MODULE 9 EXAMPLES

TAs: Nihal Reddy
Email: nireddy@ucsd.edu
OH: Thursdays 6-7pm

Slide Credits: Kira Fleischer

PROBLEM #: KEY TOPICS FROM PROBLEM

Problem setup and description.

Question

Key notes from readings/lectures needed to answer the question

Solution: written with as much detail as we expect you to give on your homework sets

PROBLEM 1: NOTATIONS

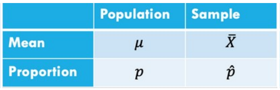
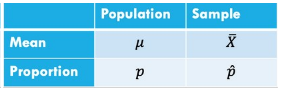
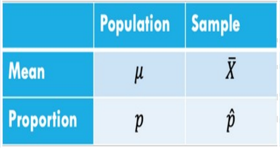
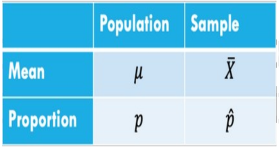
Suppose that we want to test whether the average grade of a Statistics class is 80 or not. **Discuss whether the following notations are correct in this context. If it is not correct, explain why.**

(a) μ = 80

(b) p = 80

Solution:

(a) Correct. We want to t pulation mean, which in this case repres tistics class.

	Populatio n	Sample
Mean		
Proportio n		

	Population	Sample
Mean	μ	$\bar{X}$
Proportion	p	$\hat{p}$

	Population	Sample
Mean	μ	$\bar{X}$
Proportion	p	$\hat{p}$

	Population	Sample
Mean	μ	$\bar{X}$
Proportion	p	$\hat{p}$

	Population	Sample
Mean	μ	$\bar{X}$
Proportion	p	$\hat{p}$

(b) Incorrect. We want to test a hypothesis about a population mean, which is denoted by μ. The notation p is used to represent a population proportion.

PROBLEM 2: PROBABILITY OF ERRORS

Suppose that we conduct a hypothesis test at significance level α = 0.01. **If now α is changed to be 0.10, how does the probability of Type I and Type II error change? Do they increase, decrease, or stay the same?**

P(Type I error) = P(rejecting null hypothesis when null is true) = α

P(Type II error) = P(failing to reject null when alternative is true) : typically has an inverse relationship with P(Type I error)

Solution: Since P(Type I error) = α, if α increases from 0.01 to be 0.10 now, the probability of Type I error would also increase. This would decrease the probability of Type II error.

