Chapter 4B

Inference for one sample

Statistical Inference Methods

- Statistical Inference: Drawing conclusions about a population from sample data.

- Methods
  - Point Estimation – Using a sample statistic to estimate a parameter
  - Confidence Intervals – supplements an estimate of a parameter with an indication of its variability
  - Hypothesis Tests– assesses evidence for a claim about a parameter by comparing it with observed data

Parameter	Measure	Statistic
μ	Mean of a single population	$\bar{X}$
σ^2	Variance of a single population	S^2
σ	Standard deviation of a single population	S
p	Proportion of a single population	$\hat{p}$
$\mu_1 - \mu_2$	Difference in means of two populations	$\bar{X}_1 - \bar{X}_2$
$p_1 - p_2$	Difference in proportions of two populations	$\hat{p}_1 - \hat{p}_2$

Inference for a single mean (μ)

- We will look at confidence intervals and hypothesis tests when μ is our parameter of interest.

- In demonstrating those ideas earlier in the chapter we assumed we knew σ, but this is unrealistic.

Inference about a Mean

- **σ known**– CLT for means allows us to use Z procedures if…
  - 1.) If our Population ~ N(μ,σ)
    - Then $\bar{X}$ ~ N(μ,$\frac{\sigma}{\sqrt{n}}$) for any n.
  - 2.) If our Population ~ ?(μ,σ) aka non-normal
    - Then $\bar{X}$ ~ N(μ,$\frac{\sigma}{\sqrt{n}}$) if n > 30

- **σ unknown** – We will most likely use T procedures with n-1 D.o.F. if…
  - Our sample appears to come from a normal Population
    - We see no skewness (Check Histogram)
    - We see no outliers (Check Boxplot)
    - Can check a Q-Q plot for both
  - For n > 30 we can be more flexible with assumptions.

Building a CI (σ known)

In words:

estimate $\pm$ (critical value) x (standard error)

Using: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

Then: $\bar{X} \pm Z^*(\frac{\sigma}{\sqrt{n}})$

Estimate σ

- Estimate the population standard deviation σ with the sample standard deviation, s.
- s is known to be a good estimate of σ.
- s is a statistic calculated from the sample data.

Confidence Interval

In words:

estimate $\pm$ (critical value) x (standard error)

We replace σ with s when we estimate the standard error. Thus, our confidence interval is

$$\bar{x} \pm t^{*}\left(\frac{s}{\sqrt{n}}\right)$$ where df = n – 1

The t distribution table

- The T table gives t critical values for t distributions with specific degrees of freedom.

- Each column is labeled as an upper tailed probability based on α and as a t* critical value for a specific confidence interval.

- See Table 4 (posted in Canvas).

T-Table Critical Value Example

- What Critical Value would we use for a 95% CI from a sample of size 17?

- Don't forget Degrees of freedom!
  - n=17 so df=n-1=17-1=16

- We would use: 2.120

Example

- Estimate the average height of adult males in Virginia.

- We will take a sample of size 24.

- Calculate sample mean and standard deviation.

Sample of Heights (in inches)

- Data in Canvas

- Sample statistics:

Column	n	Mean	Std. dev.
Heights	24	68.666667	2.8386566

- Population parameters are unknown so we should use t

Heights
67
70
68
70
71
70
68
68
69
66
73
73
68
65
67
74
73
68
65
66
67
63
71
68

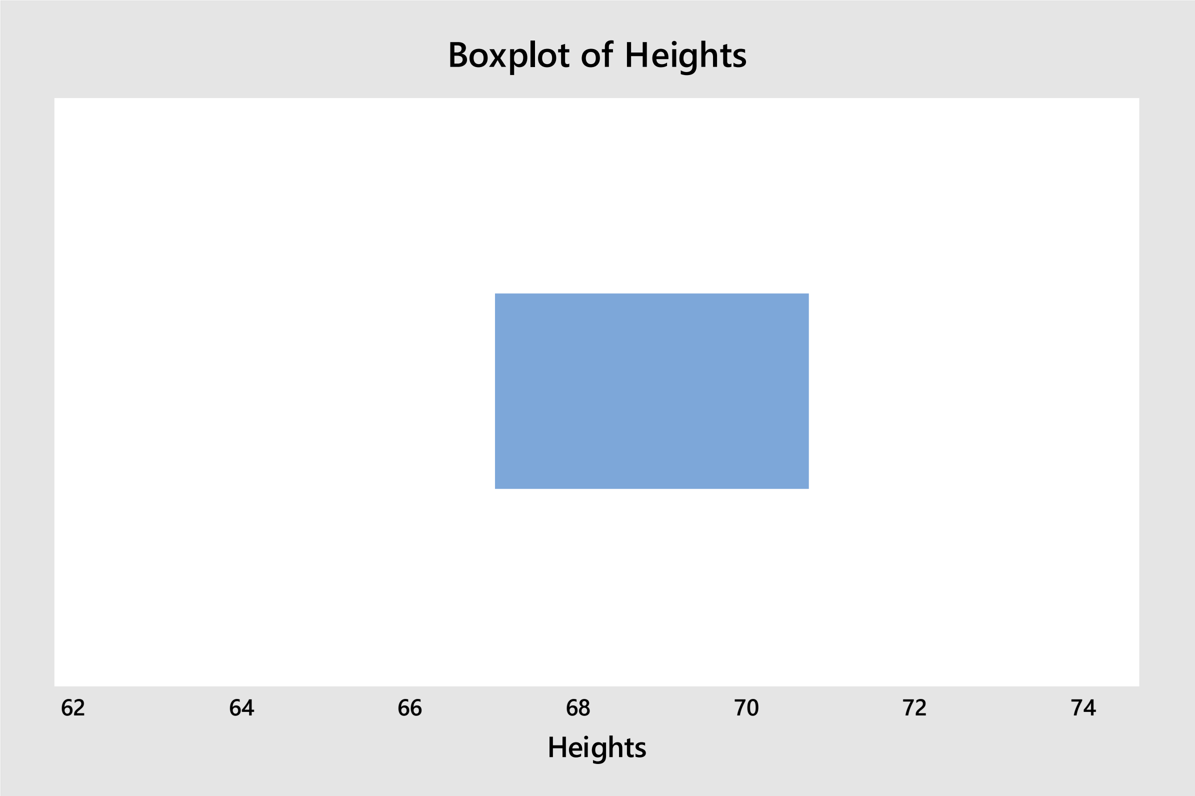
Example

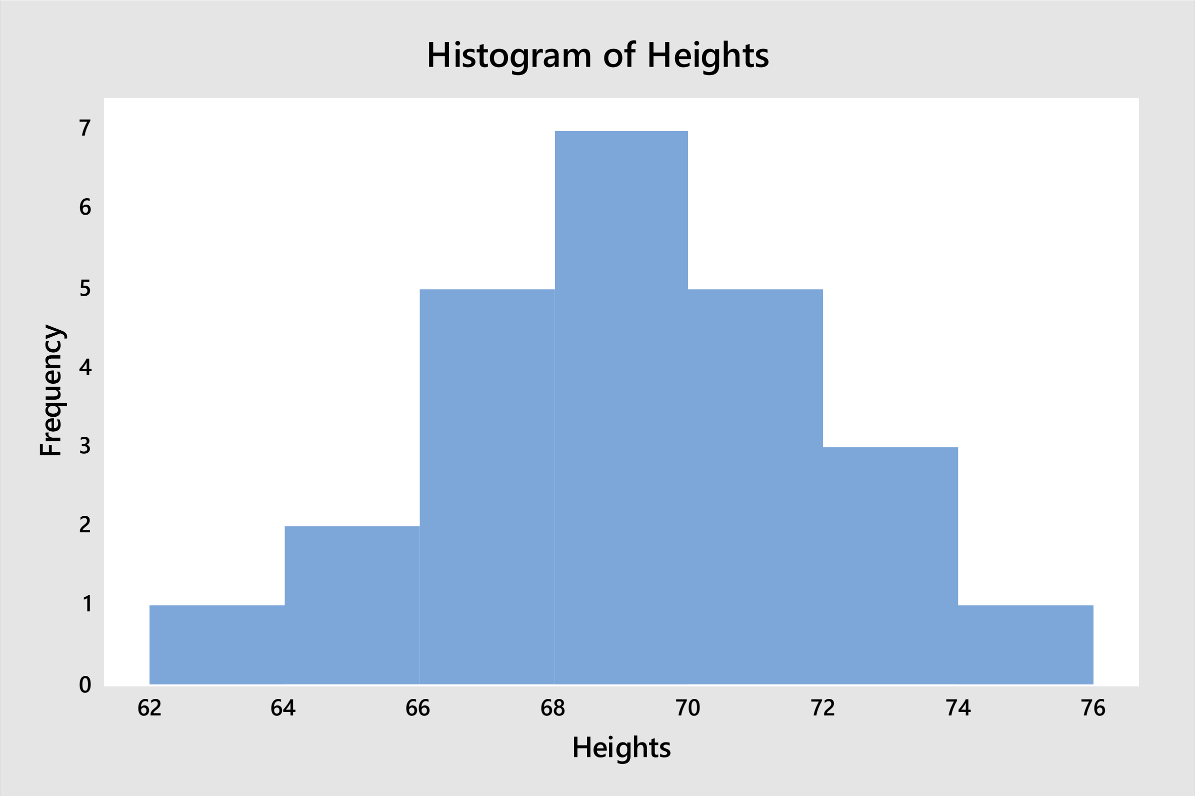
- Construct a 95% confidence interval to estimate the mean height of adult males in Virginia.

- Remember, t inference methods work well if the population is not normal as long as our data from our sample
  - Do not contain outliers
  - Is not extremely skewed

- Check these conditions by graphing the sample data (use a box plot and a histogram).

Graphs





Conditions of normality and no outliers hold

Solution

- Remember, n = 24
- We need to obtain t* from Table 4.
- Read down the first column to 23 degrees of freedom (24 – 1).
- Read across the columns until you are under the 95% confidence level.
- t* = 2.069

Solution

Use: $\bar{x} \pm t^* \left(\frac{s}{\sqrt{n}} \right)$

$$68.67 \pm 2.069 \left(\frac{2.84}{\sqrt{24}} \right)$$

$$68.67 \pm 1.20$$

$$(67.47,\ 69.87)$$

- Interpretation: We still can say that we are 95% confident that this interval captures the unknown population mean height of adult males in VA.

