

# Week 10: The Effective Use of Statistical Tests

Jessie Yeung

STA 220

Winter 2024

# Overview

- Last week, we covered an introduction to statistical testing
- We will be continue the topic of statistical tests
- Topics for this week
  - Review/Overview of last week's content
  - Significance level, Power, Type 1 and Type 2 errors
  - General advice about using tests
  - Connection between Hypothesis Tests and Confidence Intervals
- This content corresponds with [Module 9](#)

# Overview of Last Week

# Steps of a Statistical Test

1. Determine your null and alternative hypotheses. Assume the null is true.
2. Collect the data and calculate your test statistic
3. Calculate your p-value
4. Make a conclusion based on the p-value

# Null and Alternative Hypothesis

- Null hypothesis will take the form of

$$p = \text{some value}$$

$$\mu = \text{some value}$$

- Alternative hypothesis will take the form of

$$p \neq \text{some value}$$

$$\mu \neq \text{some value}$$

$$p < \text{some value}$$

$$\mu < \text{some value}$$

$$p > \text{some value}$$

$$\mu > \text{some value}$$

- Direction of the alternative hypothesis depends on whether there is a specific direction specified in the question
- To run the test, we assume that the null is true

# Calculating Test Statistic

- The test statistic is a value with a known distribution.
- Need to assume the null is true in order to calculate it
- For tests about  $p$ ,

$$Z\text{-Statistic} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

- For tests about  $\mu$  where  $\sigma$  is unknown,

$$T\text{-Statistic} = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \sim t_{n-1}$$

- For tests about  $\mu$  where  $\sigma$  is known,

$$Z\text{-Statistic} = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

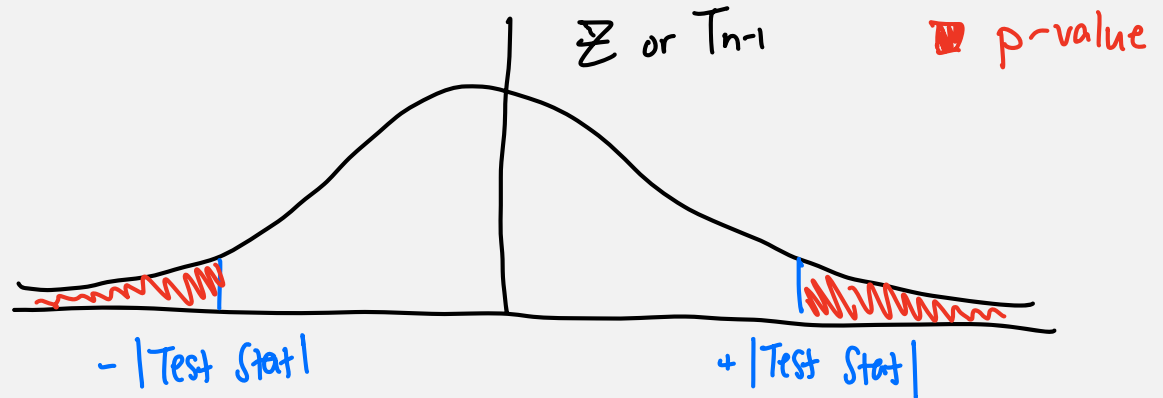
# P-value

- P-value is the probability (given that the null is true) that we see data that is as extreme or more extreme than the data we observed
- This is equivalent to how likely it is to get a test statistic that is even farther away from the centre of the distribution as the one we calculated.