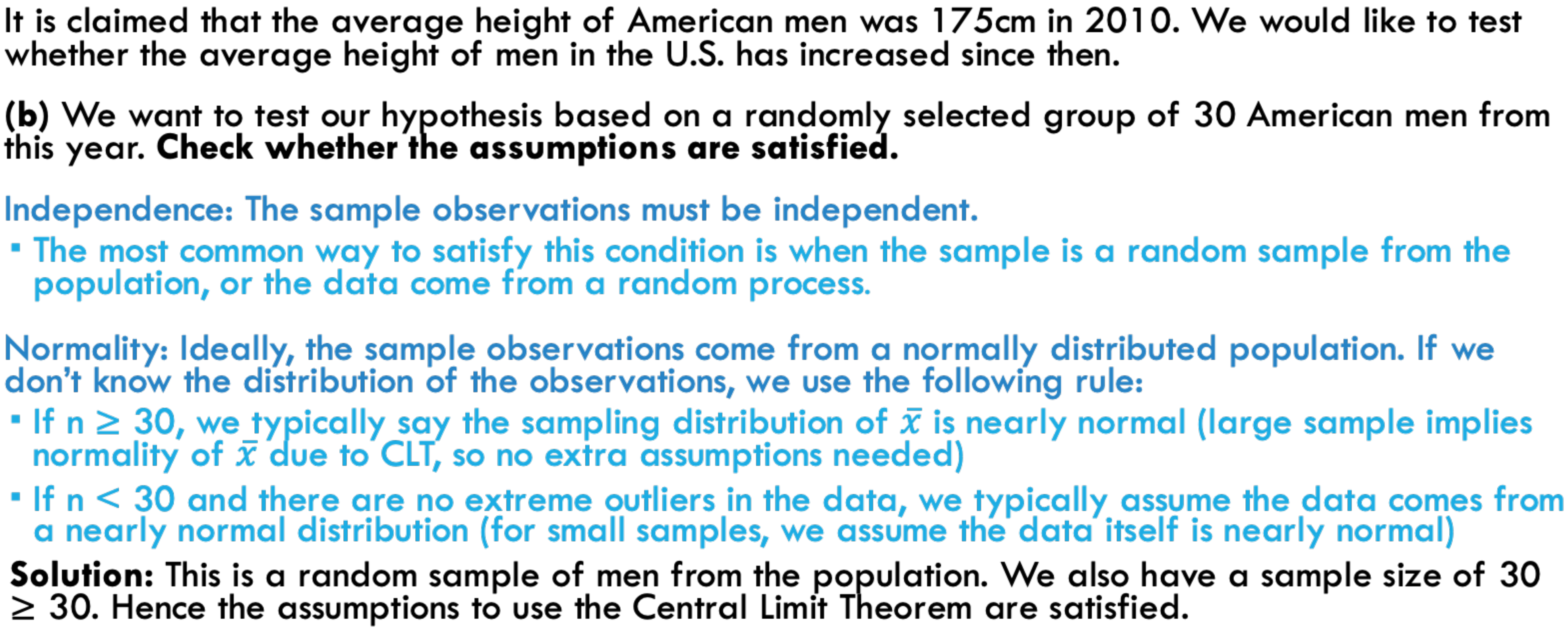
PROBLEM 3: SETTING UP A T-TEST

It is claimed that the average height of American men was 175cm in 2010. We would like to test whether the average height of men in the U.S. has increased since then.

(a) Please write down the appropriate null and alternative hypotheses.

Solution: Let μ be the average height of American men. We have H_0: $\mu = 175$ vs H_A: $\mu > 175$.

PROBLEM 3: SETTING UP A T-TEST



It is claimed that the average height of American men was 175cm in 2010. We would like to test whether the average height of men in the U.S. has increased since then.

(b) We want to test our hypothesis based on a randomly selected group of 30 American men from this year. **Check whether the assumptions are satisfied.**

Independence: The sample observations must be independent.

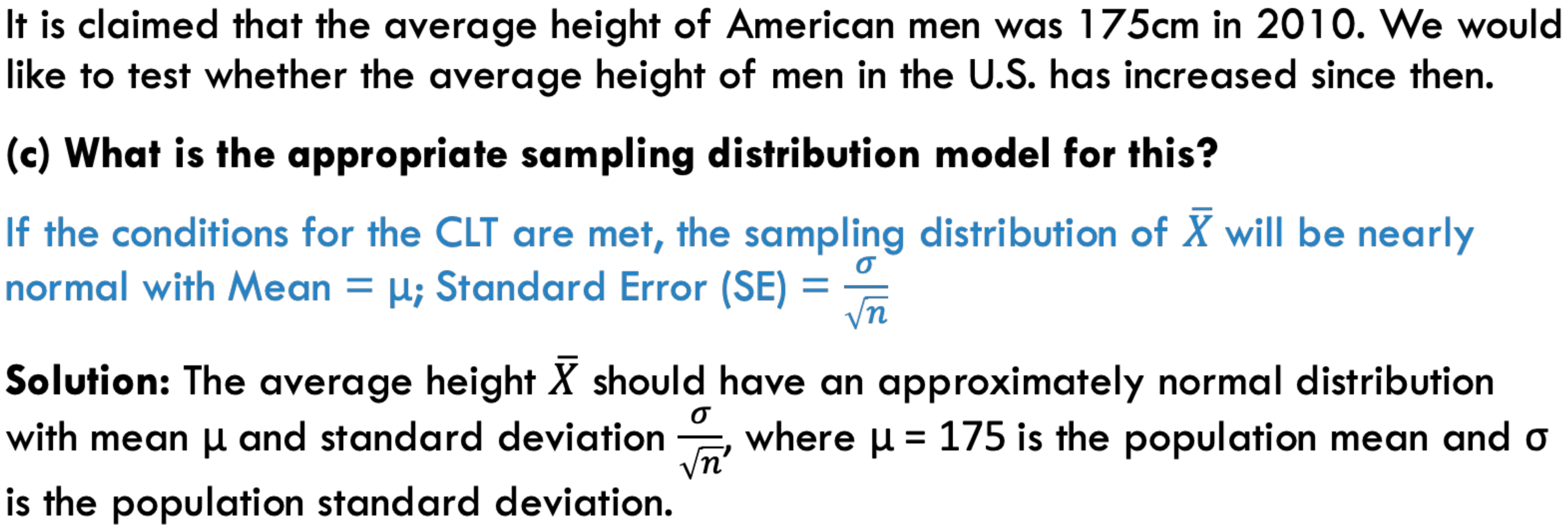
- The most common way to satisfy this condition is when the sample is a random sample from the population, or the data come from a random process.

