Correlation

Correlation

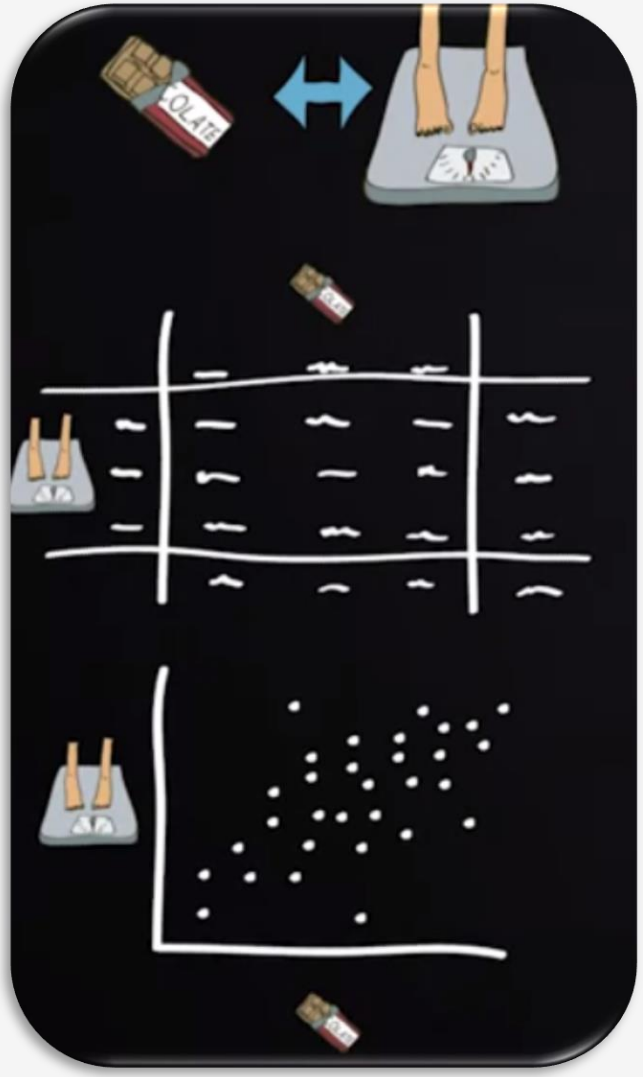
In this section we’ll discuss the concept of correlation. We'll talk about how we can display the correlation between two variables using tables and graphs. First we'll look at categorical variables and discuss **contingency tables**. In a next step we look at how we can best display the relationship between two quantitative variables. Here we'll introduce the **scatterplot**.

In the second section we'll discuss the **Pearson's** r - one of the most frequently used measures of correlation. It is an appropriate measure if the variables under analysis are measured on a quantitative level and if they are linearly related to each other. The Pearson's r expresses the direction and strength of the correlation. We'll learn how to interpret the Pearson's r and how to compute it ourself.

Correlation

Many people like eating chocolate. Yet most people are somewhat cautious with their chocolate consumption, because it might well be the case that eating a lot of chocolate increases your weight.

In this section, we'll talk about how we can display a relationship between two variables using tables and graphs. This can be very useful to help you discover if two variables are correlated or not.

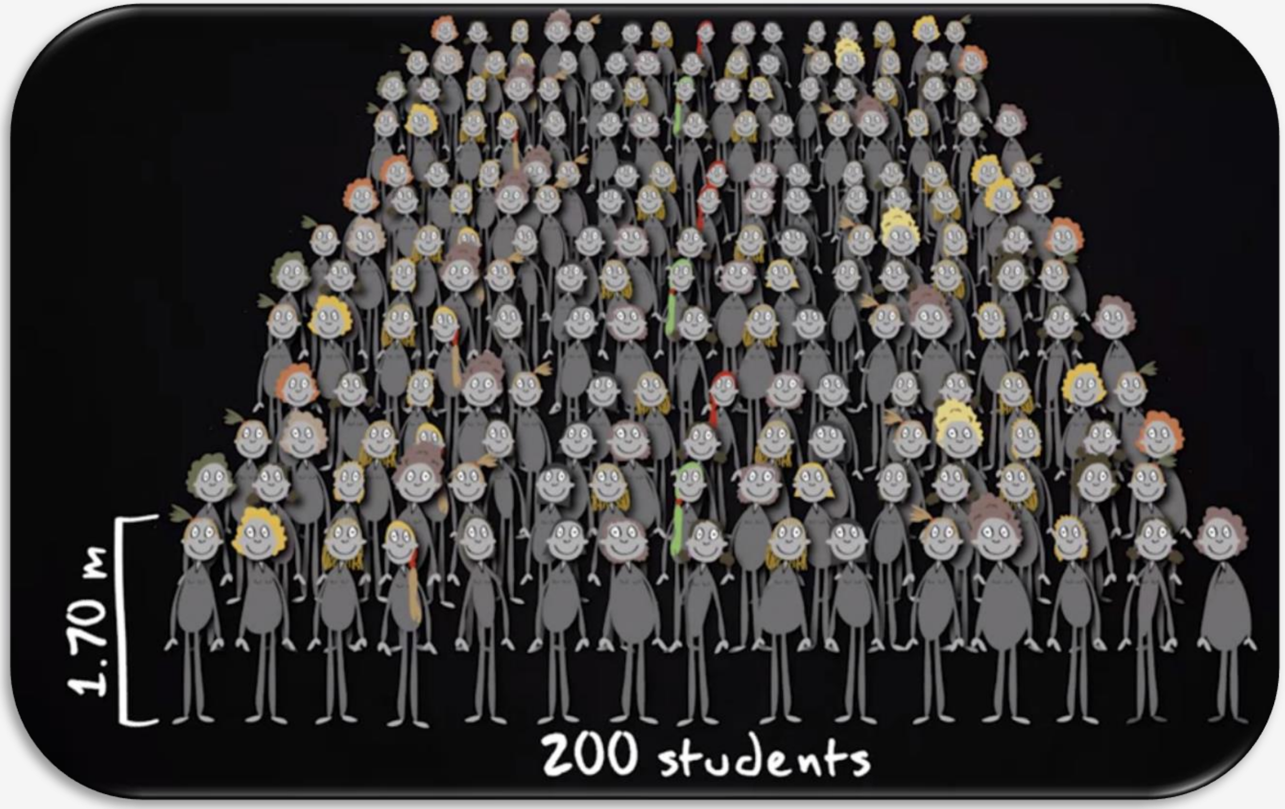


Correlation

Let us investigate the relationship between eating chocolate and body weight further. Suppose I have selected 200 female students at my university who are all 1m70 tall.

This way height is a constant and cannot account for differences in body weight (or chocolate consumption).

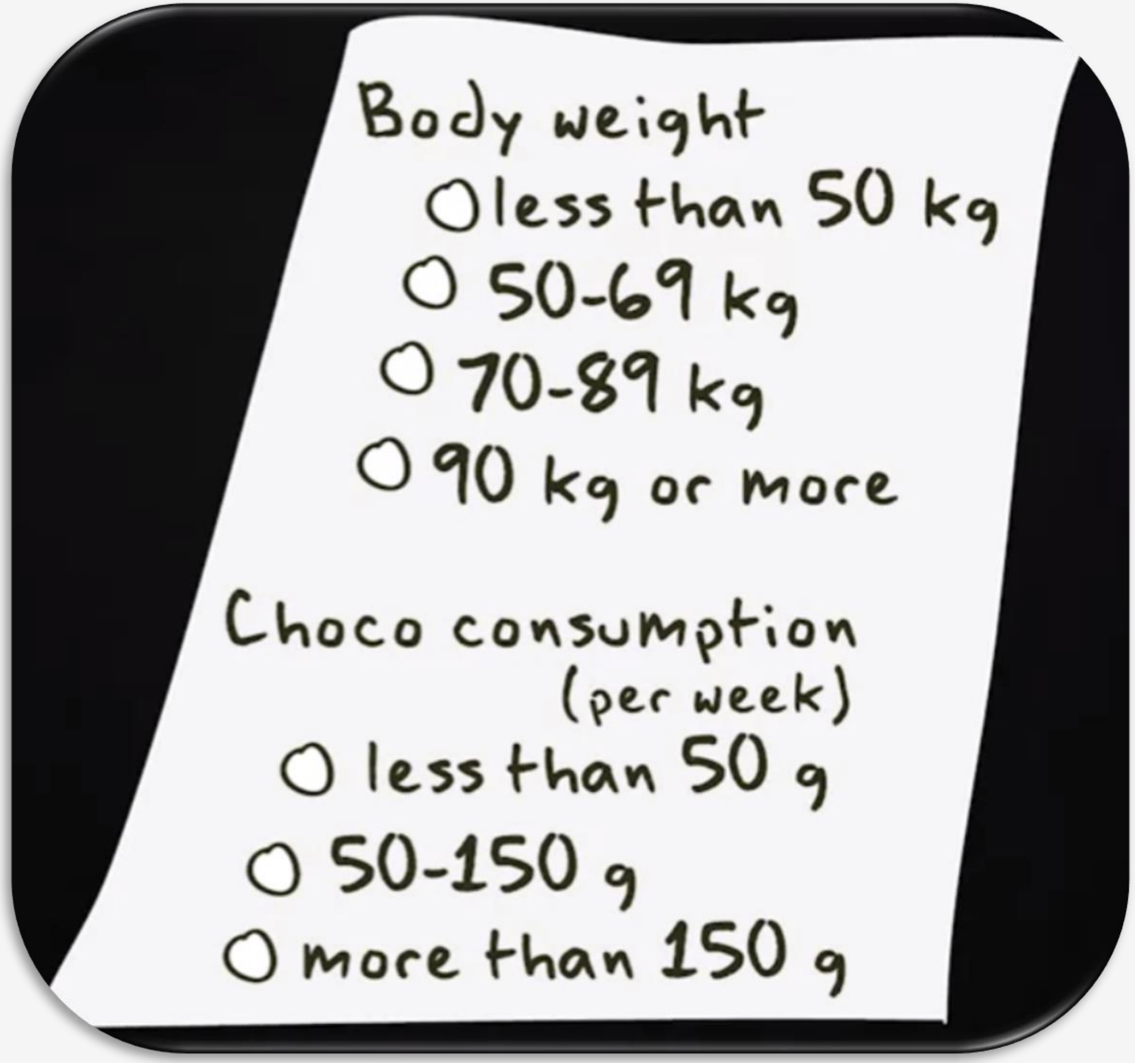
I asked the students to report their body weight and their weekly chocolate consumption



Correlation

They could choose between the categories ‘less than 50 kilograms’, ‘50 to 69 kilograms’, ‘70 to 89 kilograms’ and ’90 kilograms or more’.

They could indicate their chocolate consumption by choosing ‘less than 50 grams per week’, ‘between 50 and 150 grams per week’, and ‘more than 150 grams per week’.



Body weight

○ less than 50 kg
○ 50-69 kg
○ 70-89 kg
○ 90 kg or more

Choco consumption
(per week)

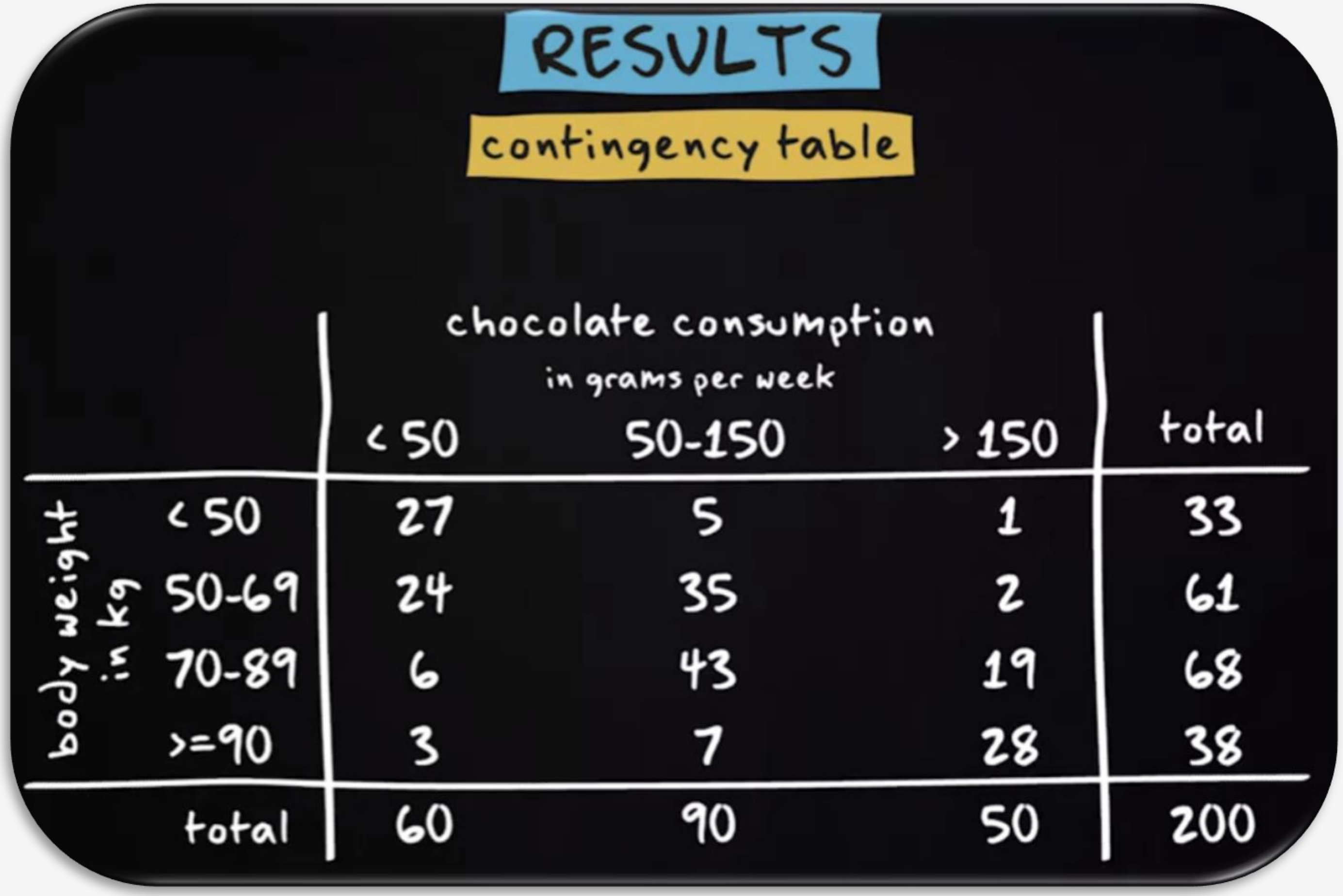
○ less than 50 g
○ 50-150 g
○ more than 150 g

Correlation

Here are the results. What you see here is a **contingency table**.

A contingency table enables you to display the relationship between two ordinal or nominal variables.

It is similar to a frequency table. But the major difference is that a frequency table always concerns only one variable, whereas a contingency table concerns two variables.



RESULTS

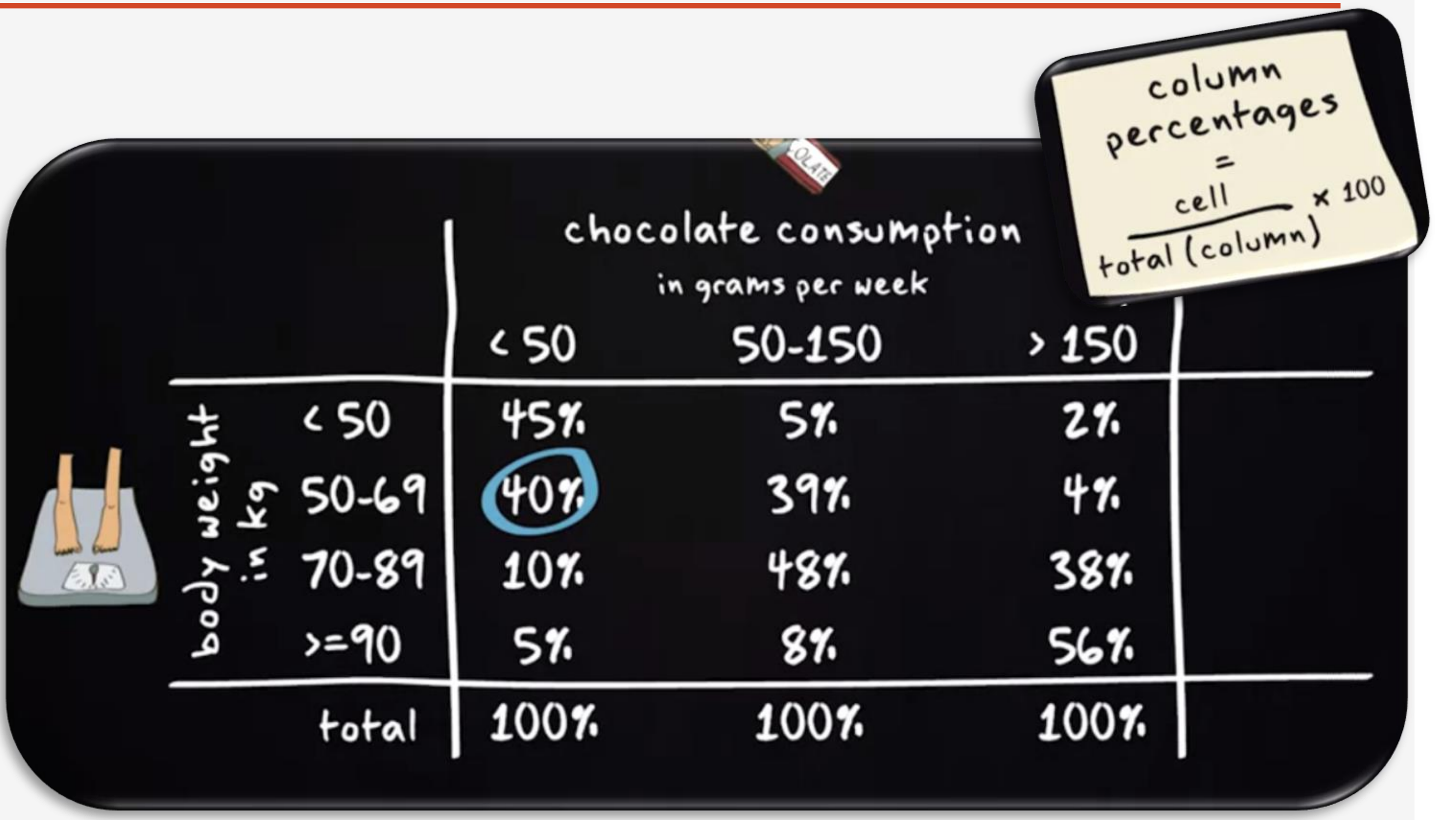
contingency table

body weight in kg	chocolate consumption in grams per week			
	< 50	50-150	> 150	total
< 50	27	5	1	33
50-69	24	35	2	61
70-89	6	43	19	68
>=90	3	7	28	38
total	60	90	50	200

Correlation

In previous table form does not tell you much yet about the correlation between the two variables, because the columns and rows contain different numbers of cases.

It provides more insight when you compute percentages. In this case we compute column percentages.



		chocolate consumption in grams per week		
		< 50	50-150	> 150
body weight in kg	< 50	45%	5%	2%
	50-69	40%	39%	4%
	70-89	10%	48%	38%
	>=90	5%	8%	56%
	total	100%	100%	100%

column percentages

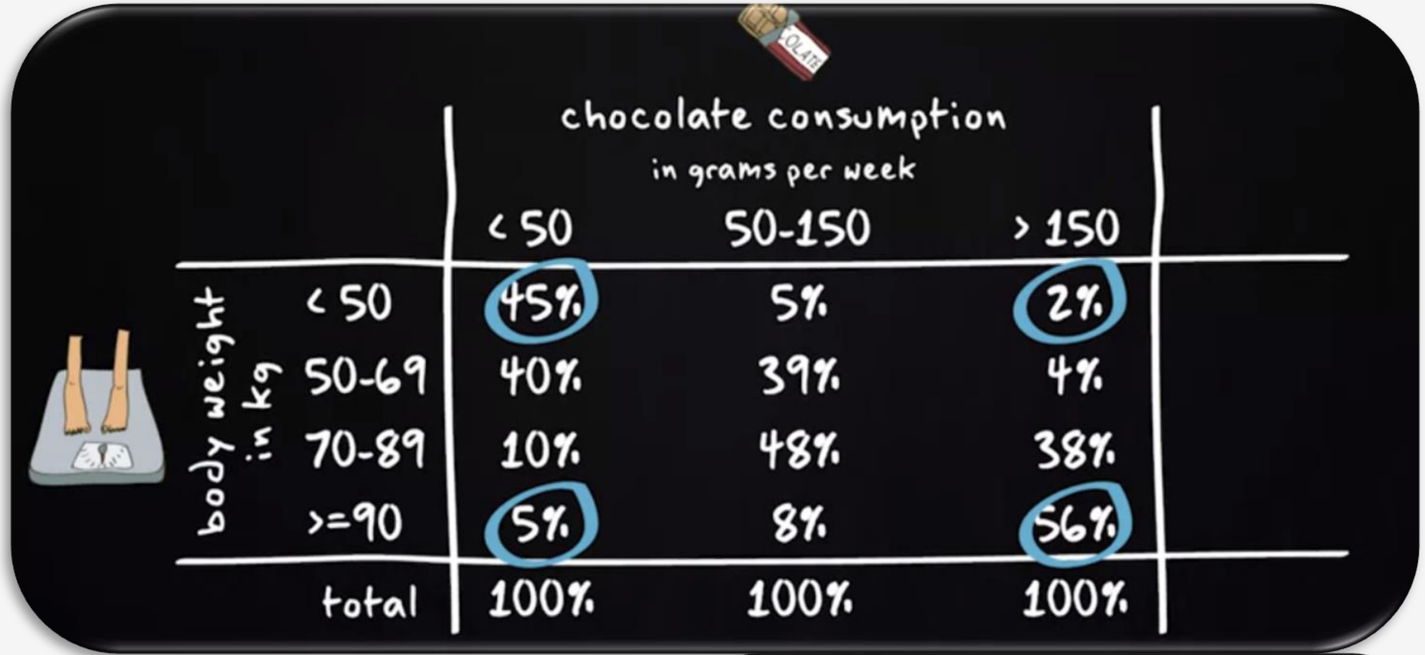
$$= \frac{\text{cell}}{\text{total (column)}} \times 100$$

Correlation

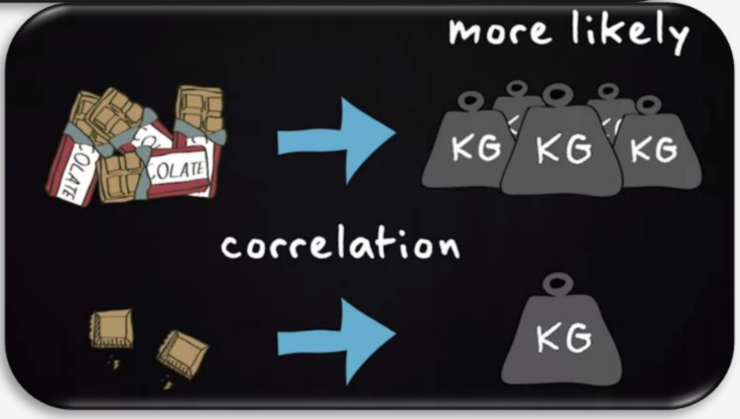
So what does this mean?

Of those who eat more than 150 grams of chocolate per week, 56 percent weighs 90 kilograms or more. Of those who eat less than 50 grams of chocolate, only 5 percent weighs 90 kilograms or more.

Also, of those who eat less than 50 grams of chocolate, 45 percent weighs less than 50 kilograms, while of those who eat more than 150 grams of chocolate only 2 percent weighs less than 50 kilograms.



		chocolate consumption in grams per week		
		< 50	50-150	> 150
body weight in kg	< 50	45%	5%	2%
	50-69	40%	39%	4%
	70-89	10%	48%	38%
	>=90	5%	8%	56%
	total	100%	100%	100%



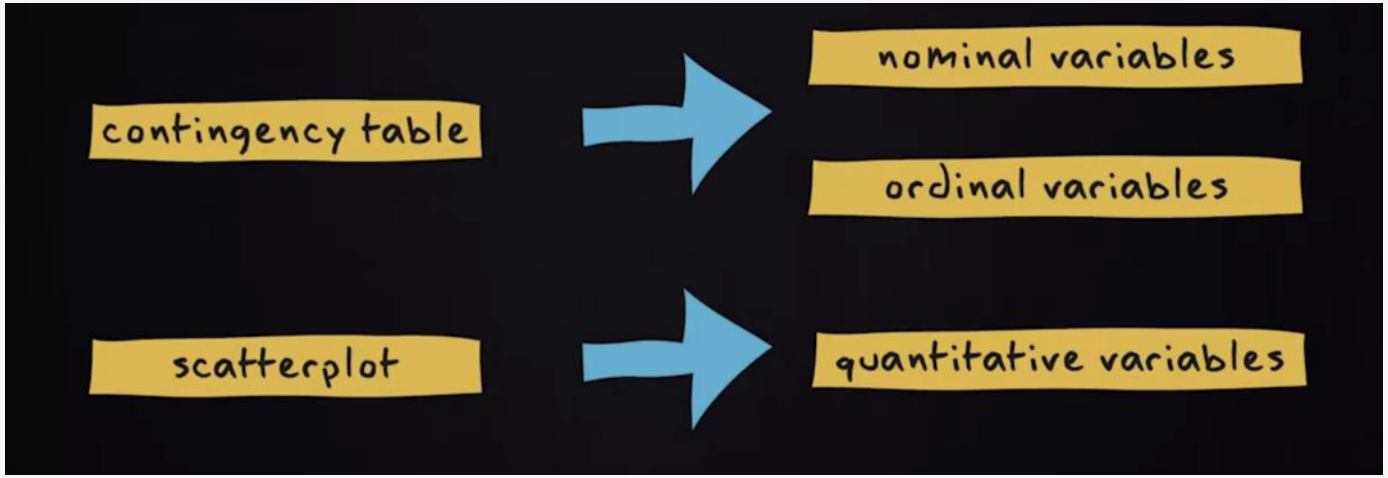
The percentages show that there is a correlation between chocolate consumption and body weight.

Correlation

A contingency table is useful for nominal and ordinal variables, but not for quantitative variables.

