CHRISTOPHER PRENER, PH.D.
FALL, 2016

WEEK 08
LECTURE 10

QUANTITATIVE ANALYSIS

DIFFERENCE OF MEANS (1)

AGENDA

1. Follow-up
2. Revisiting Distributions
3. One Sample
4. Independent Samples
5. Dependent Samples
6. Effect Sizes

1 FOLLOW-UP

2 REVISITING DISTRIBUTIONS

VARIANCE

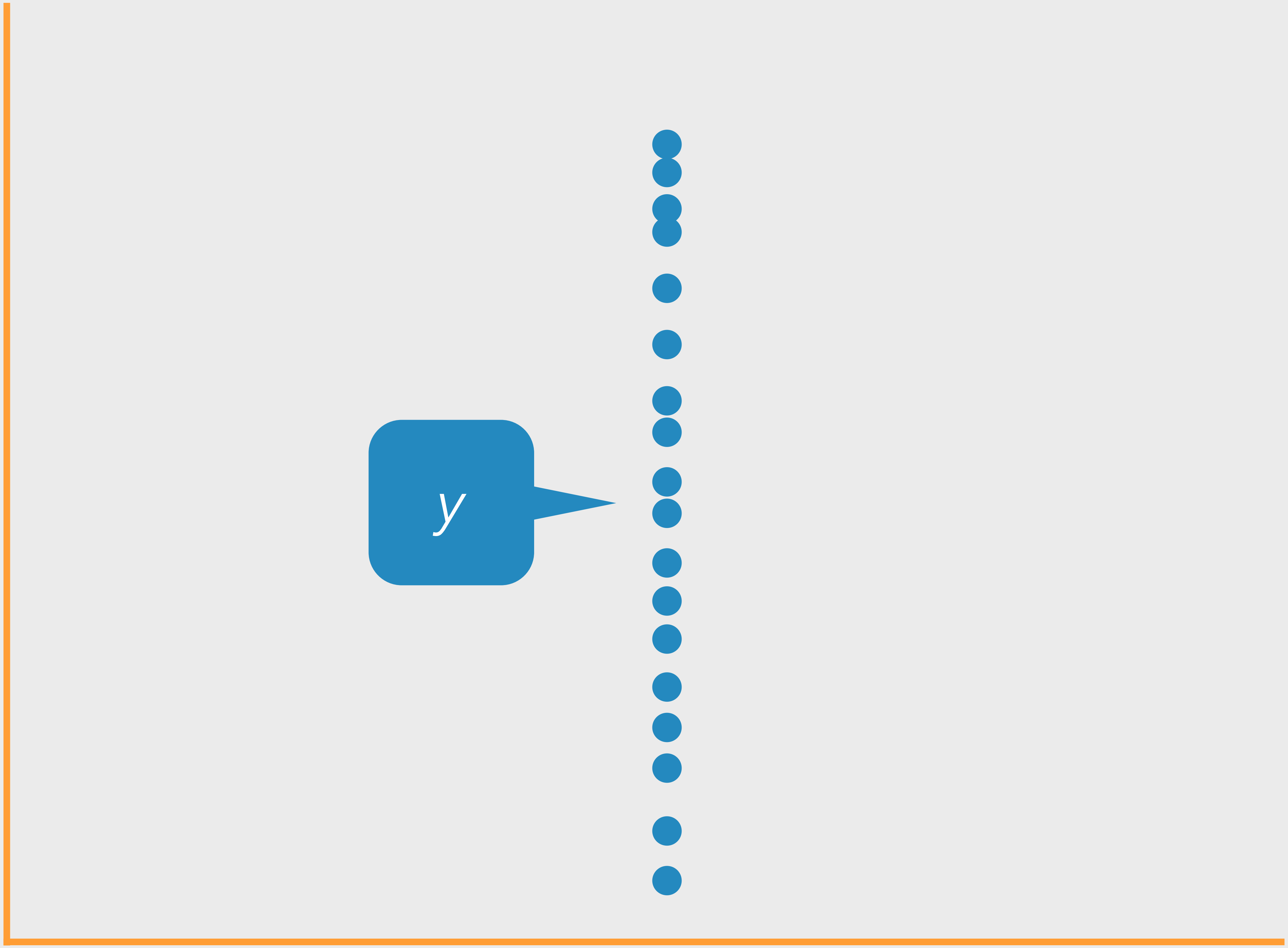
SECOND MOMENT

$$s^2 = \frac{\sum_{i=1}^{n} (x - \overline{x})^2}{n-1}$$

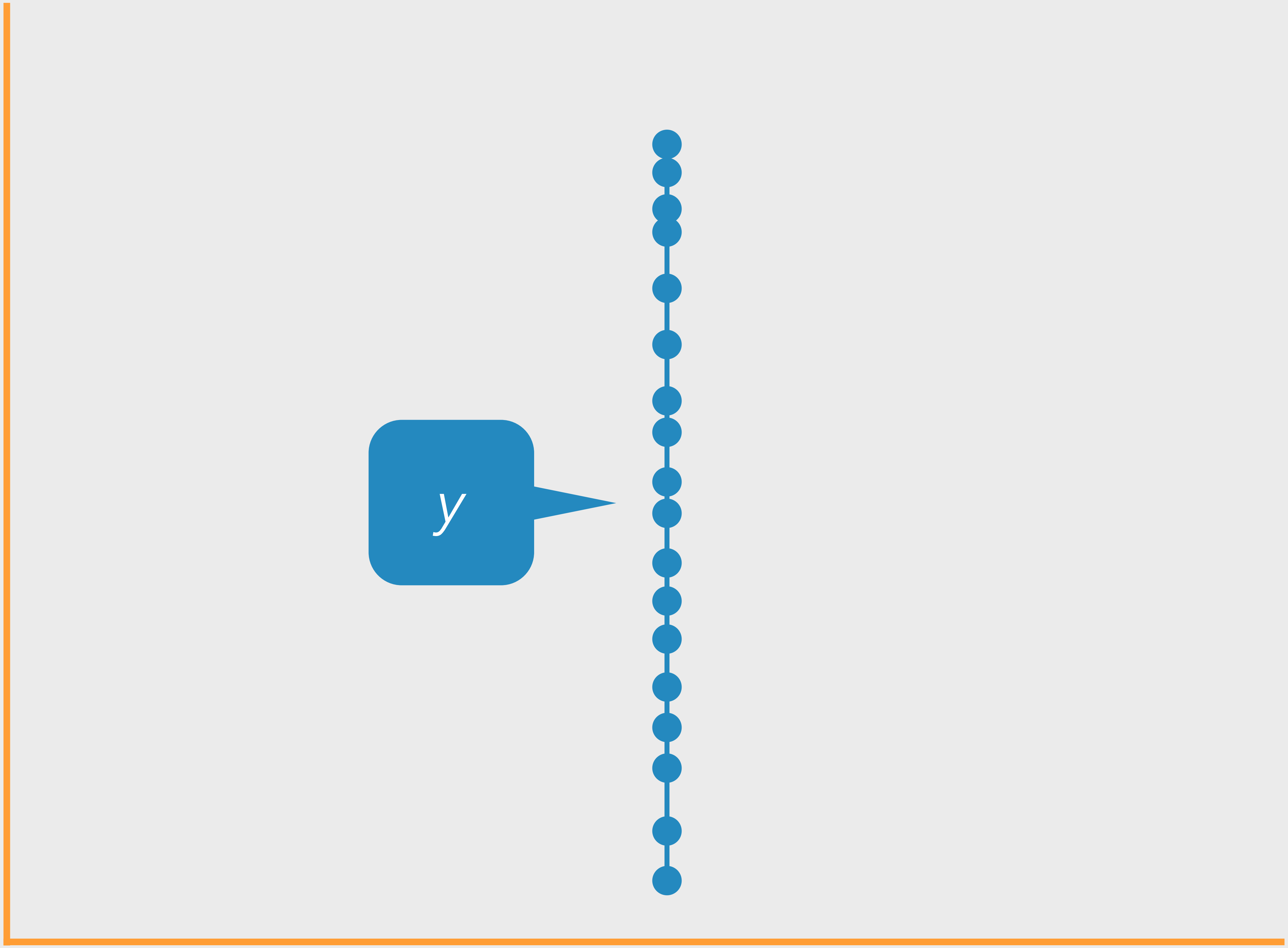
DEFINITION

SUM OF ALL DEVIANCES, SQUARED AND DIVIDED BY ONE DEGREE OF FREEDOM; EXPECTATION OF HOW DISTRIBUTION DEVIATES FROM THE MEAN

