

EDS 212: Day 4, Lecture 2

Basic probability theory, PDFs and PMFs, hypothesis testing intuition

August 6th, 2025

Basic probability concepts

Why refresh probability?

- Being familiar with the terms
- Reporting uncertainty
- Used in statistics
- Hypothesis testing

Event space, event, and probability

- ✎ • event space: the collection of all possible unique outcomes of an experiment or scenario.
Also called sample space.
 - event: a possible outcome within the event space.
 - probability of an event: how likely it is that a certain event occurs.
- If A denotes an event. Then $P(A)$ = probability of A happening.
- * $P(A) = 0$ indicates no chance of A happening.
 - * $P(A) = 1$ indicates A will happen with complete certainty.
- All probability values are between $[0, 1]$.

Event space, event, and probability

Event space: The collection of all possible unique outcomes of an experiment or scenario. Also called the *sample space*.

Event: A possible outcome within the event space.

Probability: the likelihood that an event occurs.

The probability of event A occurring is written as $P(A)$.

- $P = 0$ indicates no chance of an event happening
- $P = 1$ indicates it is a certainty it will happen

Two examples

1. Coin flip

- Experiment: flipping a coin
- Event space: {heads, tails}
- Event: A = we flip the coin and get tails
- $P(A)$ = the probability of flipping the coin and getting tails

2. Water sample contaminant test

- Experiment: testing water samples at the same location
- Event space: all possible nitrate concentration readings (mg/L)
- Event : A = the sample exceeds a certain nitrate limit
- $P(A)$ = the probability that a sample exceeds the nitrate limit

Independent vs. non-independent events

✎ Two events are independent if one occurring doesn't change the probability of the other one occurring.

If knowing A occurred does change the probability of B occurring, then the events are non-independent.

Conditional probability: A and B are events. The probability of B happening when we know A has happened is $P(B|A)$ "probability of B given A".

Independent vs. non-independent events

Two events are **independent** if one occurring doesn't change the probability of the other:

If knowing A occurred *does* change the probability of B , the events are **non-independent**.

Conditional probability: For events A and B , the probability of event B happening given that A is known to occur is denoted as $P(B|A)$.

 Are the following events independent or non independent?

- Two consecutive coin flips.
- Exceeding the nitrate limit in one day, then exceeding nitrate limit the next day.

Independent vs. non-independent events

Two events are **independent** if one occurring doesn't change the probability of the other:

If knowing A occurred *does* change the probability of B , the events are **non-independent**.

Conditional probability: For events A and B , the probability of event B happening given that A is known to occur is denoted as $P(B|A)$.

 Are the following events independent or non independent?

Two consecutive coin flips are independent events. One has no influence on the other.

Exceeding the nitrate limit in two consecutive days at the same location would likely be non-independent.

Intersection, union, and complement

Symbols:
 $\cap$ = intersection
 $\cup$ = union
' = complement

Given two events A and B :

- **Intersection:** $A \cap B$ = denotes the events A and B co-occurring. In other words, these are **AND** probability statements.
- **Union:** $A \cup B$ = denotes at least one of the events happens (could be just one, or both). In other words, these are **OR** probability statements.
- **Complement:** A' = denotes the event A does not occur.

 Given the events A and B , write in words what each of the following mean:

A = nitrate concentration in January 1st exceeds the safety limit

B = nitrate concentration in August 1st exceeds the safety limit

- $A \cap B$
- A' = nitrogen concentration doesn't exceed limit
- $A \cup B$
- (?) nitrogen concentration exceeds limit on a day that is not Jan 1.

Intersection, union, and complement

Given two events A and B :

- **Intersection:** $A \cap B =$ denotes the events A and B co-occurring. In other words, these are **AND** probability statements.
- **Union:** $A \cup B =$ denotes *at least one* of the events happens (could be just one, or both). In other words, these are **OR** probability statements.
- **Complement:** $A' =$ denotes the event A *does not* occur.

 Given the events A and B , write in words what each of the following mean:

A = nitrate concentration in January 1st exceeds the safety limit

B = nitrate concentration in August 1st exceeds the safety limit

Basic probability

Probability of the intersection

$P(A \cap B)$ = the probability of events A and B happening.

Calculation:

- If A and B are independent: $P(A \cap B) = P(A) \times P(B)$
- If A and B are not independent: $P(A \cap B) = P(A)P(B|A)$

Basic probability

Probability of the intersection

$P(A \cap B)$ = the probability of events A and B happening.

Calculation:

- If A and B are independent: $P(A \cap B) = P(A) \times P(B)$
- If A and B are not independent: $P(A \cap B) = P(A)P(B|A)$

Probability of the union

Notation: $P(A \cup B)$ = the probability of A or B happening.

Calculation: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Basic probability

Probability of the intersection

$P(A \cap B)$ = the probability of events A *and* B happening.

Calculation:

- If A and B are independent: $P(A \cap B) = P(A) \times P(B)$
- If A and B are not independent: $P(A \cap B) = P(A)P(B|A)$

Probability of the union

Notation: $P(A \cup B)$ = the probability of A *or* B happening.

Calculation: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Probability of the complement

Notation: $P(A')$ = the probability of A *not* happening

Calculation: $P(A') = 1 - P(A)$

Inference

Population/sample, parameter/statistic

- 
- population : the entire collection of things in a category you are trying to understand. You define the population.
 - Sample : a subset of the population. Goal: the sample is representative of the population.
 - parameter : a characteristic of ^{the} population
 - statistic : a characteristic of the sample

Population/sample, parameter/statistic

- **Population:** The entire collection of things in a category you are trying to understand. *You define the population.*
- **Sample:** A subset of the population, goal is to be *representative* of the population
- **Parameter:** A characteristic of the population
- **Statistic:** A characteristic of the sample

Examples

	Nitrate concentration	SB registered voters	Purple urchins (Channel Islands)
Population	All nitrate concentration readings over the year	All registered voters in Santa Barbara County	All purple urchins in the Channel Islands Marine Sanctuary
Sample	Nitrate concentration from 20 water samples collected weekly	500 voters surveyed by phone	50 urchins counted along 10 dive transects
Parameter	True average nitrate concentration in the stream over the year	True % of all SB voters supporting a ballot measure	True average urchin density (per m ²), sanctuary-wide
Statistic	Average nitrate concentration calculated from the 20 samples	% of the 500 surveyed voters supporting the measure	Average urchin density (per m ²) <u>across the 10 transects</u>

Examples



Your example

SB registered voters

Purple urchins (Channel Islands)

Population	All plankton in the ocean	All registered voters in Santa Barbara County	All purple urchins in the Channel Islands Marine Sanctuary
Sample	Sample of plankton	500 voters surveyed by phone	50 urchins counted along 10 dive transects
Parameter	true % of copepods in the ocean.	True % of all SB voters supporting a ballot measure	True average urchin density (per m ²), sanctuary-wide
Statistic	% of copepods in the sample	% of the 500 surveyed voters supporting the measure	Average urchin density (per m ²) across the 10 transects

Why do we sample?

