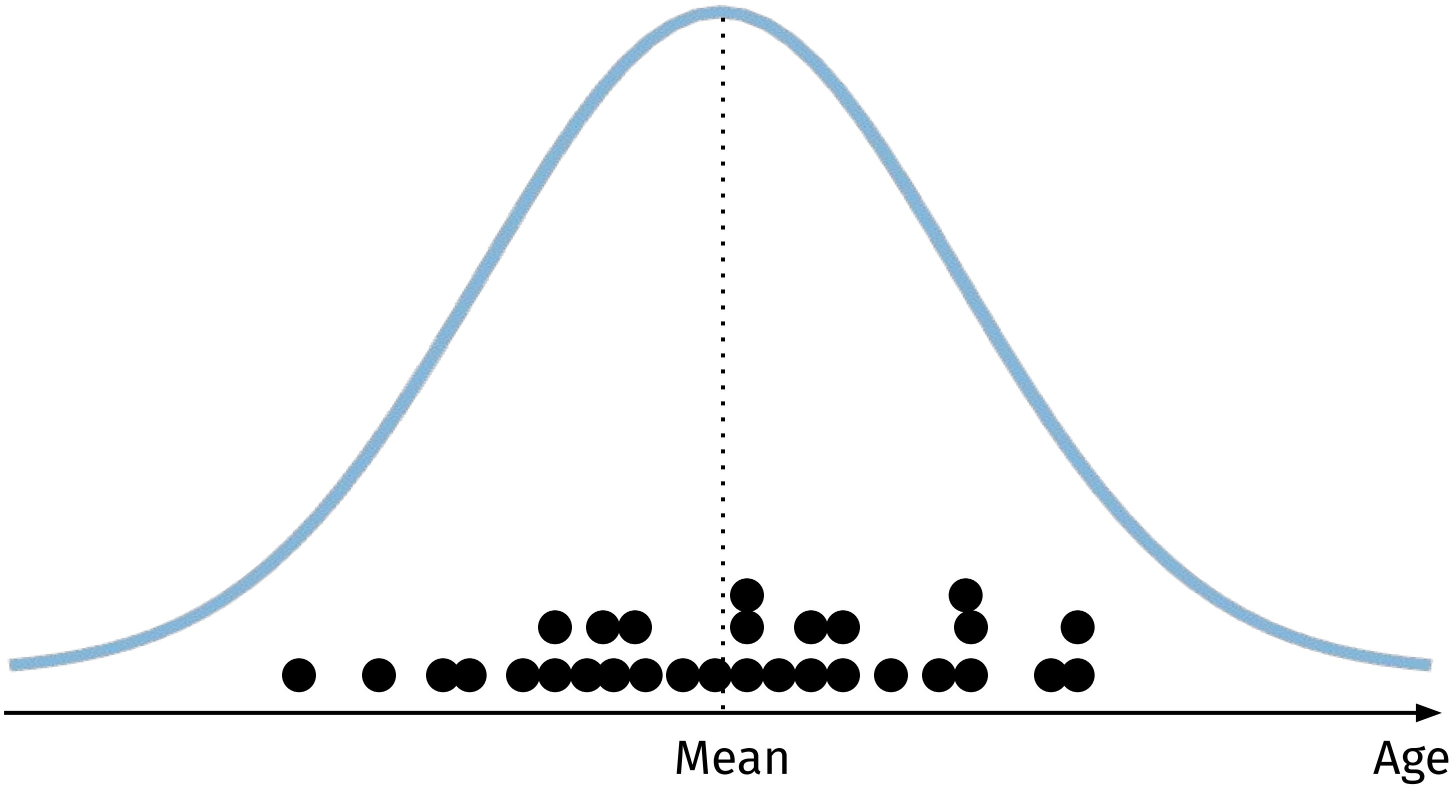
Z-scores & Probability

Lecture 5

Emma Ning, M.A.

From our last lecture...

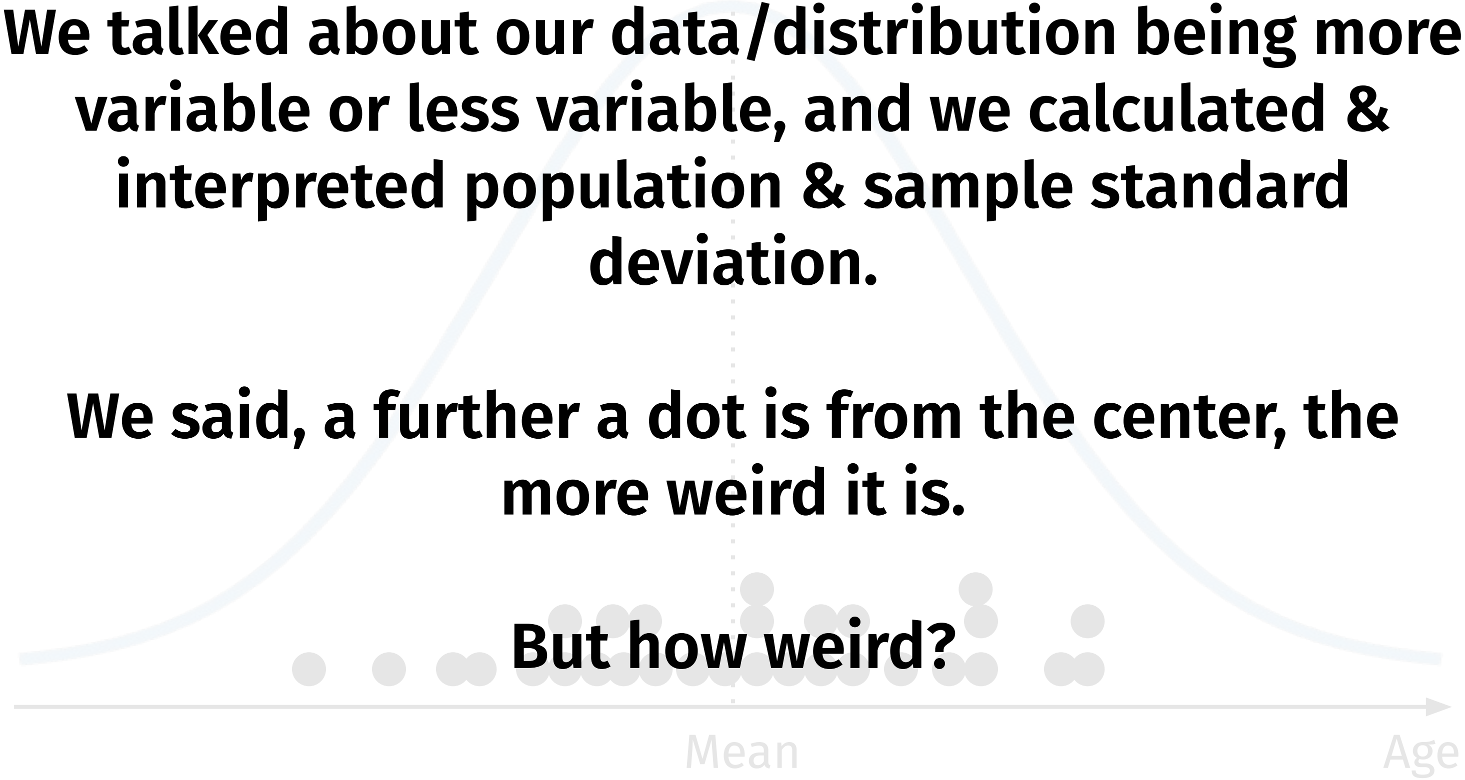


From our last lecture...

We talked about our data/distribution being more variable or less variable, and we calculated & interpreted population & sample standard deviation.

We said, a further a dot is from the center, the more weird it is.

But how weird?



