MODULE 7

PAIRED-SAMPLES T-TEST

OUTLINE

1. Within-subjects designs
2. Quick review
3. Significance testing steps
4. Statistical assumptions
5. Study questions

OUTLINE

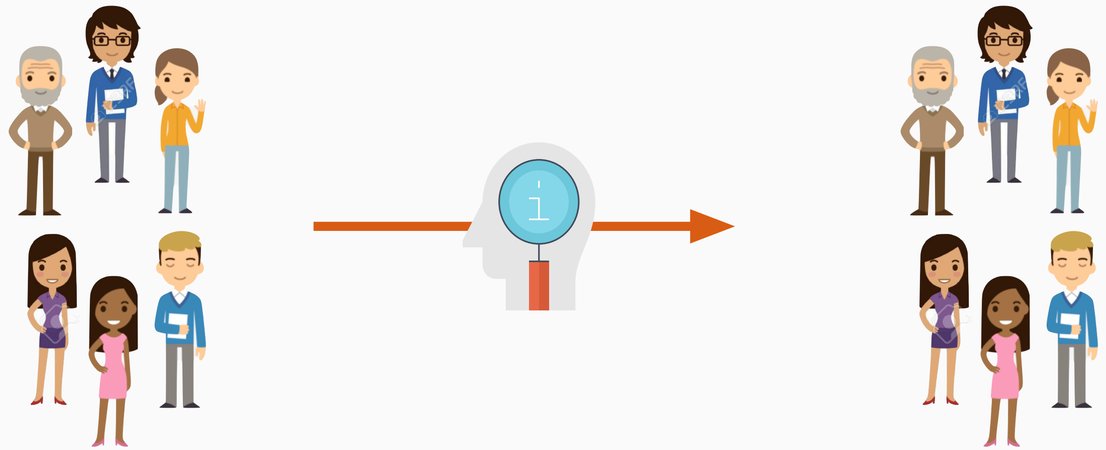
1. Within-subjects designs
2. Quick review
3. Significance testing steps
4. Statistical assumptions
5. Study questions

WITHIN-SUBJECTS RESEARCH DESIGNS

- A within-subjects research design produces two or more scores from the same participant
- A classic within-group design collects two or more repeated measurements from the same group of individuals (e.g., a pretest measure followed by a posttest)
- The scores can also come from matched pairs of scores

RANDOMIZED TRIAL APPLICATION

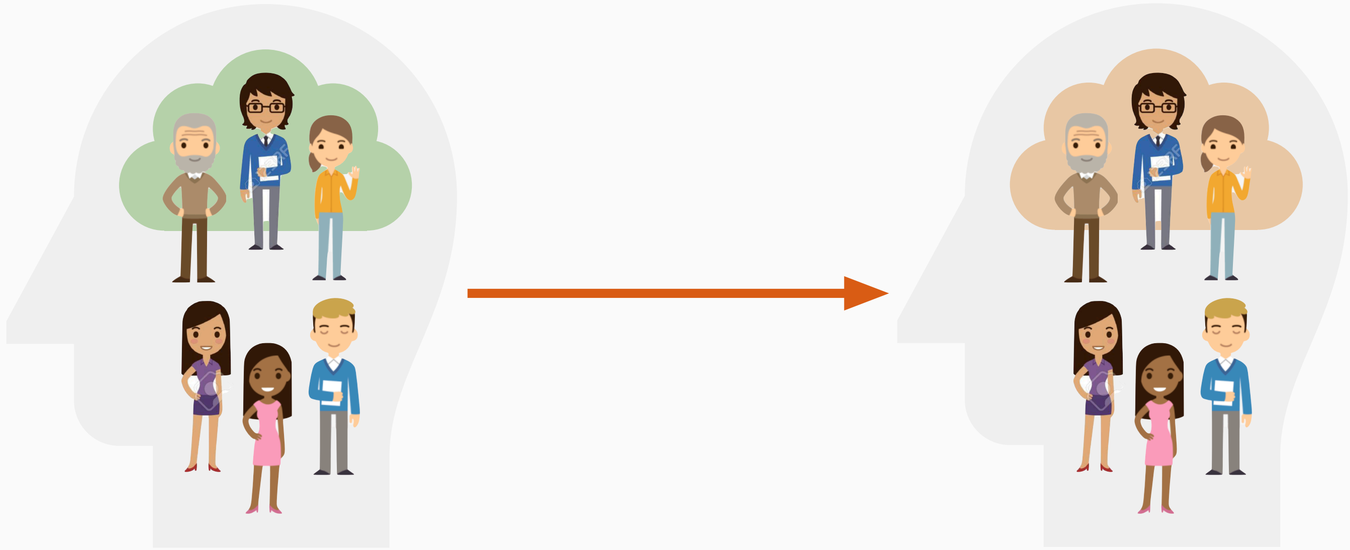
- Two scores are obtained from the same group of people before and after an intervention



Baseline Intervention Follow-up

EXPERIMENTAL APPLICATION

- The same group of people are exposed to two different experimental conditions

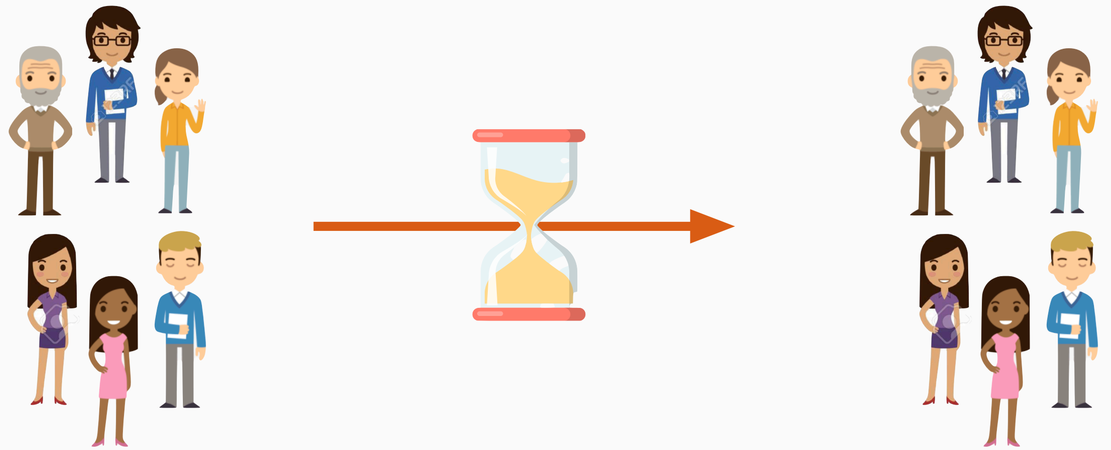


Stimuli 1

Stimuli 2

DEVELOPMENTAL APPLICATION

- The same group of people are followed over time, with the goal of examining change or development

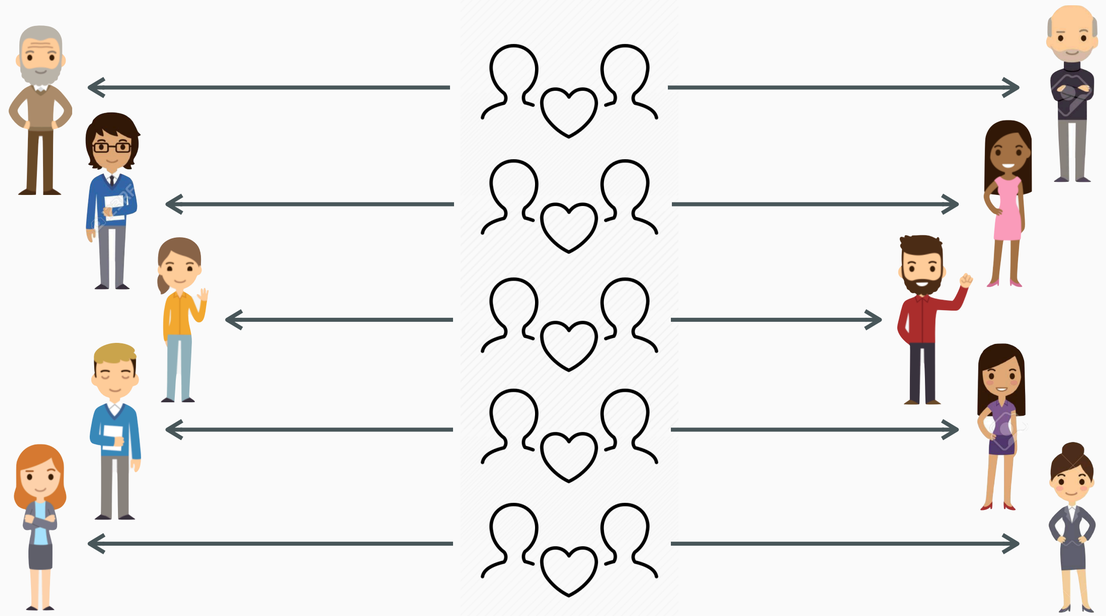


Occasion 1

Occasion 2

DYADIC APPLICATION

- Pairs of individuals forming naturally-occurring dyads (e.g., romantic partners, siblings) with linked scores

