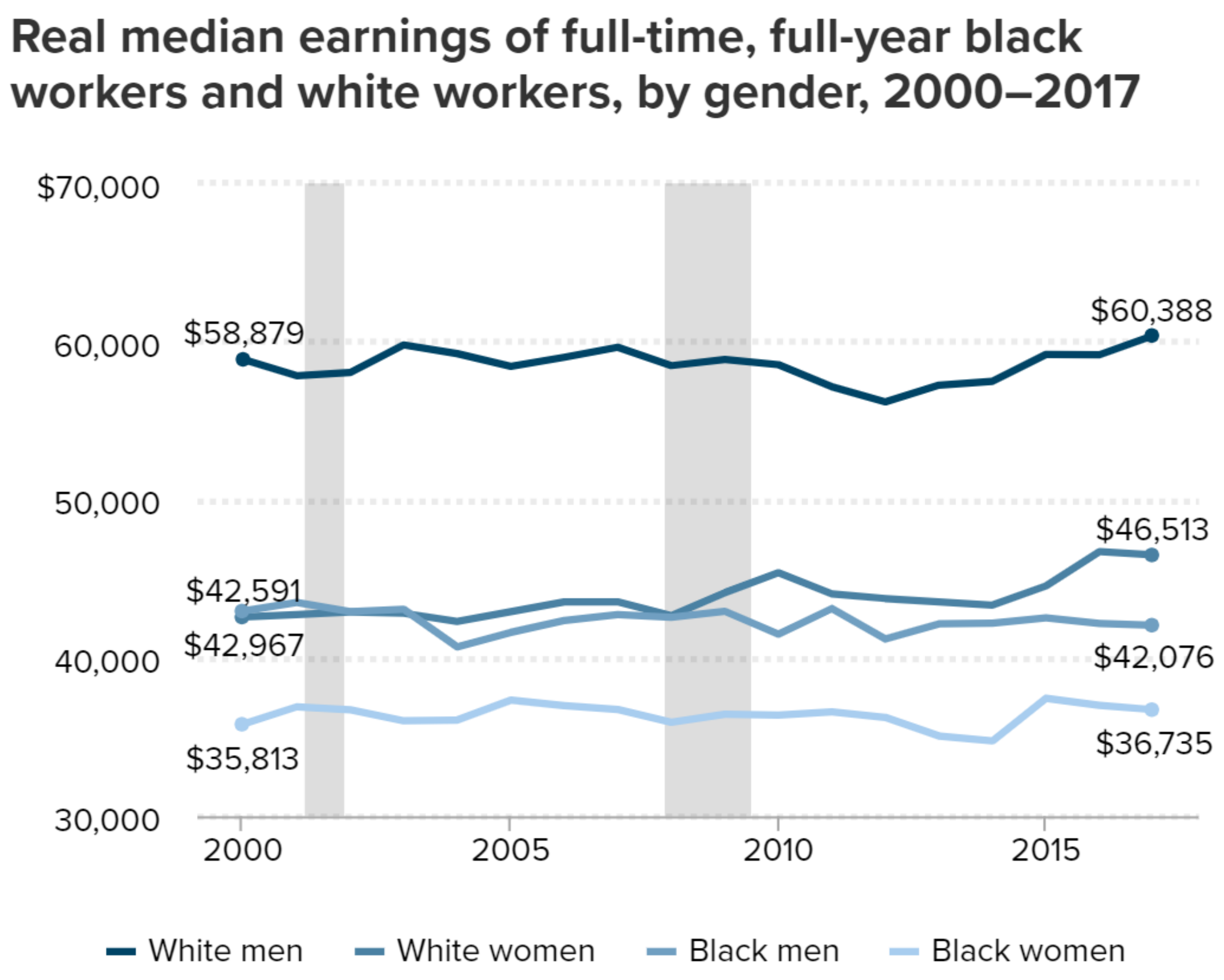
Hypothesis Test for Two Proportions

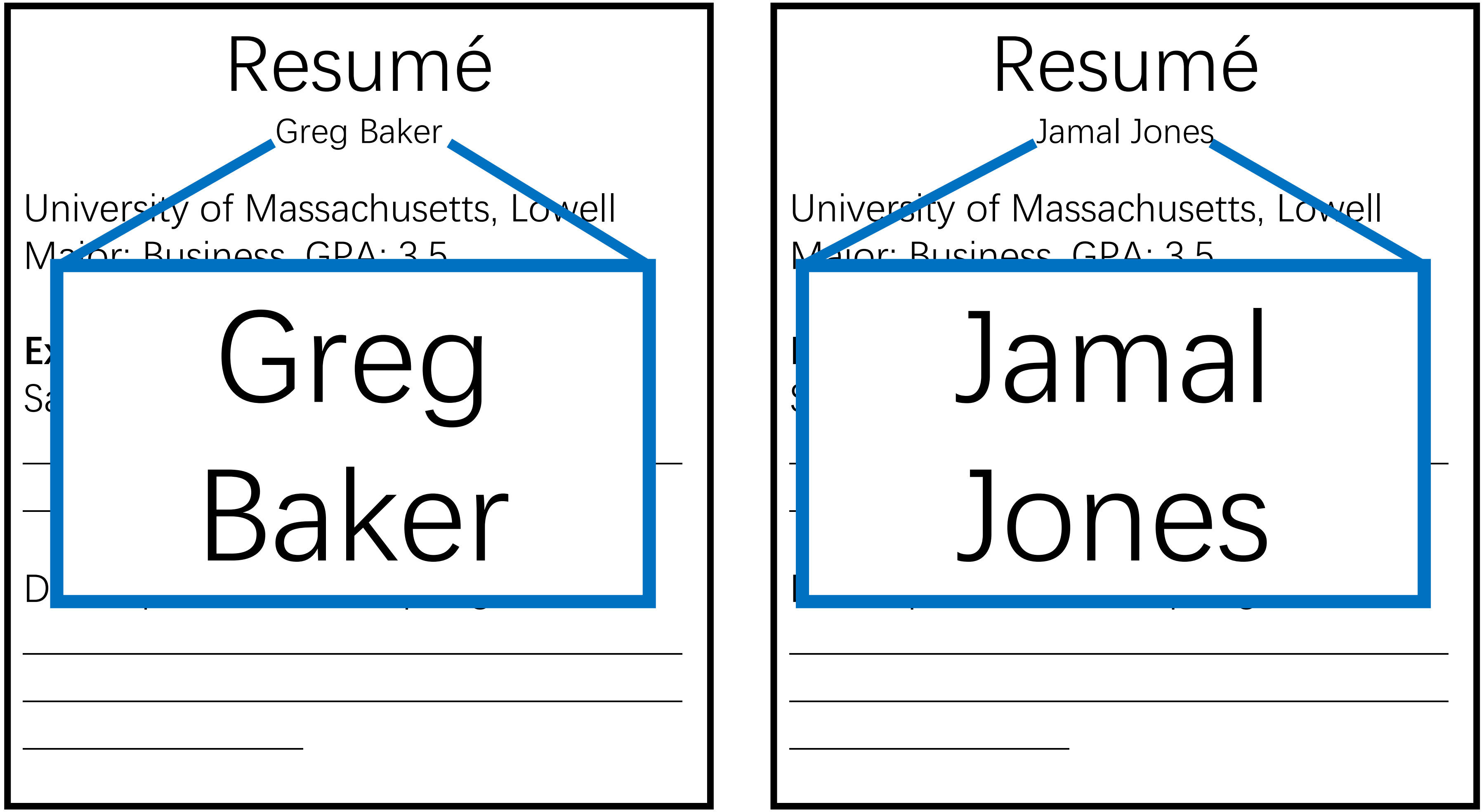


Hiring discrimination

Researchers wanted to test if hiring discrimination was a factor in labor markets

Economic Policy Institute, 2018: https://www.epi.org/blog/black-workers-have-made-no-progress-in-closing-earnings-gaps-with-white-men-since-2000/