## Two-sided Test:

$H_0$ : parameter = some value

$H_A$ : parameter  $\neq$  some value

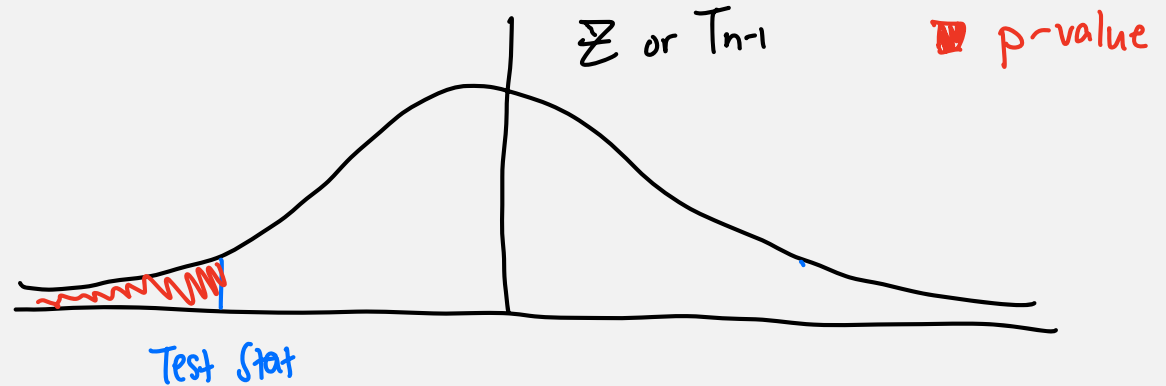


# P-value

## One-sided Test:

$H_0$ : parameter = some value

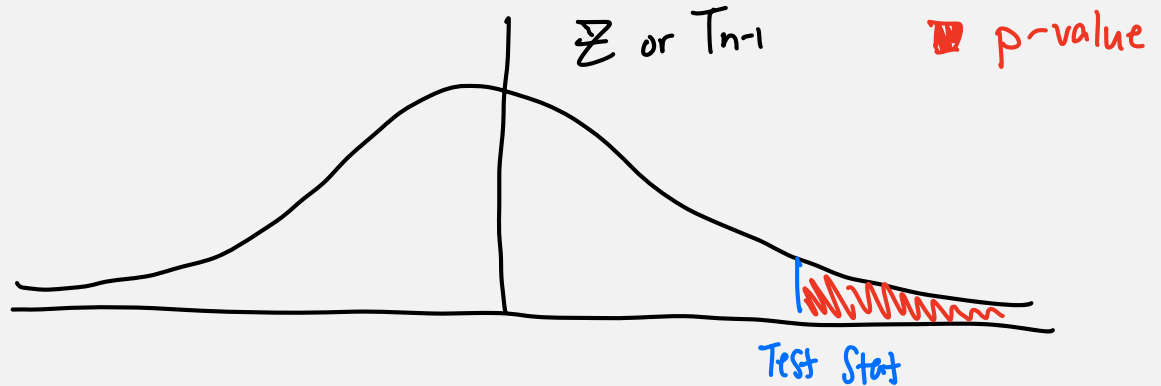
$H_A$ : parameter  $<$  some value



## One-sided Test:

$H_0$ : parameter = some value

$H_A$ : parameter  $>$  some value





# P-value

- The diagrams in the last two slides are equivalent to the rules provided in this table last week

	$H_A$ : parameter < value	$H_A$ : parameter > value	$H_A$ : parameter $\neq$ value
Proportion	$P(Z < \text{test value})$	$P(Z > \text{test value})$	$2 \times P(Z >  \text{test value} )$
Mean	$P(T_{n-1} < \text{test value})$	$P(T_{n-1} > \text{test value})$	$2 \times P(T_{n-1} >  \text{test value} )$

# Conclusion

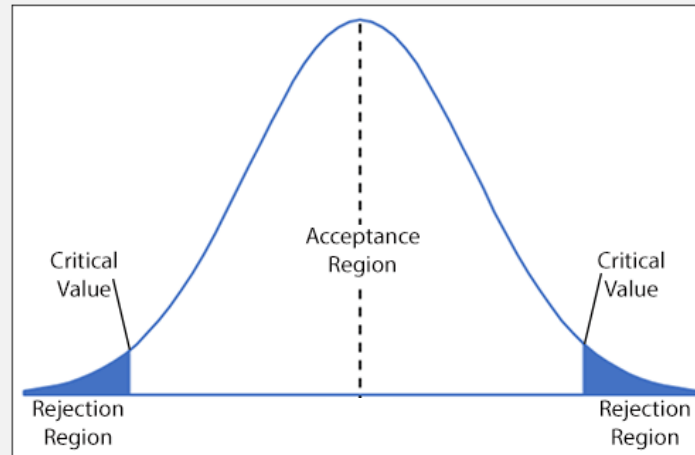
- A small p-value means that the data is very unlikely assuming that the null is true
  - Eg. Consider a null hypothesis of  $p = 0.01$ . Suppose further that we see data such that  $\hat{p} = 0.9$ . This data would be very unlikely in a world where the null is true. This would lead to a very small p-value.
- The smaller the p-value, the stronger the evidence against the null hypothesis.
- Conclusion using a threshold of 0.05:
  - If the p-value  $< 0.05$ , then we reject the null hypothesis.
  - If the p-value is  $> 0.05$ , then we fail to reject the null hypothesis

