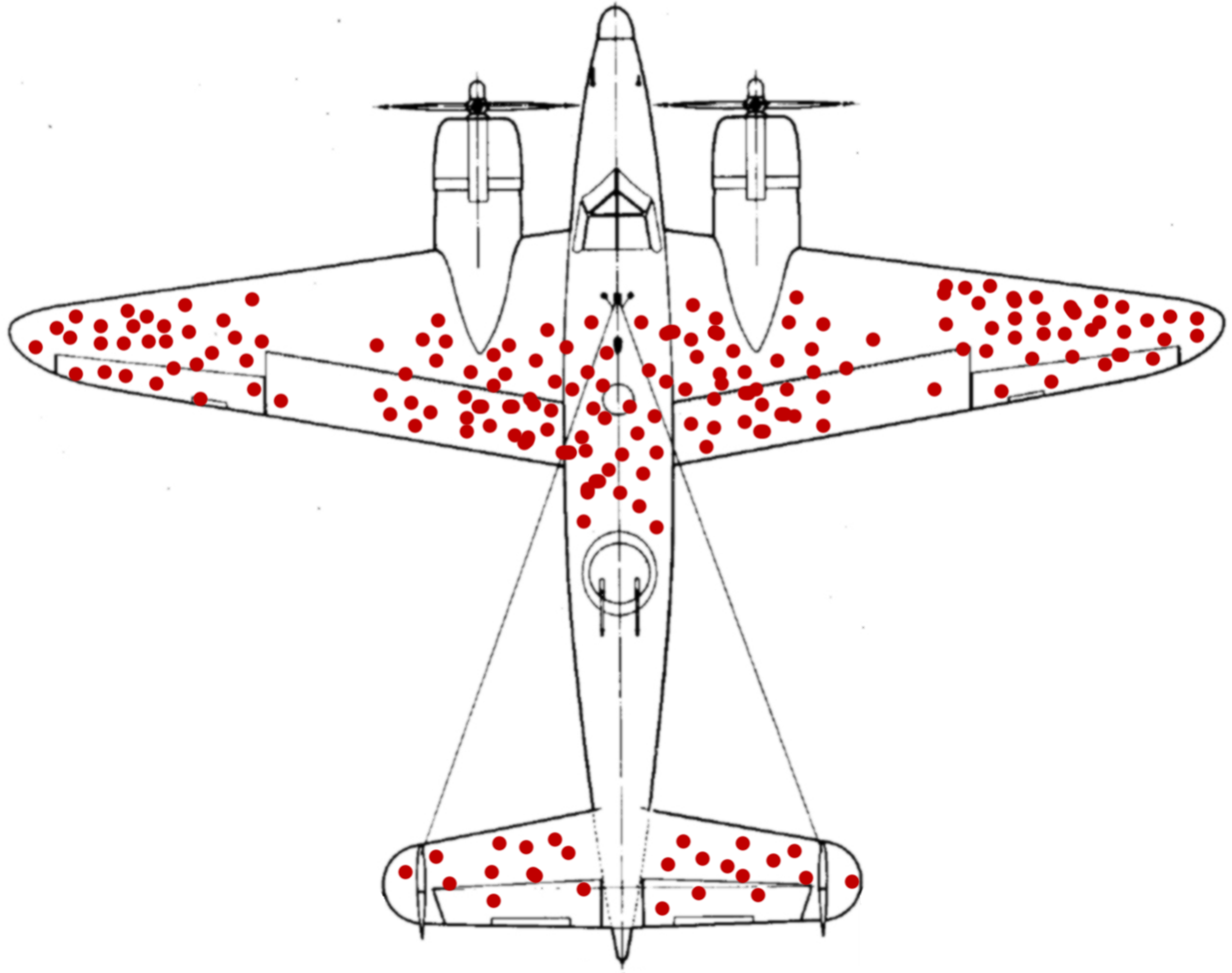
EDS 212 - Essential Math in Environmental Data Science

Day 5 Part 2 - Probability continued, intuition for hypothesis tests, and Boolean algebra

Think about your question: Abraham Wald's memo



Hypothesis testing: building intuition, continued

You'll learn about hypothesis testing in EDS 222. Let's just build a bit more intuition here.