Normality: Ideally, the sample observations come from a normally distributed population. If we don't know the distribution of the observations, we use the following rule:

- If n ≥ 30, we typically say the sampling distribution of $\bar{x}$ is nearly normal (large sample implies normality of $\bar{x}$ due to CLT, so no extra assumptions needed)
- If n < 30 and there are no extreme outliers in the data, we typically assume the data comes from a nearly normal distribution (for small samples, we assume the data itself is nearly normal)

Solution: This is a random sample of men from the population. We also have a sample size of 30 ≥ 30. Hence the assumptions to use the Central Limit Theorem are satisfied.

PROBLEM 3: SETTING UP A T-TEST



It is claimed that the average height of American men was 175cm in 2010. We would like to test whether the average height of men in the U.S. has increased since then.

(c) What is the appropriate sampling distribution model for this?

If the conditions for the CLT are met, the sampling distribution of $\bar{X}$ will be nearly normal with Mean = μ; Standard Error (SE) = $\frac{\sigma}{\sqrt{n}}$

Solution: The average height $\bar{X}$ should have an approximately normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, where μ = 175 is the population mean and σ is the population standard deviation.

PROBLEM 4: PERFORMING A T-TEST

It is claimed that the average height of American men was 175cm in 2010. We would like to test whether the average height of men in the U.S. has increased since then.

Now, if the heights of men in our sample has a mean of 176cm and a standard deviation of 4.5cm,

(a) Compute the t test statistic and p-value.

$t = \frac{\bar{X}-\mu}{SE} = \frac{\bar{X}-\mu}{s/\sqrt{n}}$; p-value= $P(T_{df} > t)$

Solution: $t = \frac{176-175}{4.5/\sqrt{30}} = 1.217$. Since df=30-1=29, the p-value is $P(T_{29} > 1.217)$.

Calculate the p-value using technology or a table to get p-value = 0.117.

Ex: R-code pt(1.217, 29, lower.tail = FALSE)

PROBLEM 4: PERFORMING A T-TEST

It is claimed that the average height of American men was 175cm in 2010. We would like to test whether the average height of men in the U.S. has increased since then.

Now, if the heights of men in our sample has a mean of 176cm and a standard deviation of 4.5cm,

(b) Explain what this p-value means.

The p-value represents the probability that, if the null hypothesis is true, you would get the same (or more extreme) sample statistic that you did, if you were to repeat the sampling procedure many times.

Solution: This means that, if the mean heights of men had not changed since 2010, then the average heights of men in a random sample of size 30 would be at least as large as it was in our sample approximately 11.7 percent of the time.