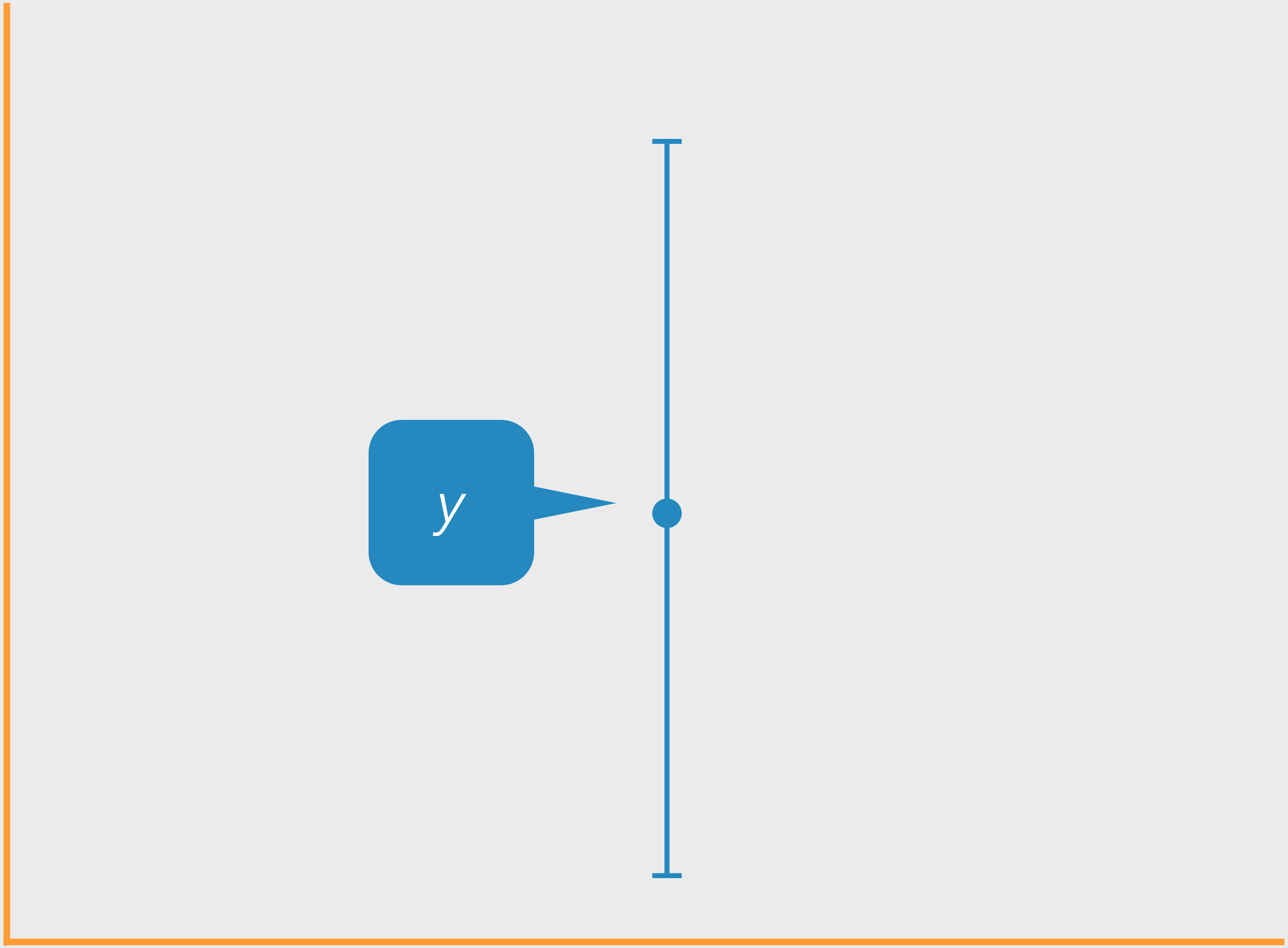
VARIANCE



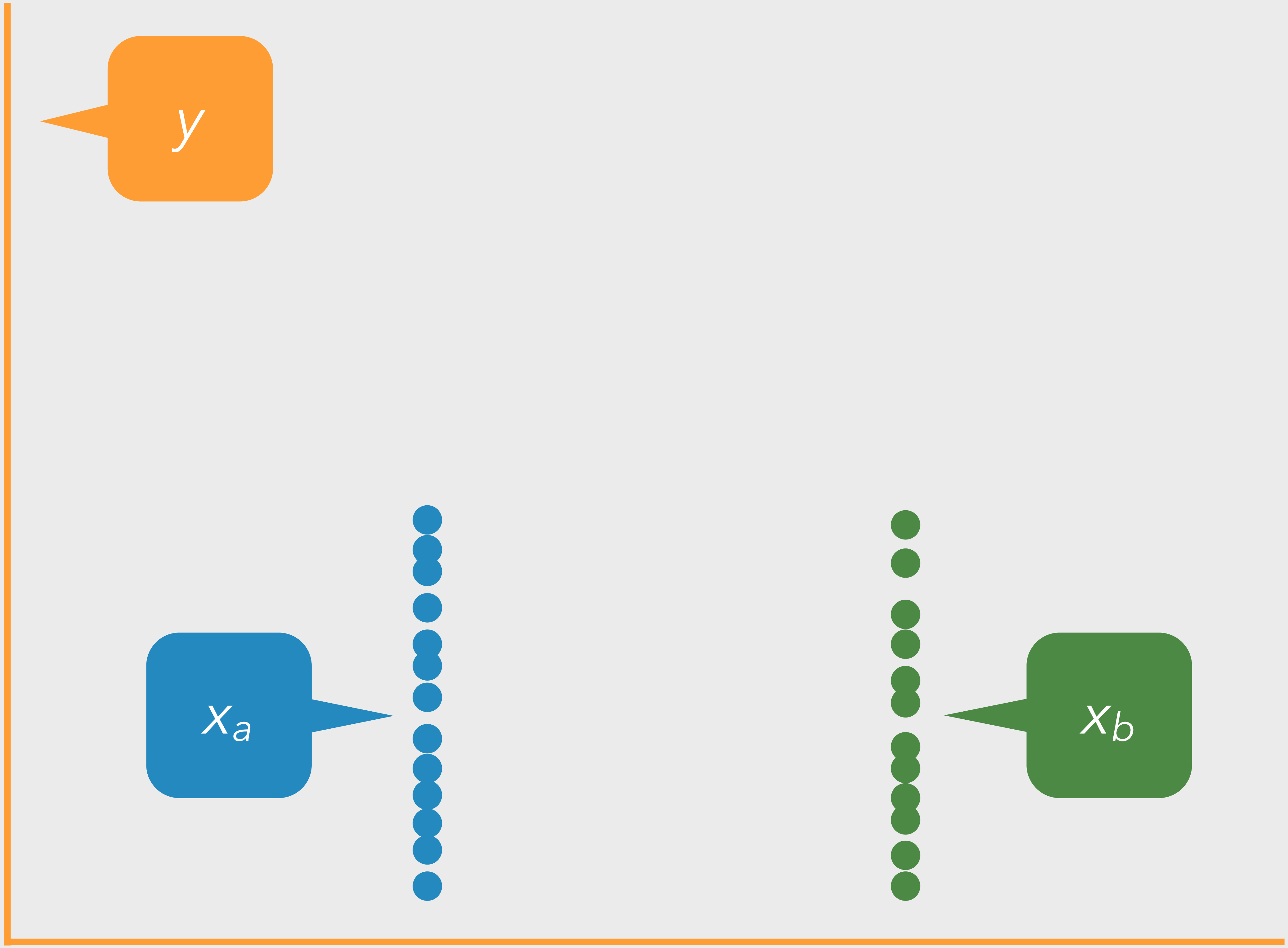
VARIANCE



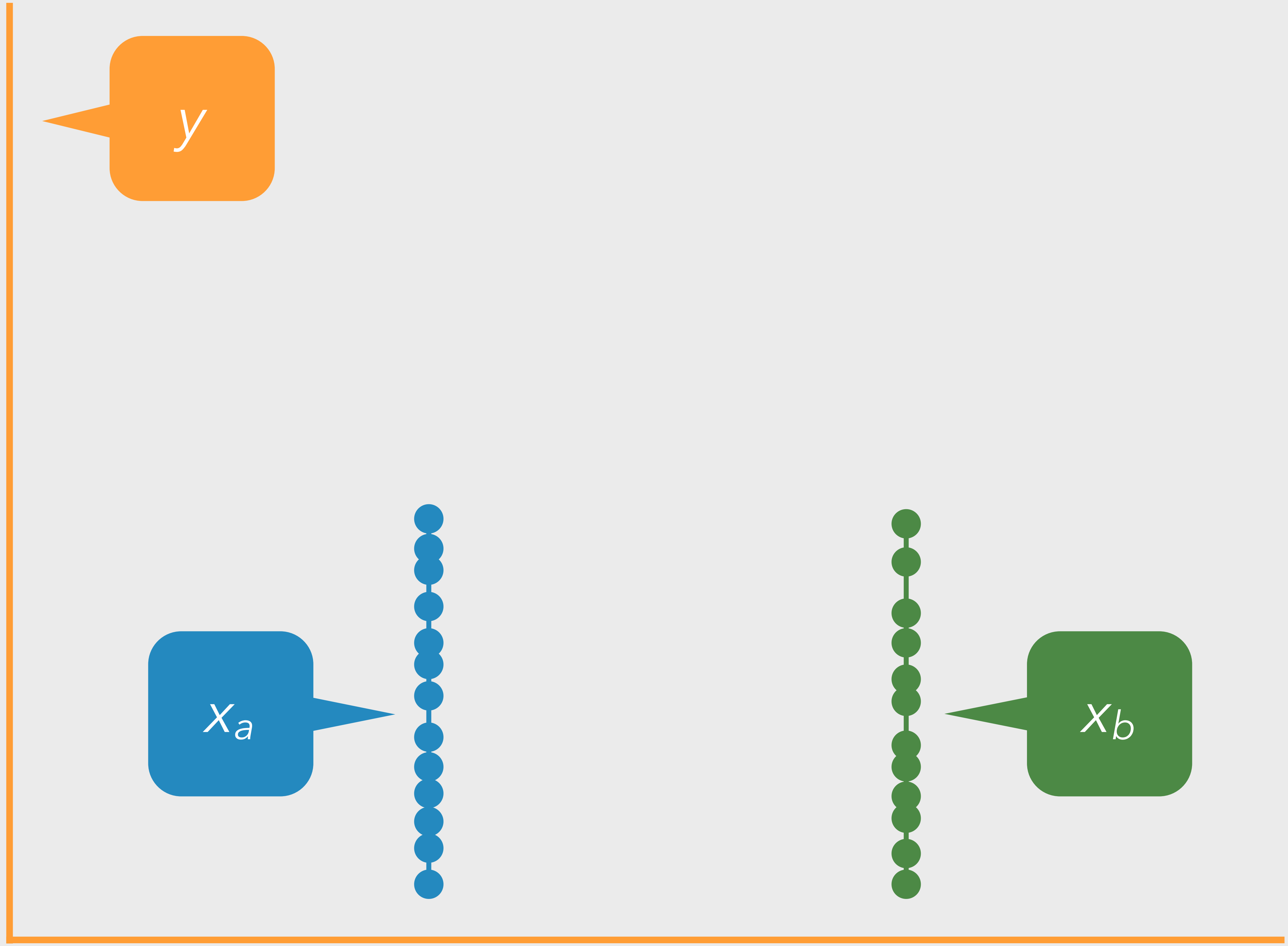
VARIANCE



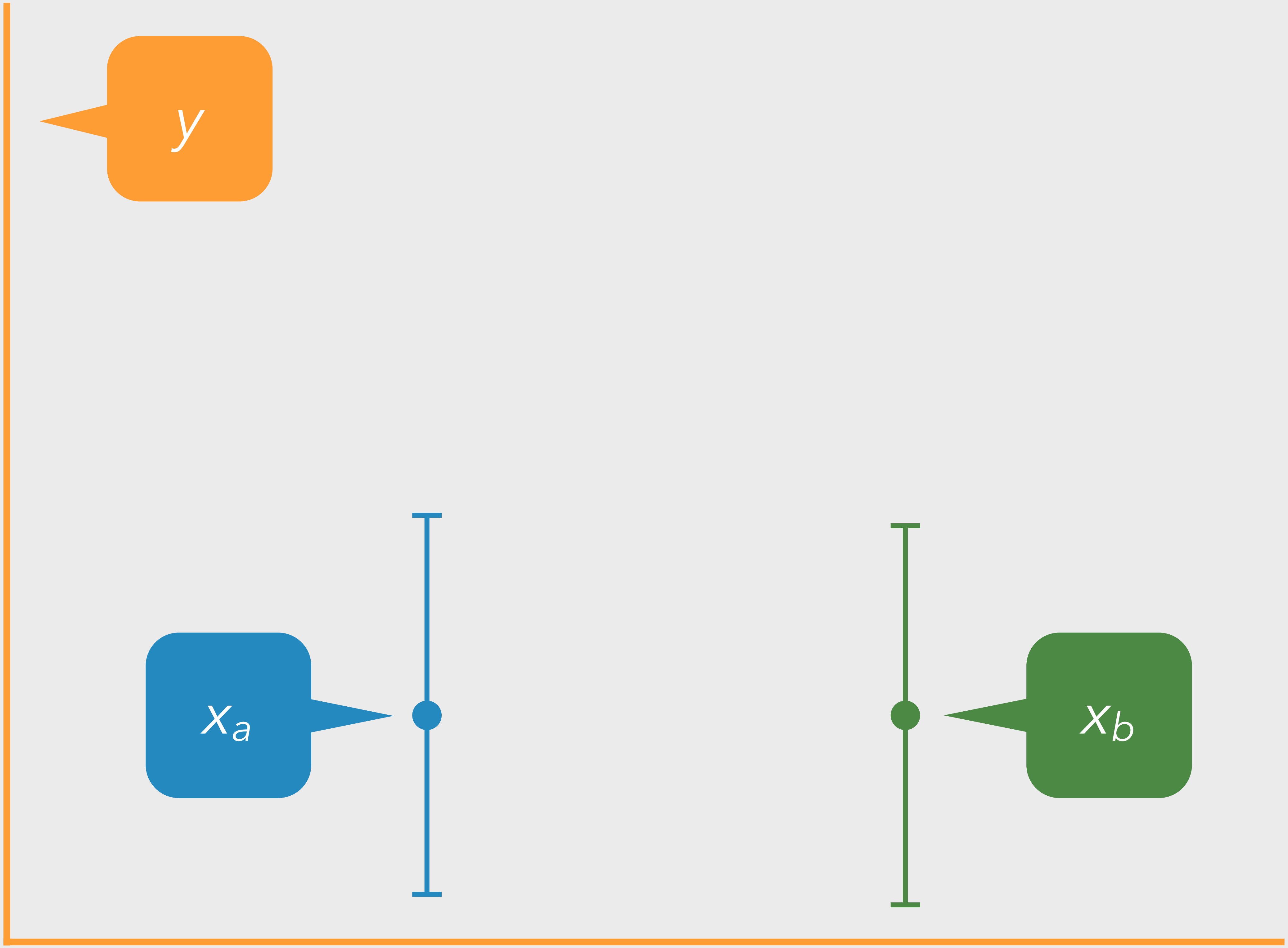
VARIANCE



VARIANCE

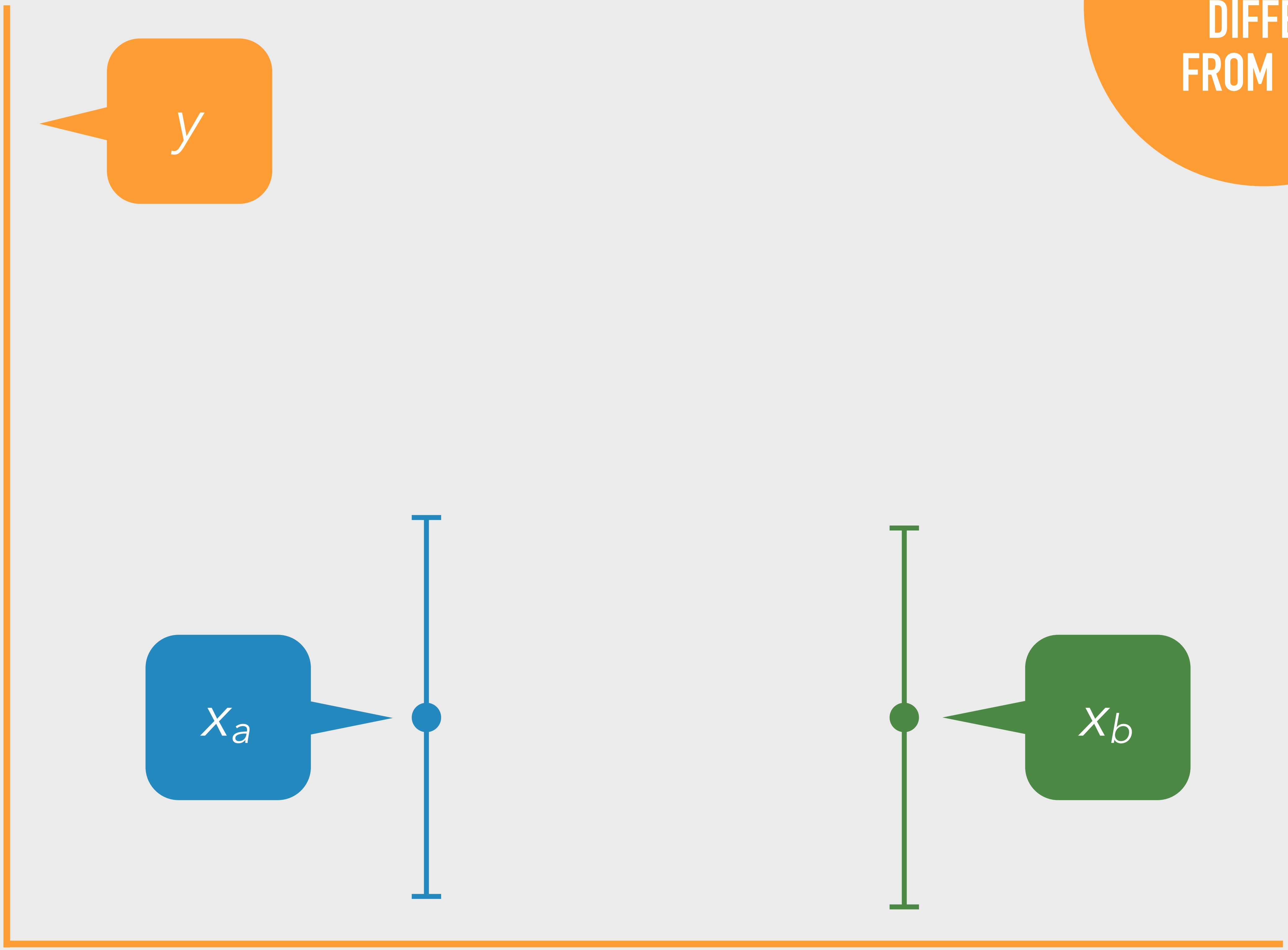


VARIANCE

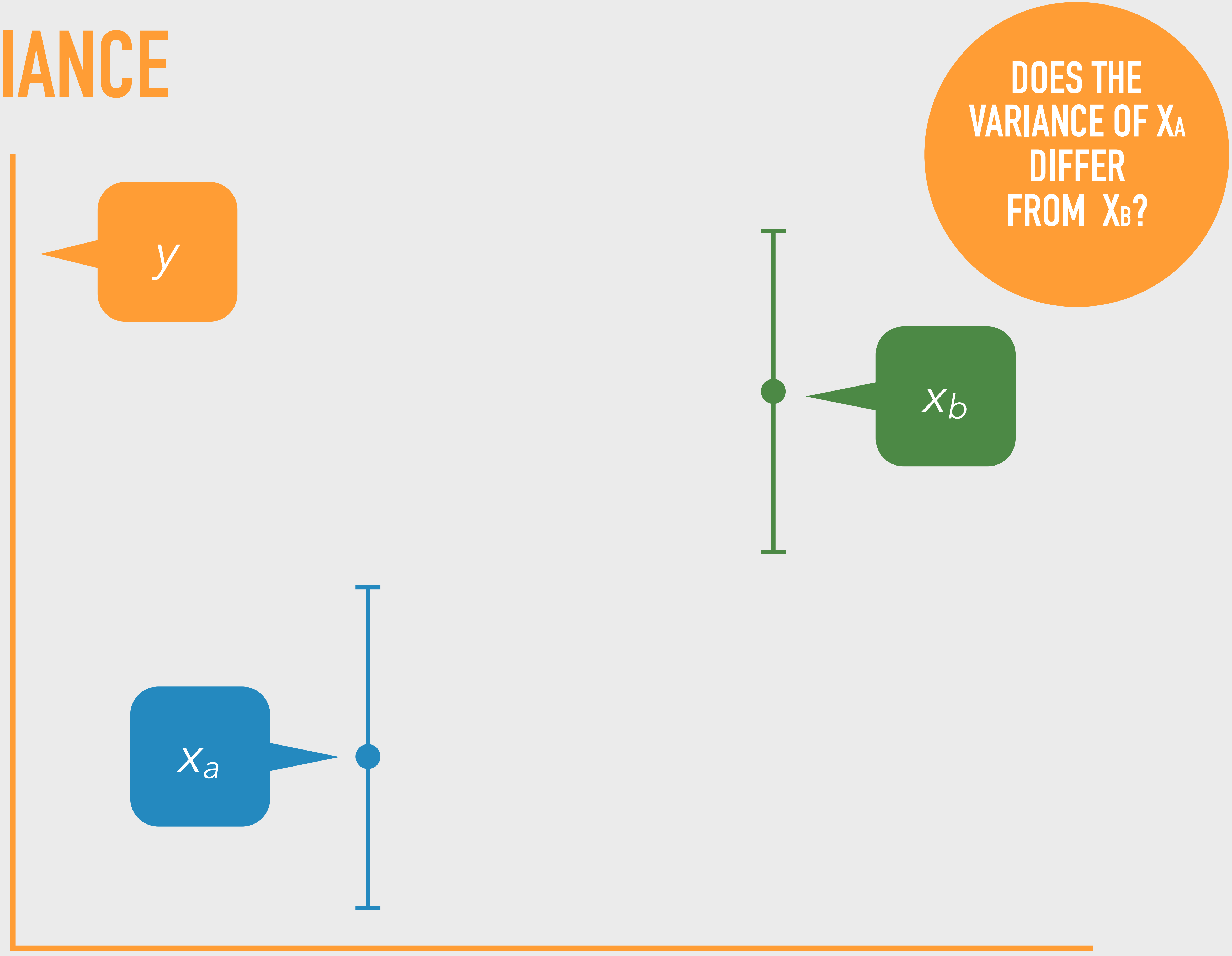


VARIANCE

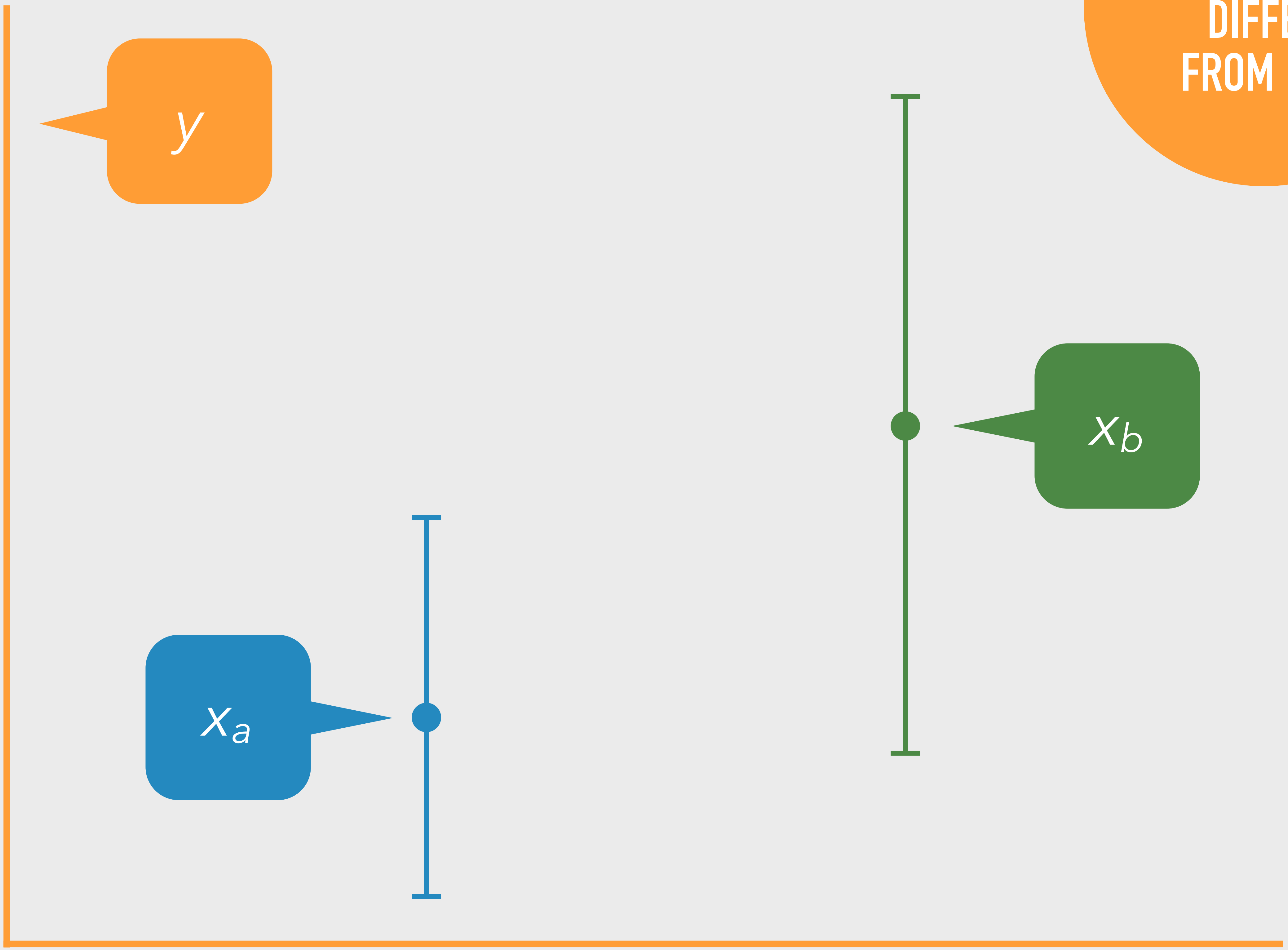
DOES THE VARIANCE OF X_A DIFFER FROM X_B?



VARIANCE



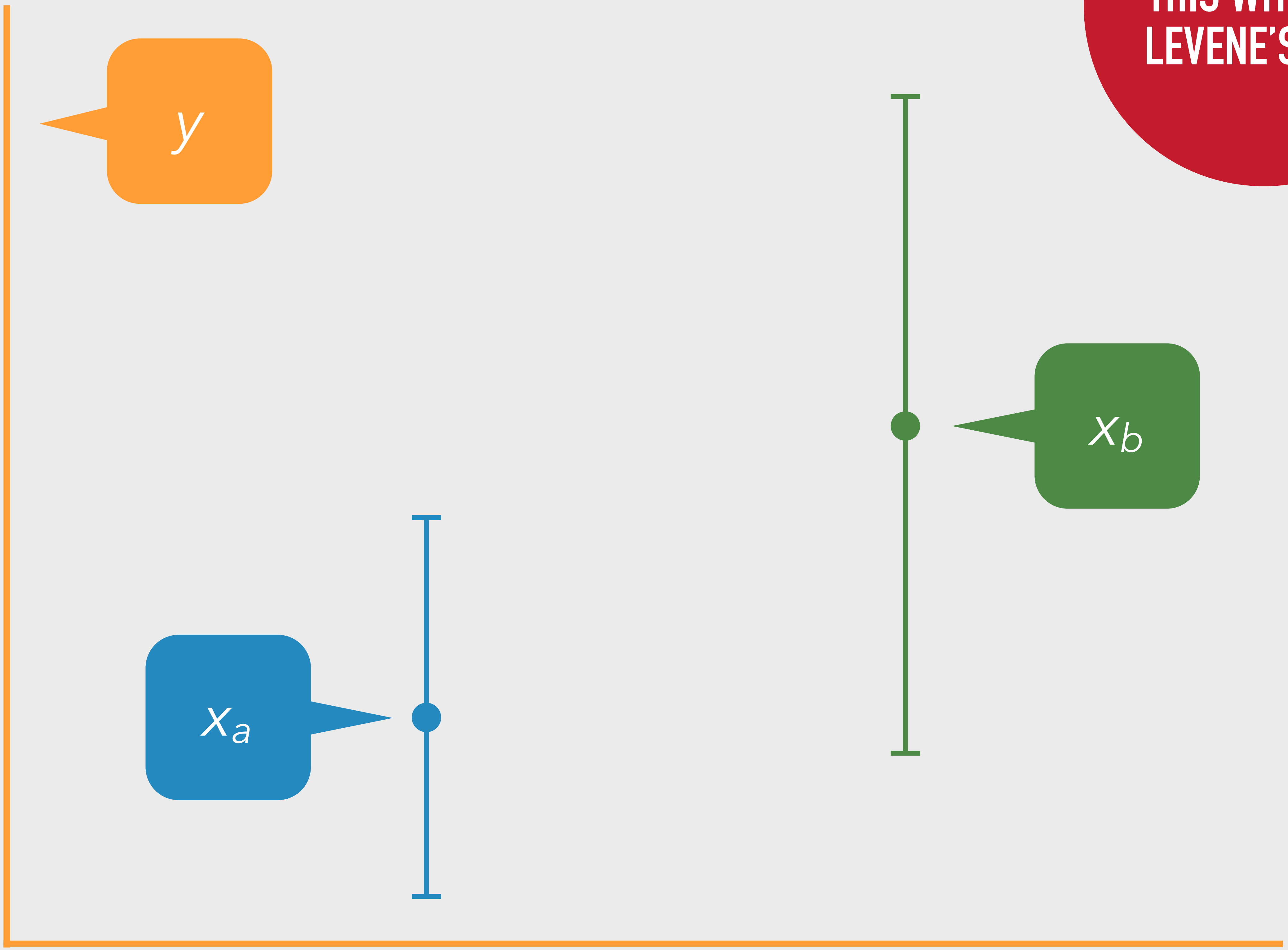
VARIANCE



DOES THE VARIANCE OF X_A DIFFER FROM X_B?