Usually, we don't have the resources (time, money, human power, etc.) to collect observations for an entire population.

- As a proxy, we try to collect a representative sample.

Then, we attempt to draw conclusions about the pop
from the sample.

↪ The statistic becomes an estimate for the pop's true parameter.

This statistical inference!

Why do we sample?

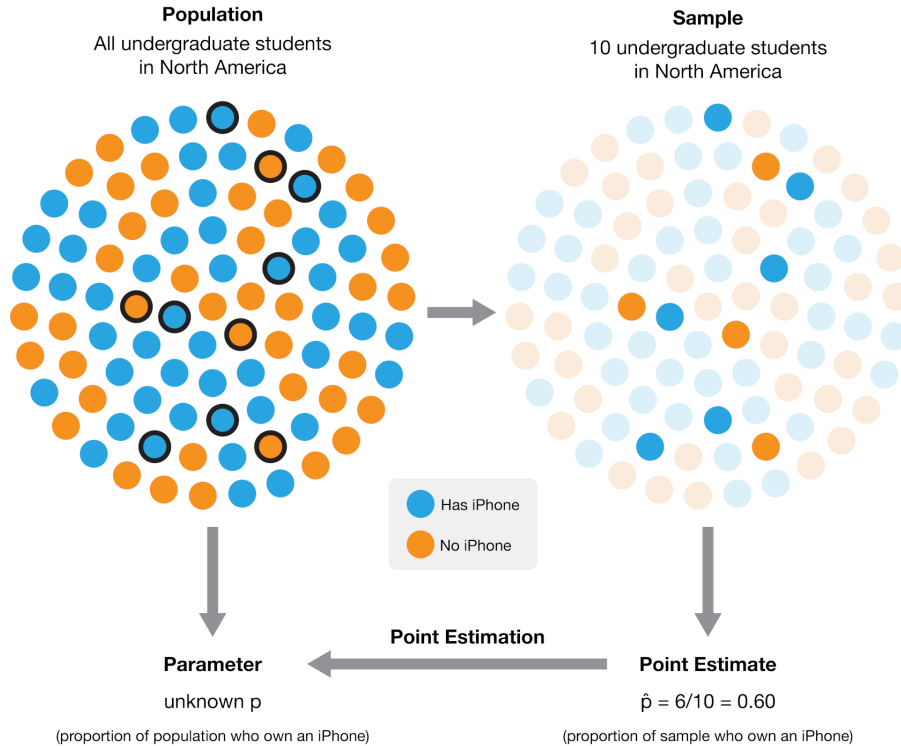
Usually, we don't have the resources (time, money, human power, etc.) to collect observations for an entire population. As a proxy, we try to **collect a representative sample**.

Then, we attempt **to draw conclusions about the populations** from which our samples were collected.

The statistic of the sample becomes an **estimate** for the true parameter in the population.

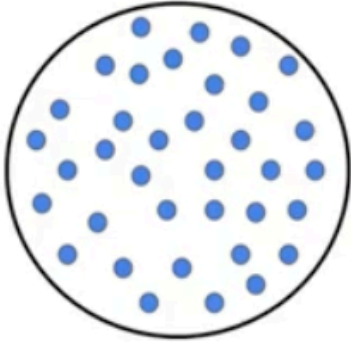
This process is called **statistical inference**.

Population vs. sample



Bessel's correction for variance

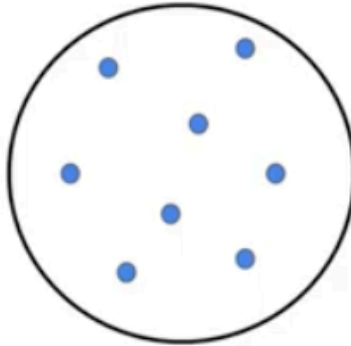
Population



$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

population mean (unknown)

Sample



$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

sample mean



Quantifying uncertainty

Once we use the sample to compute an estimate of the population's true parameter: how can we communicate how certain we are about that estimate?

One way of quantifying uncertainty around our statistics is using a...

Confidence interval: an interval based on a sample that, if we were to take multiple samples from the population and calculate the confidence interval from each, would contain the true population parameter X percent of the time.

Confidence interval example

Mean shark length is 8.42 ± 3.55 ft (mean $\pm$ standard deviation), with a 95% confidence interval of $[6.45, 10.39]$ ft ($n = 15$).

Confidence interval example

Mean shark length is 8.42 ± 3.55 ft (mean $\pm$ standard deviation), with a 95% confidence interval of [6.45, 10.39 ft] ($n = 15$).

What this **DOES NOT** mean: There is a 95% chance that the true population mean length is between 6.45 and 10.39 feet.

What this **DOES** mean: If we took a bunch of sets of samples from the population (all $n = 15$), then 95% of the time, the calculated mean would fall within this range.

This statement correctly describes the frequency with which we would expect CIs to capture the mean over many samples.

Let's take a 5 minute break



Let's meet
at 11:05.



Probability density and mass functions

Random variables

- A random variable is a variable whose value is the outcome of a random process.
Usually denoted with X, Y .
- Discrete random variable : take a countable set of values.
ex. X = the flip of a coin
 $X=1$: tails, $X=0$: heads.
- Continuous random variable : can take any value in an interval.
ex. Y = nitrate concentration (mg/L) in a water sample.

Random variables

A **random variable** is a variable whose value is a numerical outcome of a random process. Usually denoted with a capital letter (X , Y), with specific values denoted lowercase (x , y).

- **Discrete random variable:** takes on a countable set of values.

E.g. $X = 1$ if a coin flip lands tails, $X = 0$ if it lands heads.

- **Continuous random variable:** can take any value in an interval.

E.g. Y = the nitrate concentration (mg/L) in a water sample.

