

CONTENTS OF DAY 3

- I. Catch-up and review from yesterday as needed
- II. Trade-Off between Type I and Type II errors 3
- III. Deciding what significance level to use 9
- V. Power of a statistical procedure 14
 - Factors affecting power 20
 - Significance level 20
 - Sample size 21
 - Variance 24
 - Experimental design 26
 - Calculating sample size 27 and appendix
 - Detrimental effects of underpowered or overpowered studies 28
 - Winner’s Curse (AKA Statistical Significance Filter) 30
 - Type M and S errors 32
 - Caution on different meanings of “replication” 34
 - Overpowered studies 36
- VI. Common mistakes involving power 37
- VII. The File Drawer Problem (“Publication Bias”) (as time permits) 46

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

NOTES FOR SUMMER STATISTICS INSTITUTE COURSE

COMMON MISTAKES IN STATISTICS – SPOTTING THEM AND AVOIDING THEM

Day 3: Type I and II Errors, Power, and the File Drawer Problem

MAY 22 - 25, 2017

Instructor: Mary Parker

II. TRADE-OFF BETWEEN TYPE I AND TYPE II ERRORS

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.

In the following table, fill in the cells with “Correct” or “Error, as appropriate, and then indicate Type I and Type II errors.

	Truth (for population studied)	
	Null Hypothesis True	Null Hypothesis False
Decision (based on sample)	Reject Null Hypothesis	
	Don't reject Null Hypothesis	

Analogy: Results of a murder 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:

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

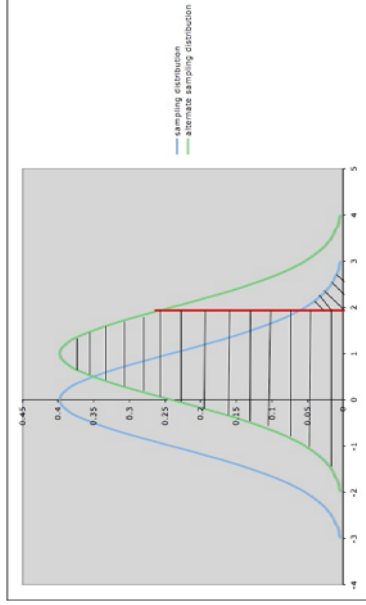
This could be more than just an analogy if the verdict hinges on statistical evidence (e.g., a DNA test), and where rejecting the null hypothesis would result in a verdict of guilty, and not rejecting the null hypothesis would result in a verdict of not guilty.

Which type of error here does the US Court system consider as the worse? (Or does it consider them to be equally bad mistakes?)

What does that imply about how the decision is made?

If we make it hard to make one kind of error, what does that do to the likelihood of making the other kind of error?

The following diagram illustrates *both* the Type I error *and* the Type II error for $H_0: \mu = 0$ and $H_a: \mu = 1$, using 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 value in the 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.

→ In the diagram on the previous page, look at the relative sizes of the Type I and Type II error. Which of these best describes it?

- Has about equal probability of making a Type I error and Type II error.
- Has a much lower probability of making a Type I error than a Type II error.
- Has a much lower probability of making a Type II error than a Type I error.

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

- increase α ?
- decrease α ?

- It is possible, of course, we are interested in hypotheses $H_0: \mu \leq 0$ and $H_a: \mu > 0$. Note, however, that to find numerical values for the Type I and Type II errors, we must choose specific values to use for the parameter value in each hypothesis, as we did on the previous page.
 - Typically, statisticians will compute power for several different values in the H_a . Typically they will only use the endpoint value in the H_0 .
 - See the discussion of power in the next section for related details.

III. DECIDING WHAT SIGNIFICANCE LEVEL TO USE

This should be done by someone familiar enough with the context of the problem to make appropriate decisions about which type of consequence is worse, and how much worse than the other. That balance is a major part of choosing a significance level.

There are two reasons you should work out this *before analyzing* the data -- *preferably before gathering the data*.

- 1) The significance level desired is one criterion in deciding on an appropriate sample size.
 - See discussion of Power in the next section.
- 2) If more than one hypothesis test is planned, additional considerations need to be taken into account.
 - See discussion of Multiple Inference in a later lecture.

The choice of significance level should be based on the consequences of Type I and Type II errors. Think back to the diagram.

- a. If the consequences of a Type I error are serious or expensive, a very _____ significance level is appropriate.
- b. If the consequences of a Type I error are not very serious (and especially if a Type II error has serious consequences), then a _____ significance level is appropriate.

Answers: a. Small b. Relatively large

Example 1: Two drugs are being compared for effectiveness in treating the same condition. Typical significance levels are 1%, 5%, or 10%.)

