MODULE 6

NULL HYPOTHESIS SIGNIFICANCE TESTING

NULL HYPOTHESIS SIGNIFICANCE TESTING

- Null hypothesis significance testing (NHST) is a procedure that uses a sample of data to determine whether an effect or association might be present at the population level
- A cornerstone of psychological research for over 100 years
- Provides a formal recipe to evaluate evidence from data, and the same steps apply to virtually any research question

SIGNIFICANCE TESTING STEPS

1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

OUTLINE

1. Quick review
2. Overview of NHST
3. Significance testing steps
4. Study questions
5. R analysis

OUTLINE

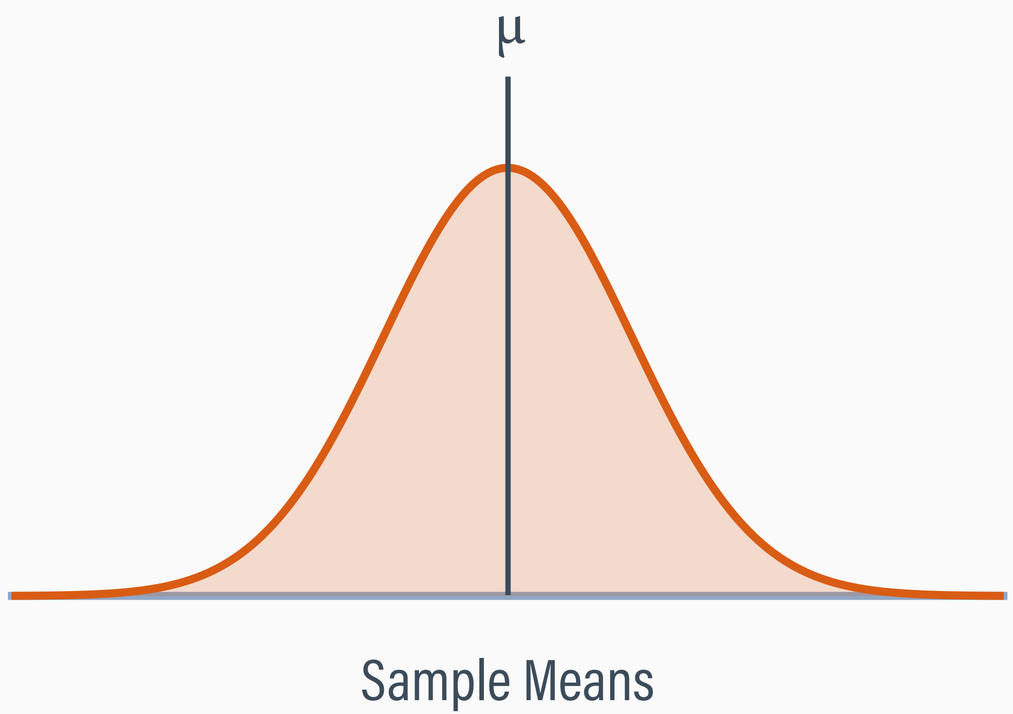
1. Quick review
2. Overview of NHST
3. Significance testing steps
4. Study questions
5. R analysis

QUICK REVIEW: SAMPLING ERROR

- The frequentist paradigm imagines a single population that spawns many hypothetical random samples of data (one parameter, many hypothetical estimates)
- The amount by which an estimate differs from the true population statistic is called sampling error
- Every hypothetical sample has a different amount of error

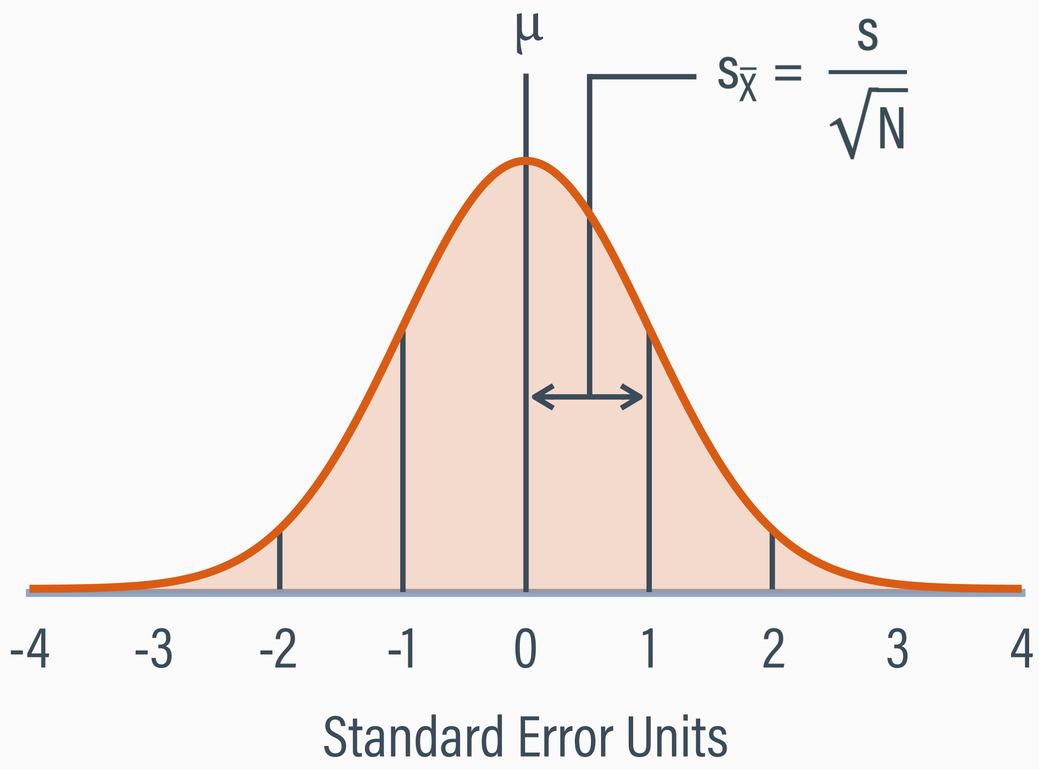
QUICK REVIEW: SAMPLING DISTRIBUTION

- The distribution of the estimates from many hypothetical samples is a sampling distribution
- With a large enough N, sample means follow a normal curve centered at the true mean
- Most estimates have small sampling errors, but a few have larger errors



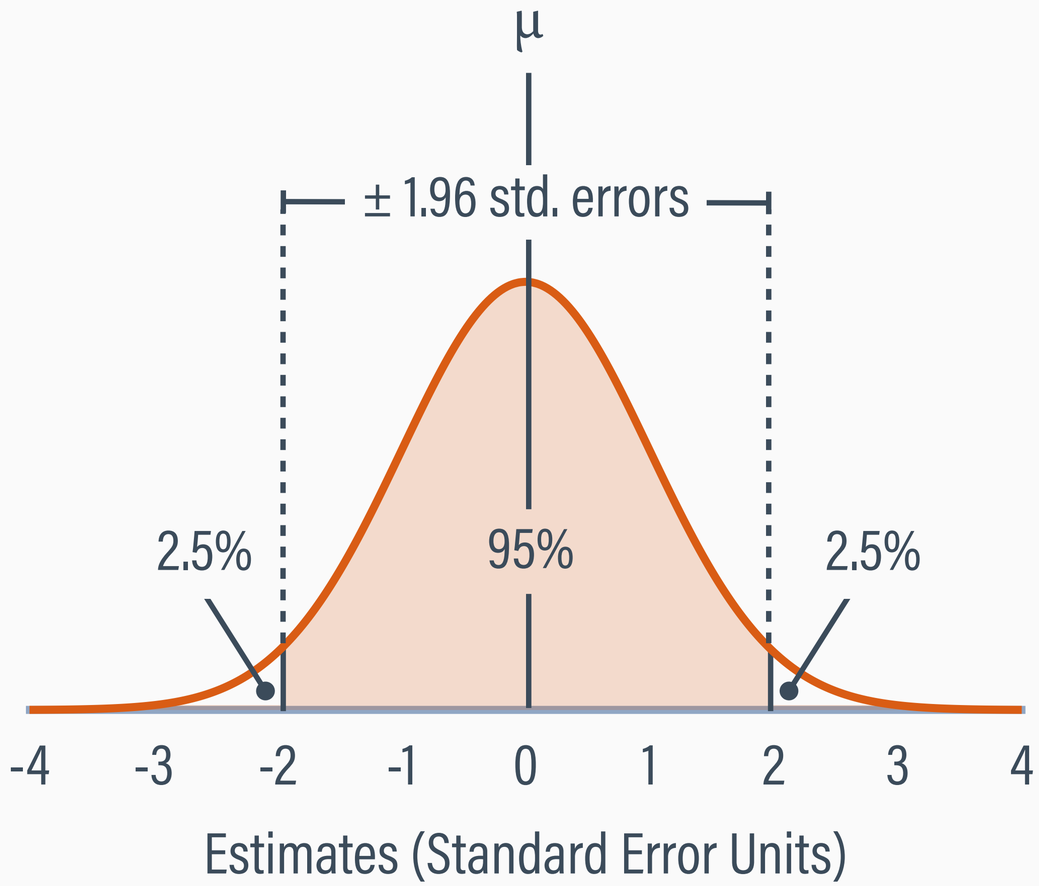
QUICK REVIEW: STANDARD ERROR

- The standard error is the average distance from a sample mean and the true mean
- $s_{\bar{X}}$ = standard deviation of the sample means
- The standard error is the average or expected amount of sampling error across many hypothetical samples



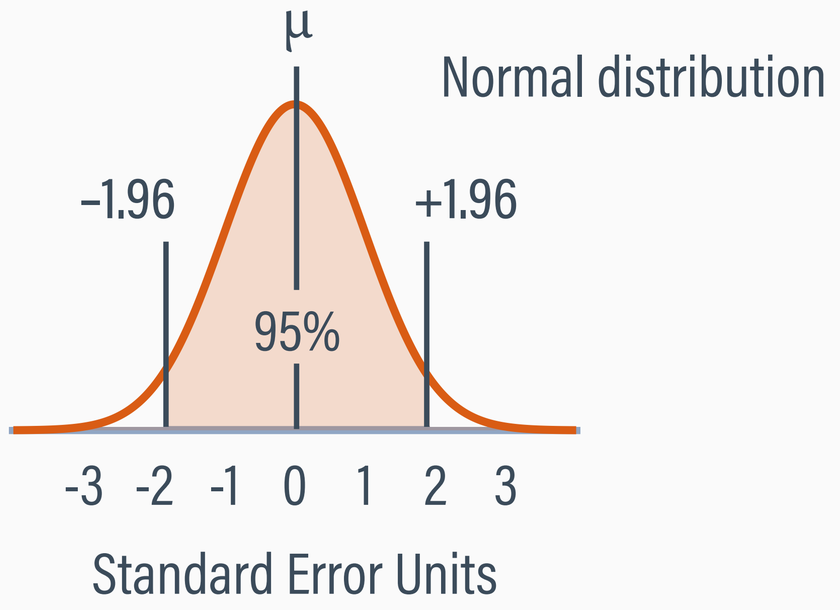
QUICK REVIEW: NORMAL CURVE RULE

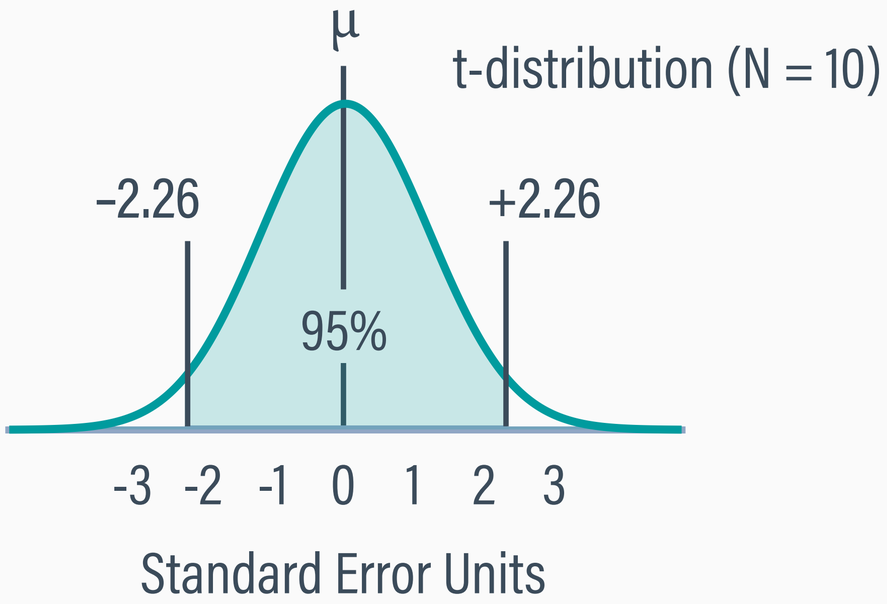
- The standard error is the standard deviation of many hypothetical sample means
- We can apply normal curve rules of thumb
- 95% of the means from large samples are within ± 1.96 standard errors of the true mean



QUICK REVIEW: T-DISTRIBUTION

- When using small samples, the normal curve is an inaccurate description of sampling error
- The t-distribution is a series of bell-shaped curves that stretch out (become more variable) as the N decreases
- Small samples are more likely to produce outlier estimates, and “stretching” the curve honors that





OUTLINE

1. Quick review
2. Overview of NHST
3. Significance testing steps
4. Study questions
5. R analysis

UVEAL MELANOMA and DEPRESSION

Uveal melanoma, a rare eye cancer, presents potential vision loss and life threat. This prospective, longitudinal study interrogated the predictive utility of visual impairment, as moderated by optimism/pessimism, on depressive symptoms in 299 adults undergoing diagnostic evaluation.