# Instead of calculating the p-value

- Instead of calculating the p-value and comparing it to a threshold, there is a *shortcut*
- Instead of calculating the p-value and comparing it to a threshold, we can just compare the test statistic to a critical value

# Rejection and Acceptance Regions for Two-Sided Tests

- To use this shortcut with a threshold of 0.05:
  - Find the critical value such that the area of both tails together is 0.05. That means the area of each tail should be 0.025.
  - Rejection Region: If the test statistic is less than (- critical value) or greater than (+ critical value), then the p-value will be less than 0.05.
  - Acceptance Region: If the test statistic is between (- critical value) and (+ critical value), then the p-value will be greater than 0.05.



# Significance Level and Power

# Why do we run a statistical test?

- We are always trying to do the same thing when we run a statistical test.
  - a claim/statement has been made regarding the possible value of a population parameter
  - when we collect data, we are trying to gather evidence in support of an alternative value for the parameter
  - this is because it is easier to show that your sample value is much too far for the hypothesized value to be true, than it is to show that your value is close enough
- But since our tests are based off data that vary between samples, there are of course chances that we will make the wrong conclusion just because of the sample we got.
  - so we will try to control for the chances that we will make a mistake.

# P-values and Significance

- Recall that our p-value accounts for the variability in our sample value due to the sampling distribution
  - we can think of it as talking about the chances that we could have computed a test statistic that was farther from the hypothesized mean/proportion than what we got
  - and we know that smaller p-values imply stronger evidence against the null hypothesis.
- We can also decide to set a cutoff/threshold for our p-value, in order to conclude whether the p-value is small enough
  - this threshold is called the **significance level**  $\alpha$
  - for some  $\alpha$ , if our p-value is smaller than it, we say that our data are **statistically significant**
  - statistical significance means we have strong enough evidence to reject the null hypothesis.
  - last week, we were using  $\alpha = 0.05$

# Significance Level

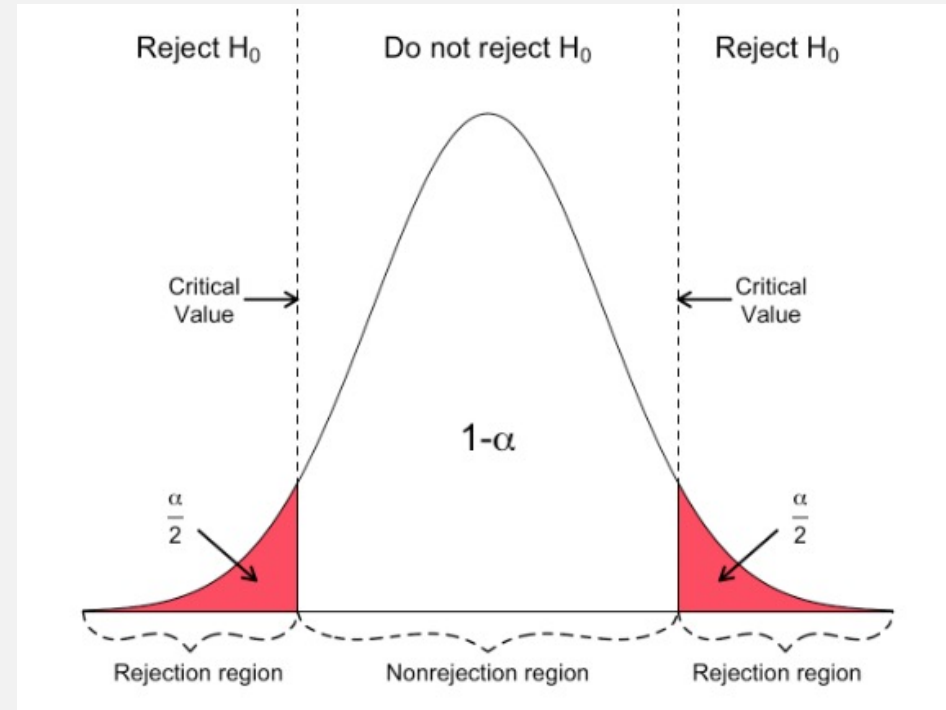
- Notice that we are using  $\alpha$  to talk about both the significance level of tests and the confidence level for intervals
  - this is not a coincidence!
- Remember that in confidence intervals,  $\alpha$  represents the tail probability
  - or the chance that our interval will not capture the true population value if we repeatedly collect samples and compute intervals
- For statistical tests,  $\alpha$  also represents a tail probability
  - specifically, if  $H_0$  is true and we repeatedly collect samples and run our test, the proportion of tests where we mistakenly reject  $H_0$  will be  $\alpha$