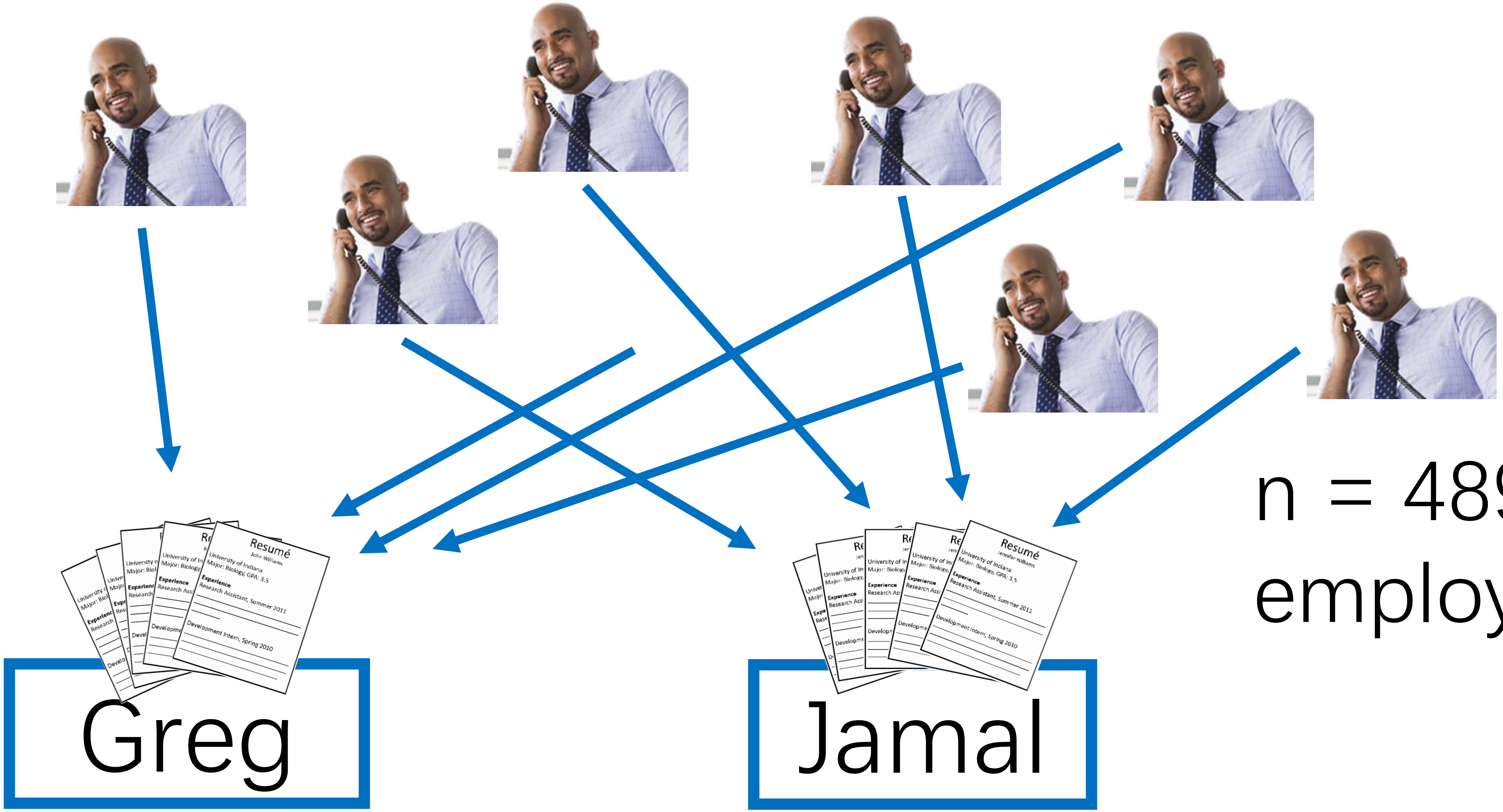
The Race/Resumé Study



The jobs

- Wide swath of jobs in the following **industries**: sales, administrative support, clerical services, and customer services
- Large range of **positions**, from "cashier work at retail establishments and clerical work in a mailroom to office and sales management positions."

The Race/Resumé Study



n = 4890 employers

The Race/Resumé Study

Measured which group got more callbacks from potential employers

Greg $n_1 = 2445$

Jamal $n_2 = 2445$

The results

	Treatment 1 Commonly-White Names	Treatment 2 Commonly-Black Names	Total
Called back	246	164	410
Not called back	2199	2281	4480
Total	2445	2445	4890

Comparing the proportion who received callbacks from both treatments.

$n_1 = 2445$ $n_2 = 2445$

$\hat{p}_1 = \frac{246}{2445} = 0.101$ $\hat{p}_2 = \frac{164}{2445} = 0.067$

Two-Sample Situation

If there's hiring discrimination, $\hat{p}_1 > \hat{p}_2$

Group 1: White

$\hat{p}_1$ = proportion of commonly-white name apps that got callback.

$>$

Group 2: Black

$\hat{p}_2$ = proportion of commonly-black name apps that got callback.

$$\hat{p}_1 = \frac{246}{2445} = \mathbf{0.101} \quad > \quad \hat{p}_2 = \frac{164}{2445} = \mathbf{0.067}$$

Are these proportions **different enough** to show discrimination, or could this difference have been a result of **chance alone**?

Hypotheses

$$H_0: p_1 = p_2$$

There is no discrimination, so the callback rate is the **same in both groups**. You're seeing if there's evidence to reject this default claim.

$$H_A: p_1 > p_2$$

There is discrimination, in which case the commonly-white named applications received a **higher rate** of callbacks.

Where:

p_1 is the proportion of **all** applicants with commonly-**white** names who'd receive callbacks when applying to jobs like the ones in this study.

p_2 is the proportion of **all** applicants with commonly-**black** names who'd receive callbacks when applying to jobs like the ones in this study.

Setting up the Hypotheses

$$H_0: p_1 = p_2$$
$$H_A: p_1 > p_2$$

OR

$$H_0: p_1 - p_2 = 0$$
$$H_A: p_1 - p_2 > 0$$

❤Preferred❤

Where:

p_1 is the proportion of **all** applicants with commonly-**white** names who'd receive callbacks when applying to jobs like the ones in this study.

p_2 is the proportion of **all** applicants with commonly-**black** names who'd receive callbacks when applying to jobs like the ones in this study.

Calculations

Since null assumes $p_1 = p_2$, so we can **combine** the proportion who got callbacks into one estimate: $\hat{p}_c$

Under certain conditions:

$$\hat{p}_1 - \hat{p}_2 \sim N(\mu = 0, \sigma = \sqrt{\frac{\hat{p}_c\,(1-\hat{p}_c)}{n_1} + \frac{\hat{p}_c\,(1-\hat{p}_c)}{n_2}})$$

Centered at zero (since null assumes **no difference** between callback rates)

$$\hat{p}_1 = \frac{246}{2445} = 0.101$$

$$\hat{p}_2 = \frac{164}{2445} = 0.067$$

Combined proportion $\hat{p}_c = \frac{246+164}{2445+2445} = 0.084$

Calculations

Since null assumes $p_1 = p_2$, so we can **combine** the proportion who got callbacks into one estimate: $\hat{p}_c$

<u>Under certain conditions</u>:

$\hat{p}_1 - \hat{p}_2 \sim \mathrm{N}(\mu = 0, \sigma = 0.0079)$

Centered at zero (since null assumes **no difference** between callback rates)

The Data:
The actual difference in callback rates from the experiment $\hat{p}_1 - \hat{p}_2 = \mathbf{0.034}$

How unlikely was our data?