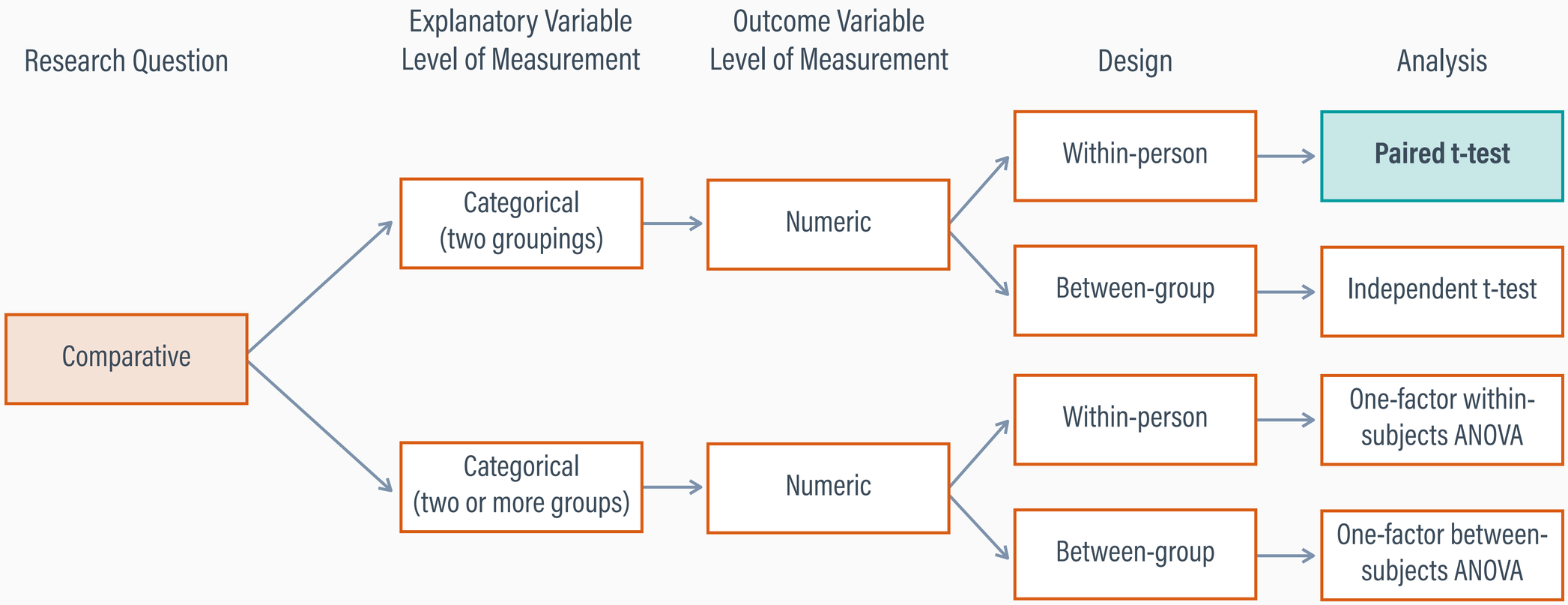


Partner 1 Partner 2

PAIRED-SAMPLES T-TEST

- The paired-samples (dependent) t-test is appropriate for within-group designs with two measurements
- Applicable to comparative research questions and hypotheses involving the difference between two means obtained from the same individuals (or nested pairs like dyads)

STATISTICAL ORG CHART



OUTLINE

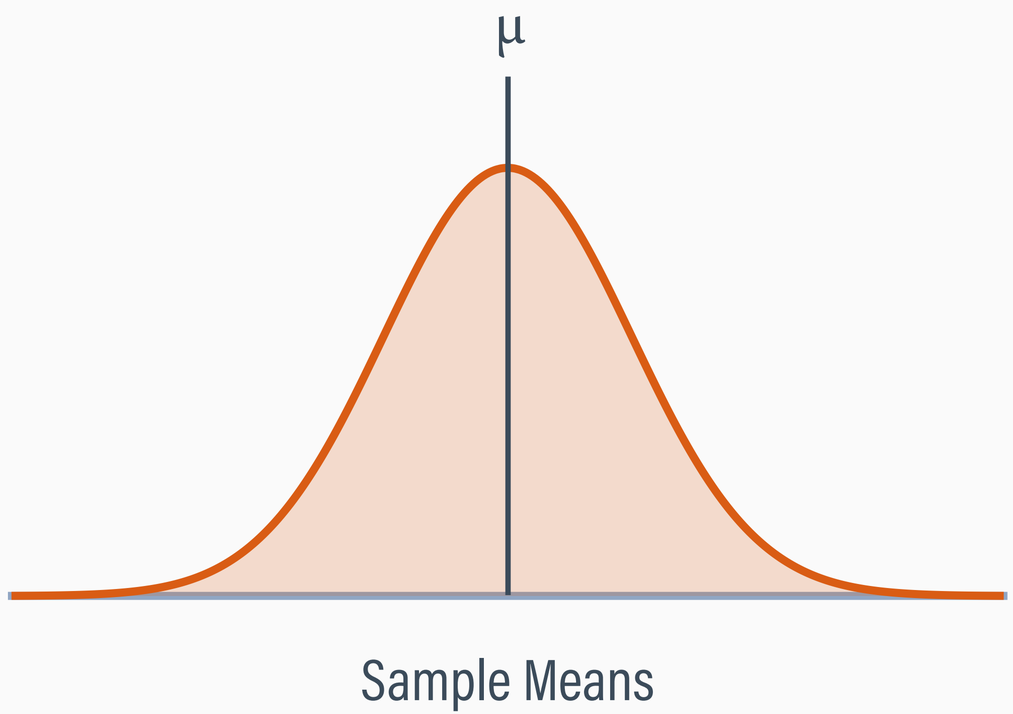
1. Within-subjects designs
2. Quick review
3. Significance testing steps
4. Statistical assumptions
5. Study questions