# Significance Level

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is True})$$


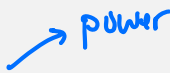
- Why is  $\alpha$  the probability of mistakenly rejecting  $H_0$  when it is in fact true?
  - If  $H_0$  is true, then the test statistic does really follow the distribution we think it does.
  - Our p-value will be less than  $\alpha$  if the test statistic falls within the rejection regions.
  - Notice that the rejection regions have a total area of  $\alpha$
  - When  $H_0$  is true, the probability of ending up in the rejection region is  $\alpha$



# Using really small $\alpha$

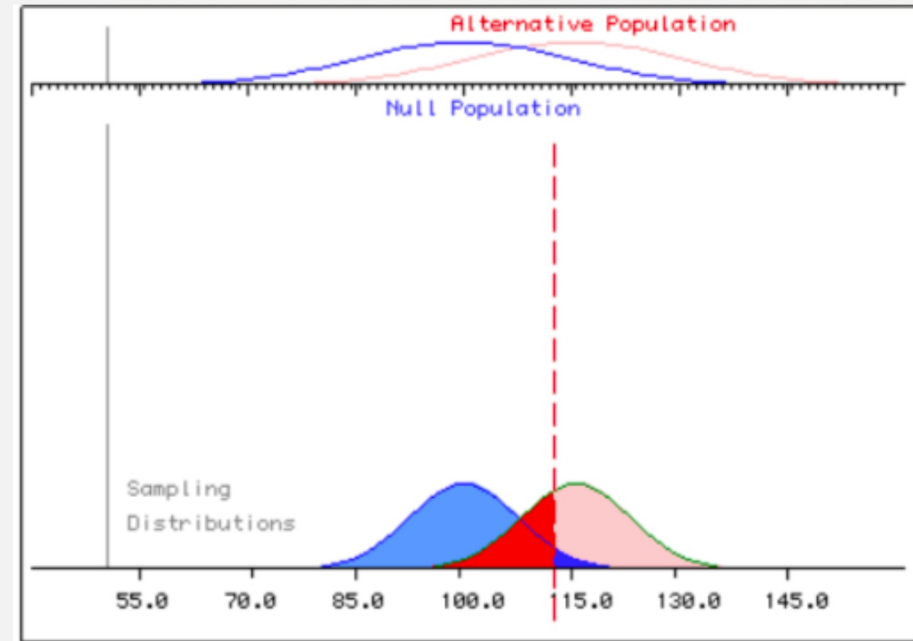
- If you really don't want to incorrectly reject  $H_0$  when it should not be rejected, you may think that picking a really small  $\alpha$  would help.
  - technically of course it would!
- $\alpha$  tells you how far away your sample value has to be from the population value in order to conclude that  $H_0$  is wrong
  - if it's too small, then you're requiring that your sample value be so very far from  $H_0$  that you may never reject the null, even when you should