VARIANCE

WE CAN TEST THIS WITH THE LEVENE'S TEST

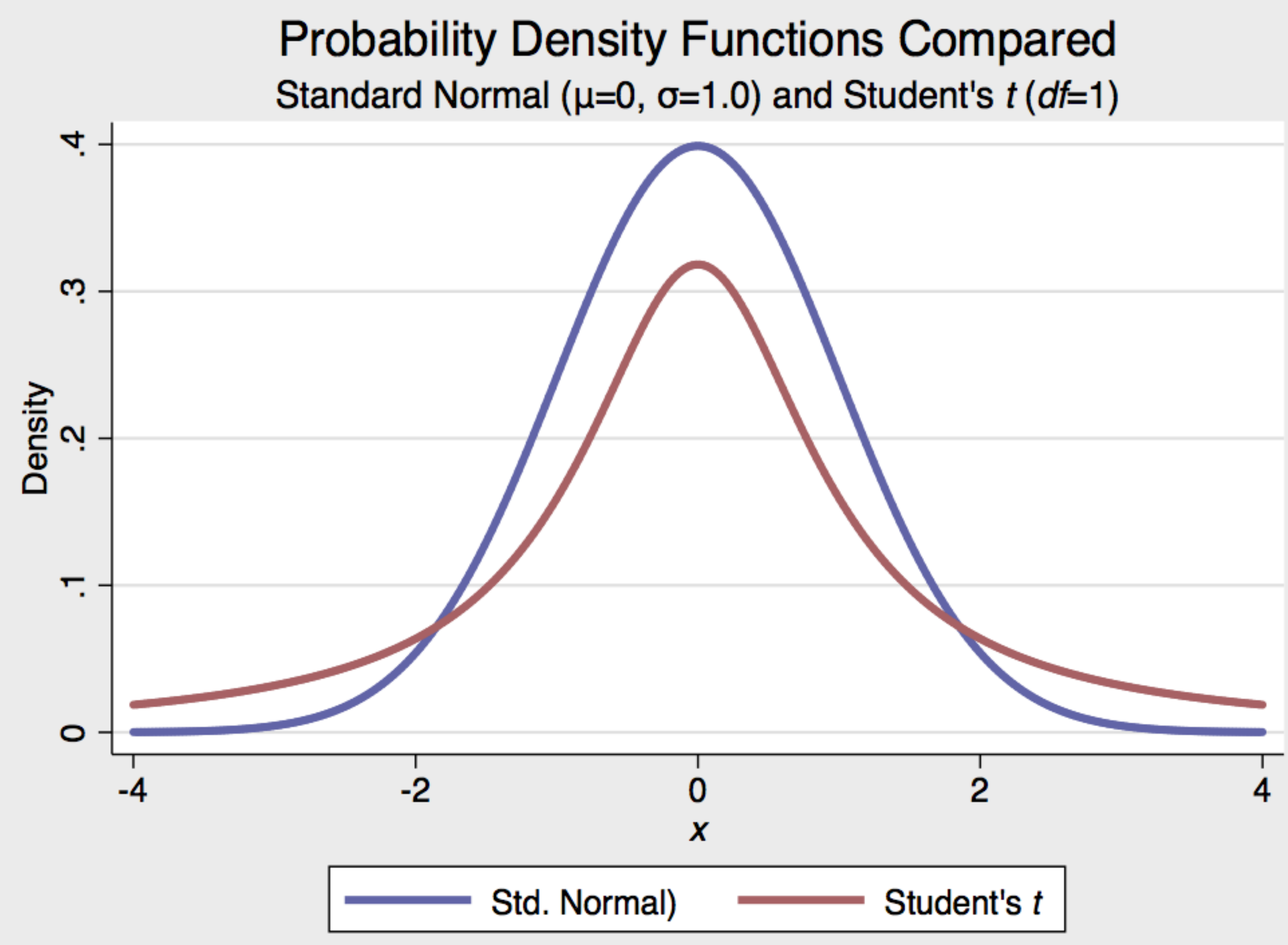


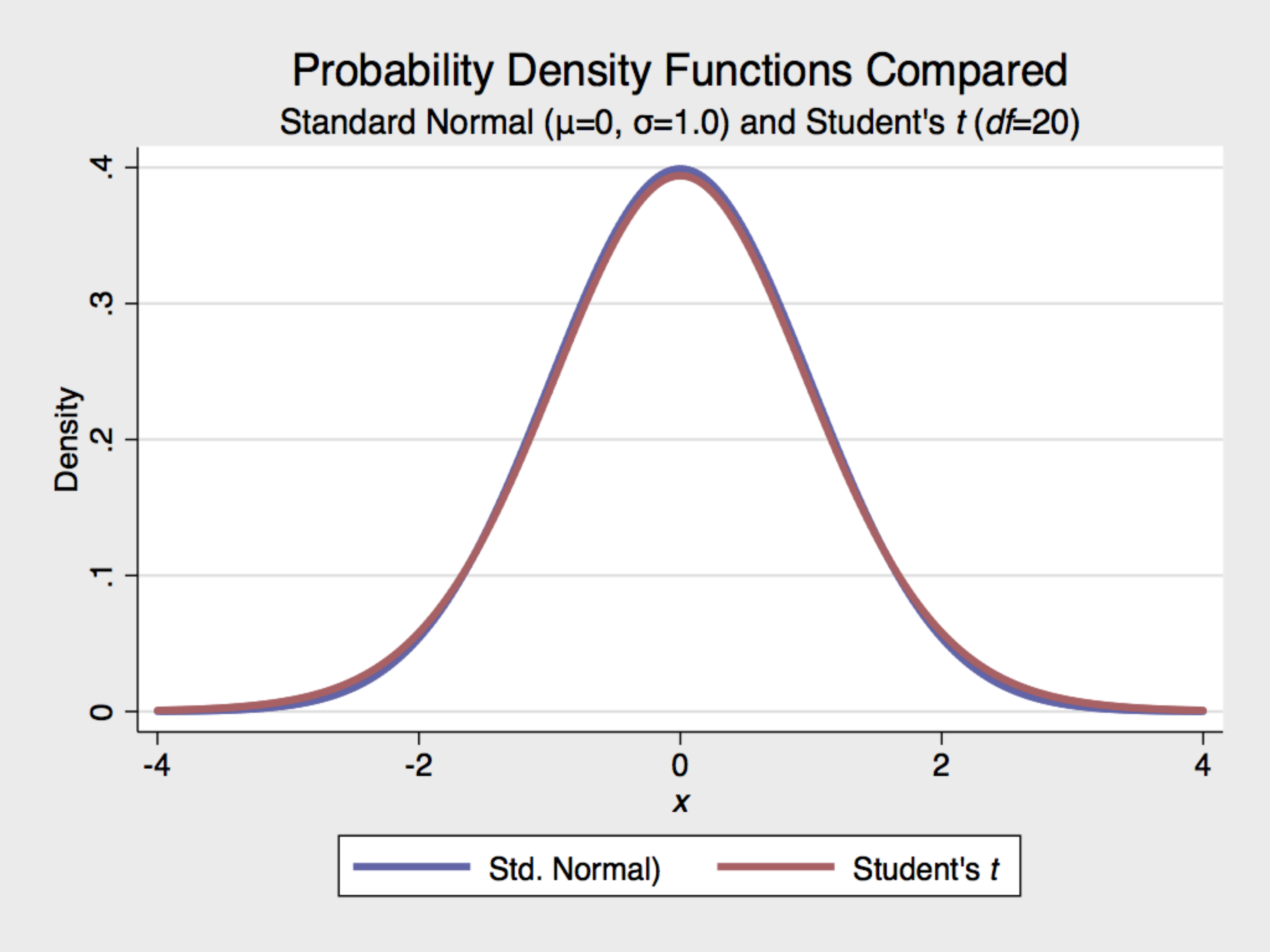
3 ONE SAMPLE

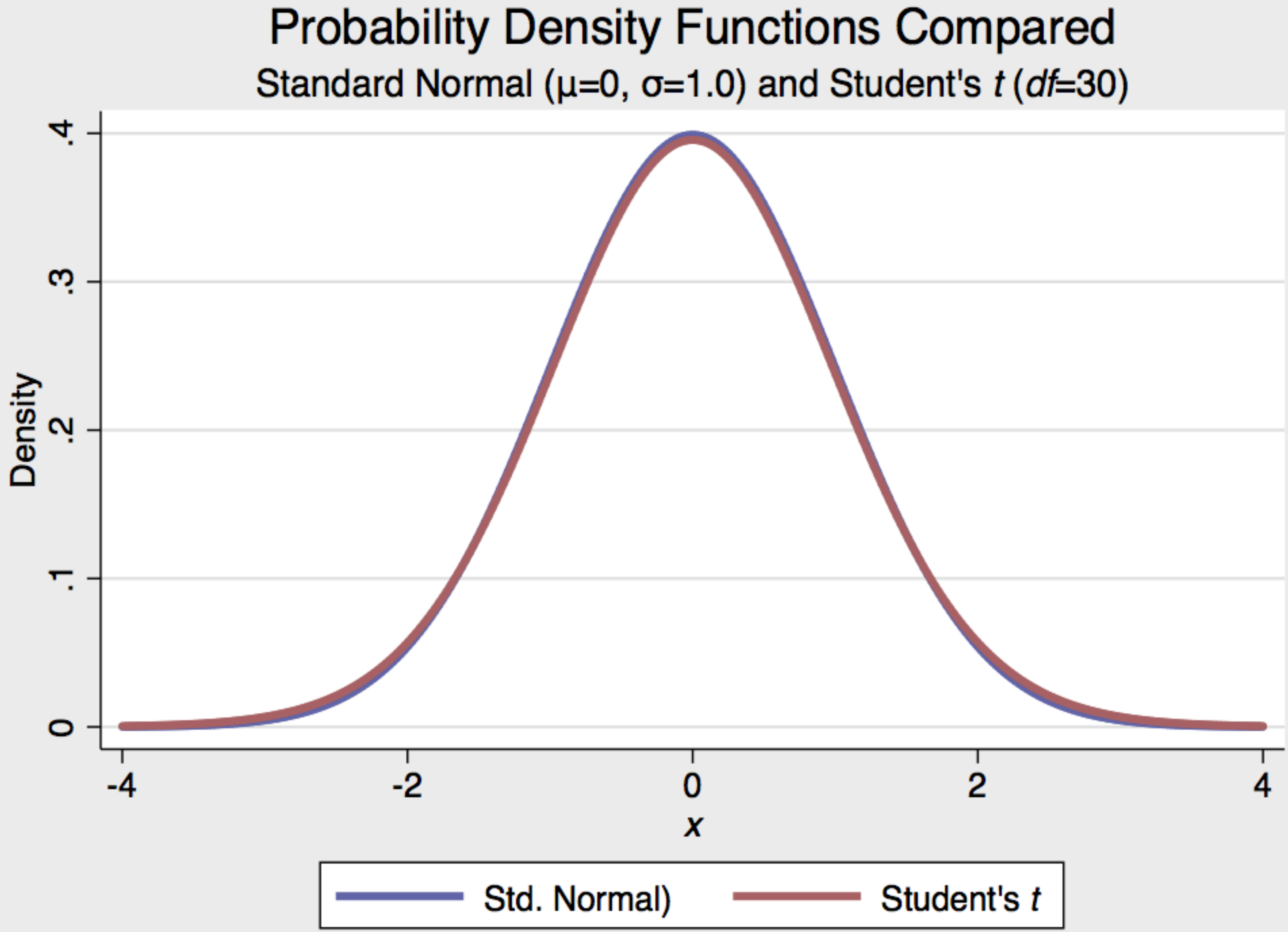
STUDENT'S T-TEST

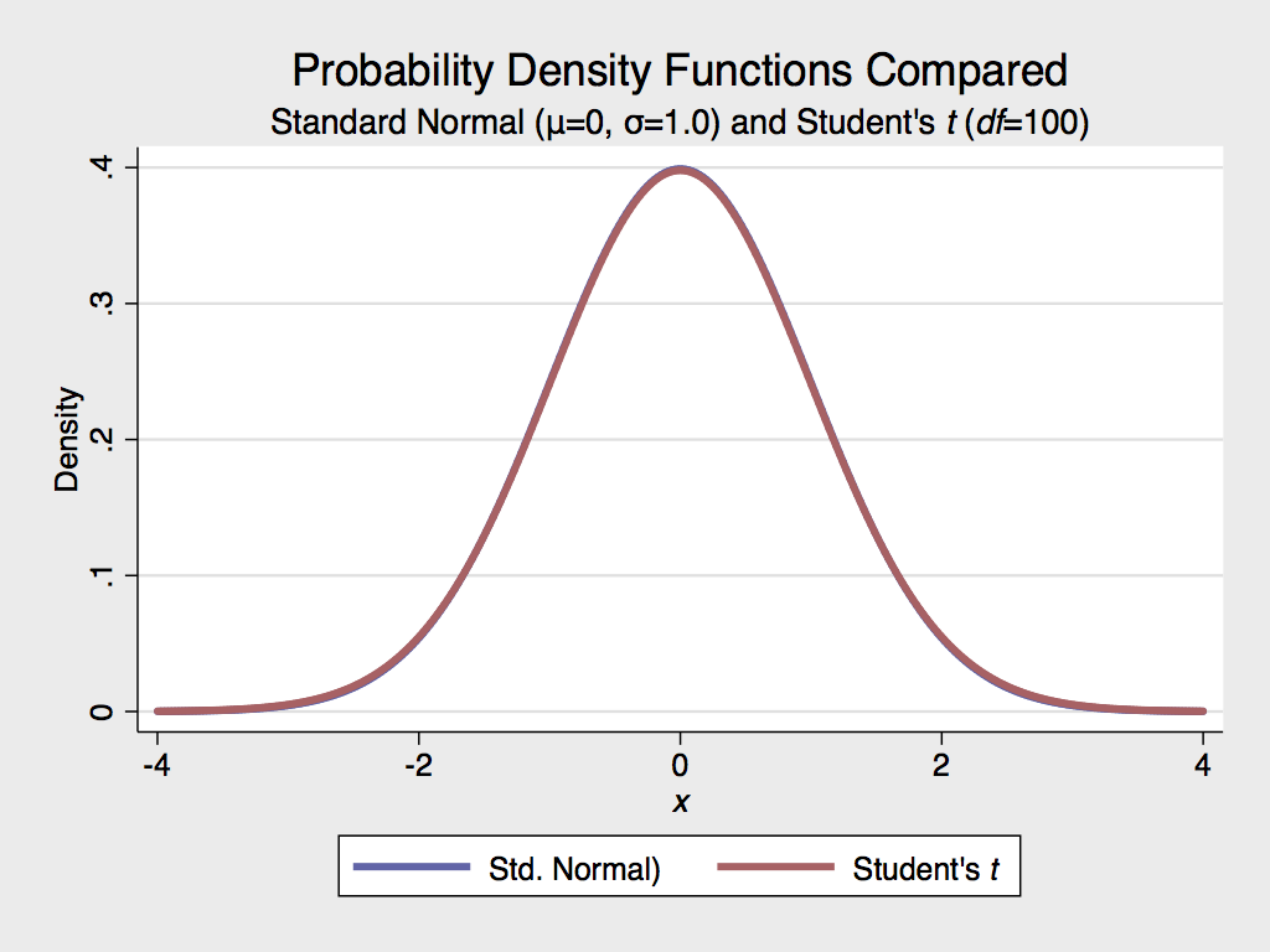
- Employee of the Guinness company who published his work under the pseudonym "Student".
- Student of Karl Pearson's while on research leaves from Guinness.
- Original t-tests were developed to conducting quality control testing on Guinness stout.

WILLIAM SEALY GOSSET (1876–1937)
"STUDENT"


Probability Density Functions Compared
Standard Normal (μ=0, σ=1.0) and Student's *t* (*df*=1)
Density
0
.1
.2
.3
.4
-4
-2
0
2
4
x
Std. Normal)
Student's *t*

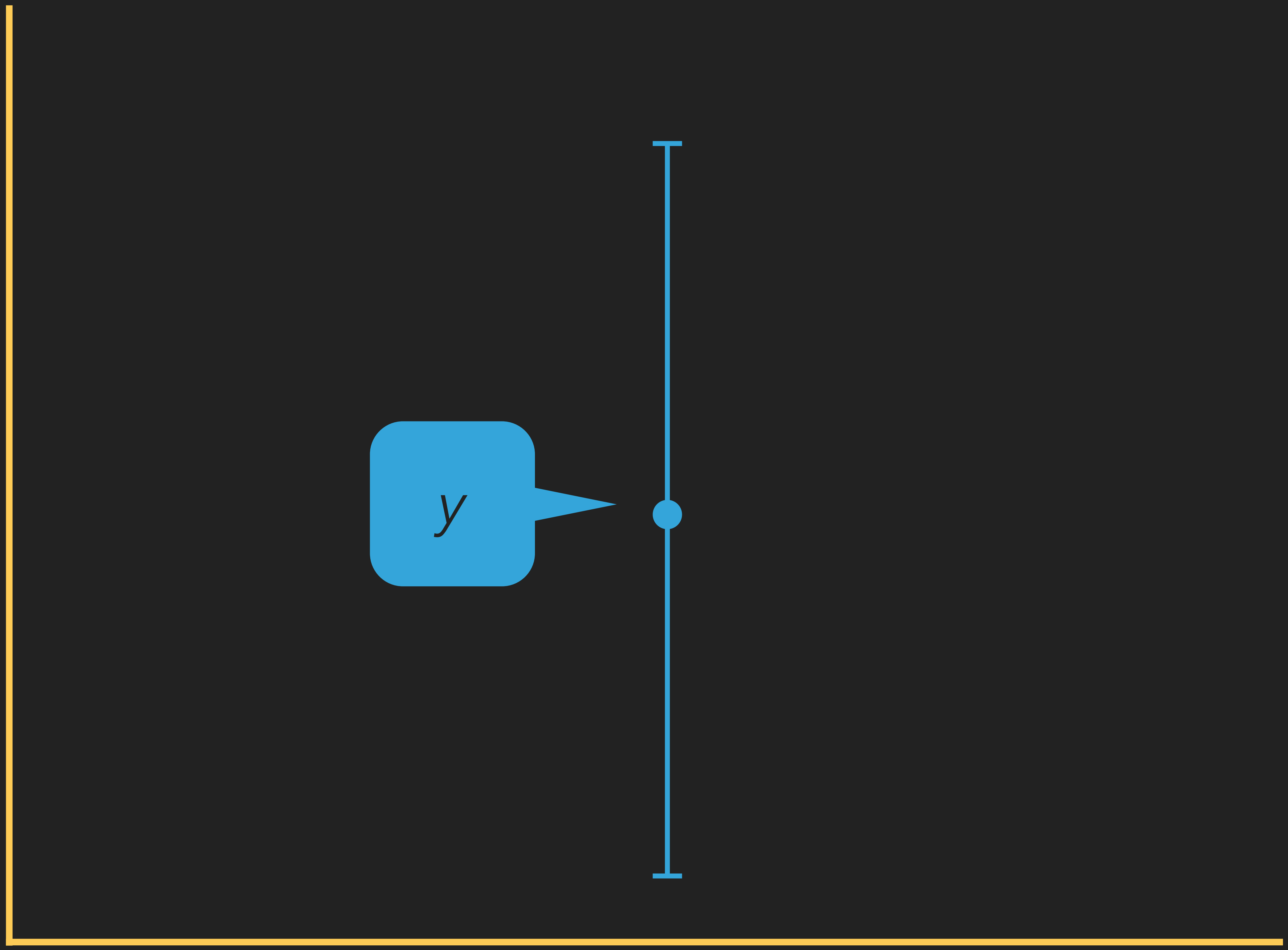

Probability Density Functions Compared
Standard Normal (μ=0, σ=1.0) and Student's t (df=20)
Density
0
.1
.2
.3
.4
-4
-2
0
2
4
x
Std. Normal)
Student's t


Probability Density Functions Compared
Standard Normal (μ=0, σ=1.0) and Student's t (df=30)
Density
0
.1
.2
.3
.4
-4
-2
0
2
4
x
Std. Normal)
Student's t



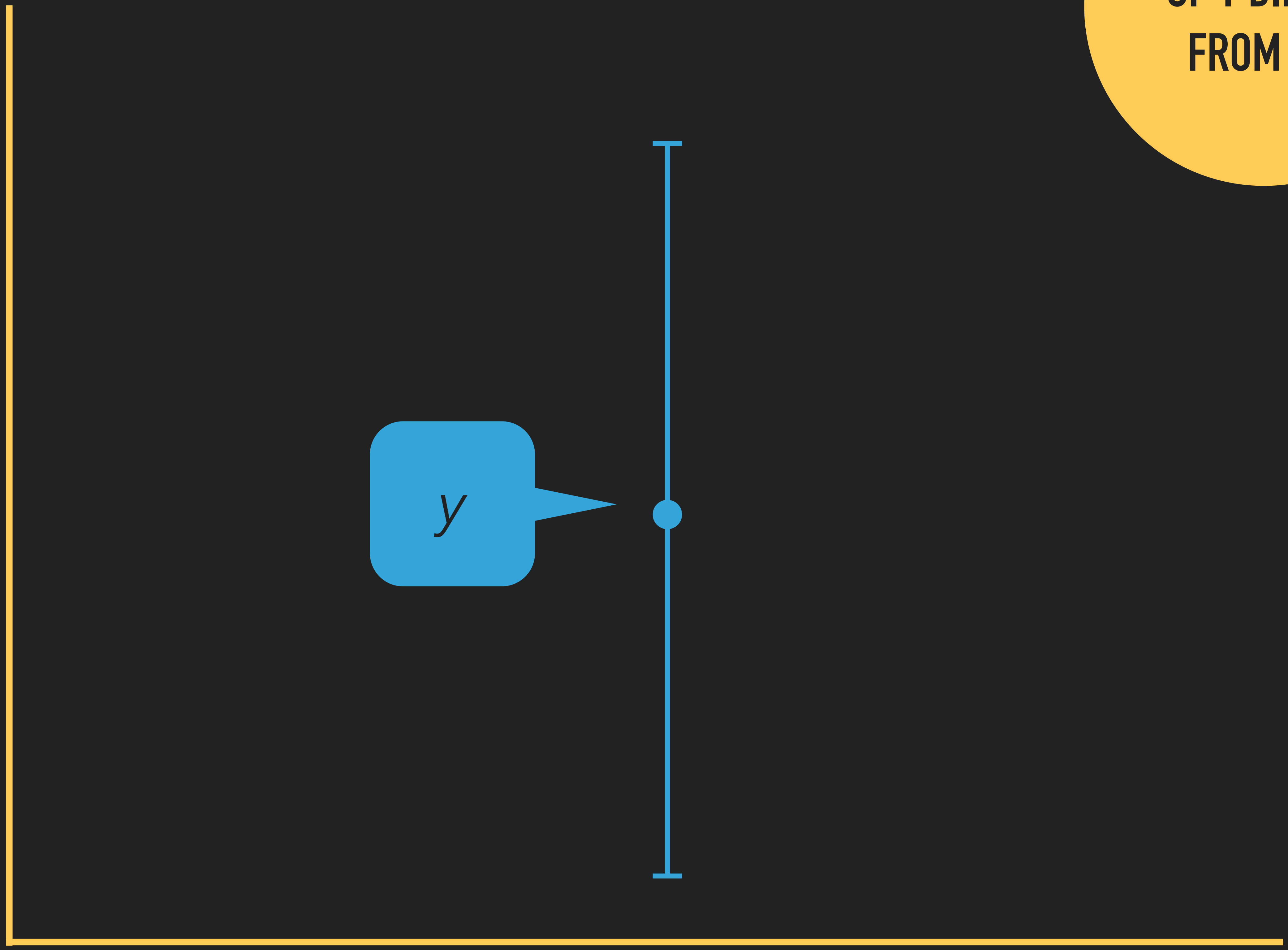
Probability Density Functions Compared
Standard Normal (μ=0, σ=1.0) and Student's t (df=100)
Density
0
.1
.2
.3
.4
-4
-2
0
2
4
x
Std. Normal)
Student's t

DIFFERENCE IN MEANS



DIFFERENCE IN MEANS

DOES THE MEAN OF Y DIFFER FROM μ?



HYPOTHESES

- H_0 = there is no significant difference between the mean of *y* and the population
- H_1 = there is a significant difference between the mean of *y* and the population

ASSUMPTIONS

- continuous data (y)
- the distribution of y is approximately normal
- degrees of freedom (v) = n-1

FORMULA

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

FORMULA

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

standard error

FIND THE PROBABILITY OF T

```
display ttail(df,t)*2
```

```
. display ttail(72,3.6308)*2

.0005255
```

```
. display ttail(72,1.6308)*2

.1072996
```

FIND THE PROBABILITY OF T

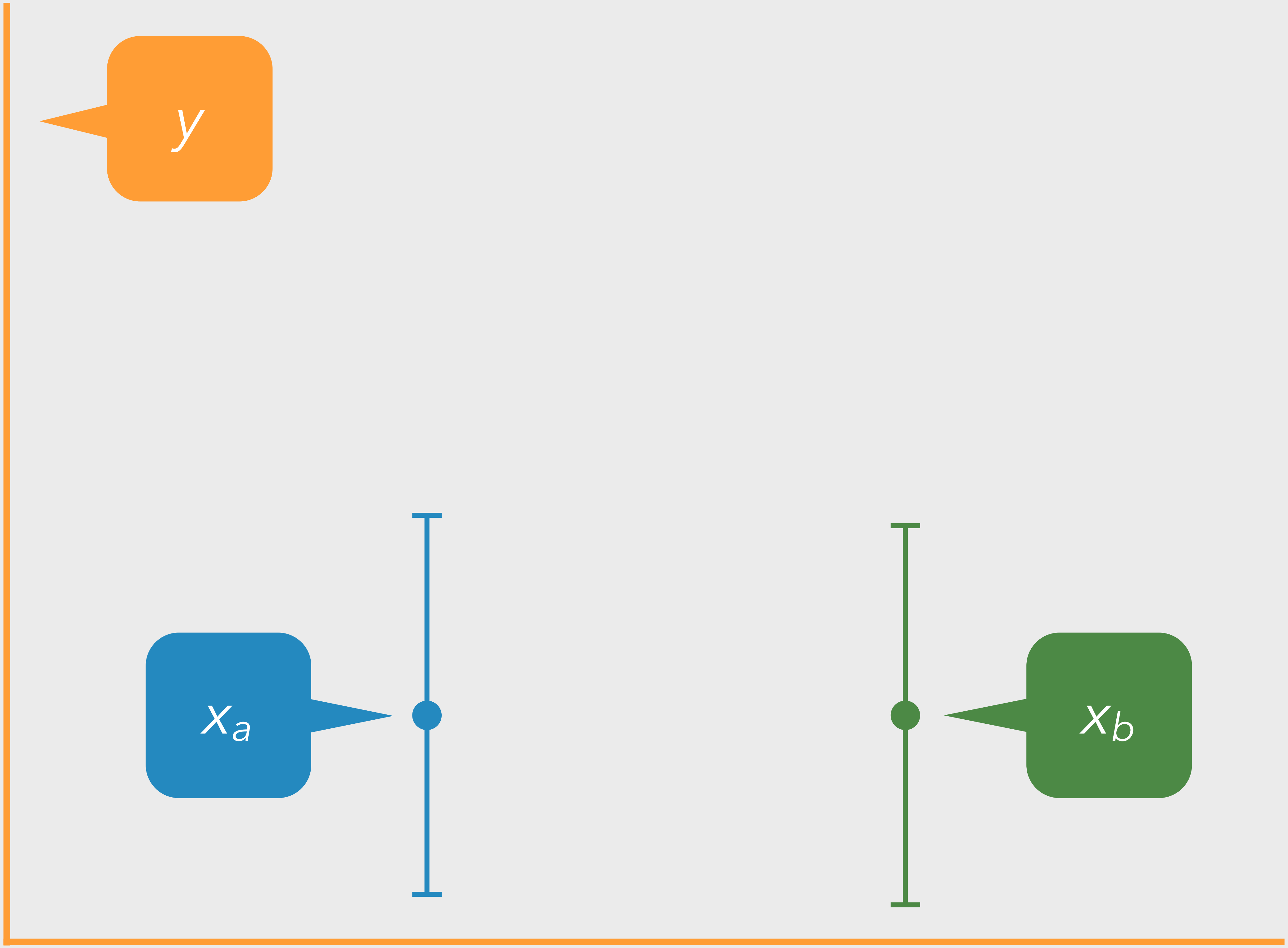
```
display (1-ttail(df,-t))*2
```

```
. display (1-ttail(72,-3.6308))*2

.0005255
```

