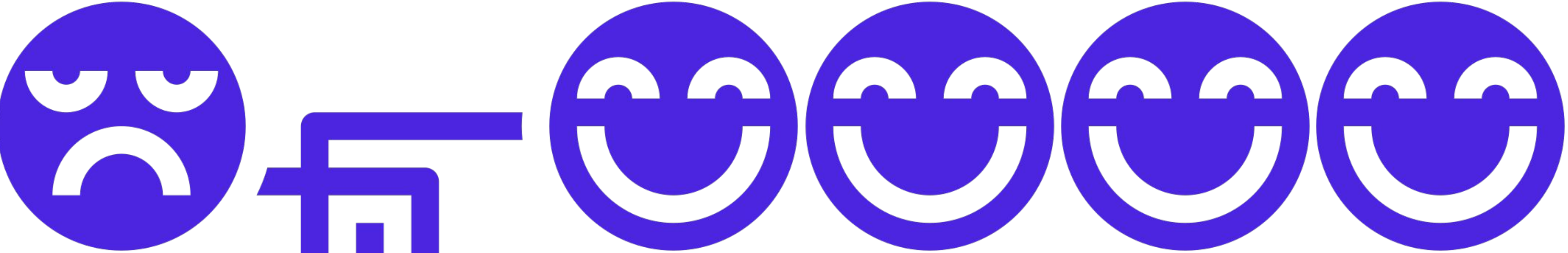
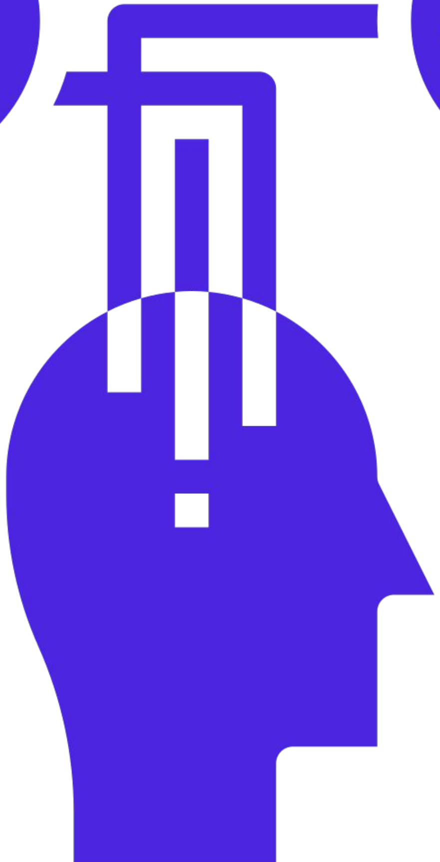
SOCY 3B LAB

//week 7//





+ how are you feeling?

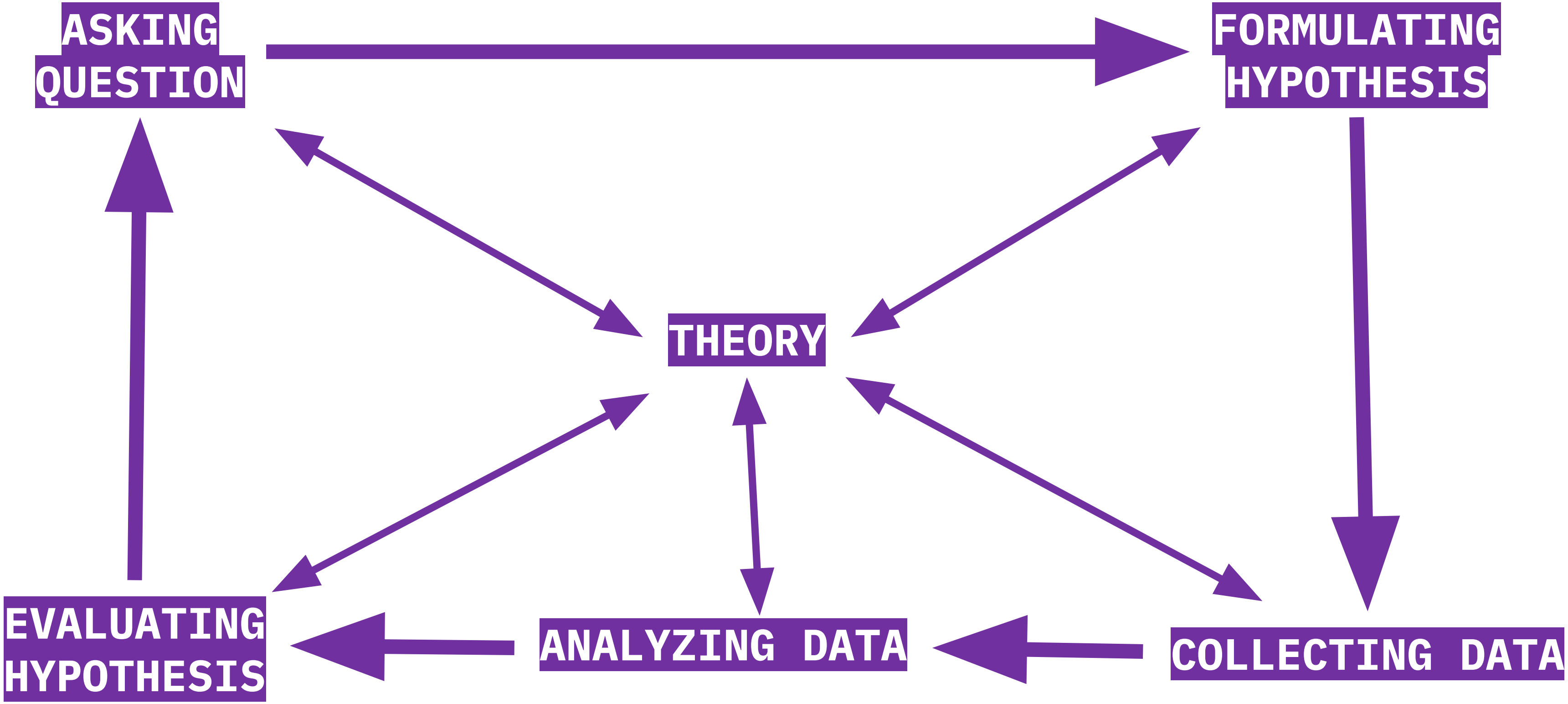
[check-in]