PROBLEM 4: PERFORMING A T-TEST

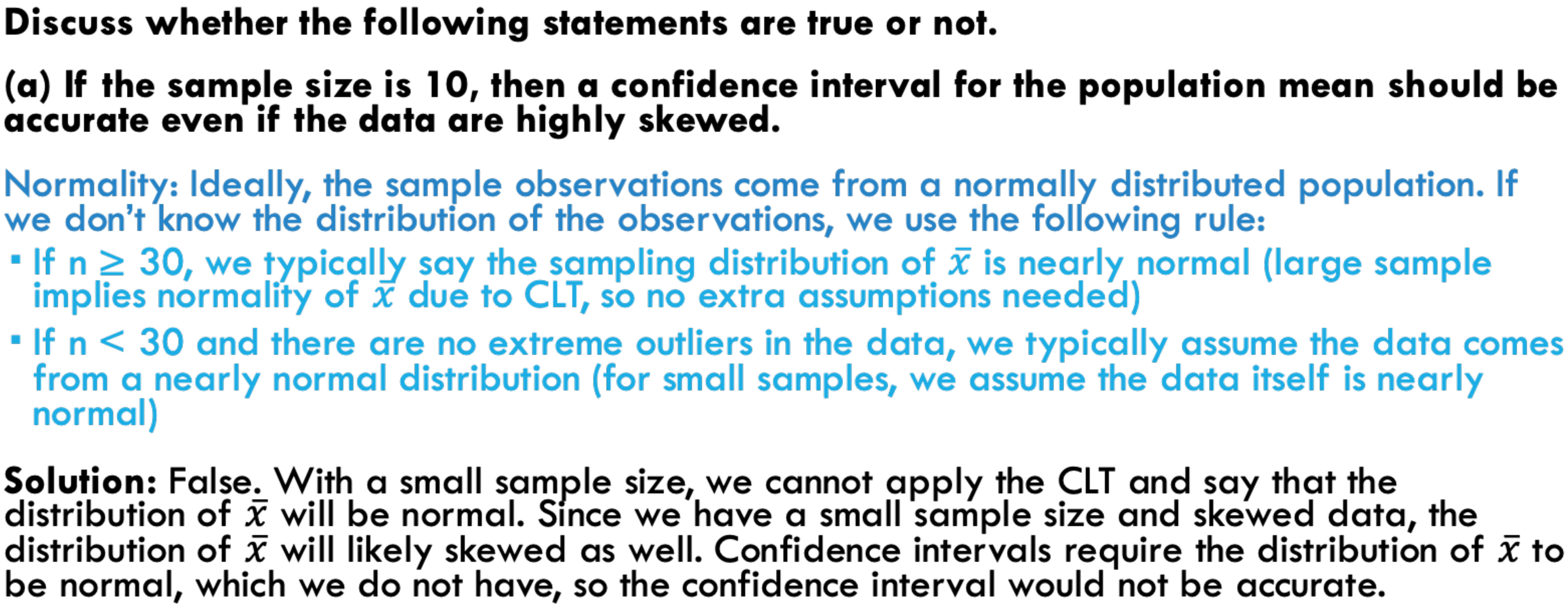
It is claimed that the average height of American men was 175cm in 2010. We would like to test whether the average height of men in the U.S. has increased since then.

Now, if the heights of men in our sample has a mean of 176cm and a standard deviation of 4.5cm,

(c) What is your final conclusion at significance level α = 0.10?

Solution: Since p-value=0.117 > α =0.10, we conclude that we do not have enough evidence to conclude that the mean heights of men in the U.S. has increased since 2010.

PROBLEM 5: NORMALITY ASSUMPTI0N FOR MODELING SAMPLE MEAN



Discuss whether the following statements are true or not.

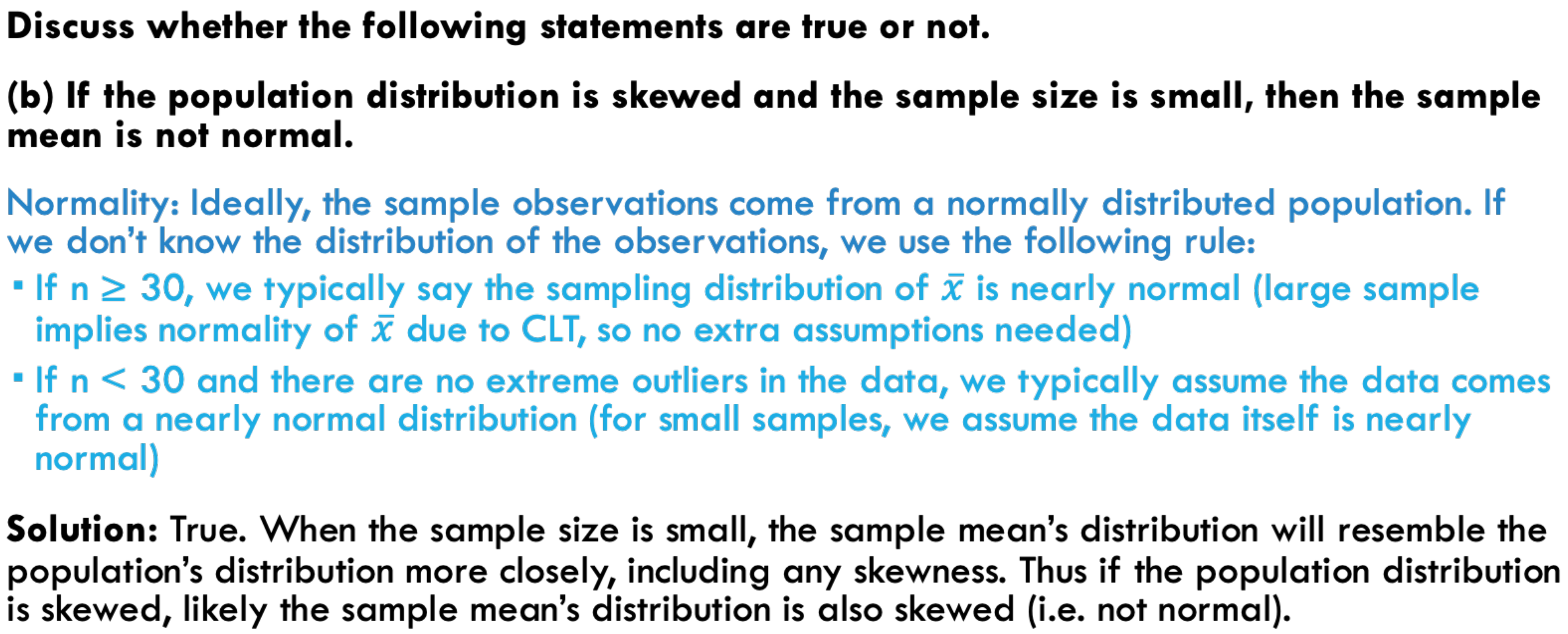
(a) If the sample size is 10, then a confidence interval for the population mean should be accurate even if the data are highly skewed.