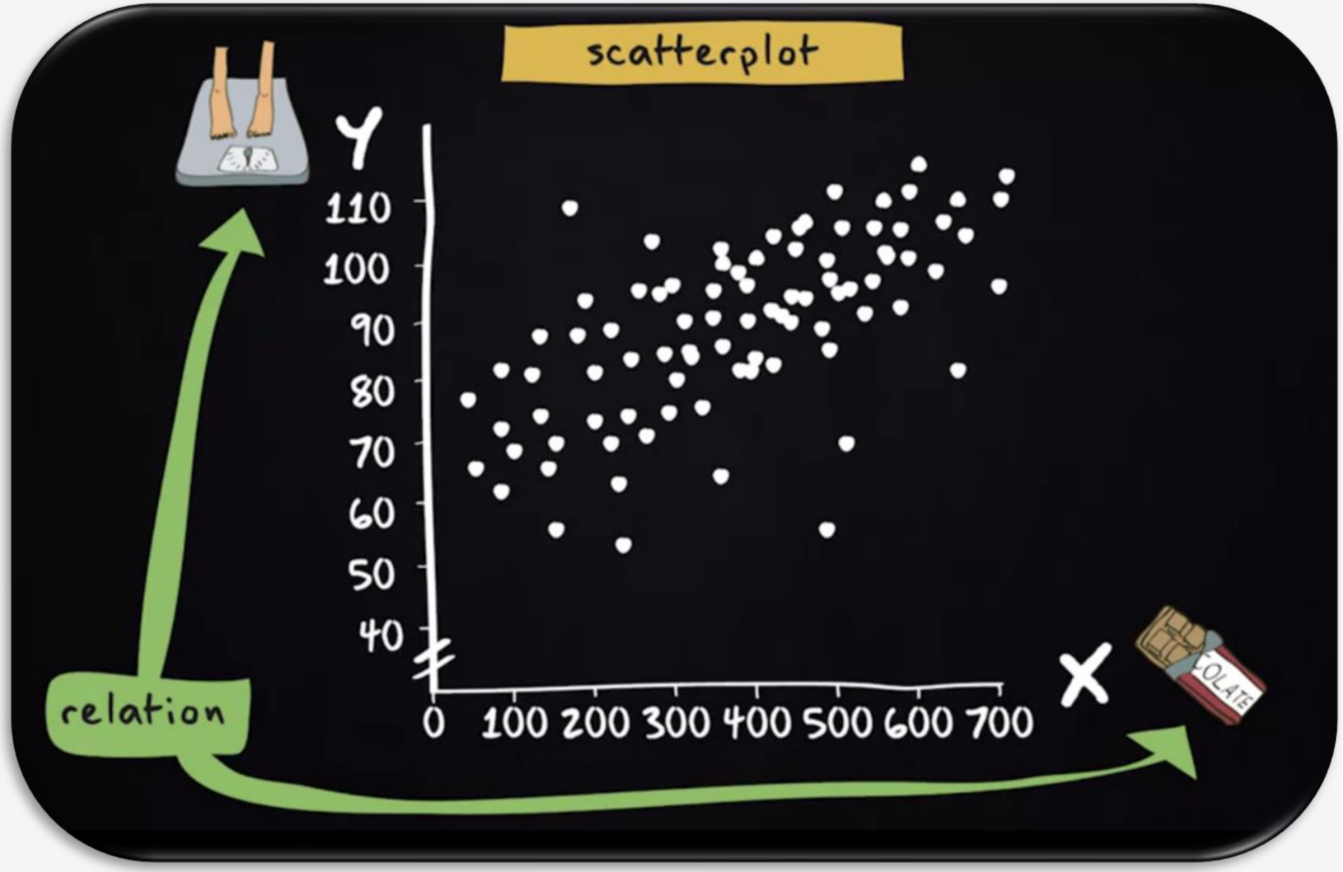
For quantitative variables a scatterplot is more appropriate.

Suppose that instead of providing categories, I asked the 200 women to give me their exact body weight; for instance 65 or 72 kilograms. Suppose I also asked them to tell me how much chocolate they eat every week. That could be, for instance, 64, or 99 grams per week. Now I have much more precise information than before. The best way to display the relationship between the quantitative variables chocolate consumption and weight is with a scatterplot.


contingency table
nominal variables
ordinal variables
scatterplot
quantitative variables

Correlation

To make a scatterplot we draw two lines, which we call axes. We call the horizontal axis the x-axis; here we display the independent variable. The vertical axis is called the y-axis, which we use to represent the dependent variable.



Correlation

The scatterplot shows at a glance that there is a strong correlation between the two variables: the more chocolate someone eats, the larger her body weight.

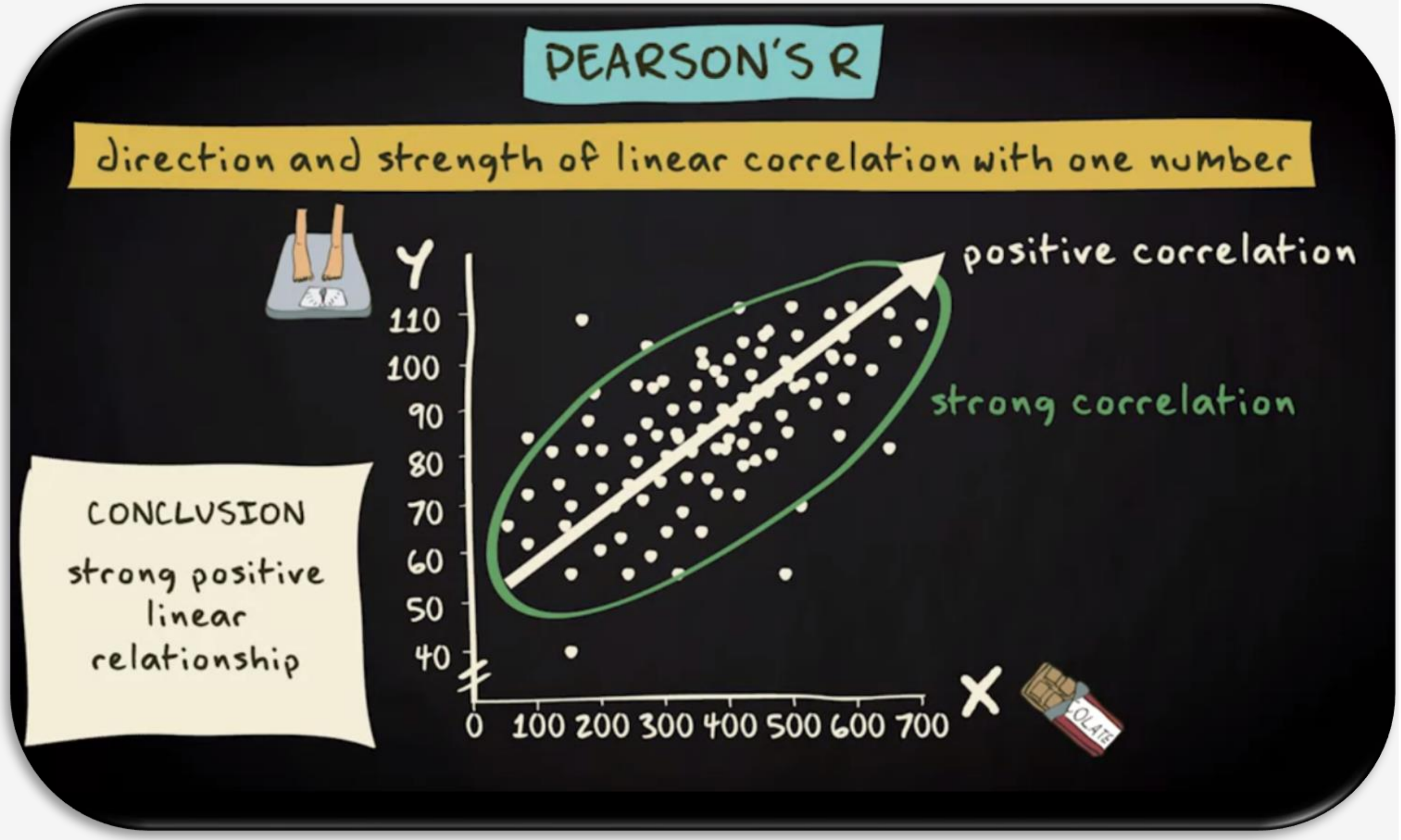
But how strong is this correlation?

We will now turn to one of the most often used measures of correlation: **Pearson's r**.

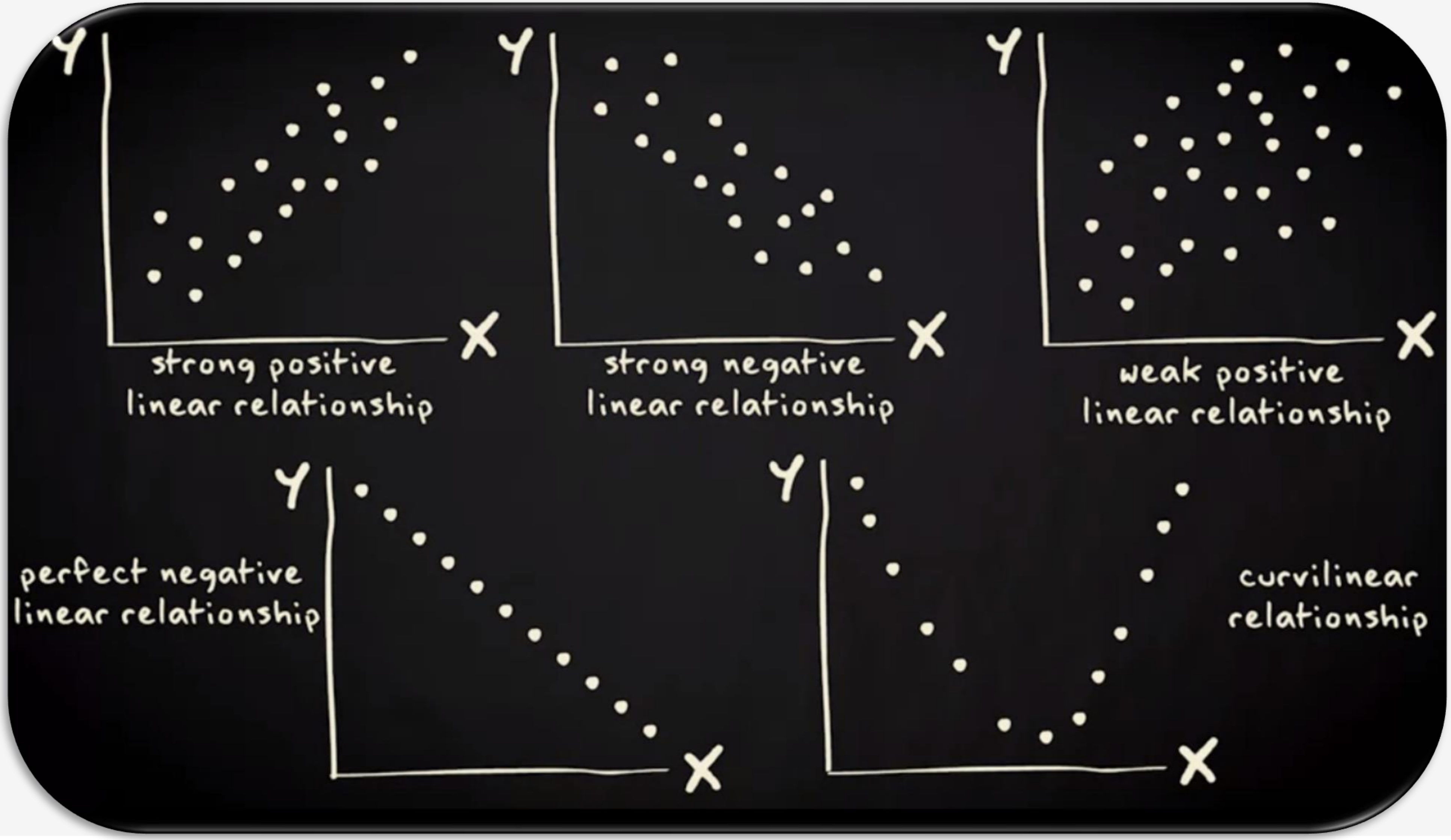
Correlation

One of the most important advantages of Pearson's r is that it expresses the direction and strength of the linear correlation between two variables with a **single number**.

However, variables could also be correlated in different ways.



Correlation

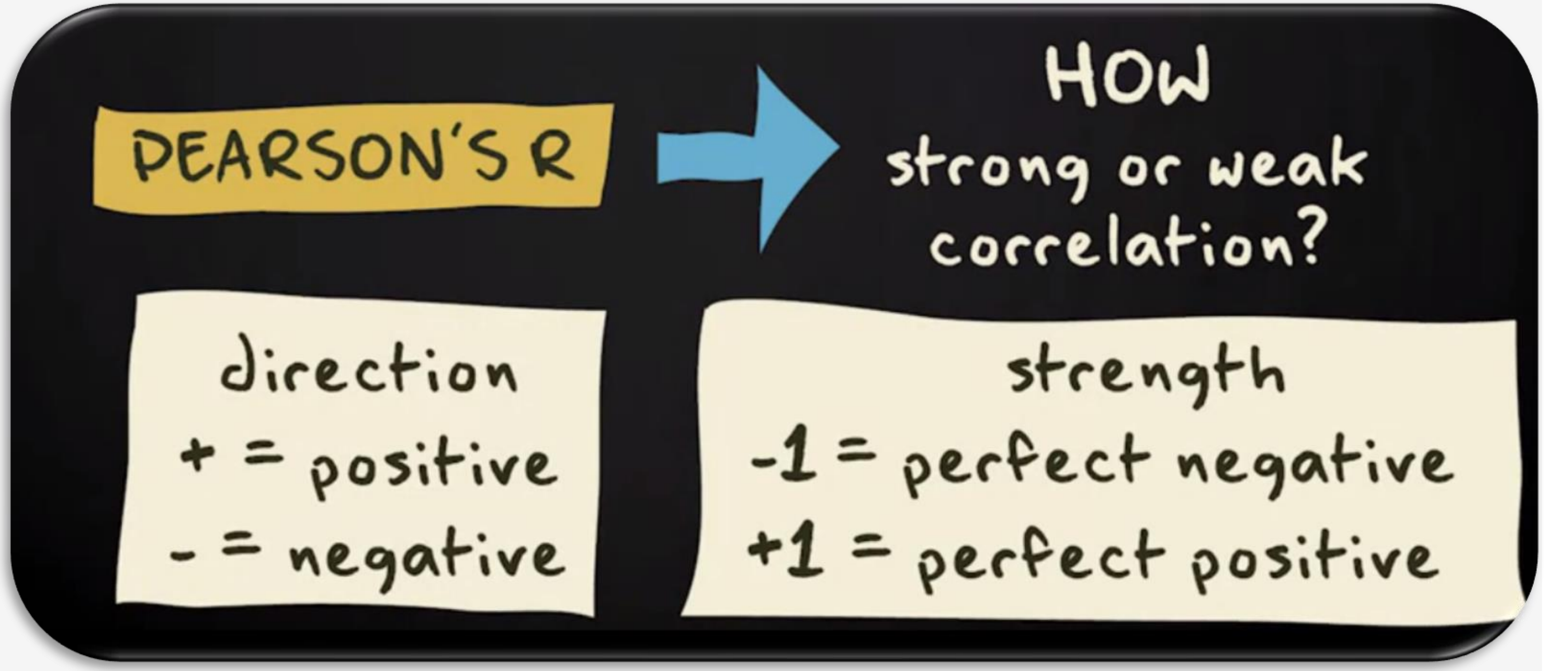


Correlation

A scatterplot helps us to broadly assess whether a correlation is strong or weak, but it does not tell us exactly how strong the relationship is.

Pearson's r is a measure that can show us exactly that.

More specifically, the **Pearson's r tells us the direction and exact strength of the linear relationship between two quantitative variables.**

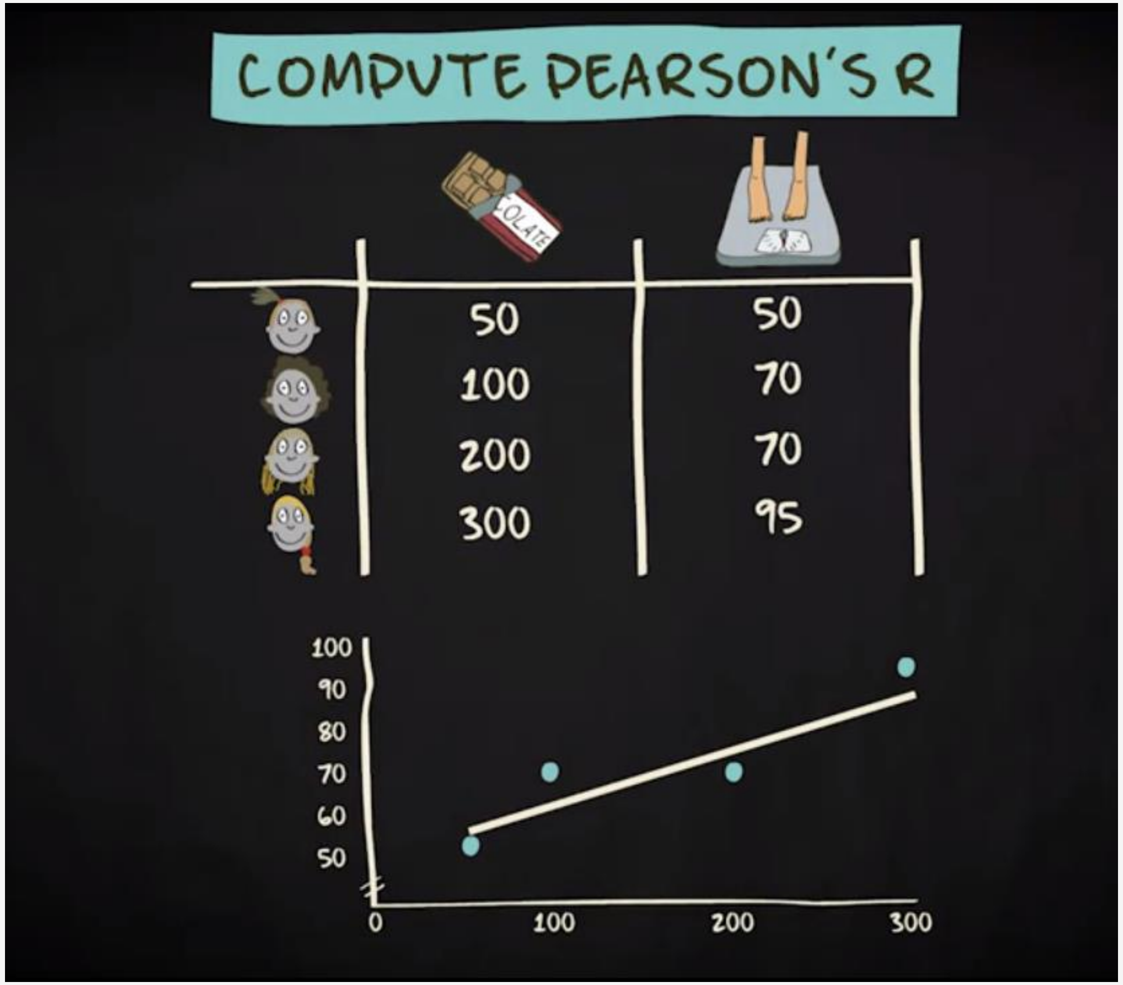


Correlation

So , how do we compute the Pearson’s r?

Imagine that the study on chocolate consumption and body weight was not based on 200 but on only 4 individuals.

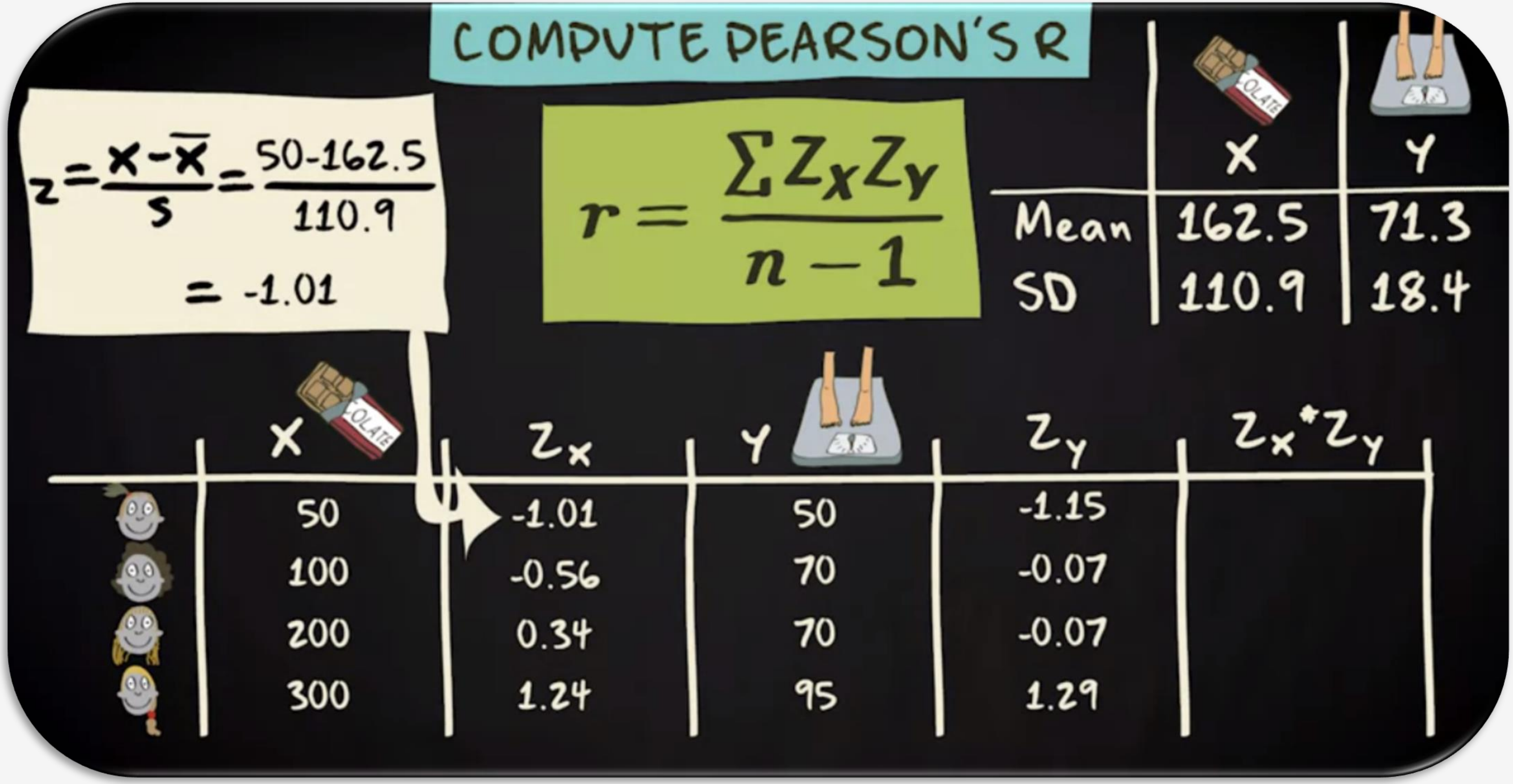
This is the data matrix and the scatterplot.



COMPUTE PEARSON'S R

50	50
100	70
200	70
300	95

Correlation – Compute Pearson’s R



COMPUTE PEARSON'S R

$$z = \frac{x - \bar{x}}{s} = \frac{50-162.5}{110.9}$$

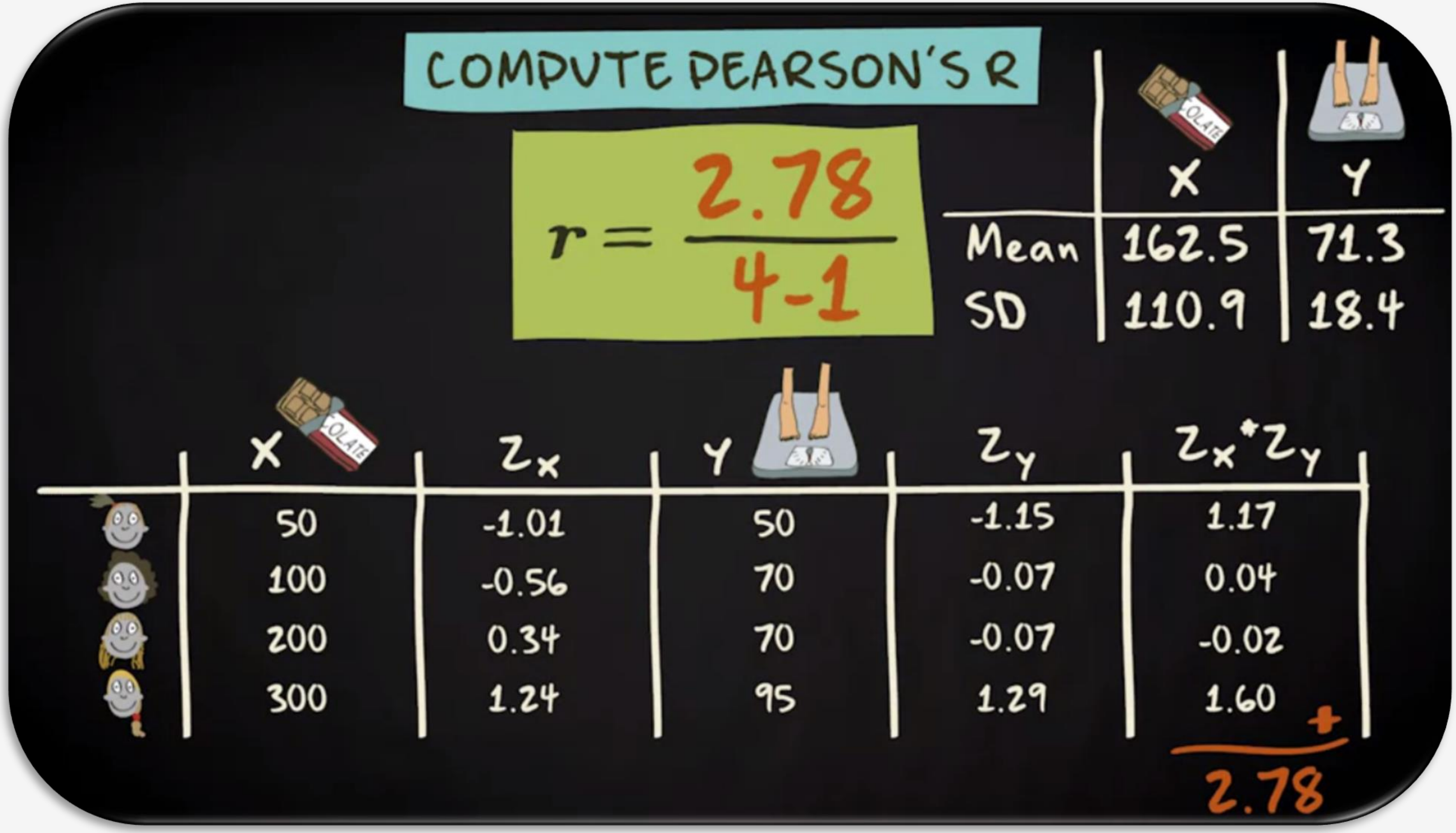
$$= -1.01$$

$$r = \frac{\sum Z_X Z_Y}{n-1}$$

	X	Y
Mean	162.5	71.3
SD	110.9	18.4

	X	Z_x	Y	Z_y	$Z_x{}^*Z_y$
	50	-1.01	50	-1.15	
	100	-0.56	70	-0.07	
	200	0.34	70	-0.07	
	300	1.24	95	1.29	