Annette Stanton

James MacDonald

MacDonald, J.J., Jorge-Miller, A., Enders, C.K., McCannel, T., Beran, T., & Stanton, A.L. (2021). Perceived and objective visual impairment predicting depressive symptoms across one year in uveal melanoma diagnostic biopsy: Optimism and pessimism as moderators. *Health Psychology*, *40*, 408-417.

KEY VARIABLES

Depression

The CES-D is a 20-item inventory that asks people to rate how often they experience depressive symptoms such as restless sleep, poor appetite, and feeling lonely.

Cancer Diagnosis

Based on a diagnostic clinical evaluation for a possible intraocular malignancy, participants were classified as having malignant or nonmalignant diagnoses.

RESEARCH QUESTION

- Studies typically attempt to answer a handful of research questions involving associations between key variables
- Do people who receive a positive cancer diagnosis experience clinical levels of depressive symptoms, or do they show no meaningful elevation?
- CES-D scores > 16 are widely viewed as indicating risk for clinical depression, and scores < 16 are in the mild range

NULL HYPOTHESIS

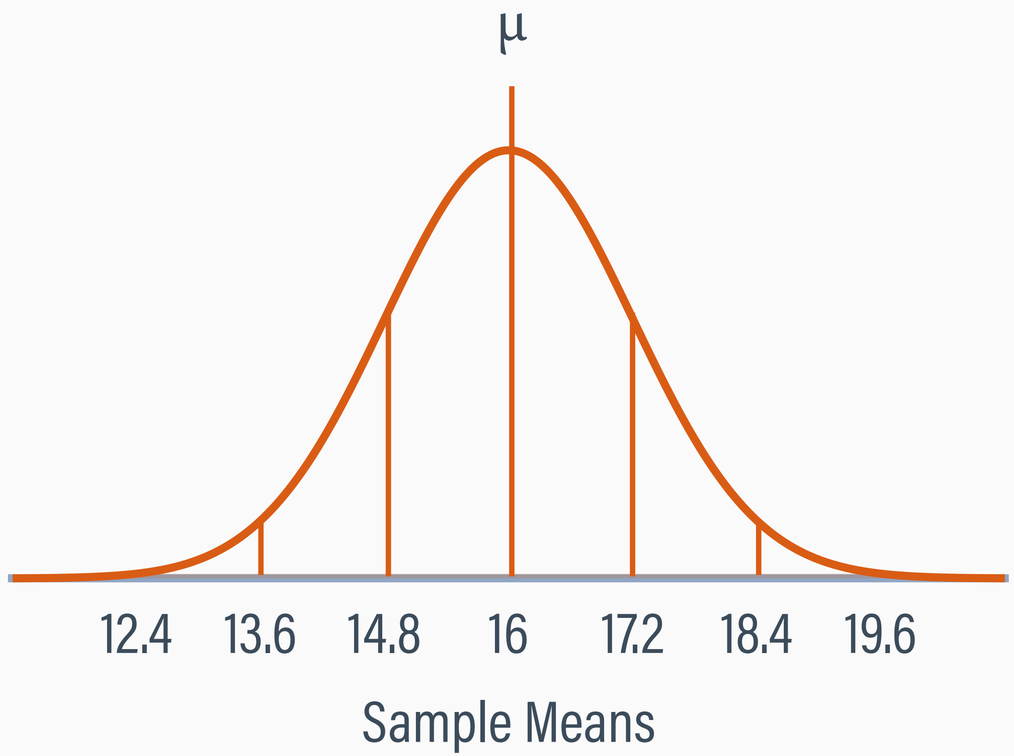
- The significance testing process starts with a null hypothesis that is counter to what we expect to find
- In the depression study, researchers want to determine whether or not people who receive a cancer diagnosis experience clinical depression
- The null that the true population mean is $\mu = 16$ is counter to expectations because researchers anticipate that the average level of depressive symptoms will exceed the clinical cutoff

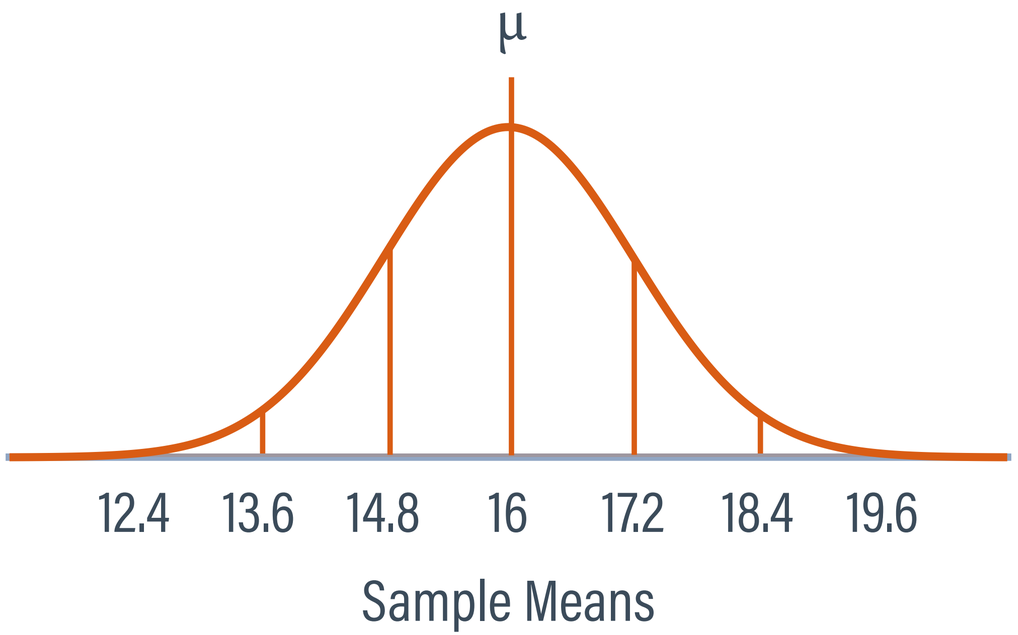
DEMONSTRATING AN EFFECT

- The null hypothesis that is counter to what we expect to find
- To demonstrate an effect (e.g., patients experience clinical symptoms), we attempt to refute the null with evidence from the data
- To refute the null that $\mu = 16$, we would need to observe a sample mean $\bar{X}$ that is much higher or much lower than 16

SAMPLING DISTRIBUTION IF NULL IS TRUE

- The sampling distribution shows the means from many hypothetical samples of N = 107 drawn out of the null population with $\mu = 16$
- To refute the null, the sample $\bar{X}$ must lie far away from 16 (either above or below)
- An $\bar{X}$ close to 16 would lend support to the null

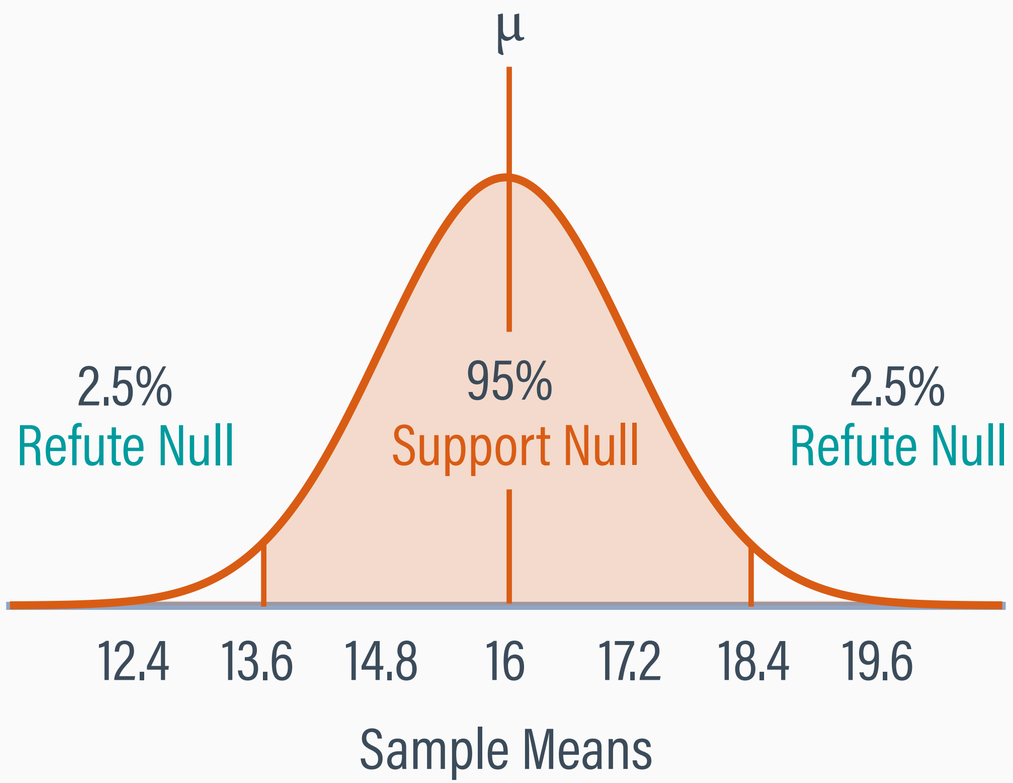




Consider the sampling distribution of sample means from a null population with $\mu = 16$. In small groups of two or three, discuss how far above or below 16 the sample mean ($\bar{X}$) would need to be for you to feel confident that the data refutes the null (i.e., demonstrates an effect).

EVALUATING THE NULL

- By convention, we refute the null if the sample $\bar{X}$ falls outside the middle 95% of the sampling distribution
- Such as sample has less than a 5% chance of originating from the null population
- Any $\bar{X}$ within the middle 95% of the sampling distribution lends support to the null



OUTLINE

1. Quick review
2. Overview of NHST
3. Significance testing steps
4. Study questions
5. R analysis

SIGNIFICANCE TESTING STEPS

1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

NULL HYPOTHESIS

- In the population, people who receive a cancer diagnosis have a mean exactly at the clinical cutoff 16

$$H_0: \mu = 16$$

- The null that $\mu = 16$ is counter to expectations because researchers anticipate that depressive symptoms could exceed (or possibly fall below) the clinical cutoff

ALTERNATE HYPOTHESIS

- The alternate hypothesis reflects the researcher's expectations about an effect
- A one-tailed (directional) alternate hypothesis predicts an effect in just one direction (e.g., people are above the clinical cutoff), and a two-tailed alternate predicts an effect in either direction (e.g., people could be above or below the clinical cutoff)
- Two-tailed tests predominate psychology applications

TWO POSSIBLE ALTERNATIVE HYPOTHESES

- One-tailed alternate: People who receive a cancer diagnosis are above the clinical cutoff for depression

$$H_A: \mu > 16$$

- Two-tailed alternate: People who receive a cancer diagnosis could be above or below the clinical cutoff for depression

$$H_A: \mu \neq 16 \ (\mu > 16 \text{ or } \mu < 16)$$

ROLE OF THE ALTERNATE HYPOTHESIS

- We do not directly test the alternative hypothesis, which represents the researcher's belief
- We provide support for the alternative by rejecting the null
- Whether the alternative is one-tailed or two-tailed is very important, as this determines what kind of evidence would lead us to reject the null