Probability mass function (PMF)

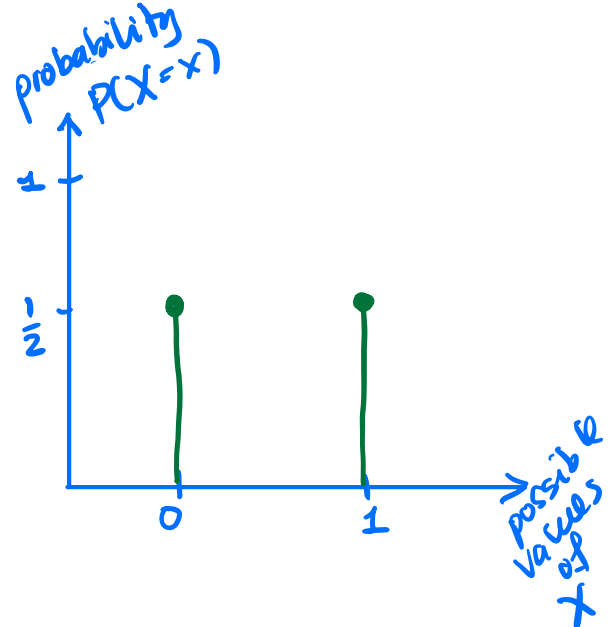
- For a discrete random variable X , the probability mass function (PMF) gives $P(X=x)$ = the probability that my random variable X takes value x .

ex: X = flipping fair coin

$X=1$: tails

$X=0$: heads

- $P(X=1)$ = prob. of coin being tails = $\frac{1}{2}$
- $P(X=0)$ = prob of coin being heads = $\frac{1}{2}$



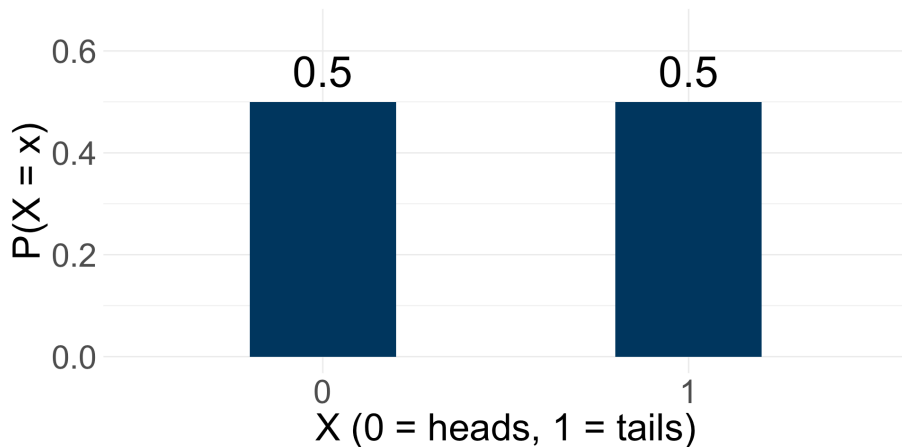
Probability mass function (PMF)

For a **discrete** random variable X , the **probability mass function (PMF)** gives $P(X = x)$ for each possible value x . The sum of all values for the PMF must be 1:

$$\sum_x P(X = x) = 1$$

Example: $X = 1$ if a fair coin lands tails, $X = 0$ if heads.

- x -axis: possible values of X
- y -axis: $P(X = x)$



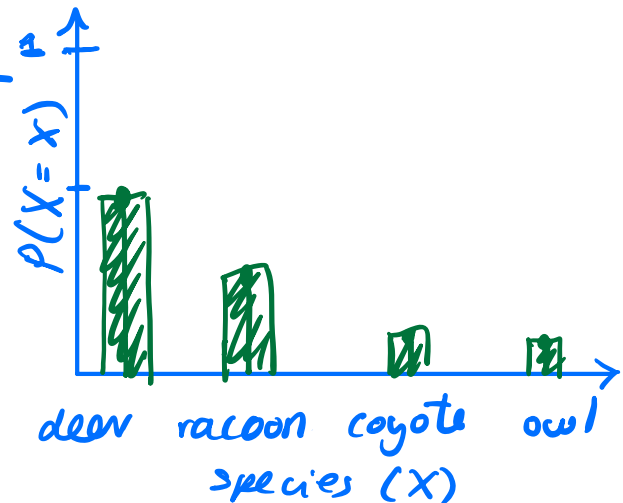
Example

✏ A wildlife camera trap records 10 animal sightings in one week: 5 deer, 3 raccoons, 1 coyote, 1 owl. Let X = the species captured in a randomly selected photo. Find $P(X = x)$ for each species and draw the PMF.

X = species in random photo

Possible outcomes for X : deer, raccoon, coyote, owl

Species	Count	$P(X=x)$
deer	5	$5/10 = 0.5$
raccoon	3	$3/10 = 0.3$
coyote	1	$1/10 = 0.1$
owl	1	$1/10 = 0.1$

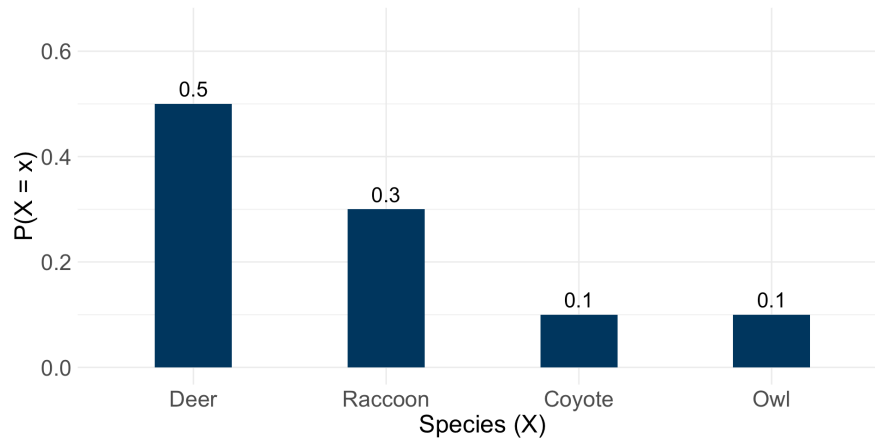


Total # of observations : 10

Solution

💡 Let's see a solution!

Species	Count	$P(X = x)$
Deer	5	0.5
Raccoon	3	0.3
Coyote	1	0.1
Owl	1	0.1



Exercise: your 212 survey

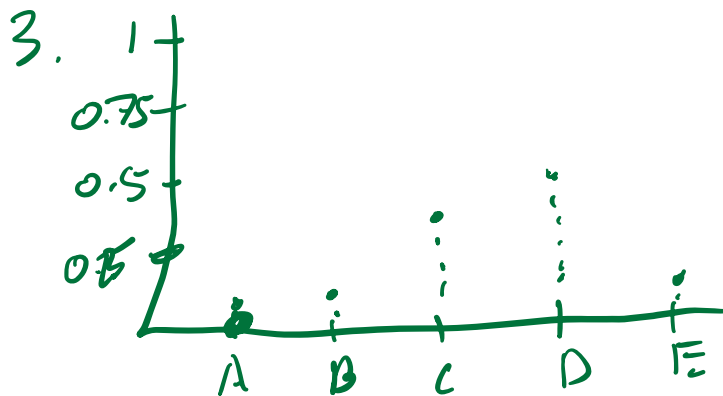
On day 1, you answered: *“What best describes your relationship with math?”*

Response	Count
My mortal enemy	0
A family friend you dislike but your parents insist you invite to everything	3
The friend of a friend whose name you can never remember	12
Respectful co-worker you can have a drink with	15
My best friend!	4

1. Response, qualitative, ordinal

2.