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

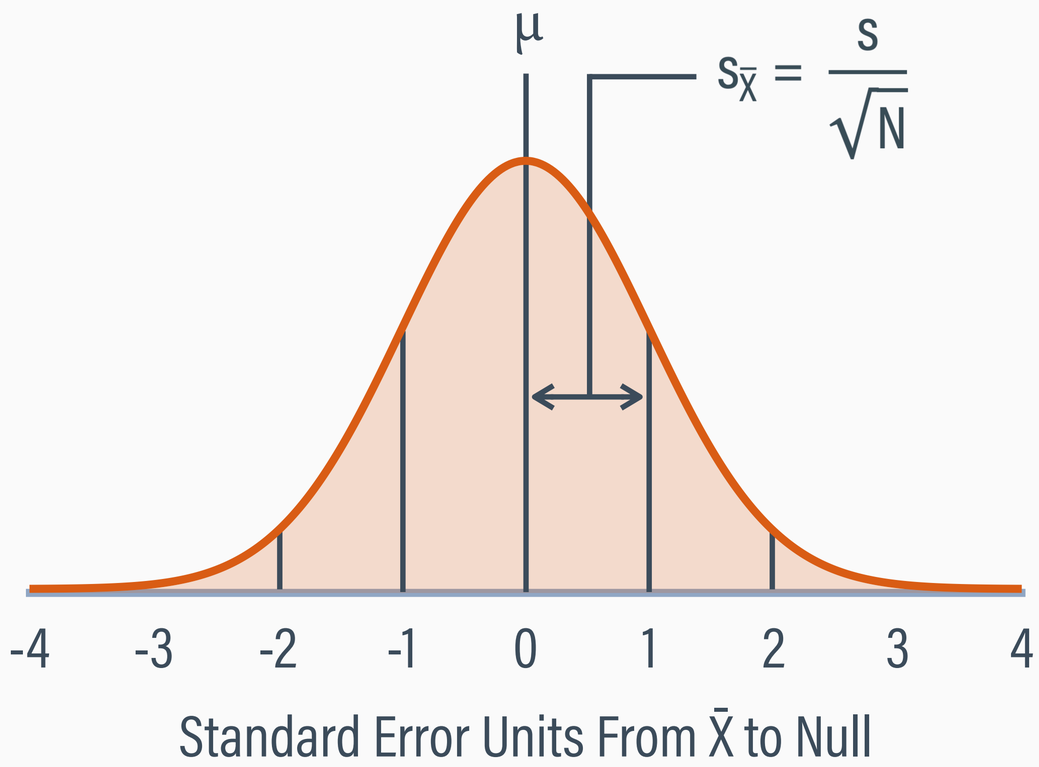
QUICK REVIEW: SAMPLING DISTRIBUTION

- 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



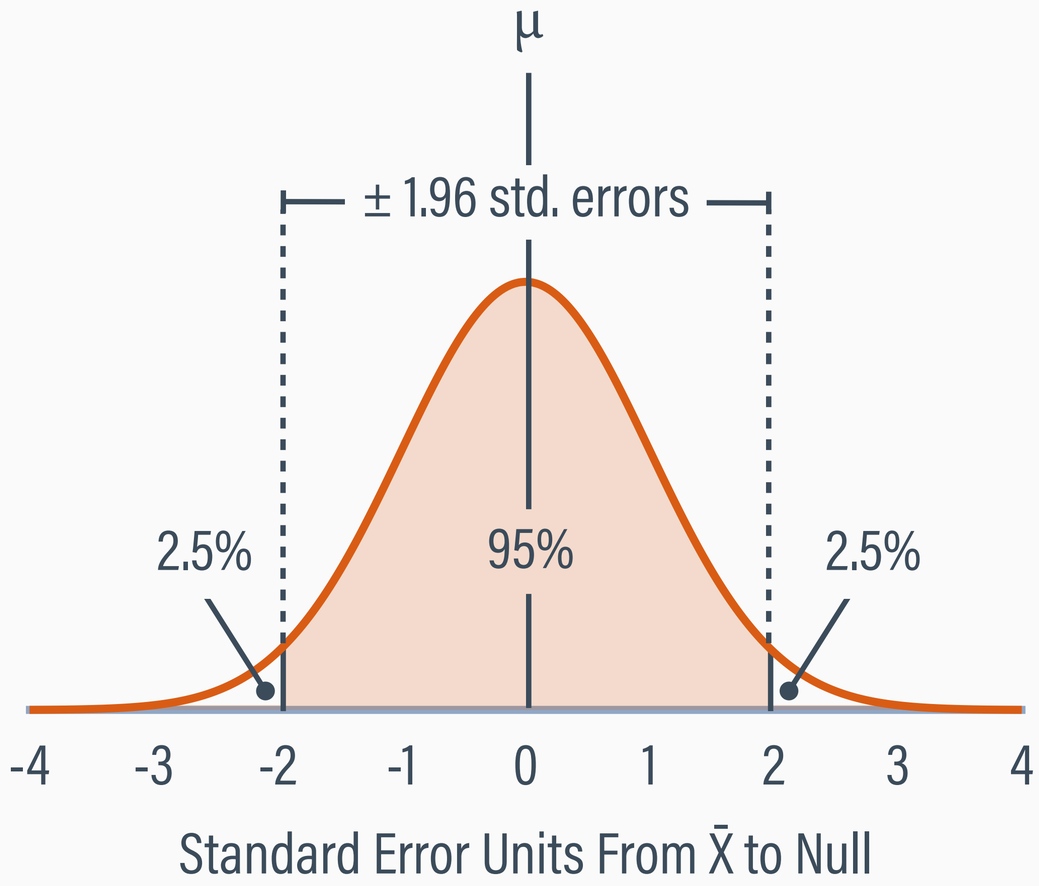
QUICK REVIEW: STANDARD ERROR

- 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



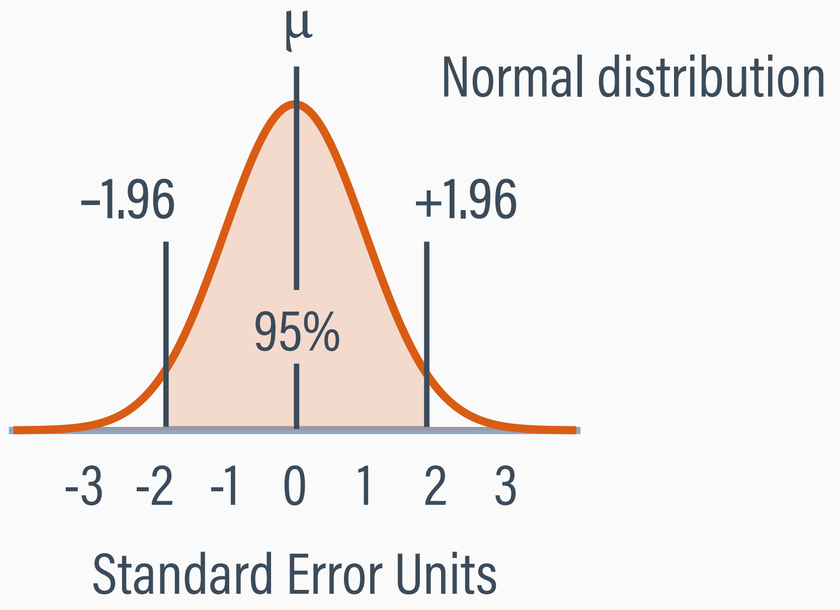
QUICK REVIEW: NORMAL CURVE RULE

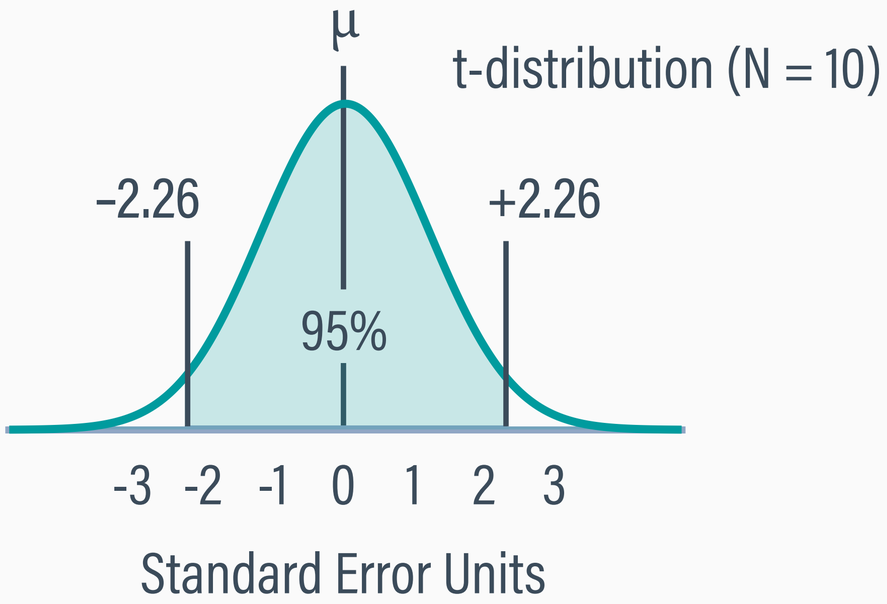
- The standard error is the standard deviation of many hypothetical sample means
- We can apply normal curve rules of thumb
- 95% of the means from large samples are within ± 1.96 standard errors of the true mean



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



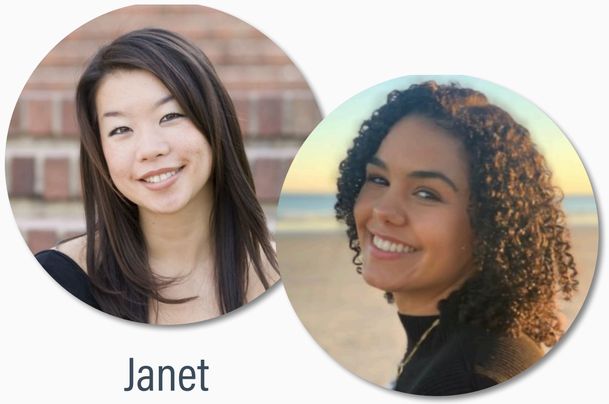


OUTLINE

1. Within-subjects designs
2. Quick review
3. Significance testing steps
4. Statistical assumptions
5. Study questions

SKIN COLOR SATISFACTION AND BINGE EATING

Although it has been demonstrated that (a) body dissatisfaction and internalization of societal appearance standards contribute to disordered eating and (b) that internalization of societal appearance standards leads to decreased skin color satisfaction among Black women, it has not been established whether skin color dissatisfaction contributes to disordered eating among Black women or girls. The objective of the present study is to determine the influence of skin color satisfaction as a potential predictor for binge eating, and its effect through body image in Black girls during the vulnerable developmental period of adolescence.



Janet Tomiyama

Jordan Parker

Parker, J.E., Enders, C.K., Mujahid, M.S., Laraia, B.A., Epel, E.S., Tomiyama, A.J. (2022). Prospective relationships between skin color satisfaction, body satisfaction, and binge eating in Black girls. *Body Image*, *41*, 342-353.

KEY VARIABLES

Body Satisfaction

Body satisfaction is the facet of self-concept associated with weight, and includes the attitudes, evaluations, and feelings an individual holds about his or her own body.

Age

The grouping variable was age. Participants were followed longitudinally, with dependent variable measured at ages 10 and 18.

RESEARCH QUESTION

- Question: Do Black girls experience a change in body satisfaction during adolescence from age 10 to 18?
- The explanatory (independent) variable, age, consists of two occasions: ages 10 and 18
- The outcome (dependent) variable, body satisfaction, is a numeric scale derived from several questionnaire items

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

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

MEAN DIFFERENCE STATISTIC

- There are population means at ages 10 and 18, μ_{10} and μ_{18}
- Hypotheses about change use a mean difference statistic that contrasts the two population means

$$\mu_{diff} = \mu_{18} - \mu_{10}$$

- The mean difference quantifies change over time in this case

NULL HYPOTHESIS

- In the population, there is no change in body satisfaction between the ages of 10 and 18

$$H_0: \mu_{diff} = 0$$

- The null that $\mu_{diff} = 0$ is counter to expectations because researchers anticipate that body satisfaction could improve or get worse during adolescence

TWO POSSIBLE ALTERNATIVE HYPOTHESES

- One-tailed alternate: Body satisfaction decreases during adolescence

$$H_A: \mu_{diff} < 0 \ (\mu_{18} < \mu_{10})$$

- Two-tailed alternate: Body satisfaction could either increase or decrease during adolescence

$$H_A: \mu_{diff} \neq 0 \ (\mu_{18} < \mu_{10} \text{ or } \mu_{18} > \mu_{10})$$

SIGNIFICANCE TESTING STEPS

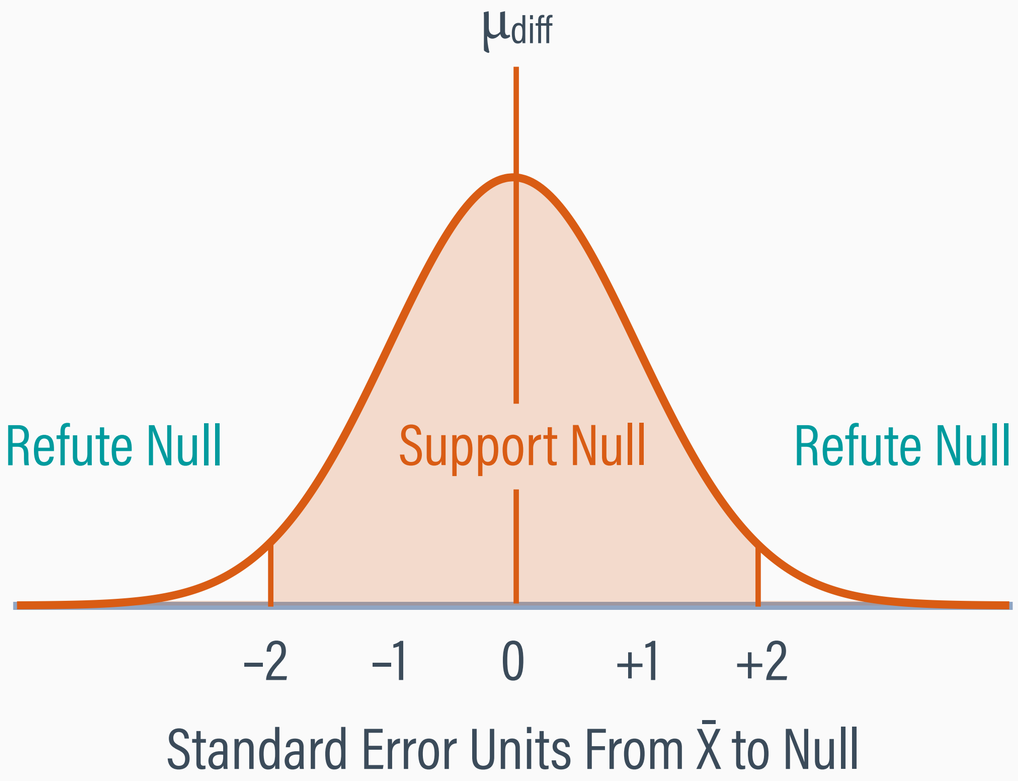
1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

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?