SIGNIFICANCE TESTING STEPS

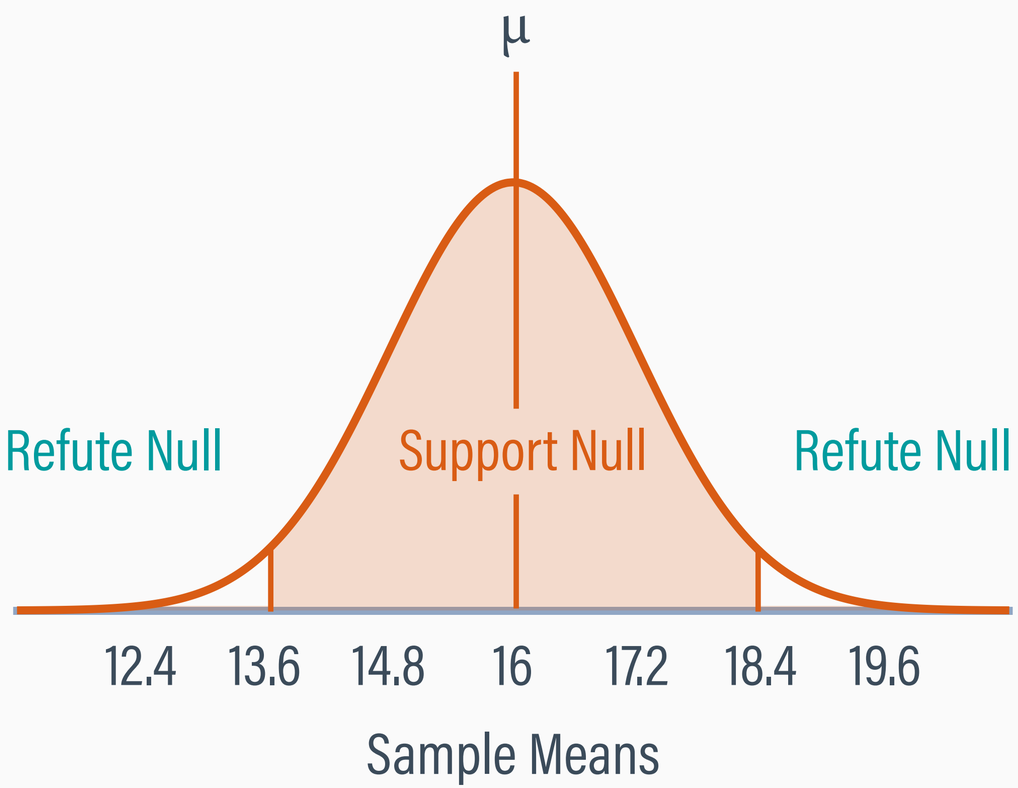
1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

STANDARD OF EVIDENCE

- The data are the evidence that we use to conclude whether the null is plausible (“innocent”) or implausible (“guilty”)
- If the sample mean from our data is very different from the null mean, then we conclude that the null hypothesis is implausible
- How big a difference do we need to observe to refute the null?

EVALUATING THE NULL

- Any $\bar{X}$ near the middle of the sampling distribution lends support to the null
- Such as sample has a high probability of originating from the null population
- We refute the null if the sample $\bar{X}$ falls far from μ
- Such as sample has a low probability of originating from the null population

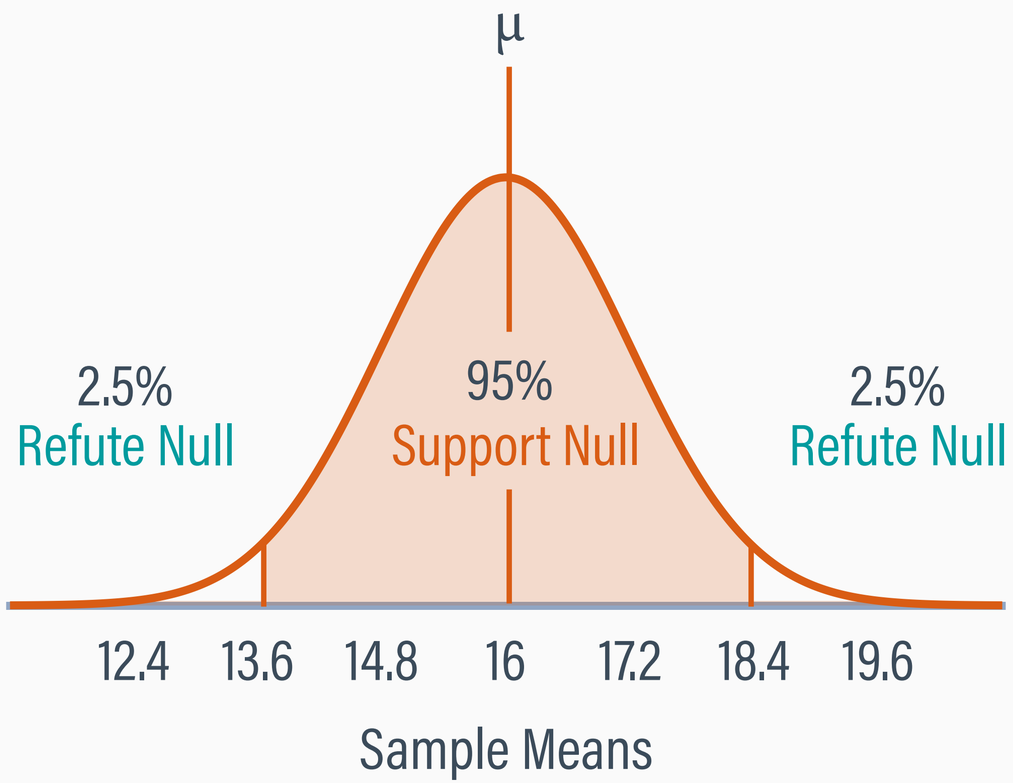


5% SIGNIFICANCE CRITERION

- Evidence against the null is presented as a probability
- R.A. Fisher — an influential biologist and statistician from the early 1900s — recommended 5% ($p < .05$) as a criterion
- If a sample mean has a 5% or less probability of originating from the null population, then we conclude H_0 is implausible

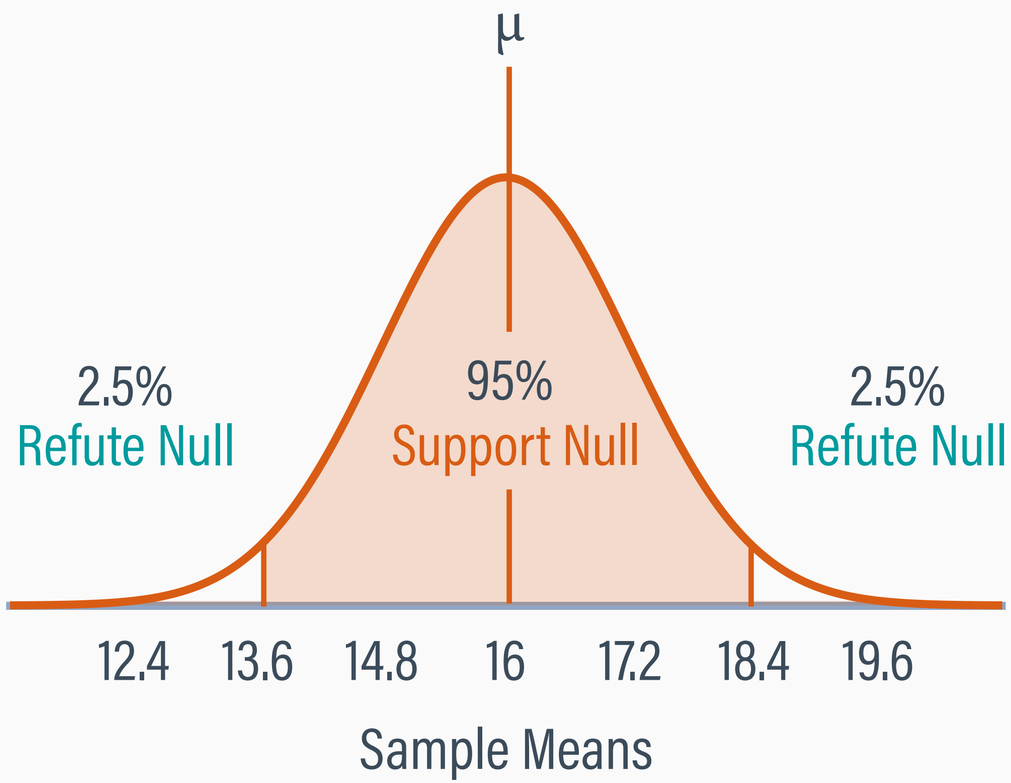
SAMPLING DISTRIBUTION

- By convention, we refute the null if the sample $\bar{X}$ falls outside the middle 95% of the sampling distribution
- Such as sample has less than a 5% chance of originating from the null population
- That is, we refute the null if the sample $\bar{X}$ is an outlier (rare) from the null population ($p < .05$)



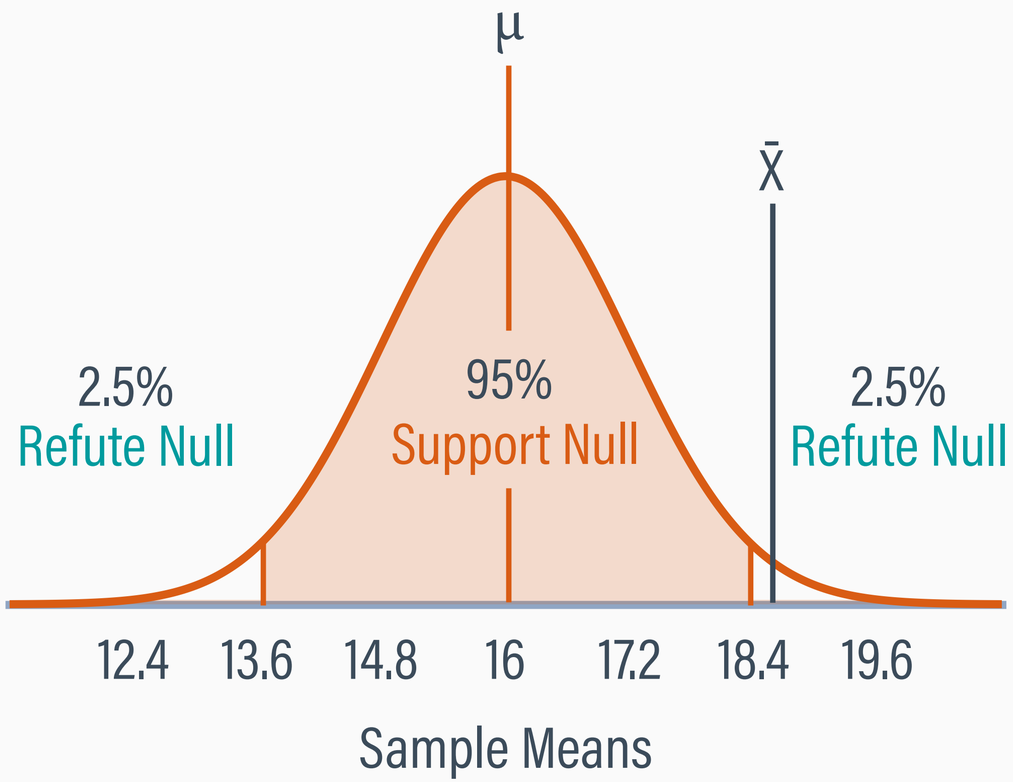
TWO-TAILED ALTERNATE HYPOTHESES

- A two-tailed alternate hypothesis predicts an effect in either direction
- e.g., People who receive a cancer diagnosis could be above or below the clinical cutoff
- The 5% rejection region (alpha level) is split in half to allow for the possibility that either an increase or a decrease provides evidence against H_0

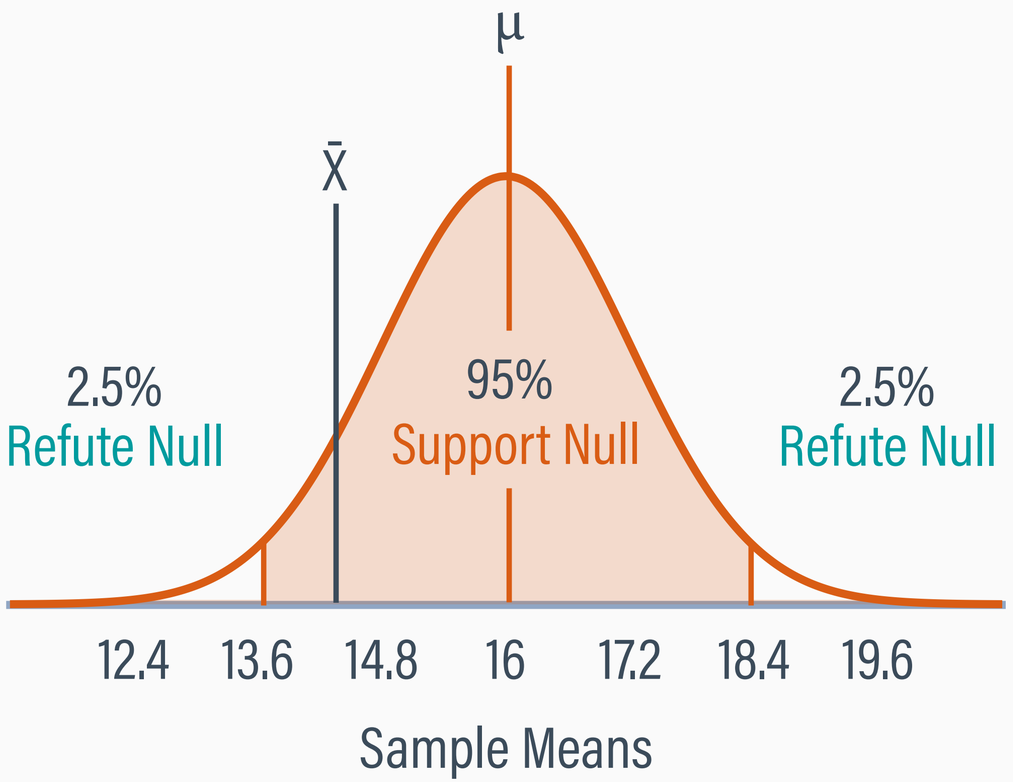


TWO POSSIBLE DECISIONS

Refute the null: $\bar{X}$ is unlikely from a population with $\mu = 16$ ($p < .05$)