Correlation – Compute Pearson’s R



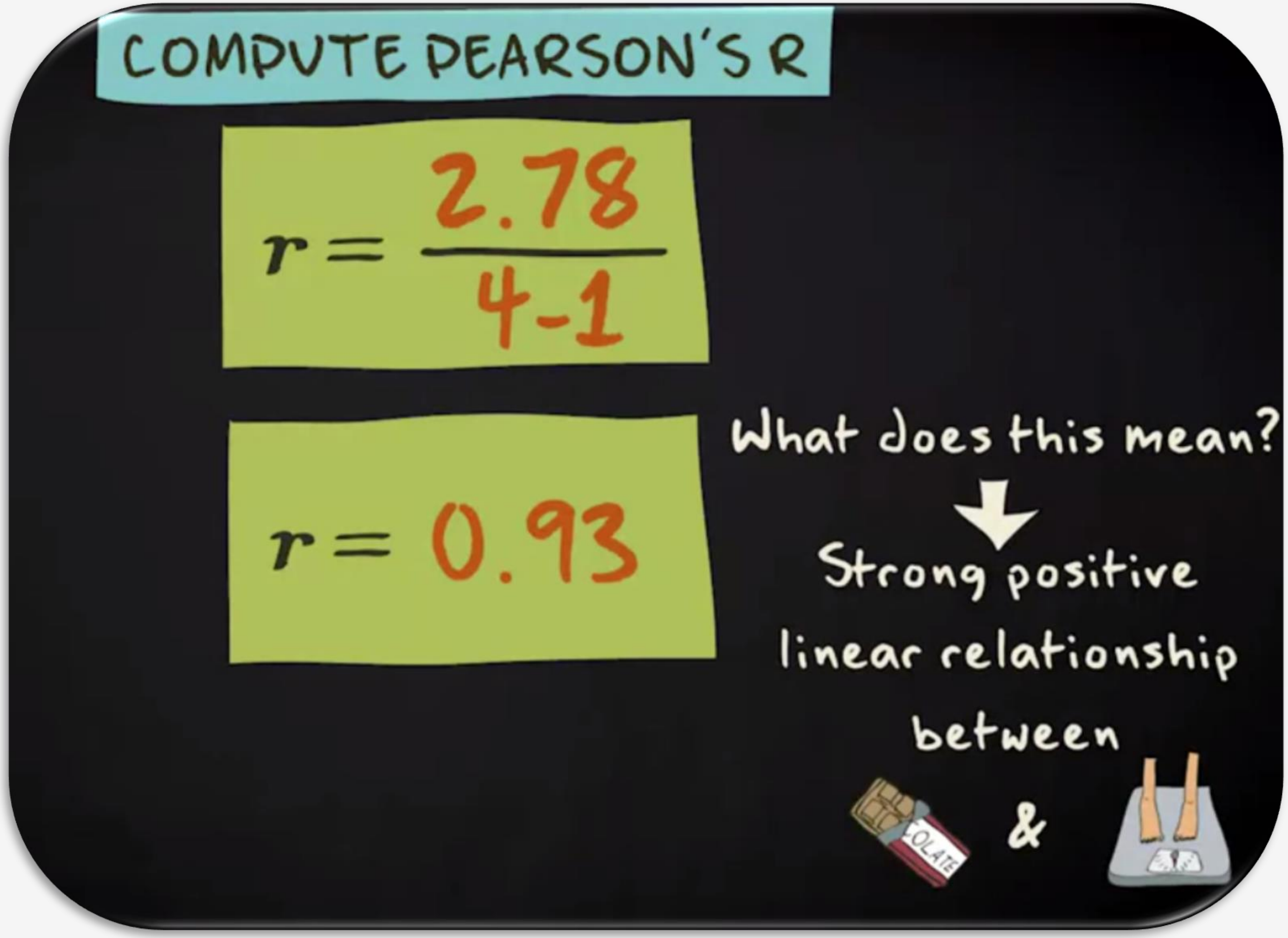
COMPUTE PEARSON'S R

$$r = \frac{2.78}{4-1}$$

	X	Y
Mean	162.5	71.3
SD	110.9	18.4

X	Z_X	Y	Z_Y	$Z_X{*}Z_Y$
50	-1.01	50	-1.15	1.17
100	-0.56	70	-0.07	0.04
200	0.34	70	-0.07	-0.02
300	1.24	95	1.29	1.60
				+ 2.78

Correlation – Compute Pearson’s R



COMPUTE PEARSON'S R

$$r = \frac{2.78}{4-1}$$

$$r = 0.93$$

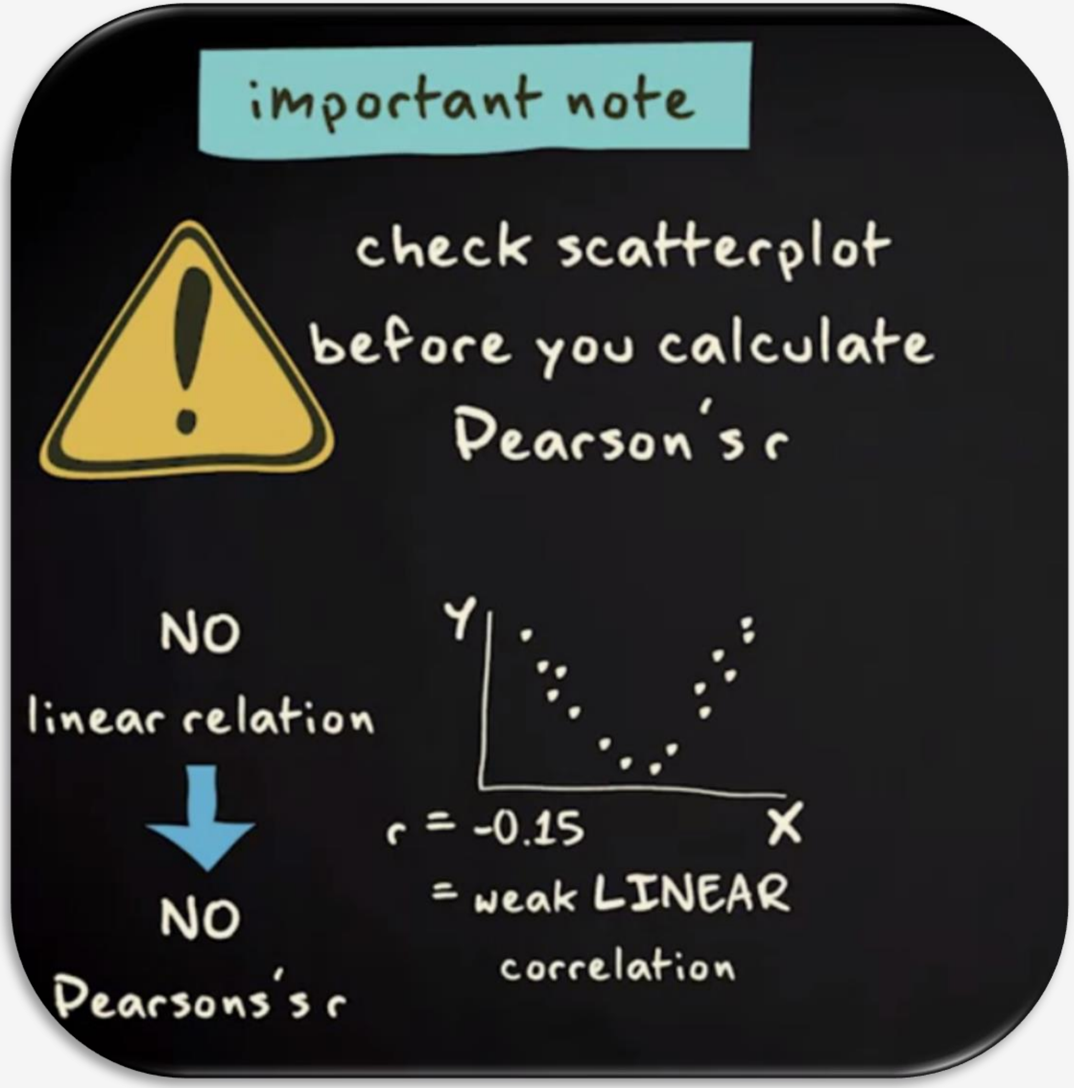
What does this mean?

↓

Strong positive linear relationship between

&

Correlation



important note

check scatterplot before you calculate Pearson's r

NO linear relation

NO Pearsons's r

r = -0.15

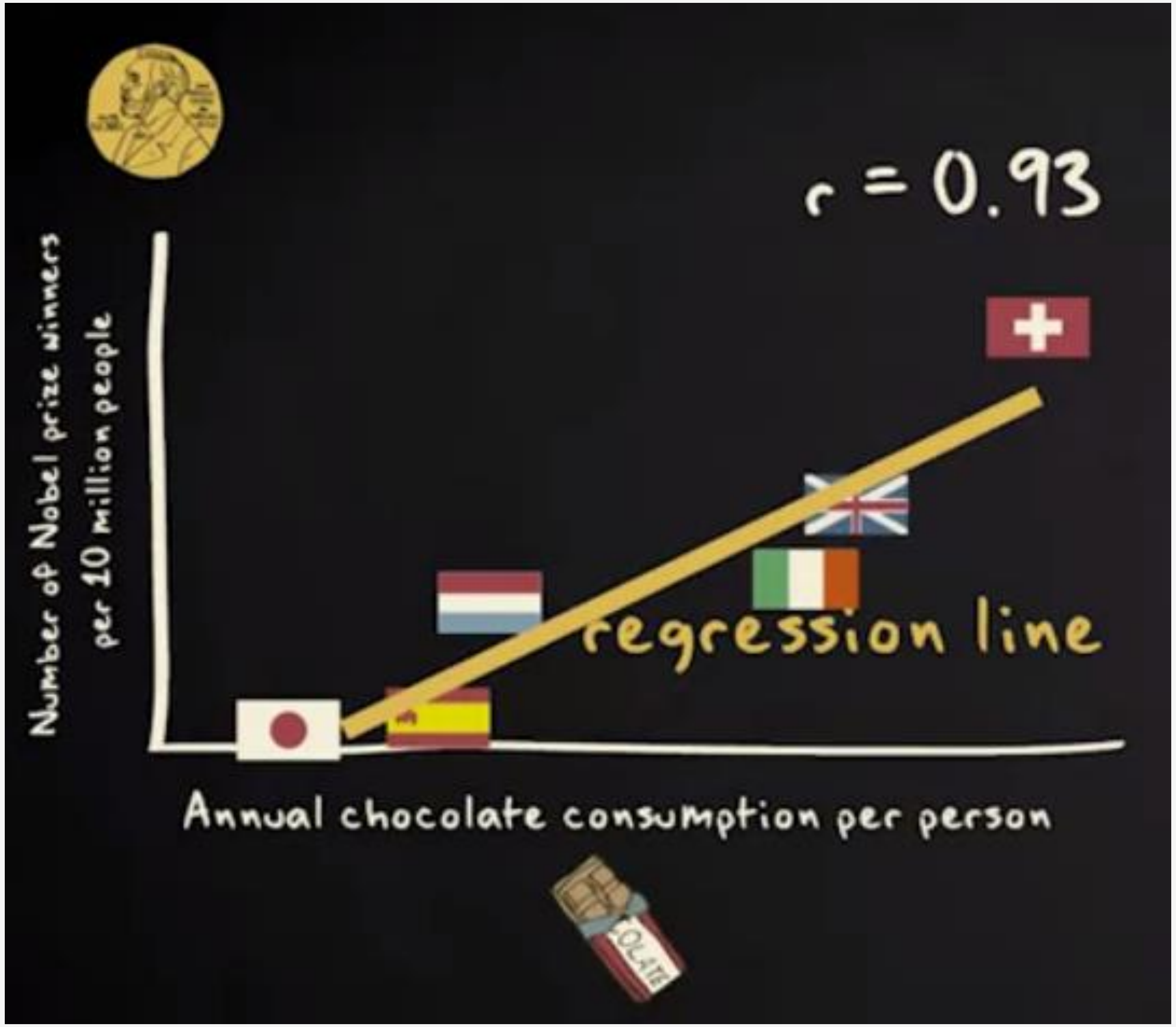
= weak LINEAR correlation

Regression

Regression

However, a recent study shows that it might actually be a good idea to eat a lot of chocolate.

This scatterplot shows that a country's annual chocolate consumption per person (so, how much chocolate someone eats in a year) is positively related to the number of Nobel Prize winners per 10 million people in a country

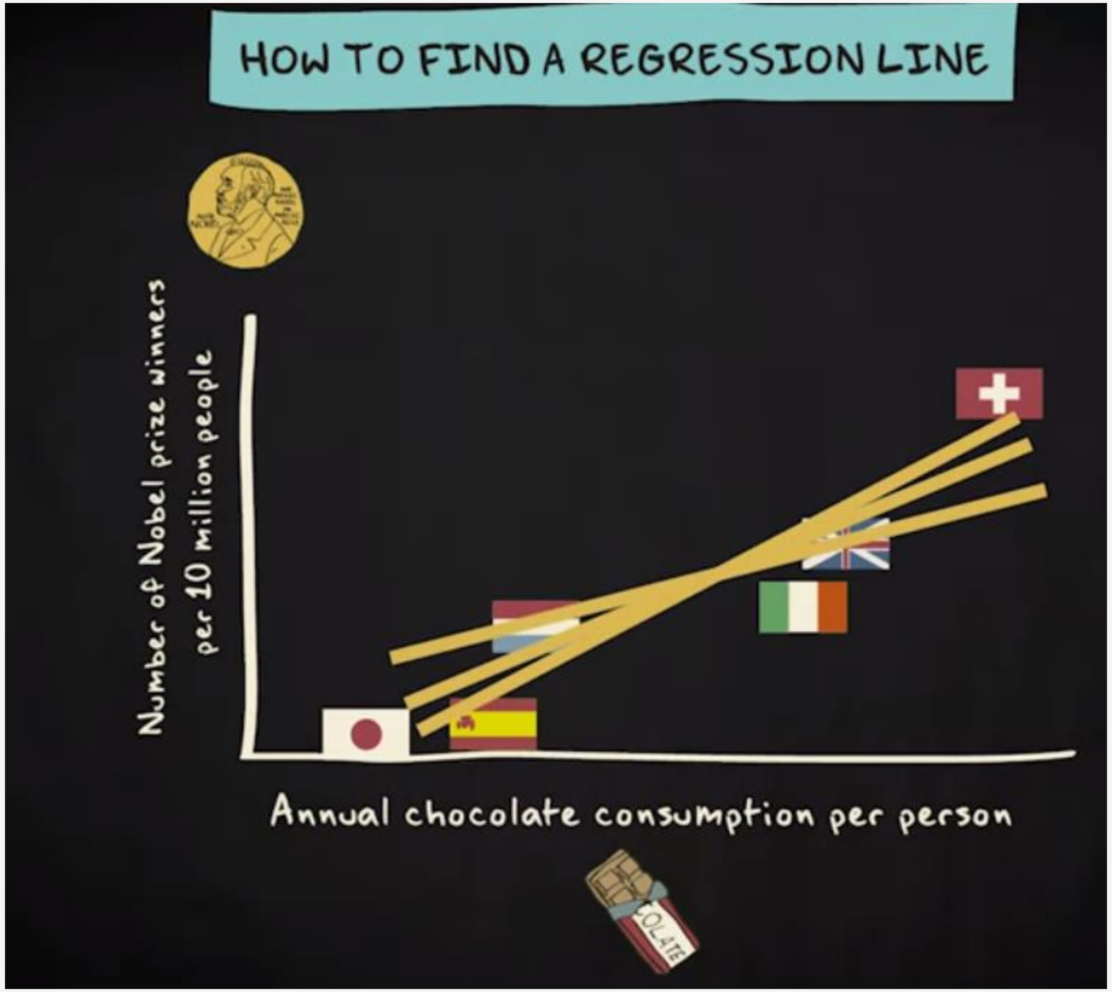


Regression

The Pearson's r tells you how strong the linear correlation between two continuous variables is.

This linear correlation can be displayed by a straight line. In our case that's this line. This is what we call the regression line and in this section we 'll discuss how we can find the regression line.

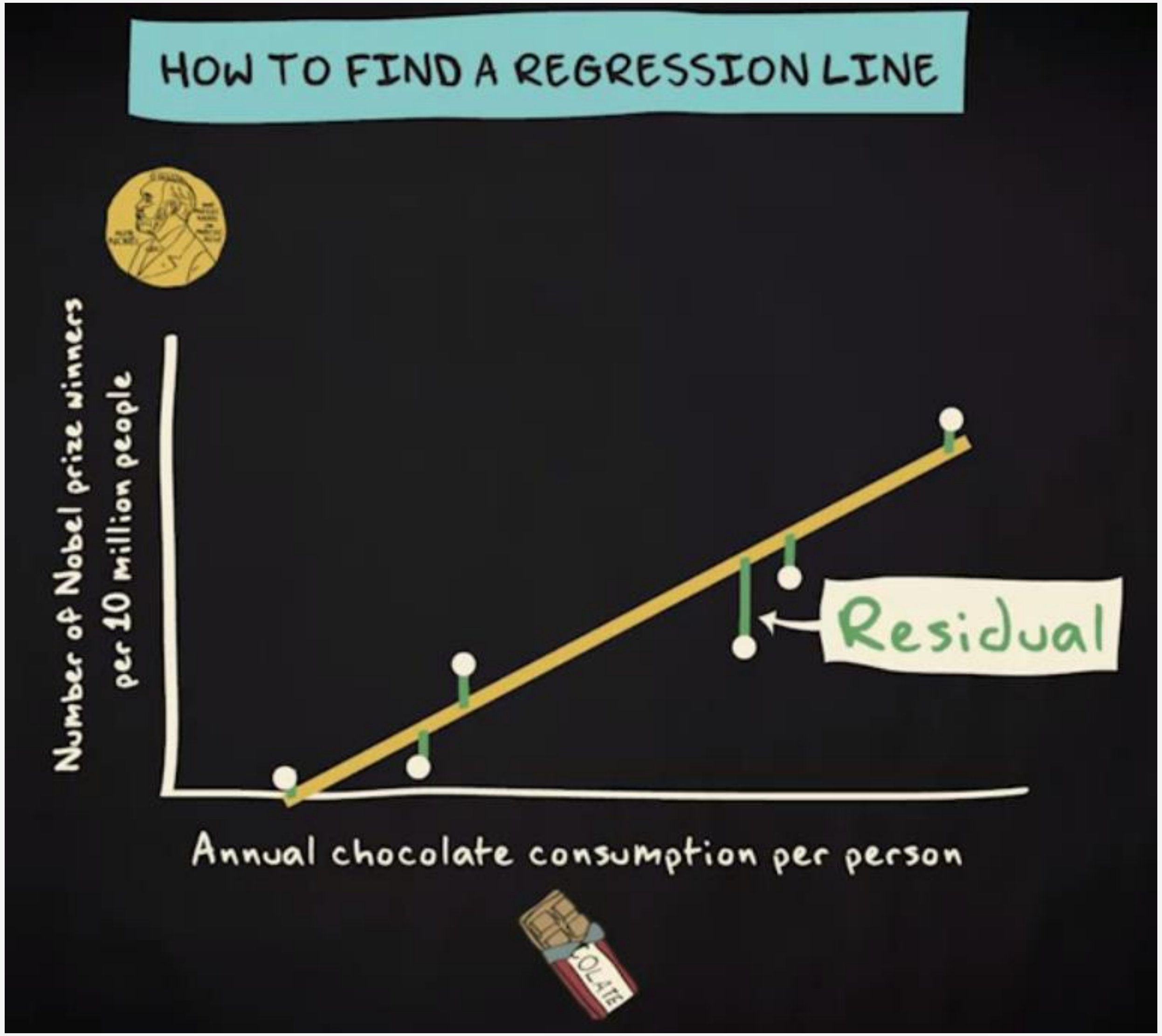
Imagine that you draw every possible straight line through this scatterplot. imagine that you have superhuman powers and that you are able to do it. Next, you measure for every possible line the distances from the line to every case (so, in this case to every flag in the scatterplot).



Regression

Let me give you an example based on a random line. You measure the vertical distance between Japan and the line, the distance between Spain and the line, and so on, until you know the distance to the line of every case in your study.

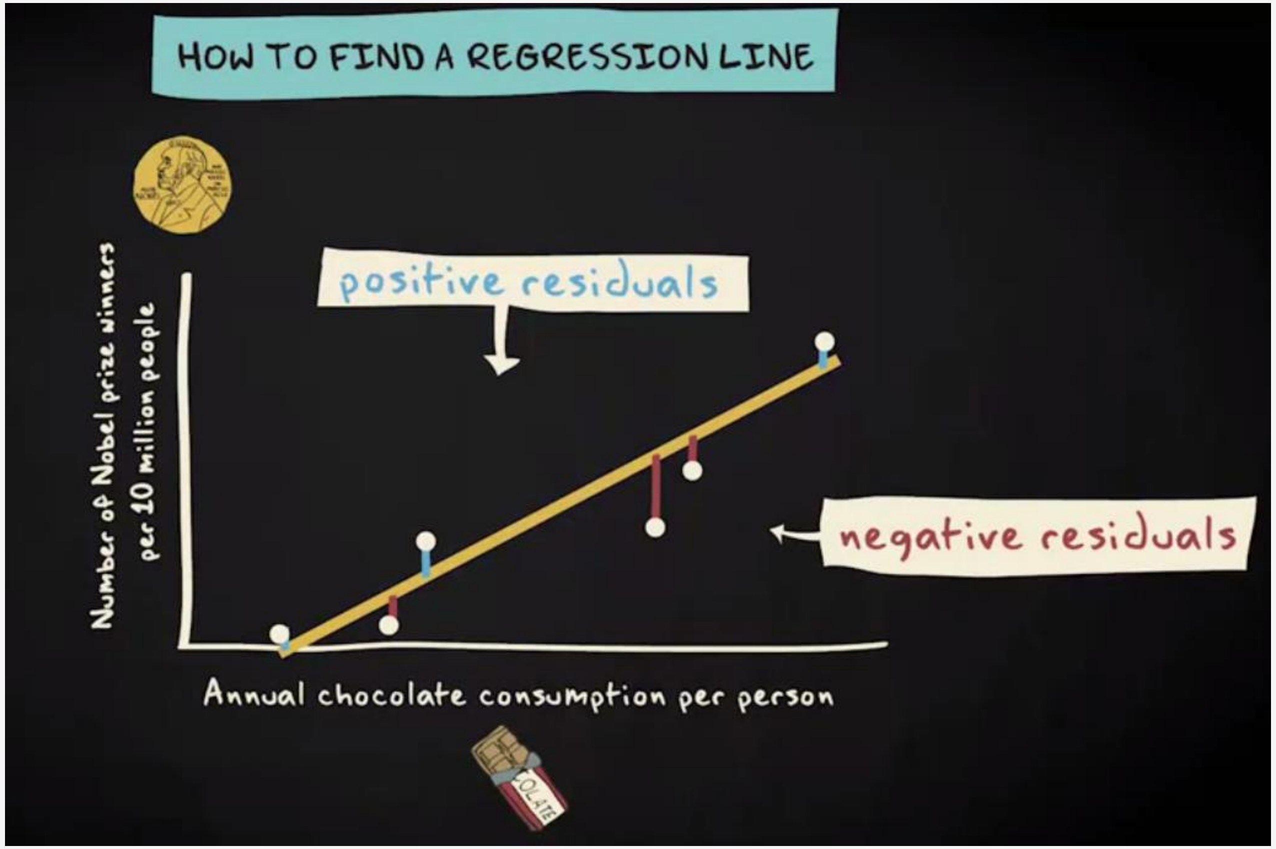
Every distance is called a **residual**.


HOW TO FIND A REGRESSION LINE
Number of Nobel prize winners per 10 million people
Residual
Annual chocolate consumption per person
OLATE

Regression

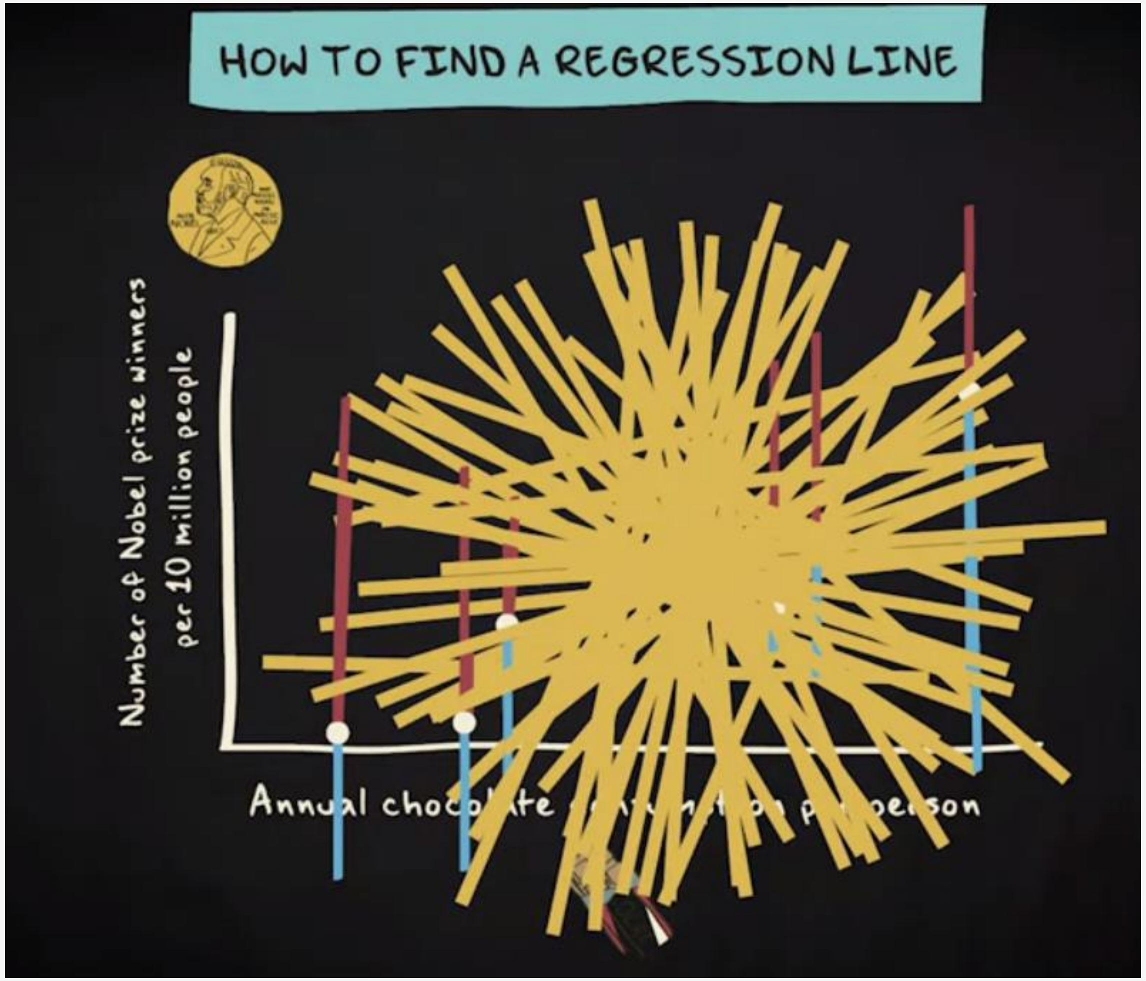
You end up with positive residuals (the distances from cases above the line to the line, displayed in blue) and negative residuals (distances from cases below the line to the line, displayed in red).

You measure these residuals for every possible line through the scatterplot.