Check the p-value!

$$\hat{p}_1 = \frac{246}{2445} = \mathbf{0.101}$$

$$\hat{p}_2 = \frac{164}{2445} = \mathbf{0.067}$$

Combined proportion $\hat{p}_c = \dfrac{246+164}{2445+2445} = \mathbf{0.084}$

Conclusion

Under my assumption that there is no difference in callback rates, the actually observed data (a 3.4% difference in callback rates among 4890 employers) is highly unlikely (**p-value = 0.00001 < alpha level of 0.05**). So, **I reject my earlier assumption**. There's convincing evidence that commonly-white named resumés receive a **higher callback rate**.

State-Plan-Do-Conclude

State: State the hypotheses, significance level, and define your parameters

$$H_0: p_1 - p_2 = 0$$
$$H_A: p_1 - p_2 > 0$$

$$\alpha = 0.05$$

Where:

p_1 is the proportion of **all** applicants with commonly-white names who'd receive callbacks when applying to jobs like the ones in this study.

p_2 is the proportion of **all** applicants with commonly-black names who'd receive callbacks when applying to jobs like the ones in this study.

State-**Plan**-Do-Conclude

Plan: Name your inference method and check conditions

We will conduct a **two-sample z-test** for $p_1 - p_2$, if all conditions are met.

Conditions

Recall: Why we check conditions

$$\hat{p} \sim \underbrace{\text{Normal}}\left(\underbrace{\mu = 0}, \sigma = \underbrace{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2}}}\right)$$

3) Large counts
→ approx. normal **shape**

1) Random condition
→ unbiased **center**

2) 10% condition
→ calculable **spread**

State-Plan-Do-Conclude

Plan: Name your inference method and check conditions

We will conduct a **two-sample z-test** for $p_1 - p_2$, if all conditions are met.

Conditions

1. **Random:**
Employers were randomly **assigned** either a commonly-white or commonly-black named resumé ✓

Only have to do **10%** when sampling. However, this is an experiment. We don't have to check this condition!

3. **Large Counts:**

$n_1\hat{p}_c \geq 10$ $\quad$ $n_2\hat{p}_c \geq 10$

$n_1(1-\hat{p}_c) \geq 10$ $\quad$ $n_2(1-\hat{p}_c) \geq 10$

State-**Plan**-Do-Conclude

Plan: Name your inference method and check conditions

We will conduct a **two-sample z-test** for $p_1 - p_2$, if all conditions are met.

Conditions

1. **Random:** Employers were randomly **assigned** either a commonly-white or commonly-black named resumé

2. **Large Counts:**

$$n_1\hat{p}_c \geq 10$$
$$(2445)(.084) \geq 10$$

$$n_1(1-\hat{p}_c) \geq 10$$
$$(2445)(1-.084) \geq 10$$

$$n_2\hat{p}_c \geq 10$$
$$(2445)(.084) \geq 10$$

$$n_2(1-\hat{p}_c) \geq 10$$
$$(2445)(1-.084) \geq 10$$

State-**Plan**-Do-Conclude

Plan: Name your inference method and check conditions

We will conduct a **two-sample z-test** for $p_1 - p_2$, if all conditions are met.

Conditions

1. **Random:** Employers were randomly **assigned** either a commonly-white or commonly-black named resumé

2. **Large Counts:**

$$n_1\hat{p}_c \geq 10$$
$$205.4 \geq 10$$

$$n_2\hat{p}_c \geq 10$$
$$205.4 \geq 10$$

$$n_1(1-\hat{p}_c) \geq 10$$
$$2239.6 \geq 10$$

$$n_2(1-\hat{p}_c) \geq 10$$
$$2239.6 \geq 10$$

State-Plan-**Do**-Conclude

Do: Perform calculations (if conditions met), report the test statistic and the p-value

$z = 4.231$

p-value = 0.00001

State-Plan-Do-**Conclude**

Conclude: Reject or fail to reject H_0 and justify

$$H_0: p_1 - p_2 = 0$$
$$H_A: p_1 - p_2 > 0$$

$\alpha = 0.05$

z = 4.231
p-value = 0.00001

Conclusions template: Because our p-value (____) is **less/greater** than our alpha level (___), we **reject/fail to reject** H_0. We **do/don't** have convincing evidence that (H_A in context).

State-Plan-Do-**Conclude**

Conclude: Reject or fail to reject H_0 and justify

$$H_0: p_1 - p_2 = 0$$
$$H_A: p_1 - p_2 > 0$$

$\alpha = 0.05$

z = 4.231
p-value = 0.00001

Because our p-value (0.00001) is **less** than our alpha level (0.05), we **reject** H_0. We **do** have convincing evidence that commonly-white name resumés get a higher callback rate for jobs similar to the ones in this study.