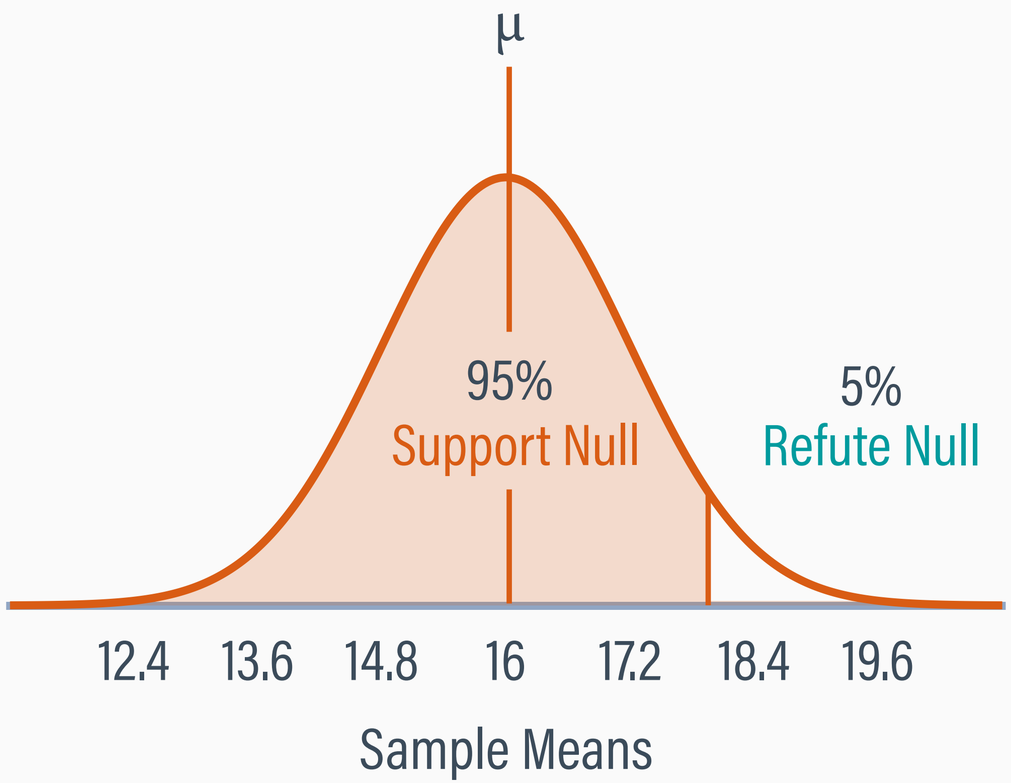


Support the null: $\bar{X}$ is likely from a population with $\mu = 16$ ($p > .05$)



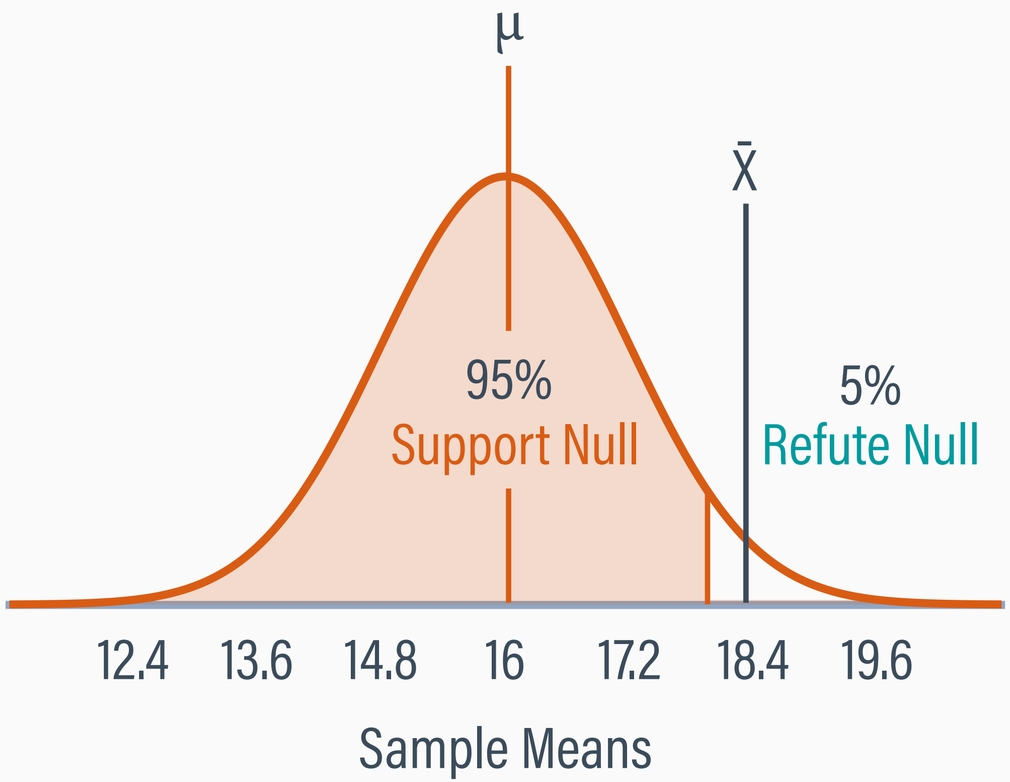
ONE-TAILED ALTERNATE HYPOTHESES

- A one-tailed alternate hypothesis predicts an effect in just one direction
- e.g., People who receive a cancer diagnosis could only be above the clinical cutoff
- The 5% rejection region (alpha level) is assigned to one tail to reflect that only a very high $\bar{X}$ can provide evidence against H_0

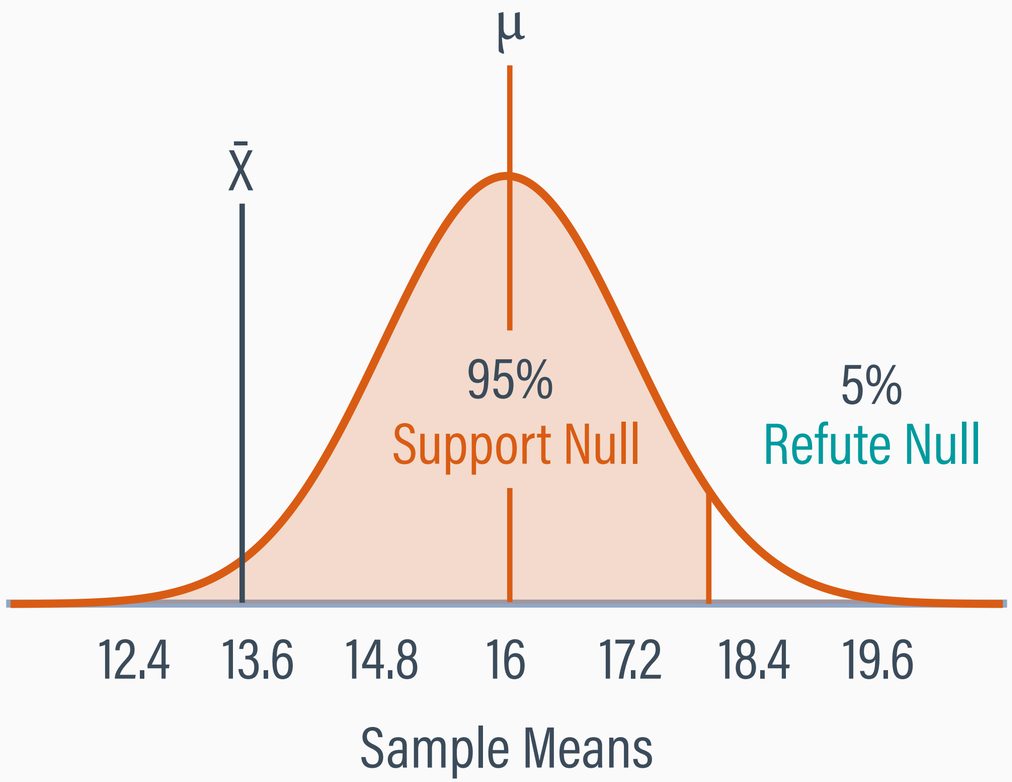


TWO POSSIBLE DECISIONS

Refute the null: $\bar{X}$ is unlikely from a population with $\mu = 16$ ($p < .05$)



Support the null: $\bar{X}$ is likely from a population with $\mu = 16$ ($p > .05$)



FALSE POSITIVES (TYPE I ERRORS)

- The 5% criterion defines a region of the distribution that contains outlier samples that are unlikely *but not impossible*
- When $\bar{X}$ falls in the rejection region (evidence against the null), there is still a 5% chance it came from the null population
- A significant result means we risk a false positive—incorrectly rejecting the null when it is actually true (called a Type I error)
- Adopting a 5% criterion fixes the Type I error risk to 5%

SIGNIFICANCE TESTING STEPS

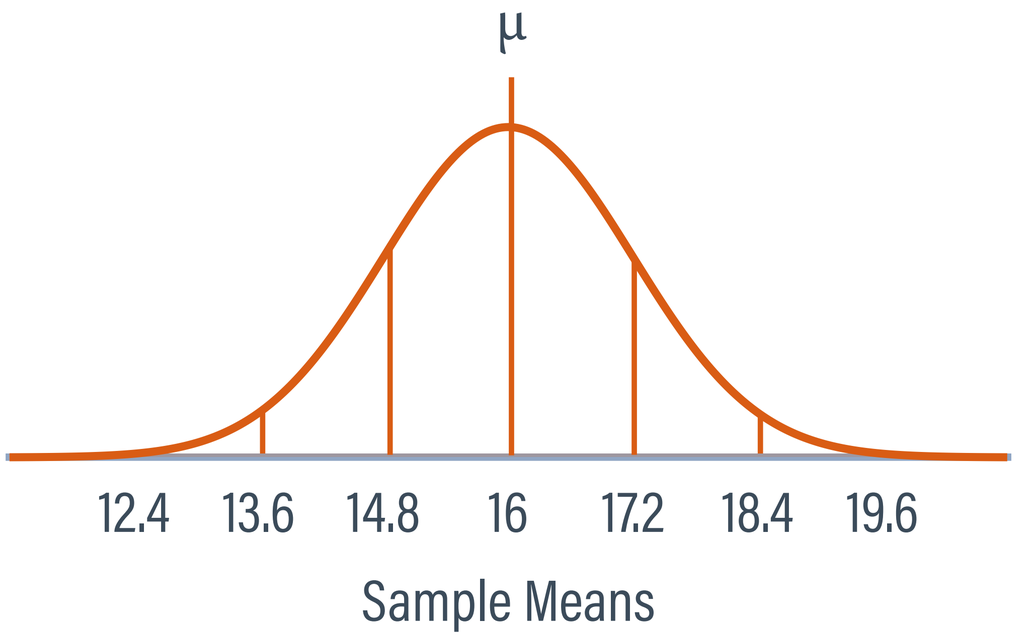
1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

ANALYSIS SUMMARY

- N = 107 participants received a cancer diagnosis
- The sample mean is 18.10
- The standard error $s_{\bar{X}} = 1.20$ indicates that the expected (average) sampling error across many samples of N = 107 is about 1.20 CES-D points

N	$\bar{X}$	SD	SE
107	18.10	12.45	1.20

$$s_{\bar{X}} = \frac{s}{\sqrt{N}} = \frac{12.45}{\sqrt{107}} = 1.20$$



Consider the sampling distribution of sample means from a null population with $\mu = 16$. In small groups of two or three, discuss the descriptive statistics and determine whether $\bar{X} = 18.10$ ($s_{\bar{X}} = 1.20$) provides evidence for or against the null hypothesis.