Normality: Ideally, the sample observations come from a normally distributed population. If we don't know the distribution of the observations, we use the following rule:

- If n ≥ 30, we typically say the sampling distribution of $\bar{x}$ is nearly normal (large sample implies normality of $\bar{x}$ due to CLT, so no extra assumptions needed)
- If n < 30 and there are no extreme outliers in the data, we typically assume the data comes from a nearly normal distribution (for small samples, we assume the data itself is nearly normal)

Solution: False. With a small sample size, we cannot apply the CLT and say that the distribution of $\bar{x}$ will be normal. Since we have a small sample size and skewed data, the distribution of $\bar{x}$ will likely skewed as well. Confidence intervals require the distribution of $\bar{x}$ to be normal, which we do not have, so the confidence interval would not be accurate.

PROBLEM 5: NORMALITY ASSUMPTI0N FOR MODELING SAMPLE MEAN



Discuss whether the following statements are true or not.

(b) If the population distribution is skewed and the sample size is small, then the sample mean is not normal.

Normality: Ideally, the sample observations come from a normally distributed population. If we don't know the distribution of the observations, we use the following rule:

- If n ≥ 30, we typically say the sampling distribution of $\bar{x}$ is nearly normal (large sample implies normality of $\bar{x}$ due to CLT, so no extra assumptions needed)
- If n < 30 and there are no extreme outliers in the data, we typically assume the data comes from a nearly normal distribution (for small samples, we assume the data itself is nearly normal)

Solution: True. When the sample size is small, the sample mean's distribution will resemble the population's distribution more closely, including any skewness. Thus if the population distribution is skewed, likely the sample mean's distribution is also skewed (i.e. not normal).

PROBLEM 6: PAIRED DATA

Discuss whether the data associated with the following scenarios are paired data or independent data.

(a) We would like to know the difference in salaries of men and women. We survey 200 men and 200 women randomly and record their annual salaries.

Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Solution: This would be independent data since the salaries of men and women are unrelated with one another.

PROBLEM 6: PAIRED DATA

Discuss whether the data associated with the following scenarios are paired data or independent data.

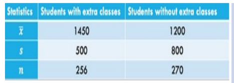
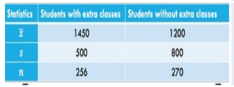
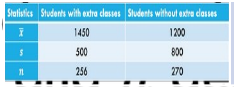
(b) We would like to know whether the pandemic has influenced the amount of money that hotels make annually. A random sample of 50 hotels are chosen in the U.S. and they reported how much they made in 2019 and 2020.

Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Solution: This would be paired data since how much hotels make in 2019 and 2020 are related with one another.

PROBLEM 7: CONFIDENCE INTERVAL FOR DIFFERENCE IN MEANS

In a certain test, researchers are interested in whether taking extra classes would improve students' performances on a standardized test. The following data is collected from a sample of students:

Statistics	Students with extra classes	Students without extra classes
	1450	1200
	500	800
	256	270

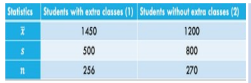
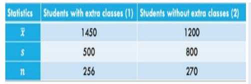
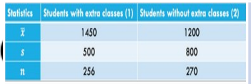
Statistics	Students with extra classes	Students without extra classes
$\bar{x}$	1450	1200
s	500	800
n	256	270

Here $\bar{x}$ of the sample, and n denotes the sample sizes.