I. **INTRODUCING:** hypothesis testing

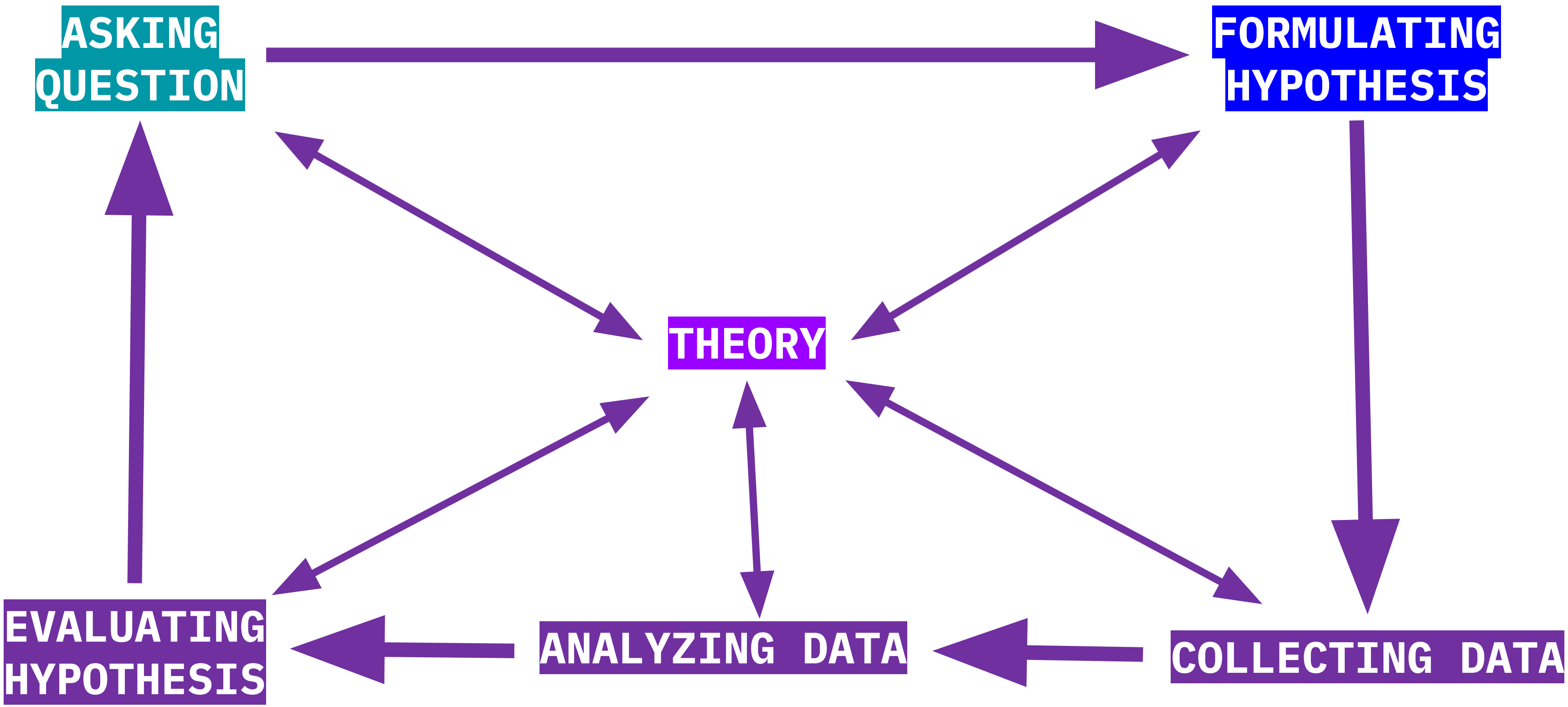
II. **STATA**: t-tests to compare group means

III. **GROUP ASSIGNMENT**

THE RESEARCH PROCESS:



THE RESEARCH PROCESS:



RESEARCH QUESTION:

Are Covid-19 rates higher in US communities depending on income?

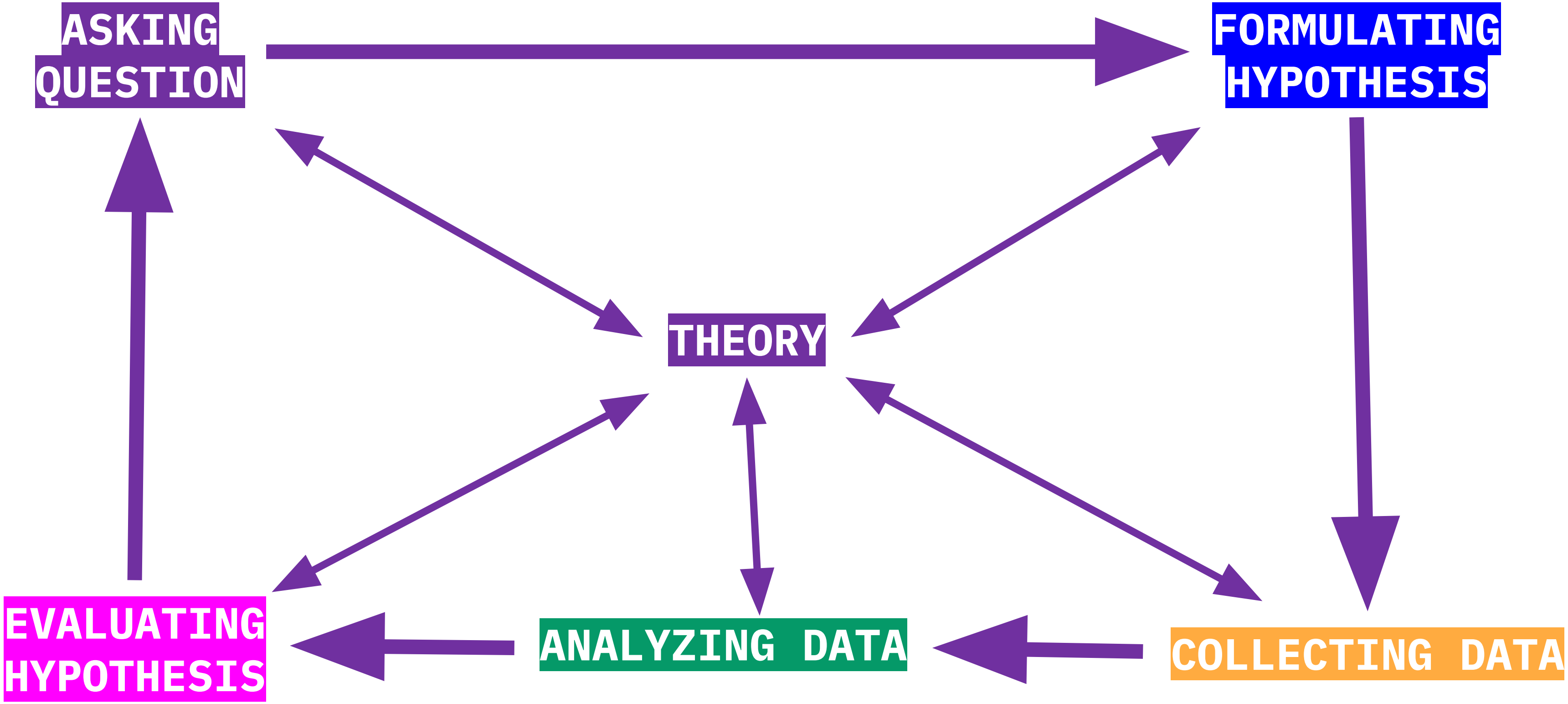
HYPOTHESIS:

Covid-19 rates in low-income communities are higher than in high-income communities.

THEORY:

“essential work”; healthcare costs; racist & classist healthcare practices and social services; pre-existing health disparities

THE RESEARCH PROCESS:



RESEARCH QUESTION:

Why is volunteering in the community less likely among low-income families?

too leading

Are families more likely to volunteer in local community events depending on their income?

great question! 🎉🎉🎉

HYPOTHESIS: low-income families are less likely to volunteer than high-income families

$$\mu \text{ low-income families} < \mu \text{ high-income families}$$

[review]

RESEARCH HYPOTHESES

1. μ group 1 > μ group 2

   Avg rent in Santa Cruz, CA **>** Avg rent in Chapel Hill, NC

2. μ group 1 < μ group 2

   Avg income of US residents who speak no English **<**
   Avg income of U.S. residents who speak English

3. μ group 1 ≠ μ group 2

   Price per gallon of gas in CA **≠** price per gallon of gas in NY

T-TEST:

What we use to test for differences between means when:

+ the **population standard deviation** is **unknown**
+ the **population standard deviation** is **estimated** using the **sample standard deviation**

$$t = \overbrace{(\bar{X}_1-\bar{X}_2)}^{\text{SIGNAL}} \,/\, \overbrace{S_{\bar{Y}_1-\bar{Y}_2}}^{\text{NOISE}}$$

$\bar{X}_1 - \bar{X}_2$ = *difference in means*

$S_{\bar{Y}_1-\bar{Y}_2}$ = *combined standard error*

ALTERNATIVE HYPOTHESIS:

The **difference between two means** is **significantly different from zero.**

H_A: μ group 1 > μ group 2

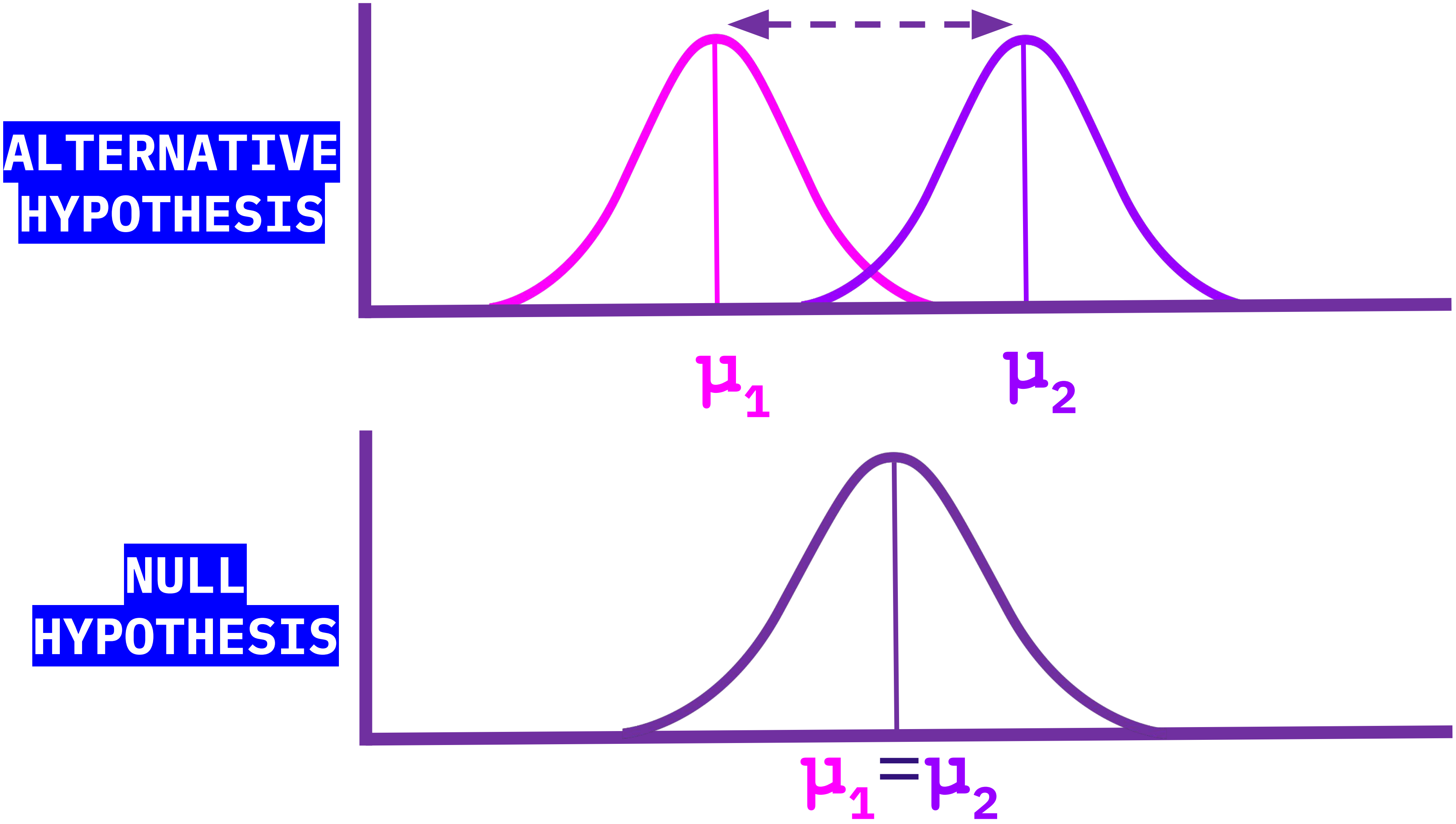
H_A: μ group 1 < μ group 2

H_A: μ group 1 ≠ μ group 2

NULL HYPOTHESIS:

The **difference between two means** is **NOT** **significantly different from zero**. The **noise** (combined standard errors) overwhelms the **signal** (difference between means).

H_0: μ group 1 = μ group 2


ALTERNATIVE HYPOTHESIS
μ_1
μ_2
NULL HYPOTHESIS
$\mu_1 = \mu_2$

ARE FAMILIES MORE LIKELY TO VOLUNTEER IN LOCAL COMMUNITY EVENTS DEPENDING ON THEIR INCOME?

H_A: [ALTERNATIVE HYPOTHESIS]

low-income families are less likely to volunteer than other families.

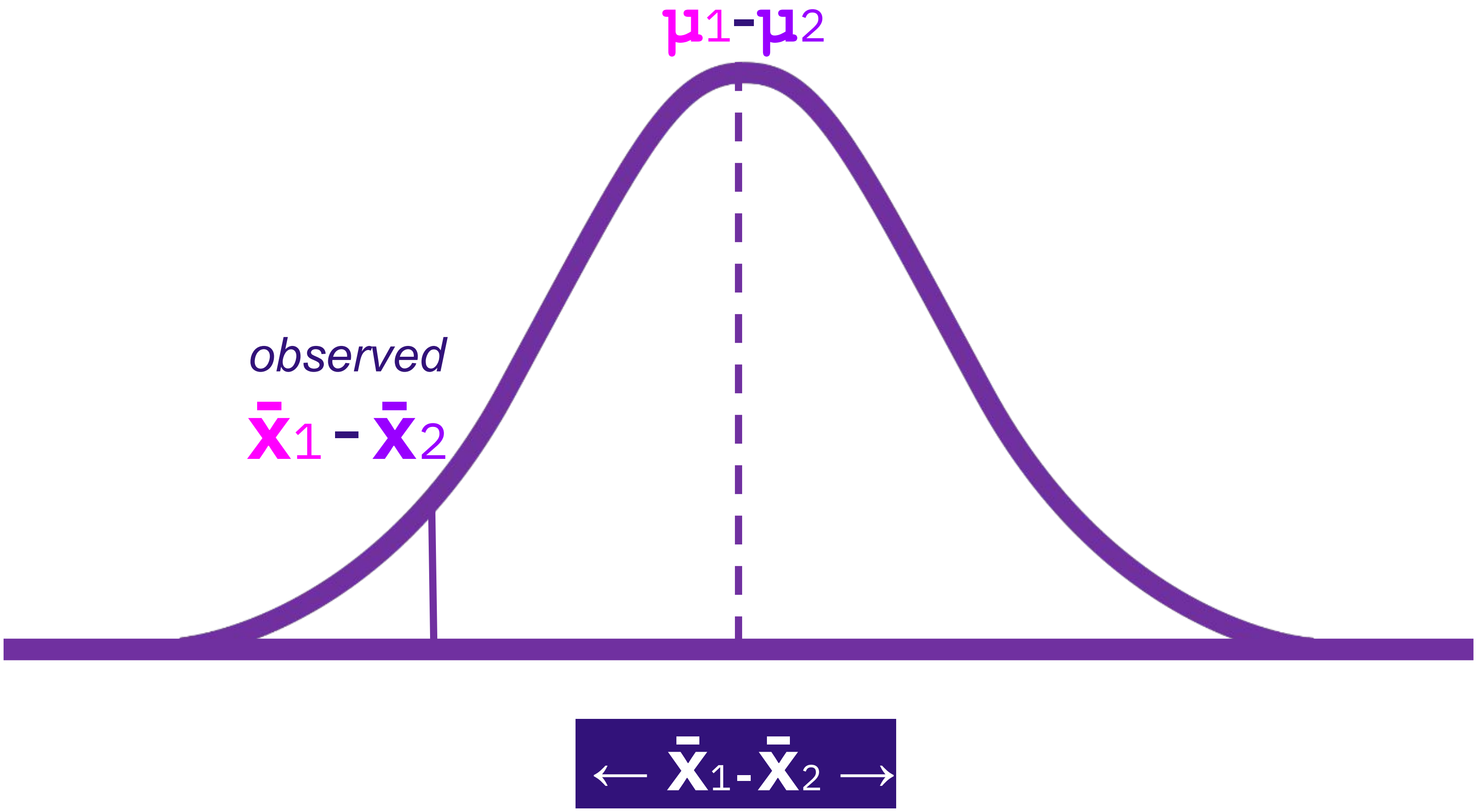
µ low-income families < µ high-income families