Hypothesis Test for Two Means

Standard deviations known

When two random samples are independently selected and when n_1 and n_2 are both large or the population distributions are (at least approximately) normal, the distribution of

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is described (at least approximately) by the standard normal (z) distribution.

Standard deviations unknown

When two random samples are independently selected and when n_1 and n_2 are both large or when the population distributions are normal, the standardized variable

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

has approximately a t distribution with

$$\text{df} = \frac{(V_1 + V_2)^2}{\dfrac{V_1^2}{n_1 - 1} + \dfrac{V_2^2}{n_2 - 1}} \quad \text{where } V_1 = \frac{s_1^2}{n_1} \text{ and } V_2 = \frac{s_2^2}{n_2}$$

The computed value of df should be truncated (rounded down) to obtain an integer value of df.


Today’s Key Analysis

Did the stent treatment work?

Measurement of Outcome

No Symptoms

Death

0 1 2 3 4 5 6

If stent works, it will move patients **down** this scale

Measurement of Outcome

No Symptoms

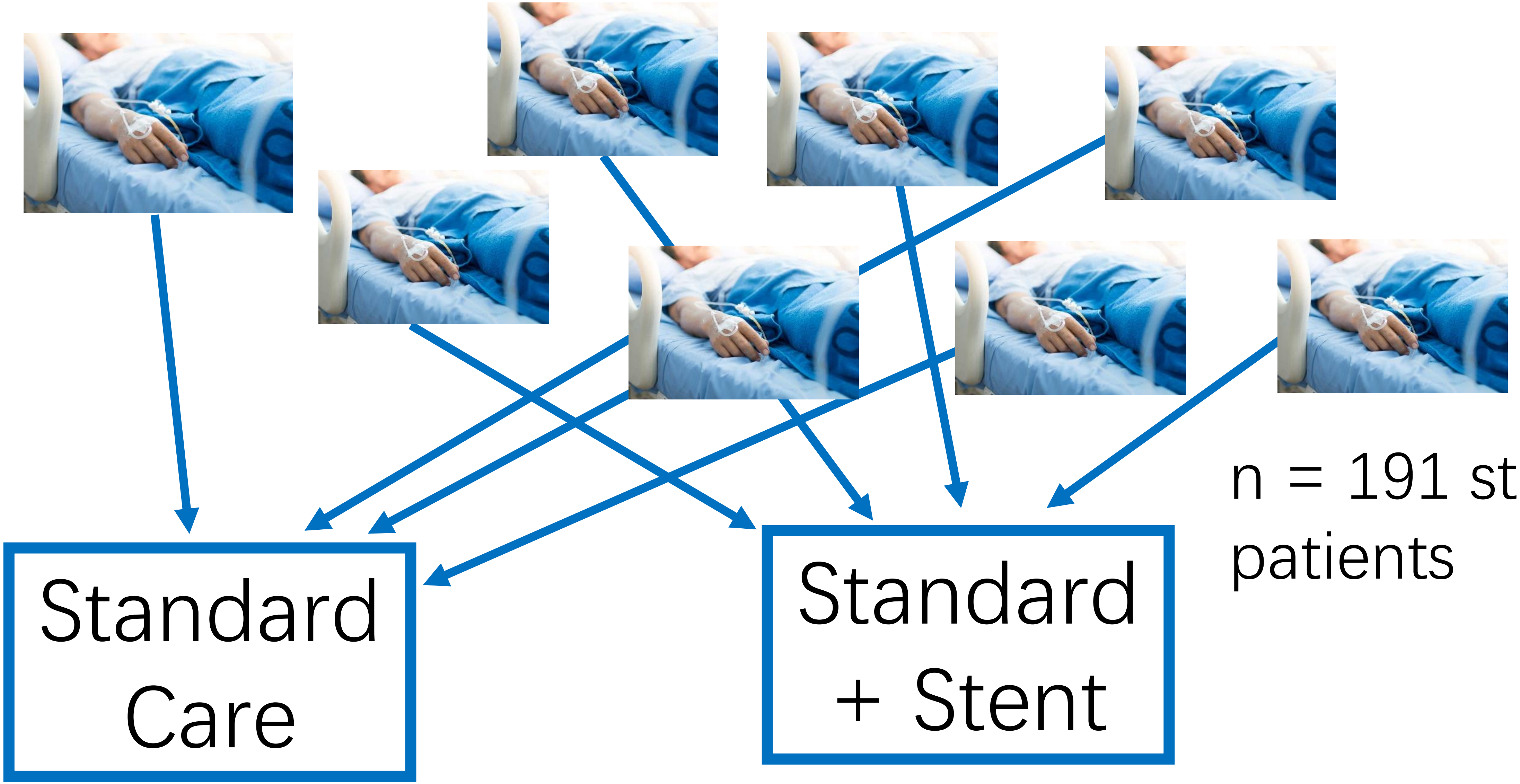
Death

0 1 2 3 4 5 6

From here on: referred to as “**disability score**”

Our Study

Random Assignment



n = 191 stroke patients

Our Study

Standard Care $n_1 = 93$

Standard + Stent $n_2 = 98$

Our Study

Measured which group had **lower mean** disability score

Standard Care

$n_1 = 93$

Standard + Stent

$n_2 = 98$

Topics

1. **Two-sample t-test for a difference of means**
2. Four step process

Setting up the Hypotheses

$$H_0: \mu_s = \mu_c$$
$$H_A: \mu_s < \mu_c$$

$$H_0: \mu_s - \mu_c = 0$$
$$H_A: \mu_s - \mu_c < 0$$

<u>Where:</u>

μ_s is the mean disability score of **all** patients who'd receive stents.