HOW TO FIND A REGRESSION LINE
positive residuals
negative residuals
Number of Nobel prize winners per 10 million people
Annual chocolate consumption per person
COLATE

Regression

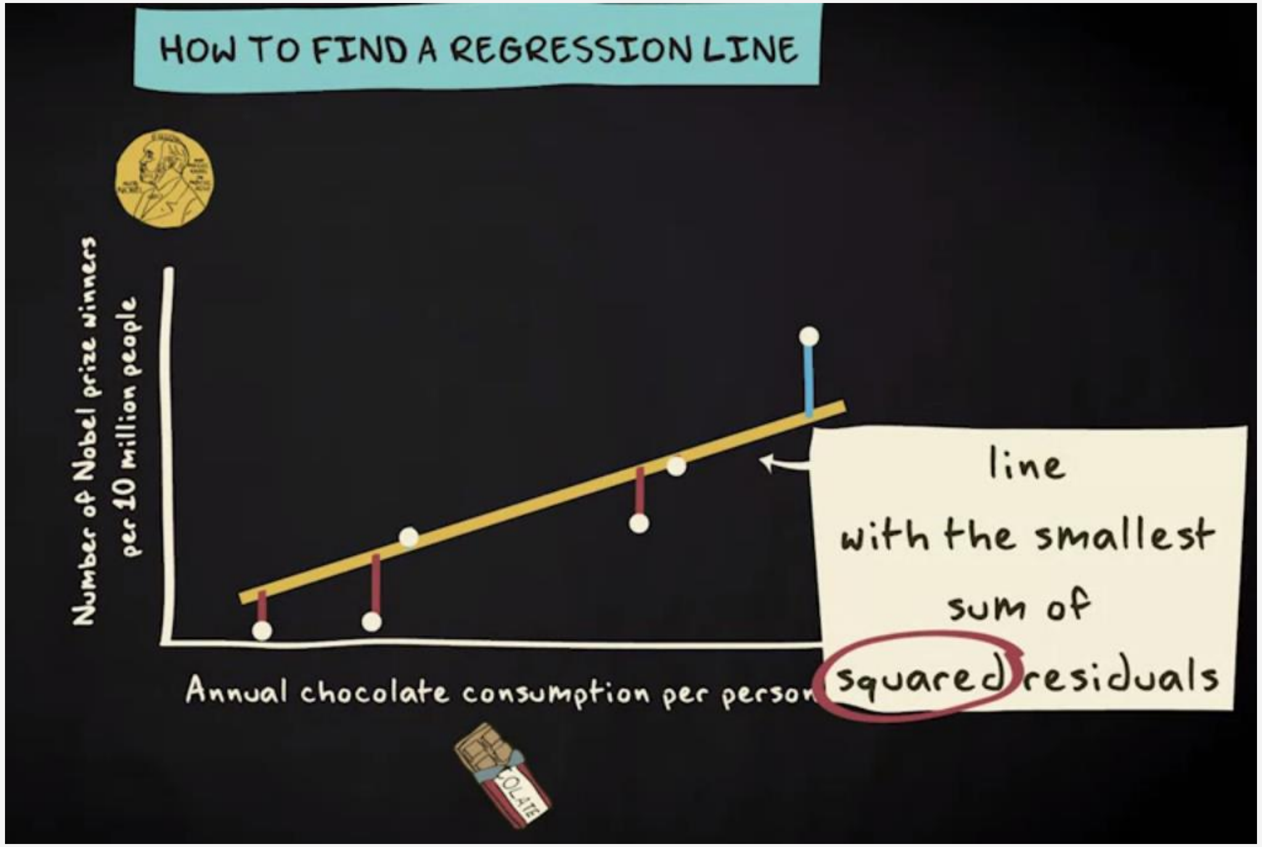
So, not only for this line, but also for this line, this line and this line. And for every other possible line through the scatterplot.


HOW TO FIND A REGRESSION LINE
Number of Nobel prize winners per 10 million people
Annual chocolate consumption per person

Regression

Eventually, you choose the line for which the **sum of the squared residuals** is the smallest. That's this one.

Why the squared residuals? Because positive and negative residuals cancel each other out: the sum of the length of the positive residuals (the blue lines) is exactly as big as the sum of the length of the negative residuals (the red lines).

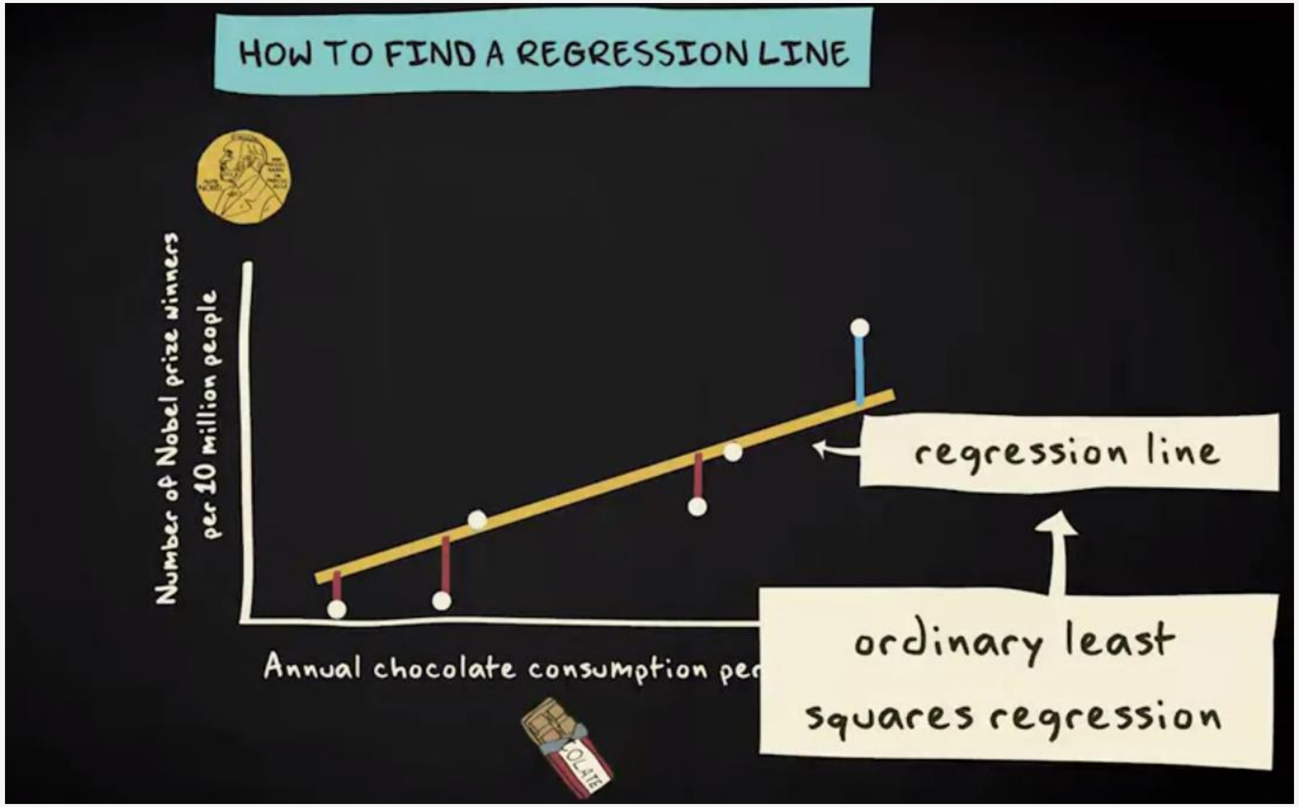

HOW TO FIND A REGRESSION LINE
Number of Nobel prize winners per 10 million people
Annual chocolate consumption per person
line with the smallest sum of squared residuals
OLATE

Regression

The best fitting line is called the **regression line**,

and

the name of the **method of analysis** is called **ordinary least squares (OLS) regression**, which refers to the way we have found the line.

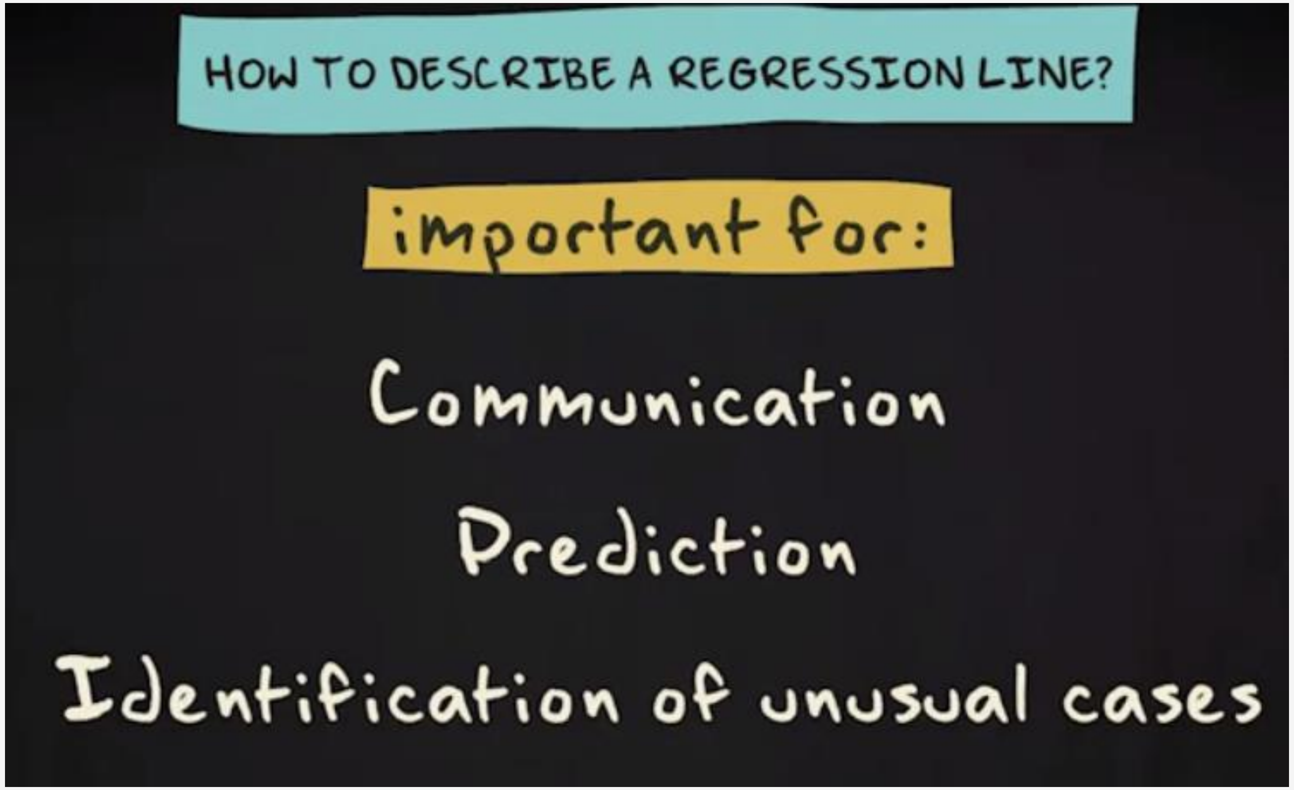


Regression

The regression line is the straight line that describes the linear relationship between the two variables best.

But how can we describe what this line looks like?

This is an important question, because by **describing the line with a formula**, we can easily communicate our regression analysis to other people, predict the number of Nobel Prize winners in other countries, and identify countries that do not fit the pattern.



HOW TO DESCRIBE A REGRESSION LINE?

important for:

Communication

Prediction

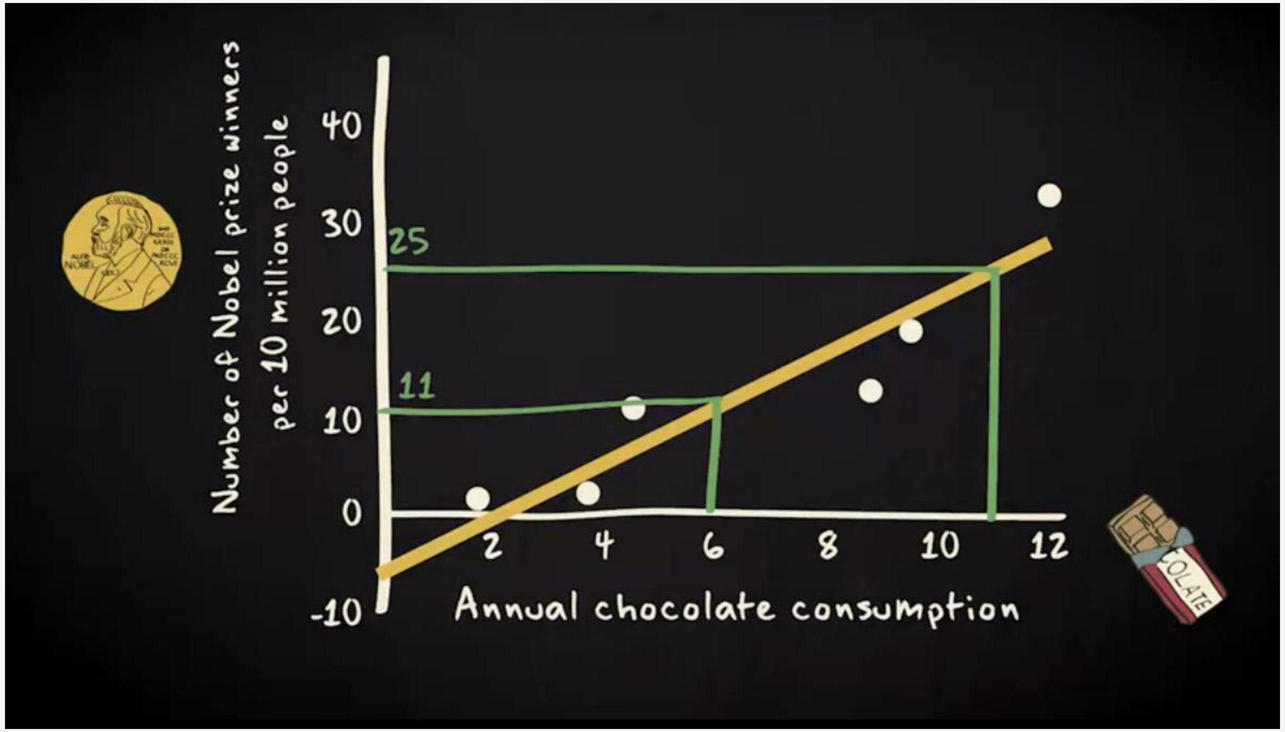
Identification of unusual cases

Regression

Based on the regression line in this scatterplot we would predict that a country with an annual chocolate consumption of **6 kilograms** per year per capita would have about **11 Nobel prize winners** per 10 million people.

Similarly, based on this same line we would predict that a state with an annual chocolate consumption of **11 kilograms** per year per capita would have about **25 prize winners** per 10 million people. For most countries this prediction will not be completely correct.

After all, most countries are not exactly on the line. However, it is the best prediction that we can make based on the information that we have.

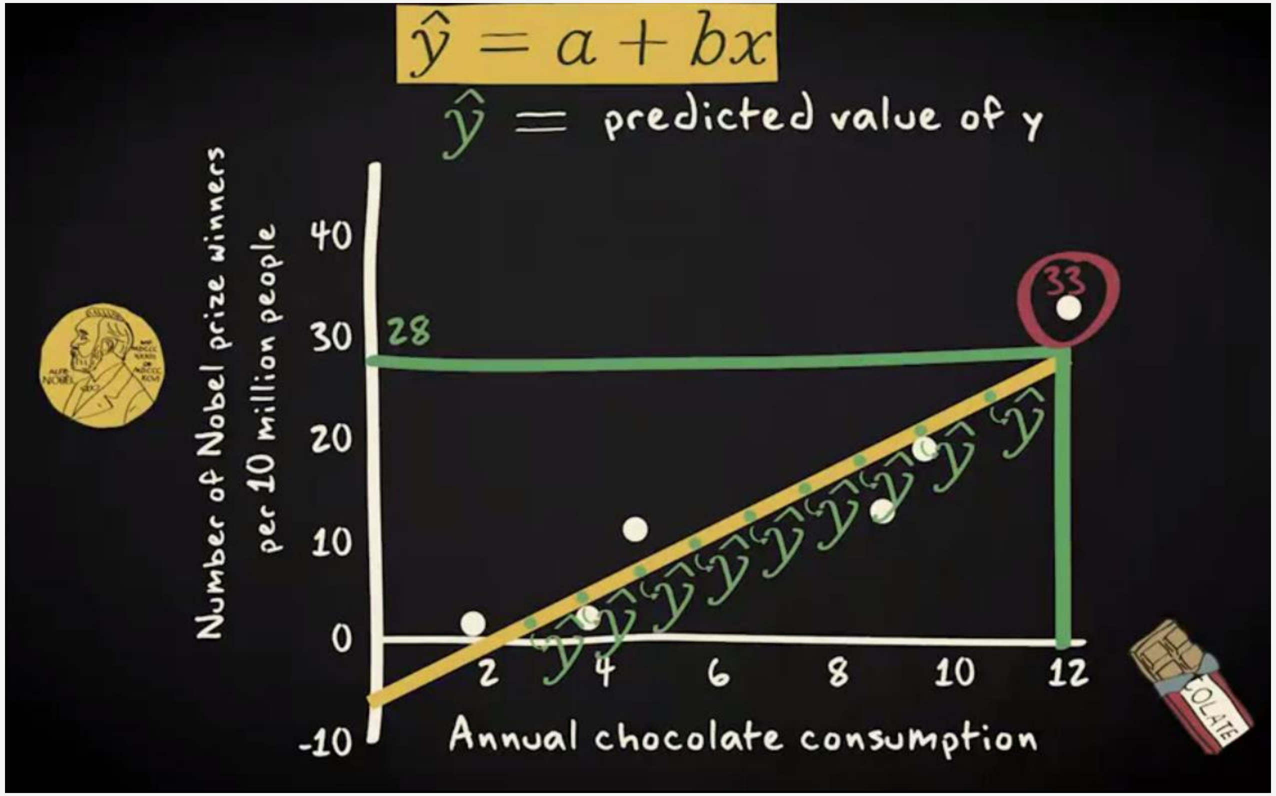


Regression

There is one simple formula with which we can describe the regression line, and that's this one:

$$\hat{y} = a + bx.$$

$\hat{y}$ is not the actual value of y, but it represents the **predicted value of y.**



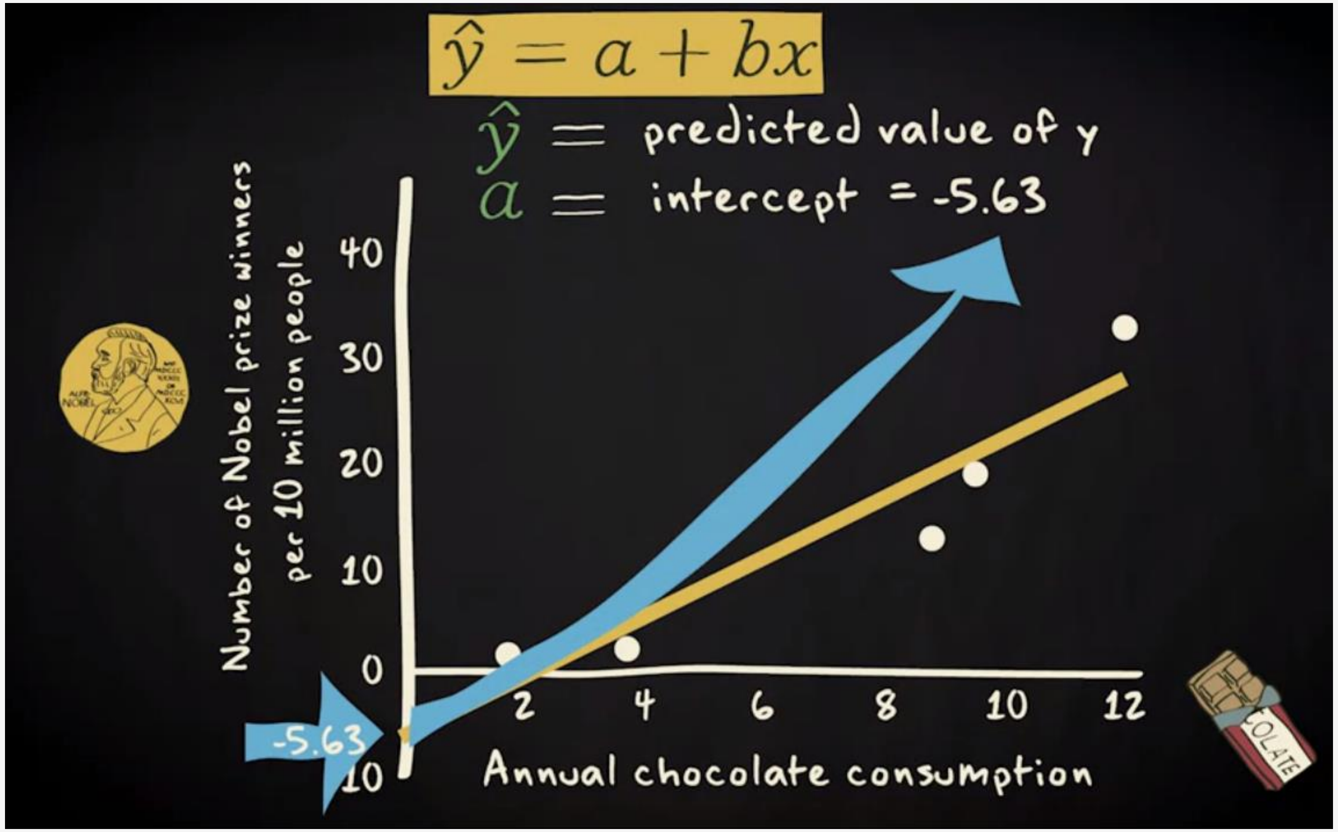
Regression

‘a’ is what we call the **intercept** or **the constant**.

It is the predicted value of y when x equals 0.

It is, in other words, the predicted value of y where the regression line crosses the y-axis and x thus equals 0.

In our case that’s -5.63. Notice that this value has no substantive meaning. It is impossible to have -5.63 Nobel prize winners per 10 million people. It only has a mathematical purpose, to describe the regression line.



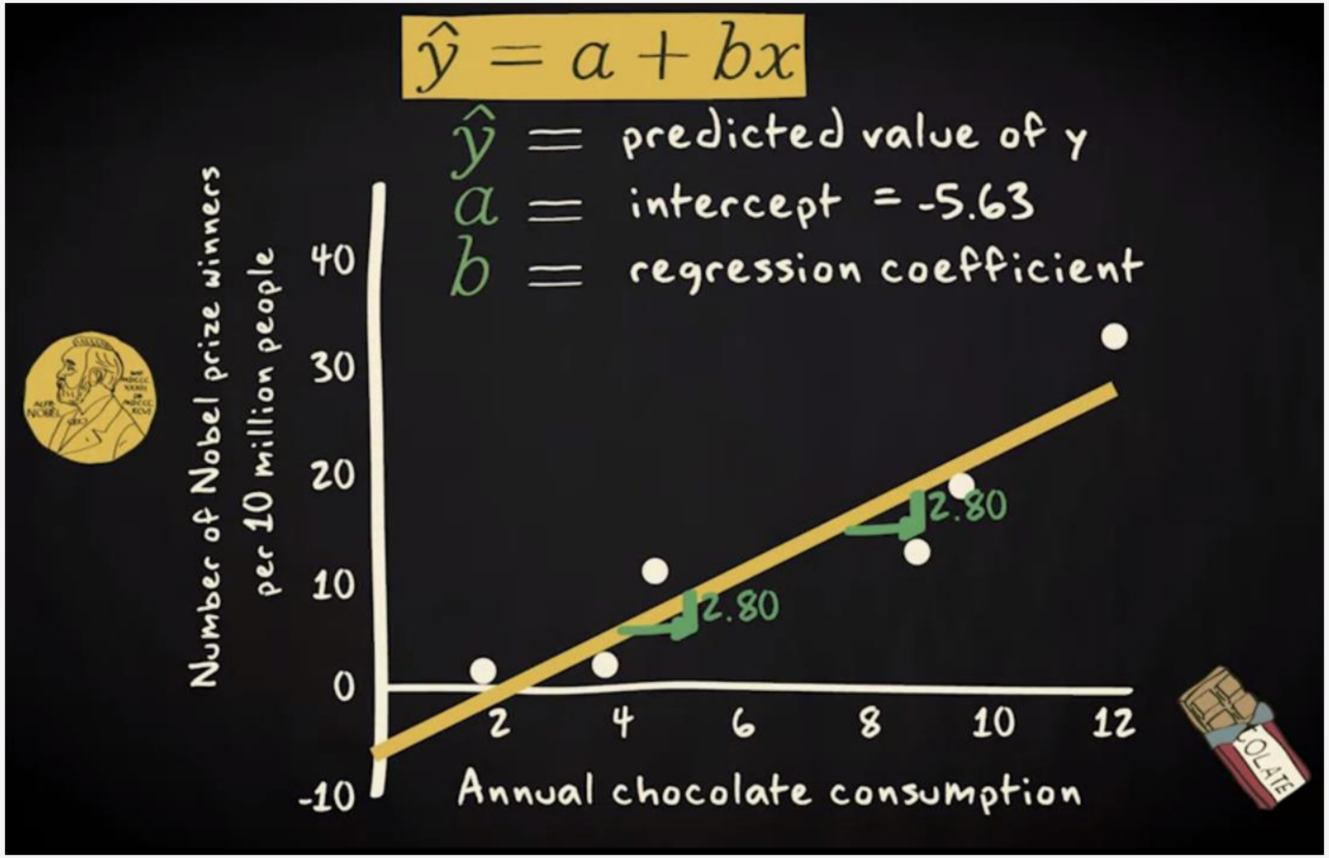
Regression

‘b’ is what we call the **regression coefficient** or **the slope**.

It is the change in $\hat{y}$ when x increases with one unit.

In our case we see that when x increases with one unit (for example from 4 to 5), the predicted value of y increases with 2.80 units. Because we have a straight line, the slope of the regression line is the same everywhere. So also if we look at what happens when x increases from 8 to 9, $\hat{y}$ increases with 2.80 units.

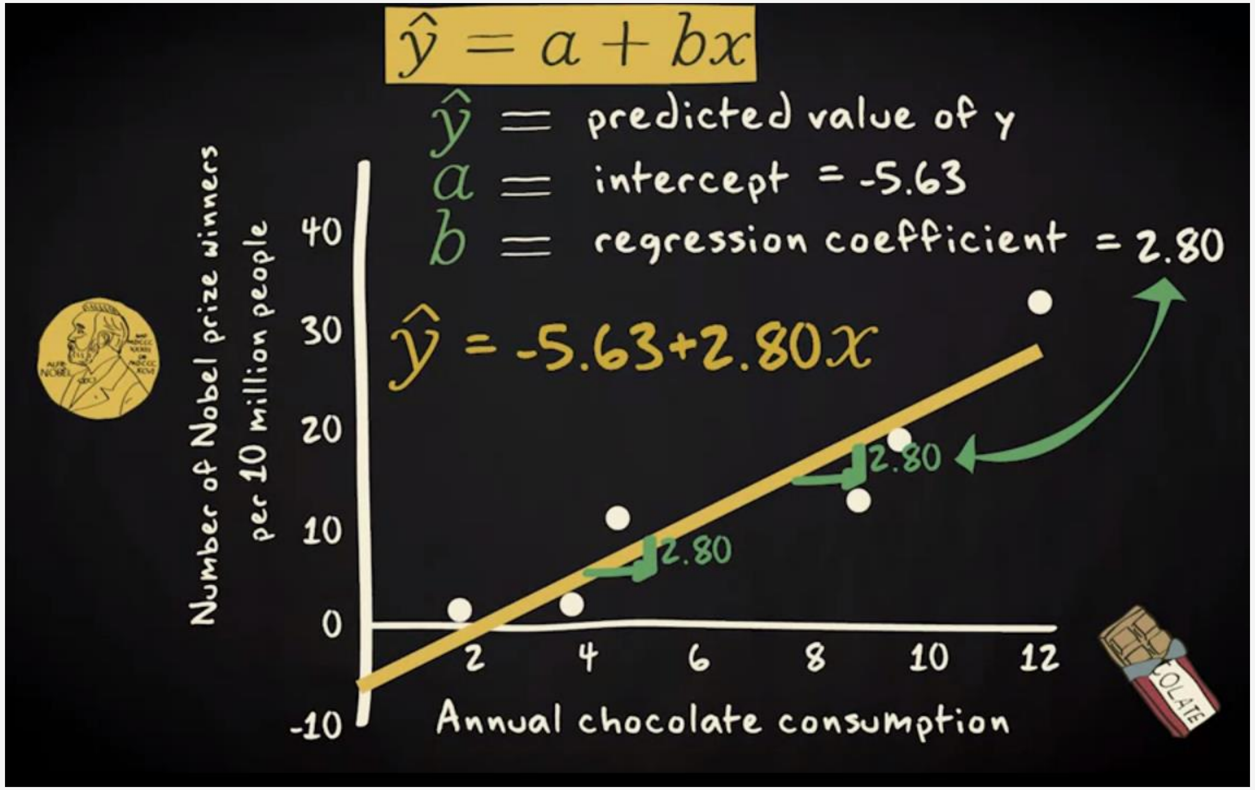
The regression coefficient in our example is 2.80.