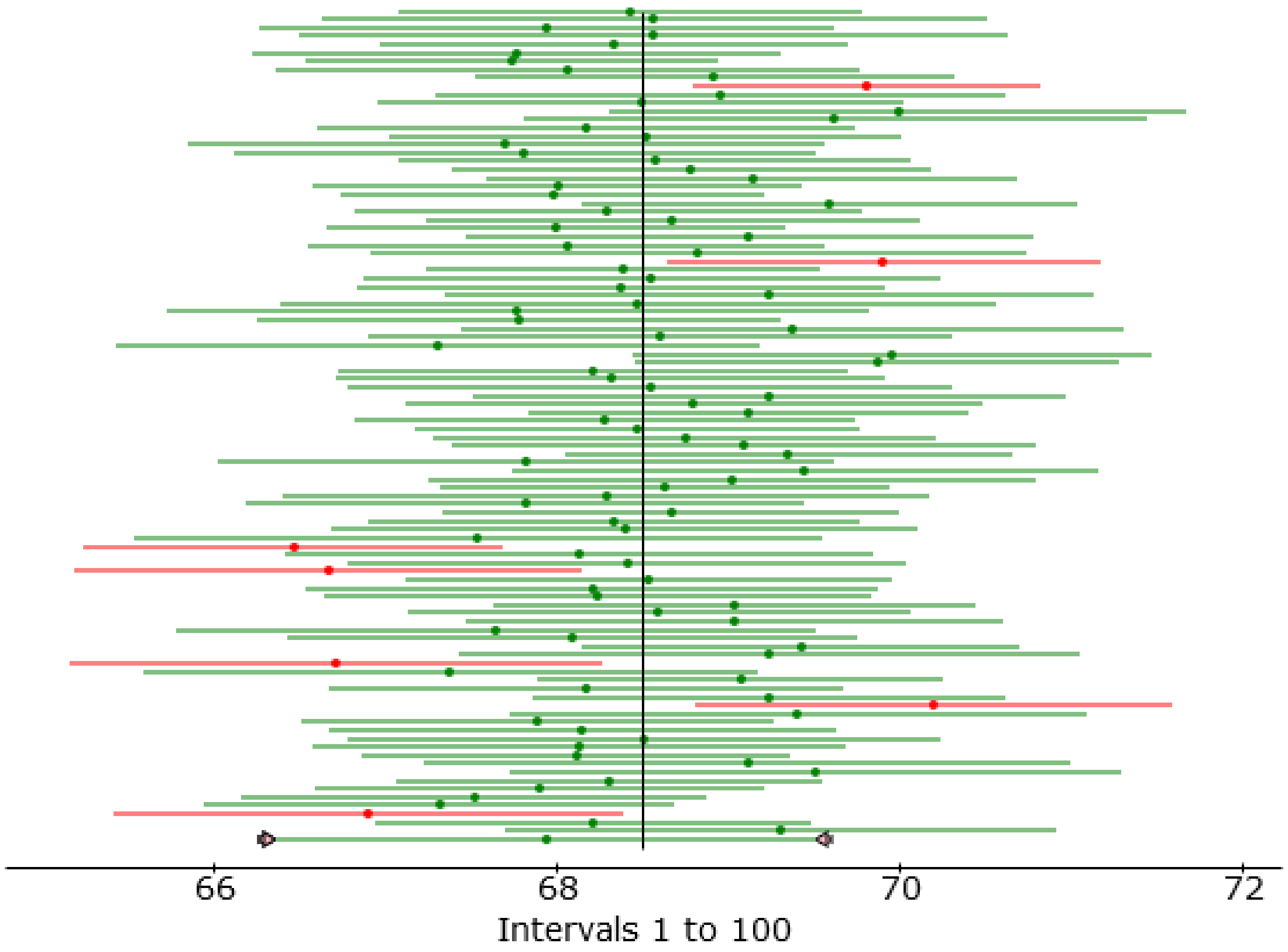
Correct Interpretation

- We can say that if we took many, many samples and constructed many, many confidence intervals, 95% of those confidence intervals will capture the true unknown population mean.

- To check this, we will simulate the process of constructing many confidence intervals.

Confidence intervals for the mean using normal values with mean(μ)=...
Sample size=24

CI Level	Containing μ	Total	Proportion
0.95	1046	1100	0.9509



Comparison

- Suppose I took two samples, both n=24, and created 2 confidence intervals

Sample 1

$$67.944 \pm 2.069\left(\frac{3.598}{\sqrt{24}}\right)$$

$$67.944 \pm 1.520$$

Larger Width

Sample 2

$$66.906 \pm 2.069\left(\frac{3.406}{\sqrt{24}}\right)$$

$$66.906 \pm 1.438$$

Smaller Width

- T Confidence intervals that use the same CL and n may have different values of s, and margins of error and widths.

Z vs T intervals

- Using a t* value creates a larger confidence interval to account for the fact that the sample standard deviations vary from sample to sample.

- If we used the z* value, a smaller amount than 95% of the intervals would capture the population mean.

Sample Size with the *t* Confidence Interval

- Determining sample size (n) is a trial-and error process since n appears in two factors.

$$E = t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

- Procedure to find a desired E:
  - Start with the $z_{\alpha/2}$ value and a guess for σ
  - Solve for n.
  - Gather sample observations & calculate s.
  - Determine $t_{\alpha/2,n-1}$ and evaluate E.
  - Gather more observations to further reduce E and repeat.

- Basically, this process gets ugly and we won't do it much by hand
- They do make software and charts capable of this

Testing a Claim About a Mean

As in the previous chapter, a test of hypotheses requires a few steps:

1) State the appropriate hypotheses
2) State the appropriate test statistic
3) State the Critical Region
4) Conduct the experiment and calculate the test statistic
5) Draw your conclusion

We still need to follow the same logic we did when creating confidence intervals for means [Z (σ known) vs. T (σ unknown)]

Test Statistic: Hypothesis Test of the population mean (σ known)

- We know the test statistic should have the structure:

$$z = \frac{\text{observed value} - \text{null value}}{SE}$$

- So if we know σ, our test statistic should be:

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

From this we substitute in s for σ and must calculate a T test statistic

One-sample t-test

- **Conditions:** SRS of size *n* from a Normal population (or a sample size n $\geq$ 30), check graphs of sample data.

- **Hypotheses:** H_0: $\mu = \mu_0$ *versus* a one- or two-sided H_a

- **Test statistic:**

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

- **P-value:**

$P(T \leq t)$	for H_a: $\mu < \mu_0$
$P(T \geq t)$	for H_a: $\mu > \mu_0$
$2P(T \geq \|t\|)$	for H_a: $\mu \neq \mu_0$

where T is $t(n - 1)$

Finding a P-value for a T-test

- We know how to find a P-value when using the Z distribution but the T presents some new challenges

- The Z table had many values and probabilities listed, but the T table only has commonly used Critical Values

- Three options:
  - Technology (find exact P-val)
  - Opt for the Critical Value method
  - Estimate a range for the p-value using the T table

Example: Gas Prices

- Information (data is from a few years back) from a gas tracking website stated the average price in the country for a gallon of regular gasoline is $3.50. You take a random sample of 21 gas stations in Virginia and want to see if the average in our state is actually lower than that (data are shown later). Assume $\alpha = 0.01$.