TODAY'S PLAN

01

Introduction to Z-Scores

02

Calculating & Interpreting Z-scores

03

Empirical Rule & Probability

04

Wrap Up

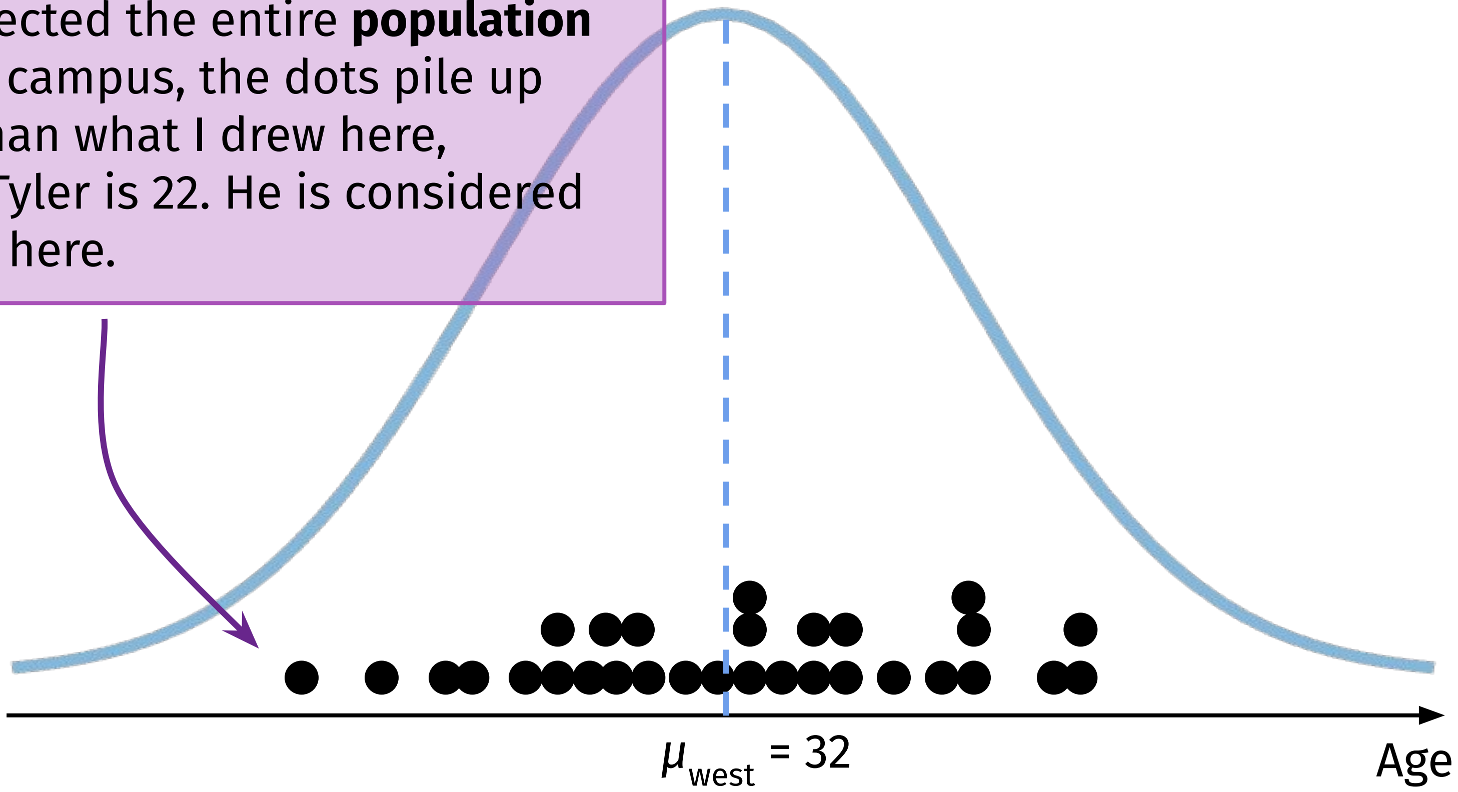
Learning objectives

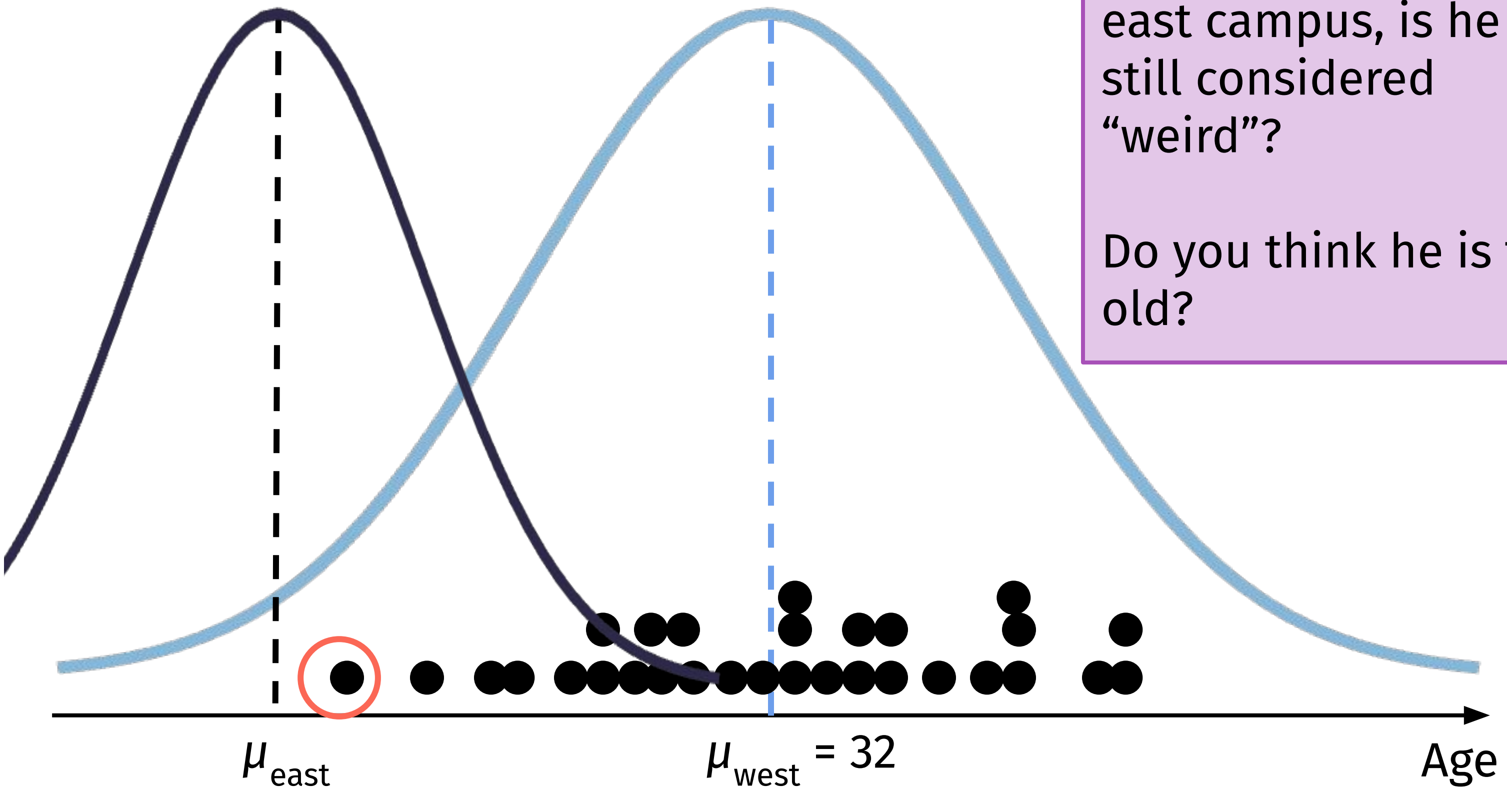
- Explain why **z-scores** are needed, and give examples of how z-scores are used
- **Transform** raw scores into z-scores and transform z-scores into raw scores
- Describe the effects of **standardizing** a distribution and explain the advantages of this transformation
- Interpret z-scores using your understanding of **probability** and the **empirical rule**

01

Z-scores

Meet Tyler. He wants to know if he is too old for school. So we compare him with our overall student ages. We collected the entire **population** of west campus, the dots pile up more than what I drew here, $\mu = 32$. Tyler is 22. He is considered "weird" here.





However, if we consider Tyler using our population from east campus, is he still considered "weird"?

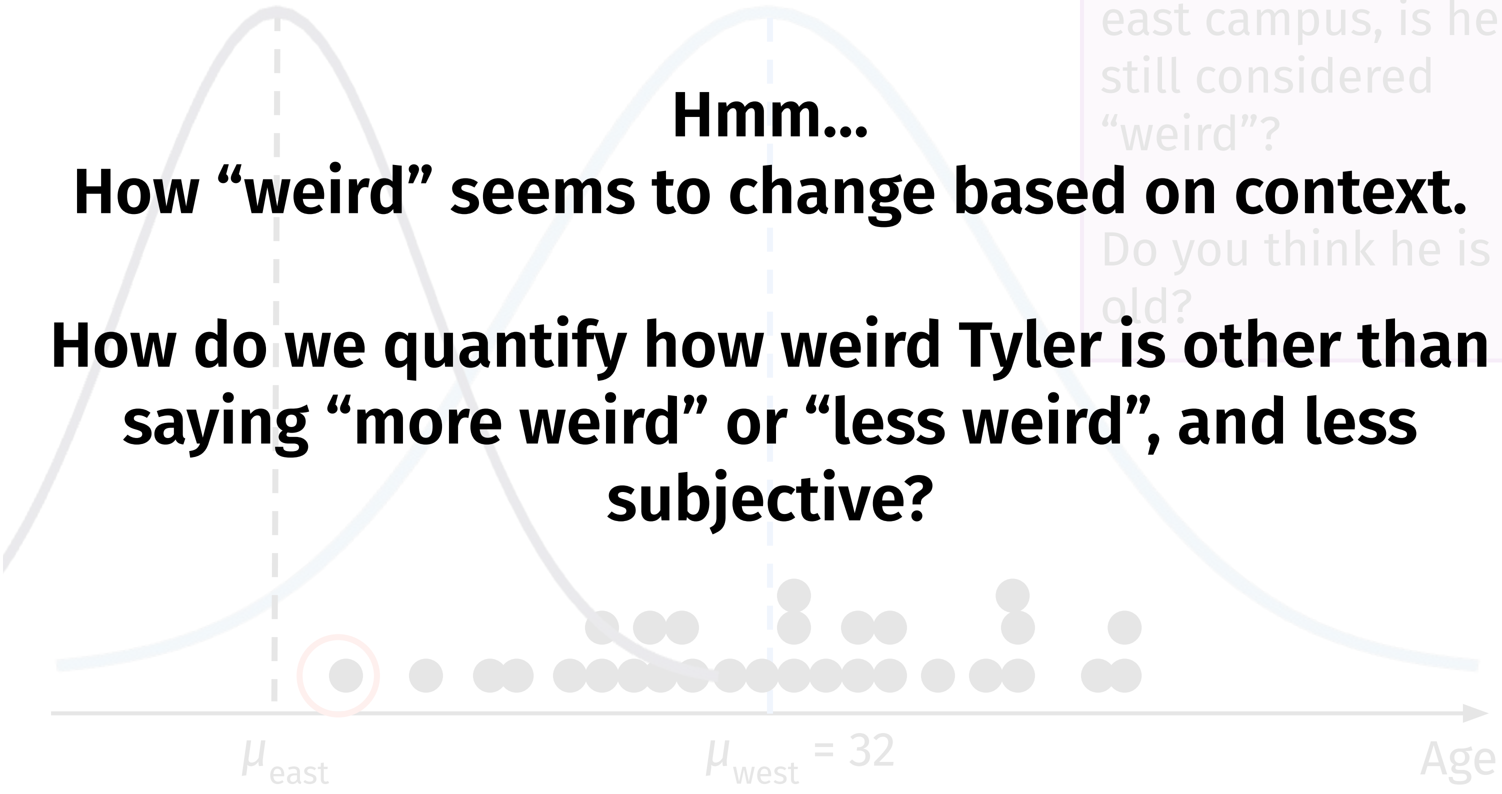
Do you think he is too old?

However, if we consider Tyler using our population from east campus, is he still considered "weird"?

Do you think he is too old?