- o Drug 1 is very affordable, but Drug 2 is extremely expensive.
- o The null hypothesis is “both drugs are equally effective.”
- o The alternate is “Drug 2 is more effective than Drug 1.”
- o Type I error, in context: _____
- o Type II error in context: _____
- o Consequences of a Type I error: _____
- o Consequences of a Type II error: _____
- o From a patient’s point of view, which consequence is worse? _____
- o From a patient’s point of view, should we choose a small, medium, or relatively larger significance level? _____

Answers:

Type I error is deciding that Drug 2 is more effective, when it really isn't. The consequence of that is that patients spend a lot more money for no additional benefit.

Type II error is deciding that Drug 2 is not more effective, when it really is. The consequence of it is that patients don't take Drug 2.

Type I error worse. Choose small significance level and accept a larger Type II error.

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."
- Type I error in context: _____
- Type II error in context: _____
- Consequences of a Type I error: _____
- Consequences of a Type II error: _____
- From a patient's point of view, which consequence is worse? _____
- From a patient's point of view, should we choose a small, medium, or relatively larger significance level? _____

Answers:

Type I error is deciding that Drug 2 is more harmful than Drug 1 when it really is not. The consequence is that you don't take Drug 2, and there is no real harm in that.

Type II error is deciding that Drug 2 is no more harmful than Drug 1 when it really is more harmful. The consequence of that is that patients use a harmful drug.

Type II error worse. Choose relatively large significance level and low probability of a Type II error.

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"
- Type I error, in context: _____
- Type II error, in context: _____
- Consequences of a Type I error: _____
- Consequences of a Type II error: _____
- From a patient's point of view, which consequence is worse? _____
- From a patient's point of view, should we choose a small, medium, or relatively larger significance level? _____

Answers:

Type I error is deciding the vaccine does produce serious disease when it really does not. The consequence is that you don't get the vaccine.

Type II error is deciding the vaccine does not produce serious disease when it really does. The consequence of that is that you take the vaccine and get a serious disease.

Type II error worse. Choose relatively large significance level and low probability of a Type II error..

- Indeed, in these cases, the Type II error rate is often set at 1%, whereas in much research, a Type II error rate of 20% 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.
 - For discussion of the types of cost-benefit analysis that might go into medical decisions, see the following and reference therein:
 - <http://andrewgelman.com/2015/09/17/26612/>
 - http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2641547
 - <http://content.healthaffairs.org/content/24/1/93.full>

IV: 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, going to more expense, or subjecting people to more risk 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 $H_0: \mu = 100$ and $H_a: \mu \neq 100$ you might be interested in the probability of rejecting H_0 when $\mu \geq 105$, or when $|\mu - 100| > 5$, etc.

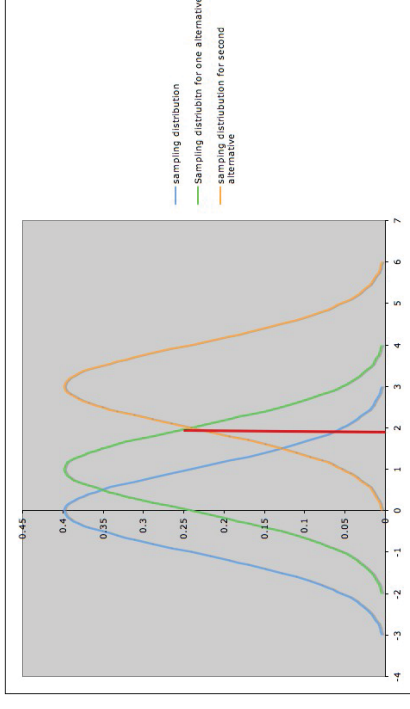
As with Type II Error, we may think of power for a hypothesis test in terms of *power against a specific alternative* rather than against a general alternative.

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 to the *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 to the *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 <https://istats.shinyapps.io/power/> [use population mean] 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

$$\begin{aligned} \beta + (\text{Power of a test against a specific alternate hypothesis}) \\ = \text{total area under sampling distribution curve} \\ = 1, \end{aligned}$$

so

$$\text{Power} = 1 - \beta$$

Illustration: <https://istats.shinyapps.io/power/> [use population mean]