- To use the one sample z test, we need to know the population standard deviation σ.

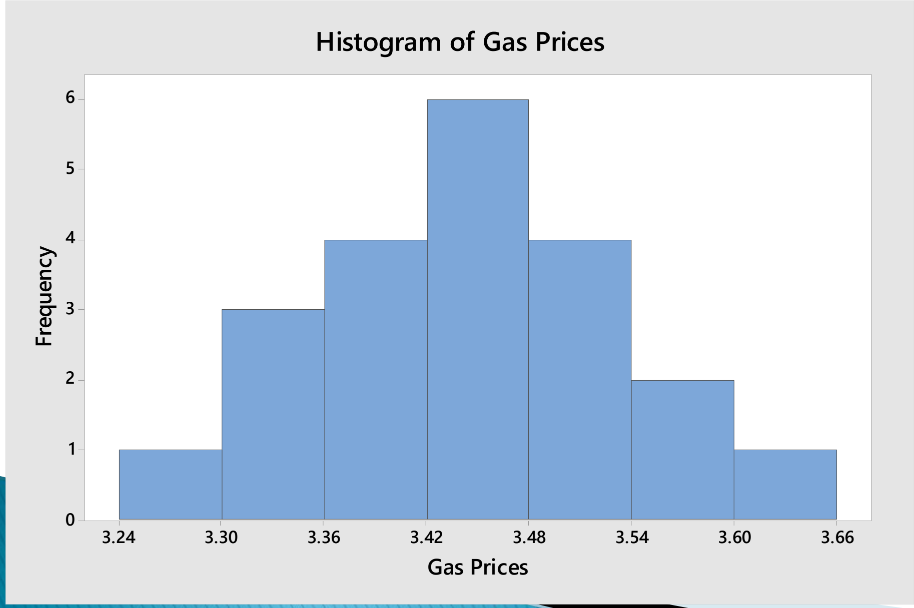
- Estimating σ with *s* introduces additional random variability so we will need to use T

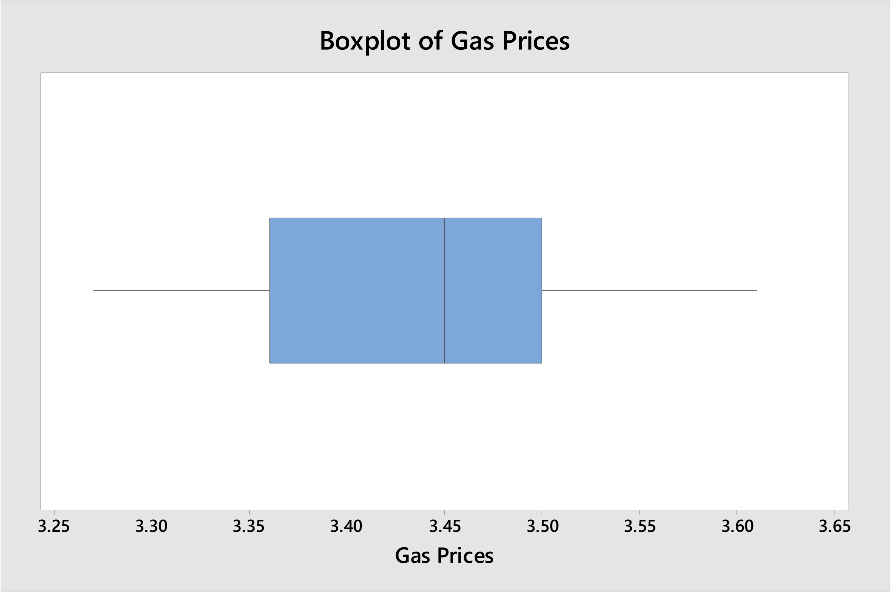
Example: Gas Prices

- Data in Canvas

Column	n	Mean	Std. dev.
Gas Prices	21	3.4381	0.0903

- Conditions:
  - SRS of size 21.
  - Normality can be checked by looking at a histogram.
  - Outliers can be checked by using a boxplot





Gas Prices
3.27
3.31
3.33
3.35
3.36
3.36
3.38
3.39
3.42
3.44
3.45
3.46
3.47
3.47
3.48
3.49
3.51
3.52
3.56
3.57
3.61

Full Solution

1) State the hypotheses.