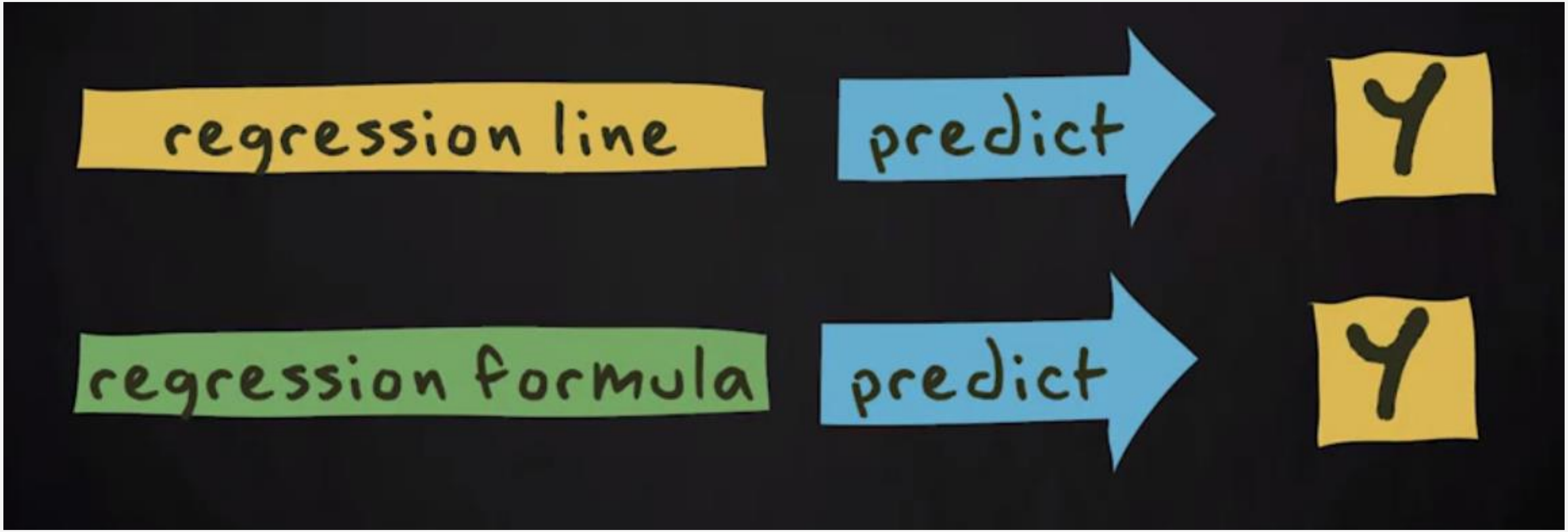
Regression

This leads to the following regression equation:

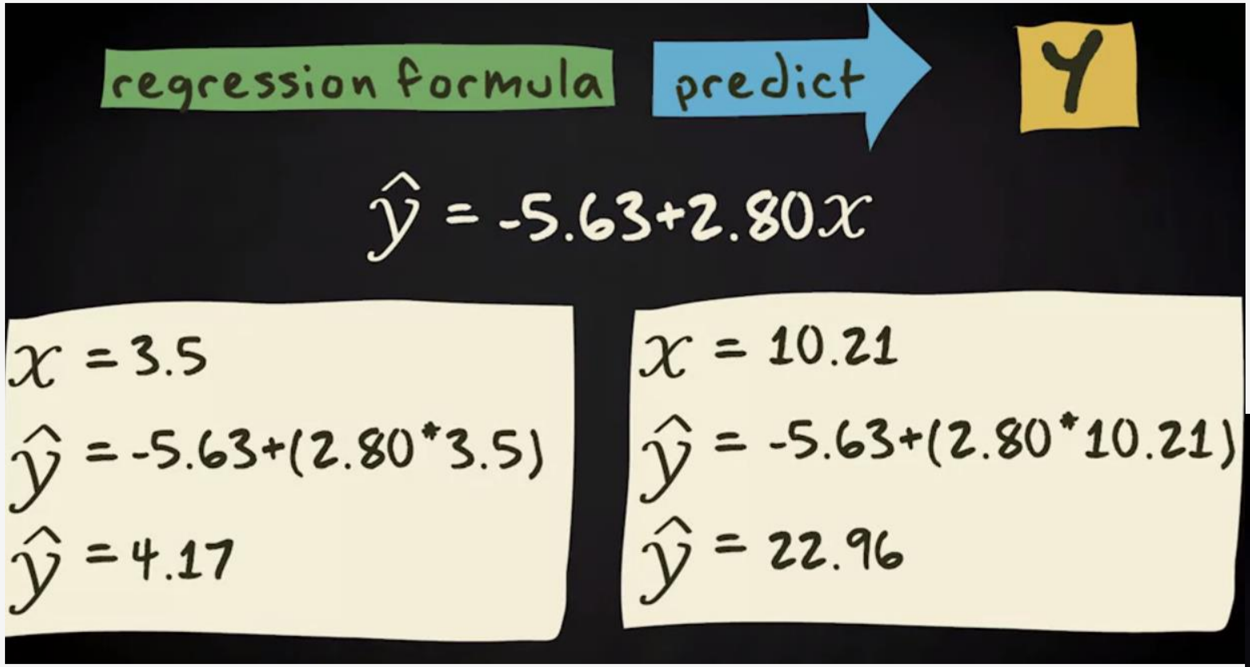
$\hat{y}$ equals -5.63 plus 2.80 times X.



Regression



Regression



regression formula → predict → Y

$$\hat{y} = -5.63 + 2.80x$$

$x = 3.5$

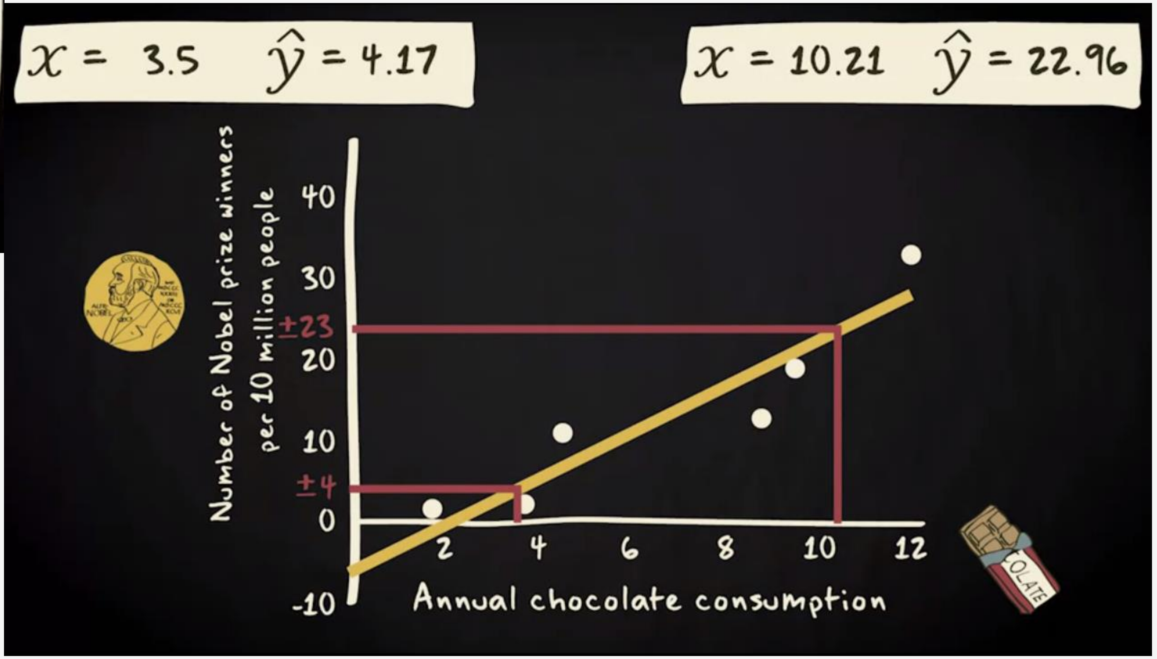
$\hat{y} = -5.63 + (2.80 * 3.5)$

$\hat{y} = 4.17$

$x = 10.21$

$\hat{y} = -5.63 + (2.80 * 10.21)$

$\hat{y} = 22.96$

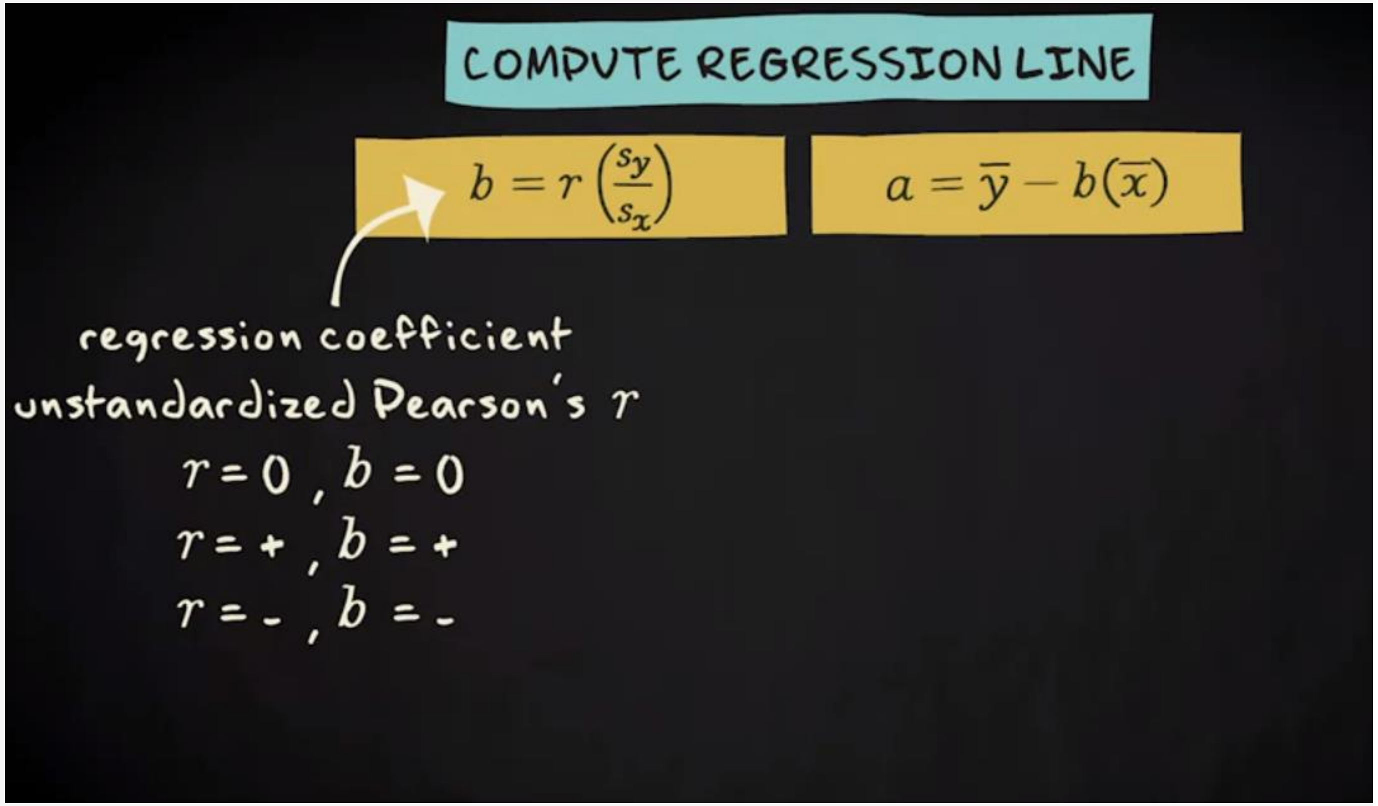


You can already see that there is one huge advantage of working with the formula: **you can make much more precise predictions.**

Regression

Usually the computer finds the regression line for you, so you don't have to compute it yourself.

However, when you know the **means** and **standard deviations** of your variables and the corresponding Pearson's r correlation coefficient, you can compute the regression equation by means of two formulas.



COMPUTE REGRESSION LINE

$b = r\left(\frac{s_y}{s_x}\right)$

$a = \overline{y} - b(\overline{x})$

regression coefficient

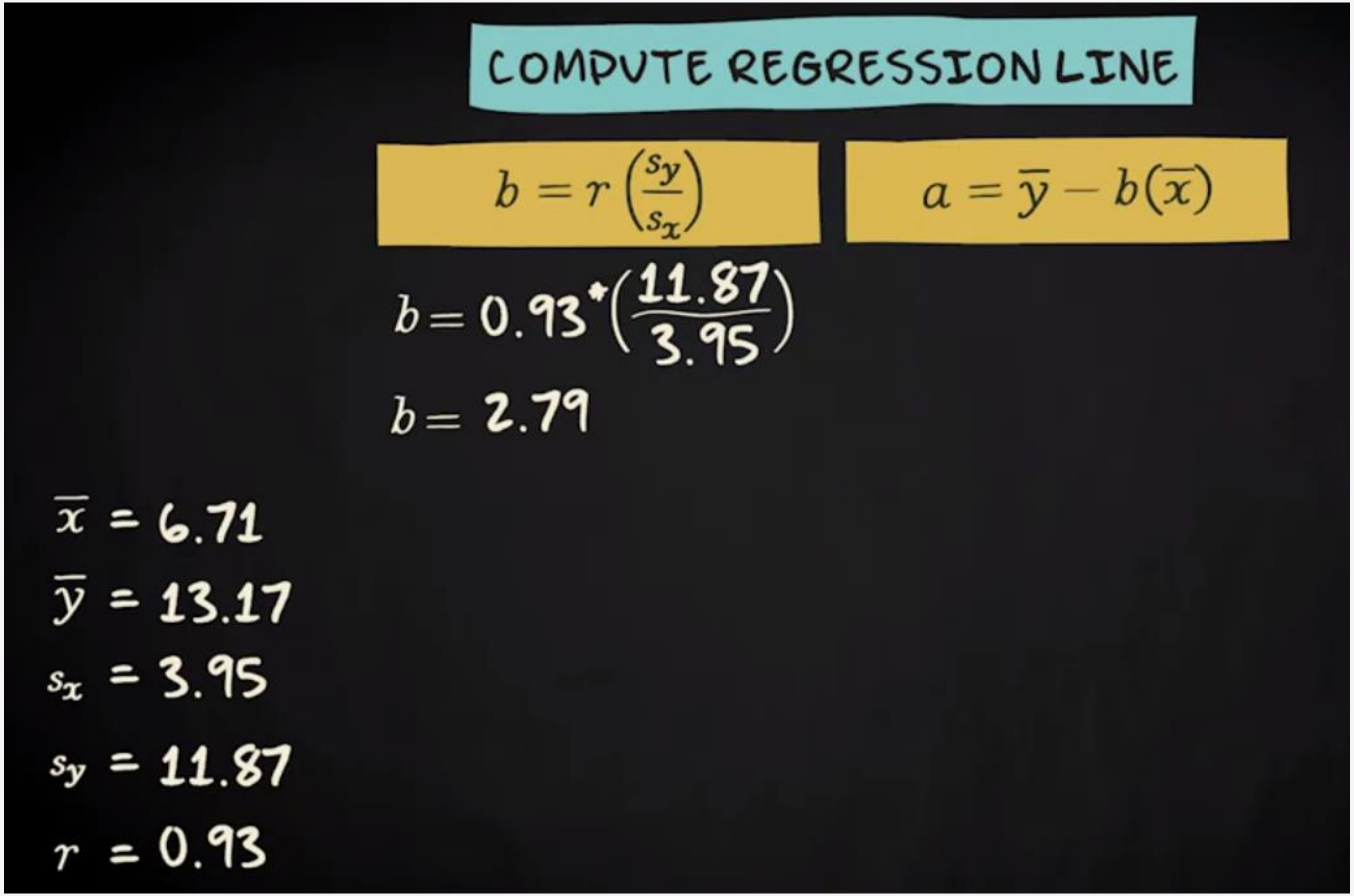
unstandardized Pearson's r

$r = 0\ ,\ b = 0$

$r = +\ ,\ b = +$

$r = -\ ,\ b = -$

Regression



COMPUTE REGRESSION LINE

$$b = r\left(\frac{s_y}{s_x}\right)$$

$$a = \overline{y} - b(\overline{x})$$

$$b = 0.93 * \left(\frac{11.87}{3.95}\right)$$

$$b = 2.79$$

$\overline{x} = 6.71$

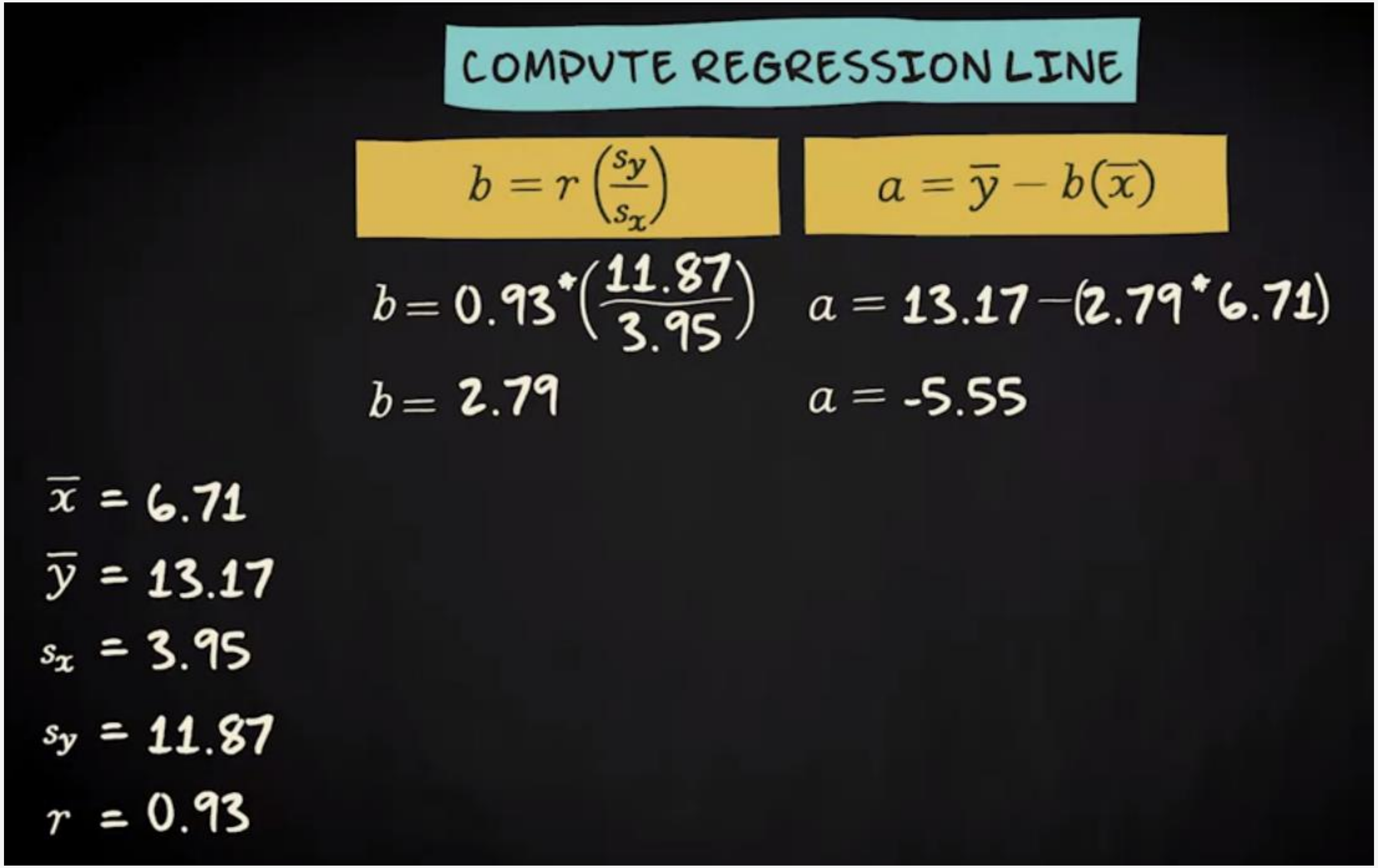
$\overline{y} = 13.17$

$s_x = 3.95$

$s_y = 11.87$

$r = 0.93$

Regression



COMPUTE REGRESSION LINE

$$b = r\left(\frac{s_y}{s_x}\right)$$

$$a = \overline{y} - b(\overline{x})$$

$$b = 0.93 * \left(\frac{11.87}{3.95}\right)$$

$$a = 13.17 - (2.79 * 6.71)$$

$$b = 2.79$$

$$a = -5.55$$

$\overline{x} = 6.71$

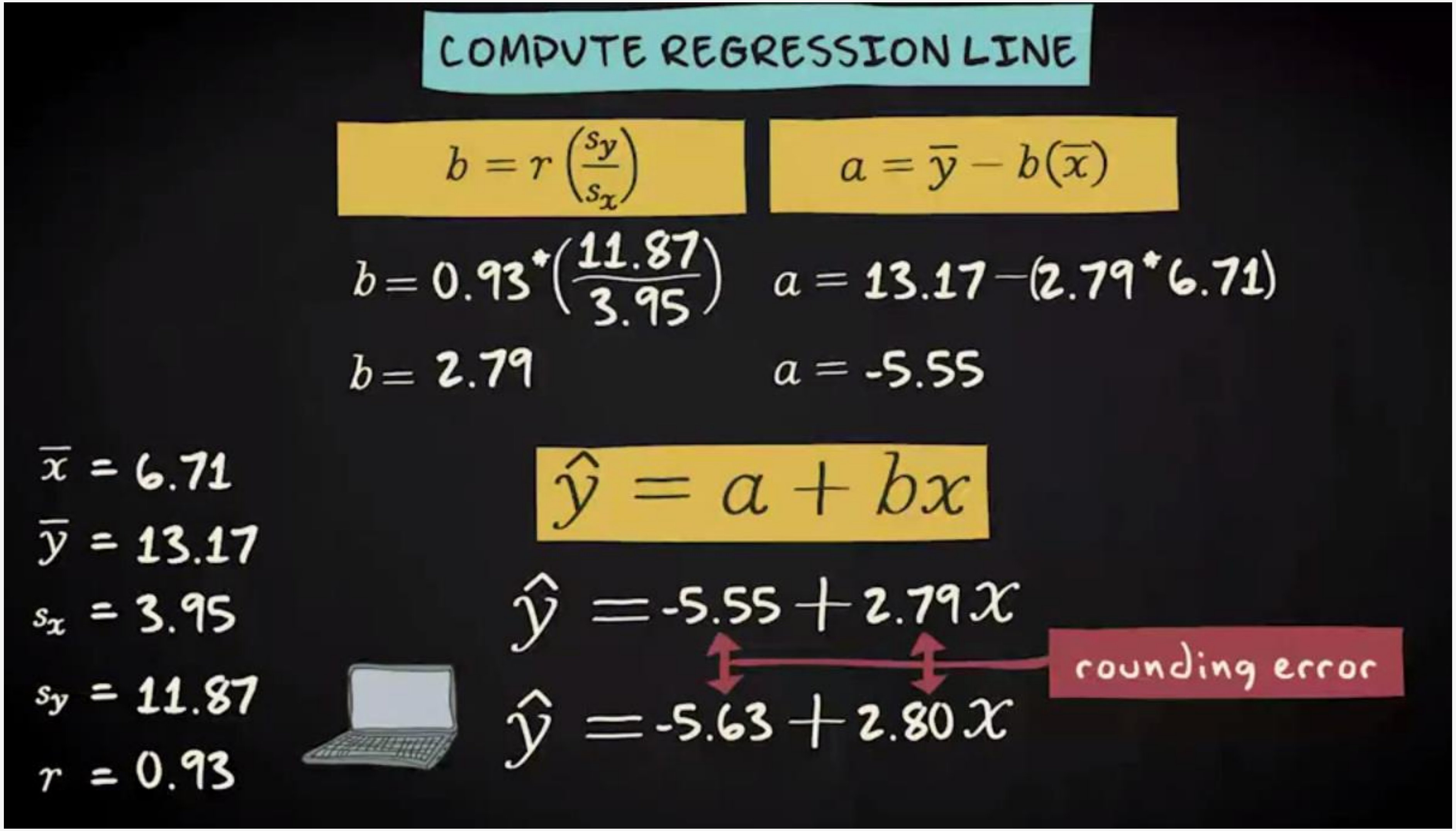
$\overline{y} = 13.17$

$s_x = 3.95$

$s_y = 11.87$

$r = 0.93$

Regression



COMPUTE REGRESSION LINE

$$b = r\left(\frac{s_y}{s_x}\right) \qquad a = \bar{y} - b(\bar{x})$$

$$b = 0.93 * \left(\frac{11.87}{3.95}\right) \qquad a = 13.17 - (2.79 * 6.71)$$

$$b = 2.79 \qquad a = -5.55$$

$\bar{x} = 6.71$

$\bar{y} = 13.17$

$s_x = 3.95$

$s_y = 11.87$

$r = 0.93$

$$\hat{y} = a + bx$$

$$\hat{y} = -5.55 + 2.79x$$

rounding error

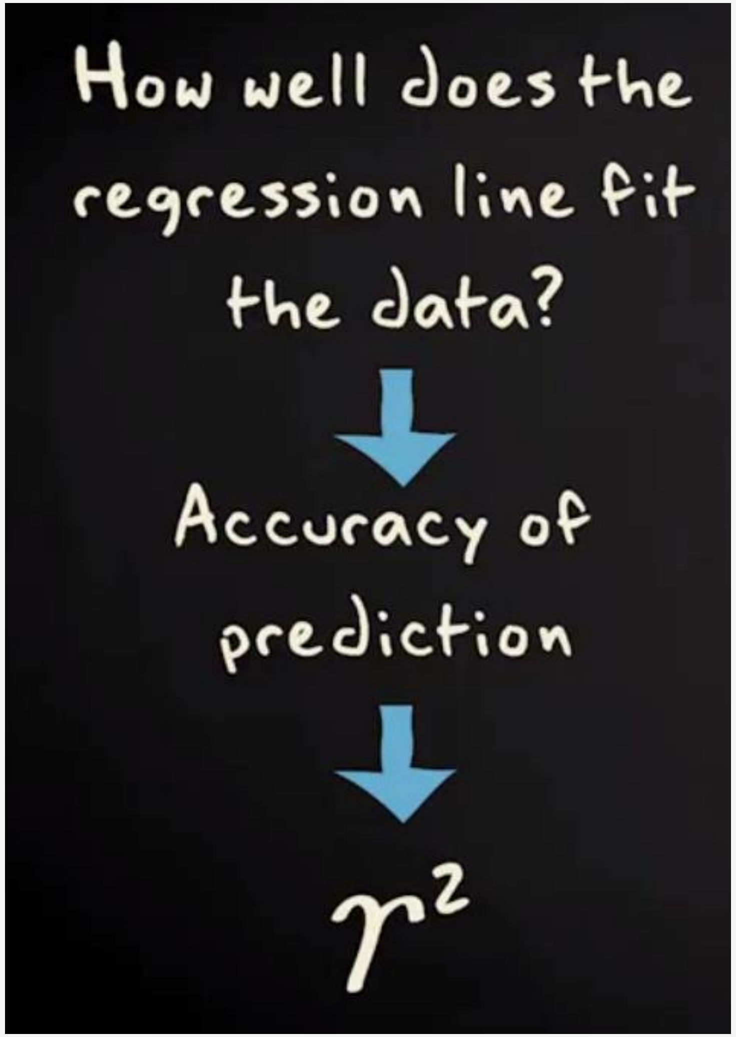
$$\hat{y} = -5.63 + 2.80x$$

Regression

In this section we will talk about the question how we can assess how well a regression line fits your data. The reason to look at how well a regression line fits, is that researchers want to know:

how accurately a regression analysis predicts the dependent variable in a study.