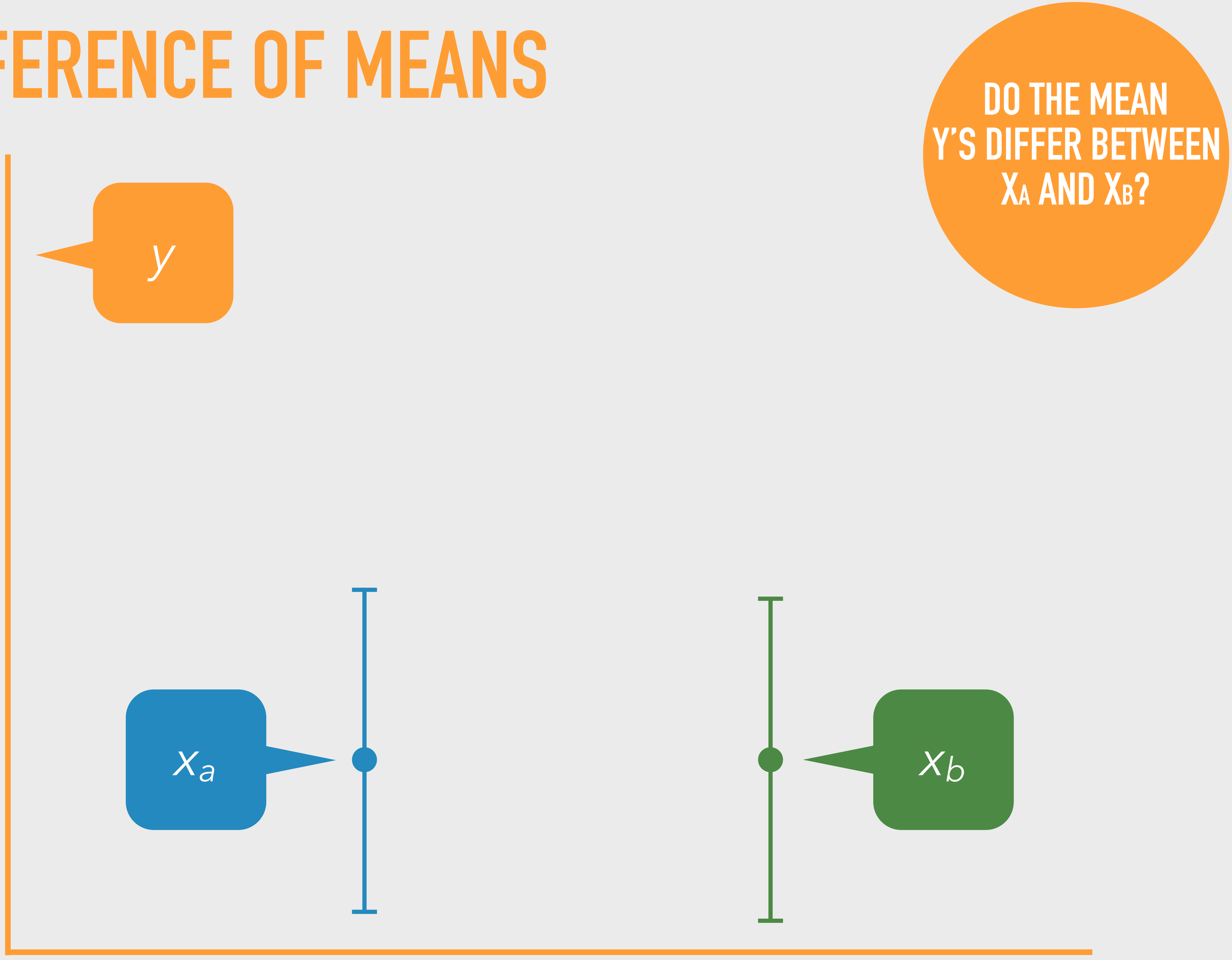
```
. display (1-ttail(72,-1.6308))*2

.1072996
```

4 INDEPENDENT SAMPLES

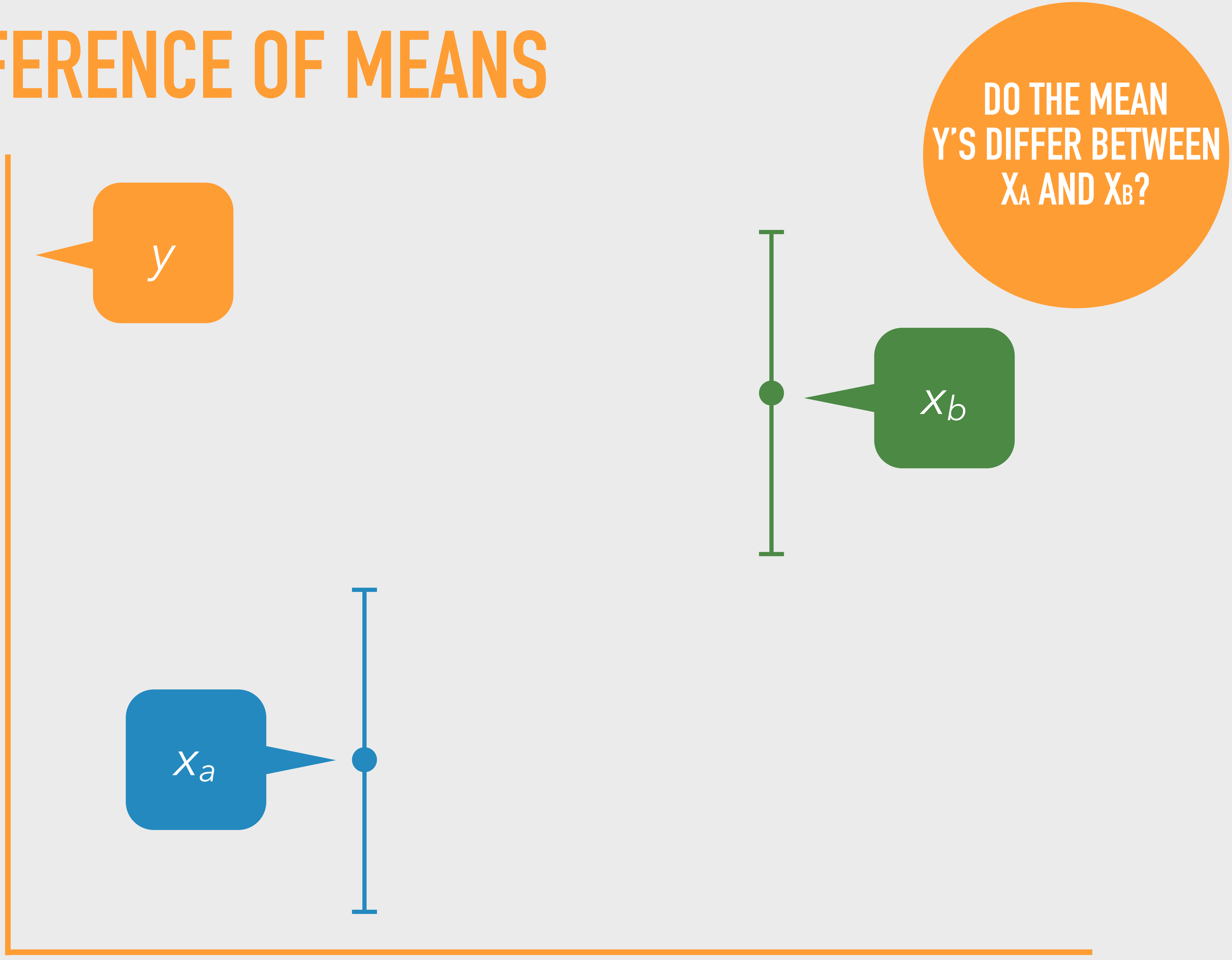
DIFFERENCE OF MEANS



DIFFERENCE OF MEANS

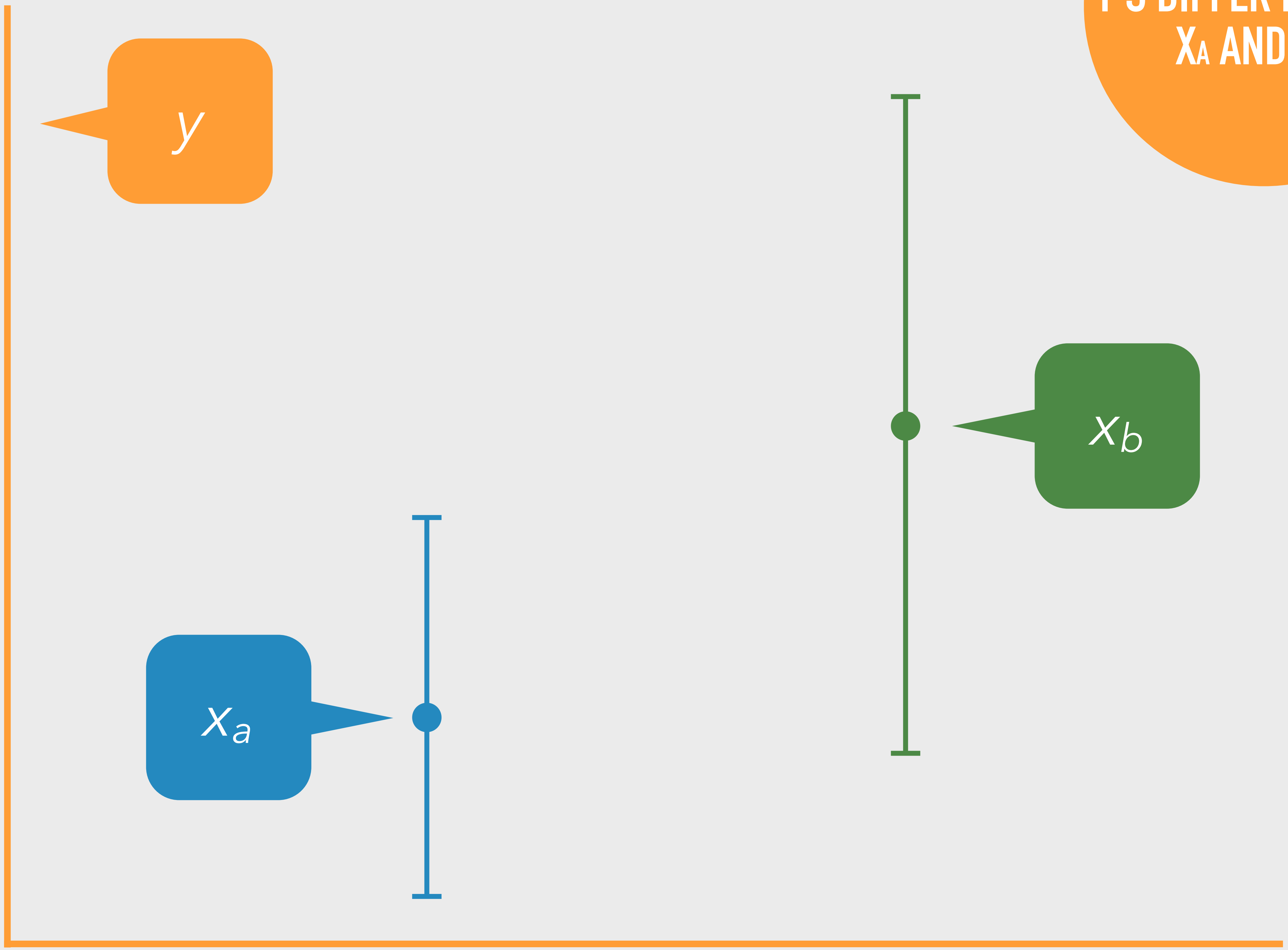


DIFFERENCE OF MEANS

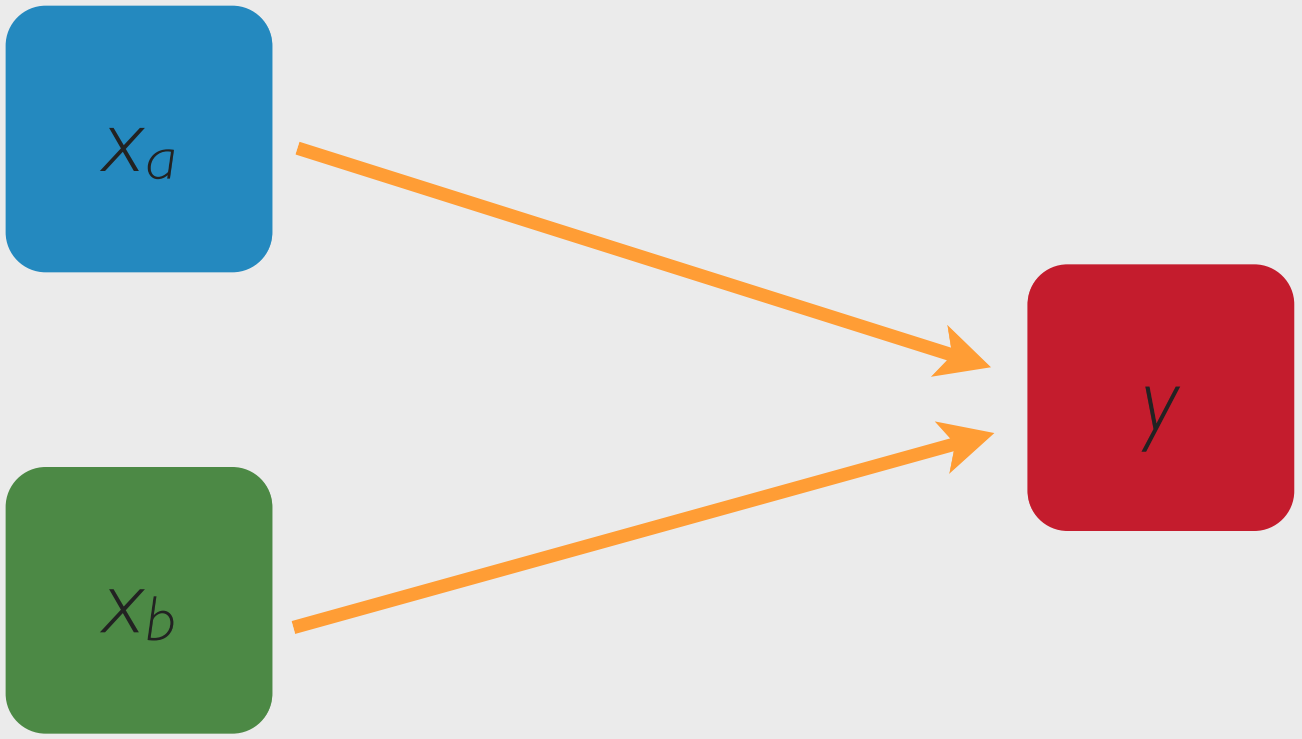


DIFFERENCE OF MEANS

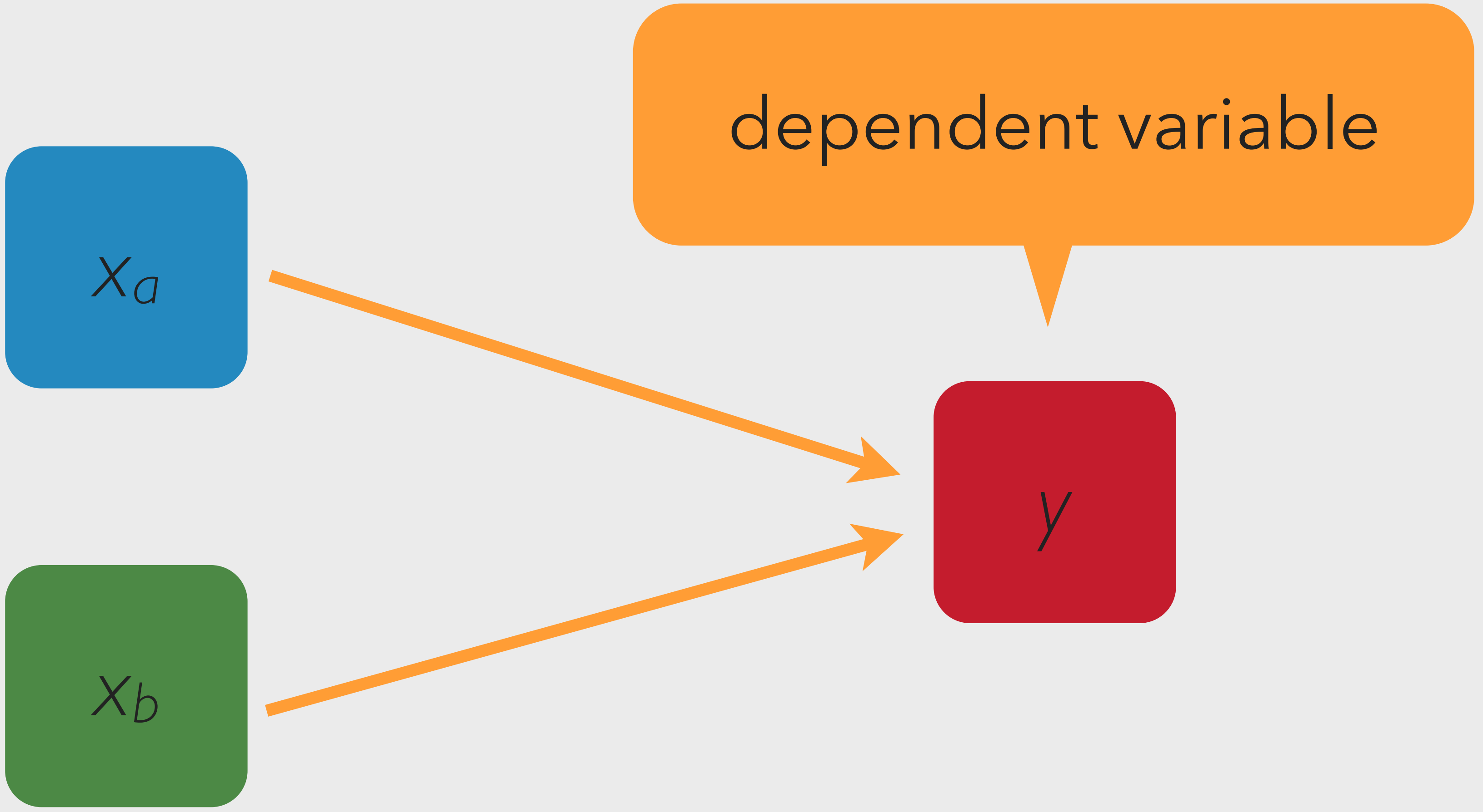
DO THE MEAN Y'S DIFFER BETWEEN X_A AND X_B?



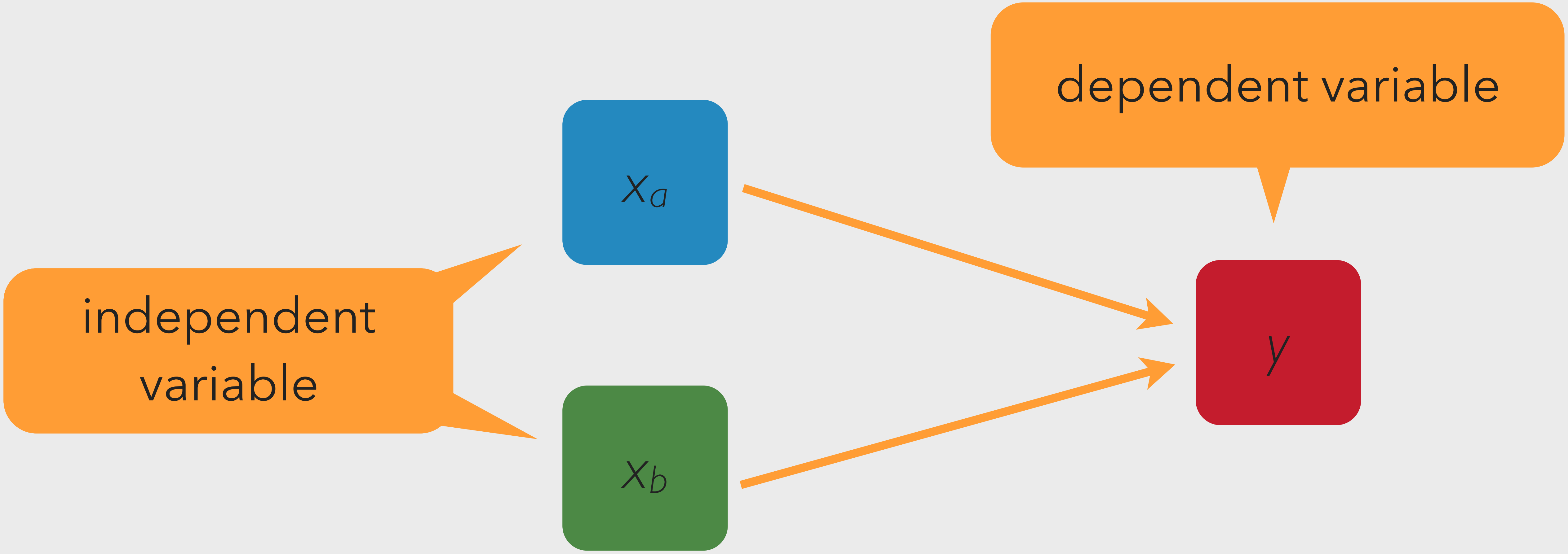
MODEL



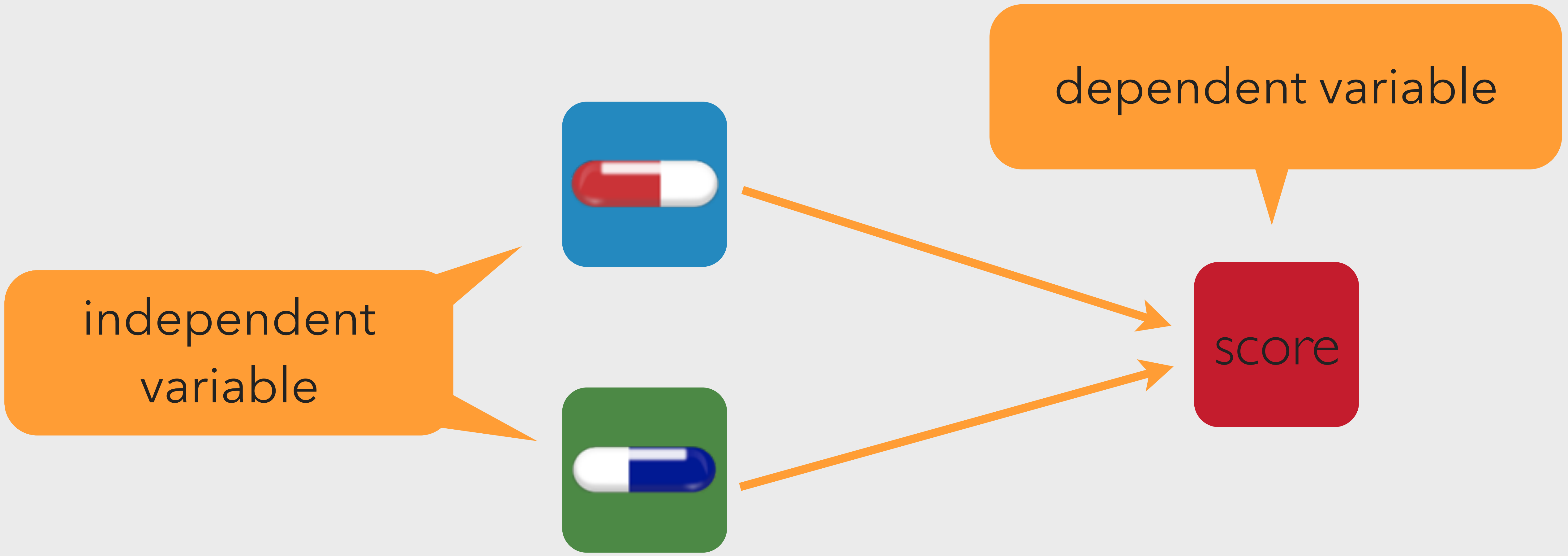
MODEL



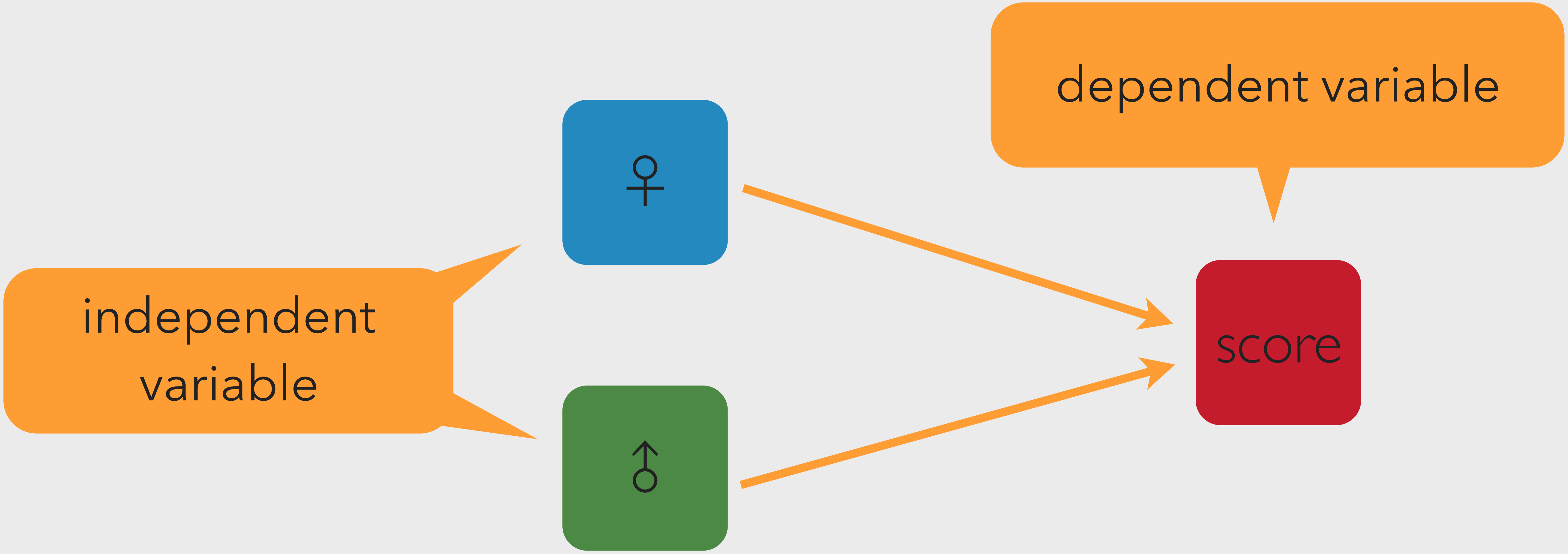
MODEL



MODEL



MODEL



HYPOTHESES

- H_0 = there is no difference in the mean of *y* between x_a and x_b

- H_1 = there is a difference in the mean of *y* between x_a and x_b

ASSUMPTIONS

- dependent variable (y) is continuous
- the distribution of y is approximately normal
- independent variable is binary (x_a and x_b)
- homogeneity of variance between x_a and x_b
- observations are independent
- $v = n_a + n_b - 2$