Hmm...
How "weird" seems to change based on context.

How do we quantify how weird Tyler is other than saying "more weird" or "less weird", and less subjective?

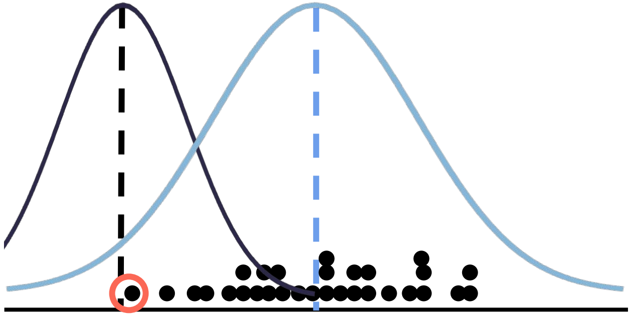


The use of z-scores

Z-scores let us compare how extreme a value is—even across different groups—by putting everything on the same scale.

East campus

μ = 20
σ = 4



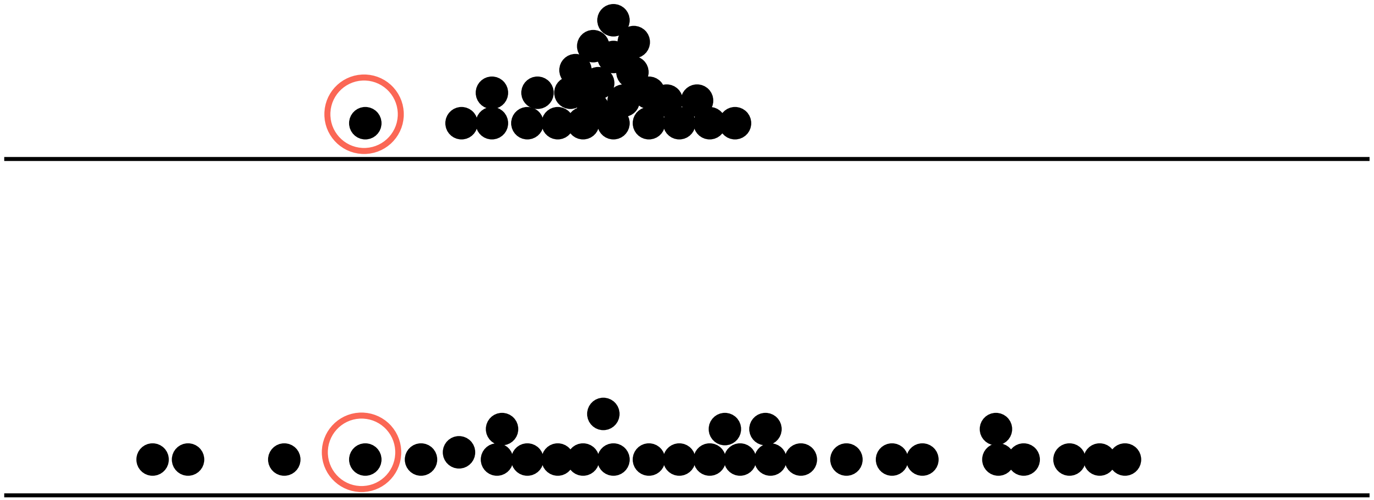
West campus

μ = 32
σ = 6

Tyler is older than average on east campus, but he is younger on west campus.

Take a guess—Is mean the only thing we need to consider to calculate z-scores?

We also need to consider SD apart from the mean to calculate z-score.
If everyone is close together in one sample, being a few distances away can mean a lot (weirdo). If everyone is far apart in the sample, then a few distances matter less (not so weird).



02

Calculating & Interpreting Z-Scores

Z-Scores

We can calculate a z-score for **each person** in a population. This score tells us how far that person is above or below the mean in standard deviation units.

“How extreme is this person compared to the population average?”

The Z-score Formula (population)

$$Z = \frac{\text{score - mean}}{\text{standard deviation}}$$

Our focus today!

$$Z = \frac{x - \mu}{\sigma}$$

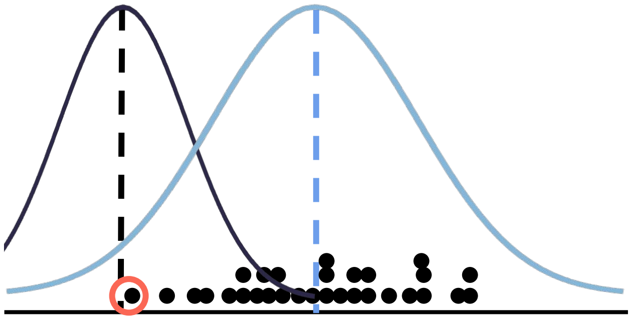
The Z-score Formula (sample)

$$Z = \frac{\text{score - mean}}{\text{standard deviation}}$$

$$Z = \frac{x - M}{s}$$

East campus

$\mu = 20$
$\sigma = 4$



West campus

$\mu = 32$
$\sigma = 6$

Tyler's age: 22

Tyler's z-score under the east campus distribution:

$$Z = \frac{\text{score - mean}}{\text{standard deviation}}$$

$$Z = \frac{x - \mu}{\sigma} = \frac{22 - 20}{4} = +0.5$$

Tyler's z-score under the west campus distribution:

$$Z = \frac{\text{score - mean}}{\text{standard deviation}}$$

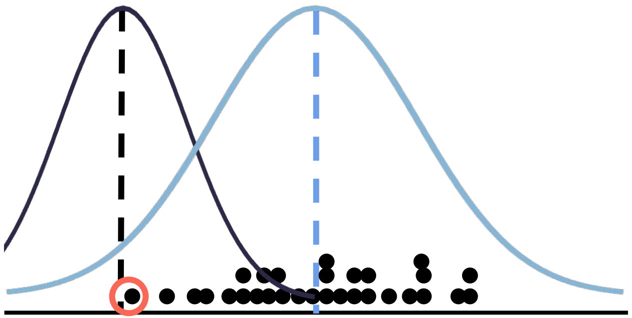
$$Z = \frac{x - \mu}{\sigma} = \frac{22 - 32}{6} = -1.7$$

East campus

$\mu = 20$

$\sigma = 4$

$z_{tyler} = \boxed{+}0.5$

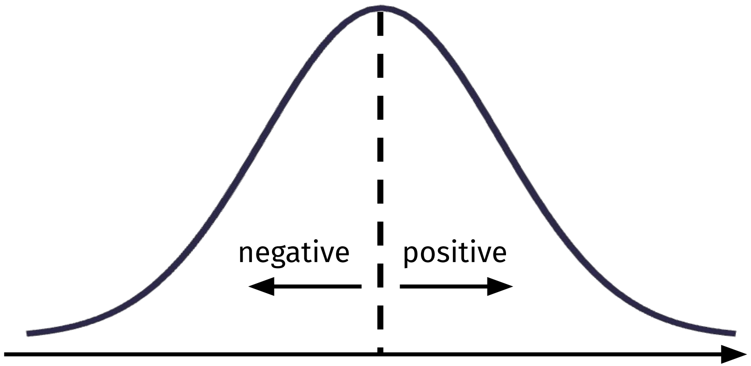


West campus

$\mu = 32$

$\sigma = 6$

$z_{tyler} = \boxed{-}1.7$



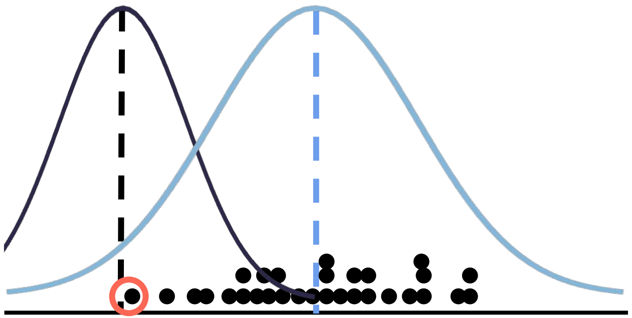
The sign of the z-score (+ or –) signifies whether the score is above the mean (positive) or below the mean (negative).

East campus

$\mu = 20$

$\sigma = 4$

$z_{tyler} = +\boxed{0.5}$



West campus

$\mu = 32$

$\sigma = 6$

$z_{tyler} = -\boxed{1.7}$

How large the value is signifies how many standard deviations the score is from the mean.

For east campus, Tyler is <u>0.5 SD units above the mean.</u>

For west campus, Tyler is <u>1.7 SD units below the mean.</u>

To interpret z-scores:

The sign tells us whether a person is above or below average (positive or negative).

We use the standard deviation like a ruler to measure how far a data point is from the mean—the absolute value of the z-score tells us that distance.

East campus

μ = 20

σ = 4

West campus

μ = 32

σ = 6

How large the value is signifies how many standard deviations the score is from the mean.

For east campus, Tyler is 0.5 SD units above the mean.

For west campus, Tyler is 1.7 SD units below the mean.

Let’s practice!

We measured the entire population of people who have last name “Wyer”. We know their population mean age is 20 years old, the population standard deviation is 15 years old. Calculate z-score for each person in our dataset!

Actual Age:	5	10	15	25	45

μ = 20 years
σ = 15 years

$$Z = \frac{x - \mu}{\sigma}$$

μ = 20 years
σ = 15 years

$$Z = \frac{x - \mu}{\sigma}$$

Actual Age:	5	10	15	25	45
Below or above average:	below	below	below	above	above
Individual Z-score Sign:	-	-	-	+	+
Deviation from the mean:	-15	-10	-5	+5	+25
Individual Z-score:	-1	-0.7	-0.3	+0.3	+1.7

Divide by σ

μ = 20 years
σ = 15 years

Actual Age:	5	10	15	25	45
Individual Z-score:	-1	-0.7	-0.3	+0.3	+1.7

Interpretation:

The first person is 1 standard deviation below the population mean.

The fourth person is 0.3 standard deviation above the population mean.