Response	#	Prob.
A	0	$0/34 = 0$
B	3	$3/34 = 0.09$
C	12	$12/34 = 0.35$
D	15	$15/34 = 0.44$
E	4	$4/34 = 0.12$



4. $19/34 = 0.56$
56%

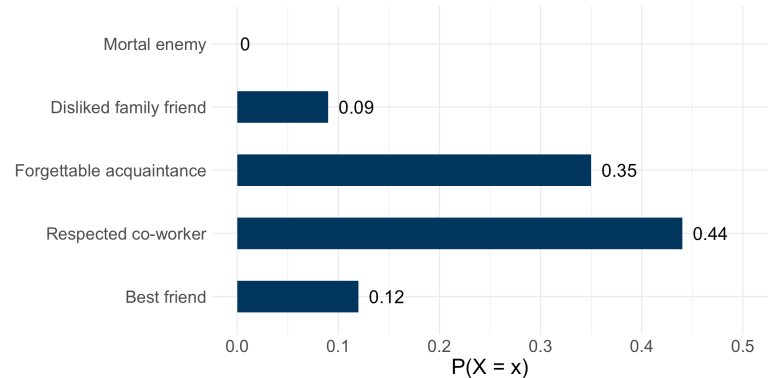
5. 63%

Exercise: solution

💡 Let's see a solution!

1. X = the response a student gave to the survey question. It's **qualitative**, and **ordinal**. The responses have an implied order, from worst to best relationship with math (much like a Likert scale).

Response	Count	$P(X = x)$
Mortal enemy	0	$0/34 = 0.00$
Disliked family friend	3	$3/34 \approx 0.09$
Forgettable acquaintance	12	$12/34 \approx 0.35$
Respected co-worker	15	$15/34 \approx 0.44$
Best friend	4	$4/34 \approx 0.12$



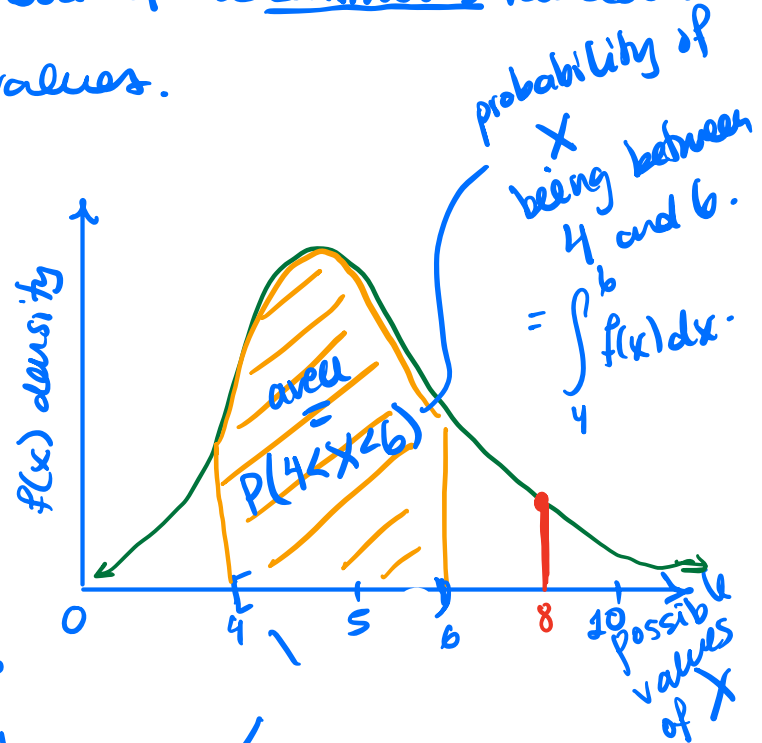
3. $P(\text{positive}) = P(\text{respected co-worker}) + P(\text{best friend}) = \frac{15+4}{34} \approx 0.56$

4. $P(\text{not neutral}) = 1 - P(\text{forgettable acquaintance}) = 1 - \frac{12}{34} \approx 0.65$

Probability density function

✎ Describes the relative likelihood of a continuous random variable taking on different values.

- x-axis : possible values of the random variable
- y-axis : the density : not a probability
- the area under the curve between two values is a probability
- The total area under the curve equals 1 : total probability



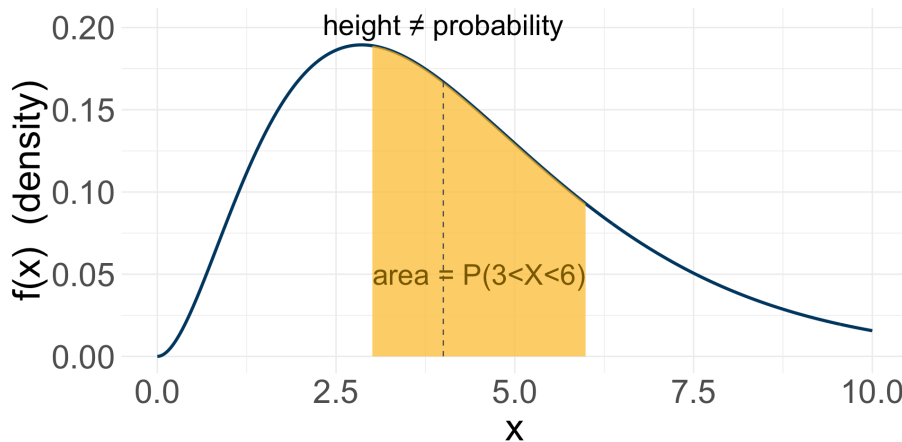
⑦ $P(4 < X < 6) = \text{prob in } [4, 6]$