# Choosing your $\alpha$

- So clearly, we can't just choose a really small  $\alpha$  because then we might incorrectly fail to reject when in fact  $H_A$  is correct.
- What we want to do is pick a value for  $\alpha$  (i.e.  $P(\text{reject } H_0 \mid H_0 \text{ true})$ ) that is small enough for us to be confident in our conclusion. 
- At the same time, we also want to make sure that we make the correct conclusion if  $H_A$  is actually true 
  - So we want to make sure  $P(\text{reject } H_0 \mid H_0 \text{ false})$  is sufficiently large
  - This probability is called the **power** of the test. We want our test to be powerful because we want our chances of being right to be high.
  - So to have a good test, we need to have a trade-off between significance level and power

# Power and Significance

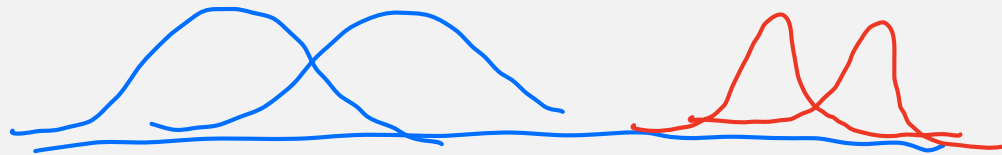
- Consider  $H_0: \mu = 100, H_A: \mu > 100$
- Left Distribution: is the sampling distribution of the sample mean if the null is true
- Right Distribution: is the sampling distribution of the sample mean if the true mean is 115 (null is false)
- The dashed line is the point corresponding to the  $\alpha$  cutoff
- Dark blue is  $\alpha$ , pink is power
- If we move the dashed line left,  $\alpha$  gets bigger (bad) but so does the power (good)
- If we move the dashed line right,  $\alpha$  gets smaller (good) but so does power (bad)
- So need to find a way to get small  $\alpha$  but a large power



# Power and Significance

- Clearly, the reason why we have a trade-off between power and significance level is partly due to how much the two possible sampling distributions overlap.
  - if the values for  $H_0$  and  $H_A$  are very close, then the curves overlap a lot this means it will be much easier to make an incorrect conclusion, i.e. we are more likely to fail to reject  $H_0$  when we should not.
    - so in this case the power will get smaller
  - obviously we could then increase the power by moving  $H_0$  and  $H_A$  farther apart
    - although in practice, you can't actually do anything about this.

# Power and Significance



- So since we can't actually move the sampling distributions, what else can we do to get higher power with our test?
- Well one important feature of sampling distributions is that their shape depends heavily on the **size of our sample**.
  - the variance of the sampling distribution will decrease as sample size increases  
$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$
  - when we take a small sample, we generally get a sampling distribution that has similar variability to the original population
  - but if we take a large sample, we get a much much narrower/skinnier distribution for the sample mean.
  - if the variability of the two sampling distributions can change, then this changes how much the null and alternative distributions overlap.

# Bigger Sample Size

- If sample size goes up, then variance of the sampling distribution goes down
- Since both the null and alternative distributions are both sampling distributions (just having different means), this means both distributions will get skinnier
- This means that even though they still sit in the same place, they will have less overlap.
  - less overlap means higher power
  - less chance of failing to reject the null, when the null is false

# Less Variability

- The last way we can ‘manipulate’ the power of the test is to consider the actual variability in the population itself.
- If I am sampling from a population that is very spread out, then my  $\sigma$  will be large
- Since my sampling distribution for the sample mean has standard deviation  $\sigma/\sqrt{n}$ , a bigger population variance means bigger sampling distribution variance too.
  - this gives us the opposite result of increasing the sample size
  - as variance goes up, power goes down because we have more overlap



# Summary of Factors Affecting Power

- Basically there are 4 factors that can affect the power of a statistical test
  1. the distance between the null value and the alternative value (greater distance means more power)
  2. higher significance level means higher power
  3. less variability in population means higher power
  4. larger sample size means higher power
- In practice though, you can really only control the sample size and your choice of significance level.
- However, because we know that all of these factors affect the power of your test, we can use them to figure out how big of a sample we should collect.
  - but we won't get into this.