SIGNIFICANCE TESTING STEPS

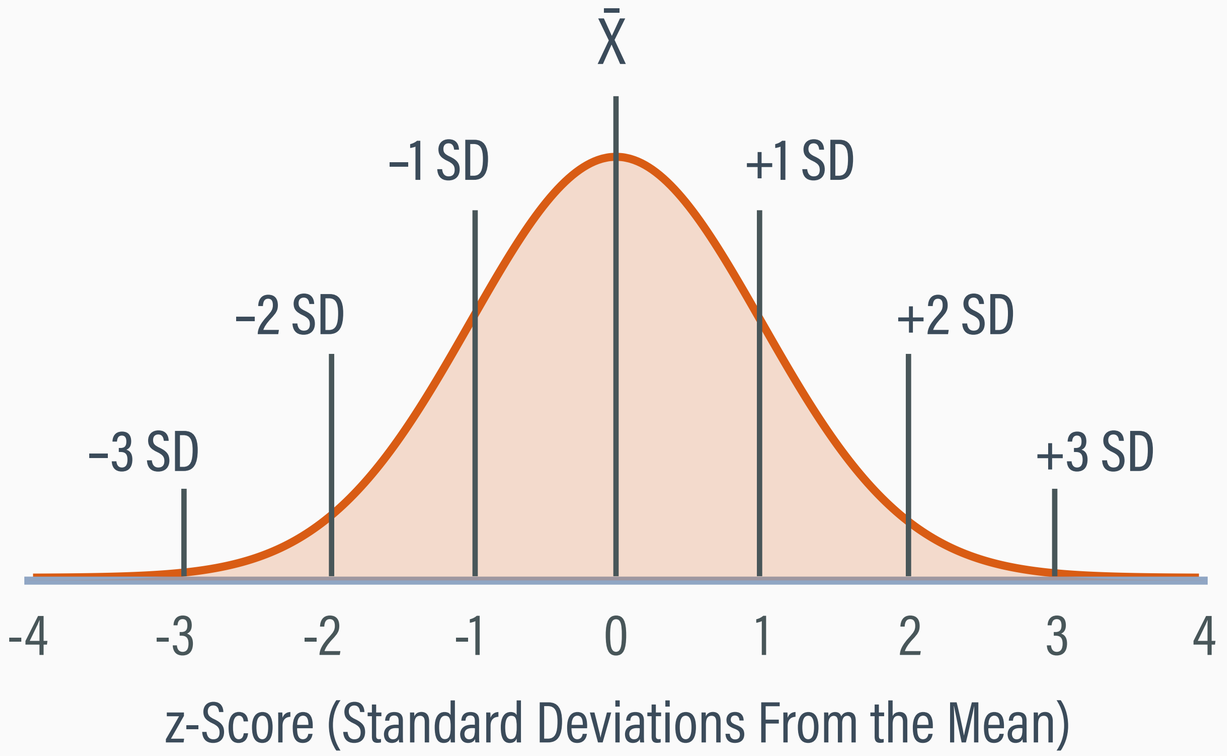
1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. **Compare data to null hypothesis**
5. Evaluate hypotheses and draw conclusion

COMPARING DATA TO THE NULL

- Two ways to determine whether the sample $\bar{X}$ is consistent (or inconsistent) with the null population mean
- The t-statistic gives a standardized distance between the sample mean and the null hypothesis mean (like a z-score)
- A p-value tells us how likely it is that hypothetical samples like our data would originate from the null population

z-SCORES REVISITED

- The z-score scale is a common standardized metric that expresses scores as standard deviation units from the mean



t-STATISTIC

- The t-statistic quantifies the number of standard error units that separate the sample mean and null hypothesis population mean

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{\text{distance from the null}}{\text{standard error (std. dev. of } \bar{X})}$$

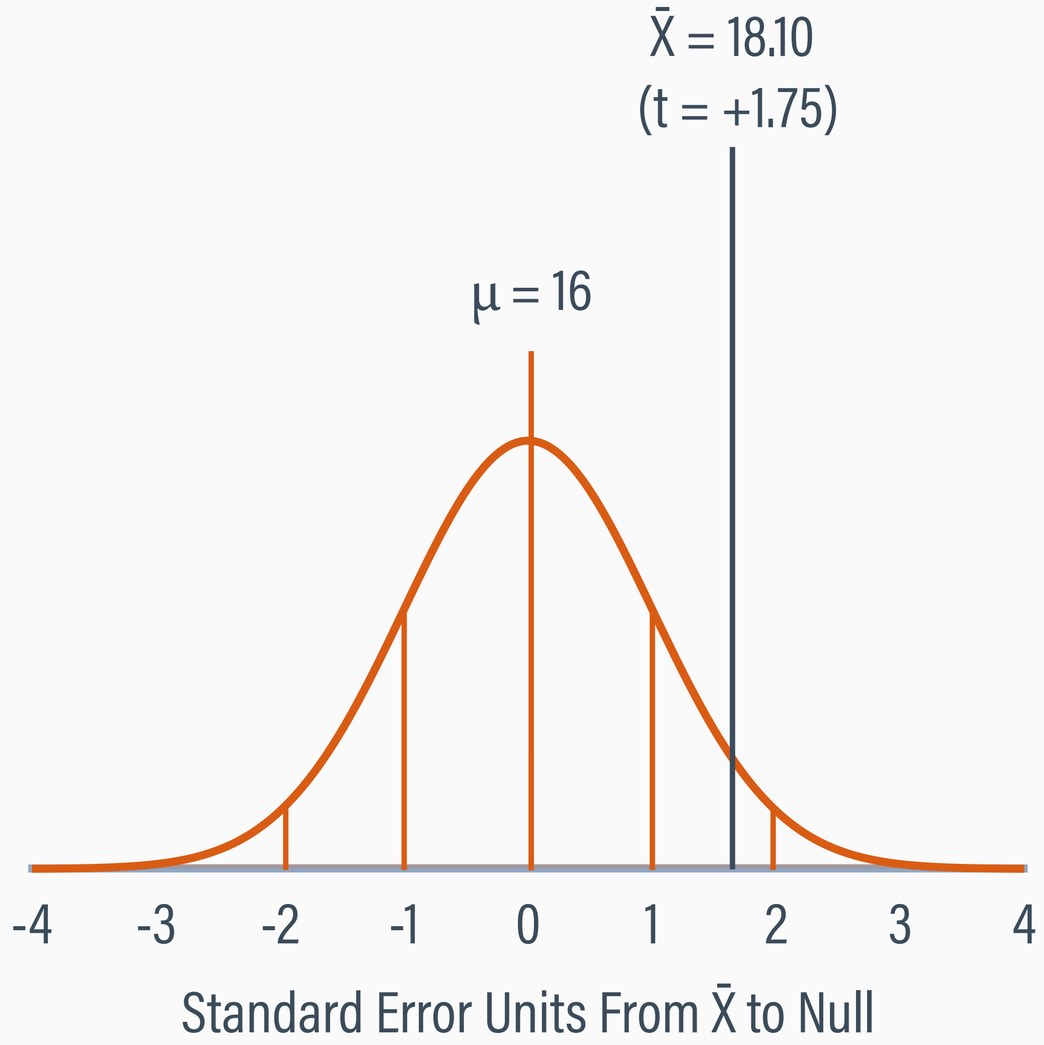
- The t-statistic is the same as a z-score (a standardized metric where distance is expressed in standard deviation units)

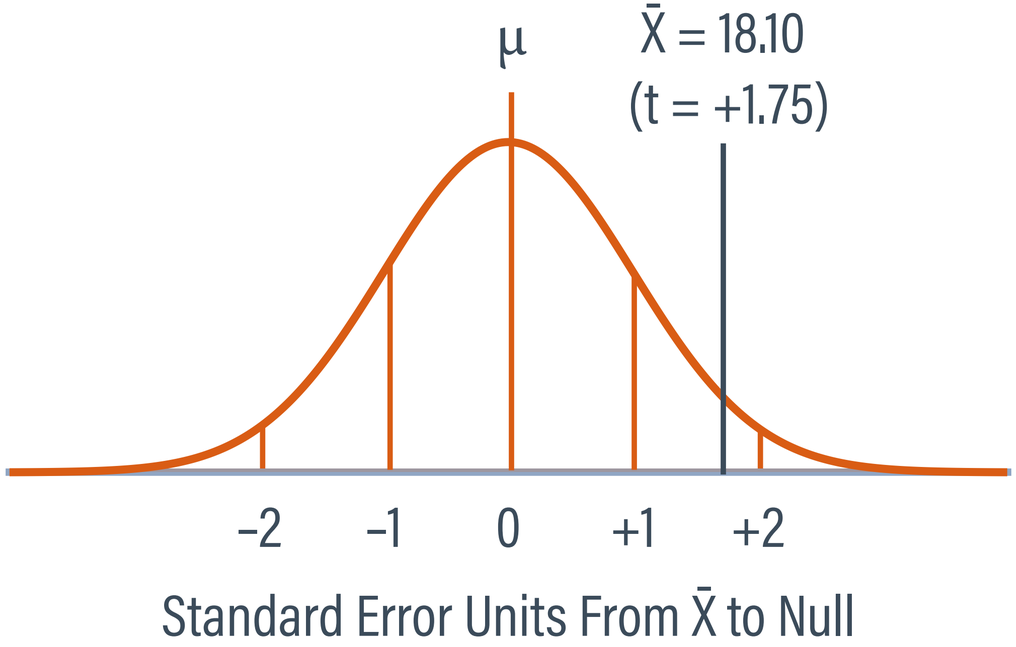
t-STATISTIC EXAMPLE

- The positive t-statistic conveys that the sample $\bar{X}$ is higher than the null ($\mu = 16$)

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{18.10 - 16}{1.20} = 1.75$$

- The numeric value indicates that $\bar{X}$ is 1.75 standard error units from the null null





Consider the sampling distribution of sample means from a null population with $\mu = 16$. The sample mean and t-statistic are $\bar{X} = 18.10$ and $t = 1.75$. In small groups of two or three, apply the normal curve rule of thumb and decide whether the sample data provide evidence for or against the null hypothesis

FREQUENTIST PARADIGM REVISITED

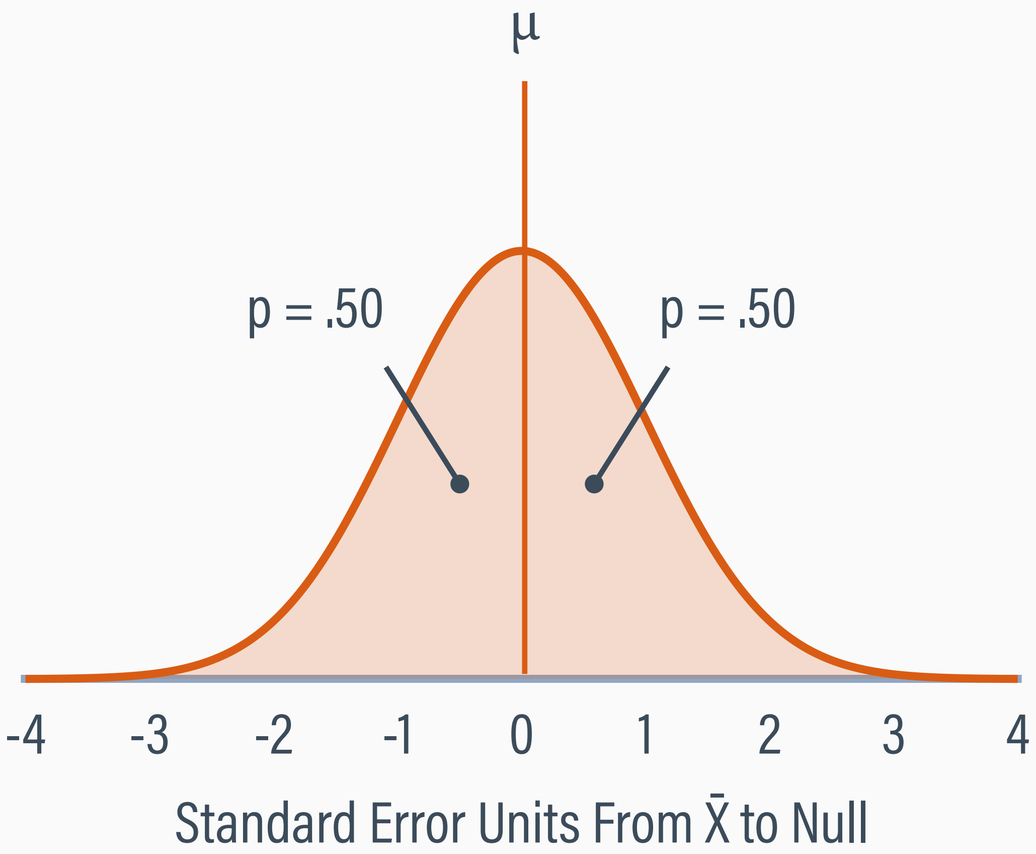
- The frequentist paradigm is defined by the idea that there is one population with unknown parameters (e.g., μ and σ)
- We imagine numerous hypothetical samples of size N from the population, each with its own unique estimates and 95% CIs
- The population-level statistics (parameters) are locked in at a single set of values, whereas the sample-level statistics (estimates, t-statistics) vary across different data sets