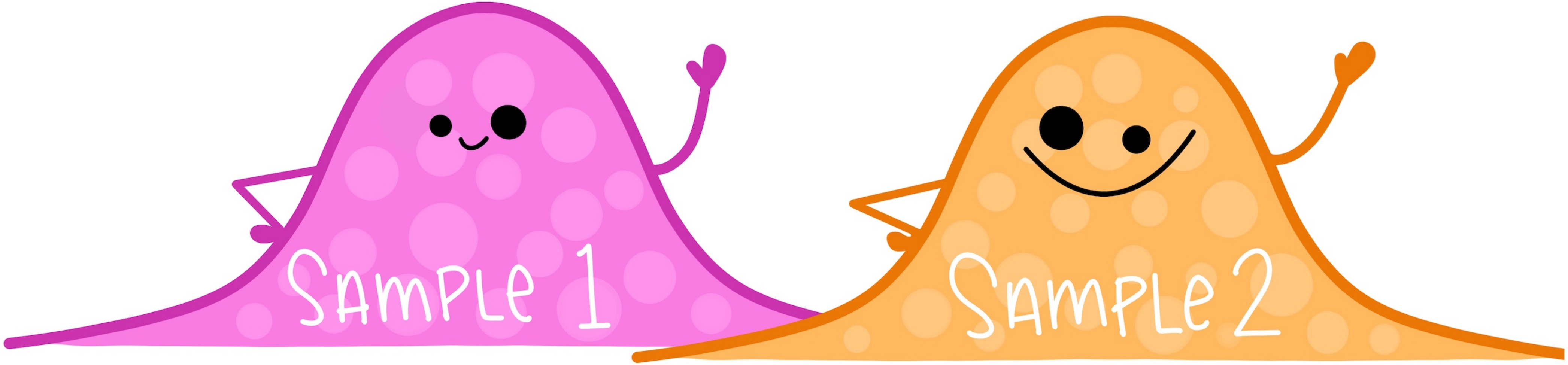
A common question: are means from two samples so different (considering data spread and sample size) that we think we have enough evidence to reject a null hypothesis that they were drawn from populations with the **same** mean?

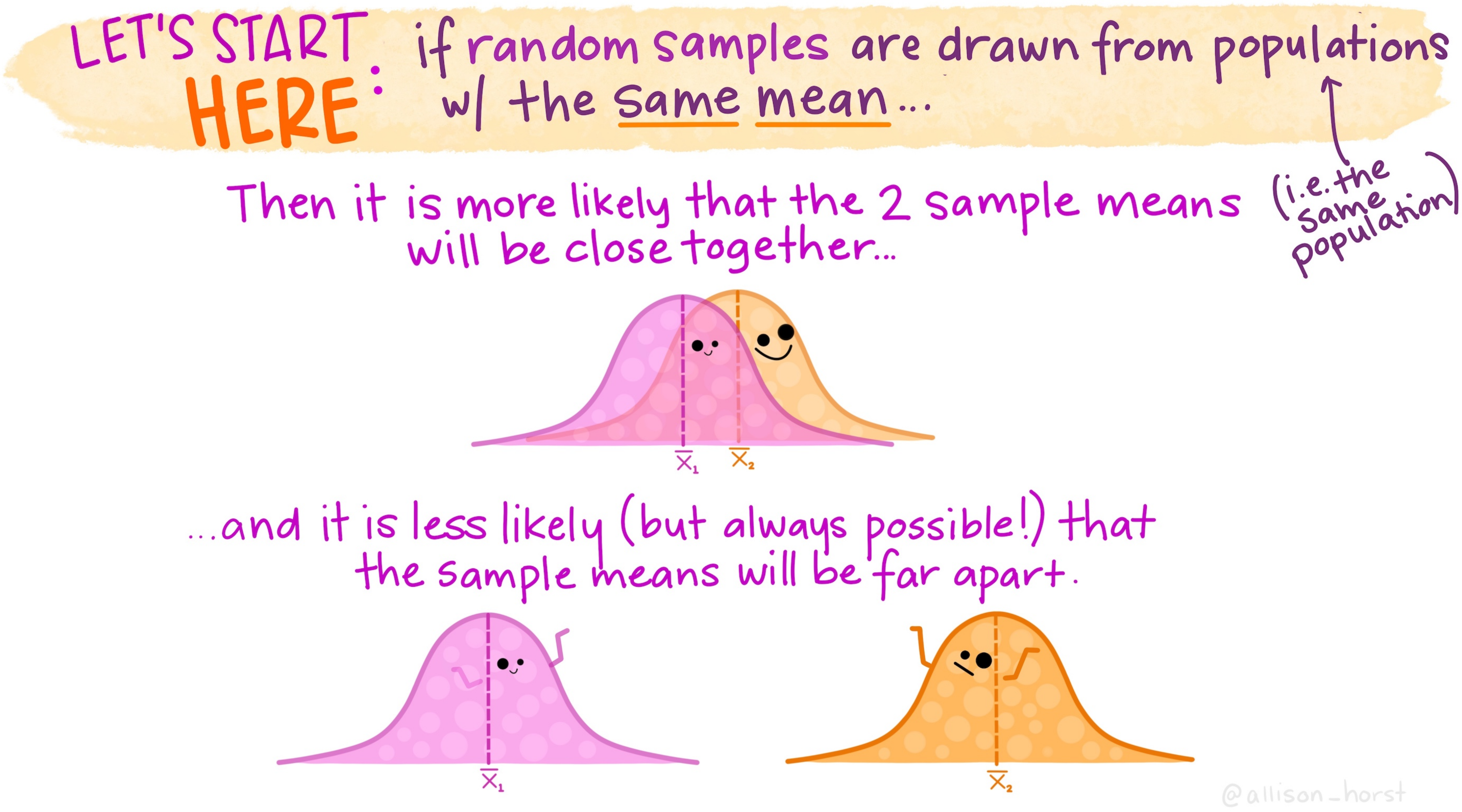
Caveat, assumptions, caveat (EDS 222)...