H_0: [NULL HYPOTHESIS]

families with different income levels are no more or less likely to volunteer

µ low-income families = µ high-income families

SAMPLING DISTRIBUTION OF DIFFERENCE BETWEEN MEANS:



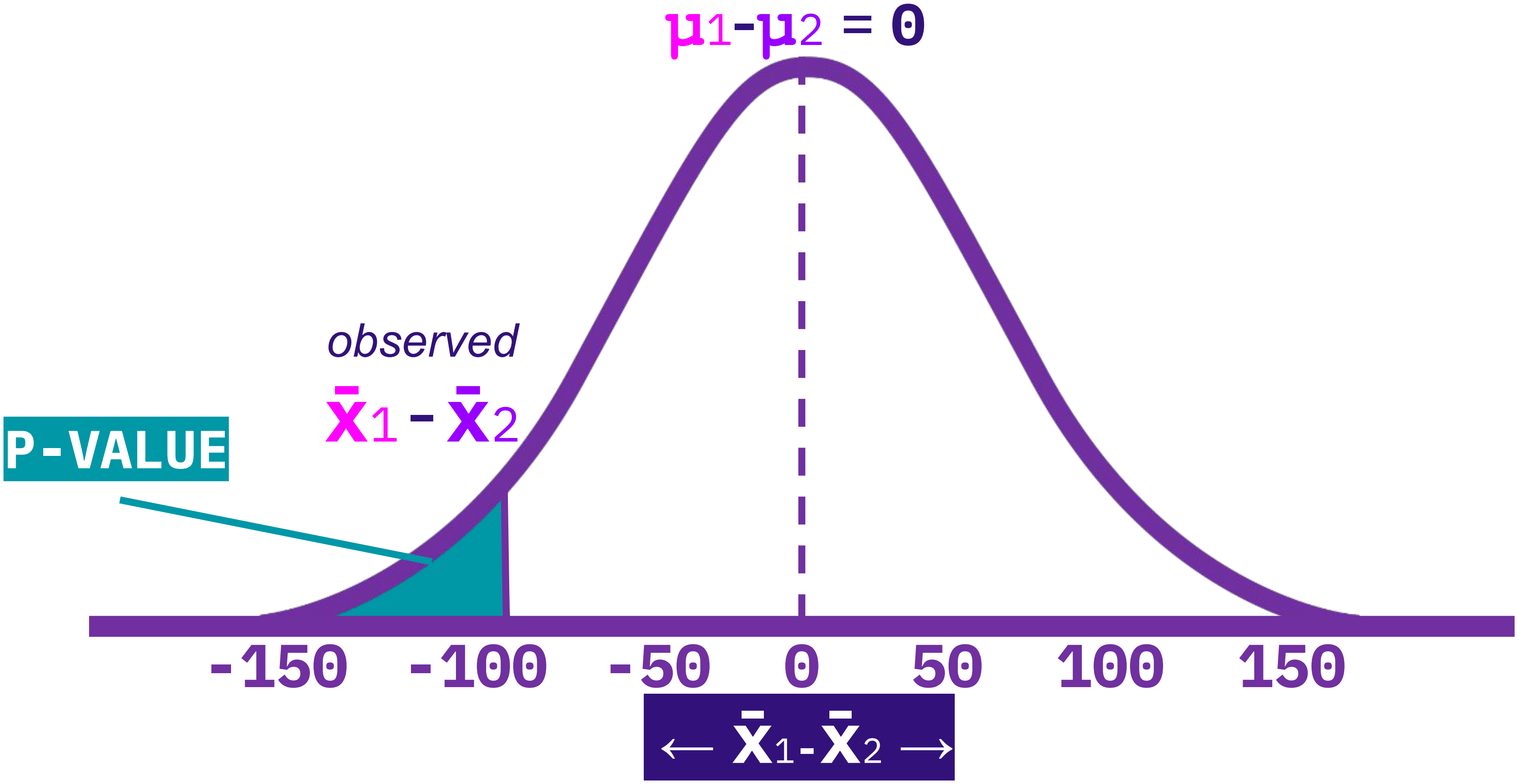
P-VALUE: ***p*** The probability of obtaining the observed difference if the null hypothesis is true (i.e., if there is no actual difference between the two groups)

*The **smaller the p-value**, the **less likely** the observed difference can be **explained by chance**. This **does provide support** for our **alternative hypothesis**.*

*The **larger the p-value**, the **more likely** the observed difference can be **explained by chance**. This **does NOT provide support** for our **alternative hypothesis**.*

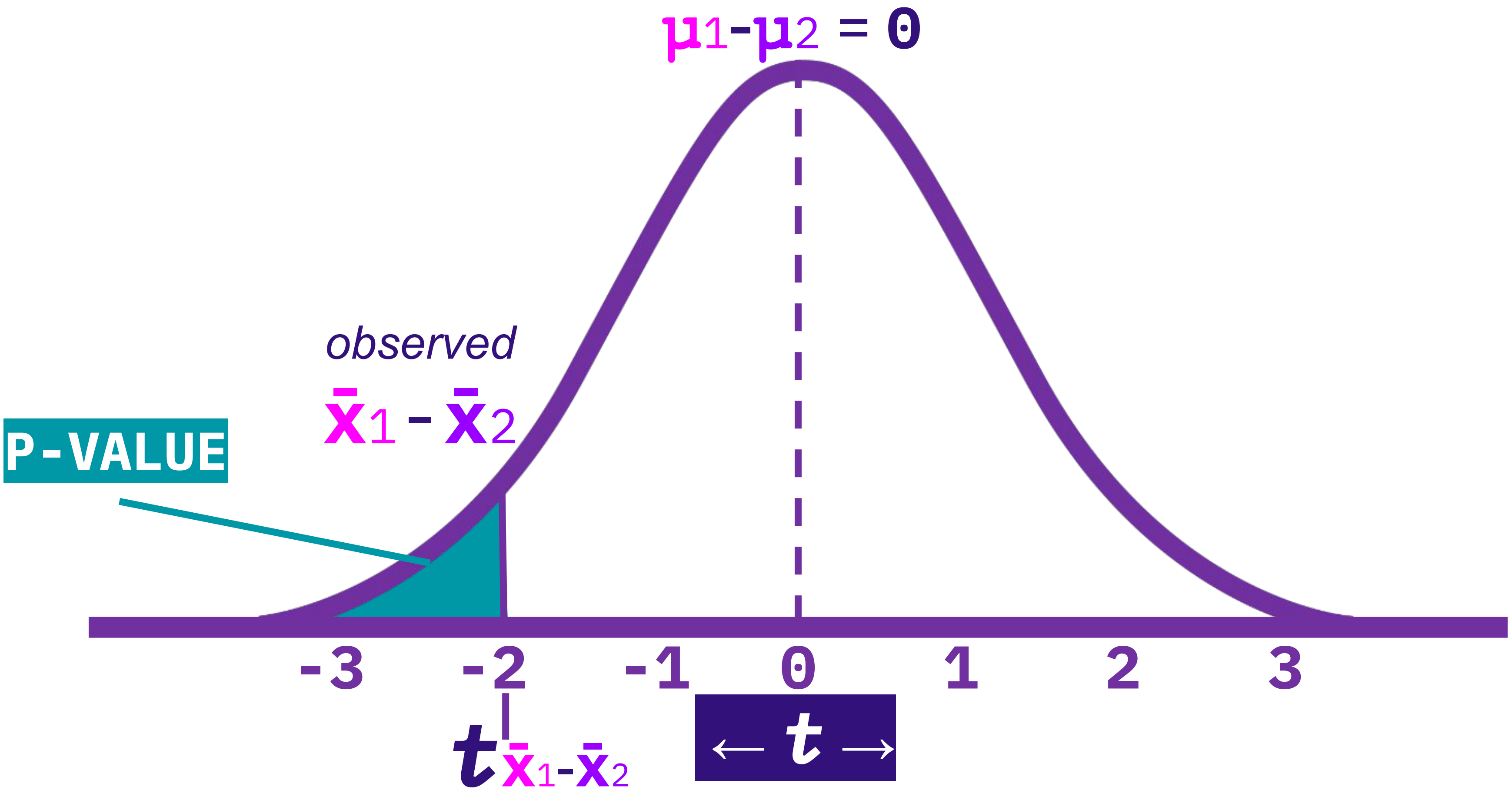
SAMPLING DISTRIBUTION OF DIFFERENCE BETWEEN MEANS:

IF THE NULL HYPOTHESIS IS TRUE



SAMPLING DISTRIBUTION OF DIFFERENCE BETWEEN MEANS:

IF THE NULL HYPOTHESIS IS TRUE



HYPOTHESES & THEIR TESTS:

H_1: μ group **1** > μ group **2**

H_1: μ group **1** < μ group **2**

ONE-TAILED TEST

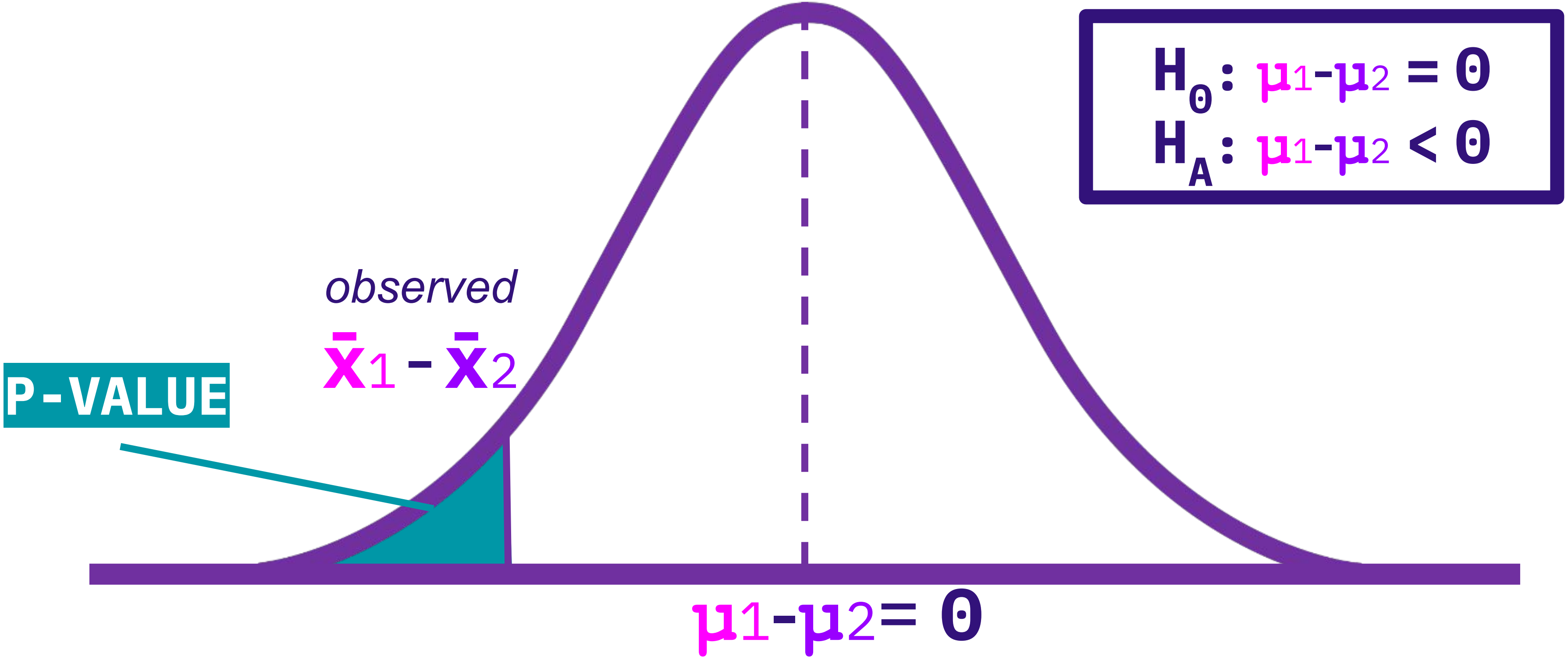
RIGHT-SIDED

LEFT-SIDED

H_1: μ group **1** ≠ μ group **2**

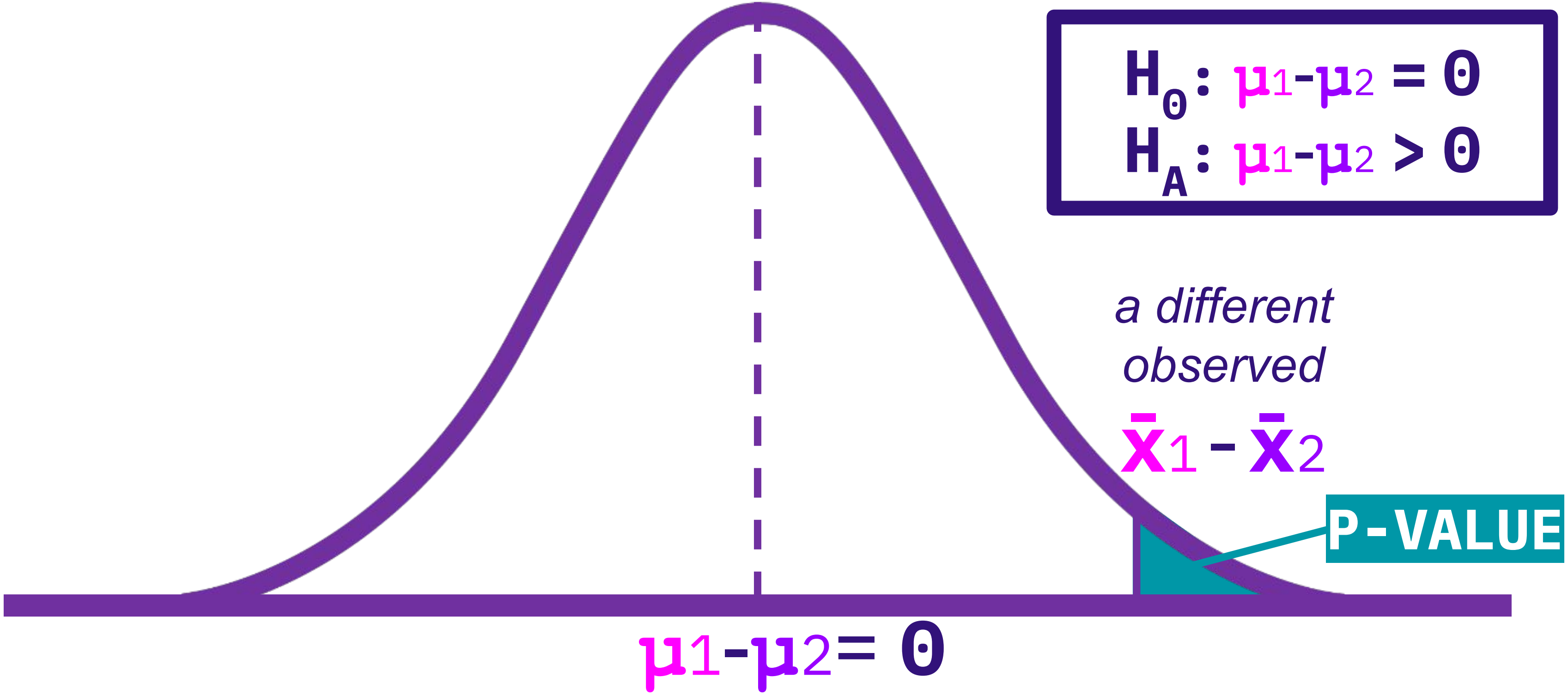
TWO-TAILED TEST

ONE-TAILED TEST (LEFT-SIDED):