EQUATION ASSUMING HOMOGENEITY OF VARIANCE

$$t = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{s_p^2}{n_a} + \frac{s_p^2}{n_b}}}$$

EQUATION ASSUMING HOMOGENEITY OF VARIANCE

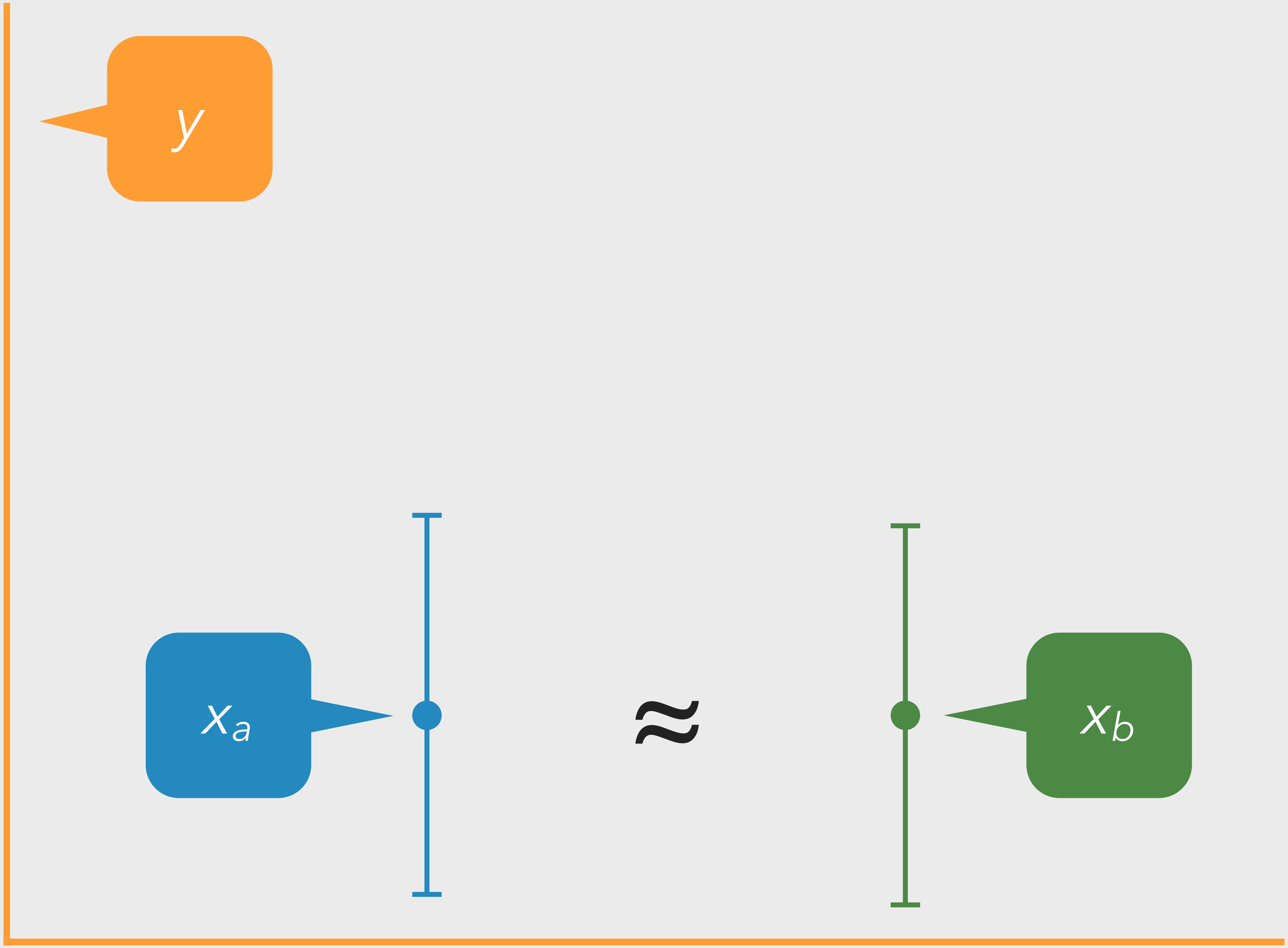
$$t = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{s_p^2}{n_a} + \frac{s_p^2}{n_b}}}$$