(a) Construct a 90% confidence interval to estimate the mean difference, and

(b) interpret the confidence interval.

PROBLEM 7: CONFIDENCE INTERVAL FOR DIFFERENCE IN MEANS

Statistics	Students with extra classes (1)	Students without extra classes (2)
	1450	1200
	500	800
	256	270

Statistics	Students with extra classes (1)	Students without extra classes (2)
$\bar{x}$	1450	1200
s	500	800
n	256	270

Statistics	Students with extra classes (1)	Students without extra classes (2)
$\bar{x}$	1450	1200
s	500	800
n	256	270

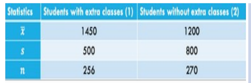
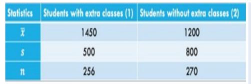
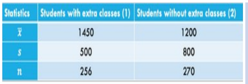
Statistics	Students with extra classes (1)	Students without extra classes (2)
$\bar{x}$	1450	1200
s	500	800
n	256	270

(a) (e

Solution: Let $\bar{x}_1$ and s_1 be the mean and standard deviation of the exam scores of those students with extra classes and $\bar{x}_2$ and s_2 be the mean and standard deviation of the exam scores of those students with no extra classes.

Let n_1 and n_2 be the number of students in each sample, respectively. Let μ_1 and μ_2 be the true mean exam scores that would be achieved by students who take extra classes and those who do not, respectively.

PROBLEM 7: CONFIDENCE INTERVAL FOR DIFFERENCE IN MEANS

Statistics	Students with extra classes (1)	Students without extra classes (2)
	1450	1200
	500	800
	256	270

Statistics	Students with extra classes (1)	Students without extra classes (2)
$\bar{x}$	1450	1200
s	500	800
n	256	270

Statistics	Students with extra classes (1)	Students without extra classes (2)
$\bar{x}$	1450	1200
s	500	800
n	256	270

Statistics	Students with extra classes (1)	Students without extra classes (2)
$\bar{x}$	1450	1200
s	500	800
n	256	270

(a) Con

CI for difference in means: $\bar{x}_1 - \bar{x}_2 \pm t^*_{df} \times SE$; $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$; df $= \min\{n_1, n_2\} - 1$

Solution (continued): df $= \min\{n_1, n_2\} - 1 = 256 - 1 = 255$

Critical t-value: $t^*_{255} = 1.651$ (using R code: qt(0.05, 255, lower.tail = FALSE))

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{500^2}{256} + \frac{800^2}{270}} = 57.852$$

$$MOE = t^*_{df} \times SE = 1.651 \times 57.852 = 95.51$$

Confidence interval:

$$(1450 - 1200 - 95.51, 1450 - 1200 + 95.51) = (154.49, 345.51)$$

PROBLEM 7: CONFIDENCE INTERVAL FOR DIFFERENCE IN MEANS

(b) interpret the confidence interval.

Solution: We are 90% confident that the average exam scores of students who took extra classes is between 154.49 and 345.51 points higher than the average score of students who did not take extra classes.

PROBLEM 8: HYPOTHESIS TEST FOR DIFFERENCE IN MEANS

With *almost* the same problem setup, now we want to do the following:

Conduct a hypothesis test and conclude whether the extra classes are improving the students' exam scores at significance level α = 0.10.

Solution: Let $\bar{x}_1$ and s_1 be the mean and standard deviation of the exam scores of those students with extra classes and $\bar{x}_2$ and s_2 be the mean and standard deviation of the exam scores of those students with no extra classes.

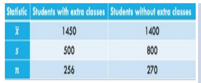
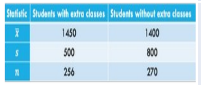
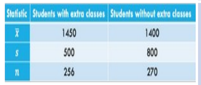
Let n_1 and n_2 be the number of students in each sample, respectively. Let μ_1 and μ_2 be the true mean exam scores that would be achieved by students who take extra classes and those who do not, respectively.