If there is no diff in population means ($\mu_1 - \mu_2 = 0$), what is the probability of seeing <u>at least</u> the observed diff in sample means ($\bar{x}_1 - \bar{x}_2$ or more <u>negative</u>)?

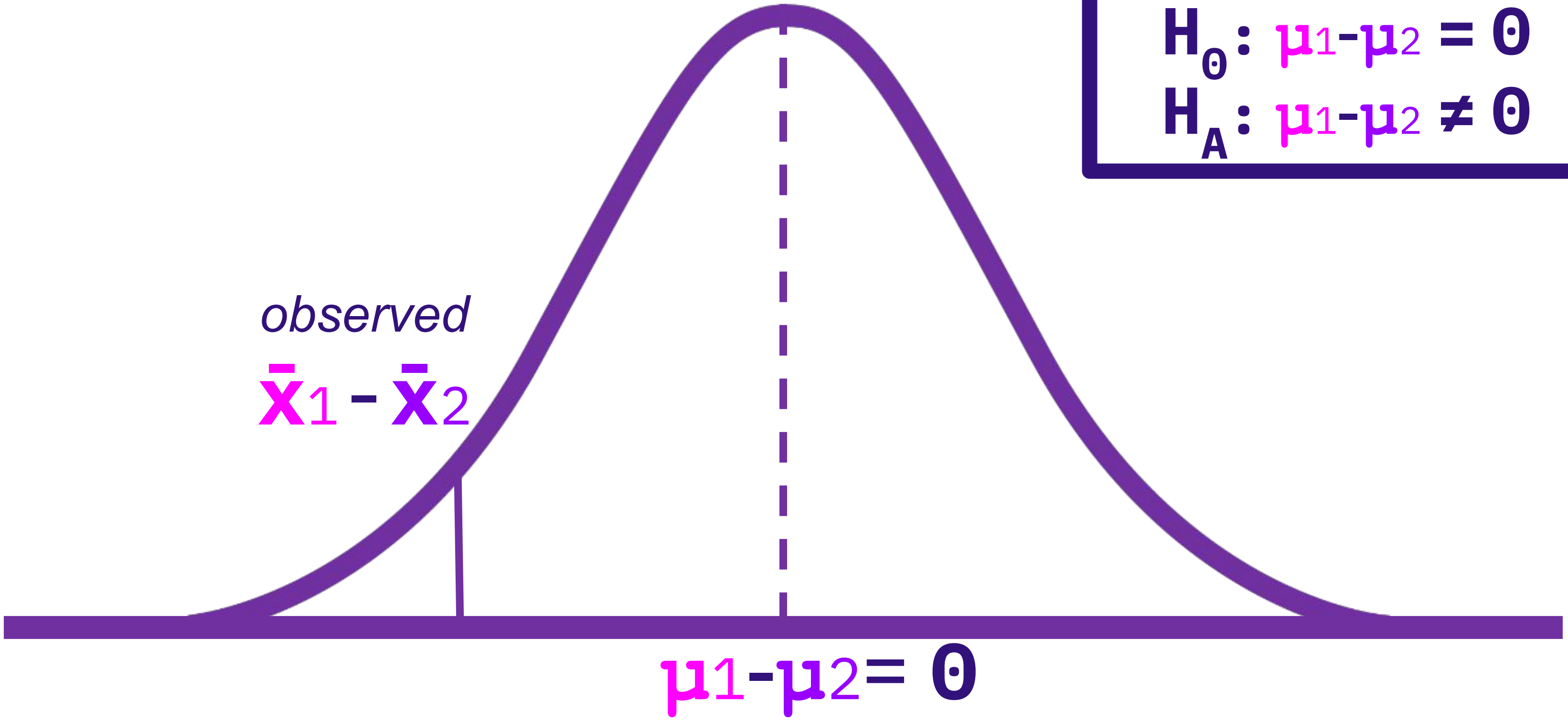
ONE-TAILED TEST (RIGHT-SIDED):



If there is no diff in population means ($\mu_1 - \mu_2 = 0$), what is the probability of seeing <u>at least</u> the observed diff in sample means ($\bar{x}_1 - \bar{x}_2$ or more <u>positive</u>)?

TWO-TAILED TEST:

$H_0: \mu_1 - \mu_2 = 0$
$H_A: \mu_1 - \mu_2 \neq 0$

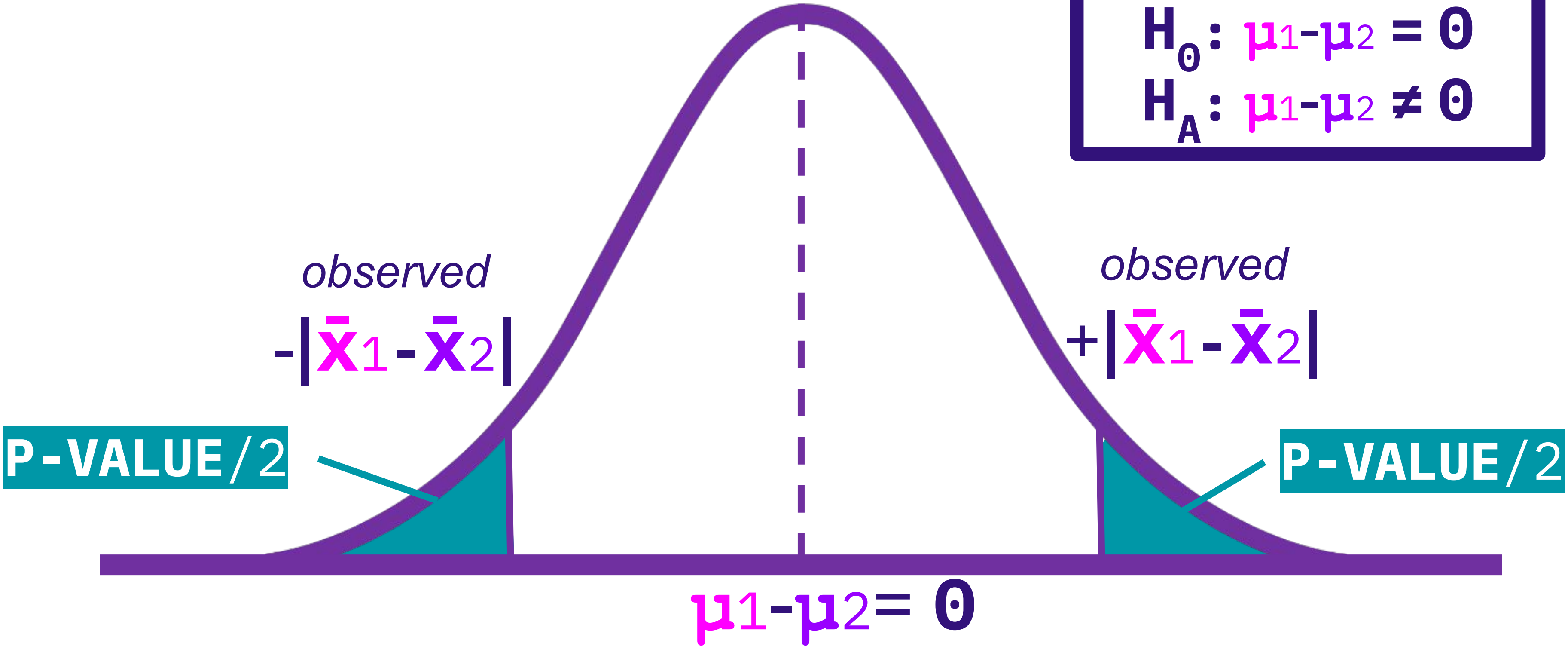


If there is no diff in population means ($\mu_1 - \mu_2 = 0$), what is the probability of seeing <u>at least</u> the observed diff in sample means ($|\bar{x}_1 - \bar{x}_2|$ or more <u>extreme</u>)?

TWO-TAILED TEST:

$H_0: \mu_1 - \mu_2 = 0$

$H_A: \mu_1 - \mu_2 \neq 0$



If there is no diff in population means ($\mu_1 - \mu_2 = 0$), what is the probability of seeing <u>at least</u> the observed diff in sample means ($|\bar{x}_1 - \bar{x}_2|$ or more <u>extreme</u>)?

ALPHA LEVEL: α Also known as the **SIGNIFICANCE LEVEL**. The level of probability (p-value) below which the null hypothesis is rejected. It is customary to set alpha at the **.05**, **.01**, or **.001** level.

$p < \alpha$ *If the p-value is __**less than**__ alpha, we __**reject**__ the null hypothesis.*

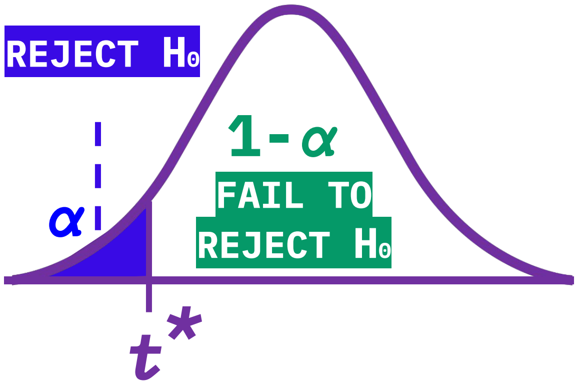
$p > \alpha$ *If the p-value is __**greater than**__ alpha, we __**fail to reject**__ the null hypothesis.*

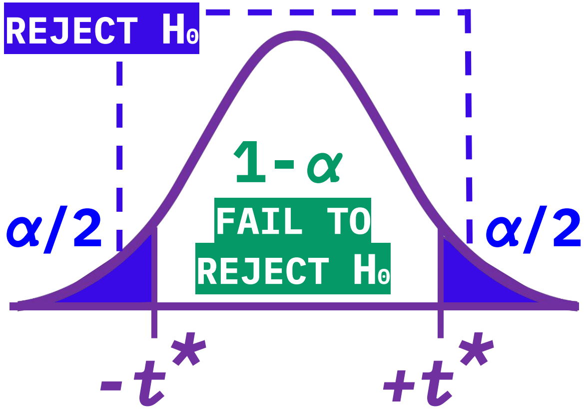
CRITICAL VALUE: t^* The t-statistic associated with the alpha level (α).

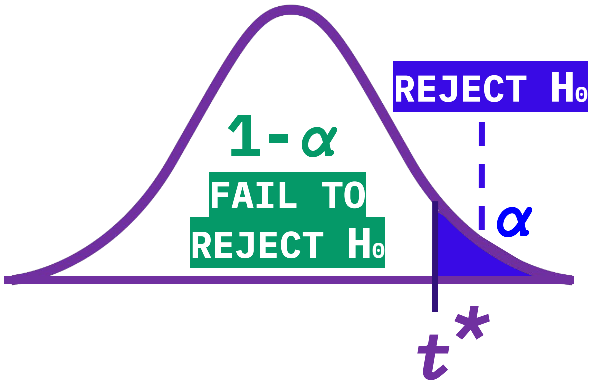
$|t| > |t^*|$ *If the observed t-score is __**more extreme**__ than the critical value, we __**reject**__ the null hypothesis.*

$|t| < |t^*|$ *If the observed t-score is __**less extreme**__ than the critical value, we __**fail to reject**__ the null hypothesis.*

ONE-TAILED TEST (LEFT-SIDED)	TWO-TAILED TEST	ONE-TAILED TEST (RIGHT-SIDED)
$H_0 : \mu_1 = \mu_2$ $H_A : \mu_1 < \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_A : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_A : \mu_1 > \mu_2$







T-TESTS & CONFIDENCE INTERVALS

When we run a t-test, we are essentially calculating a **confidence interval** for the **difference between means**:

$$\overbrace{(\bar{X}_1-\bar{X}_2)}^{\text{SIGNAL}} + t_\alpha\overbrace{(S_{\bar{Y}_1-\bar{Y}_2})}^{\text{NOISE}}$$

$\bar{X}_1-\bar{X}_2$ = *difference in means*

t_α = *critical value*

$S_{\bar{Y}_1-\bar{Y}_2}$ = *combined SE*

T-TESTS & CONFIDENCE INTERVALS

$$\boxed{\text{CONFIDENCE LEVEL} = 1 - \alpha}$$

90% confidence level corresponds to ***alpha = 0.1***

95% confidence level corresponds to ***alpha = 0.05***

99% confidence level corresponds to ***alpha = 0.01***

COMMANDS: t-tests

ttest

description: conduct two-sample t-test to compare the means of two groups.

```
ttest write, by(sex)
```

DEPENDENT VARIABLE
(what we're comparing between groups)

INDEPENDENT VARIABLE
(the variable identifying the two groups to compare)

^^^Comparing the mean **WRITING SCORE** of **FEMALE** vs. **MALE** students

Do **LATINX PEOPLE UNDER 40** have different **INCOMES** from **LATINX PEOPLE OVER 40**?

HYPOTHESIS:

LATINX PEOPLE UNDER 40 have LOWER INCOME than LATINX PEOPLE 40 AND OVER.

1. Generate the independent variable. Let's name it **Latinx_TwoGroups**.

```
generate Latinx_TwoGroups=.
```
This creates a variable with blank values

```
(1,138 missing values generated)
```

```
replace Latinx_TwoGroups = 1 if race ==4 & age<40
```
*This codes Latinx respondents who are **under 40** as "1"*

```
(192 real changes made)
```

```
replace Latinx_TwoGroups = 2 if race ==4 & age>=40
```
*This codes Latinx respondents who are **40+** as "2"*

```
(187 real changes made)
```

Do LATINX PEOPLE UNDER 40 have different INCOMES from LATINX PEOPLE OVER 40?

2. Determine type of hypothesis test.

```
tab Latinx_TwoGroups
```

LATINX PEOPLE UNDER 40

GROUP 1

GROUP 2

LATINX PEOPLE OVER 40

Latinx_TwoGroups	Freq.	Percent	Cum.
1	192	50.66	50.66
2	187	49.34	100.00
Total	379	100.00	

Do **LATINX PEOPLE UNDER 40** have different **INCOMES** from **LATINX PEOPLE OVER 40**?

2. Determine type of hypothesis test.

HYPOTHESIS:

LATINX PEOPLE UNDER 40 have a **LOWER INCOME** than **LATINX PEOPLE 40+**.