pooled variance

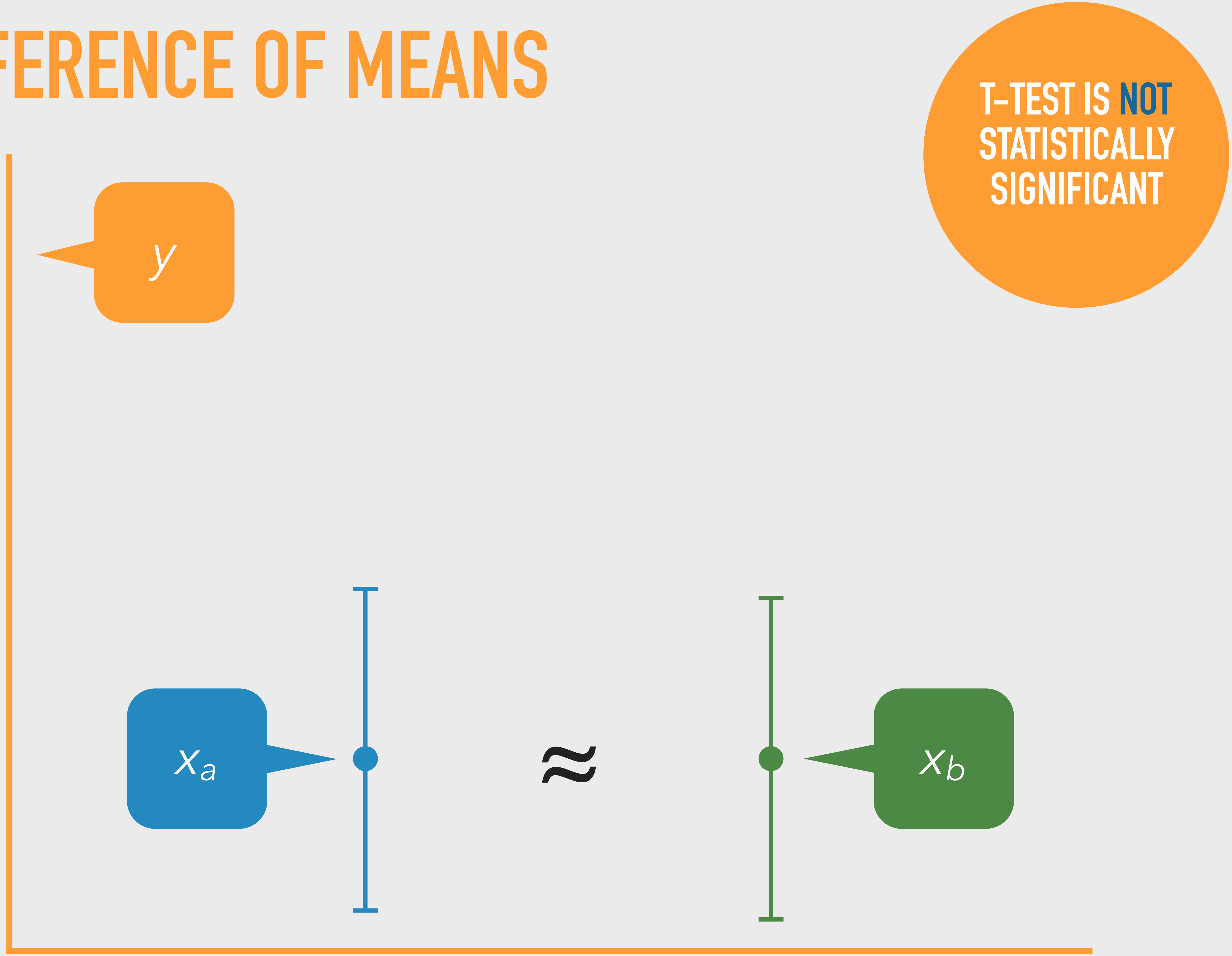
POOLED VARIANCE EQUATION

$$s_p^2 = \frac{(n_a - 1)\, s_a^2 + (n_b - 1)\, s_b^2}{n_a + n_b - 2}$$

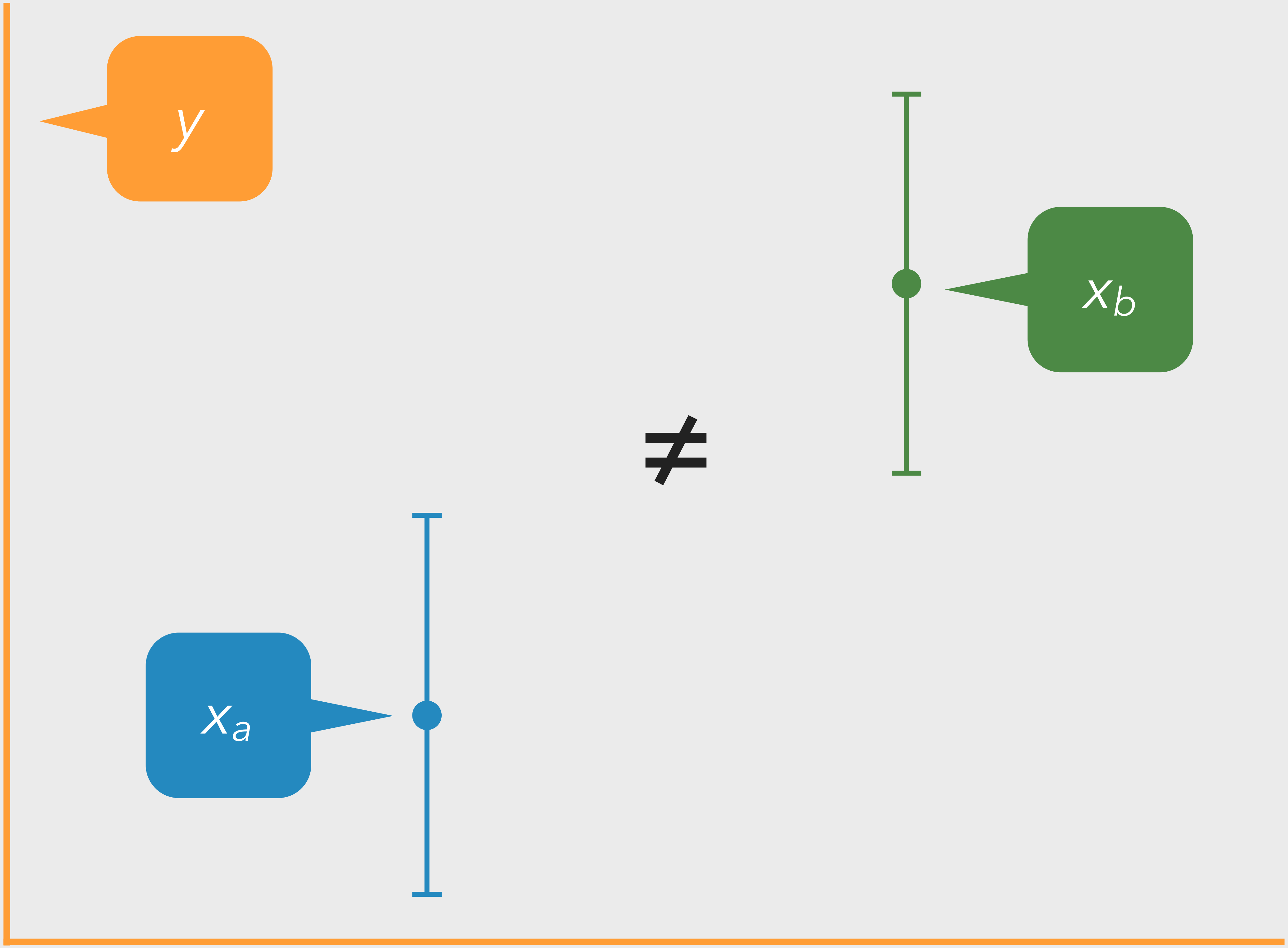
DIFFERENCE OF MEANS



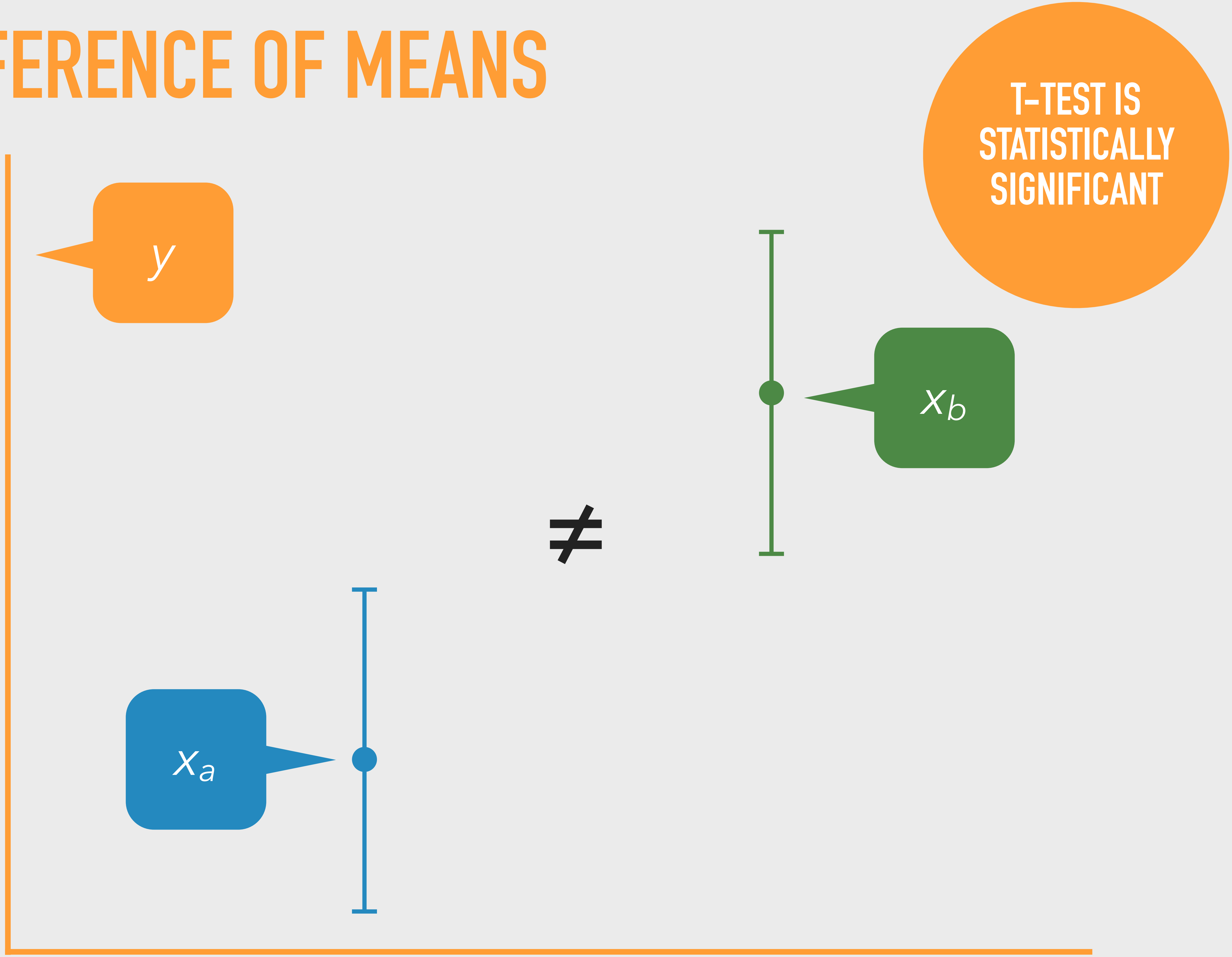
DIFFERENCE OF MEANS



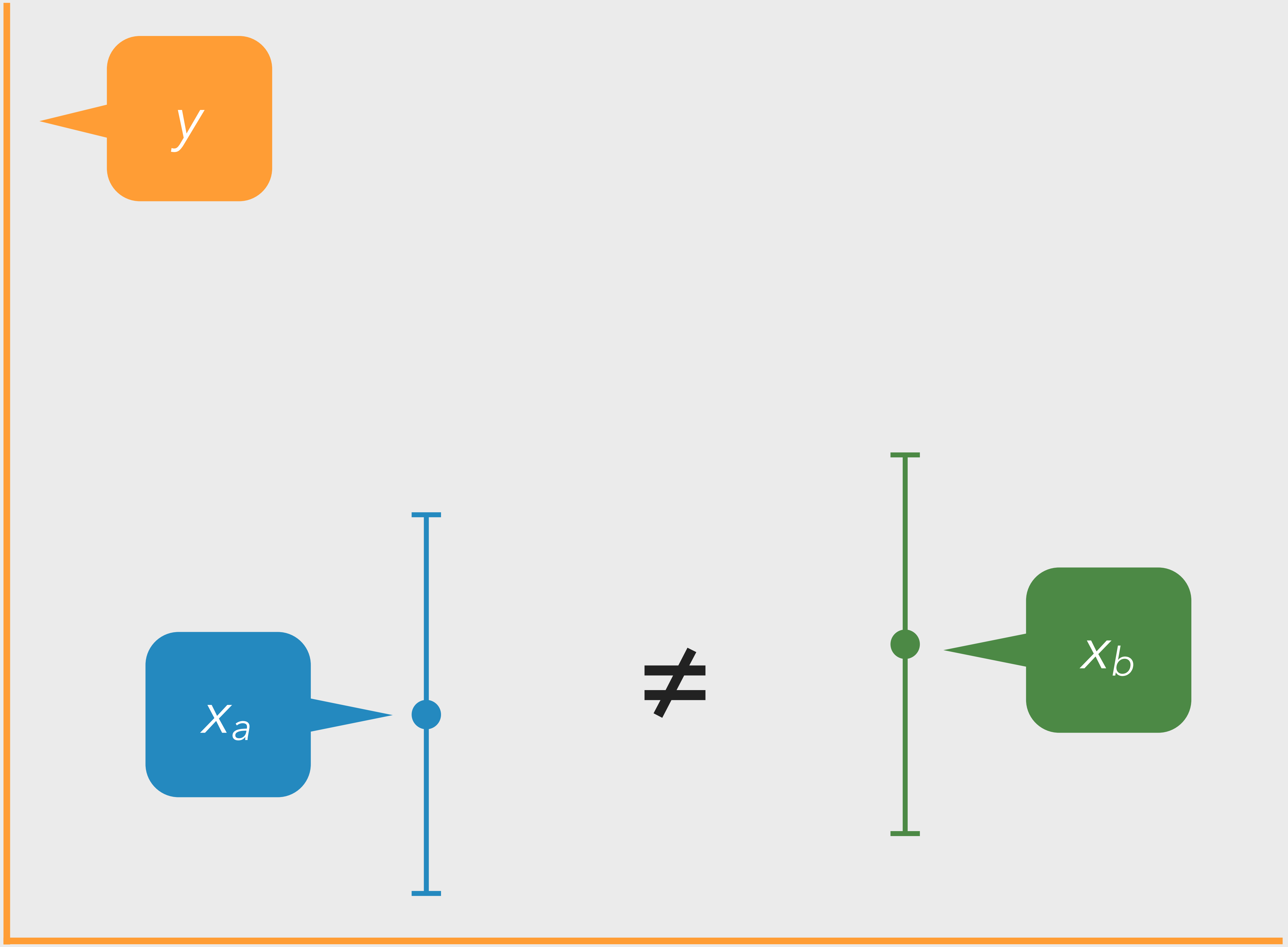
DIFFERENCE OF MEANS



DIFFERENCE OF MEANS

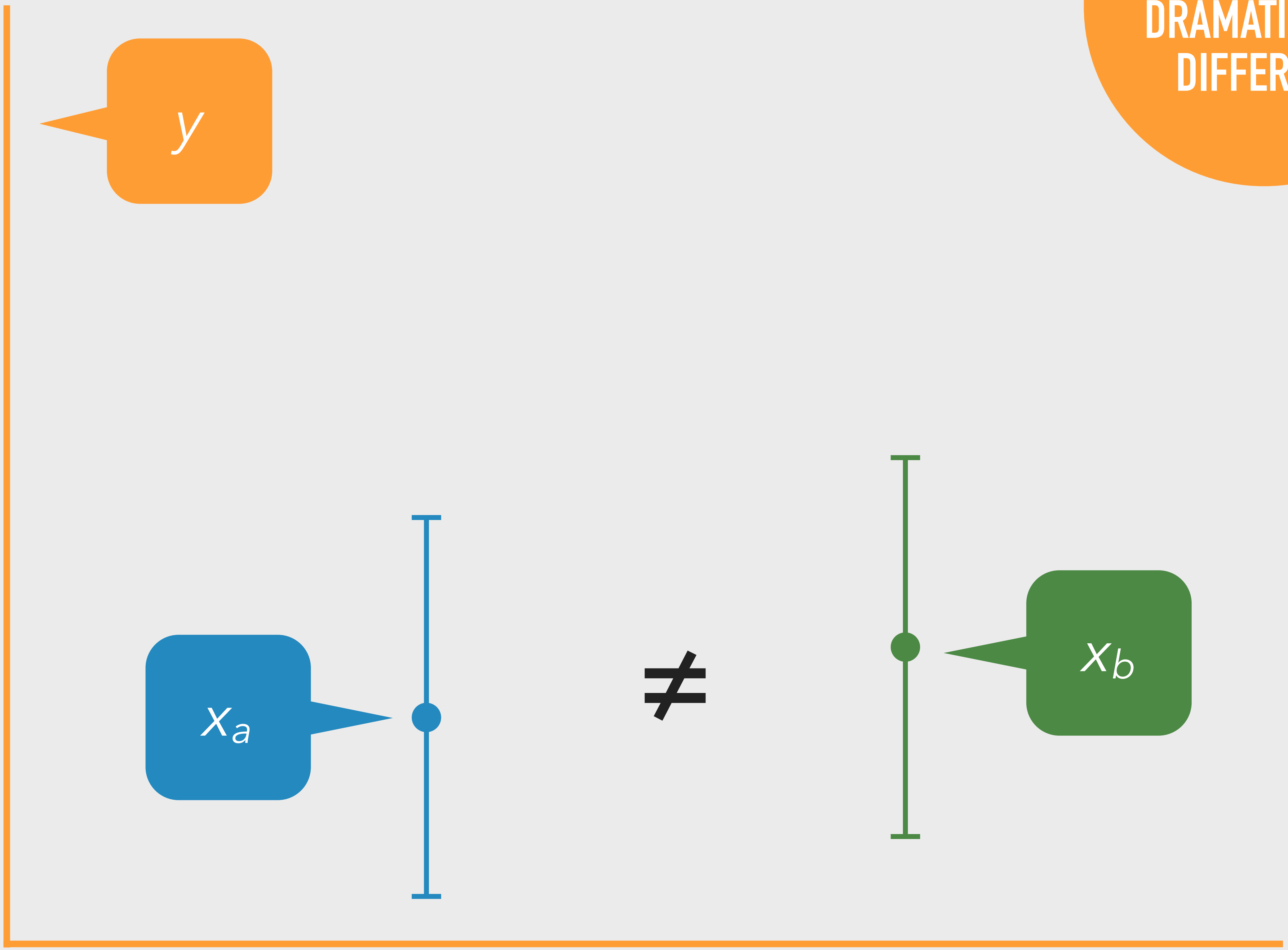


DIFFERENCE OF MEANS



DIFFERENCE OF MEANS

THE MEANS DO NOT HAVE TO BE DRAMATICALLY DIFFERENT



ASSUMPTIONS

- dependent variable (y) is continuous
- the distribution of y is approximately normal
- independent variable is binary (x_a and x_b)
- homogeneity of variance between x_a and x_b
- observations are independent
- $v = n_a + n_b - 2$

EQUATION IF HOMOGENEITY OF VARIANCE CANNOT BE ASSUMED

$$t = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

EQUATION IF HOMOGENEITY OF VARIANCE CANNOT BE ASSUMED

$$t = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

variance values for each subgroup

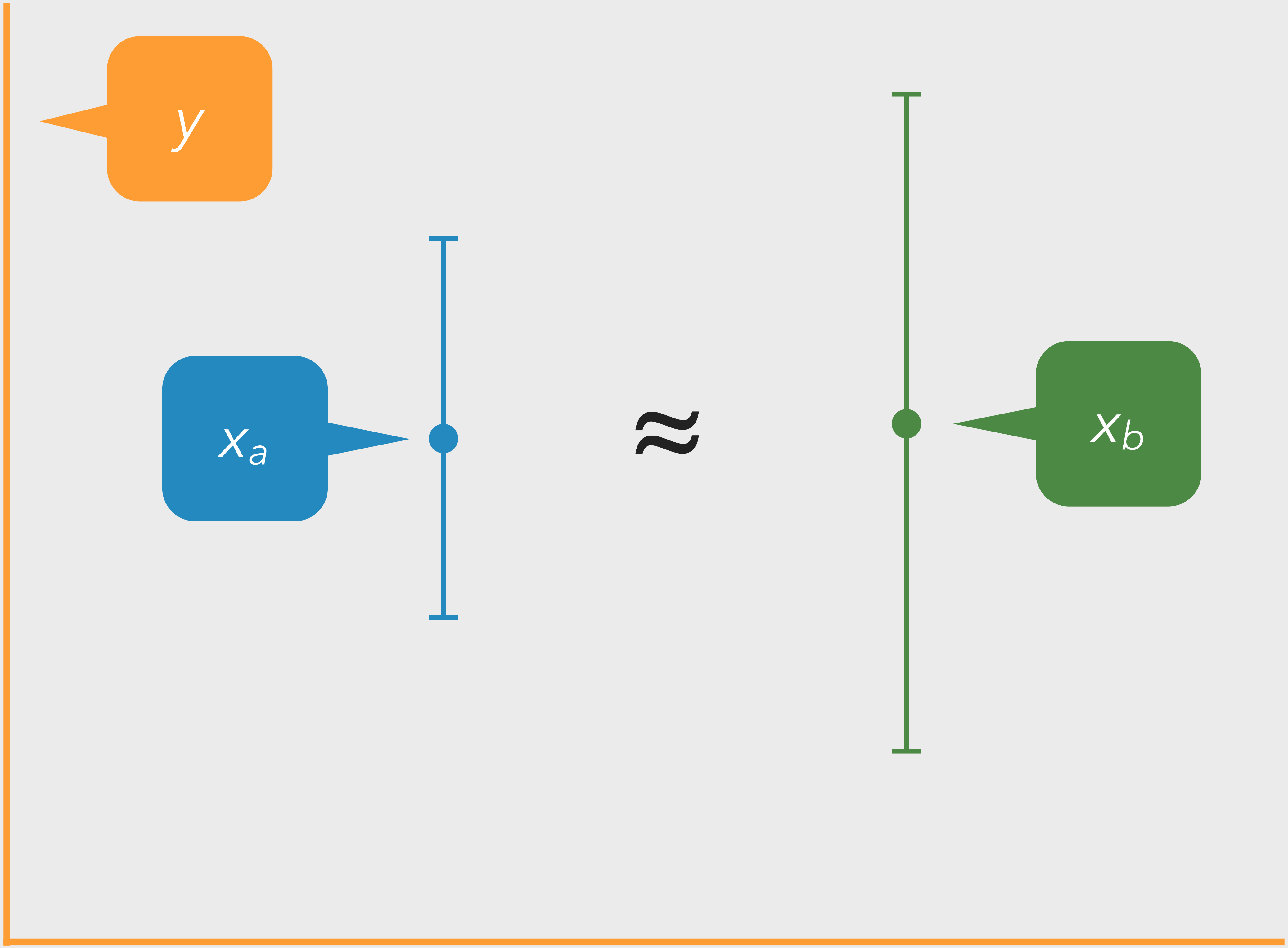
CAUTION! CAUTION! CAUTION!

$$t = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{s_p^2}{n_a} + \frac{s_p^2}{n_b}}} \neq t = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

WELCH'S CORRECTED DEGREES OF FREEDOM

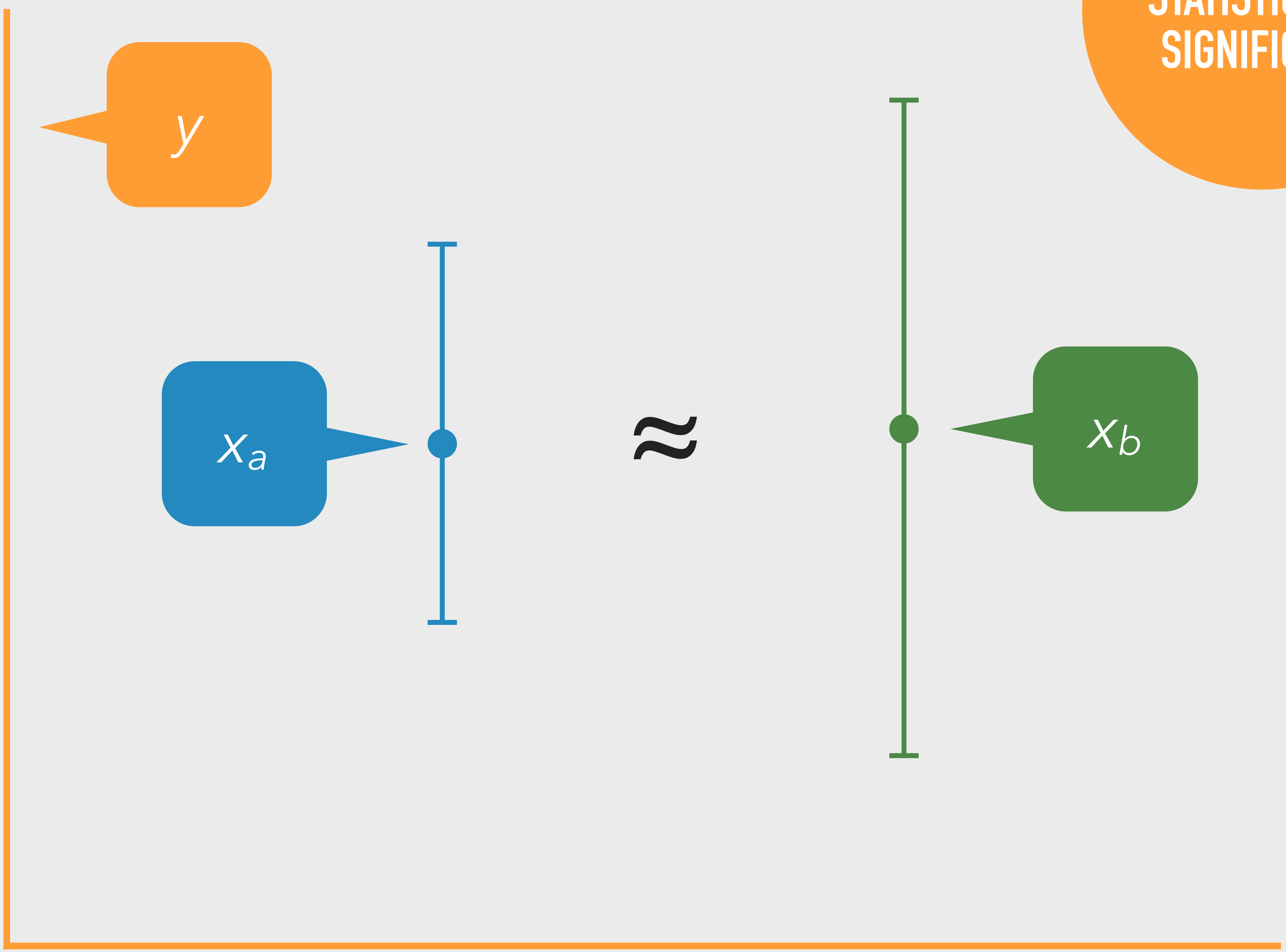
$$v \approx \frac{\left(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}\right)^2}{\frac{s_a^4}{(n_a^2)(n_a - 1)} + \frac{s_b^4}{\left(n_b^2\right)(n_b - 1)}}$$