# Errors in Statistical Tests

# Types of Errors

- When we talk about significance level and power, both of these concepts can be understood in terms of the chance of making an error in our conclusions.
- There are two types of these kinds of errors we can make:
  - we can reject the null hypothesis when in fact we should not (because the null is actually the true value) - called **Type 1 error**
  - we can fail to reject the null hypothesis when in fact we should not (because the null is actually wrong) - called **Type 2 error**
- Both of these types of errors can be represented in terms of the significance level or the power of the test.

		The Truth	
		$H_0$ True	$H_0$ False
My Decision	Reject $H_0$	Type I Error	OK
	Fail to reject $H_0$	OK	Type II Error

# Type 1 Error

- The important distinction between Type 1 and Type 2 error is the assumption we are making about which hypothesized value is true.
- In Type 1 error, the value in the null hypothesis is actually true.
  - Since we are talking about making an error, the only way that we can make the wrong conclusion if the null is actually true, is if we say the null is not true.
  - Therefore, type 1 error corresponds to rejecting when you should not.
- But we already mentioned this particular situation when we were talking about significance levels.
  - So the probability of making a type 1 error is just the significance level  $\alpha$

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true}) = P(\text{type 1 error})$$

# Type 2 Error

- Type 2 error considers the situation where the alternative hypothesis is actually the true value.
- Since we are again talking about making an error, if the alternative is true, then the only way we can make the wrong conclusion is to say that the null is true.
  - In hypothesis testing though, we don't ever actually claim the null is true, only that we have failed to reject it.
- This sounds slightly similar to the idea of power.
  - however power is the probability of correctly rejecting the null when the null is not true (i.e. when the alternative is true).
  - So  $P(\text{Type 2 Error}) = 1 - \text{power}$ , and we call  $P(\text{Type 2 Error}) = \beta$

$$\begin{aligned} P(\text{Type II Error}) &= P(\text{Fail to reject } H_0 \mid H_0 \text{ is False}) \\ &= 1 - P(\text{rejecting } H_0 \mid H_0 \text{ is False}) \\ &= 1 - \text{Power} \\ &= \beta \end{aligned}$$

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

# Example: Testing Cars

A clean air standard requires that vehicle exhaust emissions not exceed specified limits for various pollutants. Suppose government regulators double-check a random sample of cars that a suspect repair shop has certified as okay. They will revoke the shop's license if they find significant evidence that the shop is certifying vehicles that do not meet standards.

$H_0$ : Shop is conducting tests accurately

$H_A$ : Shop is certifying vehicles that do not meet standards.

# Example: Testing Cars

- a) In this context, what is type 1 error?
- b) In this context, what is type 2 error?
- c) Which type of error would the shop owner consider more serious?
- d) Which type of error might environmentalists consider more serious?
- e) What is meant by the power of the test the regulators are conducting?
- f) Will the power be greater if they test 20 or 40 cars?
- g) Will the power be greater if they use 5% or 10% level of significance



- a) Regulator claims that car shop is certifying vehicles that don't meet standards, when the shop is actually doing tests accurately.
- b) Regulator claims that the shop is conducting tests accurately when they are actually certifying vehicles that don't meet standards.
- c) Type I
- d) Type II
- e) Probability that the regulator catches the shop's bogus certifications, given that they are actually doing something wrong.
- f) Larger sample size  $\rightarrow$  Higher Power  
40 cars
- g) Higher  $\alpha$   $\rightarrow$  Higher Power  
10%

# Connection to Cls

# Connection to Confidence Intervals

- As we saw before, the fact that we use the same letter to talk about significance level in both confidence intervals and hypothesis tests is not a coincidence.
- This is because, even though the two methods differ in what we get as results, they are very related to one another.
- Since we only know how to find two-sided confidence intervals, we will just focus on the connection between two-sided intervals and hypothesis tests.
- It all comes down to the fact that we are using the z-score to measure the distance of our sample value from the hypothesized true value of the mean of the sampling distribution.