One of the Two Reasons to Use Z-Scores

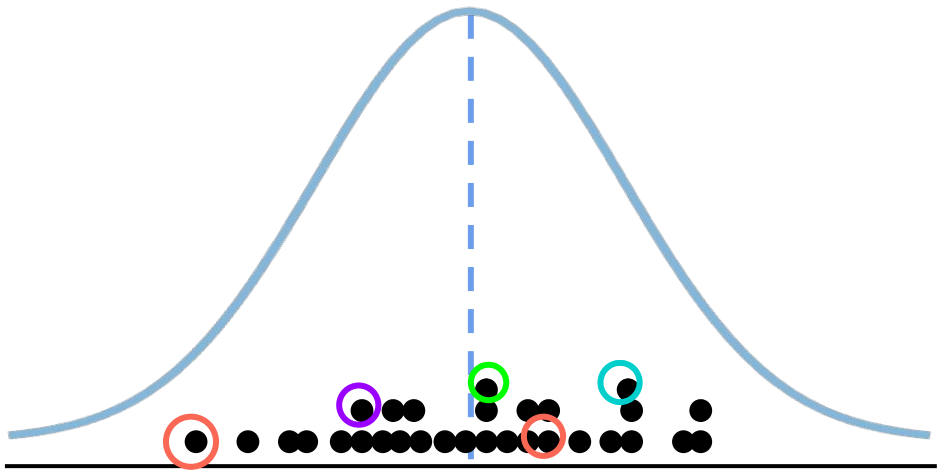
REASON 1

Z-scores can tell us a data point's **relative position** in a dataset by telling us **how many standard deviations above or below the mean** that data point is.

REASON 2

You will know soon! (:

Unlike measures of central tendency, **every** data point gets its own z-score. This is because a z score is not a summary of the whole collection, but rather a transformation of an individual score.



Unlike measures of central tendency, **every** data point gets its own z-score. This is because a z score is not a summary of the whole collection, but rather a transformation of an individual score.

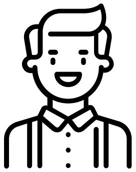
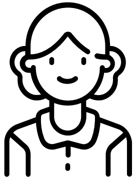
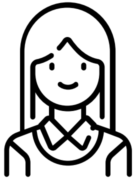
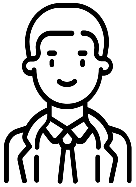
If we transform every single data point into their respective z-scores, then this is called <u>standardizing</u> our data.

For our practice above, we basically standardized our data.

Why Standardizing?

Imagine these two people took the **Beck Depression Inventory** (BDI)

Imagine these two people took the **Center for Epidemiologic Studies Depression Scale** (CES-D)

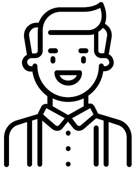
				
	Ben	Vanessa	Pooja	Hector
Raw Score	**25**	**5**	**10**	**50**

What could you conclude about each of these people? Who would you be most concerned about?

Why Standardizing?

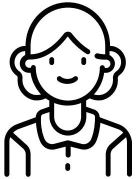
Imagine these two people took the **Beck Depression Inventory** (BDI)

Imagine these two people took the **Center for Epidemiologic Studies Depression Scale** (CES-D)



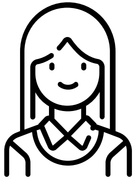
Ben

Z = +1



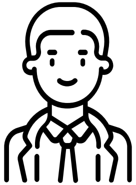
Vanessa

Z = -2.4



Pooja

Z = 0.0



Hector

Z = +2.1

What could you conclude about each of these people? Who would you be most concerned about?

Two Reasons to Use Z-Scores

REASON 1

Z-scores can tell us a data point's **relative position** in a dataset by telling us **how many standard deviations above or below the mean** that data point is.

REASON 2

Z-scores aid in data comprehension by **transforming all data to the <u>same</u> scale (standard deviations)**, which helps us compare things measured on different scales.

03

Empirical Rule & Probability

μ = 20 years
σ = 15 years

Actual Age:	5	10	15	25	45
Individual Z-score:	-1	-0.7	-0.3	+0.3	+1.7

When you say:

“The last person is 1.7 standard deviation above the population mean.”

Someone might ask you: “What’s the chance of that happening?”

μ = 20 years
σ = 15.8 years

Actual Age: 5 10 15 25 45

Individual Z-score:

Well, you would like to make a more convincing argument, right?

Then, we will need to talk about something called <u>probability</u>.

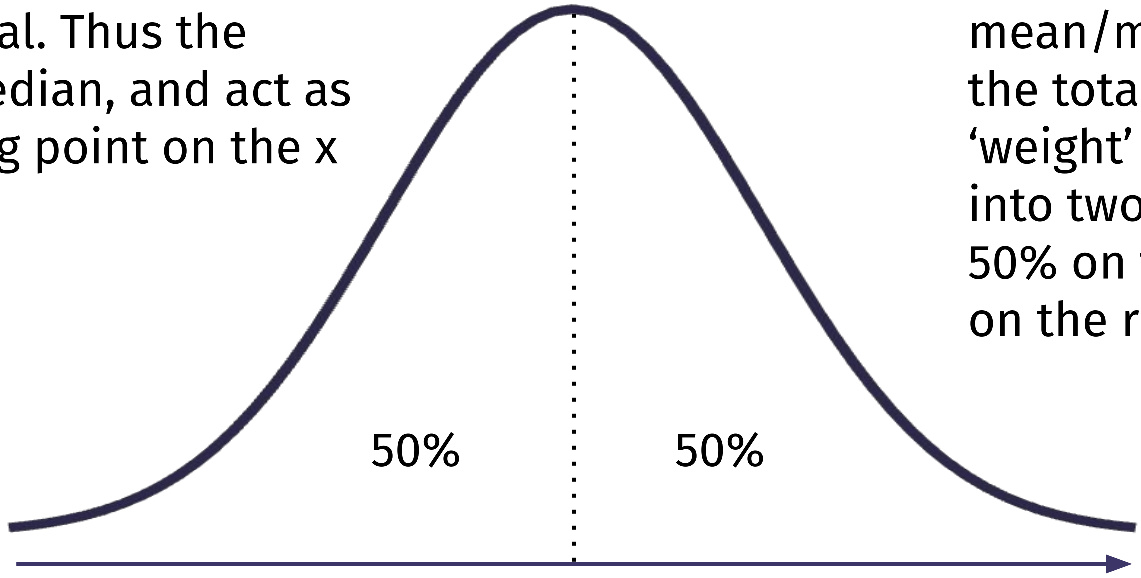
When you say:

"The last person is 1.7 standard deviation above the population mean."

Someone might ask you: "What's the chance of that happening?"

Remember our normal distribution?

We said before that this is symmetrical. Thus the mean = median, and act as a balancing point on the x axis.

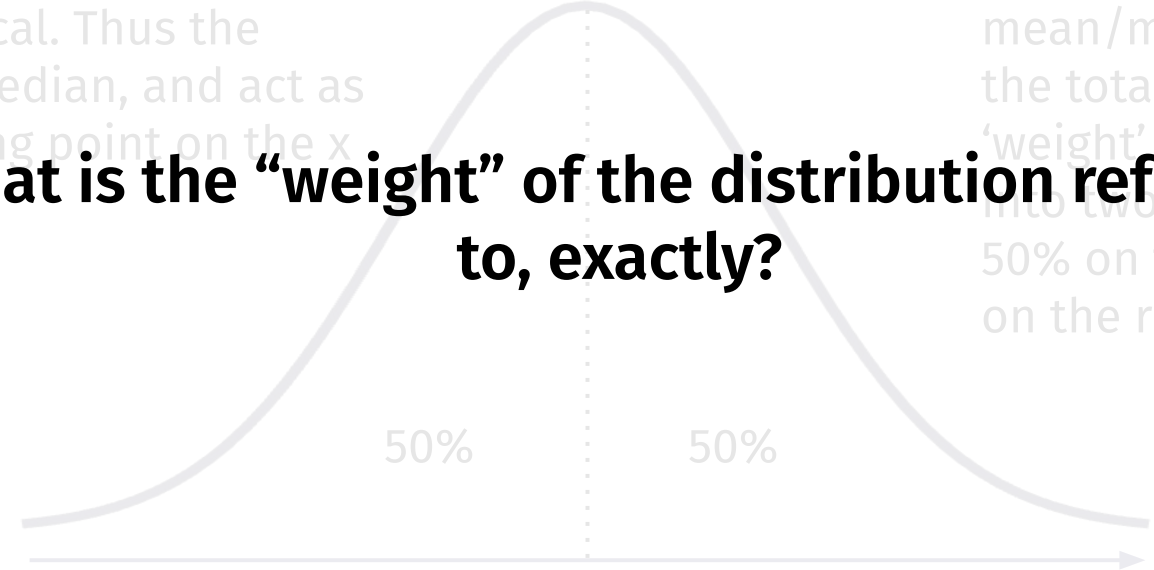


In this case, mean/median splits the total area, or the 'weight' of the curve, into two equal halves: 50% on the left, 50% on the right.

Remember our normal distribution?

We said before that this is symmetrical. Thus the mean = median, and act as a balancing point on the x axis.

In this case, mean/median splits the total area, or the 'weight' of the curve, into two equal halves: 50% on the left, 50% on the right.