DIFFERENCE OF MEANS

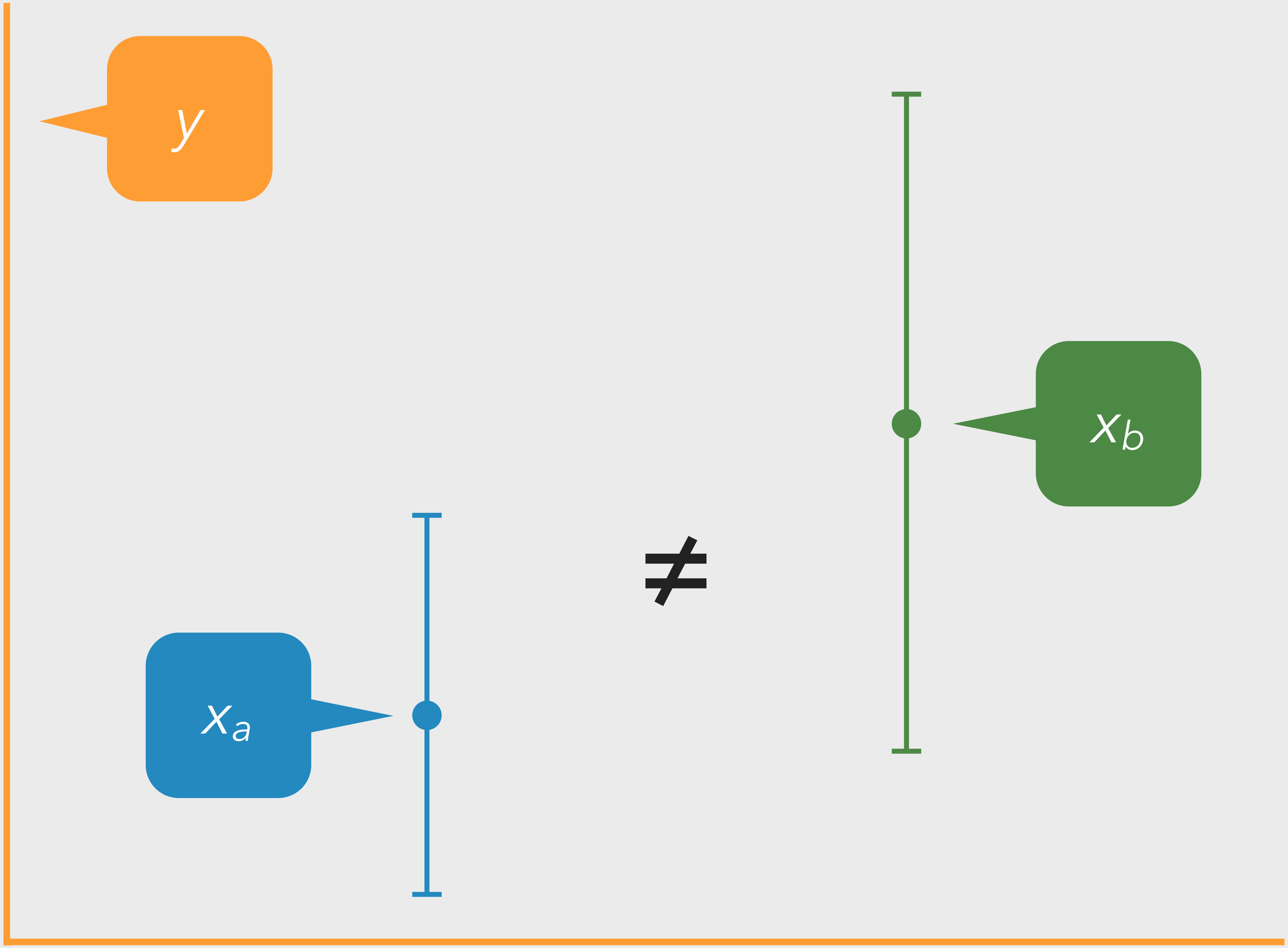


DIFFERENCE OF MEANS

T-TEST IS NOT STATISTICALLY SIGNIFICANT

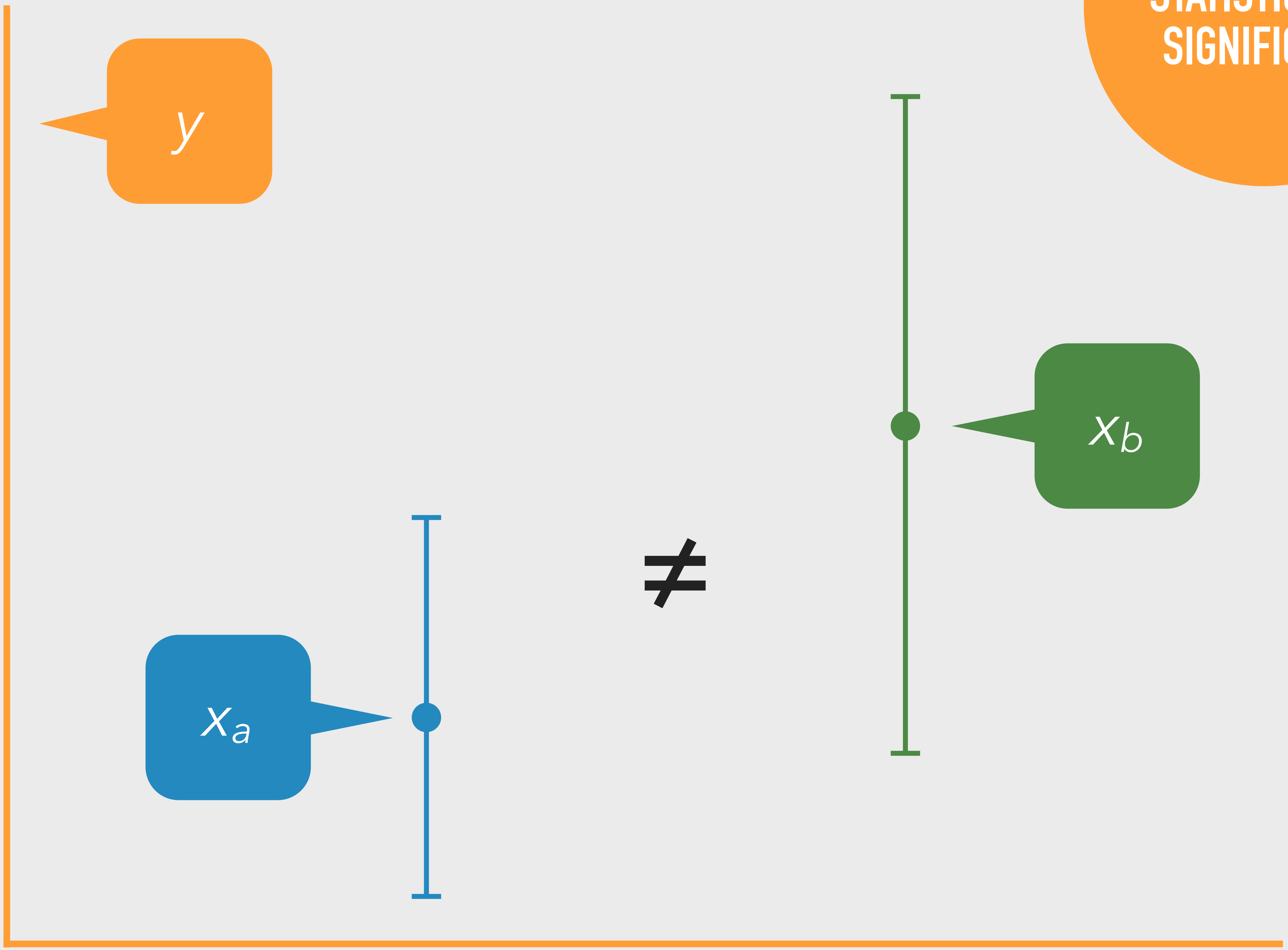


DIFFERENCE OF MEANS



DIFFERENCE OF MEANS

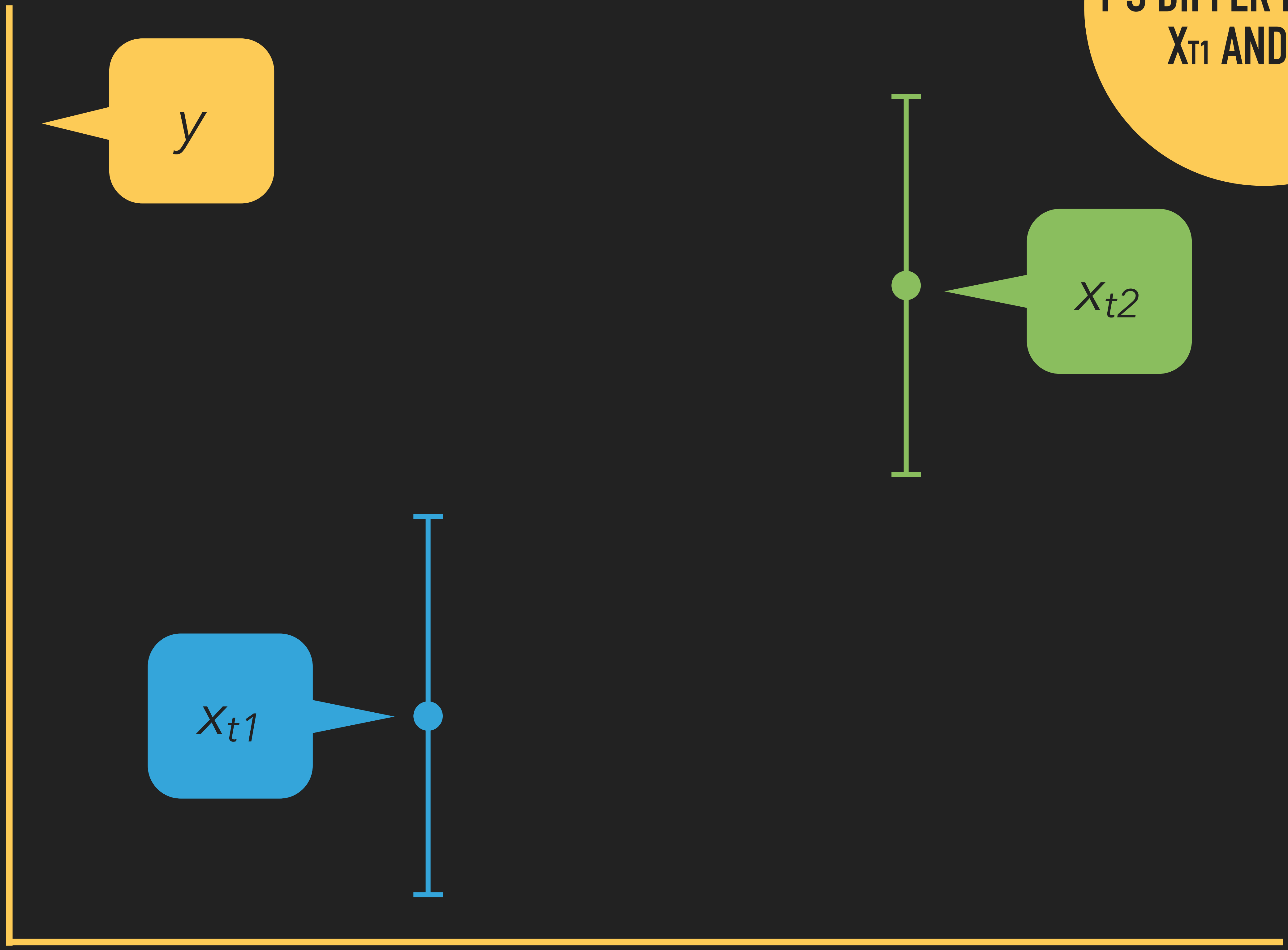
T-TEST IS STATISTICALLY SIGNIFICANT



5 DEPENDENT SAMPLES

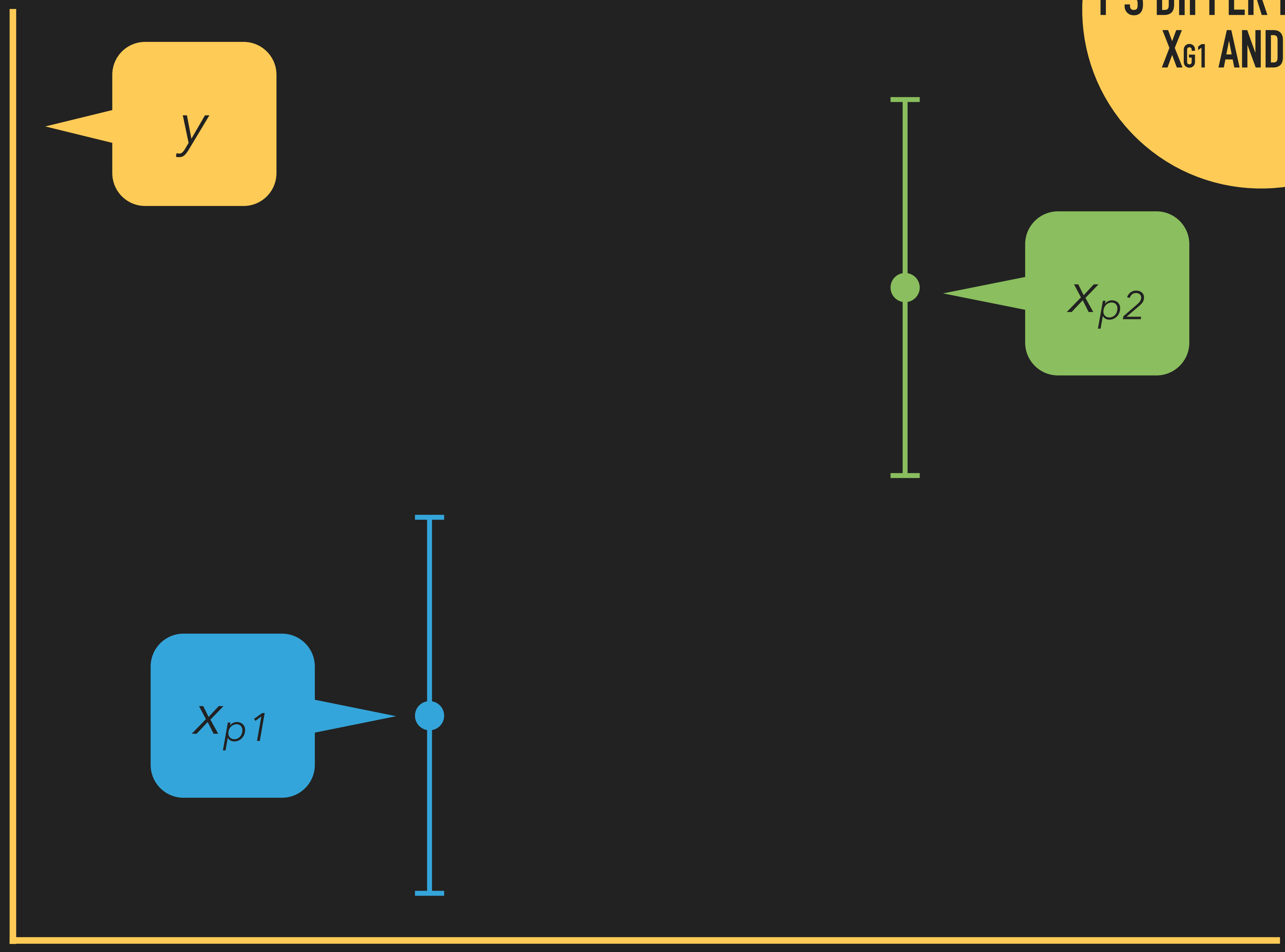
DIFFERENCE OF MEANS

DO THE MEAN Y'S DIFFER BETWEEN X_{T1} AND X_{T2}?

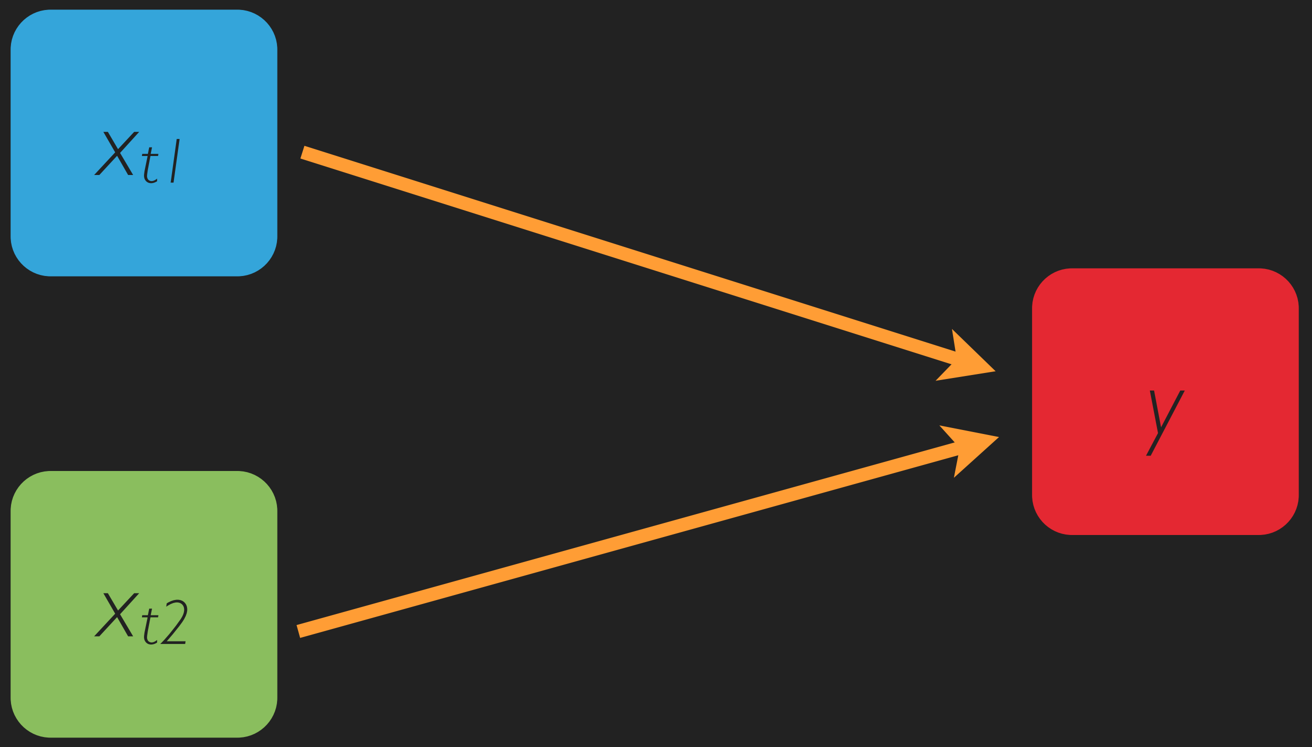


DIFFERENCE OF MEANS

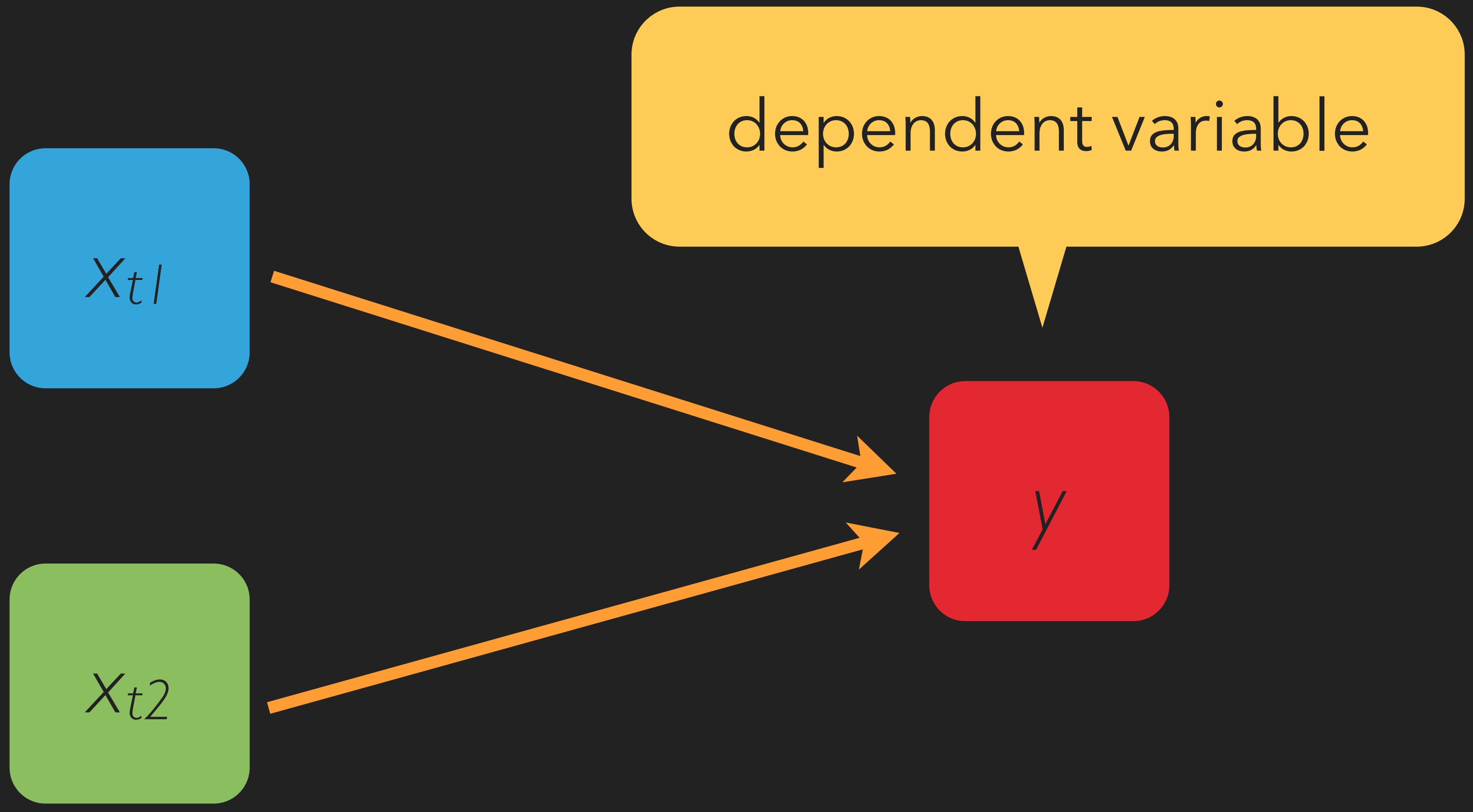
DO THE MEAN Y'S DIFFER BETWEEN X_{G1} AND X_{G2}?



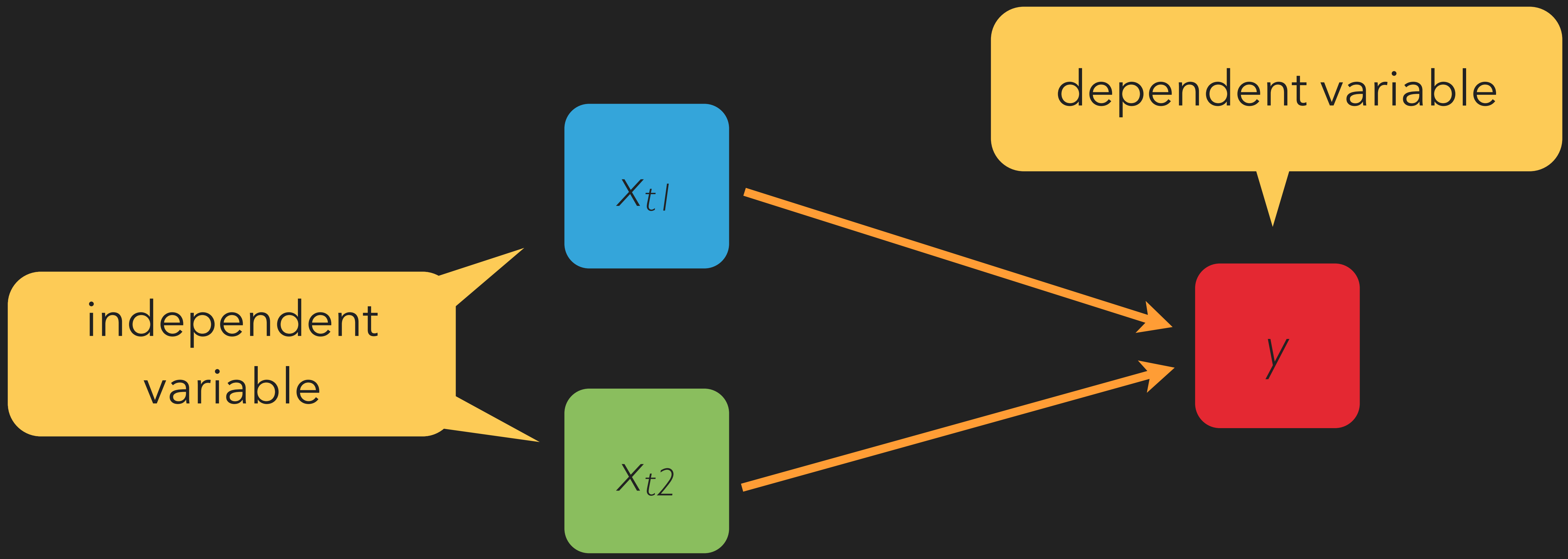
MODEL



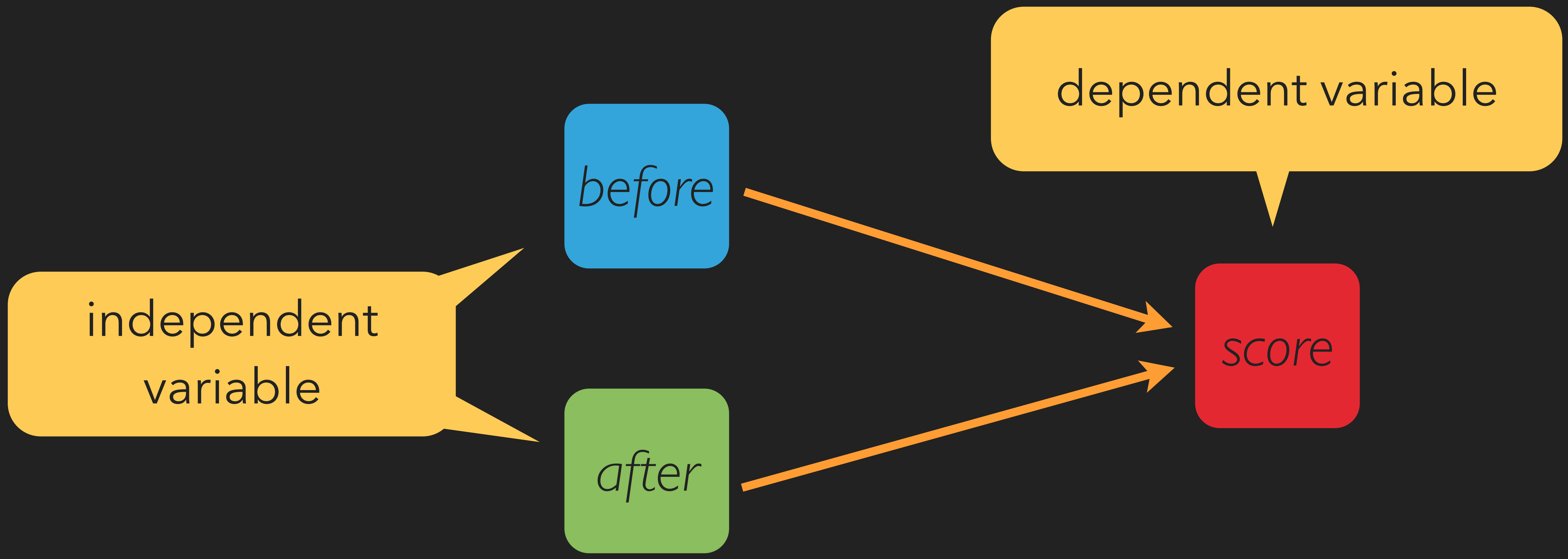
MODEL



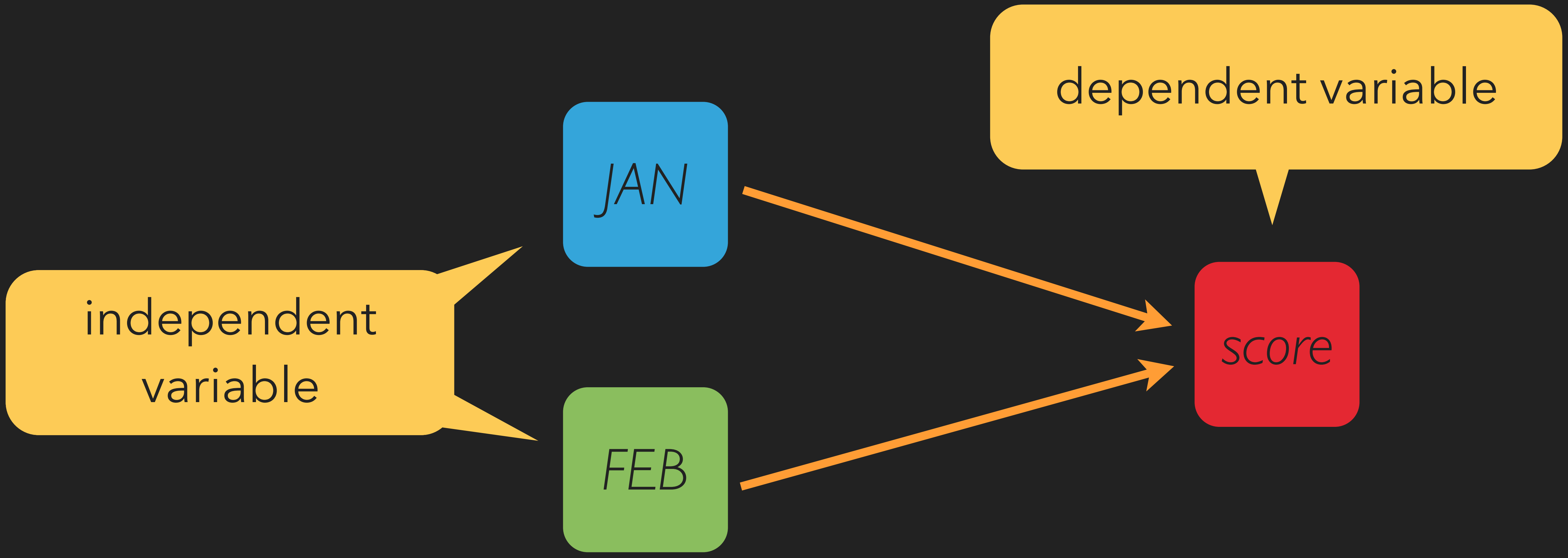
MODEL



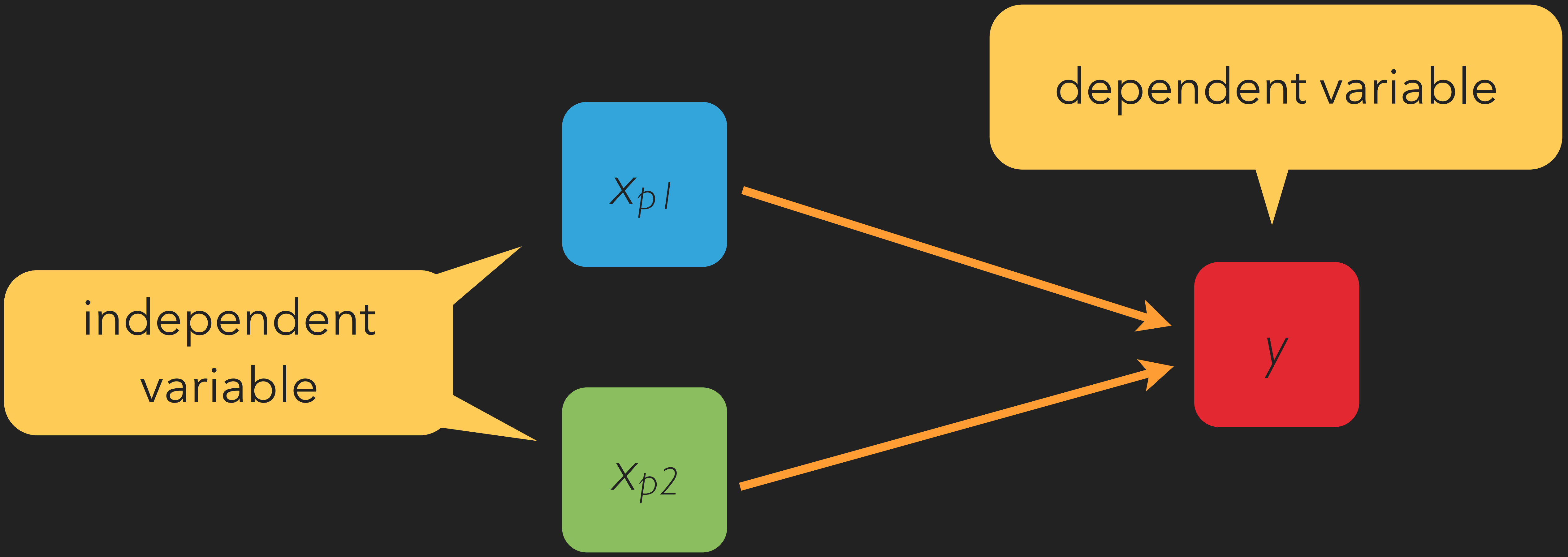
MODEL


dependent variable
before
independent variable
after
score

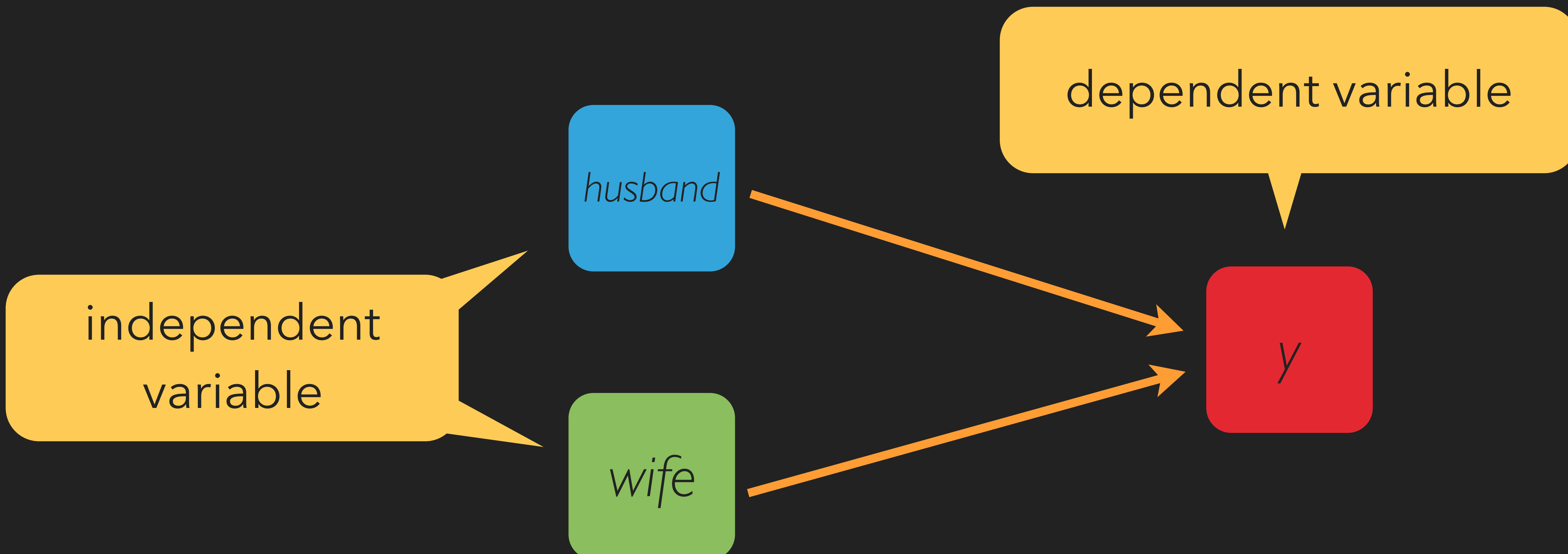
MODEL



MODEL



MODEL



HYPOTHESES

- H_0 = there is no difference in the mean of *y* between x_{t1} and x_{t2}
- H_1 = there is a difference in the mean of *y* between x_{t1} and x_{t2}

