010).

Higher Criticism (HC)

This is another alternative to bounding Type I error that's sometimes used in situations such as genome-wide testing.

- See Donoho and Jin (2015) for a review article on HC.
- See Klaus and Strimmer (2013) for a discussion of connections between HC and FDR.

Random Field Theory (RFT)

- This method is used in functional imaging data to try to account for spatial correlation when performing multiple hypothesis tests.
- See http://biostatistics.oxfordjournals.org/content/14/1/129

Multilevel Modeling

Gelman et al (2012) have proposed that in some cases multilevel modeling is a better way to address multiple inference than frequentist methods

- They point out that methods that such as Bonferroni corrections have unfortunate side effects:
  - They give large interval estimates for effects.
  - Since they require smaller cutoffs for significance, they are likely to produce Type M errors (because of The Winner's Curse).
- Multilevel modeling uses a different approach to inference that typically produces both smaller interval estimates, and more moderate point estimates of effects than standard frequentist methods, so may be a better way to approach multiple inference.

Subtleties and controversies

Bounding the overall Type I error rate (FWER) will reduce the power of the tests, compared to using individual Type I error rates.

- Some researchers use this as an argument against multiple inference procedures.
- The counterargument is the argument for multiple inference procedures to begin with: Neglecting them will produce excessive numbers of false findings, *so that the "power" as calculated from single tests is misleading*.
  - See Maxwell and Kelley (2011) and Maxwell (2004) for more details.
- Bounding the False Discovery Rate (FDR) will usually give higher power than bounding the overall Type I error rate (FWER).

Consequently, it's important to consider the particular circumstances, as in considering both Type I and Type II errors in deciding significance levels.

- *In particular, it's important to consider the consequences of each type of error in the context of the particular research*.

Examples:

1. A research lab is using hypothesis tests to screen genes for possible candidates that may contribute to certain diseases.
   - Each gene identified as a possible candidate will undergo further testing. The results of the initial screening are not to be published except in conjunction with the results of the secondary testing,
     - Case I: If the secondary screening is inexpensive enough that many second level tests can be run, then the researchers could reasonably decide to ignore overall Type I error in the initial screening tests, since there would be no harm or excessive expense in having a high Type I error rate.
     - Case II: If the secondary tests were expensive, the researchers would reasonably decide to bound either family-wise Type I error rate or False Discovery Rate.

2. Consider a variation of the situation in Example 1:
   - The researchers are using hypothesis tests to screen genes as in Example 1, but plan to publish the results of the screening *without* doing secondary testing of the candidates identified.
   - In this situation, ethical considerations warrant bounding either the FWER or the FDR -- *and* taking pains to emphasize in the published report that these results are just of a preliminary screening for possible candidates, and that these preliminary findings need to be confirmed by further testing.

The Bottom Line: No method of accounting for multiple inference is perfect, which is one more reason why ***replication of studies is important!***

Note: For more discussion of multiple inference in exploratory research, see Goeman and Solari plus discussion (2011).