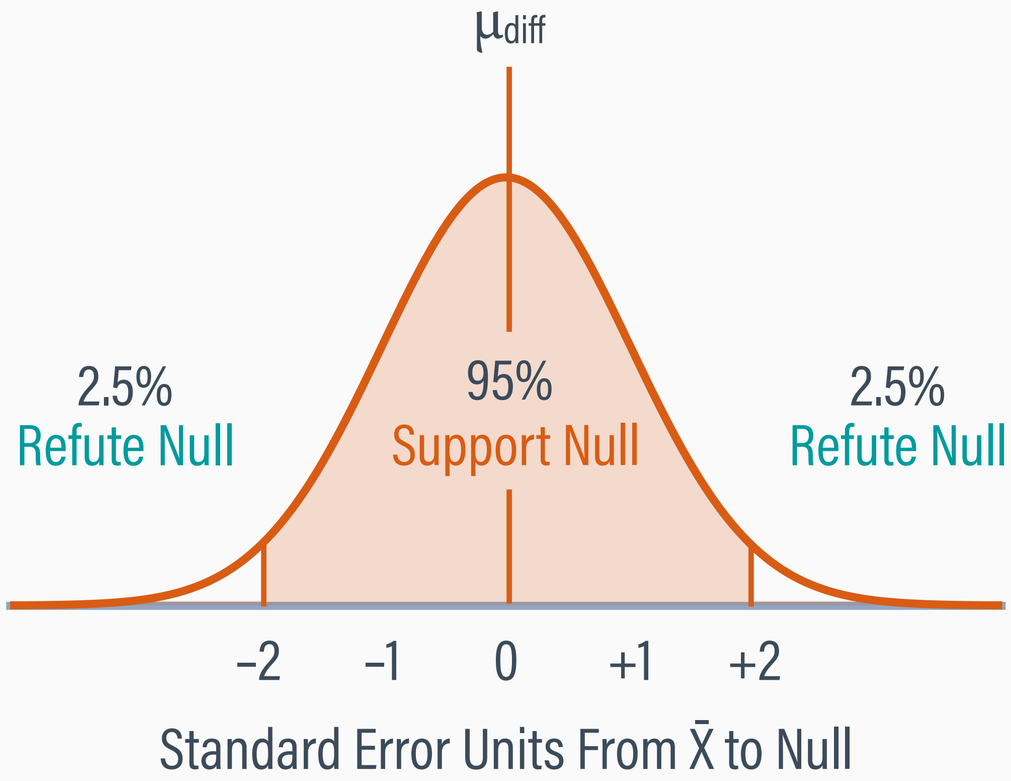
EVALUATING THE NULL

- Any $\bar{X}_{diff}$ near the middle of the sampling distribution ($\mu_{diff} = 0$) 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 $\bar{X}_{diff}$ falls far from μ_{diff}
- Such as sample has a low probability of originating from the null population



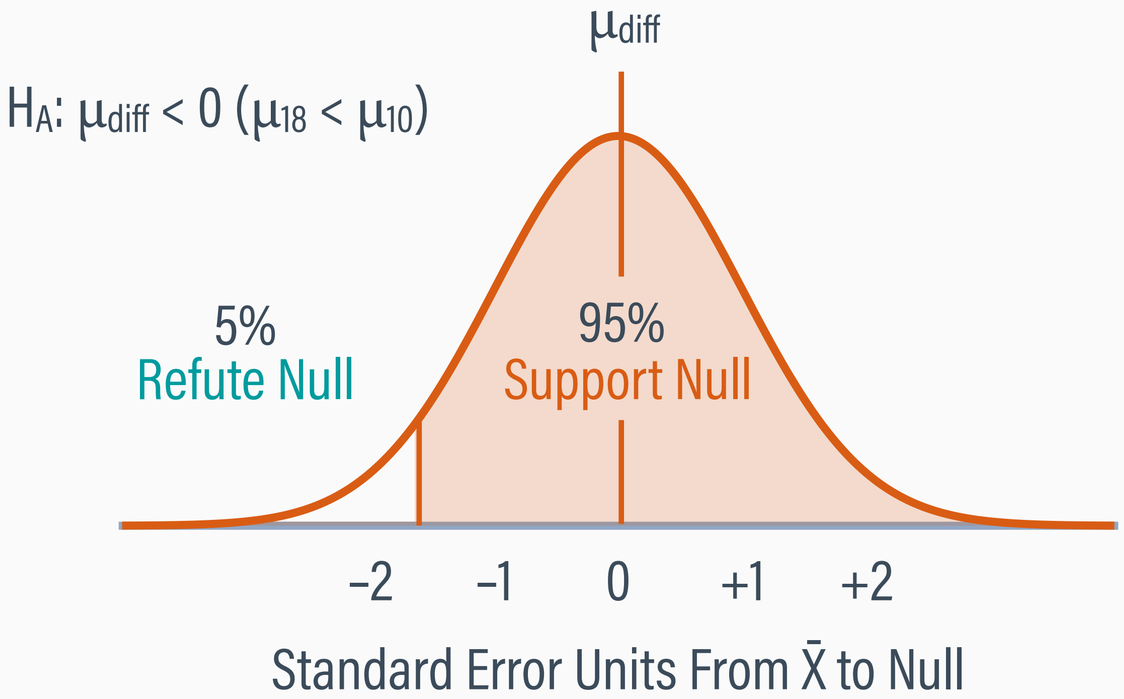
TWO-TAILED ALTERNATE HYPOTHESES

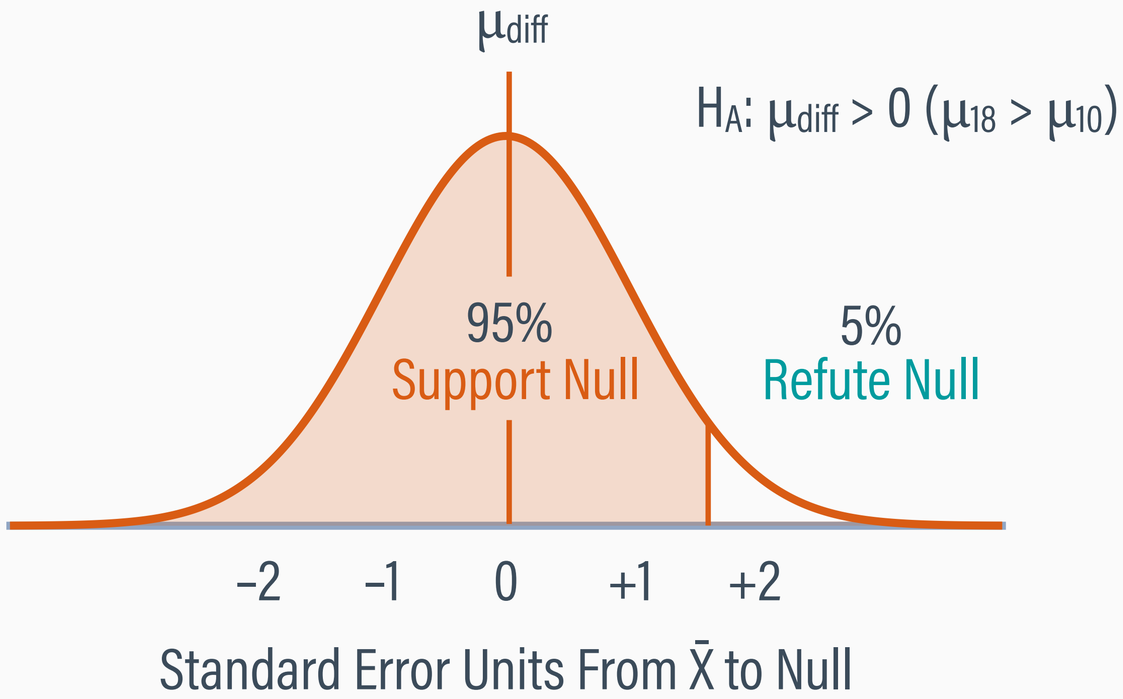
- By convention, we refute the null if the sample $\bar{X}_{diff}$ falls outside the middle 95% of the sampling distribution
- Such as sample has less than a 5% chance of originating from the null population ($p < .05$)
- The 5% rejection region (alpha level) is split in half to allow for the possibility that either an increase or a decrease provides evidence against H_0



ONE-TAILED ALTERNATE HYPOTHESES

- The 5% rejection region (alpha level) is placed in one tail, since only a positive (or negative) $\bar{X}_{diff}$ counts as evidence against H_0





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

CHANGE (DIFFERENCE) SCORES

- The paired-samples t-test converts two measurements into a single column of change (difference) scores

$$\text{BodySatCha} = \text{BodySat18} - \text{BodySat10}$$

- A positive change score indicates that scores increased, a negative scores conveys a decrease

ID	BodySat10	BodySat18	BodySatCha
1	32	32	**0**
2	29	24	**–5**
3	23	26	**3**
…	…	…	…
881	28	26	**–2**
882	25	26	**1**