μ_c is the mean disability score of **all** patients who'd receive current standard of care.

The results

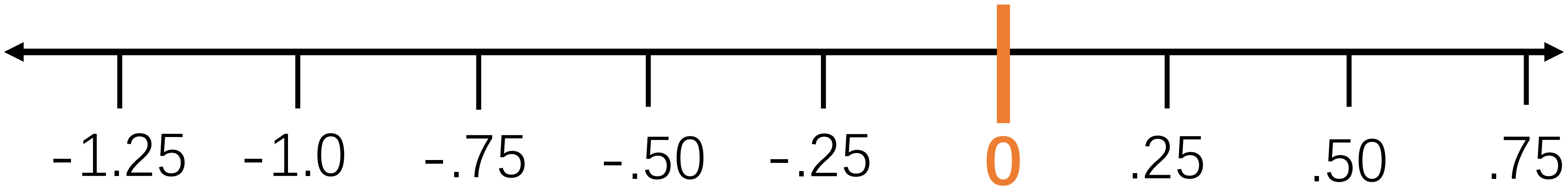
	Stent	Control
Mean Disability	2.26	3.23
Stdev. Disability	1.78	1.78
n	98	93

Saver JL, Goyal M, Bonafe A, Diener HC, Levy EI, Pereira VM, et al. “Stent-retriever thrombectomy after intravenous t-pa vs. T-pa alone in stroke.” *N Engl J Med*. 2015; 372:2285-2295 https://www.nejm.org/doi/full/10.1056/nejmoa1415061

The Calculations

1. Assume the null is true

H_0: $\mu_s - \mu_c = 0$



We assume there is no difference in average disability scores among all patients who would receive stents or the current standard of care: $\mu_s - \mu_c = 0$

The Calculations

2. **If the null is in fact true**, how likely is the data that you've gathered?

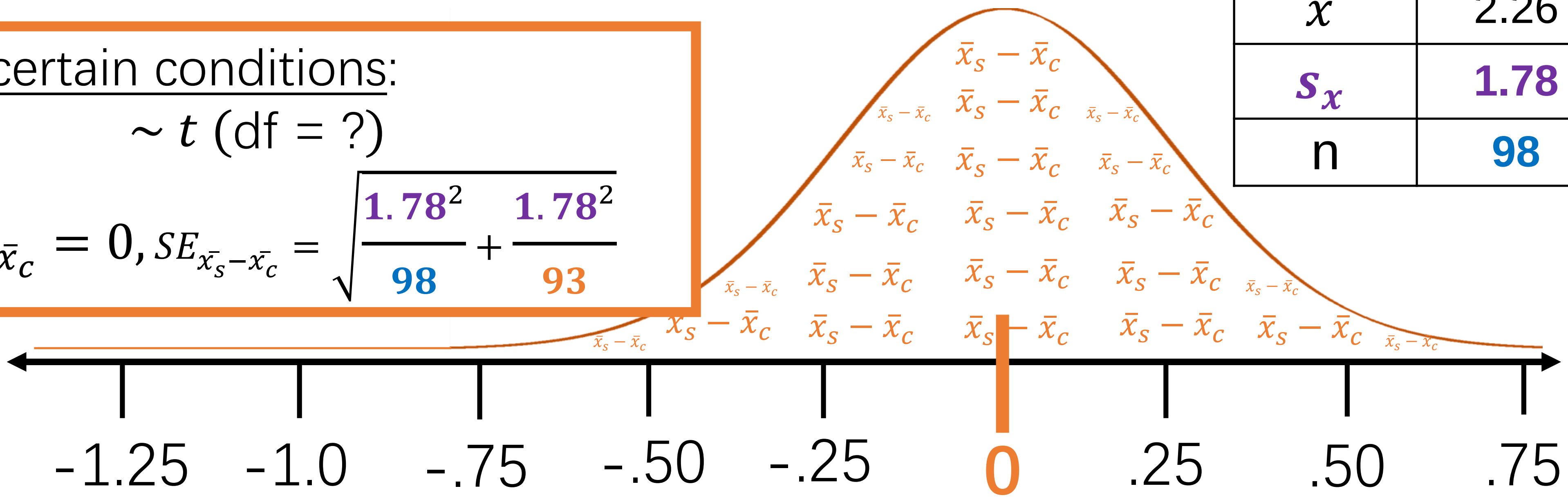
The study enrolled 191 patients (93 in control, 98 in treatment).
Let's find the sampling distribution of $\bar{x}_s$ - $\bar{x}_c$.