Probability density function

$$P(4 < X < 6) = \text{prob from } (4, 6)$$

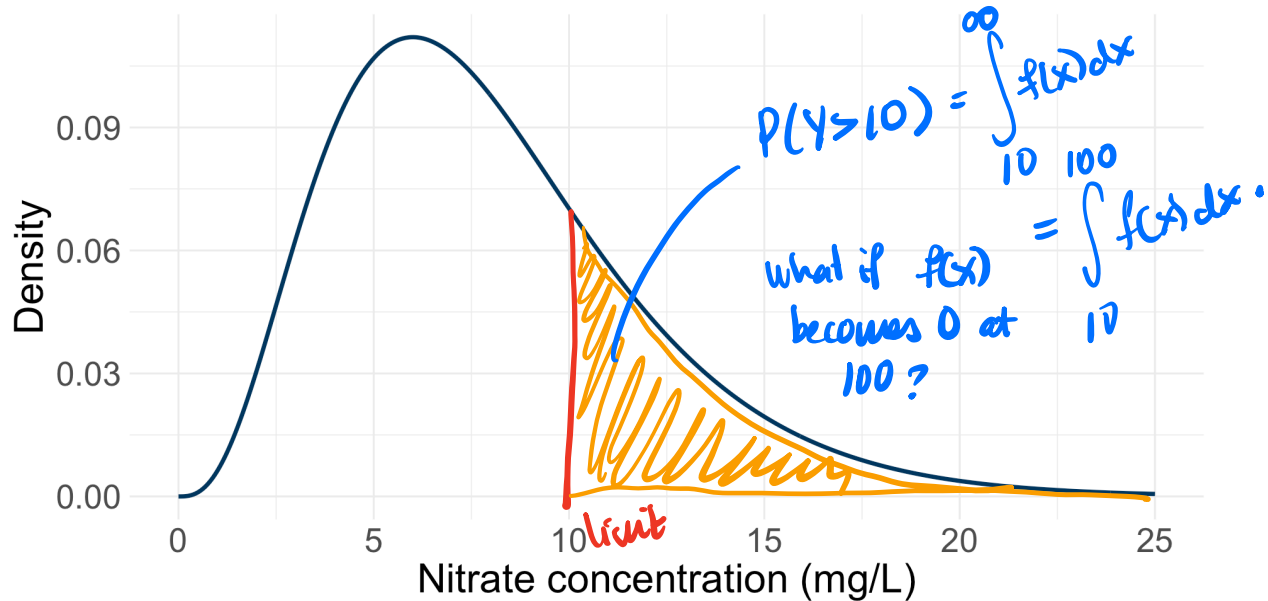
Describes the relative likelihood of a **continuous** random variable taking on different values.

- x -axis: possible values of the random variable
- y -axis: $f(x)$, the density (**not** a probability)
- The **height** of the curve is *not* a probability
- The **area under the curve** between two values *is* a probability
- The total area under the entire curve = 1



Check-in: nitrate concentration

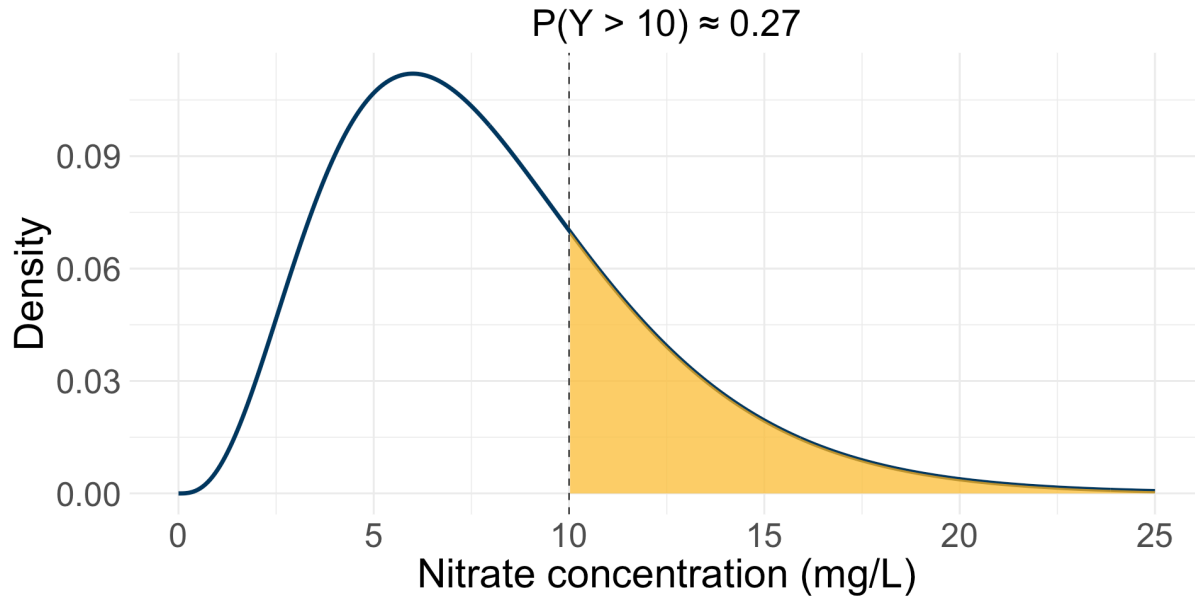
Nitrate concentration (mg/L) in stream water samples, modeled as a continuous random variable Y :



 Suppose the safety limit is 10 mg/L. How would you find $P(Y > 10)$?

Check-in: nitrate concentration

Nitrate concentration (mg/L) in stream water samples, modeled as a continuous random variable Y :

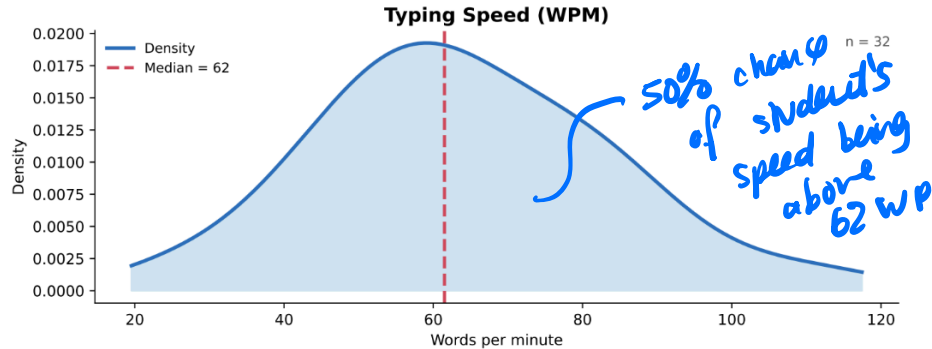


 Suppose the safety limit is 10 mg/L. How would you find $P(Y > 10)$?

The **shaded area** is the answer: there is about a 27% chance a sample exceeds the safety limit.

Exercise: your typing speed test

MEDS Keyboard Check — Distributions (2026-07-31)



1. Is this distribution symmetric or skewed? How can you tell from the shape of the curve?
2. The curve's height at 60 WPM is about 0.019. Does that mean the probability a student types at *exactly* 60 WPM is 0.019? Why or why not?
3. The dashed line marks the median at 62 WPM. What does that tell you about the area under the curve to the left of 62, compared to the right?

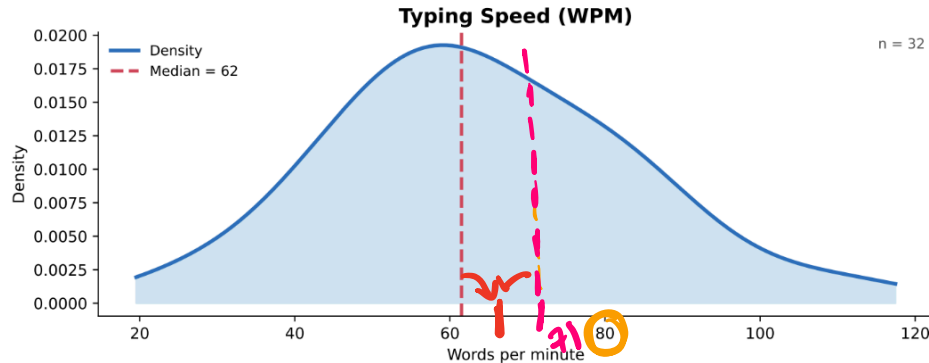
Exercise: solution

💡 Let's see a solution!