The extent to which a regression line fits the data is expressed by means of the so-called **r-squared**.



How well does the regression line fit the data?

Accuracy of prediction

r^2

Regression

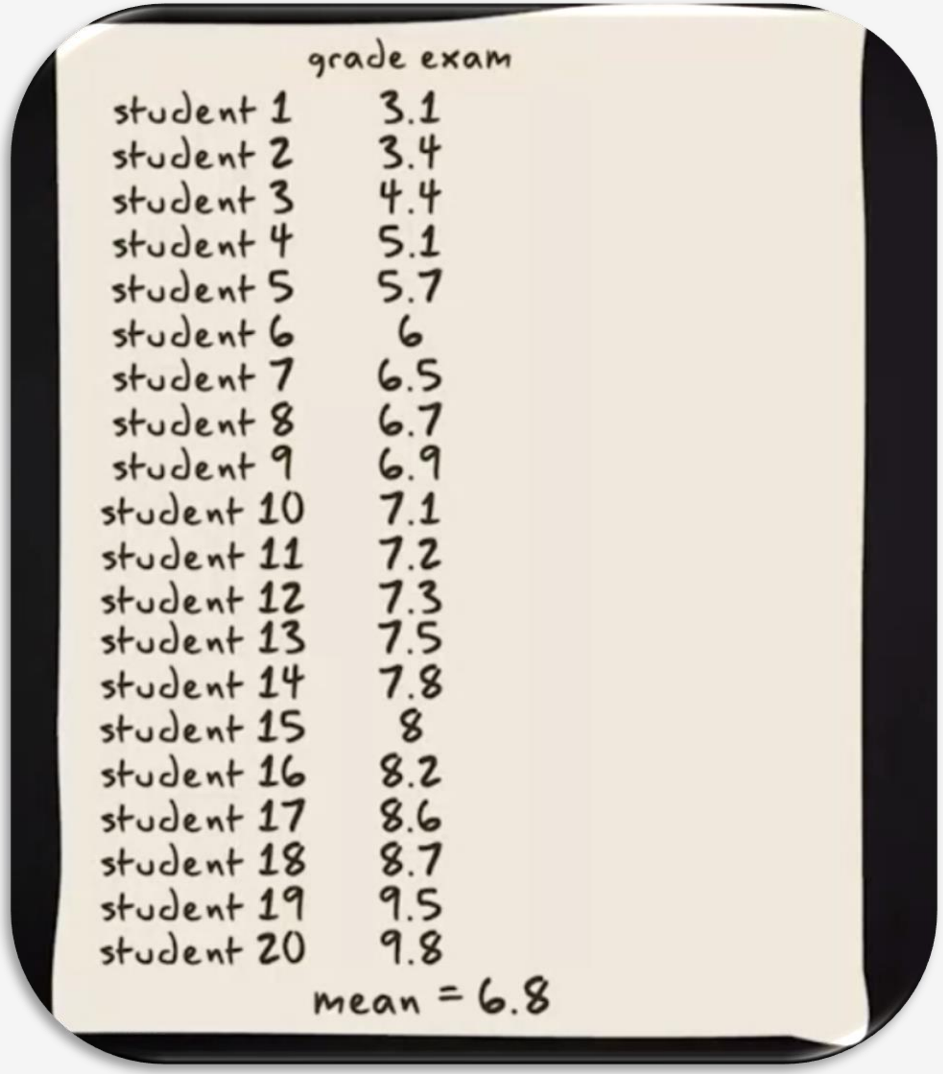
Imagine you're in a class together with 99 other students and you have just finished a statistics exam.

Your professor already has the results for 20 randomly selected students. The professor wants to share the grades of these 20 students, but she doesn't want you to know who these 20 students are.

Because the results are anonymous, you don't know which student got which grade. Note that the worst grade you could get is a 0 and the best grade is a 10.

Now, imagine you are asked to predict the grade of the student sitting next to you in the statics class. What would be your prediction?

That would, of course, be the mean of these grades. That is 6.8.



	grade exam
student 1	3.1
student 2	3.4
student 3	4.4
student 4	5.1
student 5	5.7
student 6	6
student 7	6.5
student 8	6.7
student 9	6.9
student 10	7.1
student 11	7.2
student 12	7.3
student 13	7.5
student 14	7.8
student 15	8
student 16	8.2
student 17	8.6
student 18	8.7
student 19	9.5
student 20	9.8

mean = 6.8

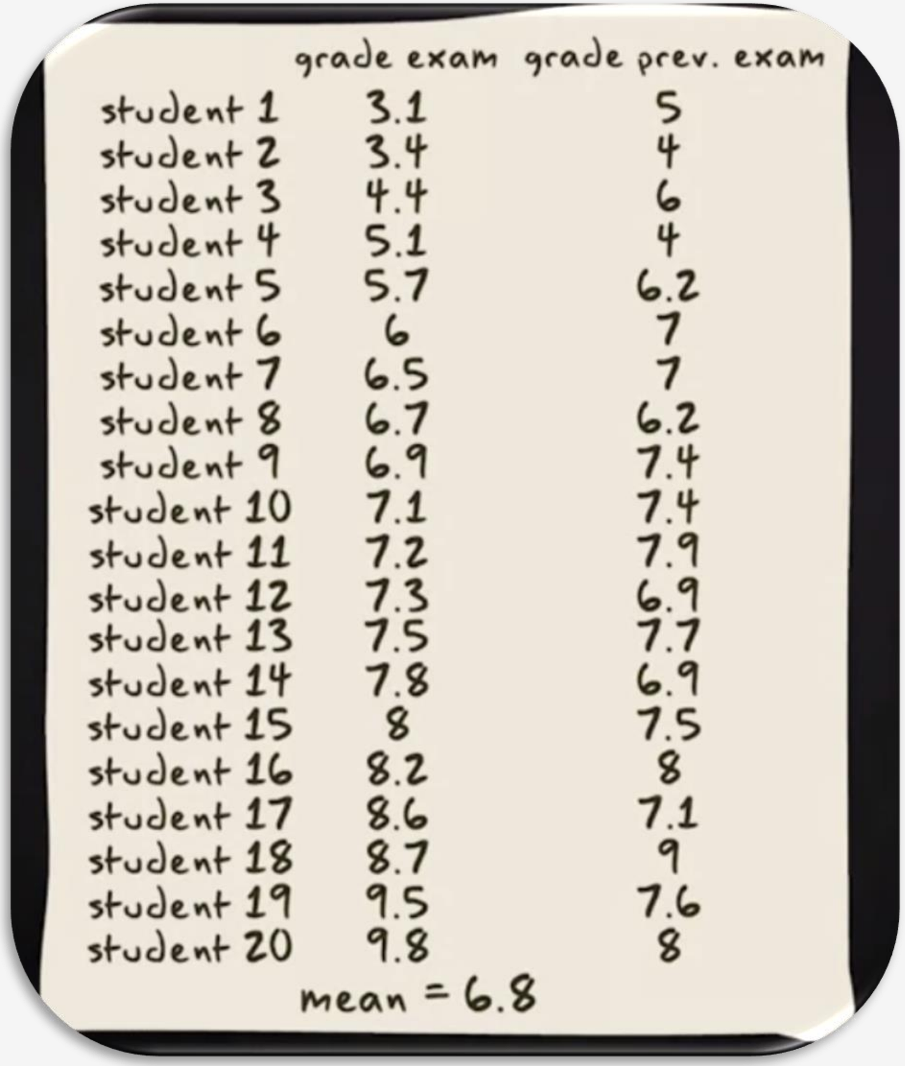
Regression

Now imagine that the professor also gives you the grades these twenty students got in a previous statistics exam, again anonymized.

How would you now predict the grade of the student sitting next to you?

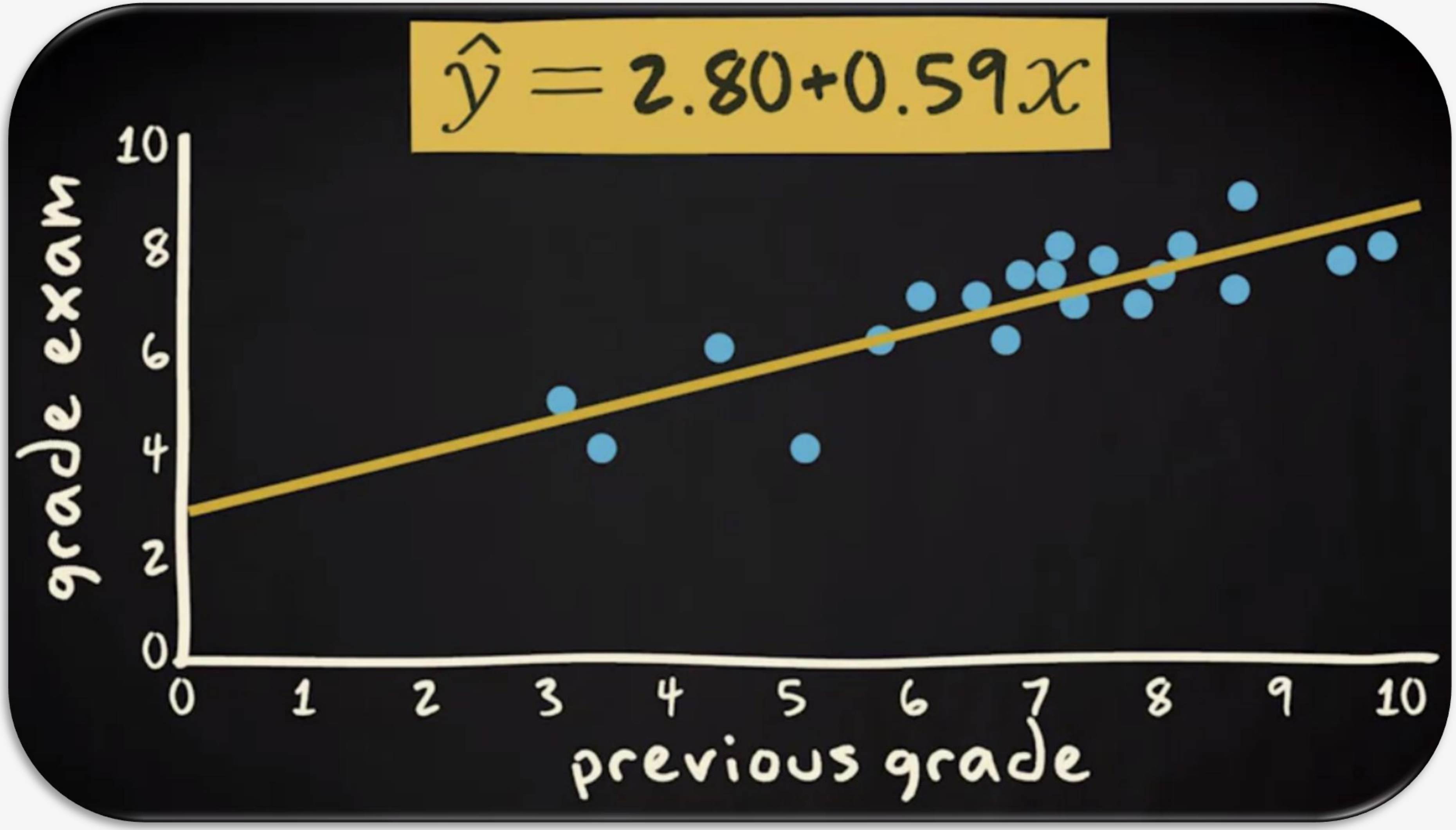
Well, now you can make use of regression analysis.

Here you see the scatterplot with the regression line and the regression equation. You see that those with a high grade for the previous exam also tend to have a high grade for the present exam. In fact, you can use the regression line and the corresponding equation to make a prediction. When you ask your neighbor what his previous grade was, you can use the regression line to predict what most likely would be his present grade.



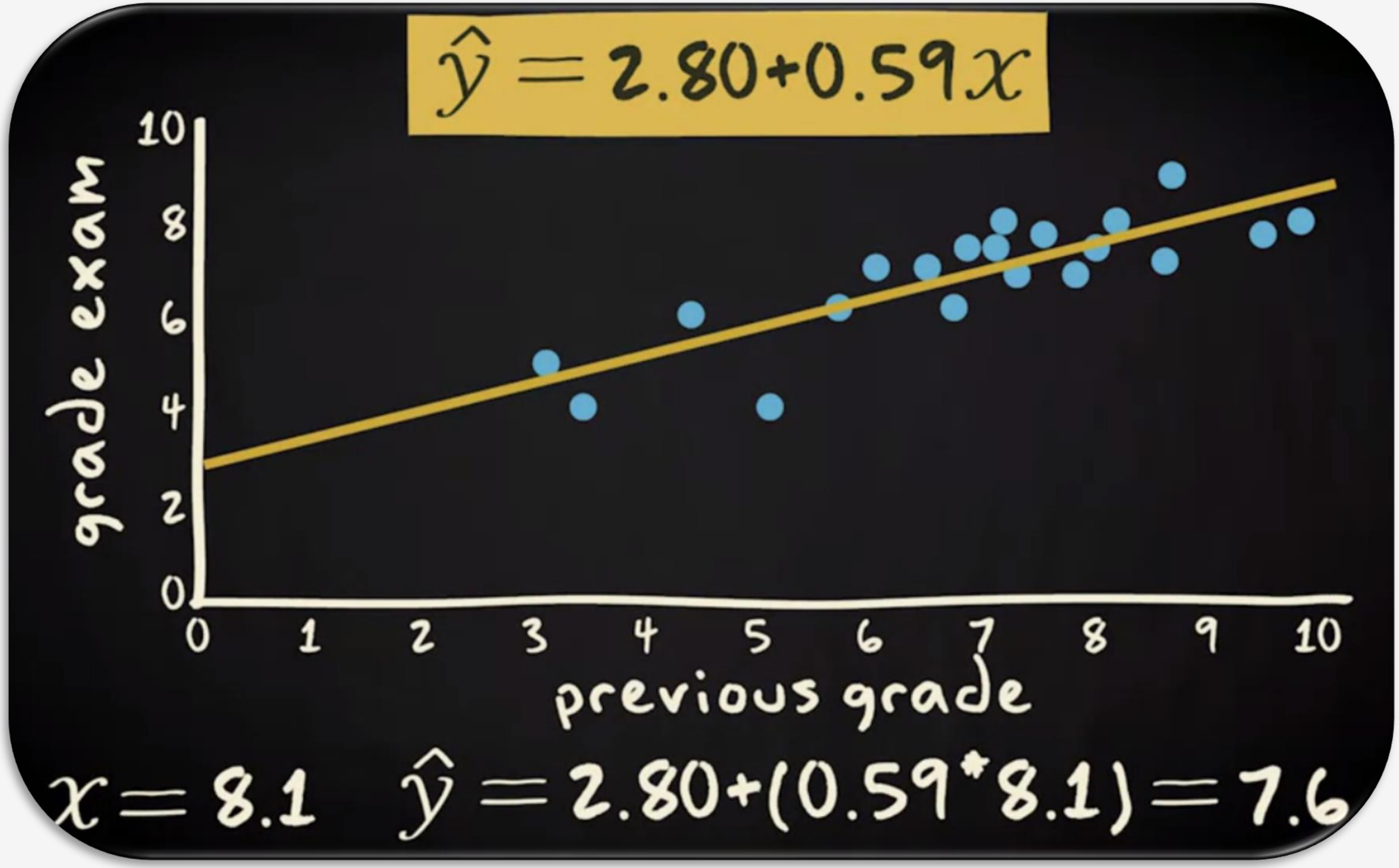
	grade exam	grade prev. exam
student 1	3.1	5
student 2	3.4	4
student 3	4.4	6
student 4	5.1	4
student 5	5.7	6.2
student 6	6	7
student 7	6.5	7
student 8	6.7	6.2
student 9	6.9	7.4
student 10	7.1	7.4
student 11	7.2	7.9
student 12	7.3	6.9
student 13	7.5	7.7
student 14	7.8	6.9
student 15	8	7.5
student 16	8.2	8
student 17	8.6	7.1
student 18	8.7	9
student 19	9.5	7.6
student 20	9.8	8
	mean = 6.8	

Regression



$\hat{y} = 2.80 + 0.59x$
grade exam
0
2
4
6
8
10
0
1
2
3
4
5
6
7
8
9
10
previous grade

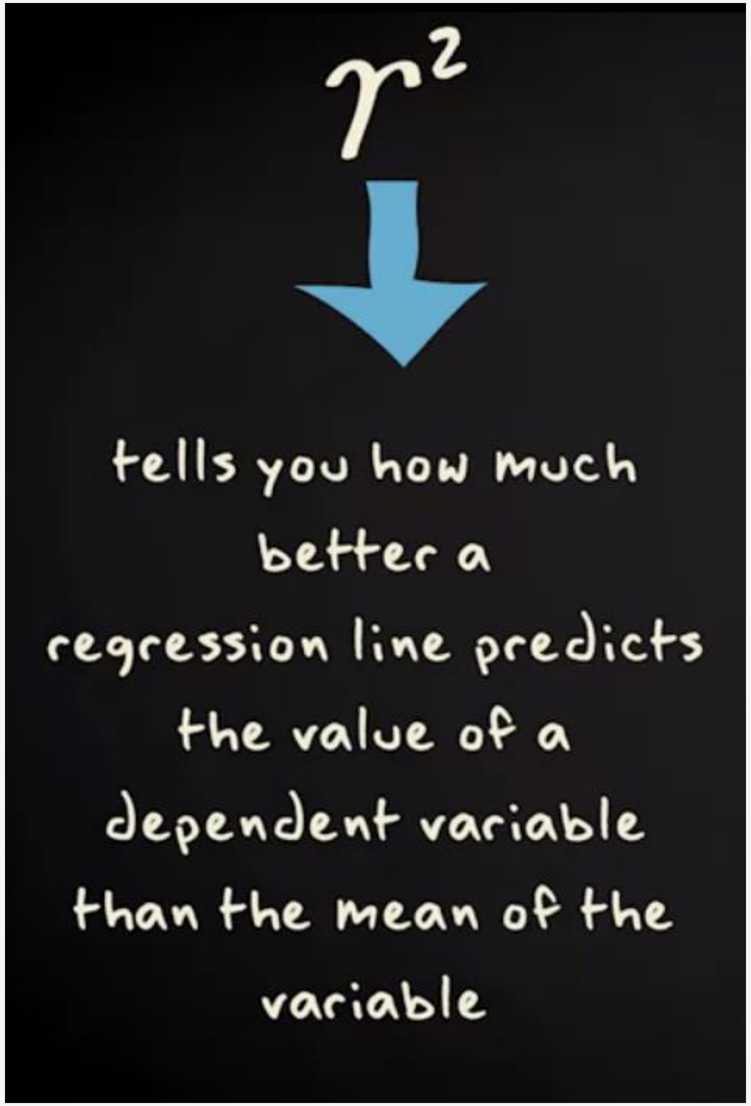
Regression


$\hat{y} = 2.80 + 0.59x$
10
8
6
4
2
0
grade exam
0
1
2
3
4
5
6
7
8
9
10
previous grade
$x = 8.1$ $\hat{y} = 2.80 + (0.59 * 8.1) = 7.6$

Regression

What does this mean? When you have information about only one variable, the predictions you make are much less accurate than when you have information about two related variables.

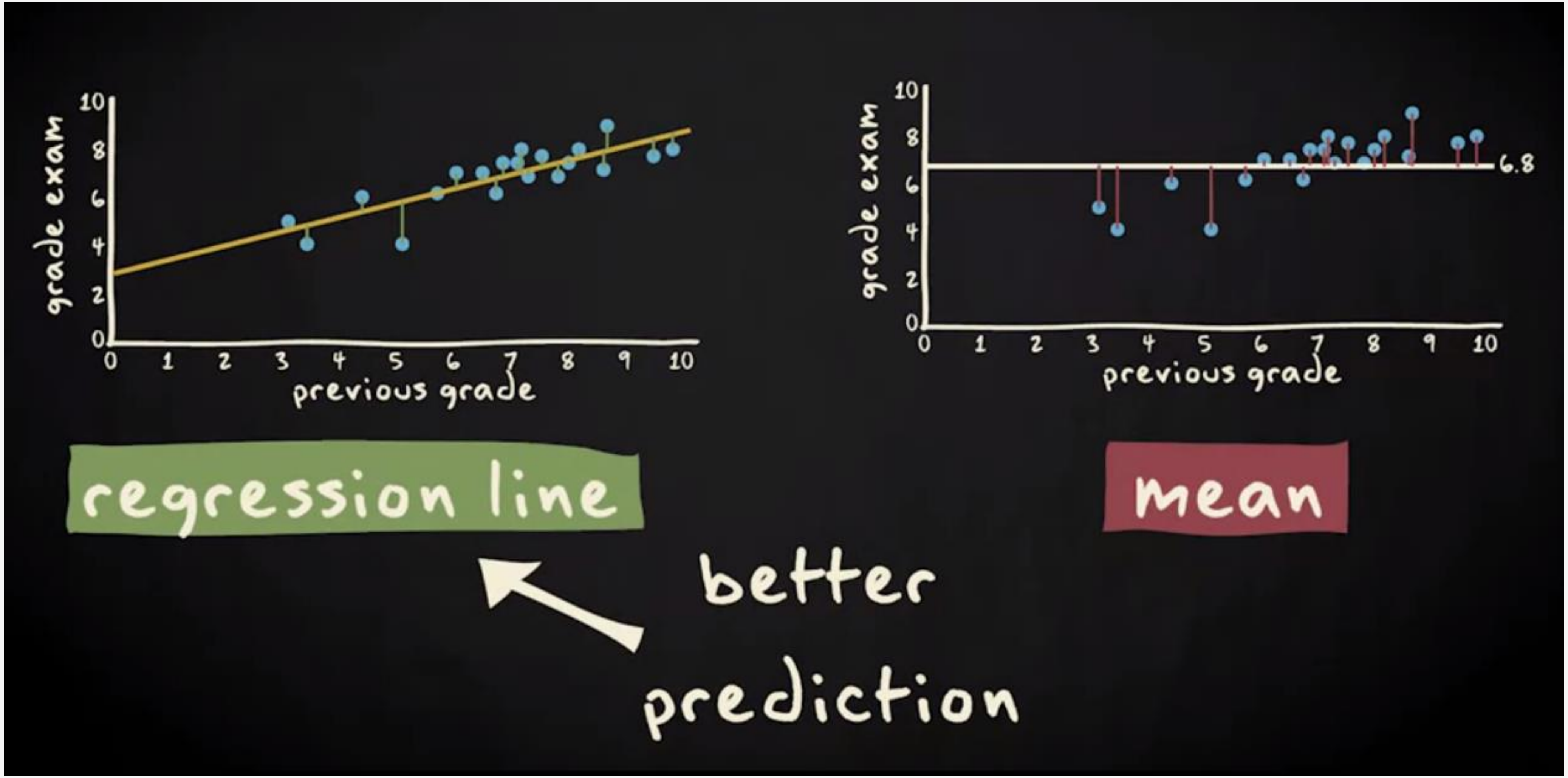
R-squared is nothing more than a number that tells you how much better a regression line predicts the value of a dependent variable than the mean of that variable.



r^2

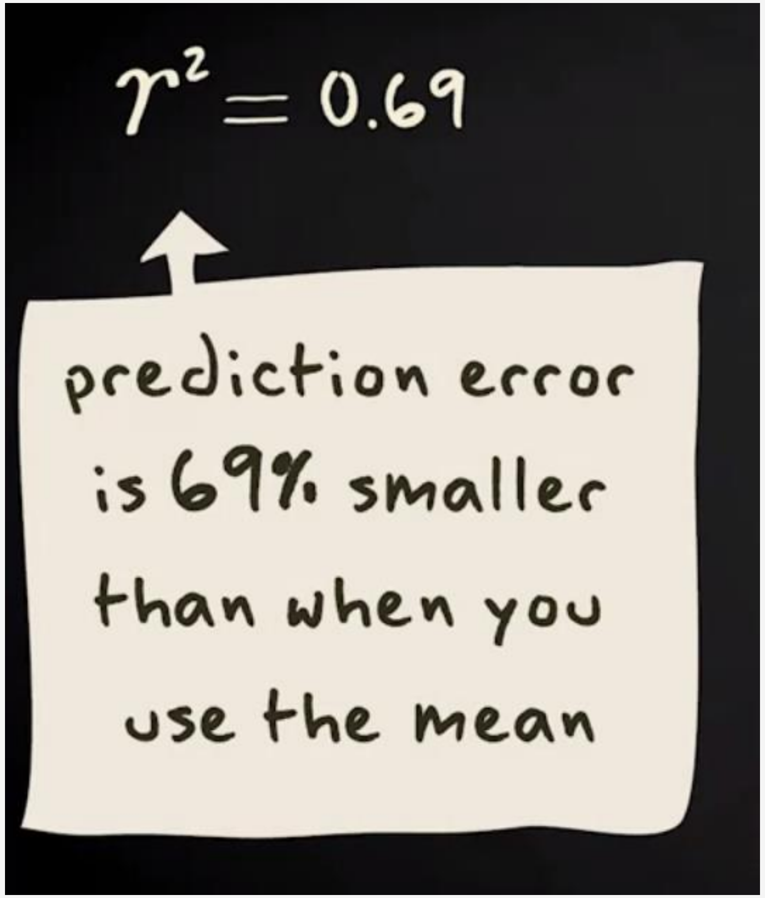
tells you how much better a regression line predicts the value of a dependent variable than the mean of the variable