ANALYSIS SUMMARY

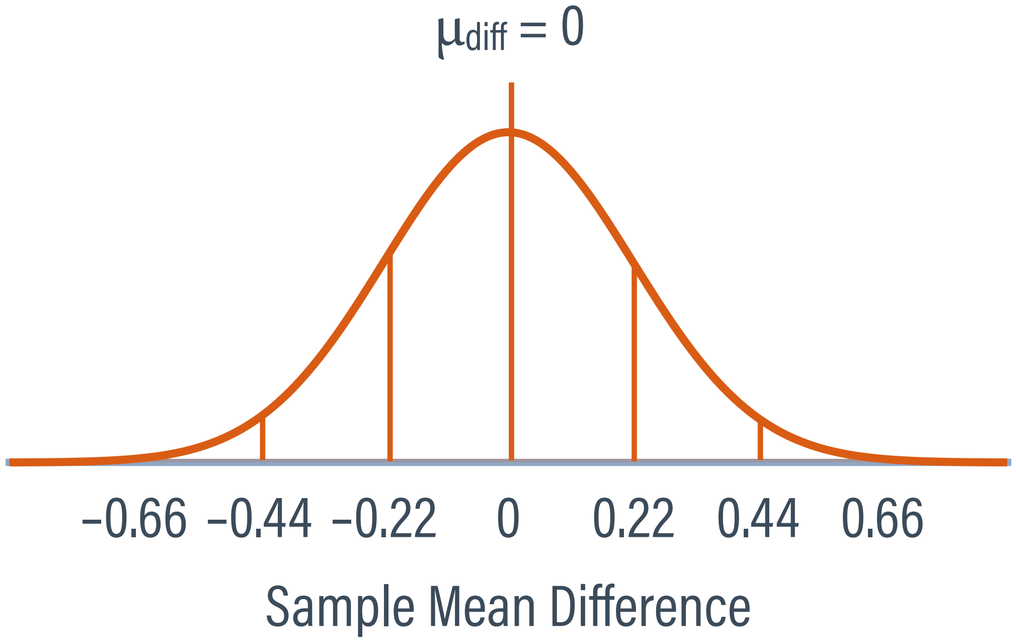
- N = 882 Black girls participated in the study
- The sample mean difference is $\bar{X}_{diff} = -3.05$, with a standard error equal to $s_{\bar{X}diff} = 0.22$
- The standard error tells us that the sample mean difference from a null population should be about ± 0.22 body satisfaction points from 0

	N	$\bar{X}$	SD	SE
Age 10	882	28.49	5.14	0.17
Age 18	882	25.44	6.06	0.20
Difference	882	−3.05	6.44	0.22

$$s_{\bar{X}_{diff}} = \frac{s_{diff}}{\sqrt{N}} = \frac{6.44}{\sqrt{882}} = 0.22$$

OUTPUT

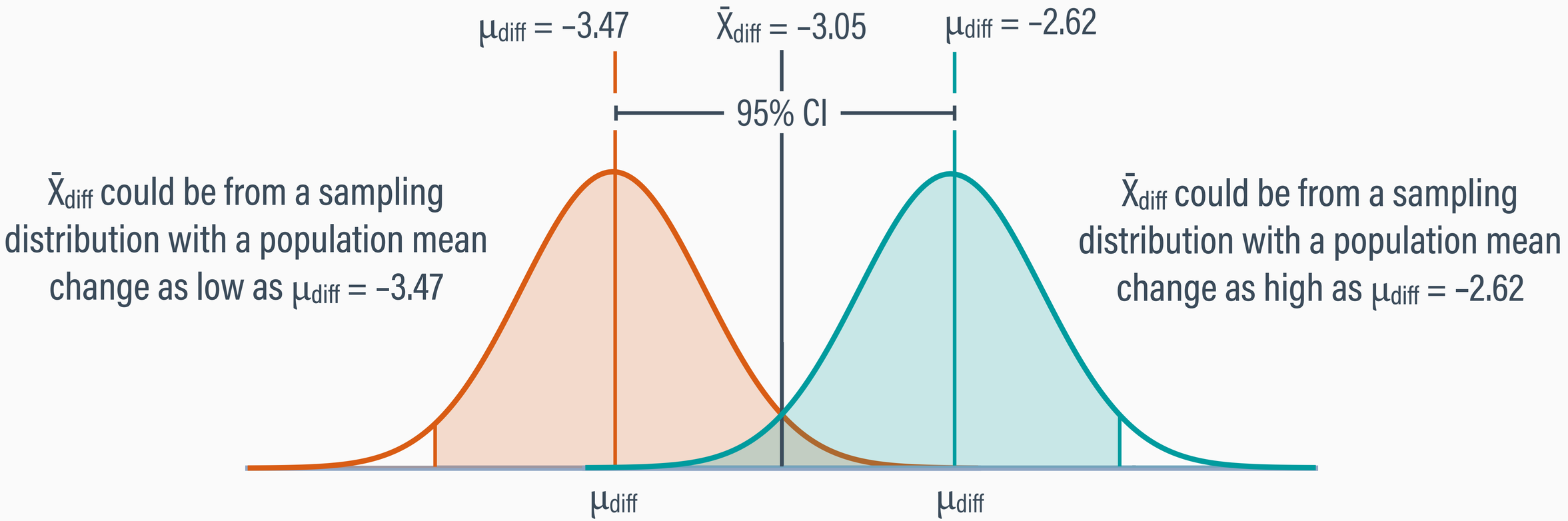
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Participant	1	882	441.50	254.76	441.50	441.50	326.91	1.00	882.00	881.00	0.00	-1.20	8.58
ParentEduc*	2	882	2.27	0.80	2.00	2.34	1.48	1.00	3.00	2.00	-0.52	-1.26	0.03
ParentIncome*	3	882	2.38	1.11	2.00	2.35	1.48	1.00	4.00	3.00	0.12	-1.33	0.04
BMI10	4	882	19.60	4.22	18.48	19.11	3.57	12.37	35.16	22.79	1.10	0.94	0.14
SkinColorSat10	5	882	3.59	0.64	4.00	3.70	0.00	1.00	4.00	3.00	-1.58	2.41	0.02
SkinColorSat18	6	882	3.50	0.65	4.00	3.58	0.00	1.00	4.00	3.00	-1.36	2.36	0.02
BodySat10	7	882	28.49	5.14	28.00	28.83	5.93	9.00	36.00	27.00	-0.61	0.36	0.17
BodySat18	8	882	25.44	6.06	26.00	25.64	5.93	9.00	36.00	27.00	-0.33	0.10	0.20
BingeEatDisorder10	9	882	2.59	1.82	2.00	2.49	1.48	0.00	8.00	8.00	0.45	-0.35	0.06
BingeEatDisorder18	10	882	1.68	1.45	1.00	1.52	1.48	0.00	8.00	8.00	0.88	0.84	0.05
BodySatCha	11	882	-3.05	6.45	-3.00	-3.09	5.93	-27.00	24.00	51.00	0.11	0.73	0.22



Consider the sampling distribution of sample means from a null population with $\mu_{diff} = 0$. A sample of 883 gave a sample mean difference of $\bar{X}_{diff} = -3.05$ ($s_{\bar{X}diff} = 0.22$). In small groups of two or three, discuss whether the data provide evidence for or against the null hypothesis.

95% CONFIDENCE INTERVAL

- The 95% confidence interval gives the two most extreme values of the population mean that could have reasonably produced these data



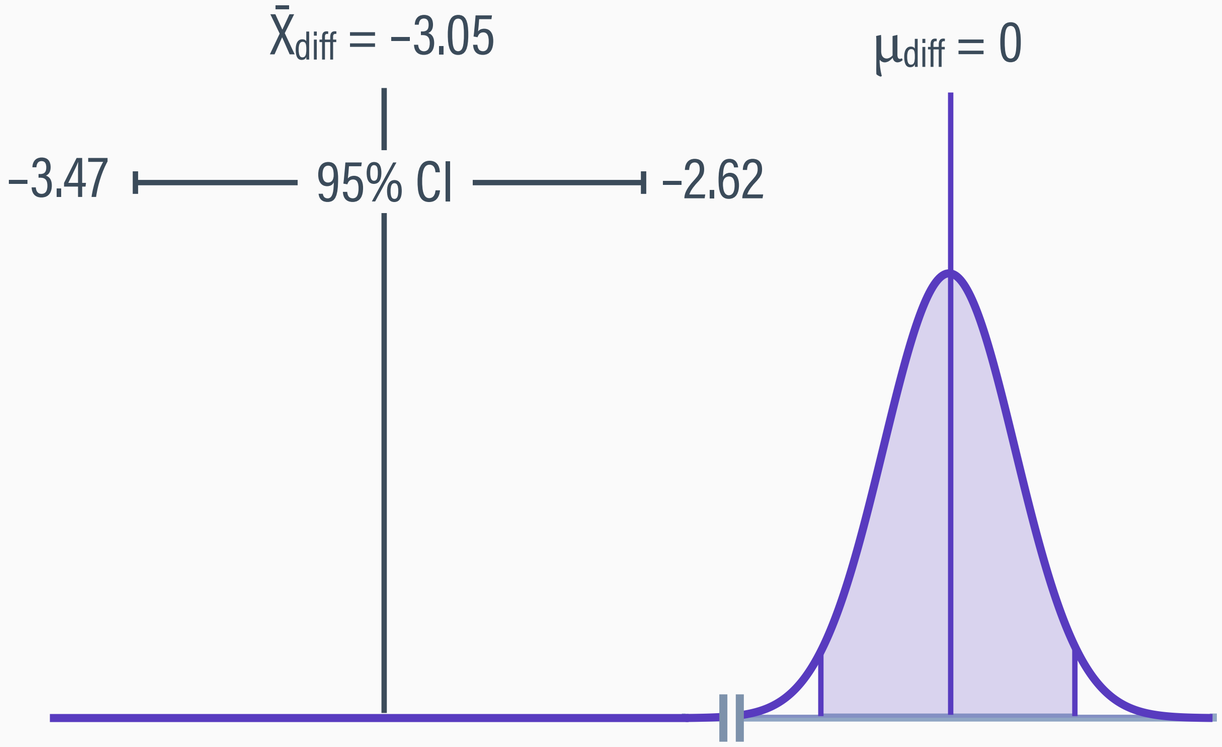
R OUTPUT

```
    Paired t-test

data:  BodySat$BodySat18 and BodySat$BodySat10
t = -14.037, df = 881, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:
 -3.472435 -2.620536

sample estimates:
mean difference
      -3.046485

standard error of mean difference: 0.2170266
```

The study produced a mean difference and 95% confidence interval of $\bar{X}_{diff} = 3.05$ and $CI_{95\%} = [-3.47, -2.62]$. In small groups of two or three, discuss whether this sample of of N = 883 participants could have reasonably originated from a population where there is truly no mean change in body satisfaction ($\mu_{diff} = 0$).