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

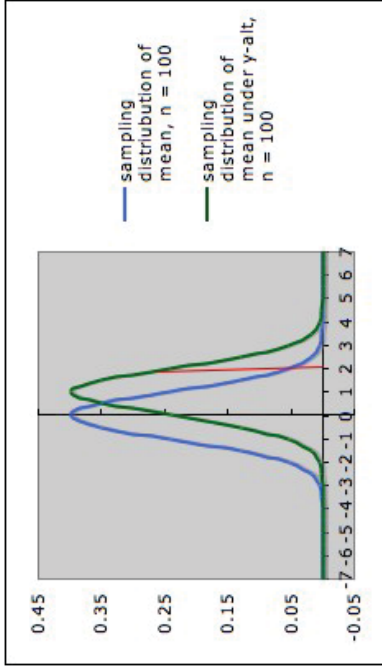
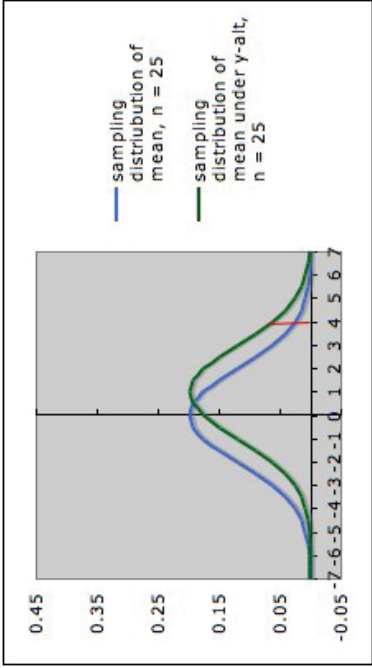
This can be seen in the diagram illustrating power:

- *Increasing* the significance level α will move the red line to the *left*, and hence will *increase power*.
- Similarly, decreasing significance level decreases power.
- See <https://istats.shinyapps.io/power/> [use population mean] for an interactive demonstration

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 <https://istats.shinyapps.io/power/> [use population mean] for an interactive demonstration of the interplay between sample size and power.

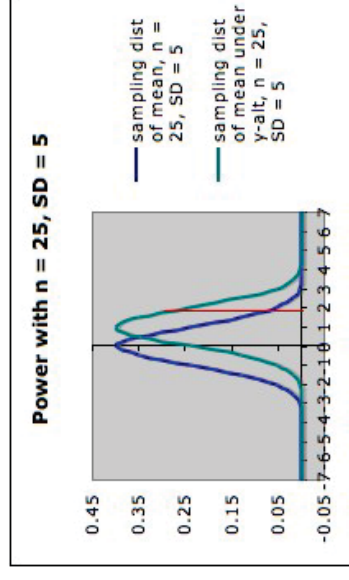
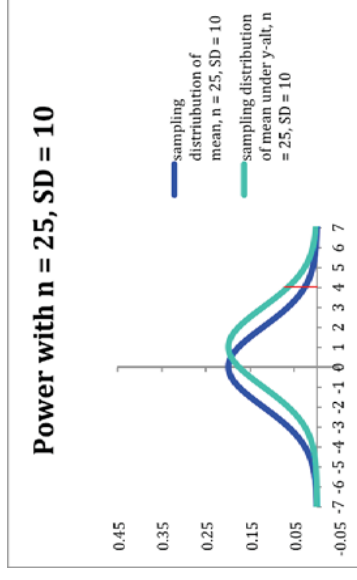
Note: Sample size needed to give desired power typically increases at an increasing rate as power increases. (e.g., in the above example, increasing the sample size by a factor of 4 increases the power by a factor of about 2; the graphics aren't accurate enough to show this well.)

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 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 <https://istats.shinyapps.io/power/> [use population mean] for an interactive demonstration. [Try mean 1, st deviations 1 and 5]

Note: Variance can sometimes be reduced by using a better measuring instrument, by restricting to a subpopulation (but then be careful not to extrapolate!), 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.
 - Calculating power or sample size for the wrong experimental design is a **common mistake**
- The statistical analysis will depend on the experimental design
- 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 the sample size 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

Underpowered 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 may not be detected*.

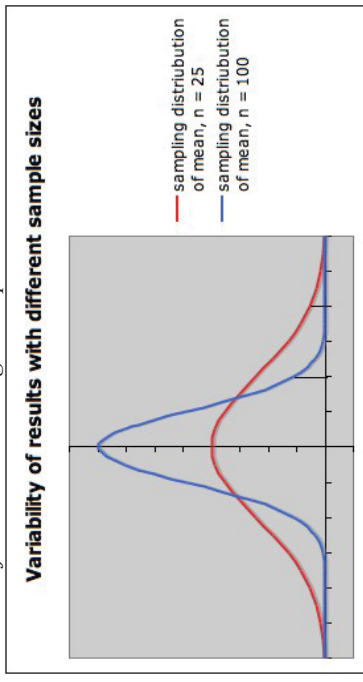
But there's another, more subtle, but important consequence:

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 $\mu = 0$ 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 exhibiting a large effect than studies with high power.*

(Just to be clear: Here: effect size = sample mean – value in H_0)

- *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 in a later section).
- Recall that low power may result from:
 - Small sample size
 - Small alpha level
 - Large variance

The effect of the statistical significance filter shows up in the recent Open Science Collaboration initiative to replicate 100 published psychology studies (Open Science Collaboration 2015):

The mean effect size of the original studies was 0.403, whereas the mean effect size of the replication studies was 0.197.