Regression



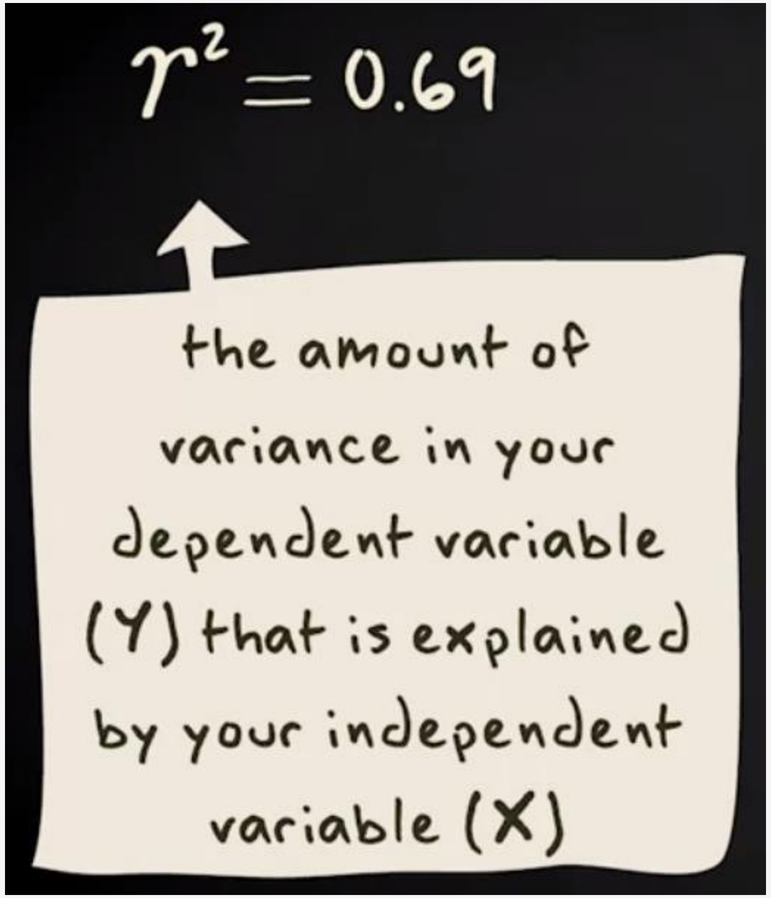
grade exam
10
8
6
4
2
0
0 1 2 3 4 5 6 7 8 9 10
previous grade
grade exam
10
8
6
4
2
0
6.8
0 1 2 3 4 5 6 7 8 9 10
previous grade
regression line
mean
better
prediction

Regression



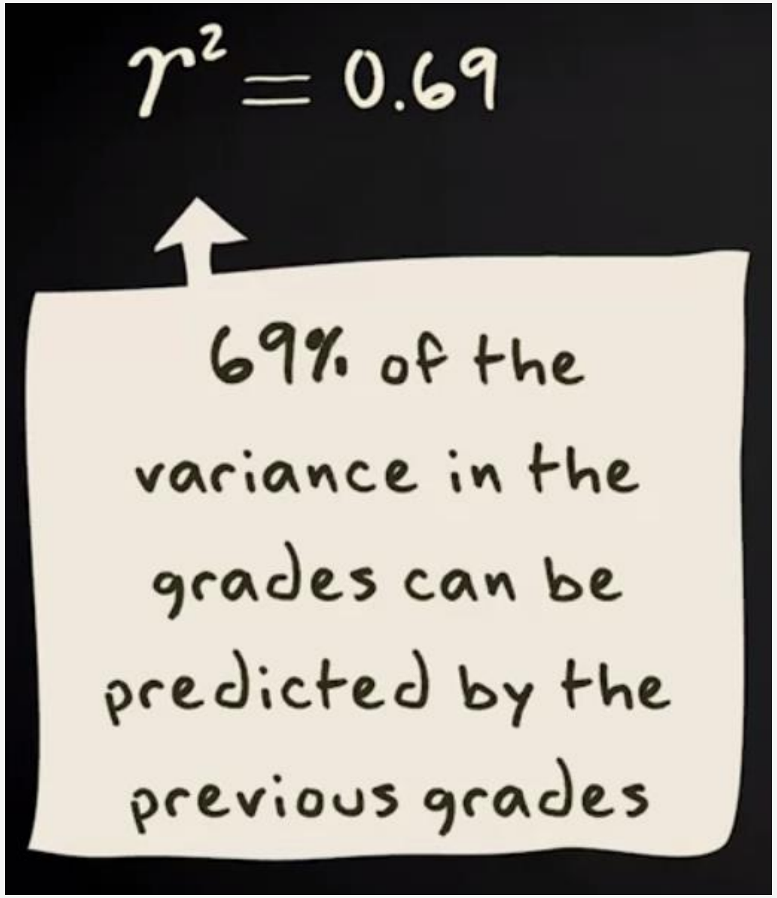
$r^2 = 0.69$

prediction error is 69% smaller than when you use the mean



$r^2 = 0.69$

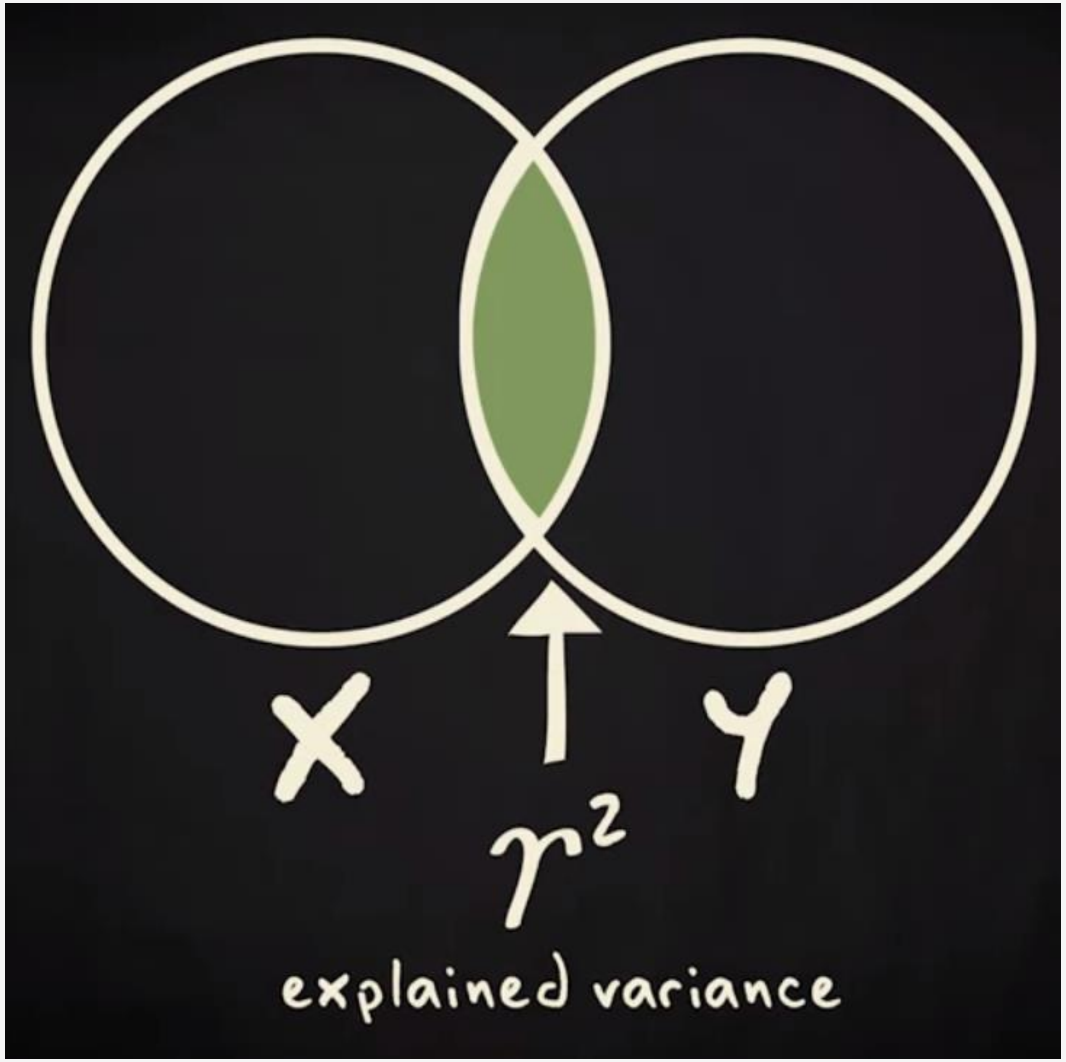
the amount of variance in your dependent variable (Y) that is explained by your independent variable (X)

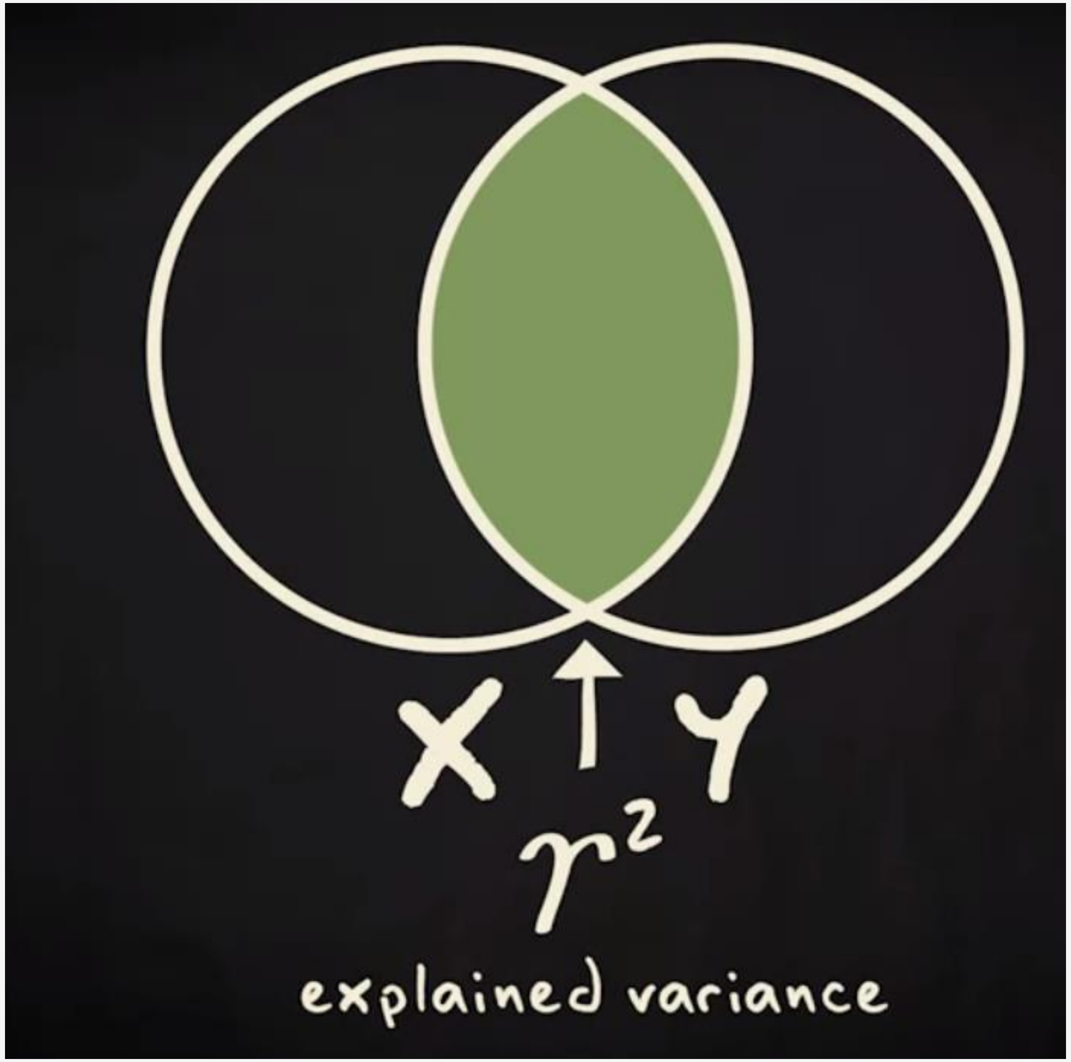


$r^2 = 0.69$

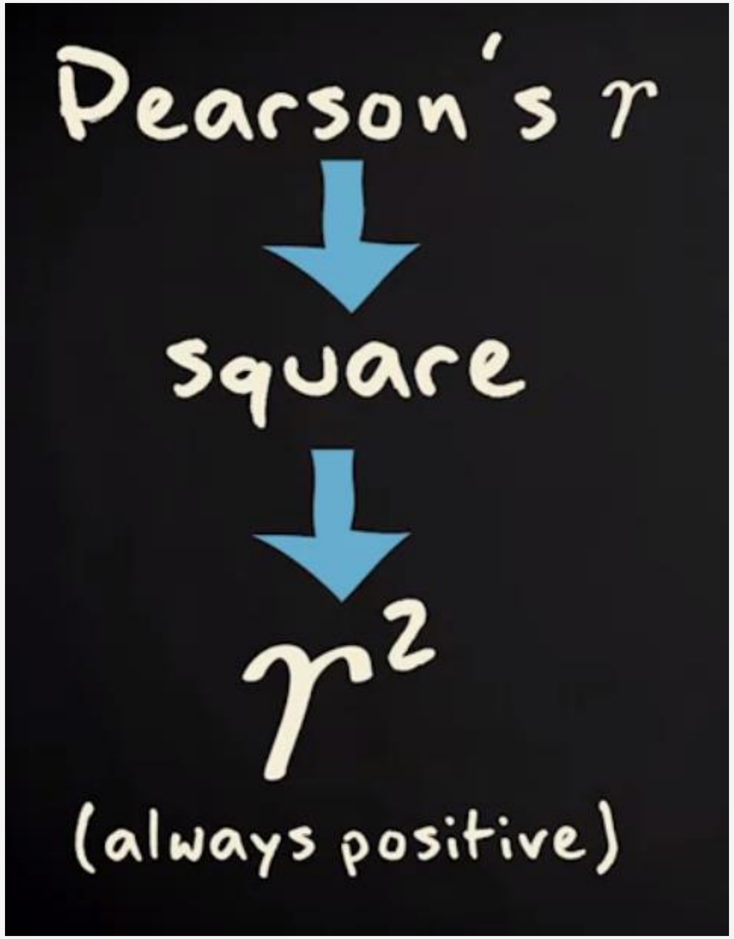
69% of the variance in the grades can be predicted by the previous grades

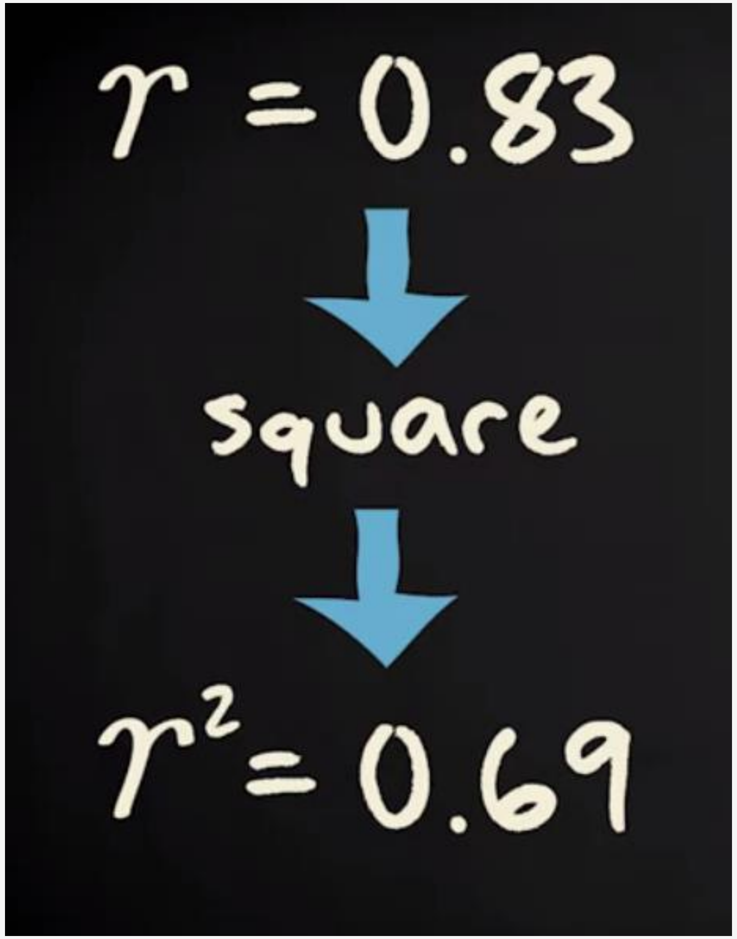
Regression



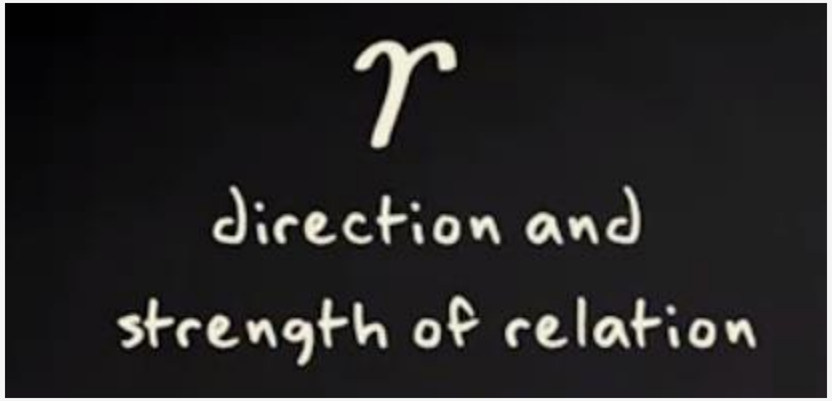


Regression



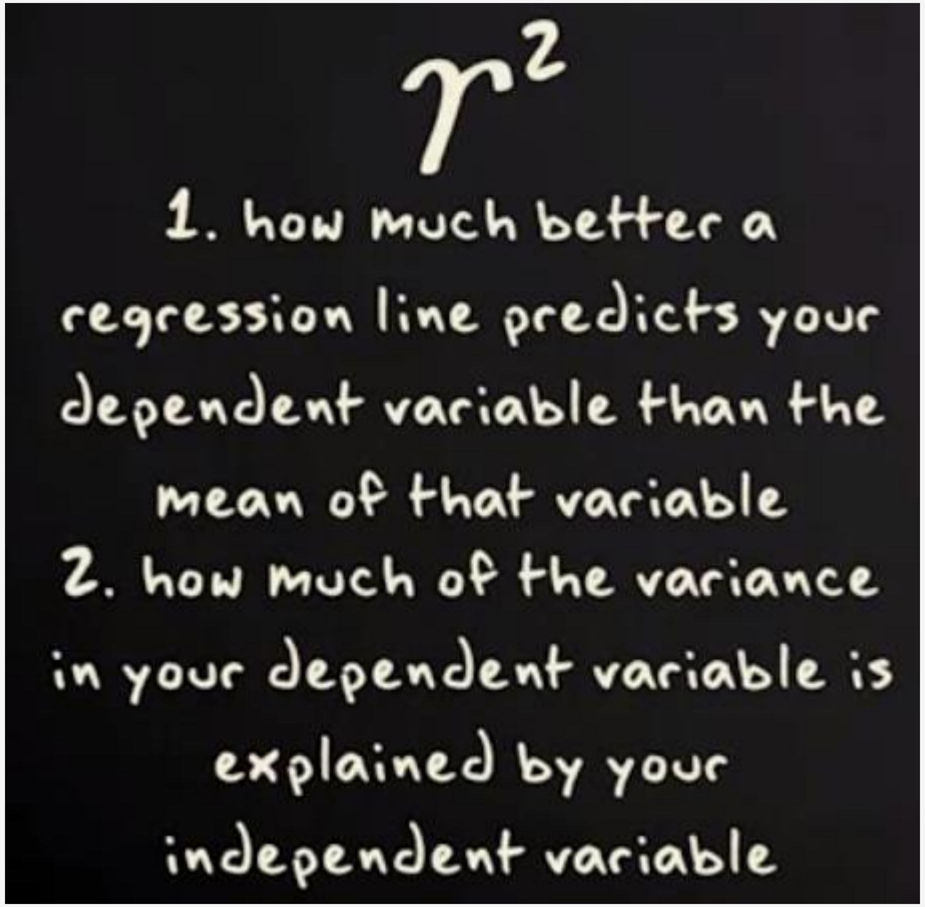


Regression



r

direction and strength of relation



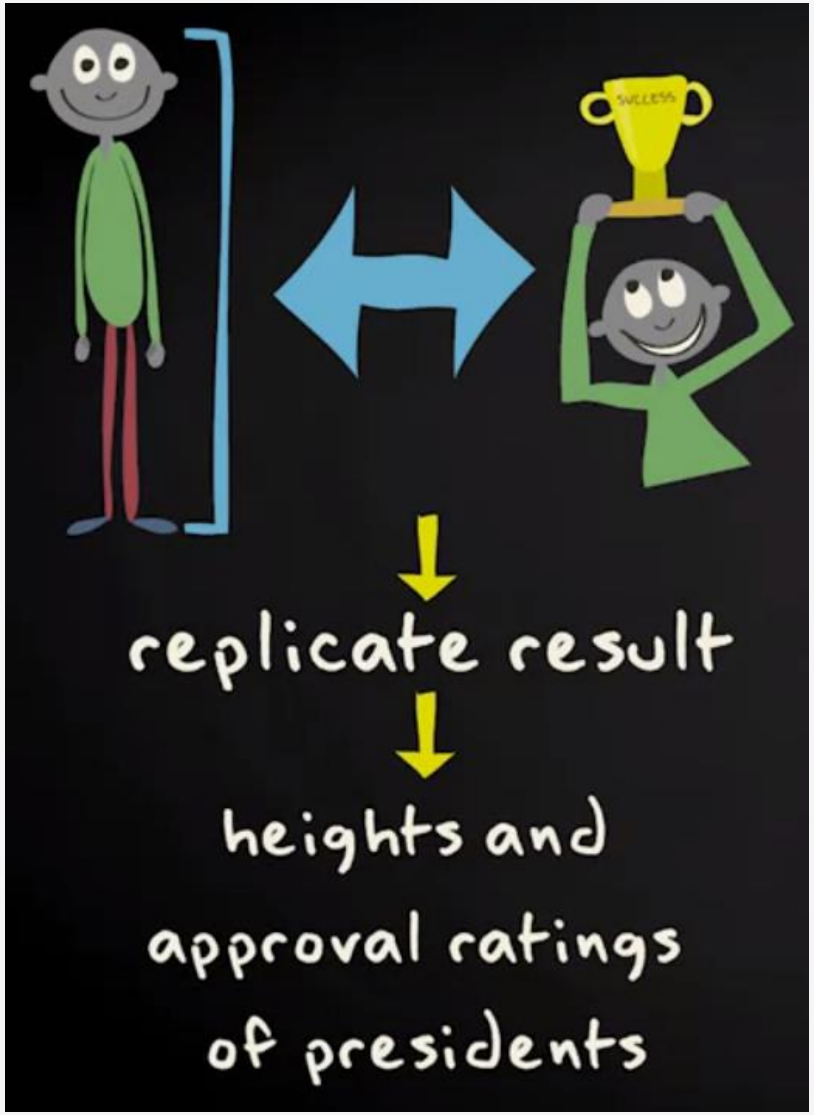
r^2

1. how much better a regression line predicts your dependent variable than the mean of that variable
2. how much of the variance in your dependent variable is explained by your independent variable

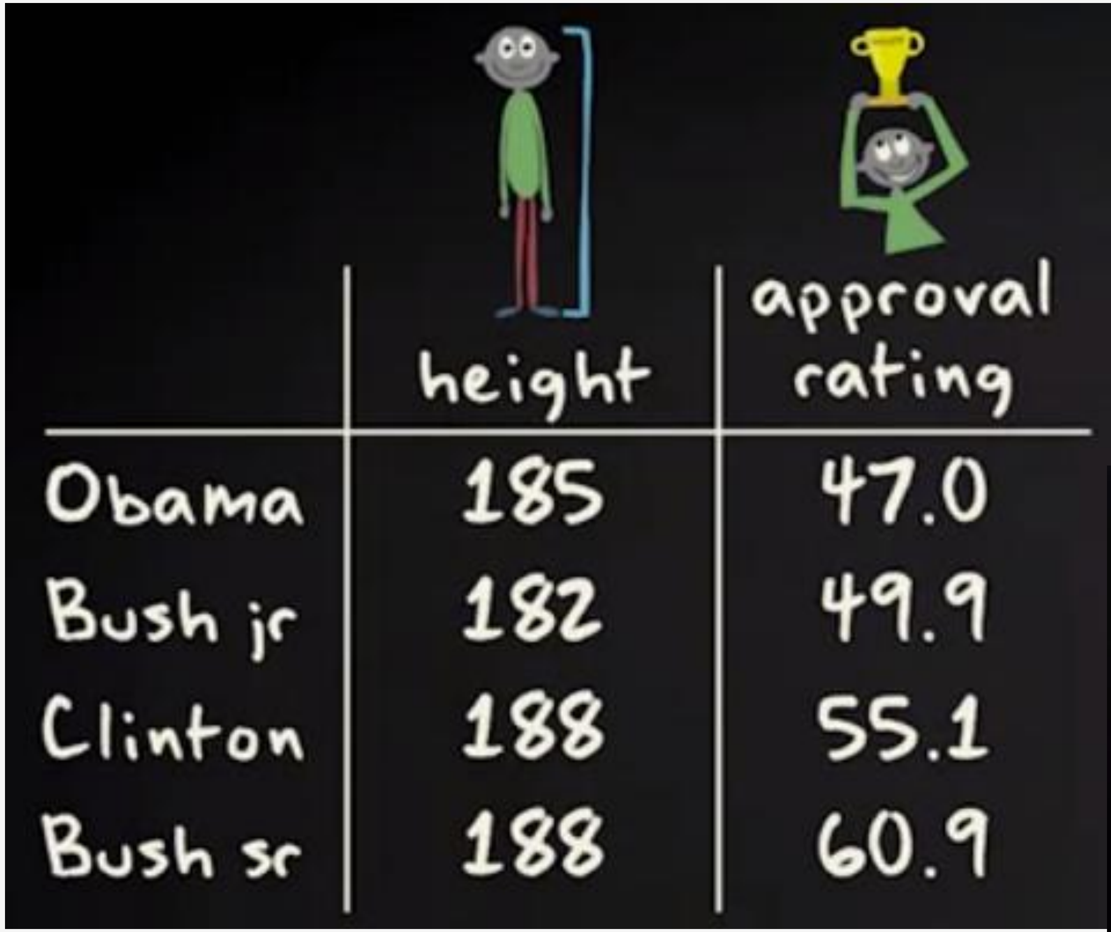
Exercise

Social scientists have shown that a leader's physical height is related to his or her success. Suppose you want to test if you can replicate this result. To do that, you look at the heights and the average approval ratings of the four most recent presidents of the United States.

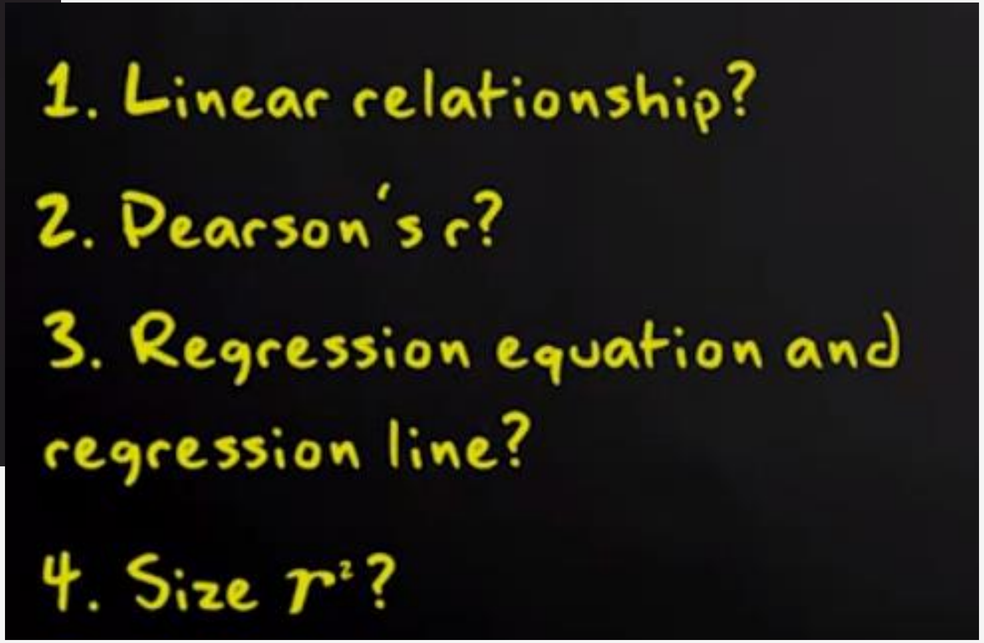
You employ this data matrix and your goal is to answer 4 related questions.