PROBABILITY VALUES (P-VALUES)

- A p-value is defined as proportion of hypothetical samples that have a t-statistic at least as large as the sample data
- Assuming the null is true, how likely is it to draw a sample with an effect at least as large as the one from our data?
- Visually, probability is an area under the curve, obtained by applying calculus integrals to the t-distribution function

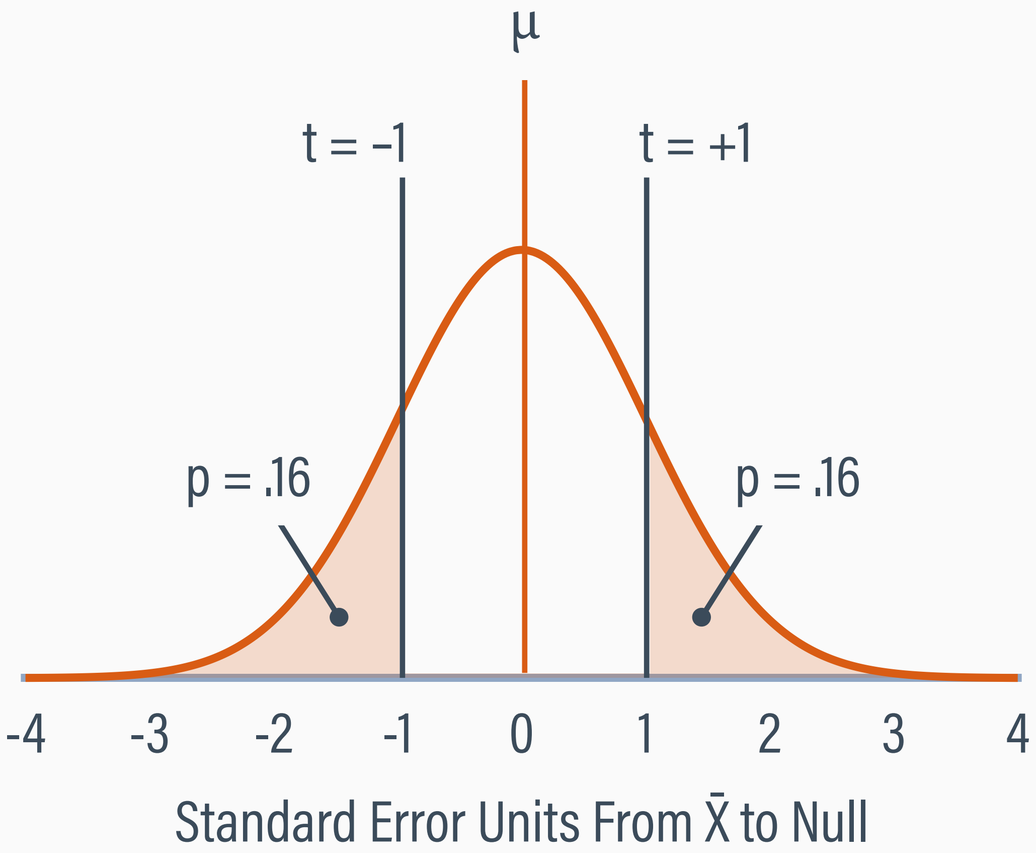
TOTAL AREA UNDER THE CURVE

- The total probability (area under the sampling distribution) in both directions is 1.0
- 50% of all hypothetical samples are above the null population mean and 50% are below



P-VALUE AT ± 1 STANDARD ERRORS

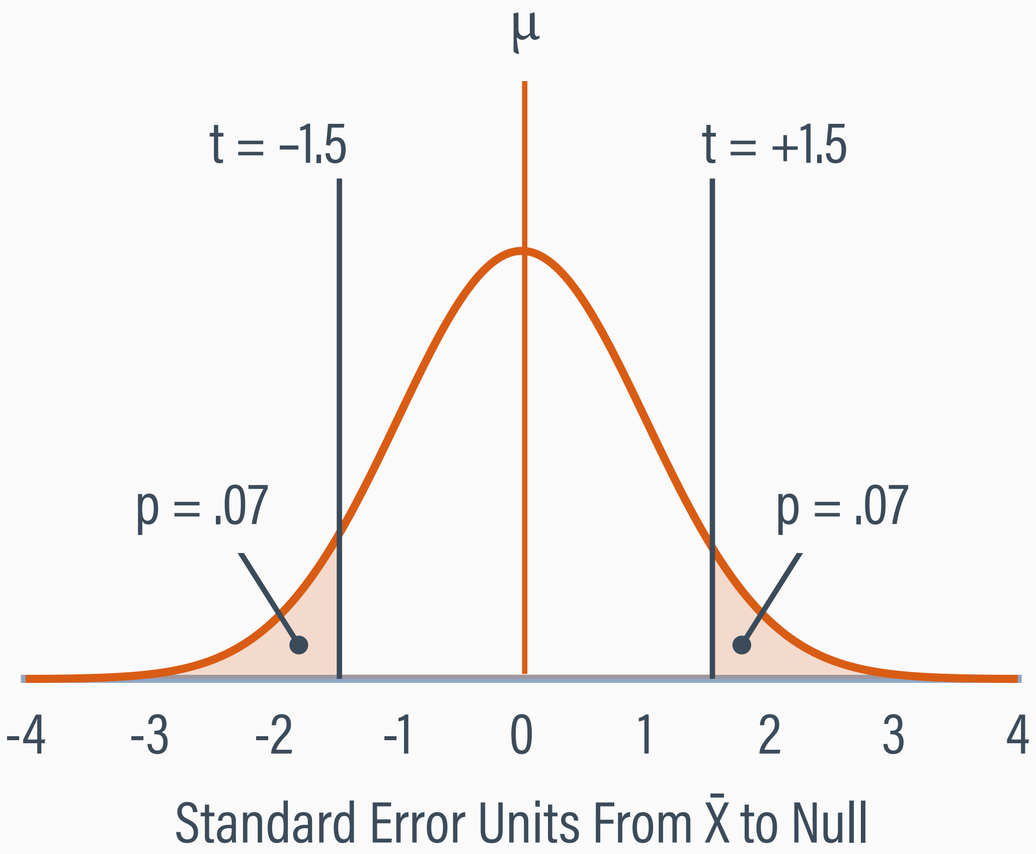
- 16% of all hypothetical samples are +1 or more standard error units above the null mean, and 16% are –1 or more standard errors below
- The two-tailed probability of drawing a sample from the null population with a t-statistic of at least ±1 is .32 (.16 positive + .16 negative)



P-VALUE AT ± 1.5 STANDARD ERRORS

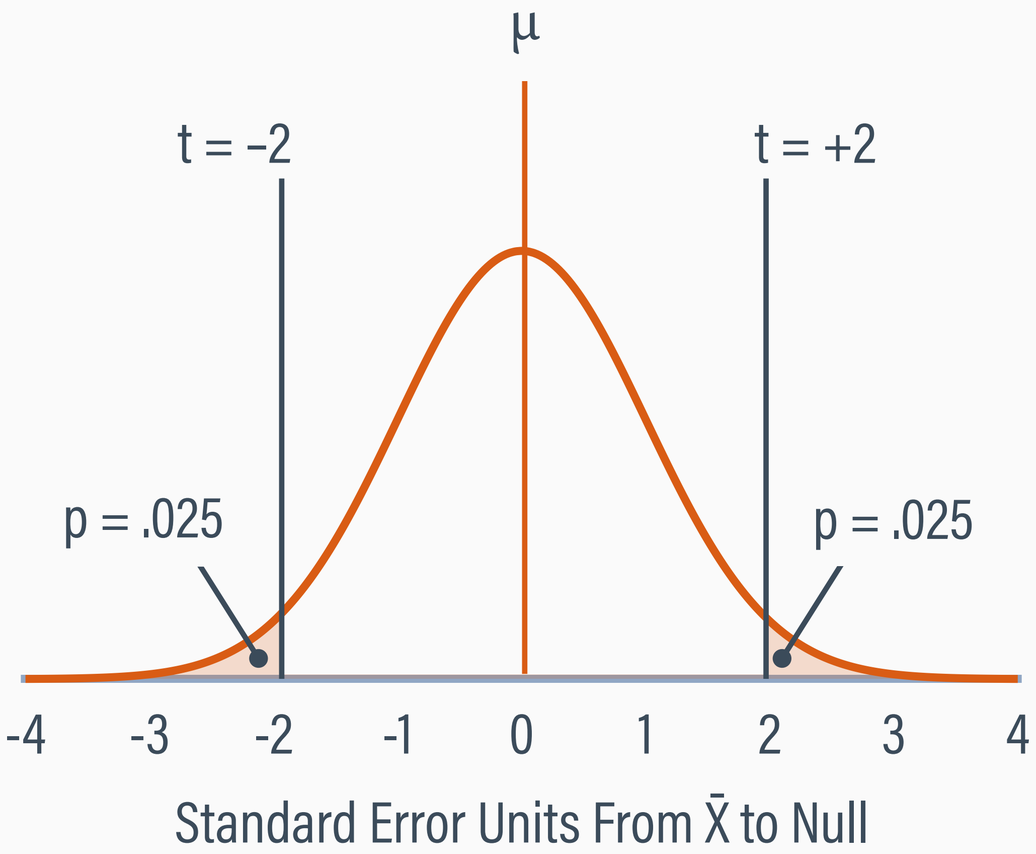
- 7% of all hypothetical samples are +1.5 or more standard error units above the null mean, and 7% are –1.5 or more standard errors below

- The two-tailed probability of drawing a sample from the null population with a t-statistic of at least ±1 is .14 (.07 positive + .07 negative)



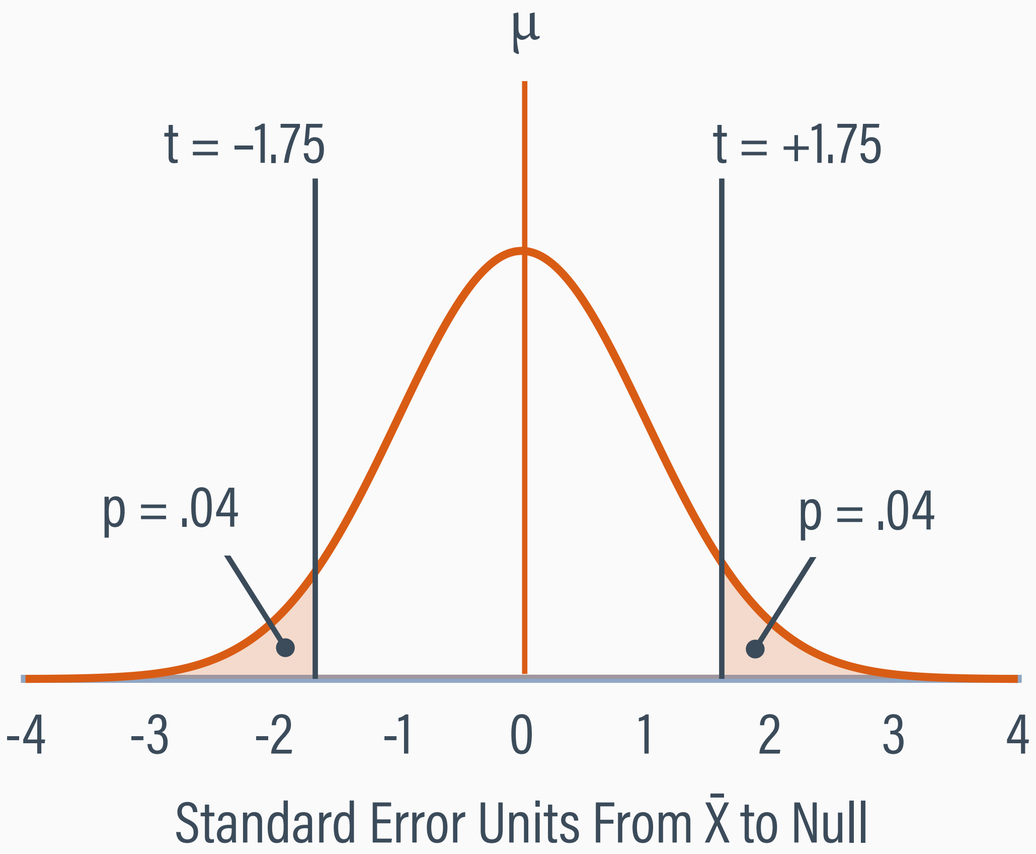
P-VALUE AT ± 2 STANDARD ERRORS

- 2.5% of all hypothetical samples are at least +1.96 standard error units above the null mean, and 2.5% are –1.96 or more below
- The two-tailed probability of drawing a sample from the null population with a t-statistic of at least ±1.96 is .05 (.025 positive + .025 negative)