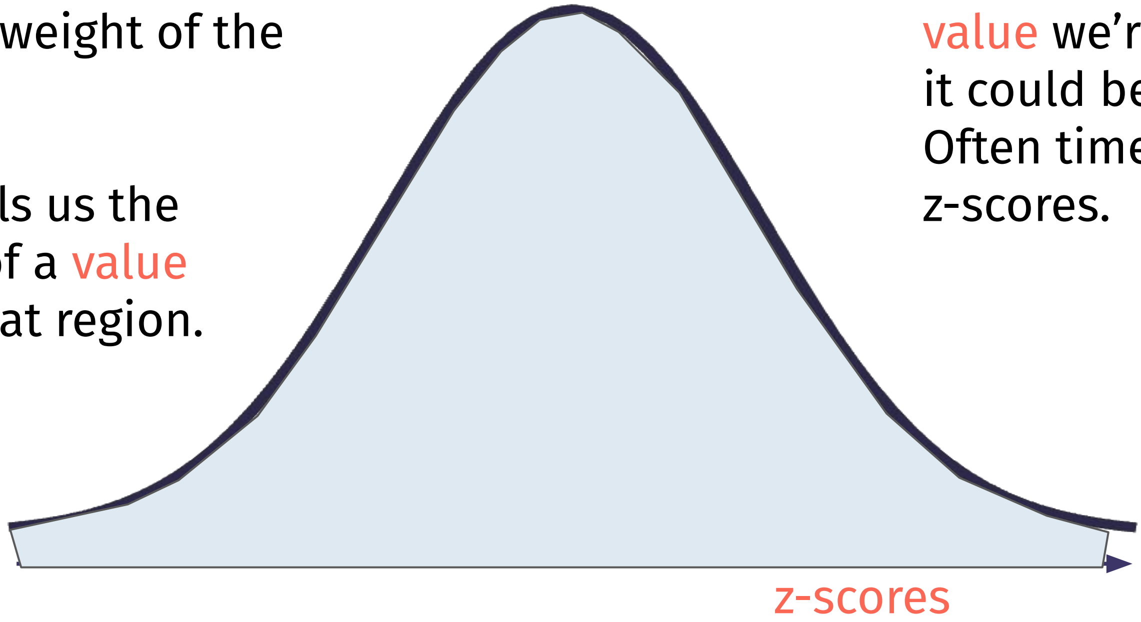


What is the "weight" of the distribution referring to, exactly?

Weight, Area Under the Curve, & Probability

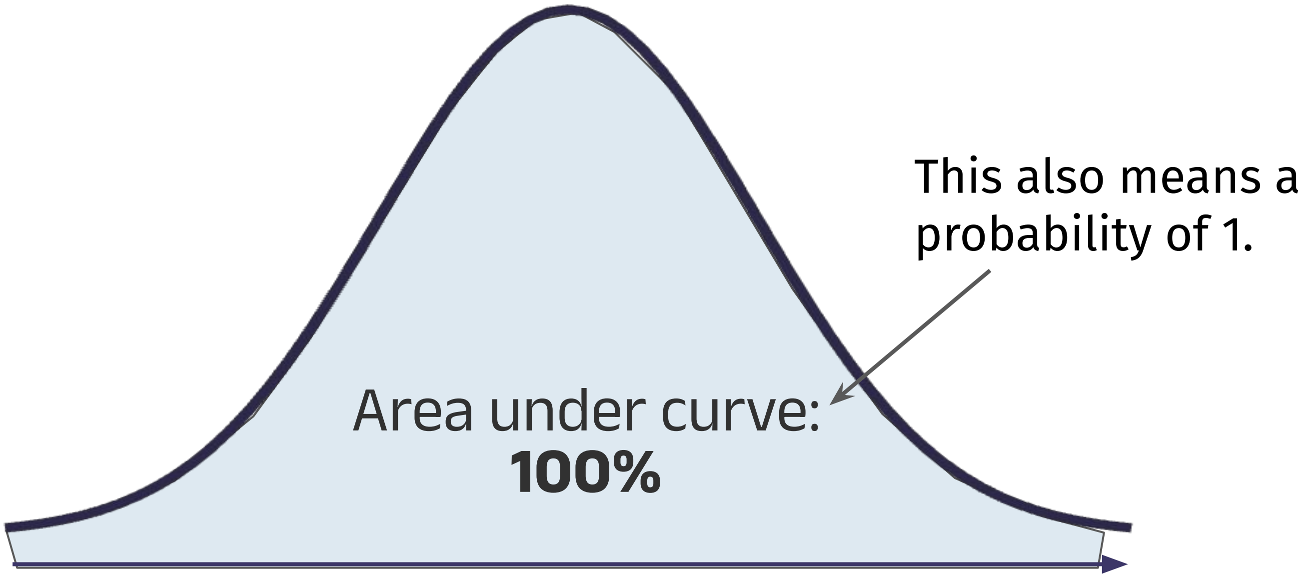
Think of the area under the curve as the weight of the distribution.

That area tells us the **probability** of a value landing in that region.



The x-axis shows what value we're looking at—it could be any number. Often times they are z-scores.

All possible values together cover everything that can happen.



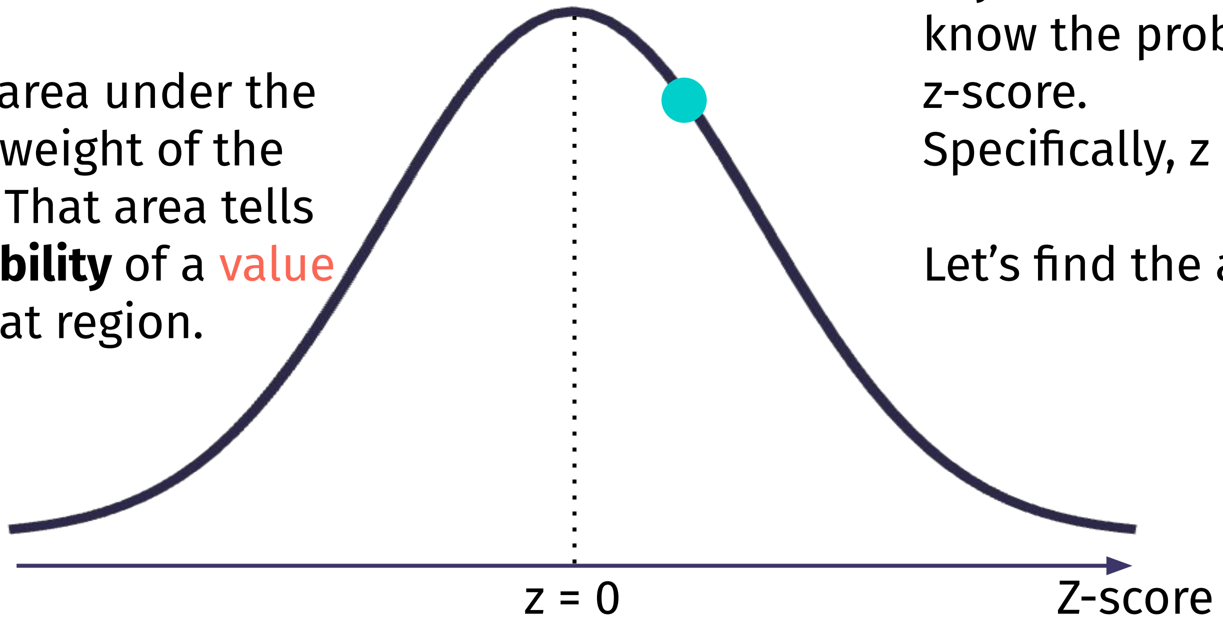
All possible values together cover everything that can happen.

Let’s look at an example!

The Normal Distribution

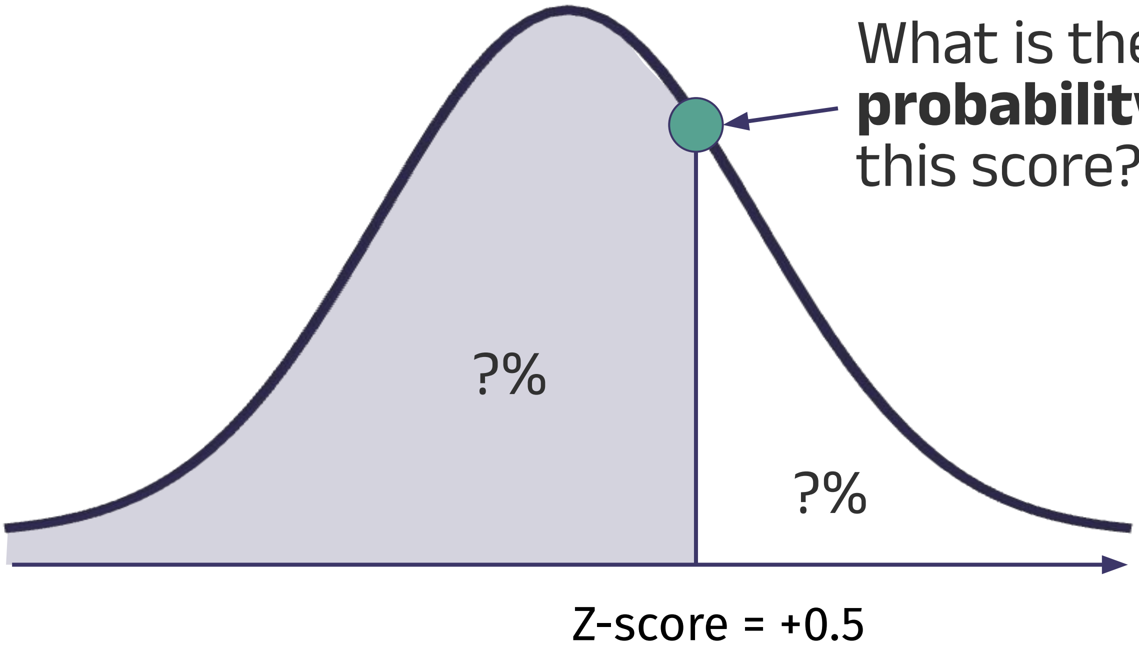
What we said earlier:

Think of the area under the curve as the weight of the distribution. That area tells us the **probability** of a value landing in that region.



Say the value we want to know the probability of is a z-score.
Specifically, $z = +0.5$.