SIGNIFICANCE TESTING WITH 95% INTERVALS

- A population with a mean change of 0 is unlikely to have produced this sample because the null mean is outside the 95% interval
- The 95% confidence interval provides the same conclusion as a two-tailed significance test with a .05 significance criterion!



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 $\bar{X}_{diff}$ 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

t-STATISTIC

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

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

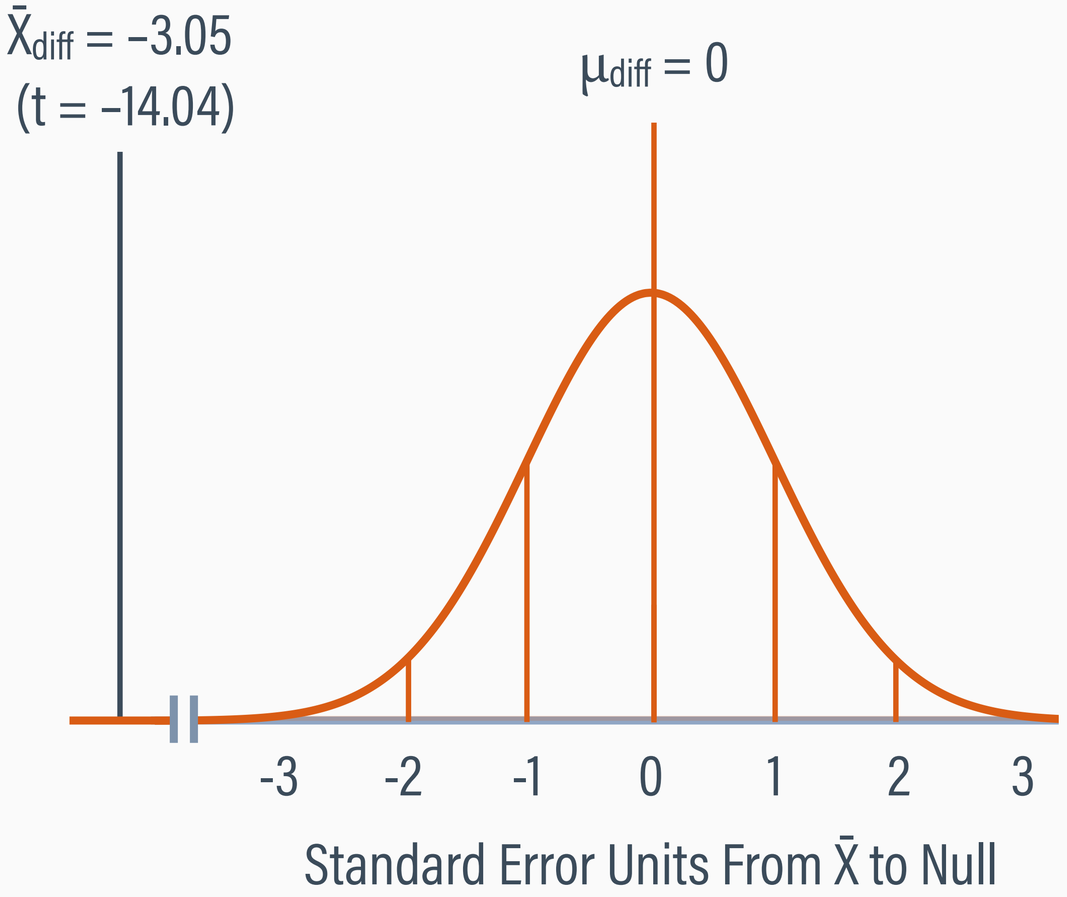
- The t-statistic is the same as a z-score (a standardized metric where distance is expressed in standard deviation units)

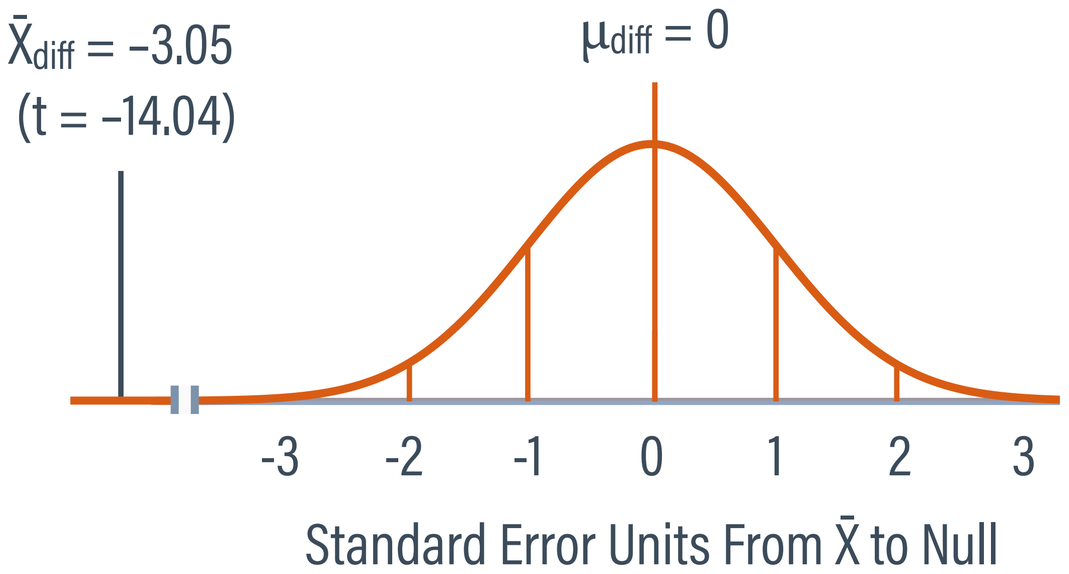
t-STATISTIC EXAMPLE

- The t-statistic indicates that $\bar{X}_{diff}$ is 14 standard error units from the null ($\mu_{diff} = 0$)

$$t = \frac{(\bar{X}_{18} - \bar{X}_{10}) - \mu_{diff}}{s_{\bar{X}_{diff}}} = \frac{-3.05 - 0}{0.22} = -14.04$$

- The negative sign means that the sample means decreased between ages 10 and 18 (satisfaction got worse)





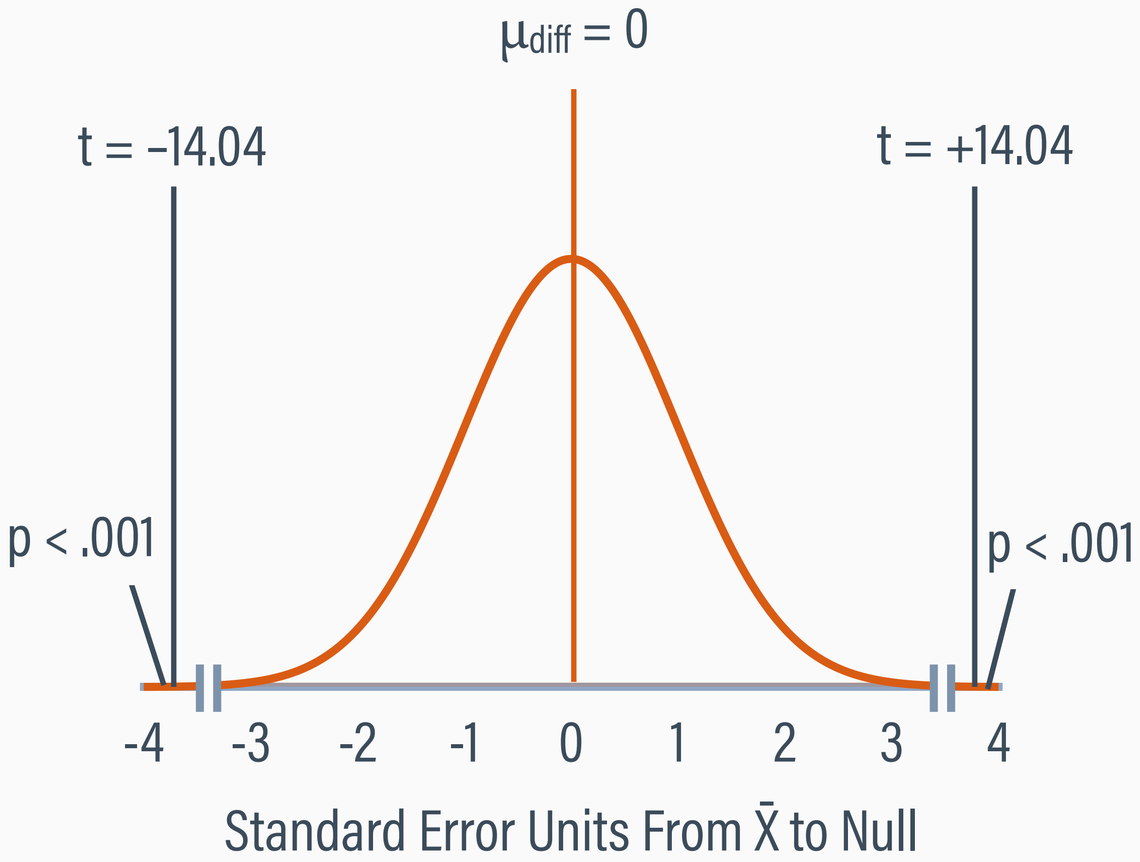
Consider the sampling distribution of sample means from a null population with $\mu_{diff} = 0$. The sample mean difference and t-statistic are $\bar{X}_{diff} = -3.05$ and $t = -14.04$. In small groups of two or three, apply the normal curve rule of thumb and decide whether the sample data provide evidence for or against the null hypothesis