TWO-TAILED P-VALUE FOR DEPRESSION DATA

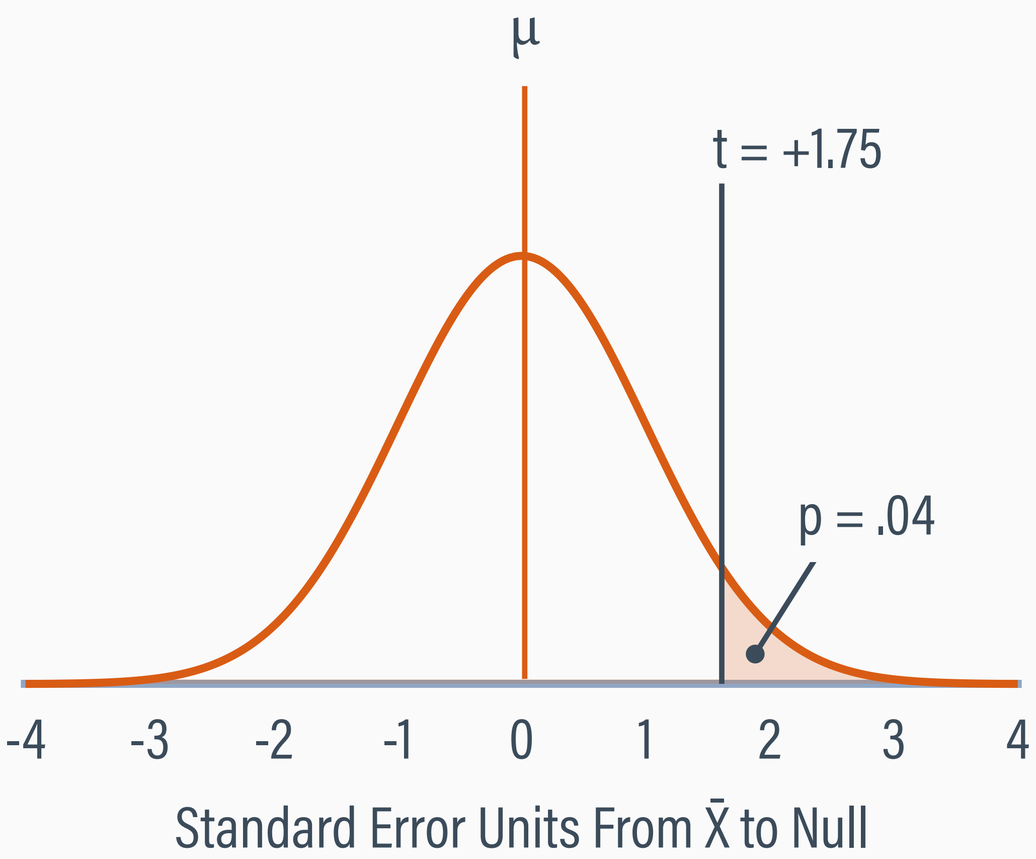
- Depress mean: $\bar{X} = 18.10$, $t = 1.75$
- The two-tailed probability of drawing a sample from the null population with a t-statistic of at least ± 1.75 is $p = .08$
- The probability of drawing a sample mean at least as extreme as ours from a null population with $\mu = 16$ is about 8%



Consider the following statement, which is a common misinterpretation about the probability value: p = 0.08 means there's a 8% chance the null is true (and 92% chance the alternative is true). In small groups of two or three, discuss why this interpretation is incorrect. Hint: Think about the frequentist statistical paradigm.

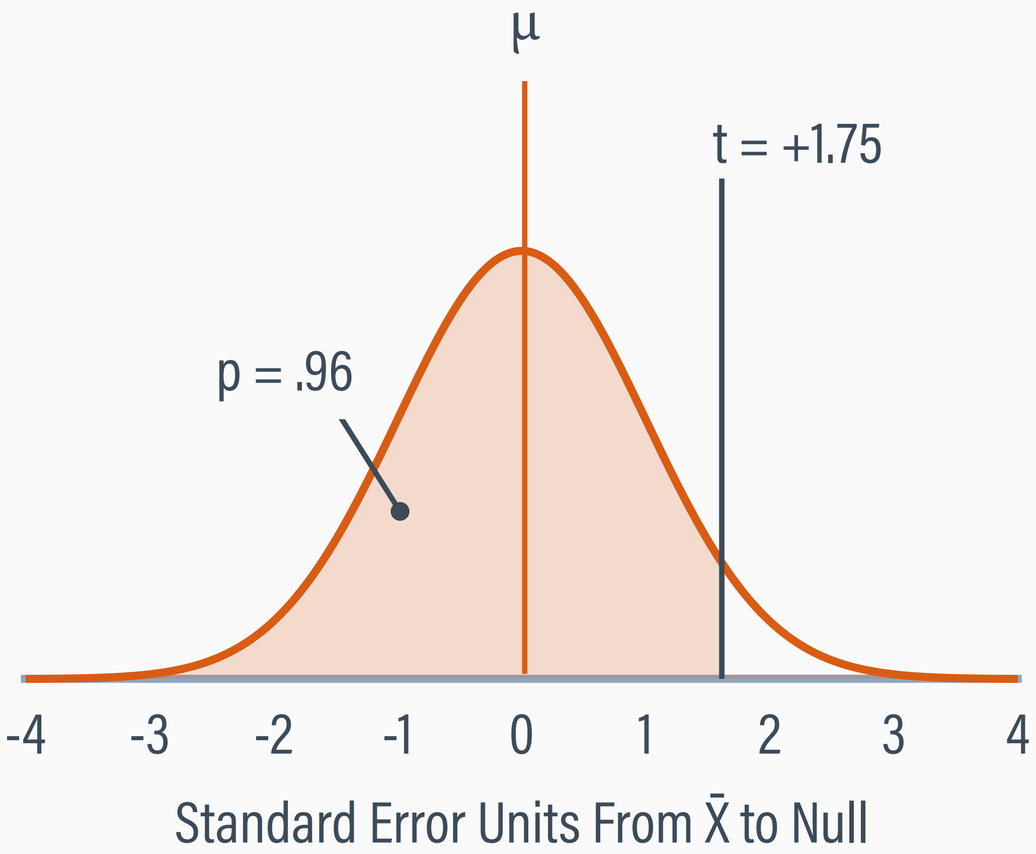
ONE-TAILED P-VALUE FOR DEPRESSION DATA

- Depress mean: $\bar{X} = 18.10$, $t = 1.75$
- Suppose the researchers specified a one-tailed hypothesis where only an increase in depression could refute the null
- The one-tailed probability of drawing a sample from the null population with a t-statistic of at least +1.75 is $p = .04$



ONE-TAILED P-VALUE FOR DEPRESSION DATA

- Depress mean: $\bar{X} = 18.10$, $t = 1.75$
- Suppose the researchers specified a one-tailed hypothesis where only an decrease in depression could refute the null
- The sample mean is in the wrong direction!
- The one-tailed probability of drawing a sample from the null population with a t-statistic *lower than* +1.75 (in the hypothesized direction) is $p = .96$



SIGNIFICANCE TESTING STEPS

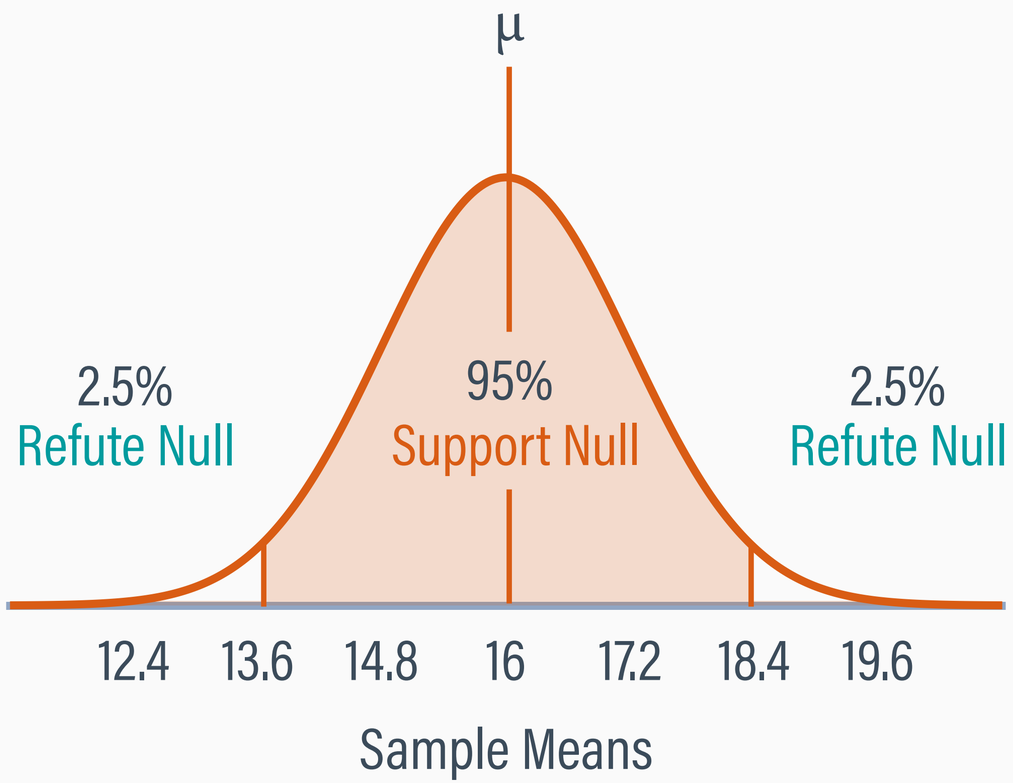
1. Specify hypotheses about population
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

RESEARCH QUESTION REVISITED

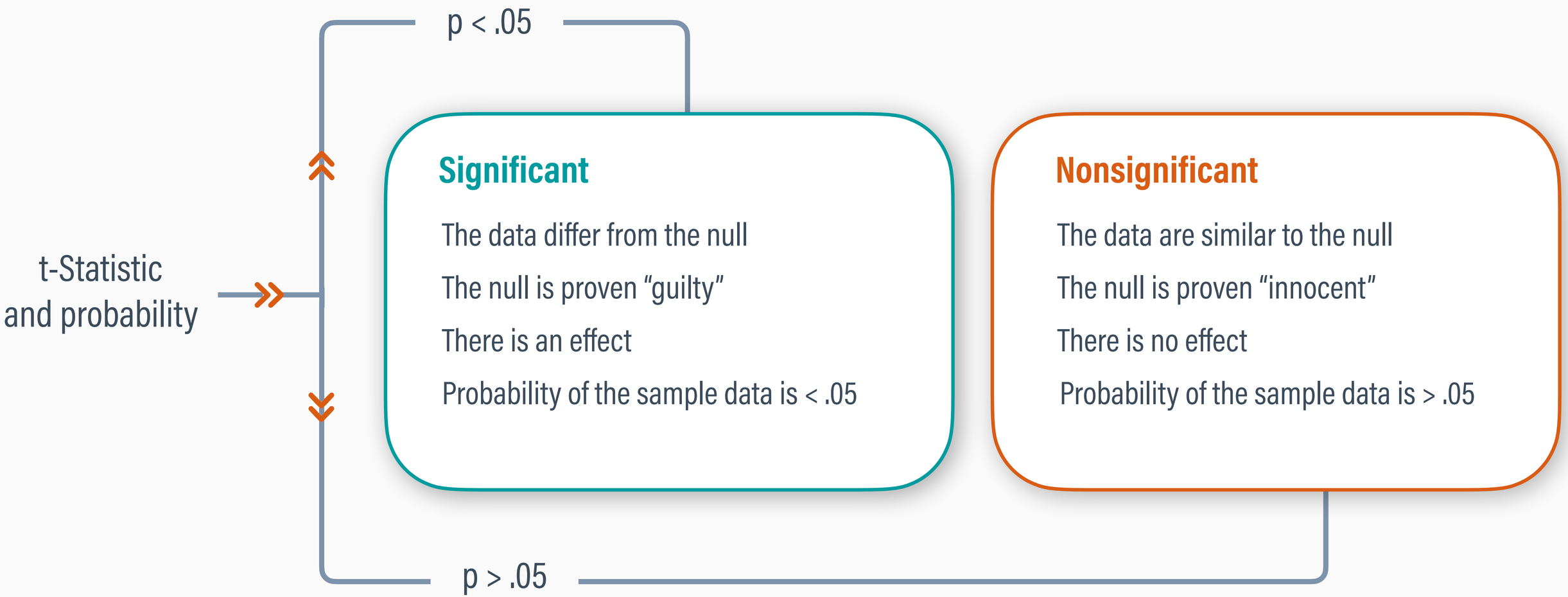
- Studies typically attempt to answer a handful of research questions involving associations between key variables
- Do people who receive a positive cancer diagnosis experience clinical levels of depressive symptoms, or do they show no meaningful elevation?
- CES-D scores > 16 are widely viewed as indicating risk for clinical depression, and scores < 16 are in the mild range