H_0: $\mu = 3.50$

H_a: $\mu < 3.50$

We see that this test is left-tailed

2) We do not know σ, so should use a T Test Statistic w/

Degrees of freedom: df = 21 – 1 = 20

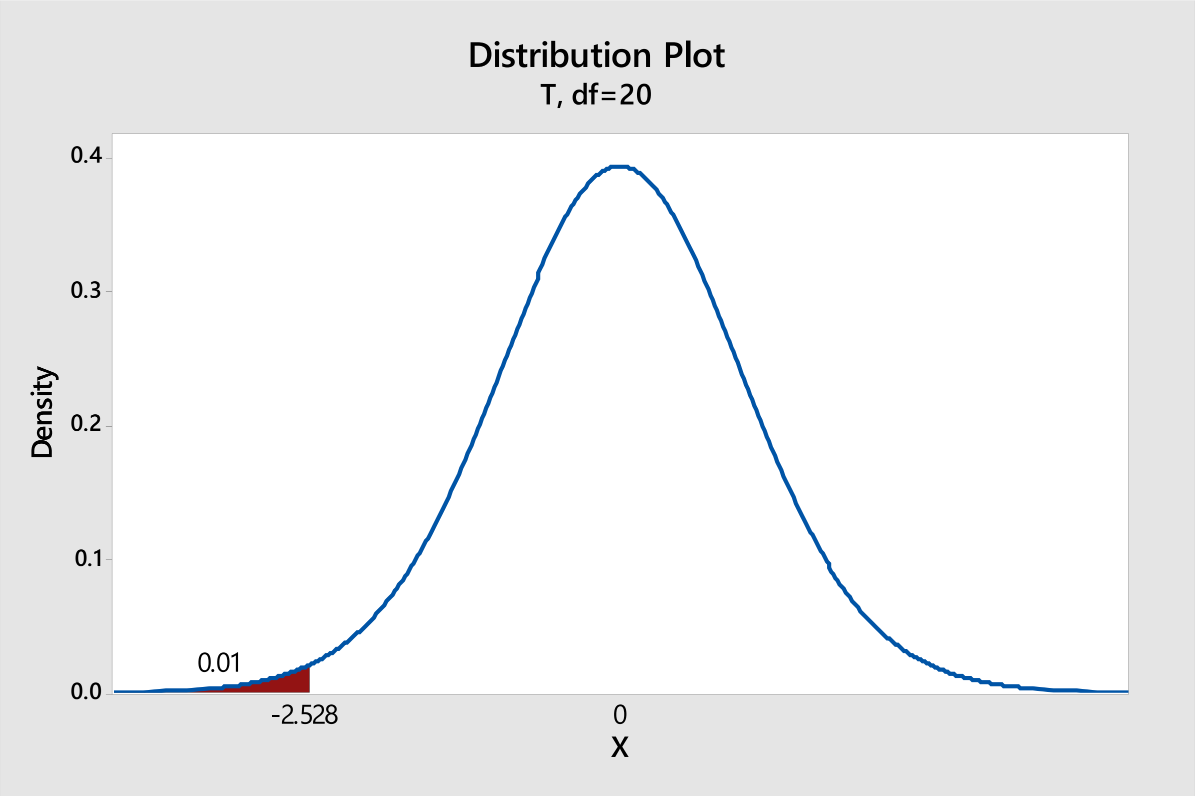
Full Solution

3) State the Critical region

We need a T Critical value with:

D.o.F= 20 and $\alpha = 0.01$ (left tailed)

Table Value:
−2.528



Full Solution

4) Conduct the experiment and calculate the test statistic

Column	n	Mean	Std. dev.
Gas Prices	21	3.4381	0.0903

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.4381 - 3.50}{0.0903/\sqrt{21}} = -3.1413$$

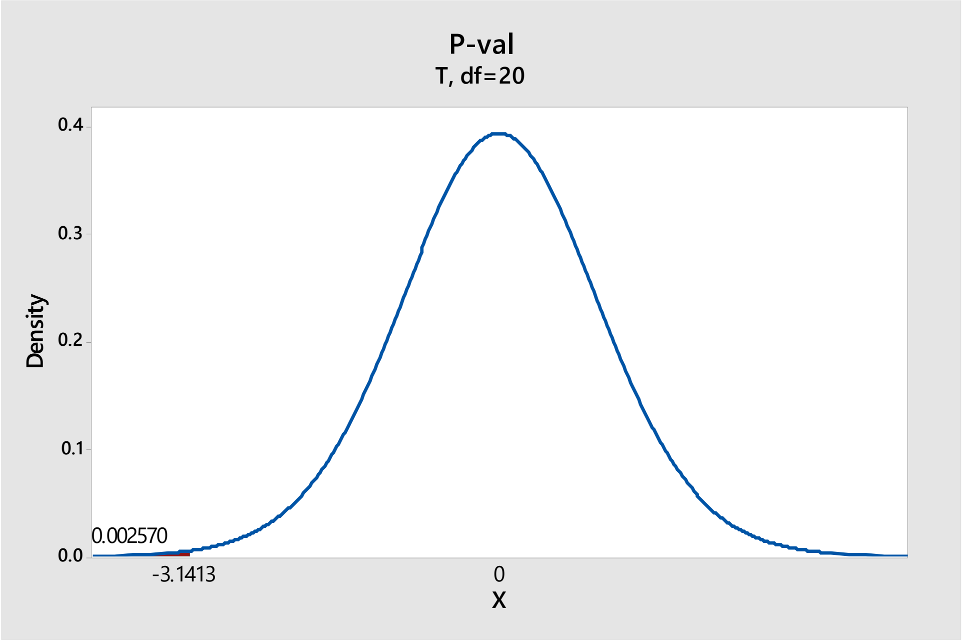
Full solution

4) Cont…

- Since it is a left tailed test we are interested in

  $P(t \leq -3.1413) = 0.00257$

- We can also estimate using the table that: $0.001 < \text{p-val} < 0.005$



Full Solution

5) Draw your Conclusion

<u>Using the Critical Value"</u>
Our Test Stat. of T=-3.14 falls in our rejection region

<u>Using the *p*-value:</u>
Comparing p-val=0.00257 to α=0.01

Both reject H_0.

Since we are rejecting H_0, we conclude that there is enough (significant) evidence to infer that the alternative hypothesis H_a is true.

Summary of Inference for the Sample Mean

- For known σ

Z

- For unknown σ but “large enough” sample ($n \geq 30$)

Z

- For unknown σ and small sample size

T

Statistical Inference Methods

- Statistical Inference: Drawing conclusions about a population from sample data.

- Methods
  - Point Estimation – Using a sample statistic to estimate a parameter
  - Confidence Intervals – supplements an estimate of a parameter with an indication of its variability
  - Hypothesis Tests– assesses evidence for a claim about a parameter by comparing it with observed data

Parameter	Measure	Statistic
μ	Mean of a single population	$\bar{X}$
σ^2	Variance of a single population	S^2
σ	Standard deviation of a single population	S
p	Proportion of a single population	$\hat{p}$
$\mu_1 - \mu_2$	Difference in means of two populations	$\bar{X}_1 - \bar{X}_2$
$p_1 - p_2$	Difference in proportions of two populations	$\hat{p}_1 - \hat{p}_2$

Categorical Variables

- Categorical variables place individuals into one of several groups

Voting on an issue:

Approve
Disapprove
Undecided

Color of Bag

Orange
Blue
Red

Analysis of Categorical data

- Calculating the mean is impossible.