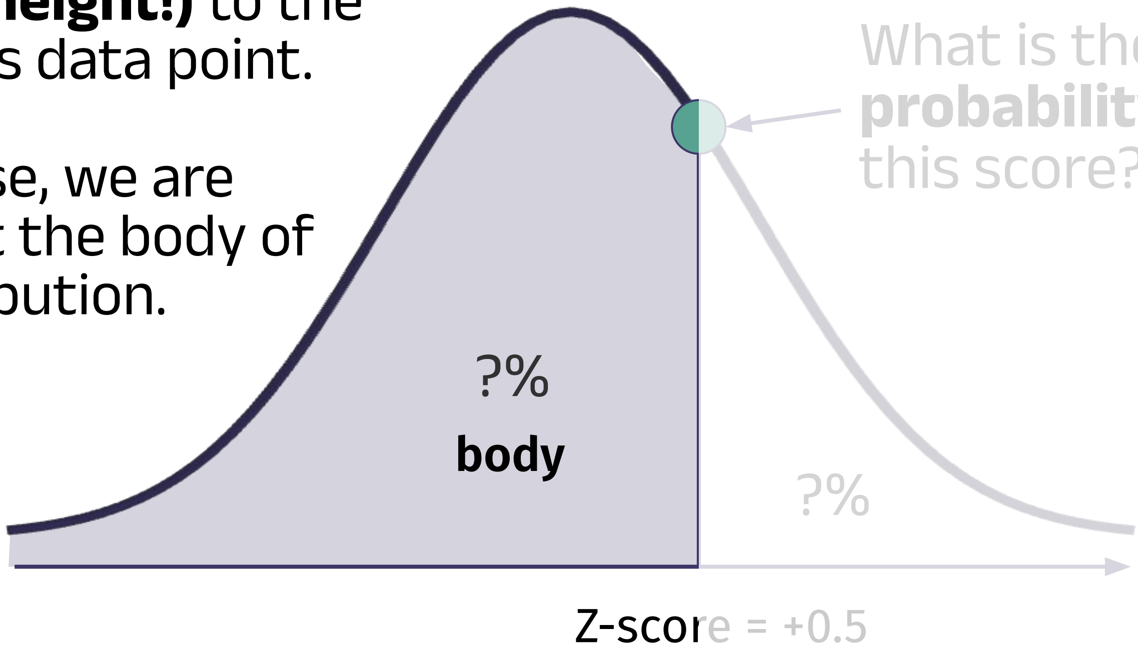
Let's find the area!



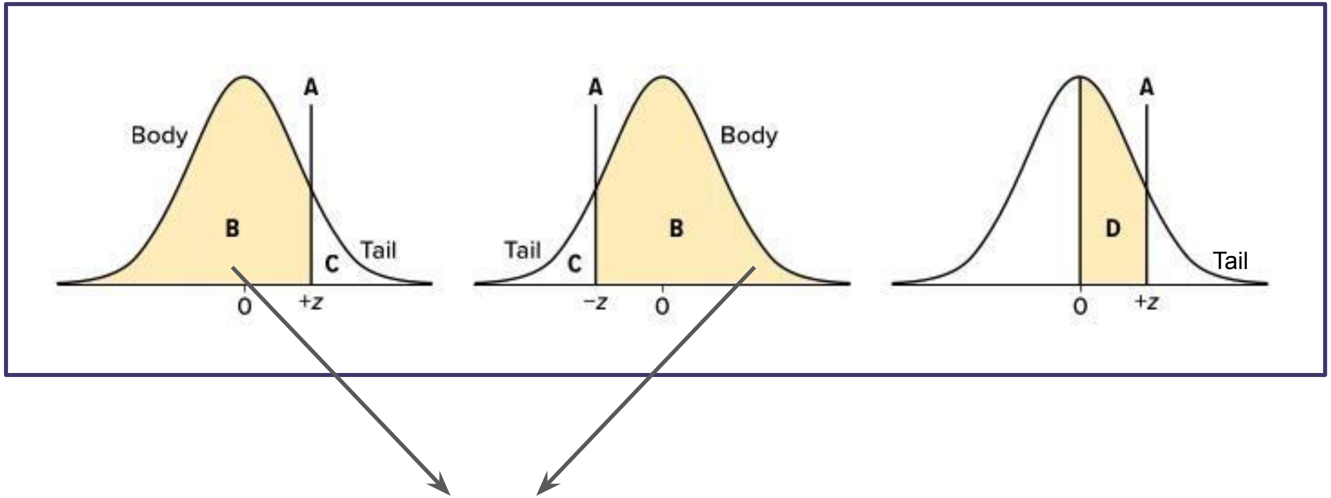
What is the exact **probability** of getting this score?

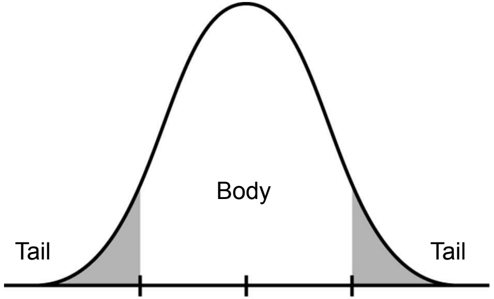
You will find the **area (not the height!)** to the left of this data point.

In this case, we are looking at the body of the distribution.



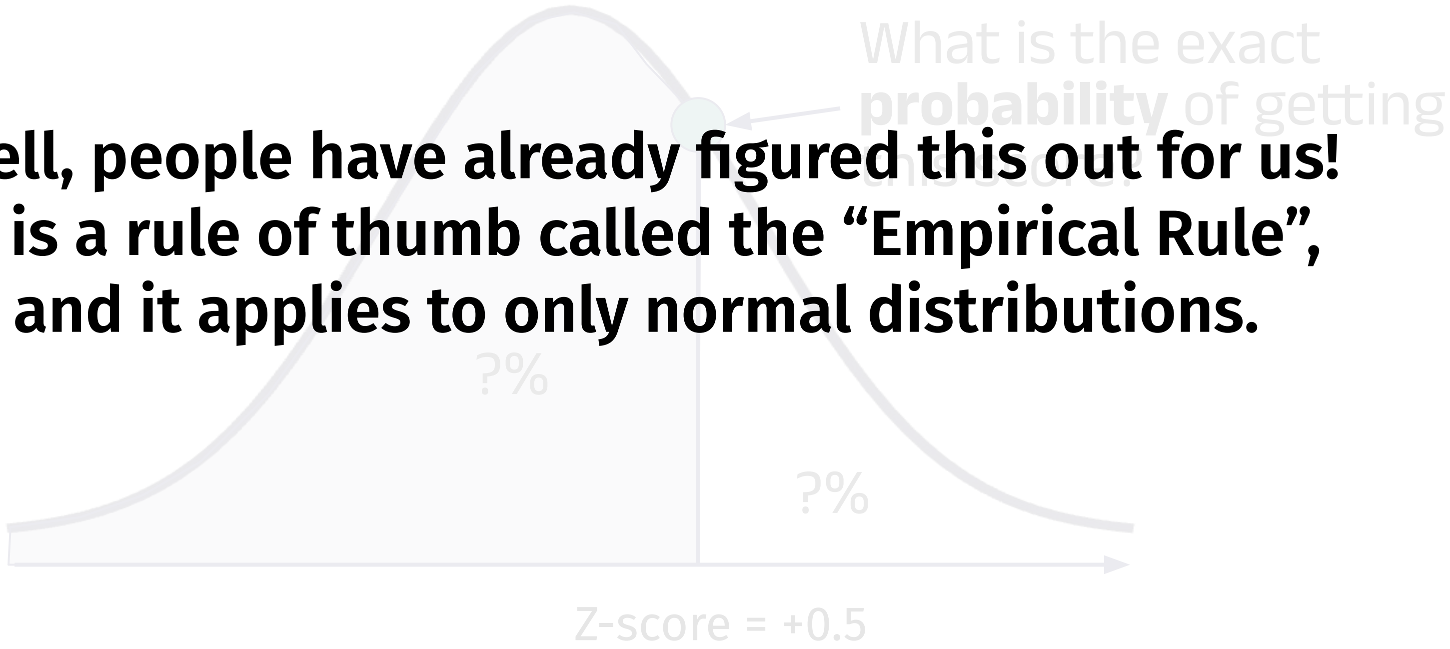
What is body and what is tail?





The area that takes up most of the curve is called the **body**; the rest is called the **tail(s)**.

Well, people have already figured this out for us!
It is a rule of thumb called the “Empirical Rule”,
and it applies to only normal distributions.



Empirical Rule (“68-95-99.7 Rule”)

~68%

of the scores will fall between **-1 and +1 SDs** of the mean.

~95%

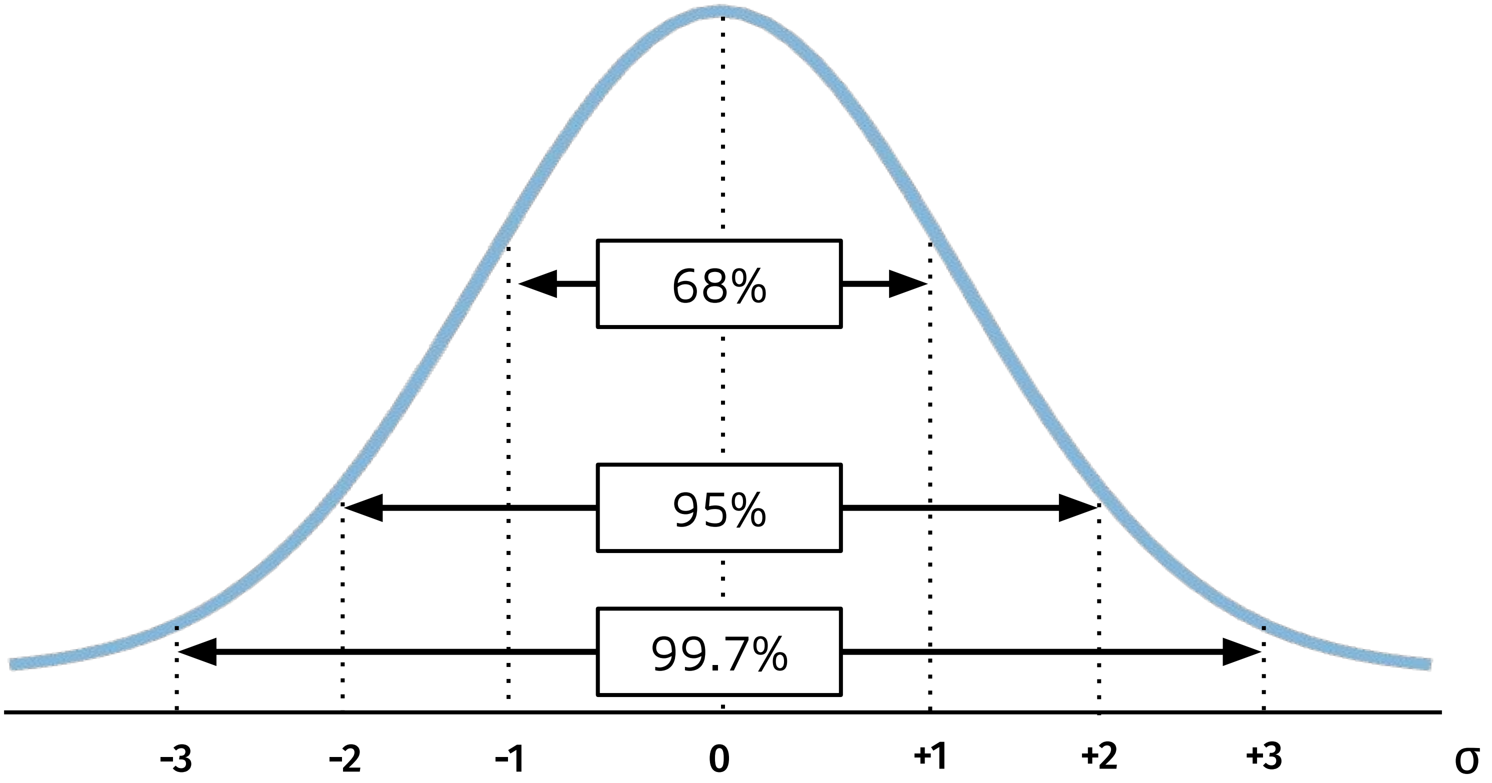
of the scores will fall between **-2 and +2 SDs** of the mean.