We test the hypotheses $H_0: \mu_1 = \mu_2$ against $H_A: \mu_1 > \mu_2$

PROBLEM 8: HYPOTHESIS TEST FOR DIFFERENCE IN MEANS

Solution (cont.): The test statistic is:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1450-1400}{\sqrt{\frac{500^2}{256} + \frac{800^2}{270}}} = 0.864$$

Statistic	Students with extra classes	Students without extra classes
	1450	1400
	500	800
	256	270

Statistic	Students with extra classes	Students without extra classes
$\bar{x}$	1450	1400
s	500	800
n	256	270

Statistic	Students with extra classes	Students without extra classes
$\bar{x}$	1450	1400
s	500	800
n	256	270

Statistic	Students with extra classes	Students without extra classes
$\bar{x}$	1450	1400
s	500	800
n	256	270

If H_0 were to be true, T would follow a t-distribution with degrees of freedom 256−1=255.

We calculate the p-value to be $P(T_{255} > 0.864) = 0.194$.

At significance level α = 0.10, we have that 0.194 > 0.10. Thus, we conclude that we do not have strong evidence that students who take extra classes would perform better than those who don’t.

PROBLEM 9: TYPE I AND II ERRORS

A hypothesis test was conducted to find out whether the average test scores in a class is higher than the national average of 1200. A random sample of 50 people was found to have an average score of 1210. The p-value is calculated to be 0.045, hence the null hypothesis is rejected. The conclusion is that their mean score is higher than that of the nation. In reality, in the population of all the students in the class, the average is 1205.

Did this study make a Type I error, Type II error, or no errors at all?

Solution: This study did not make any errors since it rejected the null hypothesis and made the conclusion that $\mu > 1200$ while the true population mean $1205 > 1200$.

PROBLEM 10: POWER OF HYPOTHESIS TEST

We are interested in the statistical power of a hypothesis test associated with whether taking extra classes would increase students' exam scores. Let μ_1 denote the mean exam scores of students who don't take extra classes; and let μ_2 denote the mean exam scores of students who do. The null and alternative hypotheses, associated with our power calculation, are

$H_0: \mu_1 - \mu_2 = 0$ vs $H_A: \mu_1 - \mu_2 = -100.$

Suppose that for the test we take sample sizes to be $n_1 = 270$; $n_2 = 256$. It is also estimated that $s_1 = 800$, $s_2 = 500$. **Calculate the statistical power of this test at significance level α = 0.05.**

Power: probability of correctly rejecting the null hypothesis when it is false. For testing a difference in means, power is the probability of detecting the "effect" between the two population groups.

$= P(\bar{x}_1 - \bar{x}_2$ lies within H_0's rejection region | H_0 is false)

$= P(\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)_0}{SE}$ outside of $[-z^*, z^*]| \ H_A$ is true), $z^* = 1.96$ for α = 0.05

$= P(\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)_0}{SE} < -z^*$ or $\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)_0}{SE} > z^* \mid H_A$ is true)

$= \mathrm{P}([\bar{x}_1 - \bar{x}_2 < (-z^* \times SE) + (\mu_1 - \mu_2)_0]$ or $[\bar{x}_1 - \bar{x}_2 > (z^* \times SE) + (\mu_1 - \mu_2)_0]| \ H_A$ is true)

$= P(\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)_A}{SE} < \frac{(-z^* \times SE) + (\mu_1 - \mu_2)_0 - (\mu_1 - \mu_2)_A}{SE}$ or $\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)_A}{SE} > \frac{(z^* \times SE) + (\mu_1 - \mu_2)_0 - (\mu_1 - \mu_2)_A}{SE})$

PROBLEM 10: POWER OF HYPOTHESIS TEST

Solution: First, we have that SE of $\bar{x}_1 - \bar{x}_2 = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{500^2}{256} + \frac{800^2}{270}} = 57.852$ (same as problem 7a).

Under H_0, we can model $\bar{x}_1 - \bar{x}_2$ as normal with mean 0 and SE = 57.852.