5% SIGNIFICANCE CRITERION REVISITED

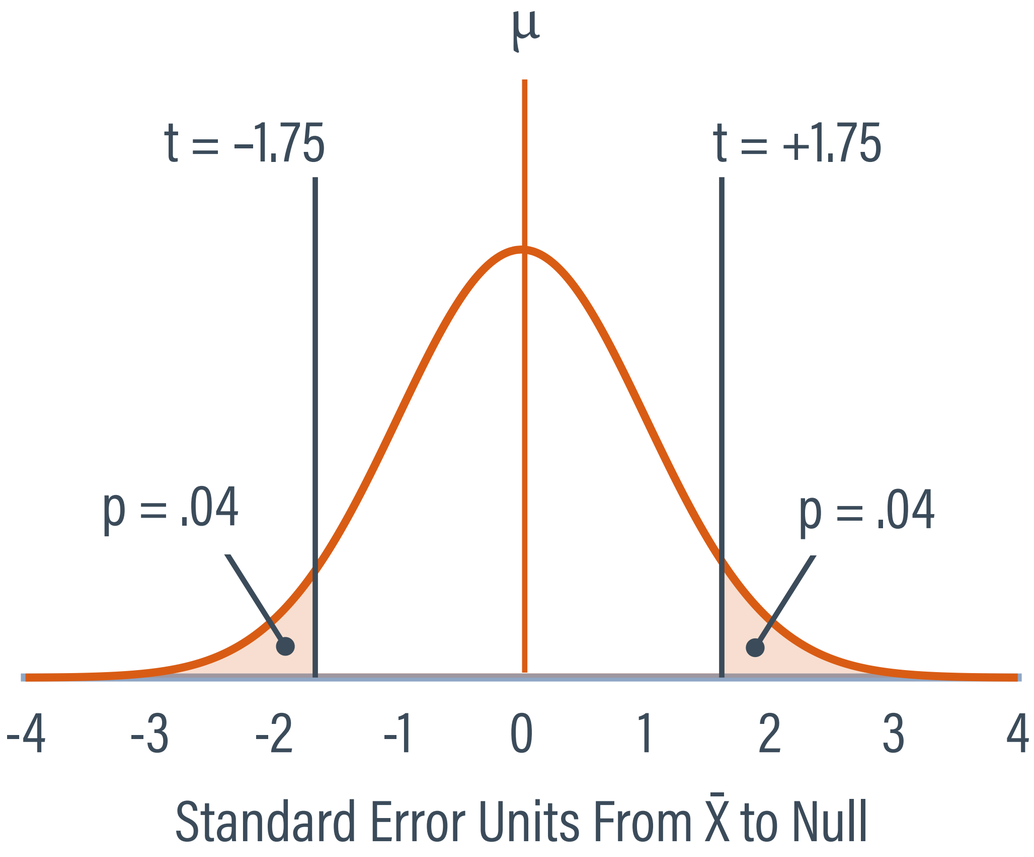
- By convention, we refute the null if the sample $\bar{X}$ falls outside the middle 95% of the sampling distribution ($p < .05$)
- Such as sample has less than a 5% chance of originating from the null population
- We deem the null implausible because our data are unlikely to originate from that population



DECISION TREE



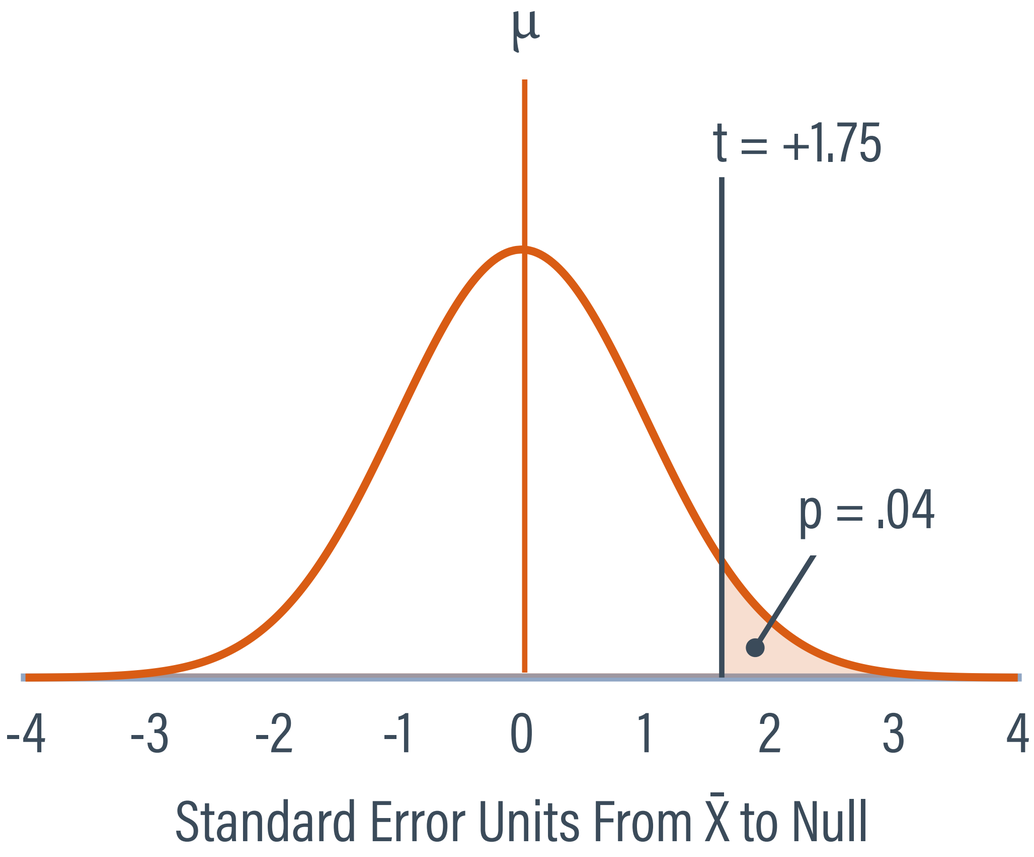
The two-tailed probability for the depression study is p = .08. In small groups of two or three, discuss your decision about the the null hypothesis. Translate your decision into a tangible statement about the effect of a cancer diagnosis on depressive symptoms.



CONCLUSION: TWO-TAILED ALTERNATE

- The p-value of .08 (8%) would lead us to support the null
- A mean as large as $\bar{X} = 18.10$ could have plausibly originated from a null population with $\mu = 16$
- There is not evidence of clinical-level depression symptoms, nor is there evidence that depression is at subclinical levels

Suppose the researchers specified a one-tailed hypothesis where only an increase in depression could refute the null. The one-tailed probability for the depression study is p = .04. In small groups of two or three, discuss your decision about the the null hypothesis. Translate your decision into a tangible statement about the effect of a cancer diagnosis on depressive symptoms.



CONCLUSION: ONE-TAILED ALTERNATE

- The p-value of .04 (4%) would lead us to refute the null
- A mean as large as $\bar{X} = 18.10$ is unlikely to have originated from a null population with $\mu = 16$
- We have evidence that a cancer diagnosis is associated with clinically elevated in depression

OUTLINE

1. Quick review
2. Overview of NHST
3. Significance testing steps
4. Study questions
5. R analysis

STUDY QUESTIONS

Use the following research scenario to answer the study questions:

Researchers want to evaluate a new medication for smoking cessation. The outcome variable, breath carbon monoxide, is a common biomarker of smoking behavior in clinical trials. The researchers administer the medication to a sample of $N = 165$ participants, and they obtain a sample mean of $\bar{X} = 5.5$. It is widely believed that breath CO levels of 5 reflect successful outcomes in clinical trials. To evaluate the medication, you will perform significance testing steps assuming a population with a true mean of $\mu = 5$ (i.e., a population where the trial was deemed a success).

STUDY QUESTIONS (1)

1. State the null hypothesis, both as a sentence and using statistical symbols.

2. State the two-tailed alternate hypothesis, both as a sentence and using statistical symbols.

STUDY QUESTIONS (2)

3. State a one-tailed alternate hypothesis of your choosing (devise one that you think makes sense), both as a sentence and using statistical symbols.

4. Explain whether you think a one- or two-tailed alternate hypothesis is most appropriate for this research scenario.

STUDY QUESTIONS (3)

5. The sampling distribution under the null hypothesis plays a vital role in hypothesis testing. Explain how the 5% significance criterion is applied to this distribution, and how it is used to decide whether to reject the null hypothesis.

6. A sample of N = 165 participants had a mean of $\bar{X} = 5.5$. The test statistic was t = +1.10. Explain what the t-statistic measures. What do the sign and the magnitude of the t-statistic indicate about the plausibility of the null hypothesis?

STUDY QUESTIONS (4)

7. Researchers report the results as “statistically significant.” What does this imply about whether a sample mean of $\bar{X} = 5.5$ is likely to have come from a null population with $\mu = 5$?

8. Researchers report the results as “statistically non-significant.” What does this imply about whether a sample mean of $\bar{X} = 5.5$ is likely to have come from a null population with $\mu = 5$?

STUDY QUESTIONS (5)

9. The two-tailed p-value was .25. Provide an interpretation of the probability value.

10. Consider the following statement, which is a common misinterpretation about the probability value: p = .25 means there's a 25% chance the null is true (and 75% chance the alternative is true). Discuss why this interpretation is incorrect.

11. Still referring to the p-value of .25, what does this tell you about whether the trial could be deemed effective?

OUTLINE

1. Quick review
2. Overview of NHST
3. Significance testing steps
4. Study questions
5. R analysis

LOAD PACKAGES AND IMPORT DATA

= data frame name
= variable name
= raw data file name

```
# LOAD R PACKAGES ----

# load R packages
library(ggplot2)
library(psych)
library(summarytools)

# READ DATA ----

# github url for raw data
filepath <-
  'https://raw.githubusercontent.com/craigenders/psych250a/main/data/CancerPositiveData.csv'

# create data frame called Cancer from github data
Cancer <- read.csv(filepath, stringsAsFactors = T)
```

SUMMARIZING DATA

= data frame name

= variable name

```
# INSPECT DATA ----

# summarize entire data frame (summarytools package)
dfSummary(Cancer)

# DESCRIPTIVE STATISTICS ----

# descriptive statistics for entire data frame (psych package)
describe(Cancer)
```

OUTPUT

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Participant	1	107	161.06	84.89	165.00	163.53	100.82	6.00	299.00	293.00	-0.13	-1.15	8.21
Diagnosis*	2	107	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00	NaN	NaN	0.00
Age	3	107	58.96	12.84	60.00	59.40	11.86	19.00	89.00	70.00	-0.40	0.51	1.24
Gender*	4	107	1.45	0.50	1.00	1.44	0.00	1.00	2.00	1.00	0.20	-1.98	0.05
Comorbids	5	107	0.94	0.75	1.00	0.93	1.48	0.00	2.00	2.00	0.09	-1.24	0.07
Optimism	6	107	8.98	2.98	9.00	9.20	2.97	0.00	14.00	14.00	-0.66	0.00	0.29
Depression	7	107	18.10	12.45	14.00	16.55	10.38	1.00	60.00	59.00	1.12	0.80	1.20
VisImpair	8	107	5.96	2.18	5.72	5.86	2.18	2.11	12.18	10.07	0.45	-0.32	0.21

T-TESTS

= data frame name

= variable name

```
# T-TEST ----

# default two-tailed test with null mean = 16 (base R)
t.test(Cancer$Depression, mu = 16, alternative = 'two.sided')

# one-tailed test in the positive direction with null mean = 16 (base R)
t.test(Cancer$Depression, mu = 16, alternative = "greater")
```

OUTPUT

```
    One Sample t-test

data:  Cancer$Depression
t = 1.7468, df = 106, p-value = 0.08357
alternative hypothesis: true mean is not equal to 16

95 percent confidence interval:
 15.71617 20.48944

sample estimates:
mean of x
  18.1028
```

t-statistic and p-value

95% confidence interval

Sample mean

OUTPUT

```
    One Sample t-test

data:  Cancer$Depression
t = 1.7468, df = 106, p-value = 0.04178
alternative hypothesis: true mean is greater than 16

95 percent confidence interval:
 16.10528      Inf

sample estimates:
mean of x
  18.1028
```

t-statistic and p-value

95% confidence interval

Sample mean