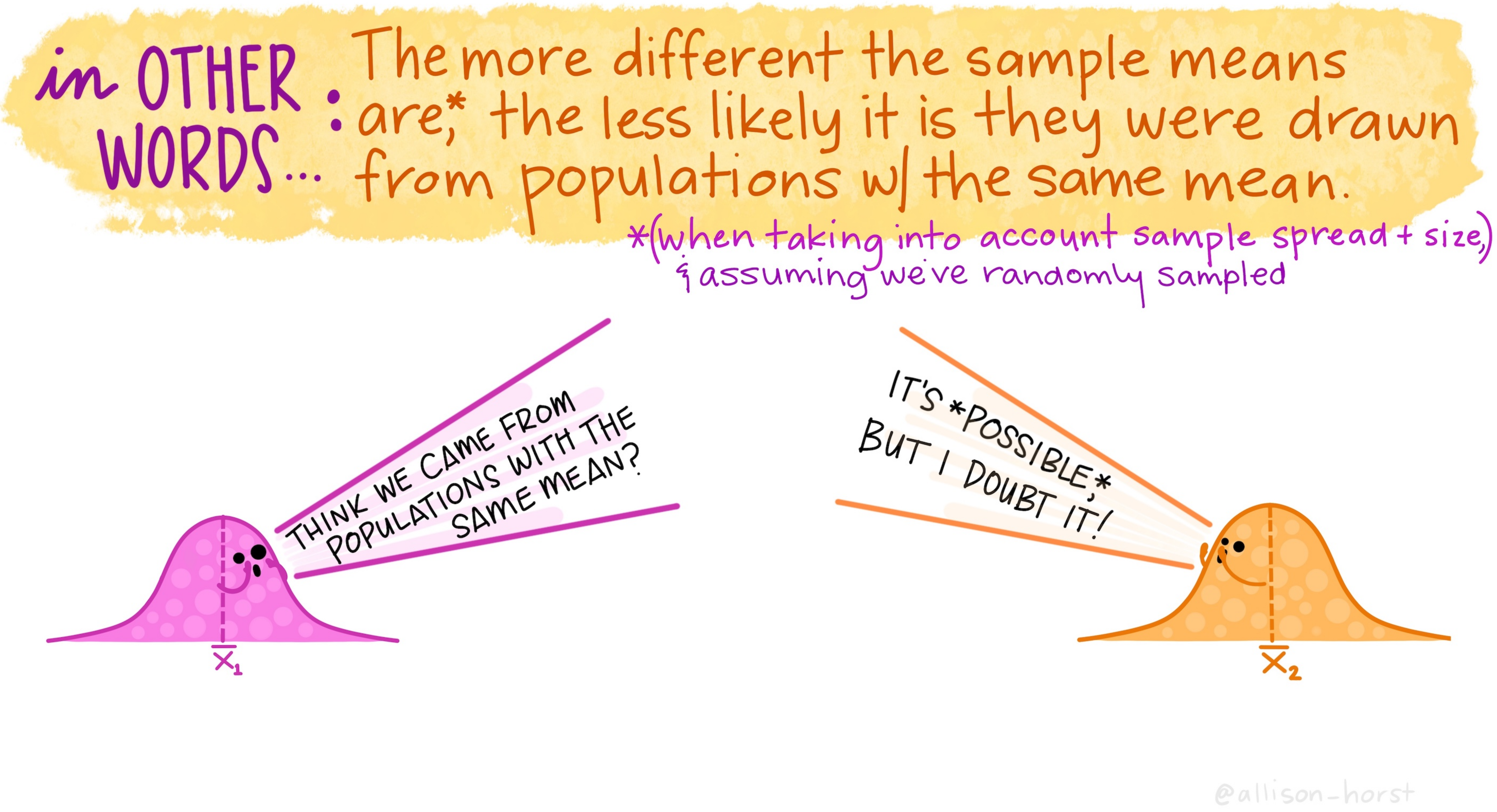
2-SAMPLE T-TESTS

@allison_horst

teaching assistants:



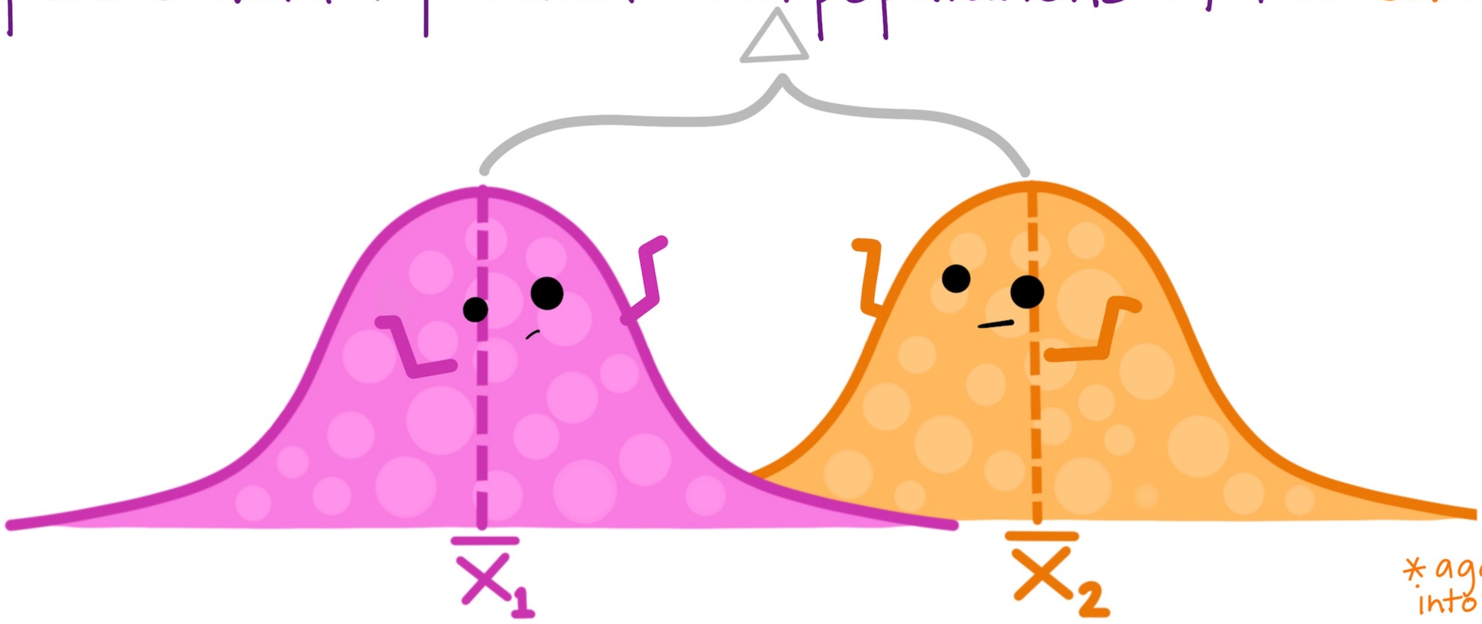

LET'S START HERE: if random samples are drawn from populations w/ the same mean...
(i.e. the same population)
Then it is more likely that the 2 sample means will be close together...
$\bar{X}_1$
$\bar{X}_2$
...and it is less likely (but always possible!) that the sample means will be far apart.
$\bar{X}_1$
$\bar{X}_2$
@allison_horst


in OTHER WORDS...: The more different the sample means are,* the less likely it is they were drawn from populations w/ the same mean.
*(when taking into account sample spread + size, & assuming we've randomly sampled)
THINK WE CAME FROM POPULATIONS WITH THE SAME MEAN?
IT'S *POSSIBLE,* BUT I DOUBT IT!
$\bar{X}_1$
$\bar{X}_2$
@allison_horst

So for our 2 random samples, we ask:

WHAT IS THE PROBABILITY OF GETTING 2 SAMPLE MEANS THAT ARE AT LEAST THIS DIFFERENT,*

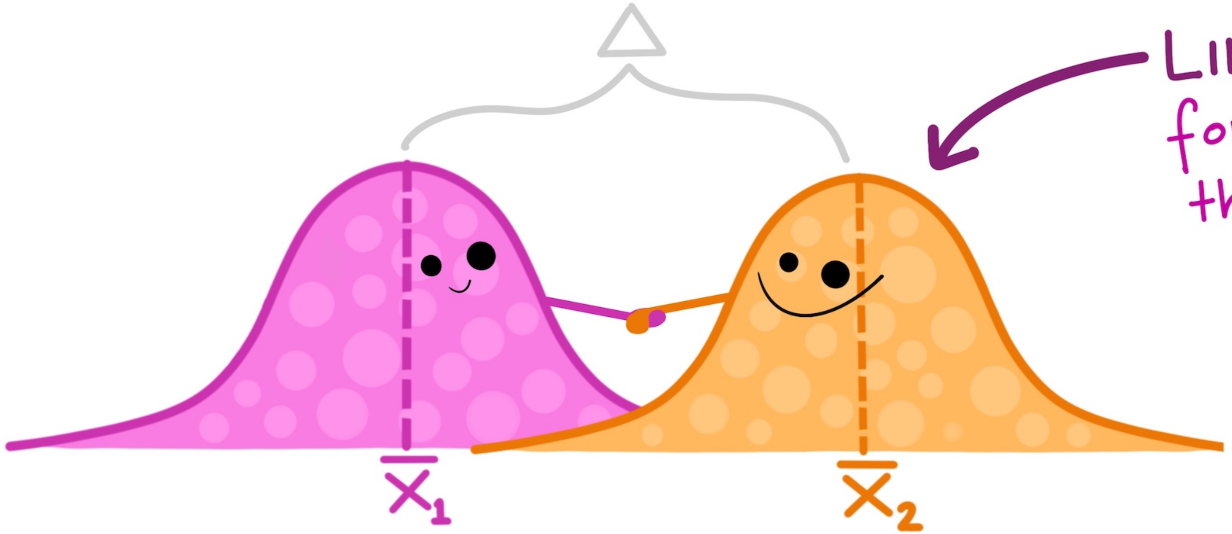
if they were actually drawn from populations w/ the same mean?



* again, when taking into account sample spread & size, + assumptions...

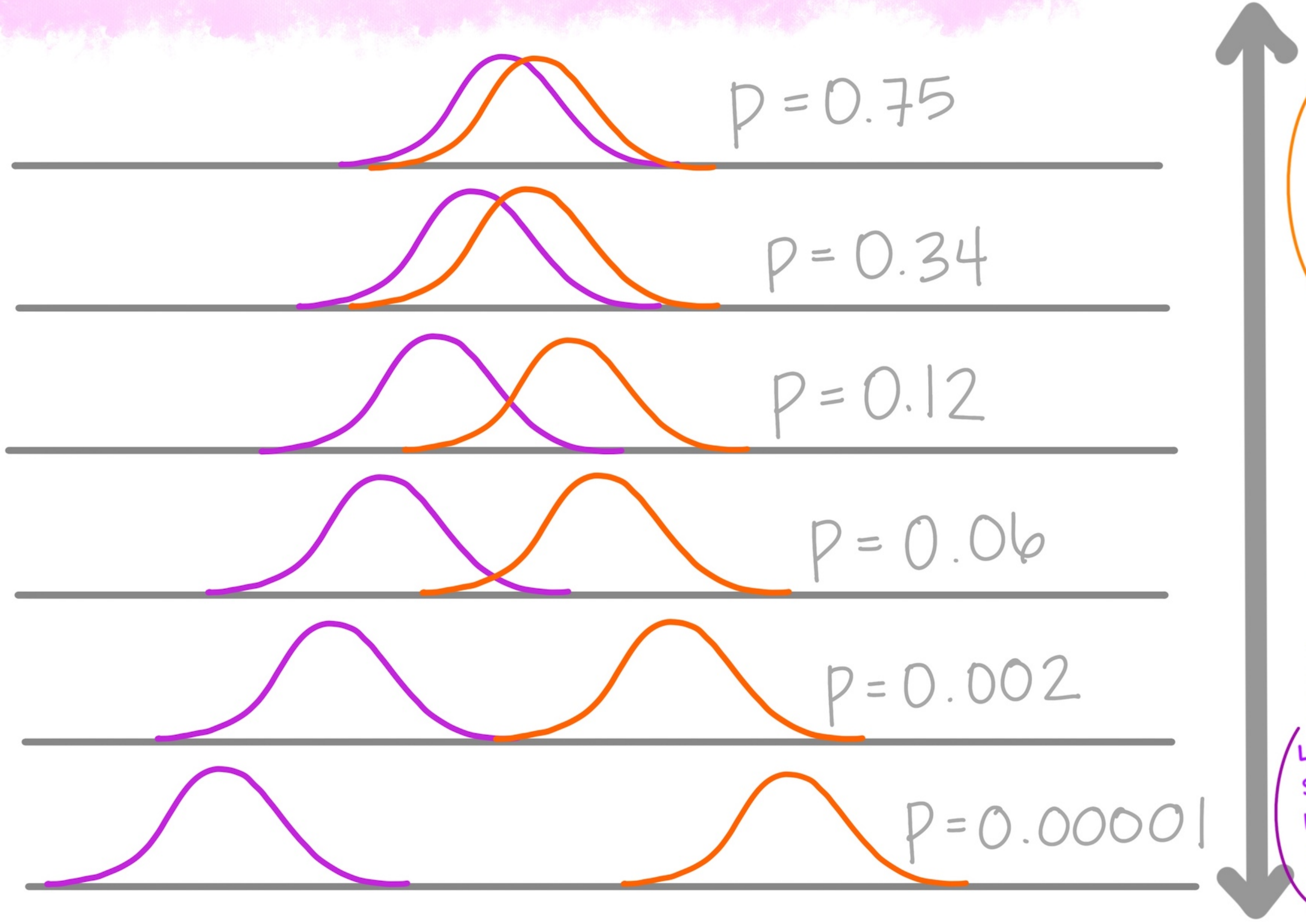
That's our p-value!