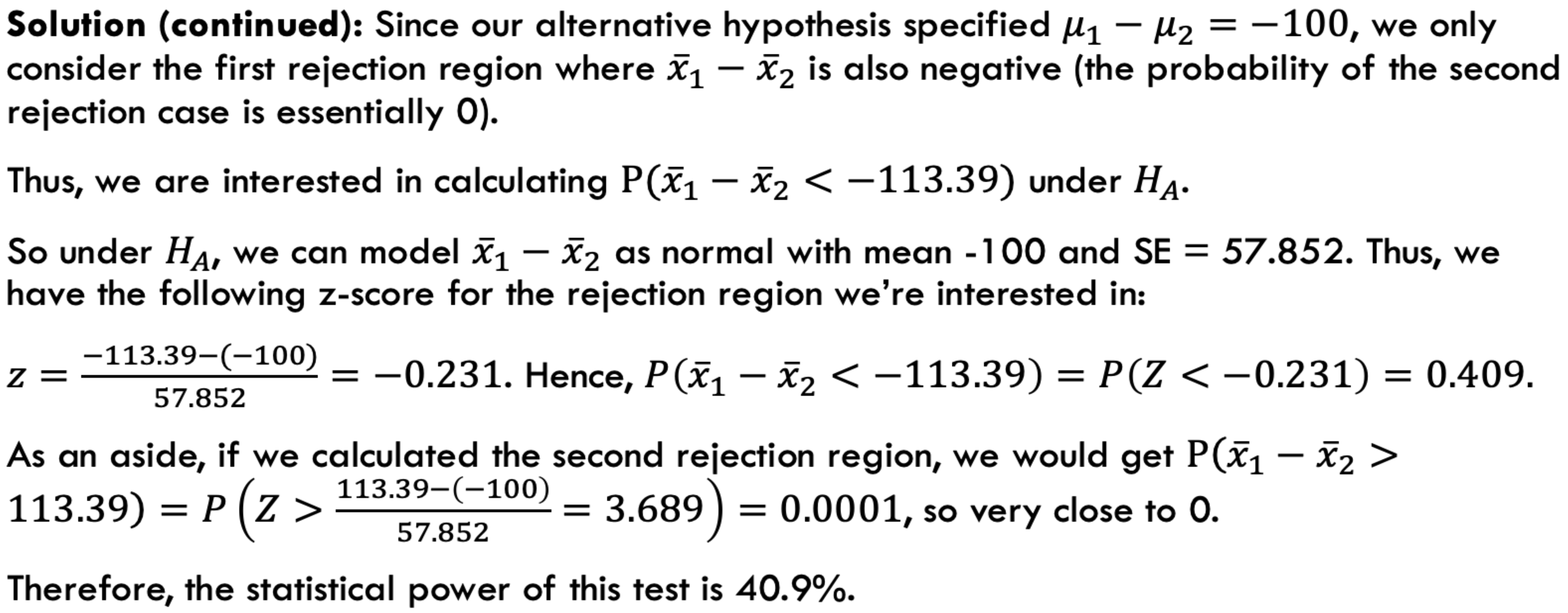
We know that rejecting H_0 occurs when $\bar{x}_1 - \bar{x}_2$ lies in the rejection region, i.e., when

$\bar{x}_1 - \bar{x}_2 < (-z^* \times SE) + (\mu_1 - \mu_2)_0$ or $\bar{x}_1 - \bar{x}_2 > (z^* \times SE) + (\mu_1 - \mu_2)_0 \Longrightarrow$

$\bar{x}_1 - \bar{x}_2 < -1.96 \times 57.852 + 0 = -113.39$, or

$\bar{x}_1 - \bar{x}_2 > 1.96 \times 57.852 + 0 = 113.39$

PROBLEM 10: POWER OF HYPOTHESIS TEST



Solution (continued): Since our alternative hypothesis specified $\mu_1 - \mu_2 = -100$, we only consider the first rejection region where $\bar{x}_1 - \bar{x}_2$ is also negative (the probability of the second rejection case is essentially 0).

Thus, we are interested in calculating $\mathrm{P}(\bar{x}_1 - \bar{x}_2 < -113.39)$ under H_A.

So under H_A, we can model $\bar{x}_1 - \bar{x}_2$ as normal with mean -100 and SE = 57.852. Thus, we have the following z-score for the rejection region we're interested in:

$z = \frac{-113.39-(-100)}{57.852} = -0.231$. Hence, $P(\bar{x}_1 - \bar{x}_2 < -113.39) = P(Z < -0.231) = 0.409.$

As an aside, if we calculated the second rejection region, we would get $\mathrm{P}(\bar{x}_1 - \bar{x}_2 > 113.39) = P\left(Z > \frac{113.39-(-100)}{57.852} = 3.689\right) = 0.0001$, so very close to 0.

Therefore, the statistical power of this test is 40.9%.