~99.7%

of the scores will fall between **-3 and +3 SDs** of the mean.

Note: This rule only applies to data that are ***normally distributed***.

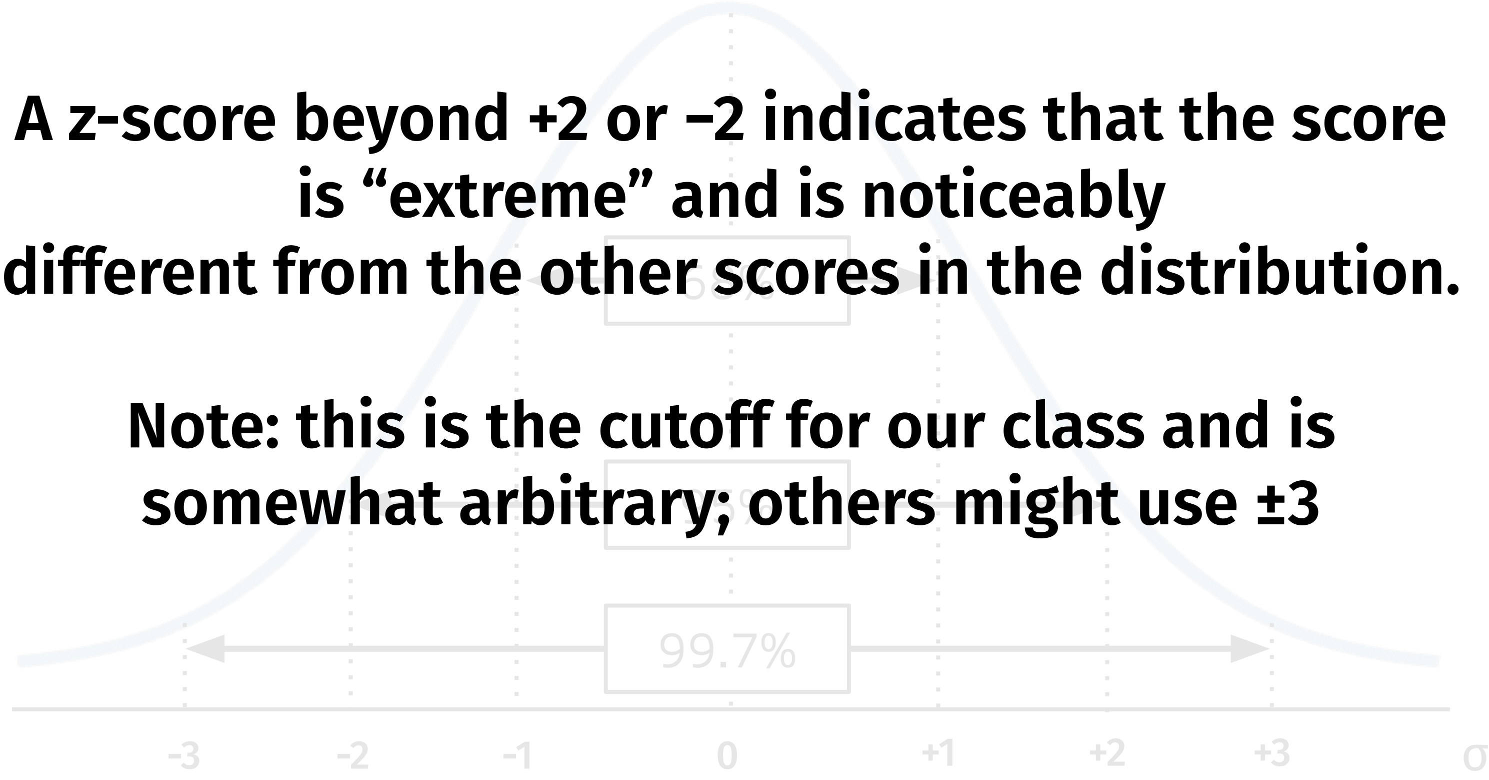
Empirical Rule ("68-95-99.7 Rule")



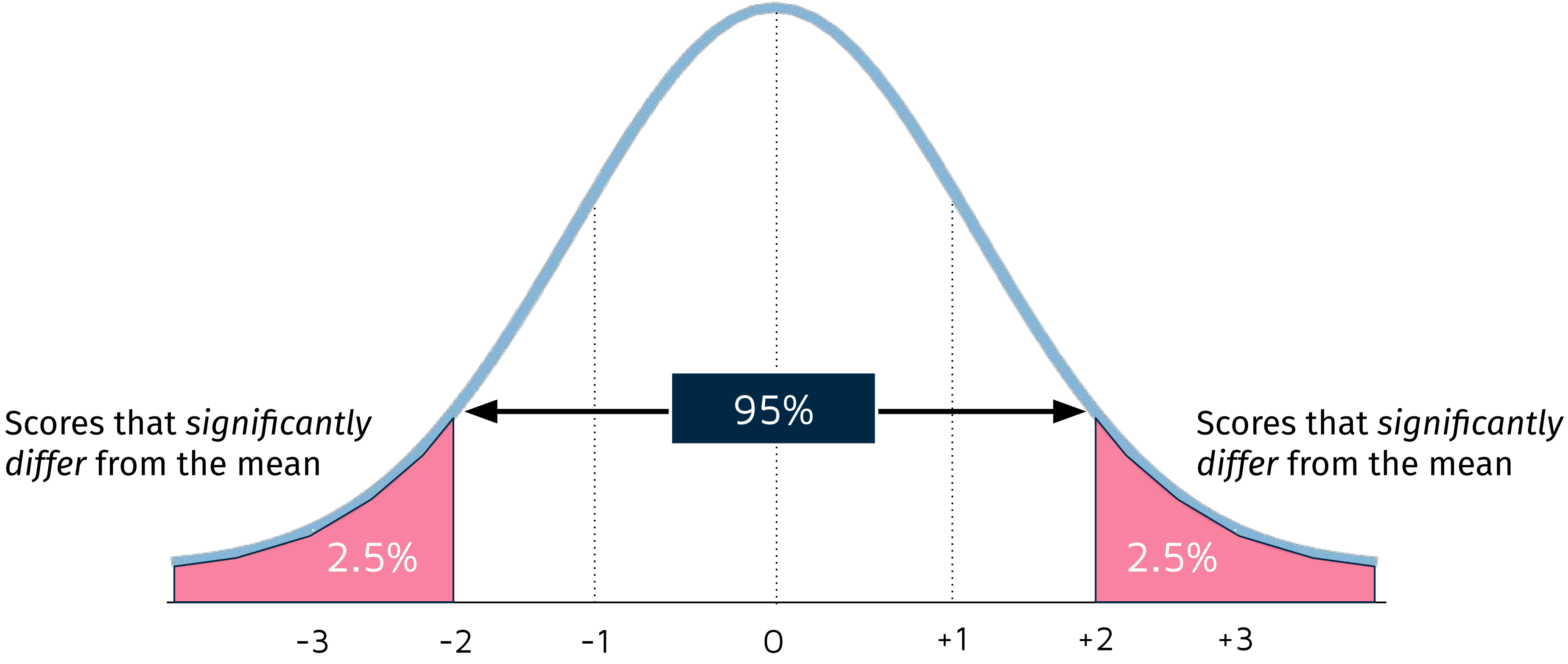
Empirical Rule ("68–95–99.7 Rule")

A z-score beyond +2 or −2 indicates that the score is "extreme" and is noticeably different from the other scores in the distribution.