- We can count the occurrences or "successes"

- From the counts we can calculate proportions…
    - Ex. Approval ratings

Read as "p-hat"

$$\hat{p} = \frac{\#\text{ who approve}}{\text{total }\#\text{ sampled}}$$

Is Inference Possible?

- Can we use a sample proportion to make inferences about a population proportion?

$$p = \text{population proportion}$$

$$\hat{p} = \text{sample proportion}$$

The Population Proportion, *p*

- If our data are categorical we can count the number of occurrences of each outcome to describe the population.
- From counts we can calculate proportions.

$$\hat{p} = \frac{x}{n}$$ is an estimate of p (sample statistic).

- We studied the binomial distribution using the count of the number of successes, X.
- We now deal with sample proportions because we want to estimate the probability of success, *p* in a population.

Conditions for CLT for Sample Proportions

Can we apply the CLT and assume normality of the sampling distribution of $\hat{p}$?

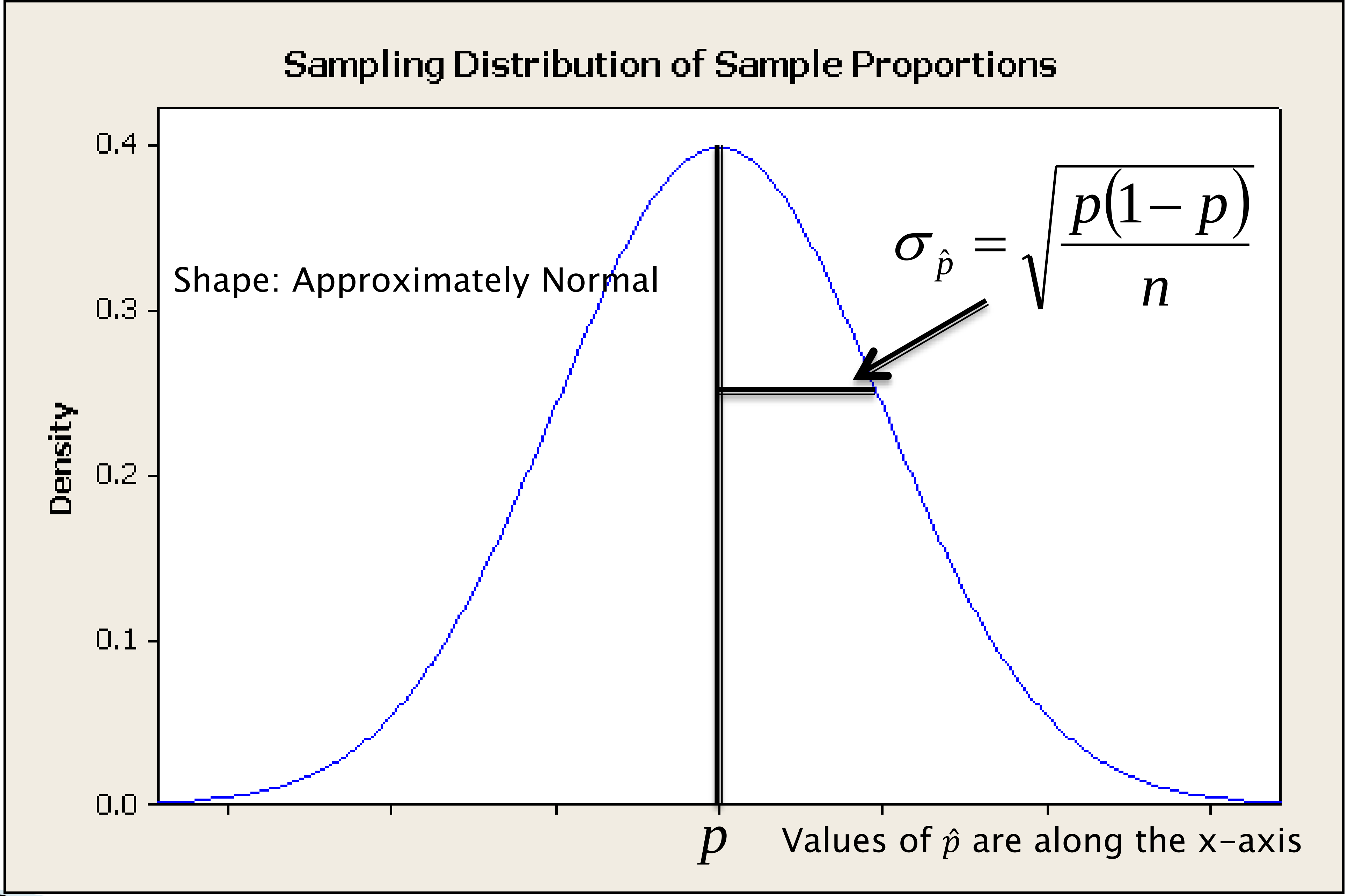
The sample size, *n*, is large enough that the sample expects at least 10 successes (yes) and 10 failures (no).

$$n\hat{p} \geq 10 \quad \text{and} \quad n(1 - \hat{p}) \geq 10$$

Central Limit Theorem (for proportions)

- When conditions hold, the sampling distribution for the sample proportion is approximately Normal, with mean *p* (the population proportion) and standard deviation defined to be the standard error given as $= \sqrt{\frac{p(1-p)}{n}}$

$$\hat{p} \text{ is} \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$



We can assume this is valid when $np \geq 10$ and $n(1 - p) \geq 10$

Confidence Interval for *p*

- In words:
  - estimate $\pm$ Margin of Error
    - Where M.o.E = C.V. x S.E.
    - So C.I. = …

estimate $\pm$ (critical value) x (standard error)

- Formula

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Estimated Standard Error

Confidence Interval on p example

In a random sample of 85 bearings, 10 have a surface finish rougher than specs. The point estimate of the proportion of faulty bearings in the population is $\hat{p} = 10/85 = 0.12$. What is a 95% *CI* of the population proportion?

$$p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$0.12 - 1.96\sqrt{\frac{0.12(0.88)}{85}} \le p \le 0.12 + 1.96\sqrt{\frac{0.12(0.88)}{85}}$$

$$0.12 - 0.07 \le p \le 0.12 + 0.07$$

$$0.05 \le p \le 0.19 \text{ which is rather wide.}$$

Choice of Sample Size

- You may need to choose a sample size large enough to achieve a specified margin of error.