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

TWO-TAILED P-VALUE

- The p-value tells how likely it is to draw a sample mean difference at least as extreme as ours from a null population with $\mu_{diff} = 0$
- The two-tailed probability of drawing a sample from the null population with a t-statistic of at least ±14.04 is $p < .001$ (less than 1 in 1000)
- Fewer than one out of every 1000 samples from a null population would have t-statistics this large



R OUTPUT

```
   Paired t-test

data:  BodySat$BodySat18 and BodySat$BodySat10
t = -14.037, df = 881, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:
 -3.472435 -2.620536

sample estimates:
mean difference
      -3.046485

standard error of mean difference: 0.2170266
```

SIGNIFICANCE TESTING STEPS

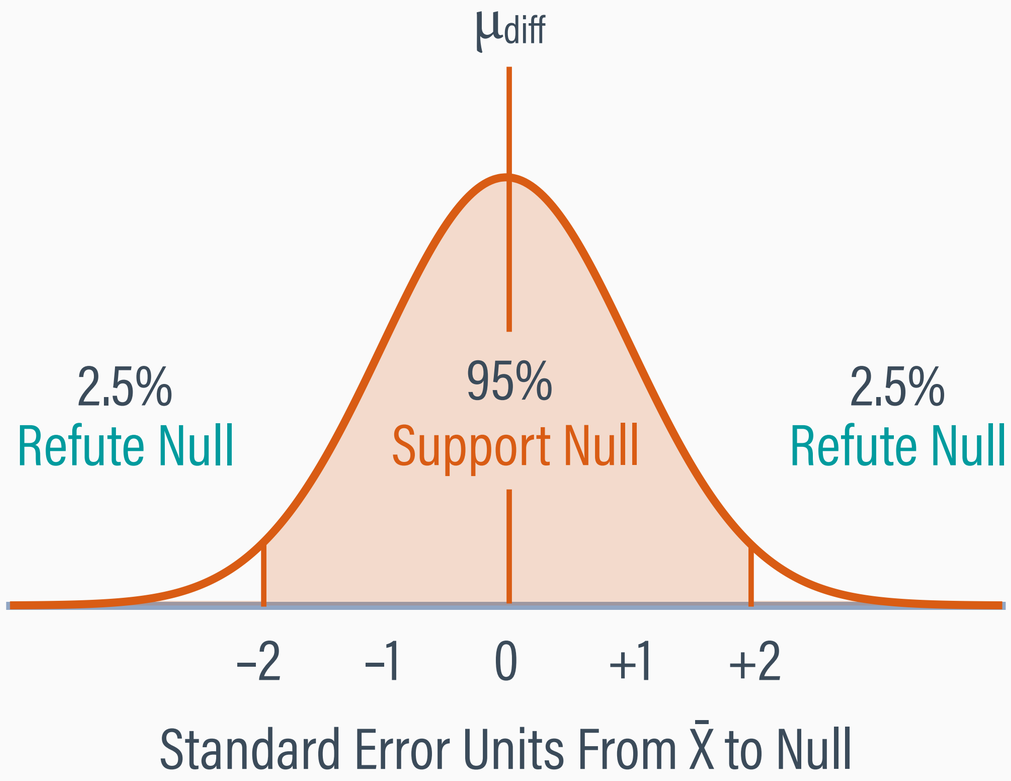
1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

RESEARCH QUESTION REVISITED

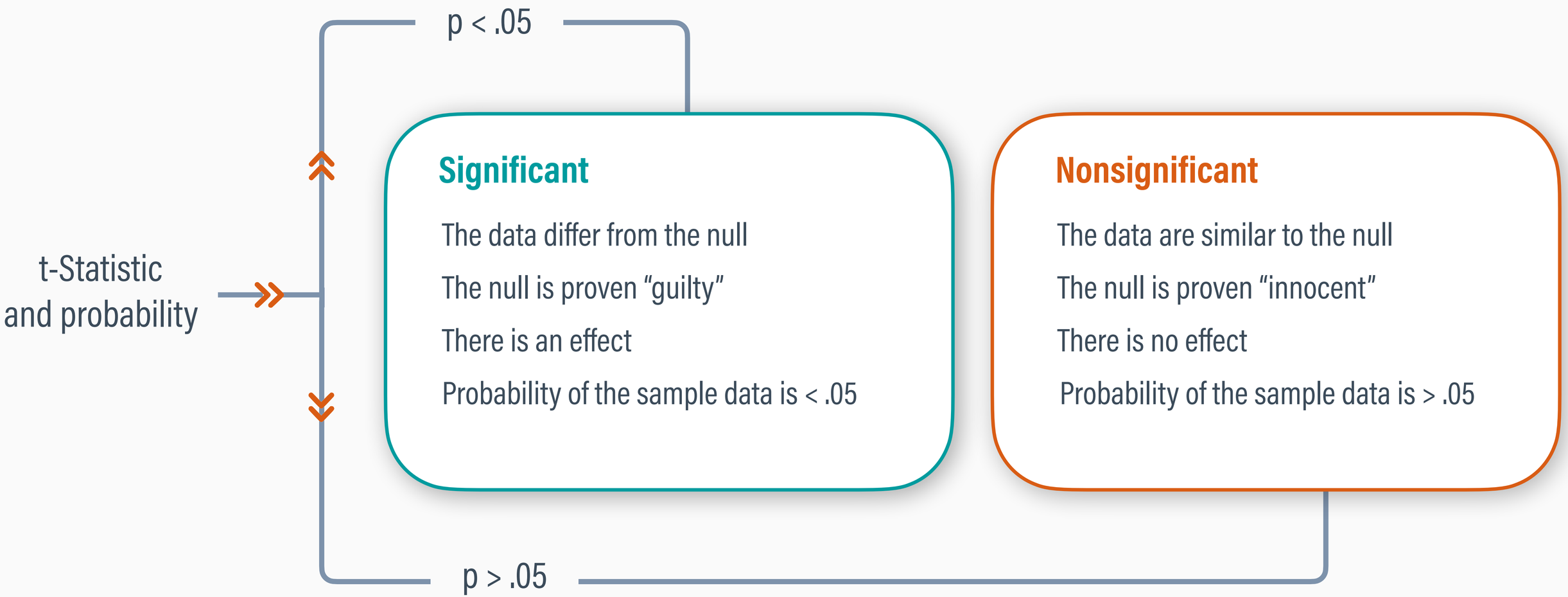
- Studies typically attempt to answer a handful of research questions involving associations between key variables
- Do Black girls experience a change in body satisfaction during adolescence from age 10 to 18?
- The null (no effect) hypothesis states that body satisfaction does not change (the population mean difference is zero)

5% SIGNIFICANCE CRITERION REVISITED

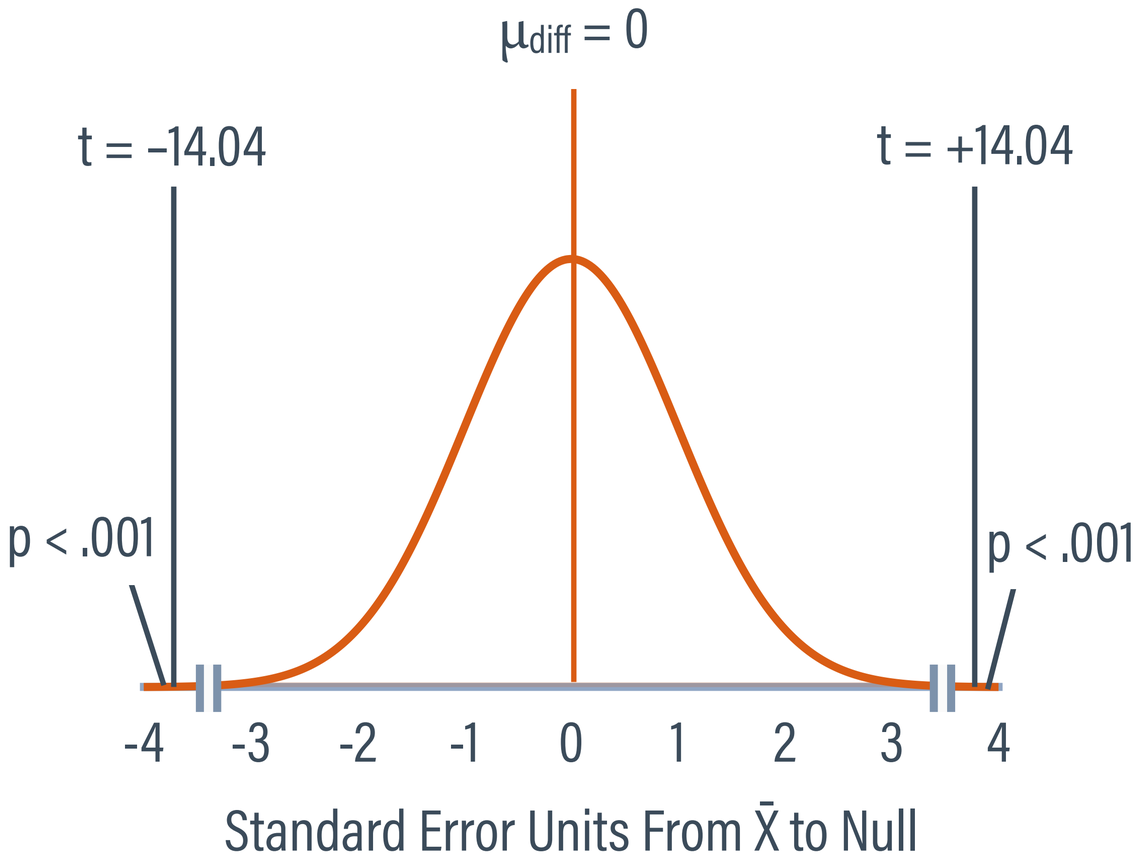
- By convention, we refute the null if the sample $\bar{X}_{diff}$ falls outside the middle 95% of the sampling distribution ($p < .05$)
- Such as sample has less than a 5% chance of originating from the null population
- We deem the null implausible because our data are unlikely to originate from that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



DECISION TREE



The two-tailed probability for the study is $p < .001$. 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 development on body satisfaction.