1. Right-skewed: the peak sits to the left of center, and the curve has a longer, thinner tail stretching toward higher WPM values (up to ~120) than toward lower ones.
2. No. For a continuous random variable, the probability of any *exact* single value is 0. The height of the curve is the **density**, not a probability. Only the **area** under the curve over a range represents a probability.
3. The median splits the total area (=1) into two equal halves: 50% of the area lies to the left of 62 WPM, and 50% lies to the right.

Exercise: comparing speeds

MEDS Keyboard Check — Distributions (2026-07-31)



 The instructor's own speed on this test was 71 WPM. Speeds of 80 WPM or higher are generally considered “fast.”

1. Is it more likely than not that the instructor types faster than a randomly selected student from the class?
2. Would “fast” typists (80+ WPM) make up a small minority, about half, or a majority of the class? Justify your answer using the shape of the curve.

Exercise: solution

💡 Let's see a solution!

- 1.** Yes. Since 71 WPM is *greater* than the median (62 WPM), more than half the area under the curve lies to the left of 71. Meaning more than half the class types slower than 71 WPM. So it's more likely than not that the instructor is faster than a randomly chosen student. Notice we can say this with confidence using *only* the median, without needing to estimate any area by eye.
- 2.** A minority. Both the median (62) and the peak of the curve sit well below 80, so the region to the right of 80 covers noticeably less than half the total area.

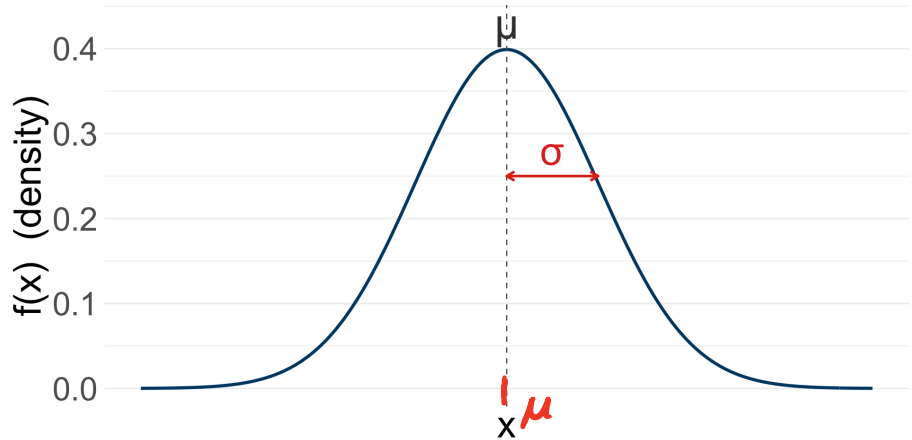
The normal distribution



The normal distribution

The **normal distribution** is a symmetric, bell-shaped PDF, fully described by two parameters:

- μ (mean): center of the distribution
- σ (standard deviation): spread of the distribution



Hypothesis testing

What is a null hypothesis?

-  Null hypothesis (H_0) claims that the effect being studied does not exist.
Proposes that there is no statistically significant difference between the data or variables being studied.
- ✱ If your null hypothesis is true, then any observed difference is due to chance alone.

What is a null hypothesis?

A null hypothesis (H_0) is the claim that the effect being studied does **not** exist. It is a hypothesis that proposes that there is **no statistically significant difference** between the data or variables being studied.

If the null hypothesis is *true*, any observed difference is due to chance alone.

What is a hypothesis test?

It asks the question

If a null hypothesis is true, what is the probability that your data outcome (e.g. mean, value, etc.) or something more extreme would have occurred by random chance?

Is that **so** unlikely (is the probability low enough), that you think you have sufficient evidence to reject the null hypothesis? Or not?

Hypothesis testing: building intuition, continued

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

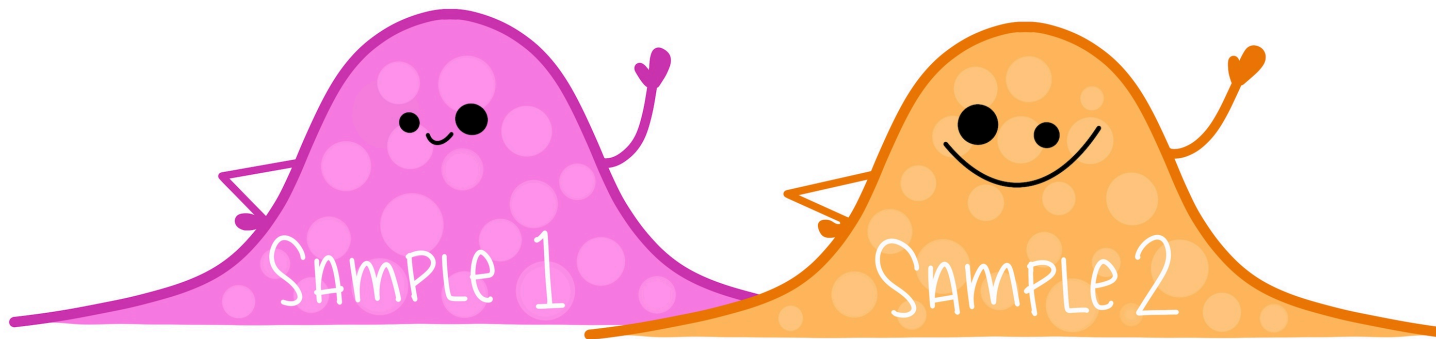
A common question: are means from two samples so different (considering data spread and sample size) that we think we have enough evidence to reject a null hypothesis that they were drawn from populations with the **same** mean?

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

2-SAMPLE T-TESTS

teaching assistants:

@allison-horst



LET'S START
HERE: if random samples are drawn from populations
w/ the same mean...

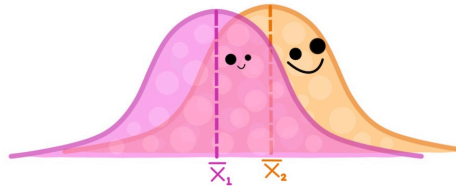
↑
(i.e. the
same
population)

Is it more likely that their
means will be similar?

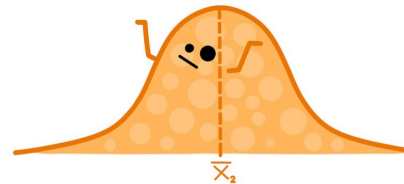
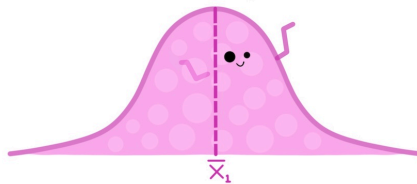
@allison_horst

LET'S START
HERE: if random samples are drawn from populations
w/ the same mean...