- Margin of Error (ME)= $z_{\alpha/2}\sqrt{p(1-p)/n}$

- Now solving for n,

$$n = p^*(1 - p^*)\left(\frac{z^*}{ME}\right)^2$$

- p may be estimated:
  - From a prior sample ($\hat{p}$)
  - Subjectively (a guess)
  - Conservatively ($p = 0.5$)

Confidence interval on P sample size example

What sample size is required to be 95% confident that the error would be equal to or less than 0.05 for an estimate of p?

With the prior estimate of $\hat{p}$,

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 p(1-p) = \left(\frac{1.96}{.05}\right)^2 0.12(0.88) = 162.3 \rightarrow 163$$

With a conservative estimate of $\hat{p}$,

$$n = \left(\frac{1.96}{.05}\right)^2 0.5(0.5) = 384.2 \rightarrow 385$$

Test Statistic: Hypothesis Test of One Population Proportion

- The test statistic has the structure:

$$z = \frac{\text{observed value} - \text{null value}}{SE}$$

- Thus our test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

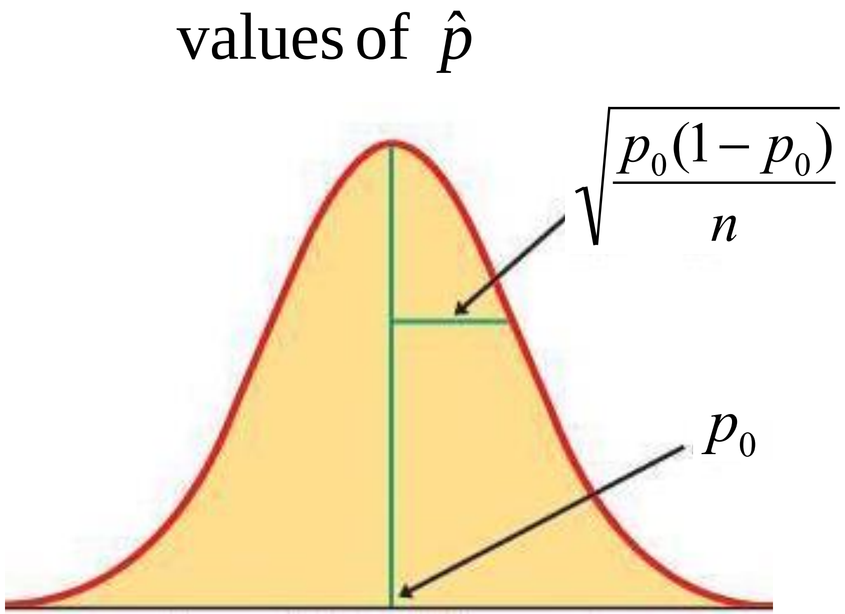
$\hat{p}$ is the sample proportion

p_0 is proportion believed to be true in the null hypothesis

$$z_o = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

Hypothesis Testing for *p*

- If H_0 is true, the sampling distribution is known.



Difference in the Standard Error

- For the confidence interval, no claim is made about the population proportion, so we must use $\hat{p}$ is the standard error.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- However, in a Hypothesis test, we assume the null hypothesis to be true and use that quantity in the standard error of the test statistic.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Proportion HT example

- A semiconductor manufacturer produces controllers for automotive engines. The customer requires that the fraction defective at a critical manufacturing step not exceed 5%, and that the manufacturer demonstrate process capability at this level of quality using $\alpha=0.05$.

- The manufacturer takes a random sample of 200 and finds 4 to be defective. Can the manufacturer demonstrate process capability to the customer?

Full Solution

1) State the hypotheses.

$$H_0 : p = 0.05 \text{ and } H_1 : p < 0.05.$$

We see that this test is left-tailed

2) We are dealing with parameter of interest, p, and we meet requirements of the CLT.