# Building Confidence Intervals

- When we create a confidence interval, we select what we want our coverage/confidence level to be
  - e.g. let's make a 95% confidence interval  $\rightarrow (1 - \alpha) = 0.95 \rightarrow \alpha = 0.05$
  - since our confidence intervals are two-sided, we divide this over the two tails of the sampling distribution
- To build the interval, we need to find the corresponding critical value  $z_{\alpha/2}$  or  $t_{\alpha/2, n-1}$  such that we only miss capturing the true parameter 5% of the time

# Missing the parameter

- When we are dealing with the confidence interval, we know that 5% of the time we will not capture the true parameter value.
- Since our interval is centred at the sample value, we say we don't capture the true value when it would fall outside our interval
- The edges of the interval are  $\pm t_{\alpha/2, n-1} \times \frac{s}{\sqrt{n}}$  from  $\bar{x}$ , so we would miss  $\mu$  if  $\mu$  is farther from  $\bar{x}$  than the edge of the interval

# CI that doesn't capture $\mu$

- Consider a situation where  $\mu = 108$  although we don't know this value
- Let  $\alpha = 0.05$
- Instead we are given  $\bar{x} = 110, n = 30, s = 5$  to build a confidence interval
- The critical value is  $t_{\alpha/2, n-1} = 2.04$  and the 95% CI is  $(108.133, 111.867)$
- Notice that the true  $\mu$  happens to not be in the confidence interval.

# CI that doesn't capture $\mu$

- Recall that  $\mu = 108$  doesn't fall in the 95% confidence interval

- If we find the z-score of using the true mean, we find that  $\frac{\bar{X} - \mu}{s/\sqrt{n}} =$

$$\frac{110 - 108}{5/\sqrt{30}} = 2.19$$

- Notice that this z-score of 2.19 does not fall between the critical values of  $\pm 2.04$
- If we were testing  $H_0: \mu = 108$ , we would reject

# CI that doesn't capture $\mu$

- Now consider if  $\mu = 109$ , which does fall in our confidence interval.

Then the z-score becomes  $\frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{110 - 109}{5/\sqrt{30}} = 1.09$

- Notice that this z-score of 1.09 does fall between the critical values of  $\pm 2.04$
- If we were testing  $H_0: \mu = 109$ , we would not reject



# Hypothesis Tests and P-values

- In general, the true population parameter will fall in the  $(1 - \alpha)\%$  CI if and only if the z-score is between the critical values
- But what we did on the previous slide (i.e., compute a z-score/test statistic and compare it to a critical value) is exactly what we do for hypothesis testing.

# CI and Hypothesis Tests

- Specifically,  $p\text{-value} \leq \alpha$  (i.e., reject  $H_0$ ) if and only if  $|\text{test statistic}| \geq |\text{critical value}|$
- Suppose you are conducting a two-sided test where  $H_0: \mu = \mu_0$  and  $H_A: \mu \neq \mu_0$  for some  $\alpha$ . You can either:
  - Use CI: Check if the hypothesized true value  $\mu_0$  is outside the  $(1 - \alpha)\%$  CI (which is the same as checking if your test statistic is bigger than the critical value for the tails)
  - Do the test: Check if the probability of having observed data that is better evidence against the null (i.e., calculating the p-value) is smaller than the cutoff of  $\alpha$

# Example

You are interested in estimating  $p$ . A 95% confidence interval is (0.67, 0.71). Which of the following hypothesis tests would be statistically significant at the  $\alpha = 0.05$  level?

a)  $H_0: p = 0.6$  and  $H_A: p \neq 0.6$

b)  $H_0: p = 0.7$  and  $H_A: p \neq 0.7$

a) Reject  $H_0: p = 0.6$  since 0.6 is not in the 95% CI  
 $\Rightarrow$  Statistically significant

b) Do not reject  $H_0: p = 0.7$  since 0.7 is in the 95% CI  
 $\Rightarrow$  Not statistically significant.

# Common Pitfalls of Statistical Testing

# Common Pitfalls of Statistical Tests

- We will now go through some of the common issues that may arise when conducting statistical tests
- These should be things to look out for and be able to identify should you come across them
- Since in general statistical tests are quite simple to conduct, they are used in many fields but are sometimes used incorrectly.
  - as a researcher, it is important to be able to recognize when conclusions made or methods used may not be appropriate to the problem at hand

# Misinterpreting P-values

- The meaning of the p-value is so important and has caused a lot of debate in recent years due to its misinterpretation and misuse.
- **Incorrect:** how likely it is that the null hypothesis is true
- **Correct:** how likely it is to get the observed data (or something more extreme) given the null hypothesis were true.
  - the p-value is making reference to the sampling distribution, and therefore reflects the chances of observing your data assuming it is centred at the value of the null.
- Tips:
  - Always report your p-value, instead of whether or not it is significant
  - It reflects the strength of your evidence and has more value as a number.