GROUP 1

GROUP 2

H_0: μ group 1 = μ group 2

H_1: μ group 1 < μ group 2 → **LEFT-TAILED TEST**

Do **LATINX PEOPLE UNDER 40** have different **INCOMES** from **LATINX PEOPLE OVER 40**?

2. Determine type of hypothesis test.

IF THE GROUP CODES WERE REVERSED:

LATINX PEOPLE UNDER 40 have a **LOWER INCOME** than **LATINX PEOPLE 40+**.

GROUP 2

GROUP 1

H_0: μ group 1 = μ group 2

H_1: μ group 1 > μ group 2

RIGHT-TAILED TEST

Do **LATINX PEOPLE UNDER 40** have different **INCOMES** from **LATINX PEOPLE OVER 40**?

3. Run t-test.

```
ttest incwage, by(Latinx_TwoGroups)
```

H_1: μ group 1 < μ group 2

Two-sample t test with equal variances

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
1	192	49234.9	3200.404	44346.1	42922.22	55547.57
2	187	67639.04	5183.168	70878.75	57413.68	77864.39
Combined	379	58315.57	3060.778	59586.98	52297.28	64333.85
diff		-18404.14	6056.479		-30312.85	-6495.43

```
    diff = mean(1) - mean(2)                                      t =  -3.0388
H0: diff = 0                                     Degrees of freedom =      377

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0013         Pr(|T| > |t|) = 0.0025          Pr(T > t) = 0.9987
```

Do **LATINX PEOPLE UNDER 40** have different **INCOMES** from **LATINX PEOPLE OVER 40**?

3. Run t-test.

```
ttest incwage, by(Latinx_TwoGroups)
```

H_1: µ group 1 < µ group 2

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       1 |     192     49234.9    3200.404     44346.1    42922.22    55547.57
       2 |     187    67639.04    5183.168    70878.75    57413.68    77864.39
---------+--------------------------------------------------------------------
Combined |     379    58315.57    3060.778    59586.98    52297.28    64333.85
---------+--------------------------------------------------------------------
    diff |           -18404.14    6056.479               -30312.85    -6495.43
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =  -3.0388
H0: diff = 0                                     Degrees of freedom =      377

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0013         Pr(|T| > |t|) = 0.0025          Pr(T > t) = 0.9987
```

µ group 1 - µ group 2 < 0

Do **LATINX PEOPLE UNDER 40** have different **INCOMES** from **LATINX PEOPLE OVER 40**?

4. Interpret results.

CONFIDENCE INTERVAL: -$**30312.85** ---> -$**6495.43**

The 95% CI for the difference in means **DOES NOT OVERLAP 0.**

T-STATISTIC: **-3.0388**

P-VALUE: **0.0013** 0.0013 is **LESS THAN** 0.05

CONCLUSION: We **REJECT THE NULL HYPOTHESIS** at an alpha level of 0.05. There is a **STATISTICALLY SIGNIFICANT DIFFERENCE** in the **INCOME** of **LATINX PEOPLE <40** vs. **LATINX PEOPLE 40+**. On average, **LATINX PEOPLE < 40** **MAKE LESS** than **LATINX PEOPLE 40+**.

Is there a difference in the **WEEKLY HOURS WORKED** by **IMMIGRANT ASIAN MEN** vs. **NON-IMMIGRANT ASIAN MEN**?

HYPOTHESIS:

IMMIGRANT ASIAN MEN work MORE HOURS WEEKLY than NON-IMMIGRANT ASIAN MEN.

1. Generate the independent variable. Let's name it **APImen**.

```
generate APImen=.
```

This creates a variable with blank values

```
(1,138 missing values generated)
```

```
replace APImen=1 if race==3 & sex==1 & immigrant==1
```

*This codes Asian men who are **immigrants** as "1"*

```
(105 real changes made)
```

```
replace APImen=2 if race==3 & sex==1 & immigrant==0
```

*This codes Asian men who are **not immigrants** as "2"*

```
(33 real changes made)
```

Is there a difference in the **WEEKLY HOURS WORKED** by **IMMIGRANT ASIAN MEN** vs. **NON-IMMIGRANT ASIAN MEN**?

HYPOTHESIS:

IMMIGRANT ASIAN MEN work **MORE HOURS WEEKLY** than **NON-IMMIGRANT ASIAN MEN**.

2. Determine type of hypothesis test.

```
tab APImen
```

IMMIGRANT ASIAN MEN

GROUP 1

APImen	Freq.	Percent	Cum.
1	105	76.09	76.09
2	33	23.91	100.00
Total	138	100.00	

GROUP 2

NON-IMMIGRANT ASIAN MEN

Is there a difference in the **WEEKLY HOURS WORKED** by **IMMIGRANT ASIAN MEN** vs. **NON-IMMIGRANT ASIAN MEN**?

2. Determine type of hypothesis test.

HYPOTHESIS:

IMMIGRANT ASIAN MEN work **MORE WEEKLY HRS** than **NON-IMMIGRANT ASIAN MEN**

GROUP 1

GROUP 2

H_0: μ group 1 = μ group 2

H_1: μ group 1 > μ group 2 → **RIGHT-TAILED TEST**

Is there a difference in the **WEEKLY HOURS WORKED** by **IMMIGRANT ASIAN MEN** vs. **NON-IMMIGRANT ASIAN MEN**?

3. Run t-test.

```
ttest uhrs, by(APImen)
```

H_1: μ group 1 > μ group 2

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       1 |     105    41.82857    .7193629    7.371276    40.40205     43.2551
       2 |      33     39.9697    1.742128    10.00776     36.4211     43.5183
---------+--------------------------------------------------------------------
Combined |     138    41.38406    .6877541    8.079281    40.02407    42.74404
---------+--------------------------------------------------------------------
    diff |            1.858874    1.610402               -1.325794    5.043543
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =   1.1543
H0: diff = 0                                     Degrees of freedom =      136

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.8748         Pr(|T| > |t|) = 0.2504          Pr(T > t) = 0.1252
```

Is there a difference in the **WEEKLY HOURS WORKED** by **IMMIGRANT ASIAN MEN** vs. **NON-IMMIGRANT ASIAN MEN**?

3. Run t-test.

```
ttest uhrs, by(APImen)
```

H_1: μ group 1 > μ group 2

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       1 |     105    41.82857    .7193629    7.371276    40.40205    43.2551
       2 |      33     39.9697    1.742128    10.00776     36.4211    43.5183
---------+--------------------------------------------------------------------
Combined |     138    41.38406    .6877541    8.079281    40.02407    42.74404
---------+--------------------------------------------------------------------
    diff |            1.858874    1.610402               -1.325794    5.043543
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =   1.1543
H0: diff = 0                                     Degrees of freedom =      136

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.8748         Pr(|T| > |t|) = 0.2504          Pr(T > t) = 0.1252
```

μ group 1 - μ group 2 > 0

Is there a difference in the **WEEKLY HOURS WORKED** by **IMMIGRANT ASIAN MEN** vs. **NON-IMMIGRANT ASIAN MEN**?

4. Interpret results.

CONFIDENCE INTERVAL: **-1.326 ---> 5.044**

The 95% CI for the difference in means **OVERLAPS 0**.

T-STATISTIC: **1.1543**

P-VALUE: **0.1252** 0.1252 is **GREATER THAN** 0.05

CONCLUSION: We **FAIL TO REJECT THE NULL HYPOTHESIS** at an alpha level of 0.05. There is **NOT ENOUGH STATISTICAL EVIDENCE** to conclude that there is a difference in the **WEEKLY HOURS WORKED** by **IMMIGRANT ASIAN MEN** vs. **NON-IMMIGRANT ASIAN MEN**.