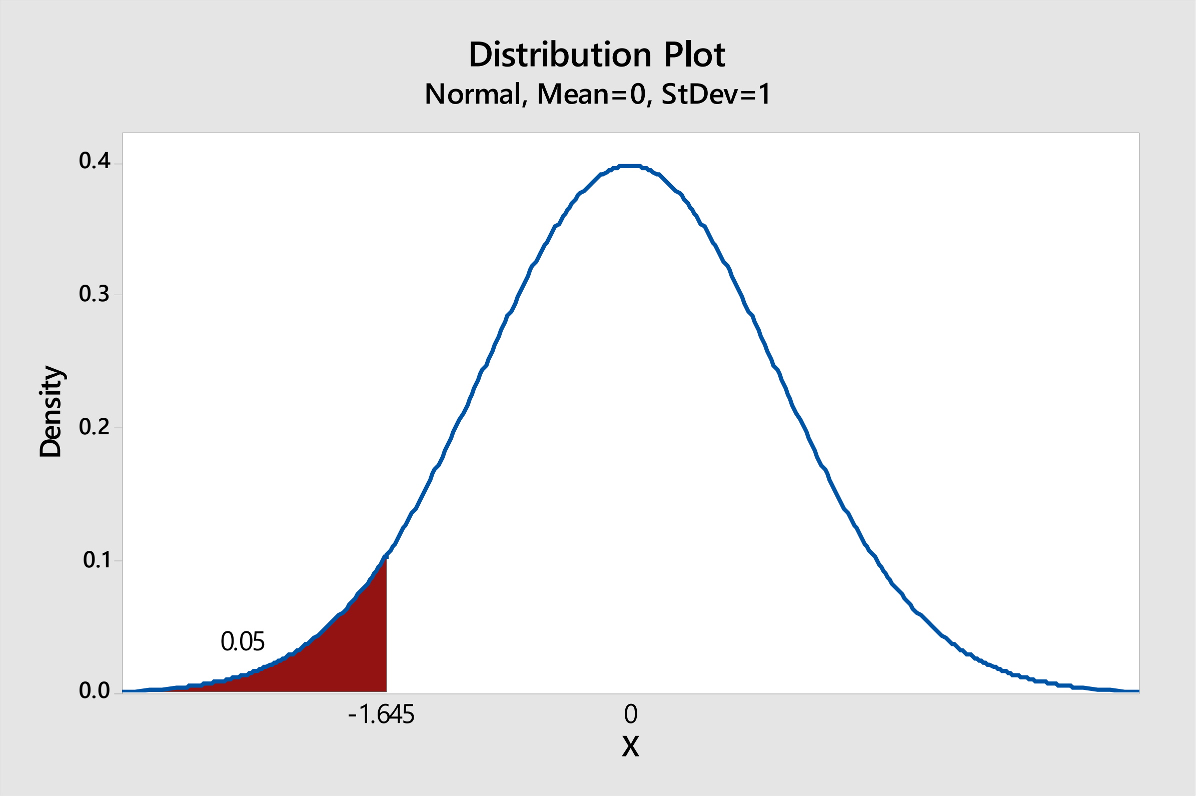
Full Solution

3) State the Critical region

$\alpha = 0.05$

Table Value:

-1.645



Full Solution

4) Conduct the experiment and calculate the test statistic

x = 4 n= 200 so $\hat{p} = \frac{4}{200} = 0.02$

$$z_o = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{4 - 200(0.05)}{\sqrt{200(0.05)(0.95)}} = -1.95$$

OR:

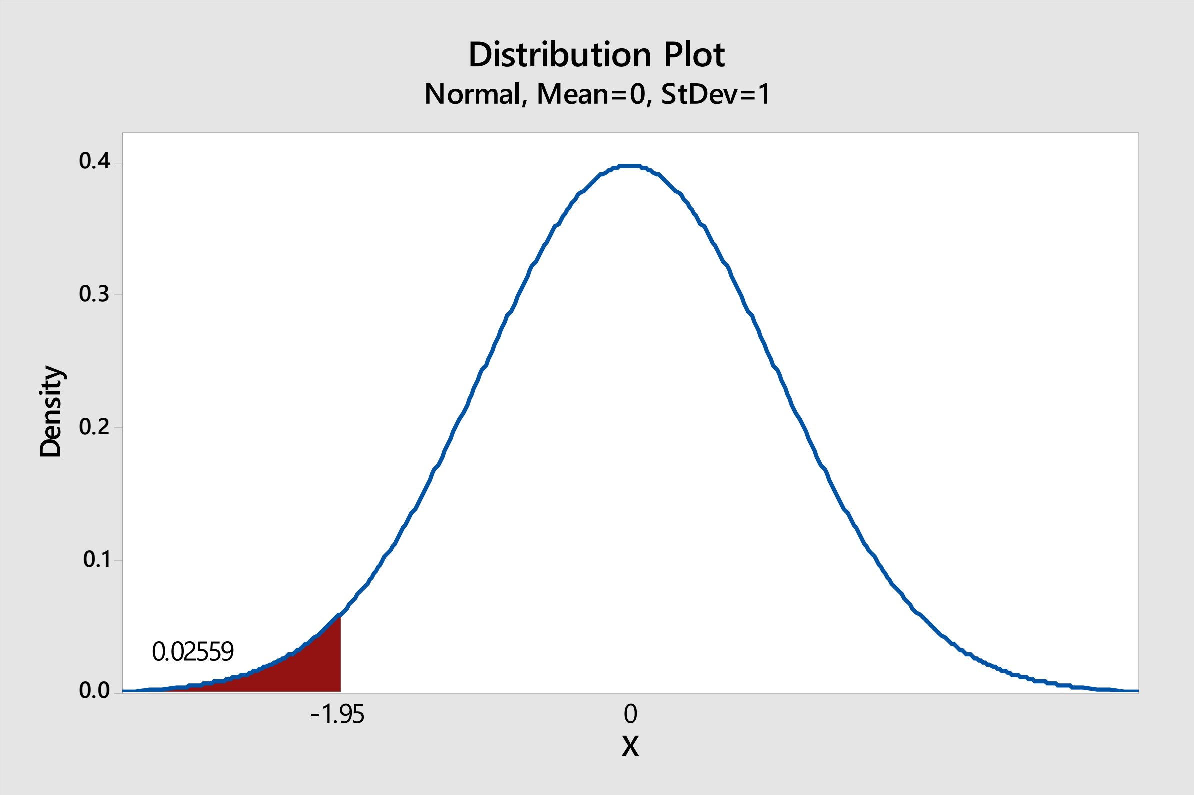
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Full solution

4) Cont… find p-val

- Since it is a left tailed test we are interested in

$$P(z \leq -1.95) = 0.0256$$



Full Solution

5) Draw your Conclusion

<u>Using the Critical Value”</u>
Our Test Stat. of T=-1.95 falls in our rejection region

<u>Using the *p*-value:</u>
Comparing p-val=0.0256 to α=0.05

Both reject H_0.

Since we are rejecting H_0, we conclude that there is enough (significant) evidence to infer that the alternative hypothesis H_a is true.