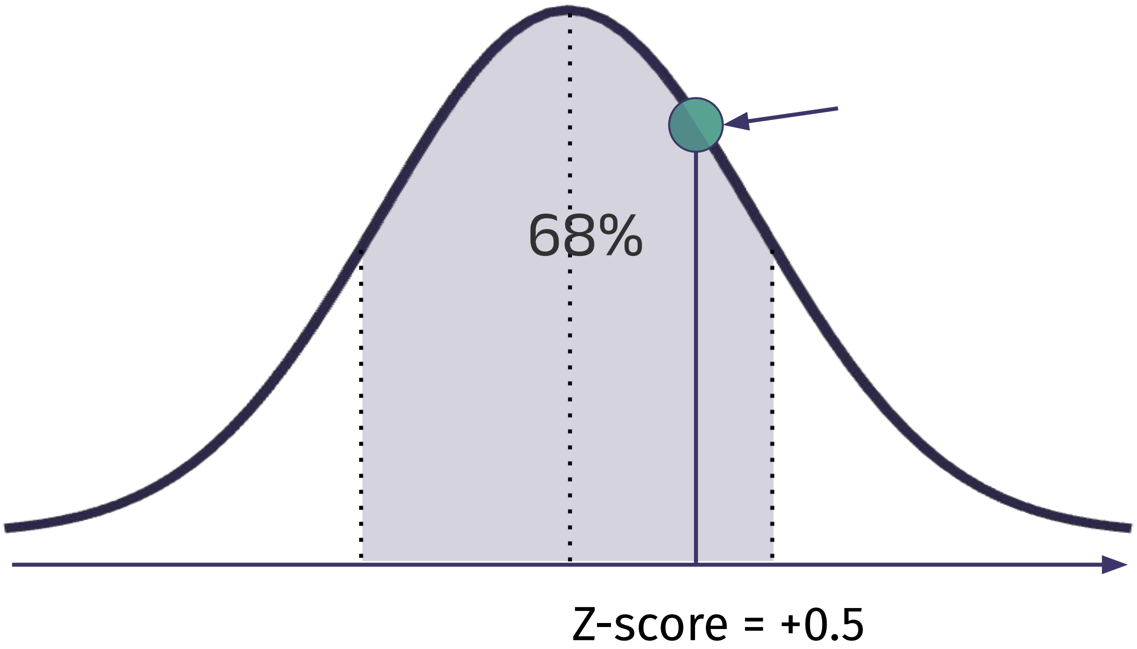
Note: this is the cutoff for our class and is somewhat arbitrary; others might use ±3



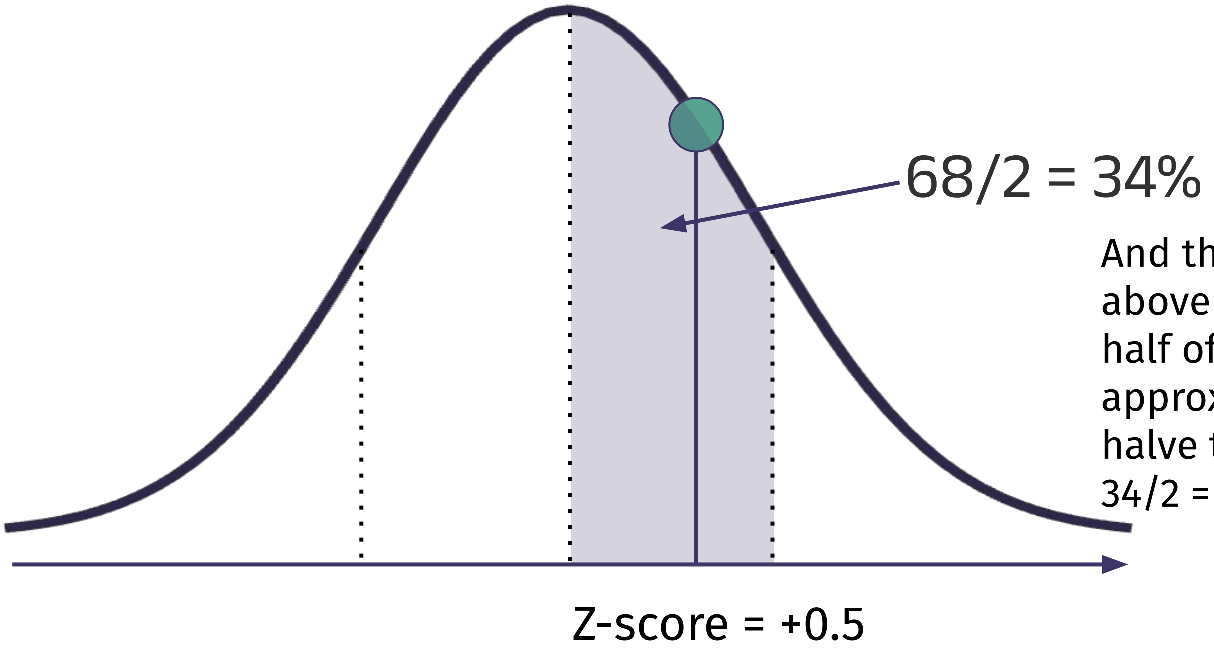
“Extreme” Scores



Answering our question using the Empirical Rule



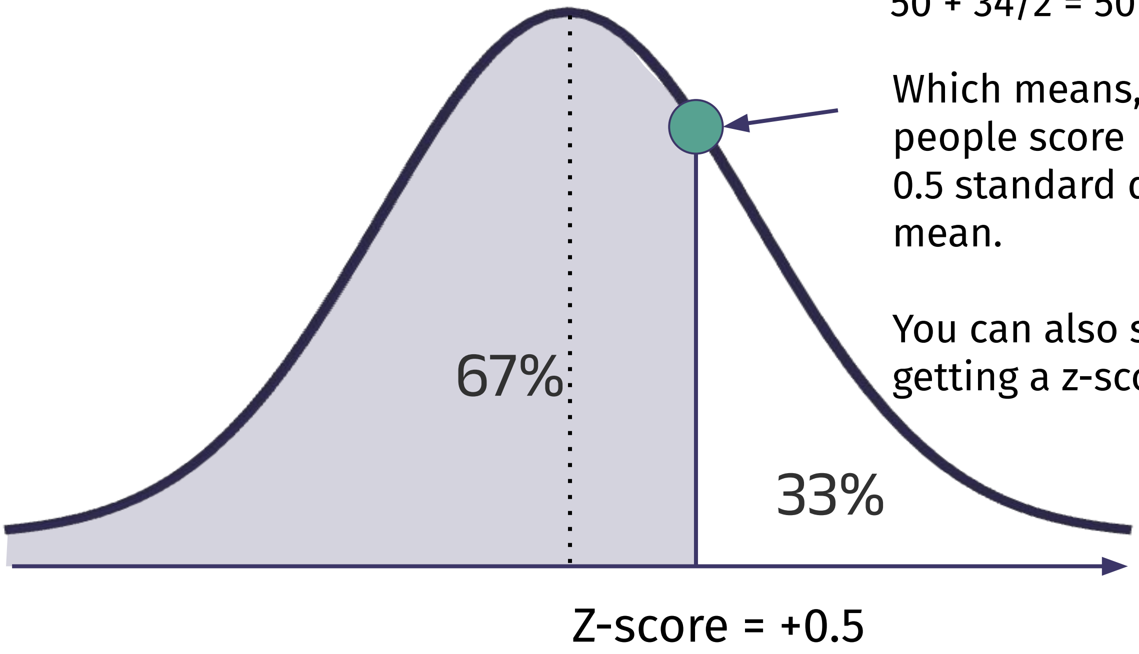
Answering our question using the Empirical Rule



And that's for 1 SD above the mean. +0.5 is half of that. So to approximate that, we halve this area again, 34/2 = 17%

Answering our question using the Empirical Rule

So, it will be approximately:
50 + 34/2 = 50 + 17 = 67%



Which means, approximately 67% of people score below a value that is 0.5 standard deviations above the mean.

You can also say: the probability of getting a z-score below 0.5 is 0.67.

We can also use the Z-table instead of the Empirical Rule

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
...	...	...	...
0.48	.6844	.3156	.0040
0.49	.6879	.3121	.0080
0.50	.6915	.3085	.0120
0.51	.6950	.3050	.0160
0.52	6985	.3015	.0199
...	...	...	...

Our estimated value is 0.67 for the body!

[Textbook Z-table]()

We can also use the Z-table instead of the Empirical Rule

That is pretty close to our approximation using the empirical rule!

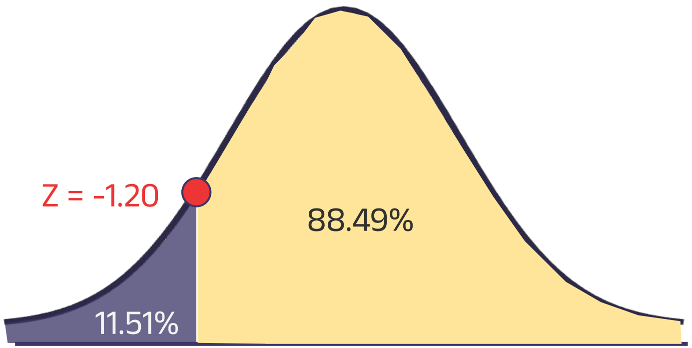
You can use either to figure out the proportion in body/tails. However, the z-table is more flexible, it has the probability that corresponds to A LOT of z-values.

(A)	(B)	(C)	(D)
[illegible]	[illegible]	[illegible]	[illegible]
...	...	...	...
0.48	.6844	.3156	.0040
0.49	.6879	.3121	.0080
[illegible]	[illegible]	[illegible]	[illegible]
0.51	.6950	[illegible]	.0160
0.52	6985	.3015	.0199
...	...	...	...

Our estimated value is 0.67 for the body!

Textbook Z-table

One last note



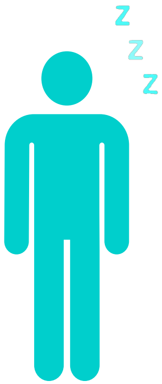
(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
...	...	...	...
1.19	.8830	.1170	.3830
1.20	**.8849**	**.1151**	**.3849**
1.21	.8869	.1131	.3869
...	...	...	...

Despite z = -1.2, you can look up the absolute value, aka z = 1.2. All you need to do is to draw it, flip the body & tail!

IN CLASS ACTIVITY (ICA 5)

HOW MUCH SLEEP DID YOU GET LAST NIGHT?

1. Record each person's **score** (X) in your group.
2. **Calculate a z-score** for each person at your table.
3. Draw a normal distribution and label your table's z-scores on it.
4. Complete **ICA 5** on Blackboard.



$\mu = 8$

$\sigma = 1$

https://umsystem.pressbooks.pub/isps/back-matter/appendix-a/

04

Wrap Up

Please complete this survey about Exam 1 Review Planning by Thursday!

https://forms.gle/ppxcCFfrTaWXDXsR9