(For an interesting popular press article about the Open Science Collaborative replication study, see McArdle (2015))

Maxim: The statistical significance filter suggests the maxim:

If a result is surprising, one study is not enough to provide convincing evidence. Replication is important.

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.
 - In this language, the results of the Open Science Collaboration paper showed that Type M errors were common in the original studies, and in the direction predicted by theory.
- A *Type S error* occurs when the effect size estimate has a different sign than the true effect size. . .
- 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.

Caution: The word “replication” is used in more than one way. Possibilities include:

1. Repeating the original analysis with the same data
 - This *could* detect errors in performing the analysis.
 - It *would not* give any evidence on how much the result depends on the particular data set.
 - However, data and code are often not accessible. Example: Chang and Li (2015) tried to replicate 67 published economics studies.
 - 6 papers used confidential data, so were not replicable
 - 2 required software not available to the authors, so were not replicable by the authors
 - 29 of 35 papers in journals requiring data and code to be made available actually provided data and code
 - 11 of 26 papers in journals not requiring data and code to be made available actually provided data and code
 - The authors successfully replicated the main quantitative result of 22 of the 67 papers without assistance from the original authors; 7 more were successfully replicated with assistance from the original authors.

2. Repeating the original analysis with a new data set

- This would help check how much the result depends on the particular data set.
 - But if the original choice of analysis has flaws, this would be silly to do.
3. Applying an appropriate (possibly different) analysis to a new data set.
- If the original choice of analysis has flaws, this plus applying the better analysis to the old data set is the best option.
 - See <http://www.ma.utexas.edu/blogs/mks/2014/06/22/beyond-the-buzz-on-replications-part-i-overview-of-additional-issues-choice-of-measure-the-game-of-telephone-and-twwadi/> and following posts for an example where this would be appropriate (but was not done).

Overpowered studies

1. They 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)

2. They can produce a “statistically significant” result when the result is not practically significant.
 - Remember that, if the sample size is quite large, then a sample statistic value very close to the H_0 value may provide significant evidence against the H_0 , because the variability in the sampling distribution is so small.
 - But a sample statistic value very close to the H_0 value is not likely to be practically significant.

V: 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, without taking power into account.*
 - Power decreases with decreasing sample size.
 - Thus a small sample size may not be able to detect a 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 Wiecezorek)

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 can be even more complex.

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

7. Confusing retrospective power and prospective power.

- Power as defined above for a hypothesis test is also called prospective or a priori power.

It's a conditional probability, $P(\text{reject } H_0 \mid H_0)$, calculated without using the data to be analyzed.

In fact, it's 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 (or might not) be legitimate to use to estimate the power and sample size for a *future* study, but *cannot 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.

VI. 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.

There are various ways in which study results might influence the publication decision, including:

- A result that is not statistically significant might not be published or even announced.
- A result that is not practically significant might not be published or announced.
- A result that does agree with the hopes or expectations of the researcher or sponsor might not be published.

Publication bias is also called the **file drawer problem**, especially when the nature of the bias is that studies that 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.

Cartoon:

<https://mchankins.files.wordpress.com/2013/04/filedrawer1.jpg?w=940>

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.

One common form of partial publication bias is **outcome switching**: pre-specified outcomes are not reported in the final report, or new outcomes were added.

The COMPare project (<http://compare-trials.org/>) checked all clinical trials published in the top five medical journals from October 2015 to January 2016 and compared published outcomes with those pre-specified in the registry or protocol. In the 67 trials checked, they found a total of

- 300 pre-specified outcomes not reported, and
- 357 not-pre-specified outcomes added.

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 Cousin-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, (at least) 5% of studies 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 (“Winner’s curse”), 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 FSD) 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 Zarlin, 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 pharmaceutical 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).

- Further examples to explore:
 - Doshi et al (2013) A call for and discussion of publishing results of “invisible and abandoned trials”.
 - Le Noury et al (2015) A reanalysis of clinical trial data on efficacy and harm of Paxil use for treatment of depression in adolescents. (An example of the type of study called for in the preceding article.)
 - Der Spiegel interview with whistleblower Peter Wilmschurst
<http://www.spiegel.de/international/zeitgeist/spiegel-interview-with-whistleblower-doctor-peter-wilmschurst-a-1052159.html>
- 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.