Then it is more likely that the 2 sample means
will be close together... (i.e. the
same population)



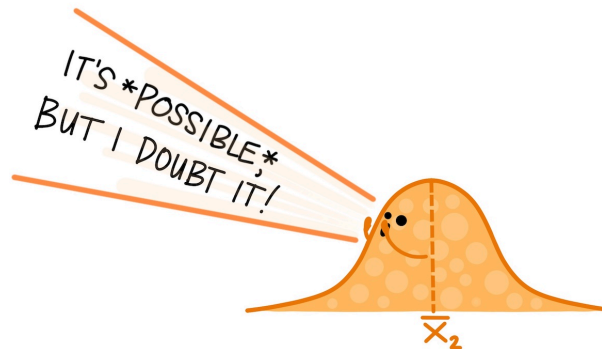
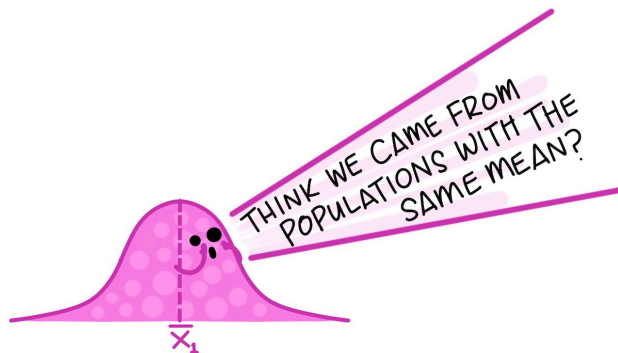
...and it is less likely (but always possible!) that
the sample means will be far apart.



@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



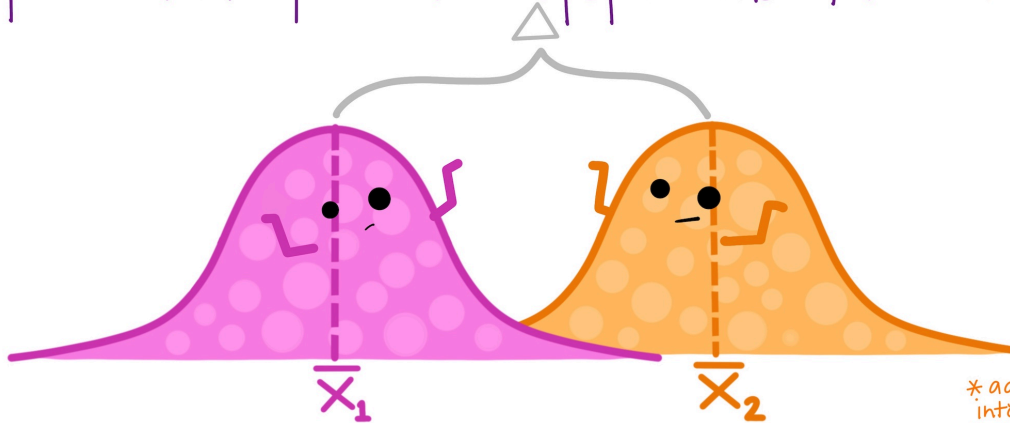
@allison-horst

So for our 2 random samples, we ask:

WHAT IS THE PROBABILITY OF GETTING 2 SAMPLE

MEANS THAT ARE AT LEAST THIS DIFFERENT,*

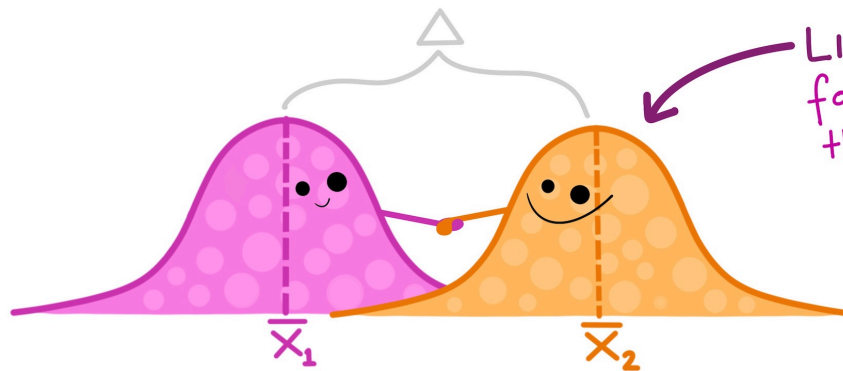
if they were actually drawn from populations w/ the same mean?



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

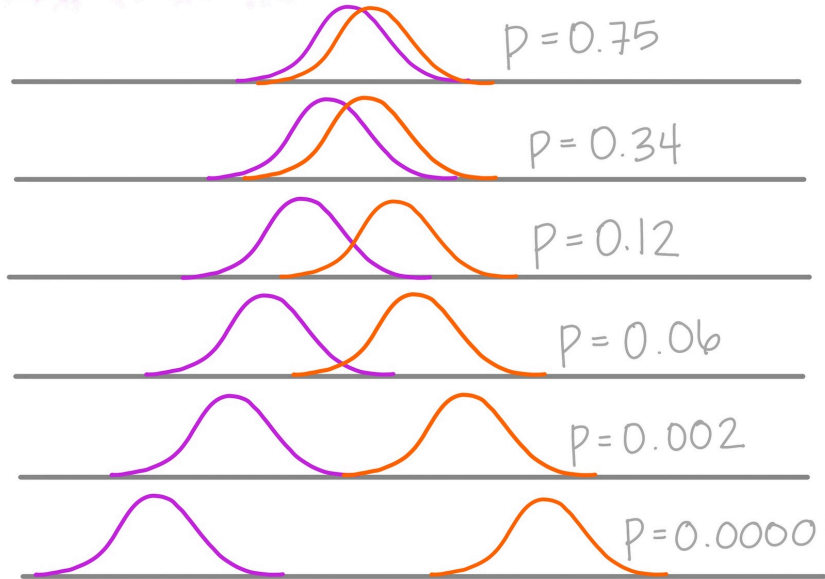
That's our p-value!

WHAT IS THE 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?