HYPOTHESES

- H_0 = there is no difference in the mean of *y* between x_{g1} and x_{g2}
- H_1 = there is a difference in the mean of *y* between x_{g1} and x_{g2}

ASSUMPTIONS

- dependent variable (*y*) is continuous
- independent variable is binary (x_{g1} and x_{g2})
- homogeneity of variance between x_{g1} and x_{g2}
- the distribution of the differences between x_{g1} and x_{g2} is normally distributed
- scores are dependent

EQUATION

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}}$$

EQUATION

mean of difference between groups

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}}$$

EQUATION

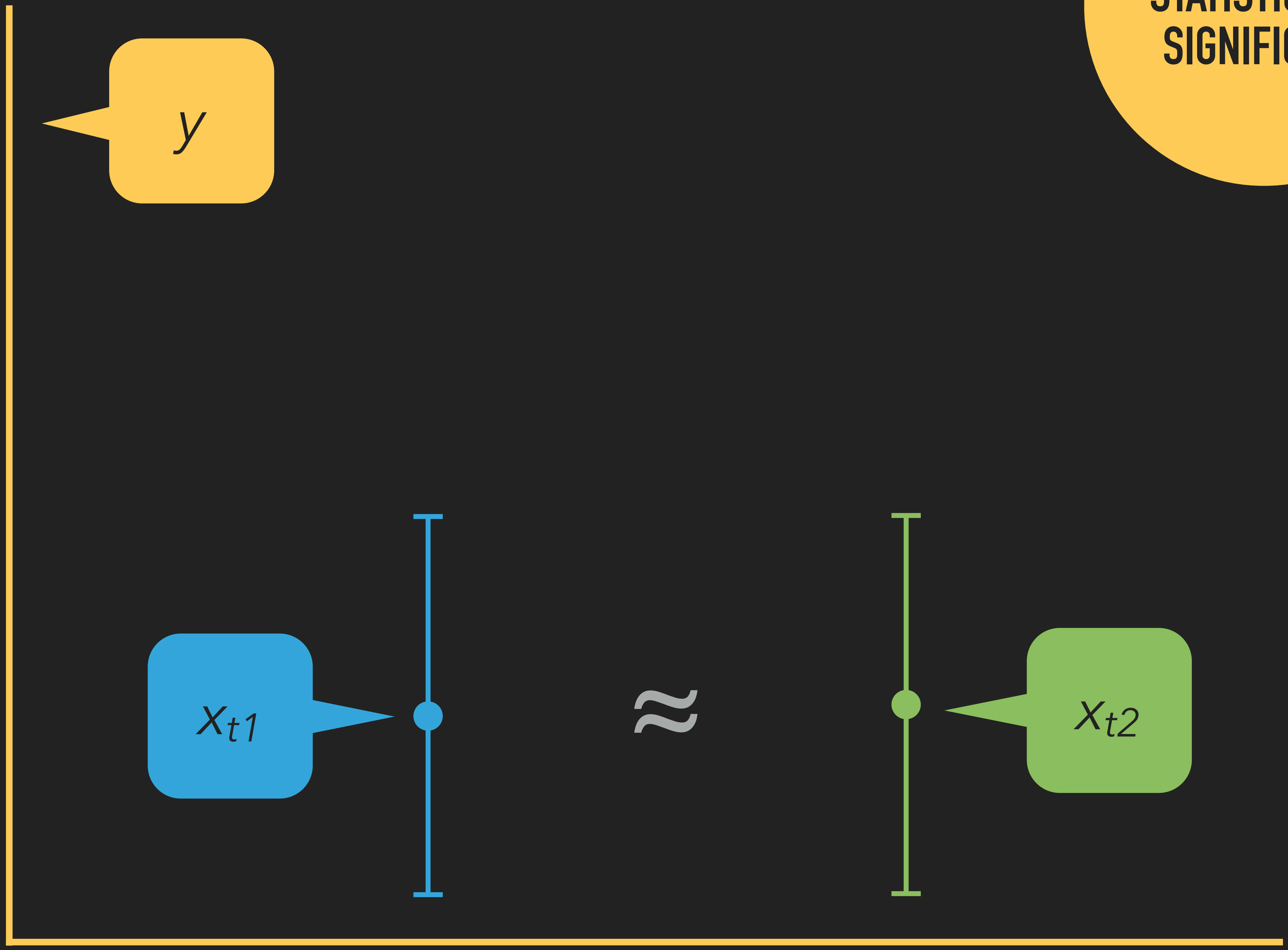
mean of difference between groups

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}}$$

variance of difference between groups

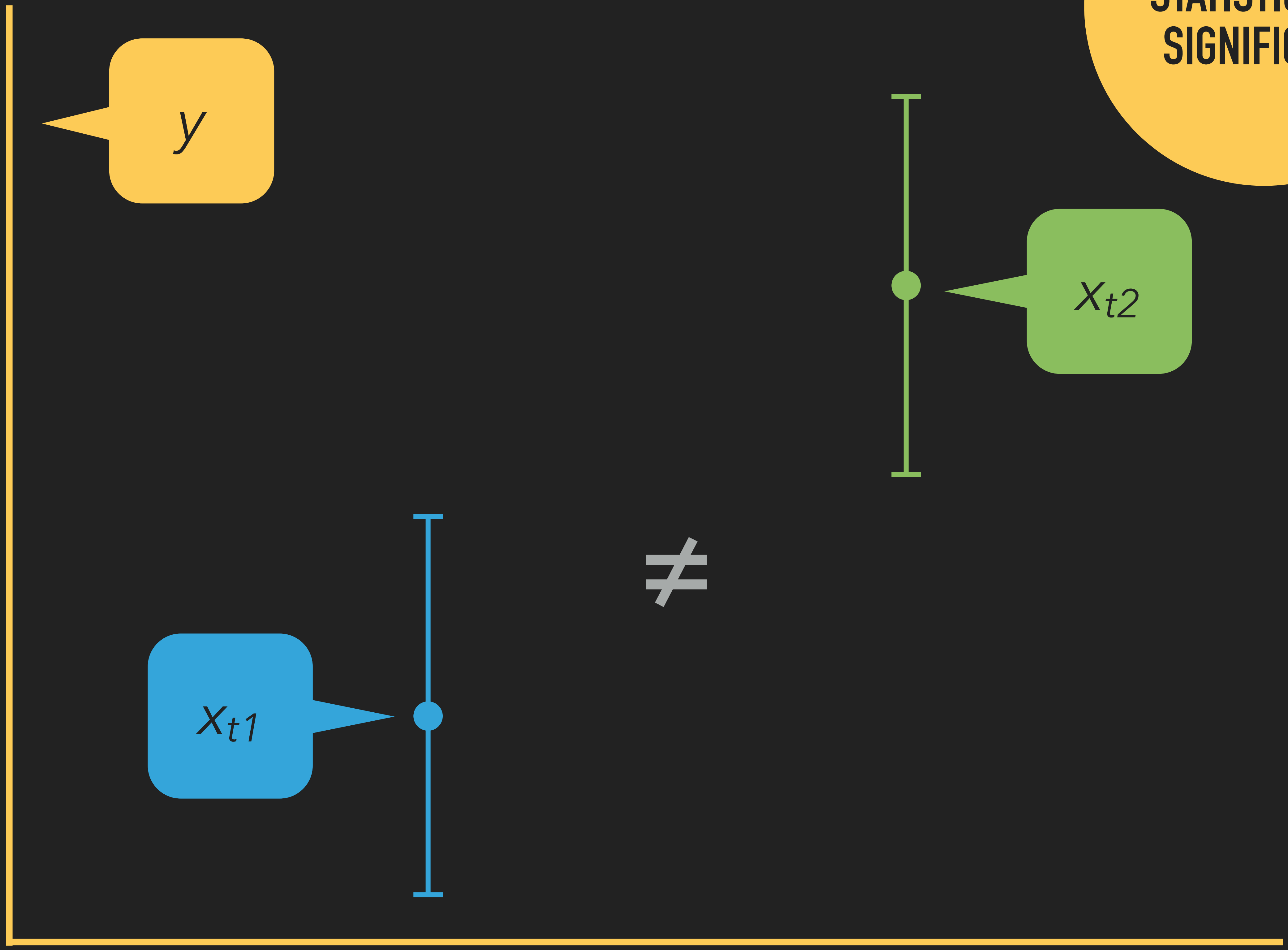
DIFFERENCE OF MEANS

T-TEST IS NOT STATISTICALLY SIGNIFICANT



DIFFERENCE OF MEANS

T-TEST IS STATISTICALLY SIGNIFICANT



LONG DATA

participant	*score*	*timePoint*
jane	10	before
jane	12	after
john	15	before
john	14	after

WIDE DATA

participant	*score1*	*score2*
jane	10	12
john	15	14
joe	12	12
jessica	8	11

RESHAPING DATA

participant	*score*	*timePoint*
jane	10	before
jane	12	after
john	15	before
john	14	after
joe	12	before
joe	12	after
jessica	8	before
jessica	11	after

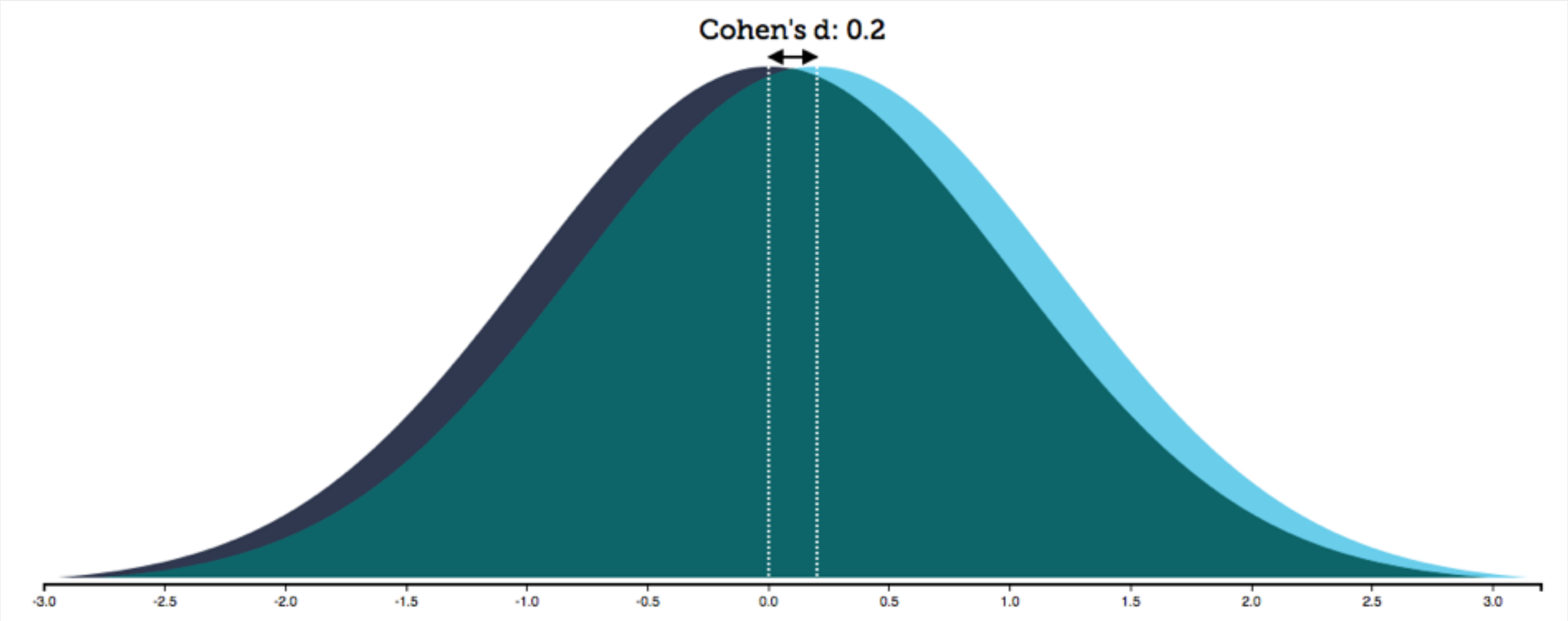
participant	*score1*	*score2*
jane	10	12
john	15	14
joe	12	12
jessica	8	11

6 EFFECT SIZES

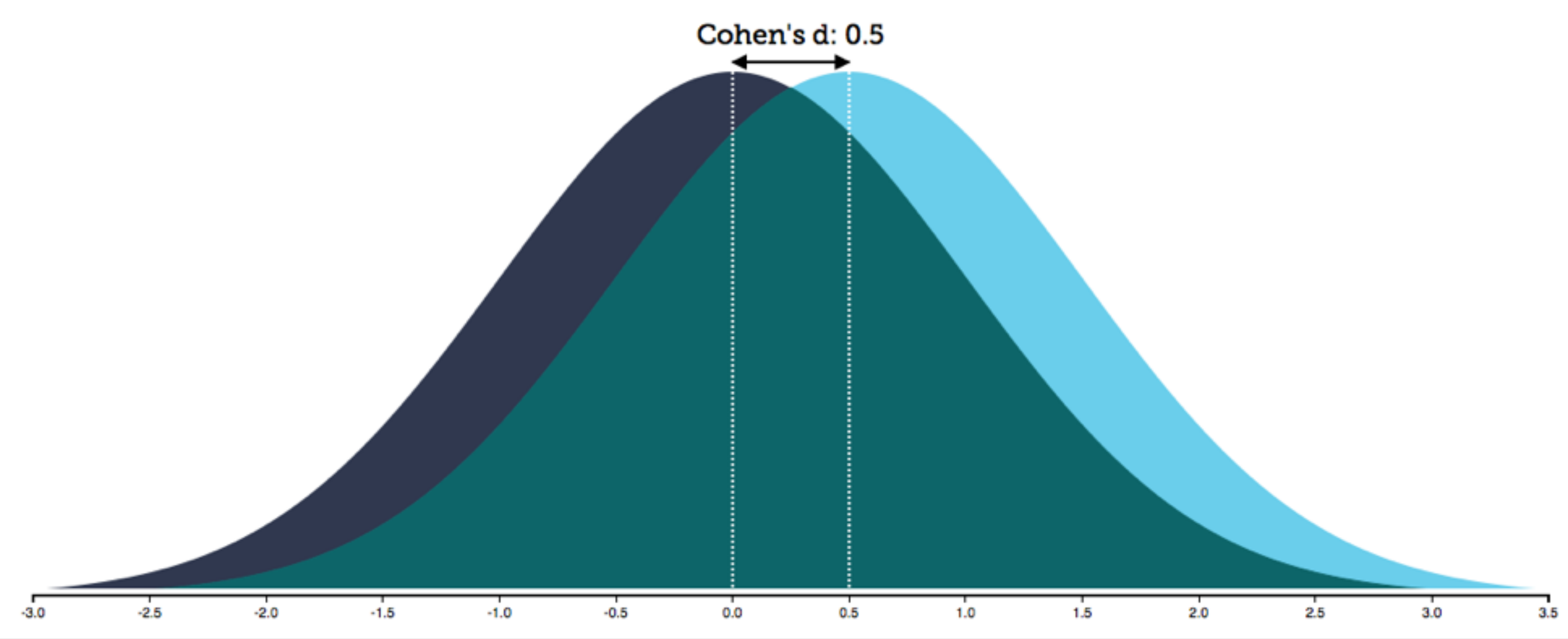
PROBLEM

Statistical Significance $\neq$ Real World Significance

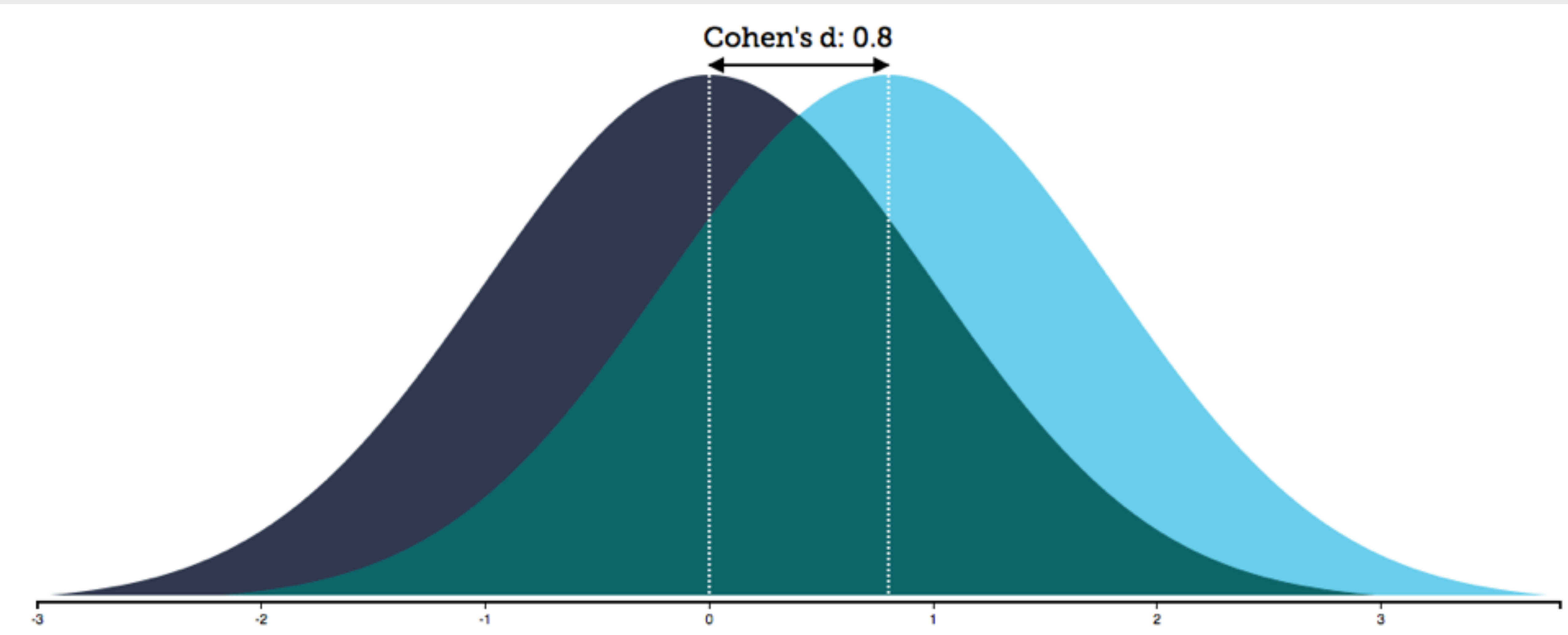
COHEN'S D INTERPRETATION



COHEN'S D INTERPRETATION



COHEN'S D INTERPRETATION



COHEN'S D EQUATION

$$d = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{n_a + n_b - 2}}}$$

COHEN'S D EQUATION

$$d = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{n_a + n_b - 2}}}$$

pooled variance

COHEN'S D EQUATION

$$d = \frac{M_t - M_c}{\sqrt{\frac{(n_t - 1)s_t^2 + (n_c - 1)s_c^2}{n_t + n_c - 2}}}$$

COHEN'S D EQUATION SIMPLIFIED

$n_a = n_b$

$$d = \frac{2t}{\sqrt{v}}$$

COHEN'S D EQUATION SIMPLIFIED

$n_a = n_b$

$$d = \frac{2t}{\sqrt{v}}$$

$n_a \neq n_b$

$$d = \frac{t\left(n_t + n_c\right)}{\sqrt{v}\left(\sqrt{n_t + n_c}\right)}$$

DOCUMENT DETAILS

Document produced by Christopher Prener, Ph.D for the Saint Louis University course SOC 5050: QUANTITATIVE ANALYSIS - APPLIED INFERENTIAL STATISTICS. See the course wiki and the repository README.md file for additional details.