	Stent	**Control**
$\bar{x}$	2.26	3.23
s_x	**1.78**	**1.78**
n	**98**	**93**

Under certain conditions:

$\sim t$ (df = ?)

$$\mu_{\bar{x}_s-\bar{x}_c} = 0, SE_{\bar{x}_s-\bar{x}_c} = \sqrt{\frac{1.78^2}{98} + \frac{1.78^2}{93}}$$



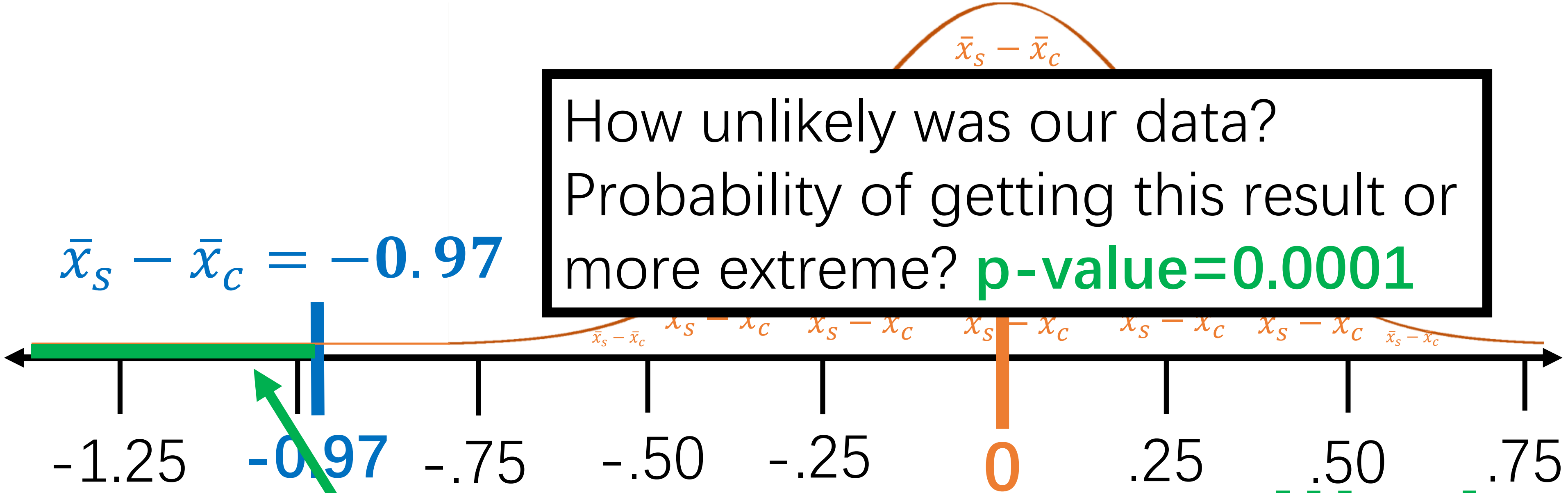
H_0: $\mu_s - \mu_c = 0$

The Data: The actual difference in mean disability score from the experiment $\bar{x}_s - \bar{x}_c = -0.97$

	Stent	Control
$\bar{x}$	2.26	3.23
s_x	1.78	1.78
n	98	93

How unlikely was our data? Probability of getting this result or more extreme? p-value=0.0001

$\bar{x}_s - \bar{x}_c = -0.97$



This probability

Very unlikely!

Recap: The Test

Is there **convincing statistical evidence** that the stent lowered the average disability from stroke?

$$H_0: \mu_s - \mu_c = 0$$
$$H_A: \mu_s - \mu_c < 0$$

$\mu_s = 2.26$, $\mu_c = 3.23$

$$\bar{x}_s - \bar{x}_c = -0.97$$

Recap: The Test

Is there **convincing statistical evidence** that the stent lowered the average disability from stroke?

$\boldsymbol{H_0: \mu_s - \mu_c = 0}$

$H_A: \mu_s - \mu_c < 0$

$\mu_s = 2.26, \mu_c = 3.23$

$\bar{x}_s - \bar{x}_c = -0.97$

1. Assume null is true

Recap: The Test

Is there **convincing statistical evidence** that the stent lowered the average disability from stroke?

$$\boldsymbol{H_0: \mu_s - \mu_c = 0}$$

$$H_A: \mu_s - \mu_c < 0$$

$\mu_s = 2.26, \mu_c = 3.23$

$\boldsymbol{\bar{x}_s - \bar{x}_c = -0.97}$

1. Assume null is true

2. How unlikely was our sampled data?

p-value: 0.0001

Recap: The Test

Is there **convincing statistical evidence** that the stent lowered the average disability from stroke?

$$\boldsymbol{H_0: \mu_s - \mu_c = 0}$$

$$H_A: \mu_s - \mu_c < 0$$

$\mu_s = 2.26,\ \mu_c = 3.23$

$\boldsymbol{\bar{x}_s - \bar{x}_c = -0.97}$

Our sample data was **very unlikely**, if we assume the null is true.