WHAT IS THE $\boxed{\text{PROBABILITY}}$ OF GETTING 2 SAMPLE MEANS THAT ARE AT LEAST THIS DIFFERENT, if they were actually drawn from populations w/ the same mean?



LIKE: If a 2-sample t-test for these samples yields $p = 0.03$, that means there is a 3% chance of getting means that are at least this different, if they're drawn from populations with the same mean.

P-VALUES, SCHEMATICALLY:



Higher p-values

(HIGHER PROBABILITY OF 2 SAMPLE MEANS BEING AT LEAST THIS DIFFERENT, IF DRAWN FROM POPULATIONS WITH THE SAME MEAN) = LESS EVIDENCE OF DIFFERENCES BETWEEN POPULATION MEANS

Lower p-values

(LOWER PROBABILITY OF 2 SAMPLE MEANS BEING AT LEAST THIS DIFFERENT, IF DRAWN FROM POPULATIONS WITH THE SAME MEAN) = MORE EVIDENCE OF DIFFERENCES BETWEEN POPULATION MEANS

RECAP: Thinking about p-values

THE MORE DIFFERENT* OUR 2 SAMPLE MEANS ARE, THE LESS LIKELY IT IS THAT THEY'RE FROM POPULATIONS WITH THE SAME MEAN...

...WHICH MEANS THE P-VALUE IS GETTING SMALLER...

...AND EVIDENCE FOR SAMPLES BEING DRAWN FROM POPULATIONS WITH DIFFERENT MEANS IS GOING UP.

*See t-statistic slides

Question:

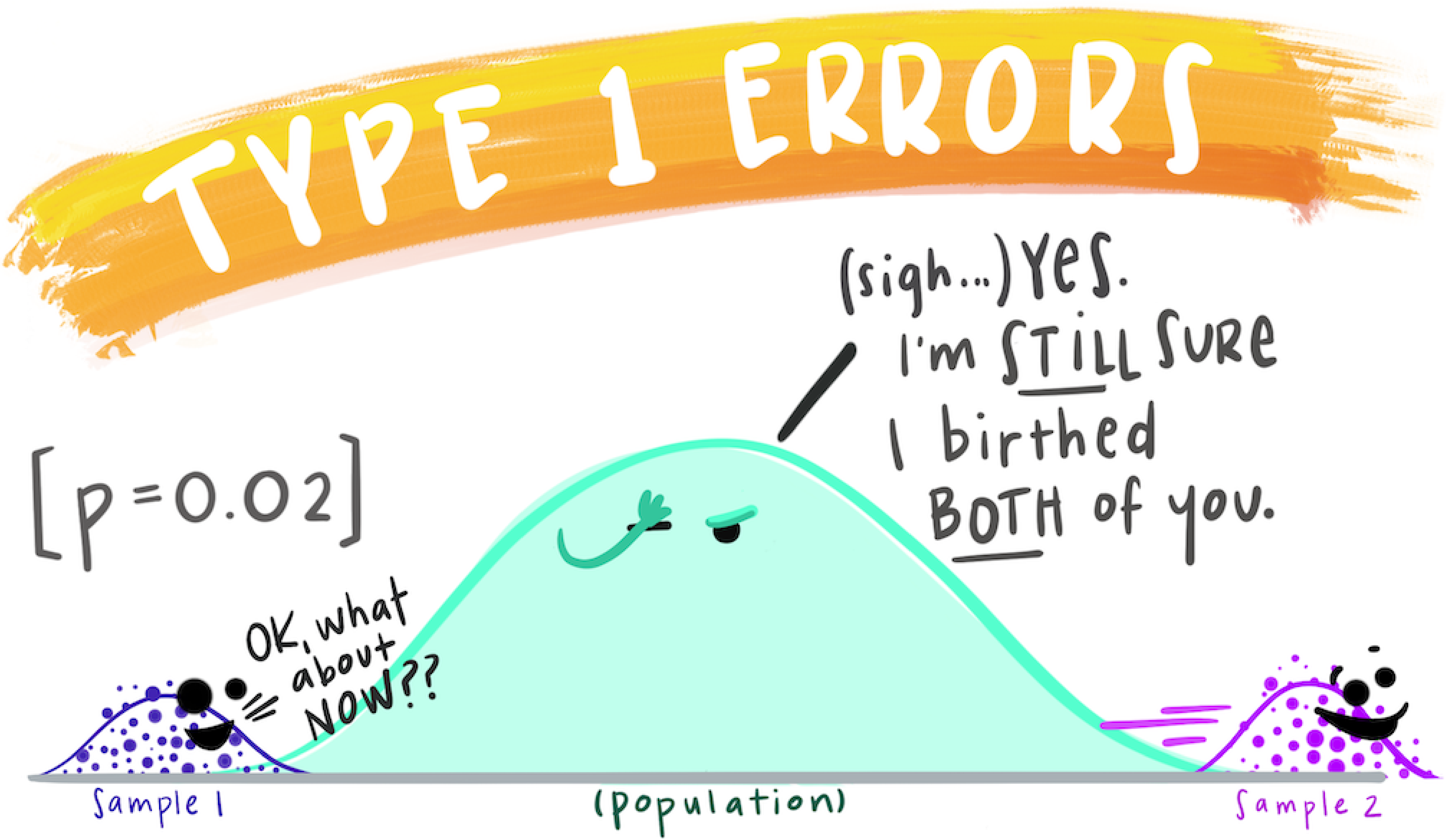
WHEN DO WE HAVE ENOUGH EVIDENCE TO SAY THERE IS A SIGNIFICANT DIFFERENCE?

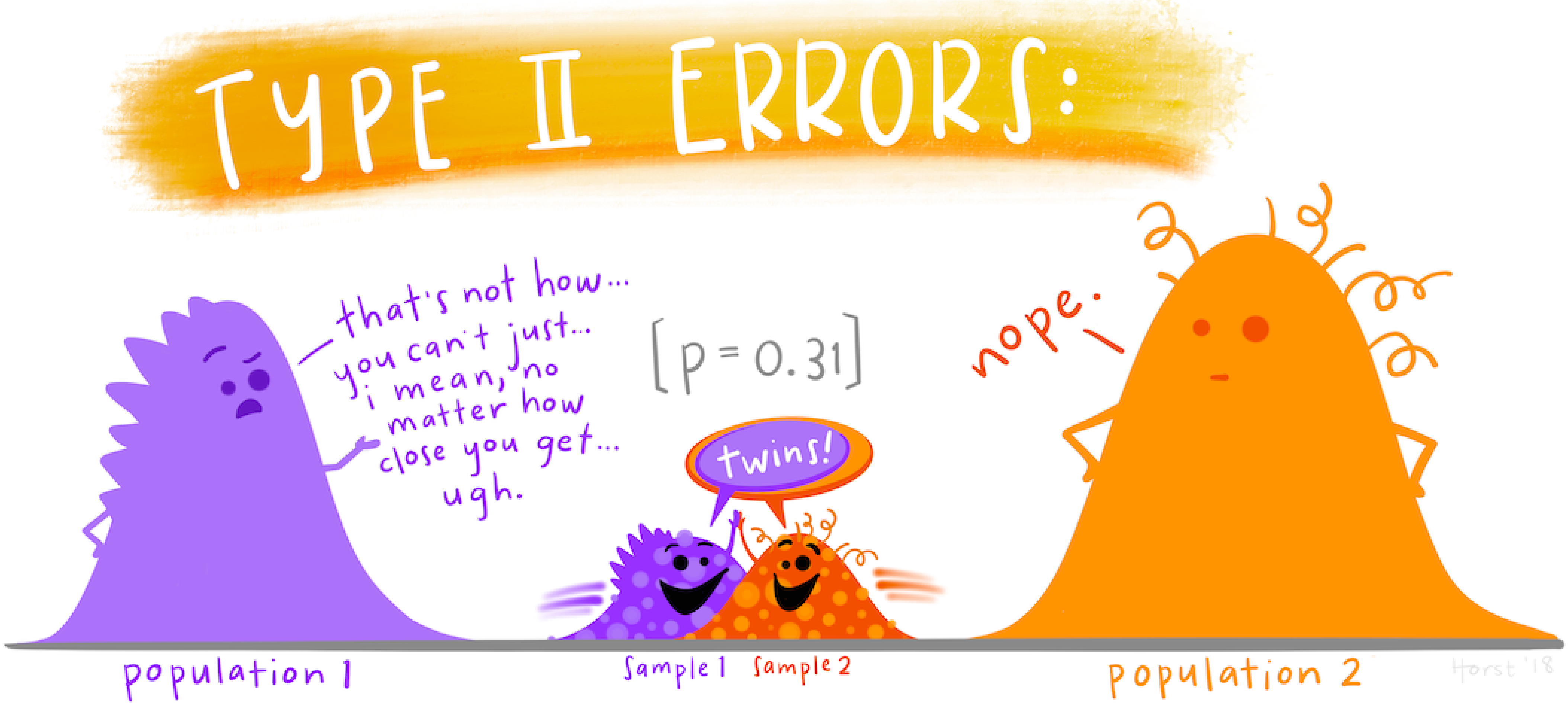
Answer:

WHEN OUR P-VALUE IS BELOW OUR SELECTED SIGNIFICANCE LEVEL (α), USUALLY (BUT NOT ALWAYS) = 0.05.

Which means:

IF THE PROBABILITY (p-value) OF FINDING AT LEAST OUR DIFFERENCE IN SAMPLE MEANS (IF THEY WERE DRAWN FROM POPULATIONS WITH THE SAME MEANS) IS LESS THAN 5%, THAT'S ENOUGH EVIDENCE FOR US TO DECIDE THEY ARE LIKELY FROM POPULATIONS WITH UNEQUAL MEANS.


TYPE 1 ERRORS
(sigh...) Yes. I'm STILL sure I birthed BOTH of you.
[p=0.02]
OK, what about NOW??
Sample 1
(population)
Sample 2


TYPE II ERRORS:
that's not how...
you can't just...
i mean, no
matter how
close you get...
ugh.
[p = 0.31]
twins!
nope.
population 1
sample 1
sample 2
population 2
Horst '18

Boolean logic

Definition: "In mathematics and mathematical logic, Boolean algebra is the branch of algebra in which the values of the variables are the truth values true and false, usually denoted 1 and 0, respectively" (Wikipedia)

Computer think

How would a computer order the objects in the following statements?

- **Nothing** is better than a **burrito**
- A **loaf of bread** is better than **nothing**

Mathematically…

- Nothing > burrito: TRUE
- Loaf of bread > nothing: TRUE

To a computer: Loaf of bread > nothing > burrito.

In environmental data science

- Conditional statements
- Filtering, subsetting, searching
- Checking classes and verification
- Testing

Logical operators

- Logical "and": `&`
- Logical "or": `|`
- Logical "negate": `!`

Comparison operators

- Is equal to? `==`
- Is less than? `<`
- Is less than or equal to? `<=`
- Is greater than? `>`
- Is greater than or equal to? `>=`
- Is not equal to? `!=`

A computer evaluates these and the outcome is either `TRUE` or `FALSE`, and proceeds accordingly.

An important distinction:

==: This is…equal to?

=: This IS equal to.

> 5 == 4
>
> FALSE

Examples:

Elements of a vector are tested separately, and the outcome is returned in a vector:

```
marmot <- c(1,2,3)
marmot == 2
```

```
## [1] FALSE  TRUE FALSE
```

```
pika <- c(1,2,5,9,10,15)
pika == 1 | pika >= 9
```

```
## [1]  TRUE FALSE FALSE  TRUE  TRUE  TRUE
```

Checking data classes works similarly:

More on data types & structures in EDS 221!

```
bear <- c(1,4,3, NA, 6) # Create a vector
is.na(bear) # Check element by element for == NA?
```

```
## [1] FALSE FALSE FALSE  TRUE FALSE
```

```
is.numeric(bear) # Checks entire *class* of vector
```

```
## [1] TRUE
```

Another we'll see often: `%in%`

- `%in%`: check for matching elements (not in order)*

Example: We have two vectors, and we want to know if any values in `vole` are also in `mouse`

```
vole <- c(1,3,4,7,10,15)
mouse <- c(0,2,4,0,0,NA,11,15,20)

vole %in% mouse
```

```
## [1] FALSE FALSE  TRUE FALSE FALSE  TRUE
```

*Keep this in mind - the distinction between `%in%` and `==` is major and important.