# Testing does not make up for bad samples

- Even if you run the correct test and make all the right conclusions based on your test, your results can still be unreliable.
- This is because just employing the right test will not correct for any flaws that may have occurred in the data collection process.
- These flaws include:
  - sample was not randomly selected
  - if running an experiment,
    - treatments were not assigned randomly
    - a control group was not employed
- These will all lead to issues like bias, confounders and inappropriate causal conclusions

# Deciding between 1 or 2-sided tests

- Rule of thumb: always use a two-sided test
  - exception: unless you are completely sure that one direction is of no interest
- Main problem of using a one-sided test when it's not appropriate:
  - p-value will only be half that of a two-sided test (assuming your test statistic fell in a specific direction) so you will be more likely to say that your result is significant
- Never peak at your data and decide between 1-/2-sided test based on what you see
  - this is called data snooping
  - hypotheses should only be determined prior to seeing the data
  - changing your hypothesis after data collection is dishonest research practices.



# Statistical Significance $\neq$ Practical Significance

- There are many reasons why you might get a significant result for your test.
- Important: significant results are not always practically important
  - just because your p-value is small does not mean that your results are useful or practical in the real world
  - this is often because the size of the effect is not large enough to be helpful
  - e.g. 6 month weight loss program study gives a significant result when testing if average amount lost > 0lbs. But the 95% CI is [0.1, 0.15] lbs. Is this amount of weight lost practically significant?
- often very small effects are highly significant because the sample was very large.

# Why do we get small p-values?

- There are many reasons your test may produce a small p-value that can lead you to reject the null hypothesis:
  - it can happen purely by **chance** - you just happened to get the one sample that would give you this result
  - it can happen with **bad data collection** - as seen earlier, we may create bias in our data which can give you a significant p-value
  - it can happen because it was **inappropriate to use that test** - you always need to make sure that the conditions of the test are satisfied
    - so far our tests require independent observations and Normal sampling distributions
  - it can happen because the **null hypothesis is actually false** - this is what we are hoping to get

# Beware of Multiple Testing

- Be suspicious of research papers that have a large number of tests on the same data.
- Remember that the significant level as well as the p-value all represent statements about the variability of the results due to the sample collected.
  - so an  $\alpha = 0.05$  means that we would expect to get significant results from tests 5% of the time, even when the null is true!
- So doing many tests in the same study can result in significant test results purely by coincidence.
- This can't always be avoided, but should be acknowledged as a limitation

# How can we avoid these?

- Knowing what conditions are necessary for your test to be appropriate
  - for proportions, large sample sizes are needed when true  $p$  is far from 0.5
  - for means, by using the t-distribution, we can trust our results even when our sample is small - except when we have extreme skew or outliers
- Having good data practices: after you have collected data, you should
  - explore your data using plots and summary statistics
  - useful for identifying problems in the data (e.g, skew, outliers)
  - by plotting the data, it can help you decide whether your results is practically significant (if you can't see it, is it really there?)
- Unfortunately if the data collection was flawed, there is nothing you can do.
  - make sure you do everything you can to avoid bias, confounding

# Exercise

Suppose that 10 different research centres are investigating the same null hypothesis, but with different groups of subjects. Each tests the same hypothesis at the 5% level. If the null is true, what is the chance of one or more centres producing significant results (at the 5% level)?

$X$  is the # of centres with significant results.

$$X \sim \text{Bin}(n=10, p=\alpha=0.05)$$

$$P(X \geq 1) = 1 - P(X = 0) = 1 - 0.95^{10} = 0.401$$

# Practice Problems

- In this lecture, we covered all of Module 9
- Practice problems are posted

# Next Week

- Only 2 more weeks of class after today
  - Next week we will see different types of tests
  - The week after, we will cover simple linear regression