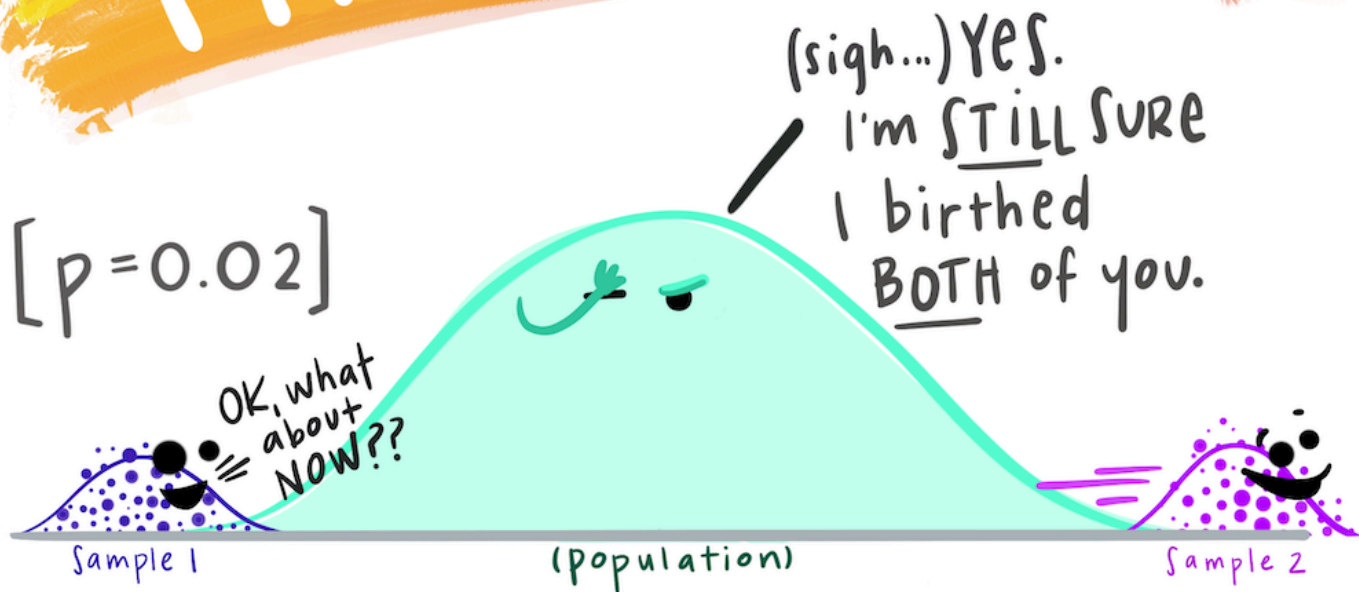
Answer:

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

Which means:

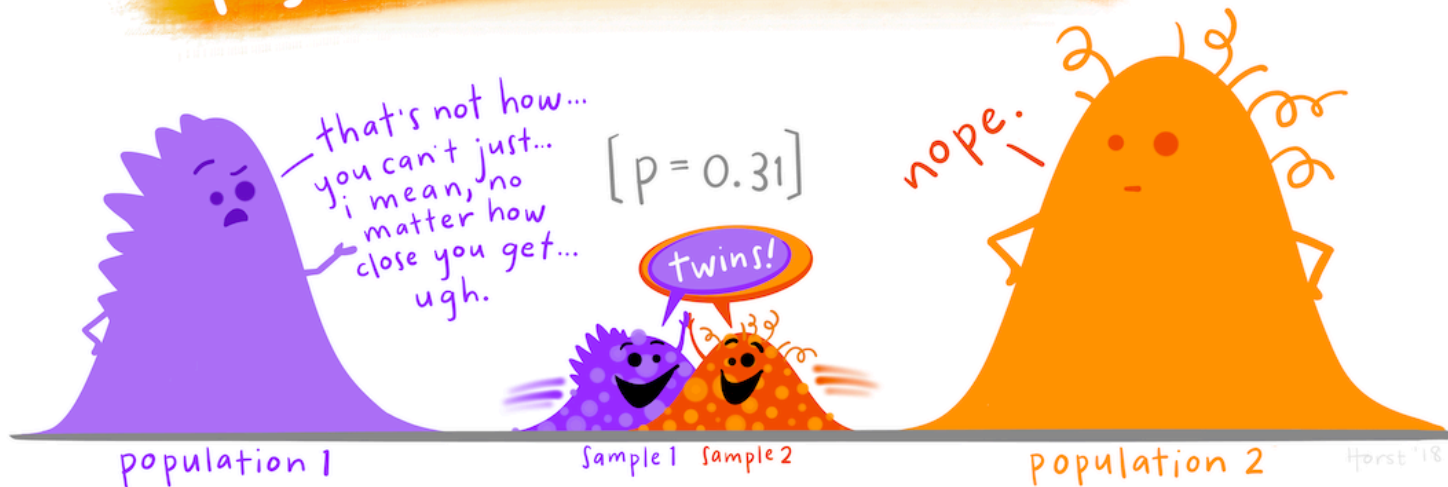
IF THE PROBABILITY (p-value) OF FINDING AT LEAST OUR
DIFFERENCE IN SAMPLE MEANS (IF THEY WERE DRAWN
FROM POPULATIONS WITH THE SAME MEANS) IS
LESS THAN 5%, THAT'S ENOUGH EVIDENCE FOR
US TO DECIDE THEY ARE LIKELY FROM POPULATIONS
WITH UNEQUAL MEANS.

TYPE 1 ERRORS



Horst '18

TYPE II ERRORS:



Wrapping up...

This course is an iterative effort



Allison Horst



Ruth Oliver



Nathan Grimes



Carmen Galaz García



Sam Shanny-Csik



Alessandra Vidal Meza



Cella Schnabel

We covered a lot!

-  **Algebra review**
-  **Solving equations**
-  **Exponents**
-  **Polynomials**
-  **Units**
-  **Definition of functions**
-  **Graphing**
-  **Reading graphs**
-  **The number e and $\exp()$**
-  **Logistic growth**
-  **Natural logarithms**
-  **Linear functions**
-  **Slope as a rate of change**
-  **Limits**
-  **Derivatives and rate of change**
-  **Tangent lines**
-  **Rules for differentiation**
-  **Constant rule**
-  **Power rule**
-  **Sum and difference rule**
-  **Higher order derivatives**
-  **Partial derivatives**
-  **Integrals**
-  **Riemann sums**
-  **Rules of integration**
-  **Initial value problems**
-  **Definite and indefinite integrals**
-  **Area between two curves**
-  **A bit of ODEs**
-  **Practice all day 1 and 2 topics**
-  **A bit of ODEs**
-  **Summation notation**
-  **Vector sum and product**
-  **Dot product**
-  **Matrix algebra**
-  **Data types**
-  **Data visualization**
-  **Quartiles & IQR**
-  **Central tendency**
-  **Variance & standard deviation**
-  **Why visualizing data matters**
-  **Confidence intervals**
-  **Communicating data summaries**
-  **Basic probability**
-  **Random variables**
-  **Probability mass functions**
-  **Probability density functions**
-  **The normal distribution**
-  **Hypothesis testing intuition**
-  **Two-sample t-tests**
-  **Type 1 & Type 2 errors**



Pleaes fill out the course evals!



ESSENTIAL MATH EDS