2. How unlikely was our sampled data?

p-value: 0.0001

Conclude

Note: more concise conclusion template provided later

3. Draw a conclusion:

$$\cancel{H_0: \mu_s - \mu_c = 0}$$

$$H_A: \mu_s - \mu_c < 0$$

Under my assumption that stents provide no added benefit, the actually observed data (0.97 point decline in disability in stent group) is highly unlikely (**p-value = 0.0001**). So, I **reject** my earlier assumption. There's convincing evidence that the **stent lowers disability from stroke.**

Topics

1. Two-sample t-test for a difference of means
2. **Four step process**

The Four Steps for Inference

A suggested way to **organize** your work so that you get full credit on FRQ's!

State: hypotheses, significance level, and define your parameters

Plan: Name your inference method and check conditions

Do: Perform calculations (if conditions met), report the test statistic and the p-value

Conclude: Reject or fail to reject H_0 and justify

State-Plan-Do-Conclude

State: State the hypotheses, significance level, and define your parameters

$$H_0: \mu_s - \mu_c = 0$$
$$H_A: \mu_s - \mu_c < 0$$

$\alpha = 0.05$

Where:

μ_s is the mean disability score of all patients who'd receive stents.

μ_c is the mean disability score of all patients who'd receive current standard of care.

State-**Plan**-Do-Conclude

<u>Plan:</u> Name your inference method and check conditions

We will conduct a **two-sample t-test** for $\mu_s - \mu_c$.

$\frac{(\overline{X_s} - \overline{X_c}) - 0}{\sqrt{\frac{{s_s}^2}{n_s} + \frac{{s_c}^2}{n_c}}} \sim t(df)$ if the following conditions are met. if all conditions are met.

df=?

State-Plan-Do-Conclude

Plan: Name your inference method and check conditions

We will conduct a **two-sample t-test** for $\mu_s - \mu_c$, if all conditions are met.

Conditions

1. **Random:** Patients were randomly **assigned** either to receive current standard care or stent treatment.

2. **Normal/Large Samples:**

$n_s \geq 30$	$n_c \geq 30$
$98 \geq 30$	$93 \geq 30$

State-Plan-**Do**-Conclude

Do: Perform calculations (if conditions met), report the test statistic and the p-value

t = (formula) = -3.76

p-value = (formula) = 0.0001

State-Plan-Do-**Conclude**

Conclude: Reject or fail to reject H_0 and justify

Conclusions template: Because our p-value (____) is **less/greater** than our alpha level (___), we **reject/fail to reject** H_0. We **do/don't** have convincing evidence that (H_A in context).

State-Plan-Do-**Conclude**

<u>Conclude:</u> Reject or fail to reject H_0 and justify

Because our p-value (<u>0.0001</u>) is **less** than our alpha level (<u>0.05</u>), we **reject** H_0. We **do** have convincing evidence that stents lower disability from stroke.

Procedure: A two-sample *t*-interval for $\mu_{NRW} - \mu_{RW}$, the difference in population means of BPA body concentrations in nonretail workers and retail workers

Checks: It is given that these are random samples. It is reasonable to assume the samples are independent, both samples sizes ($528 \geq 30$ and $197 \geq 30$) are large enough so that the CLT applies, and we assume the sample sizes are less than 10% of the populations.

Mechanics: Calculator software gives (−0.9521, −0.7479) with $df = 332.3$.

Conclusion in context: We are 99% confident that the difference in true means of BPA body concentrations in all nonretail and retail workers (nonretail mean minus retail mean) is between −0.75 and −0.95 μg/L.

Because 0 is not in the interval of plausible values for the difference of population means and the entire interval is negative, the interval does support the belief that retail workers carry higher amounts of BPA in their bodies than nonretail workers.

Parameters: Let μ_{NFL} represent the mean attendance of the population of NFL games. Let μ_{10} represent the mean attendance of the population of Big Ten football games.

Hypotheses: H_0: $\mu_{NFL} - \mu_{10} = 0$ (or $\mu_{NFL} = \mu_{10}$) and H_a: $\mu_{NFL} - \mu_{10} < 0$ (or $\mu_{NFL} < \mu_{10}$)

Procedure: A two-sample *t*-test for means.

Checks: Independent random samples (given); both samples sizes, $n_{NFL} = 35 \geq 30$ and $n_{10} = 30 \geq 30$, are large enough for the CLT to apply; and the sample sizes, 35 and 30, are less than 10% of all NFL and Big Ten football games, respectively.

Mechanics: The population SDs are unknown, so we use a t-distribution. Calculator software (such as 2-SampTTest on the TI-84 or 2-Sample tTest on the Casio Prizm) gives $t = -0.8301$, $df = 49.3$, and $P = 0.2052$.

Conclusion in context with linkage to the P-value: With a P-value this large, $0.2052 > 0.05$, there is not sufficient evidence to reject H_0; that is, there is not sufficient evidence that the true mean attendance at Big Ten Conference football games is greater than that at NFL games.