CONCLUSION: TWO-TAILED ALTERNATE

- The p-value of < .001 (less than 1 out of 1000) would lead us to refute the null
- A mean difference as large as $\bar{X}_{diff} = \pm 3.05$ (or a t larger than ± 14.04) is very unlikely to have originated from a null population with $\mu_{diff} = 0$
- There is evidence that Black girls experience a decrease in body satisfaction during adolescence

FALSE POSITIVES (TYPE I ERRORS)

- The 5% rejection region is an area of the distribution that contains outlier samples that are unlikely *but not impossible*
- When $\bar{X}_{diff}$ falls in the rejection region (evidence against the null), there is still a 5% chance it came from the null population
- We conclude there is a change, while acknowledging that there is a 5% chance of a false positive—incorrectly rejecting the null when it is actually true (a Type I error)

APA-STYLE ANALYSIS SUMMARY

We used a paired-samples *t*-test to examine the change in body satisfaction between the ages of 10 and 18. Table 1 gives the descriptive statistics. The analysis revealed a statistically significant decrease in body satisfaction scores, $t(881) = -14.04$, $p < .001$. The mean difference was approximately three points, and the 95% confidence interval for the mean difference ranged from –3.47 to –2.62. Finally, the standardized mean difference was just below Cohen's medium effect size benchmark ($d = 0.47$), indicating a salient developmental change.

OUTLINE

1. Within-subjects designs
2. Quick review
3. Significance testing steps
4. Statistical assumptions
5. Study questions

STATISTICAL ASSUMPTIONS

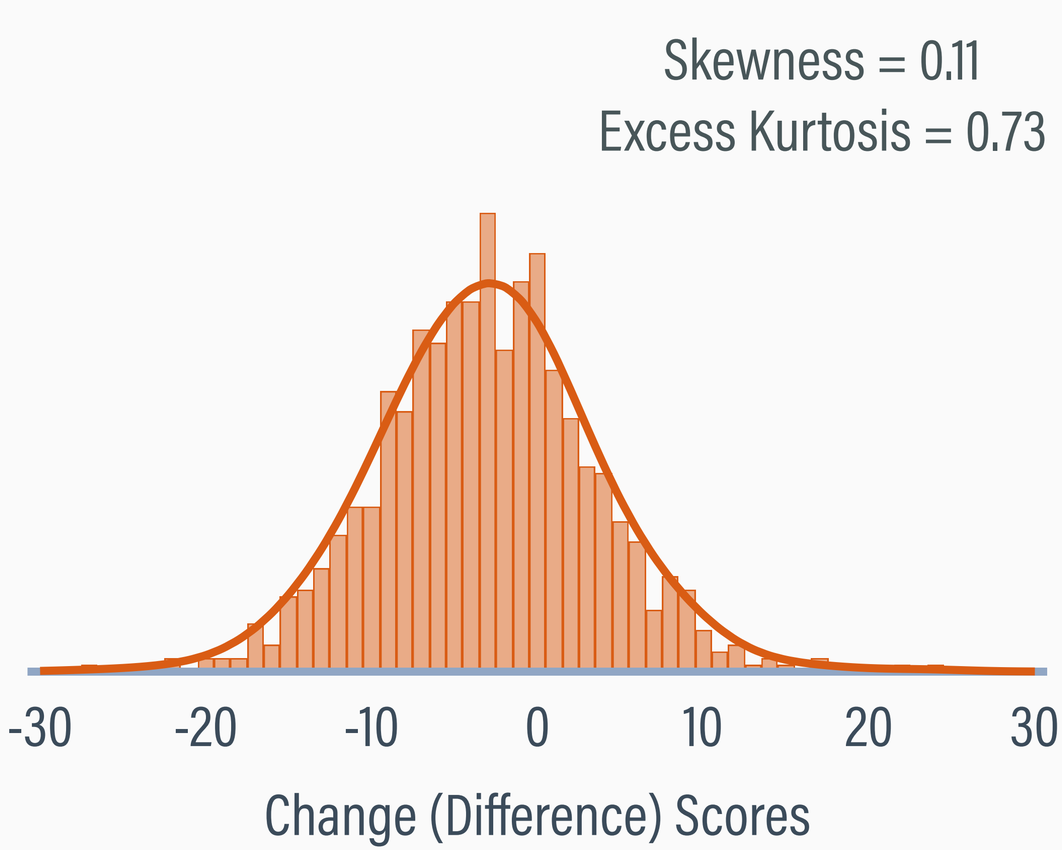
- The accuracy of t-tests (and other statistics) depends on certain conditions in the data being true (e.g., normality)
- Violations of assumptions can bias estimates, inflate or deflate standard errors, and distort significance tests
- Always check reasonableness of assumptions before drawing conclusions

PAIRED T-TEST ASSUMPTIONS

- Paired observations on a numeric (approximately continuous) dependent variable
- Normality of difference scores (not raw scores)
- Each pair of observations is independent of every other pair (e.g., one person's scores do not influence another person's)

DIFFERENCE SCORE DISTRIBUTION

- The raw satisfaction scores were quite skewed, but the change scores were relatively normal
- In small samples, normality violations can artificially inflate or deflate standard errors, thus distorting significance tests
- Normality is less of a concern if the sample size is large enough (e.g., Ns > 40 to 50)



OUTPUT

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Participant	1	882	441.50	254.76	441.50	441.50	326.91	1.00	882.00	881.00	0.00	-1.20	8.58
ParentEduc*	2	882	2.27	0.80	2.00	2.34	1.48	1.00	3.00	2.00	-0.52	-1.26	0.03
ParentIncome*	3	882	2.38	1.11	2.00	2.35	1.48	1.00	4.00	3.00	0.12	-1.33	0.04
BMI10	4	882	19.60	4.22	18.48	19.11	3.57	12.37	35.16	22.79	1.10	0.94	0.14
SkinColorSat10	5	882	3.59	0.64	4.00	3.70	0.00	1.00	4.00	3.00	-1.58	2.41	0.02
SkinColorSat18	6	882	3.50	0.65	4.00	3.58	0.00	1.00	4.00	3.00	-1.36	2.36	0.02
BodySat10	7	882	28.49	5.14	28.00	28.83	5.93	9.00	36.00	27.00	-0.61	0.36	0.17
BodySat18	8	882	25.44	6.06	26.00	25.64	5.93	9.00	36.00	27.00	-0.33	0.10	0.20
BingeEatDisorder10	9	882	2.59	1.82	2.00	2.49	1.48	0.00	8.00	8.00	0.45	-0.35	0.06
BingeEatDisorder18	10	882	1.68	1.45	1.00	1.52	1.48	0.00	8.00	8.00	0.88	0.84	0.05
BodySatCha	11	882	-3.05	6.45	-3.00	-3.09	5.93	-27.00	24.00	51.00	0.11	0.73	0.22

OUTLINE

1. Within-subjects designs
2. Quick review
3. Significance testing steps
4. Statistical assumptions
5. Study questions

STUDY QUESTIONS

Use the following research scenario to answer the study questions:

Researchers studying married couples want to determine whether husbands and wives differ in their relationship satisfaction, which is quantified by a numeric scale based on several questionnaire items. The sample mean difference from a sample of N = 100 couples is $\bar{X}_{diff} = \bar{X}_{wives} - \bar{X}_{husbands} = -5$. To evaluate whether partner differences exist, you will perform significance testing steps assuming a population with a true mean difference of $\mu_{diff} = 0$.

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.

3. Explain why a paired-samples t-test is the appropriate statistical analysis for this scenario.

STUDY QUESTIONS (3)

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

5. The sample mean difference from a sample of N = 100 couples is $\bar{X}_{diff} = \bar{X}_{wives} - \bar{X}_{husbands} = -5$, and the standard error is $s_{\bar{X}diff} = 3$. Explain what the standard error measures. How does it help you gauge whether a mean difference of –5 is similar or different from the null?

STUDY QUESTIONS (4)

6. The t-statistic is t = –1.67. 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?

7. Researchers report the results as “not statistically significant.” What is your decision about the the null hypothesis. Translate your decision into a tangible statement about the difference in marital partner relationship satisfaction.

STUDY QUESTIONS (5)

8. The two-tailed p-value was .09. Provide an interpretation of the probability value.

9. The sample mean difference from a sample of N = 100 couples is $\bar{X}_{diff} = \bar{X}_{wives} - \bar{X}_{husbands} = -5$, and the 95% confidence interval limits are $CI_{95\%} = [-11, +1]$. Provide an interpretation of the confidence interval (I am not asking about its statistical properties). Discuss whether the confidence interval supports or refutes the hypothesis that marital partners have equal relationship satisfaction in the population.