Exercise



	height	approval rating
Obama	185	47.0
Bush jr	182	49.9
Clinton	188	55.1
Bush sr	188	60.9



1. Linear relationship?
2. Pearson's r?
3. Regression equation and regression line?
4. Size r^2?

Sample and sampling

Sample and sampling

By now you know that we can do all kinds of **univariate analyses** (e.g., compute modes, means, and standard deviations) and **bivariate analyses** (e.g., compute Pearson's r correlation coefficients or do regression analyses). Usually, all these analyses are fully based on your sample. In general, the methods for analyzing sample data are called methods of descriptive statistics.

Yet in real life we're often not so much interested in samples but in populations. We therefore often use data obtained from a sample to draw conclusions about an entire underlying population. If we employ sample data to draw inferences about a population we are using methods of **inferential statistics**. We use the computed **sample statistics** to draw inferences about the corresponding **population parameters**.

Sample and sampling

It is therefore of essential importance that you know how you should draw samples. In this section we’ll pay attention to good sampling methods as well as some poor practices. We'll show you how you can draw a **simple random sample** and we'll pay attention to various forms of **bias** you could encounter along the way. We'll also discuss two alternatives to simple random sampling that are almost as good: **random multi-stage cluster sampling** and **stratified random sampling**.

Sample and sampling

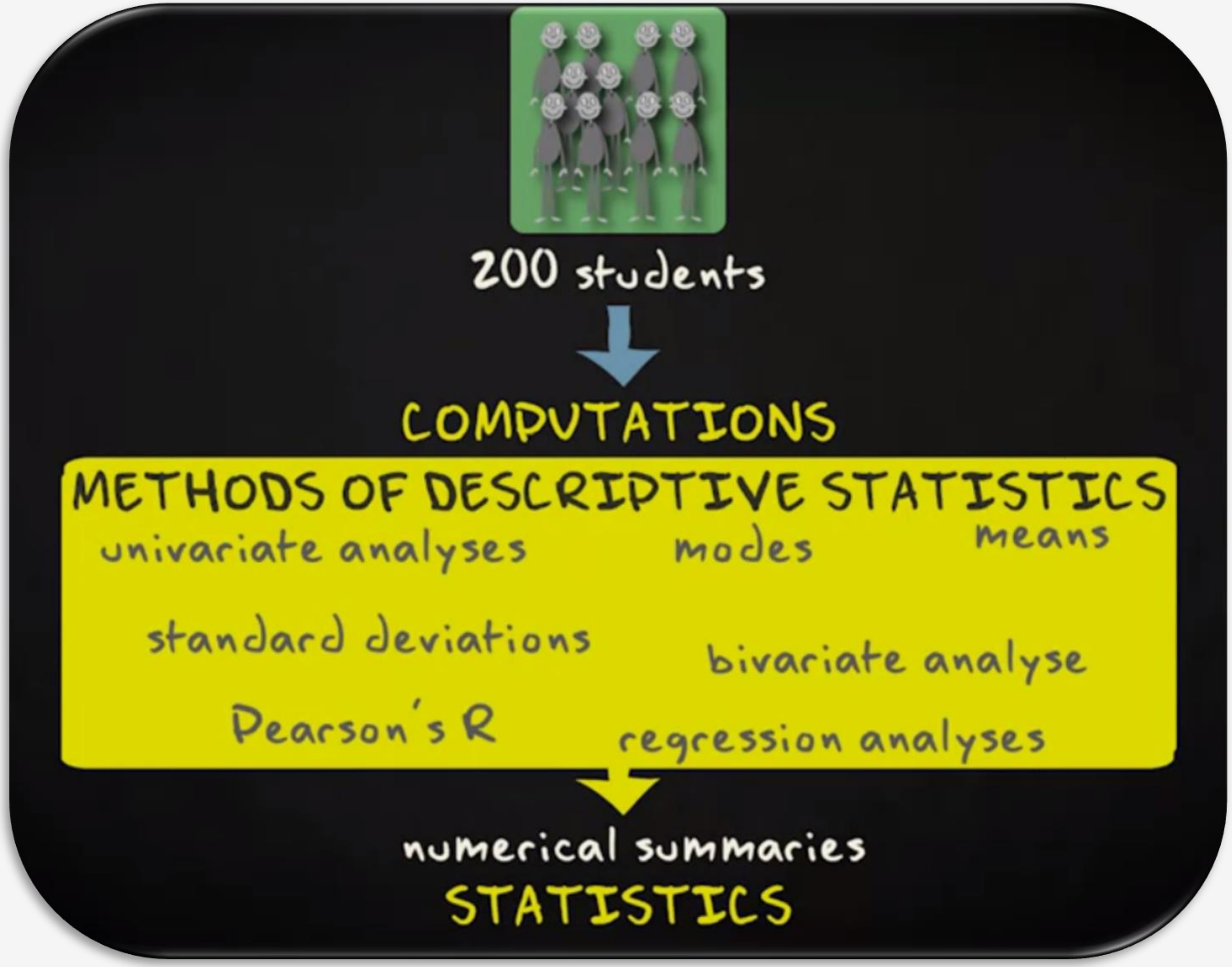
Almost all statistical studies are based on samples. Imagine you want to know to what extent students in London identify themselves as Hipsters.

It is almost impossible to ask all students, so you decide to draw a sample of, say, 200 respondents, and you assess to what extent they see themselves as Hipsters.

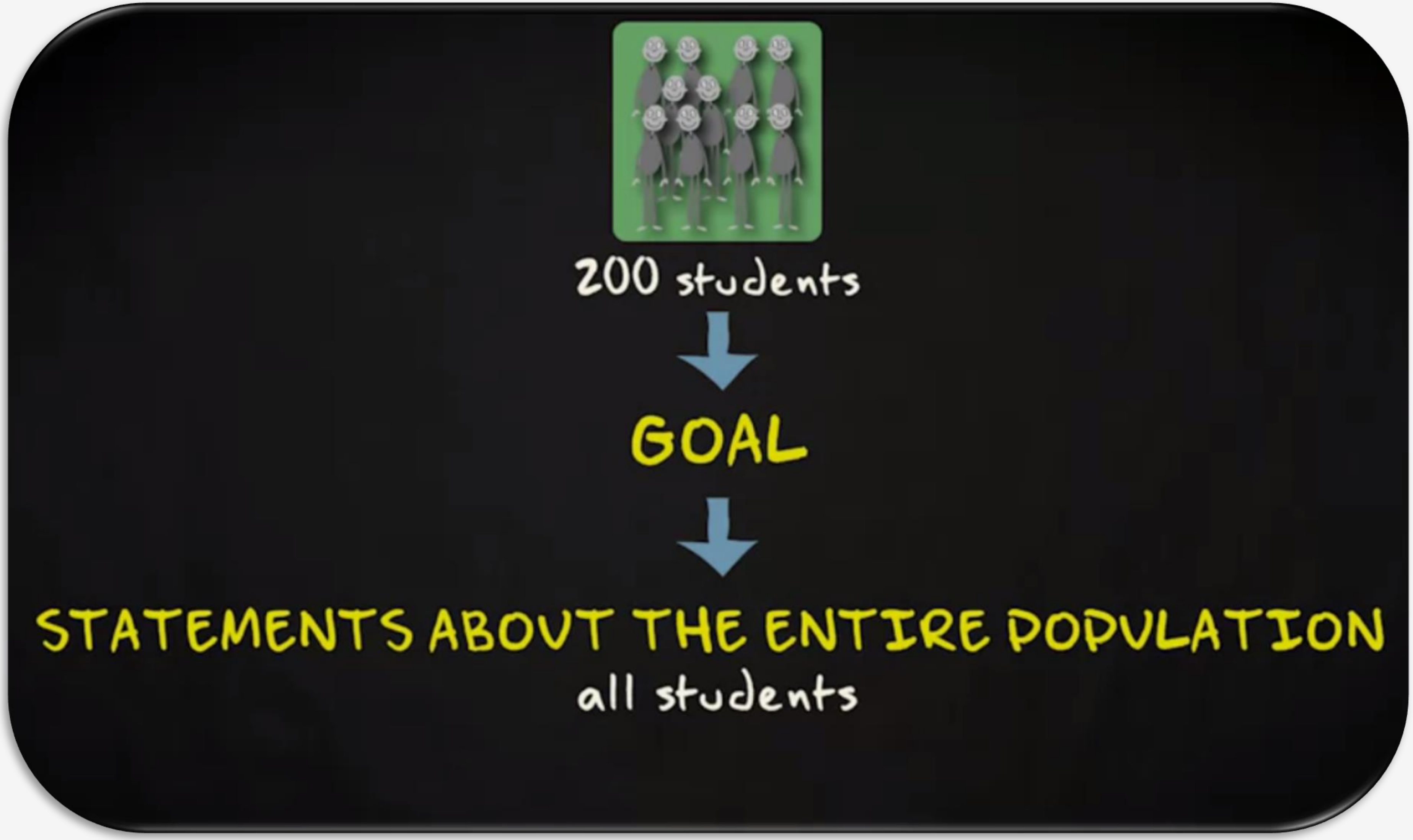
The great thing about statistics is that it can help you to draw conclusions about all students in London (which is the population), based on an analysis of only these 200 respondents (which is the sample)

HIPSTER?
all 300,000 students
POPULATION
HIPSTER?
200 students
SAMPLE

Sample and sampling



Sample and sampling



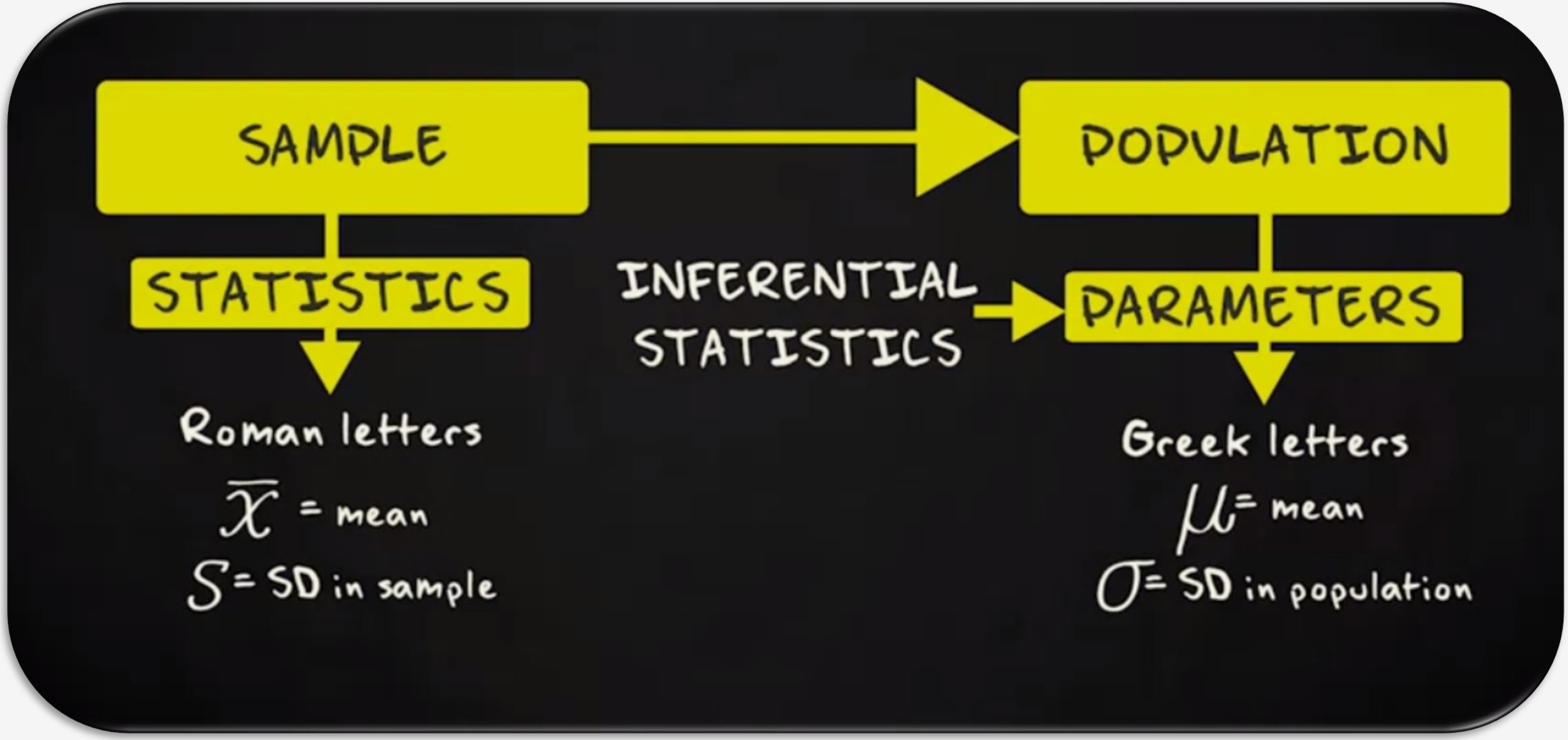
200 students

GOAL

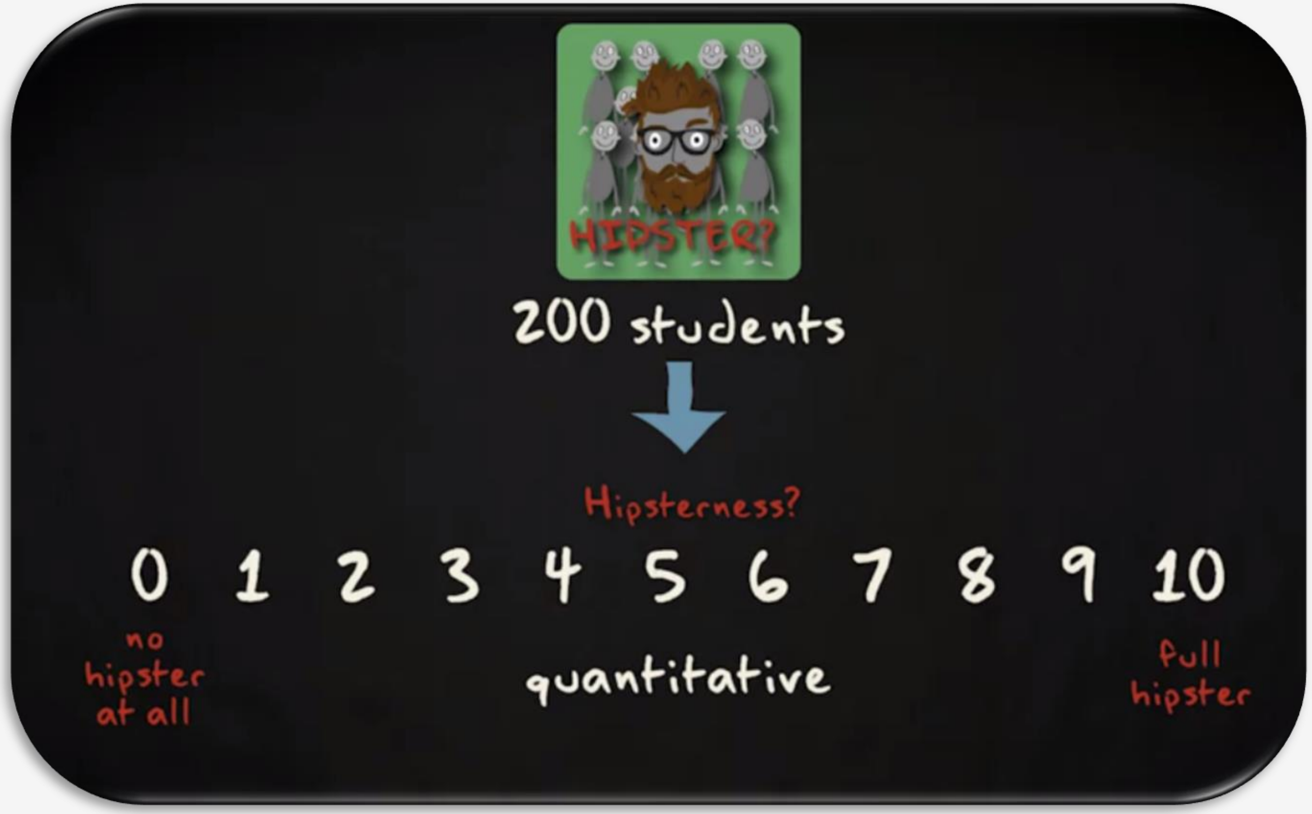
STATEMENTS ABOUT THE ENTIRE POPULATION

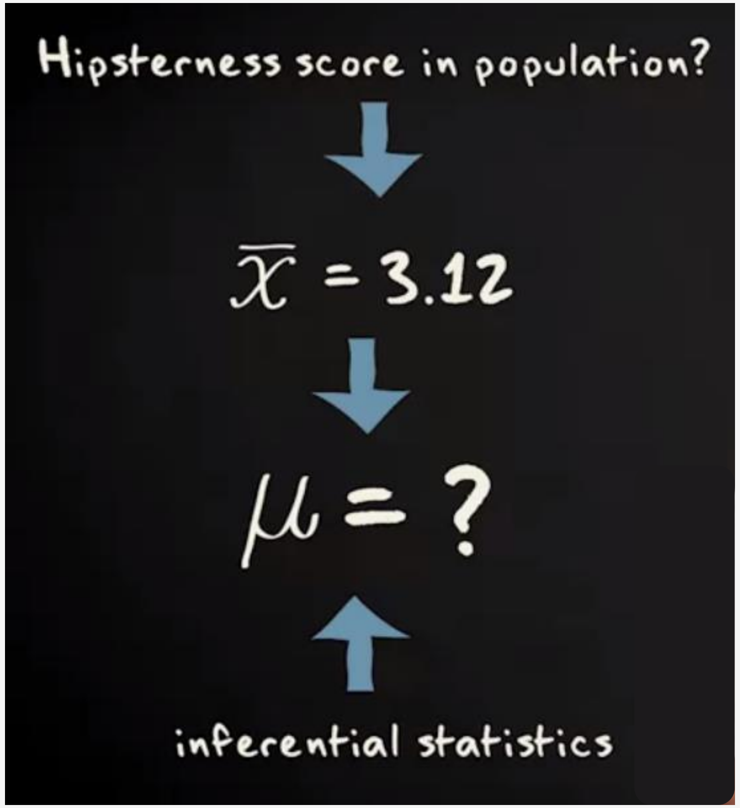
all students

Sample and sampling

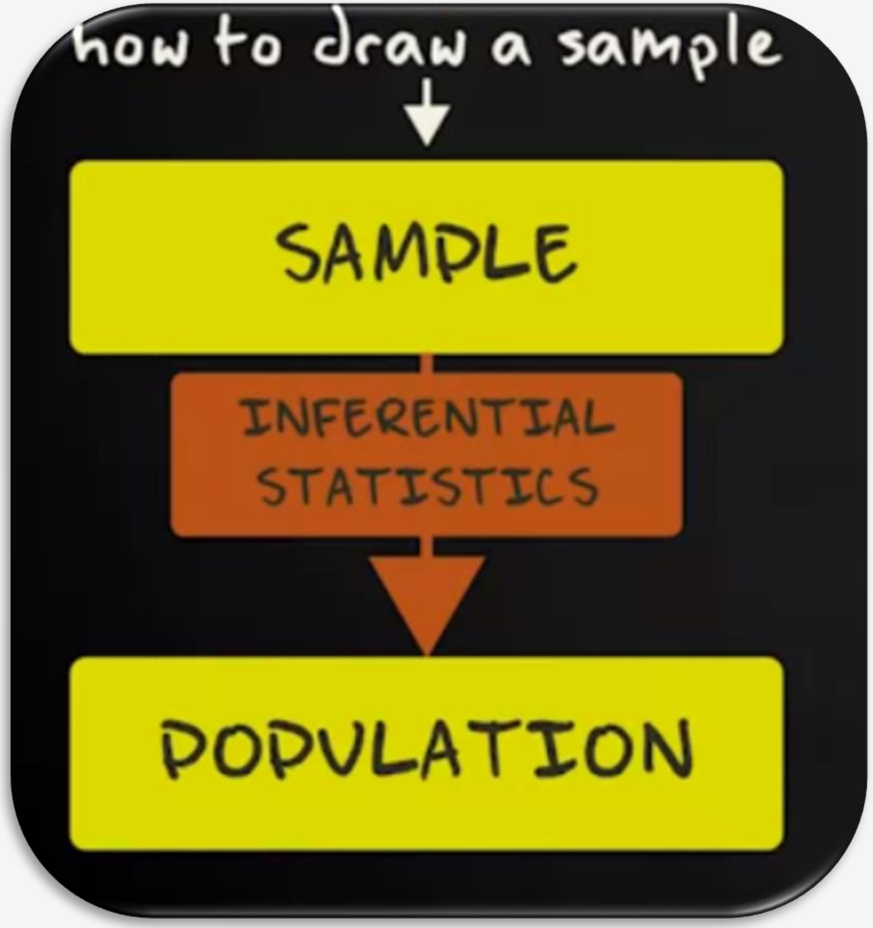


Sample and sampling



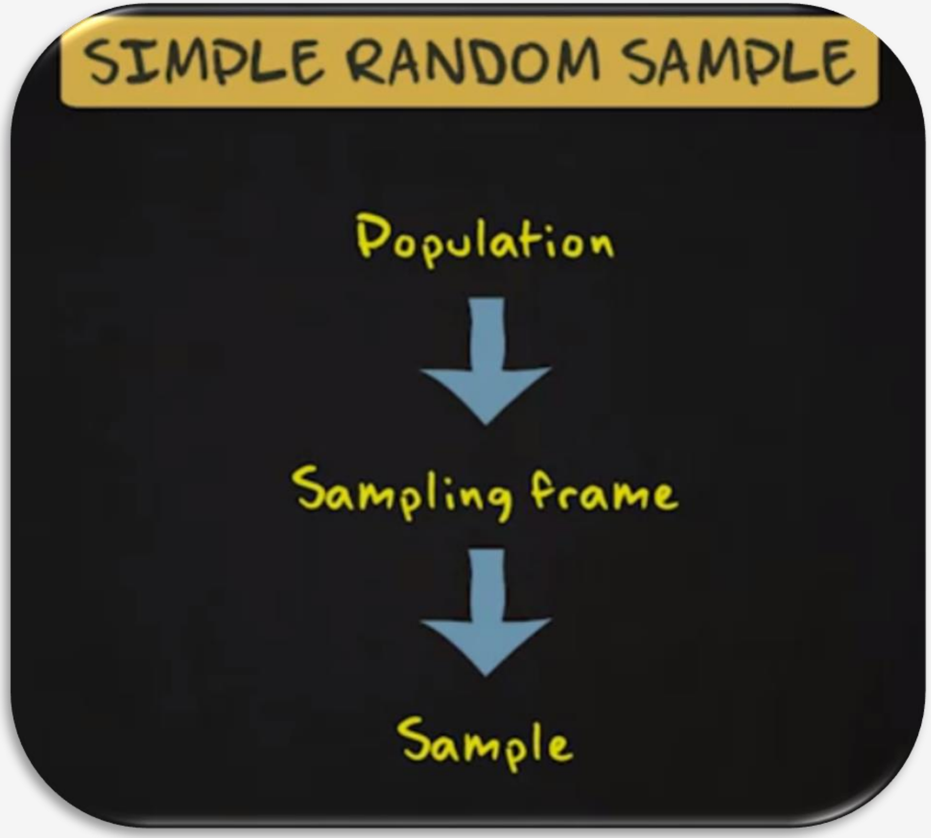


Sample and sampling



Sample and sampling

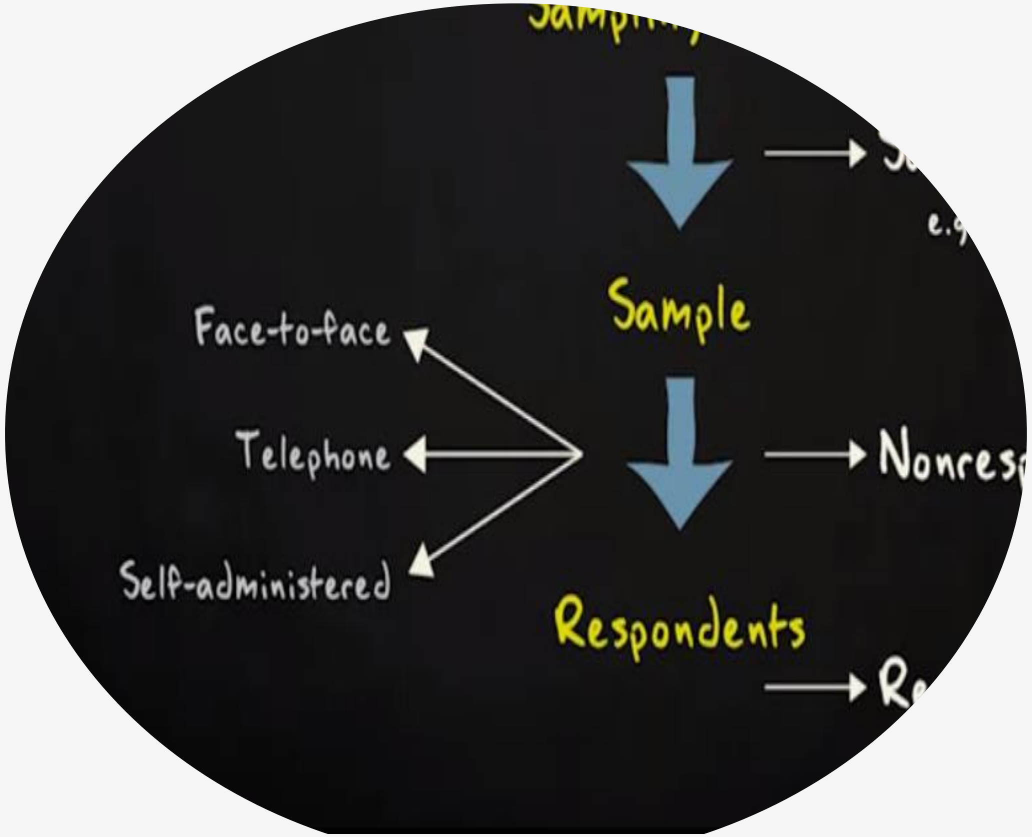
A sample is nothing more than a subset of a population. Yet, for methods of inferential statistics, not every sample is appropriate. What you want is a **representative sample**. A good way to achieve that goal is to draw a **simple random sample**. That means that you make sure that each subject in a population has the same chance of being selected.



Imagine that there is an organization in London that has an overview of all students, including their contact details. Moreover, this organization is willing to share this list with you. You ask a computer to randomly select 200 students out of this list. That’s it. There you have your simple random sample.

Sample and sampling

The next step is to decide how you're going to approach your 200 respondents

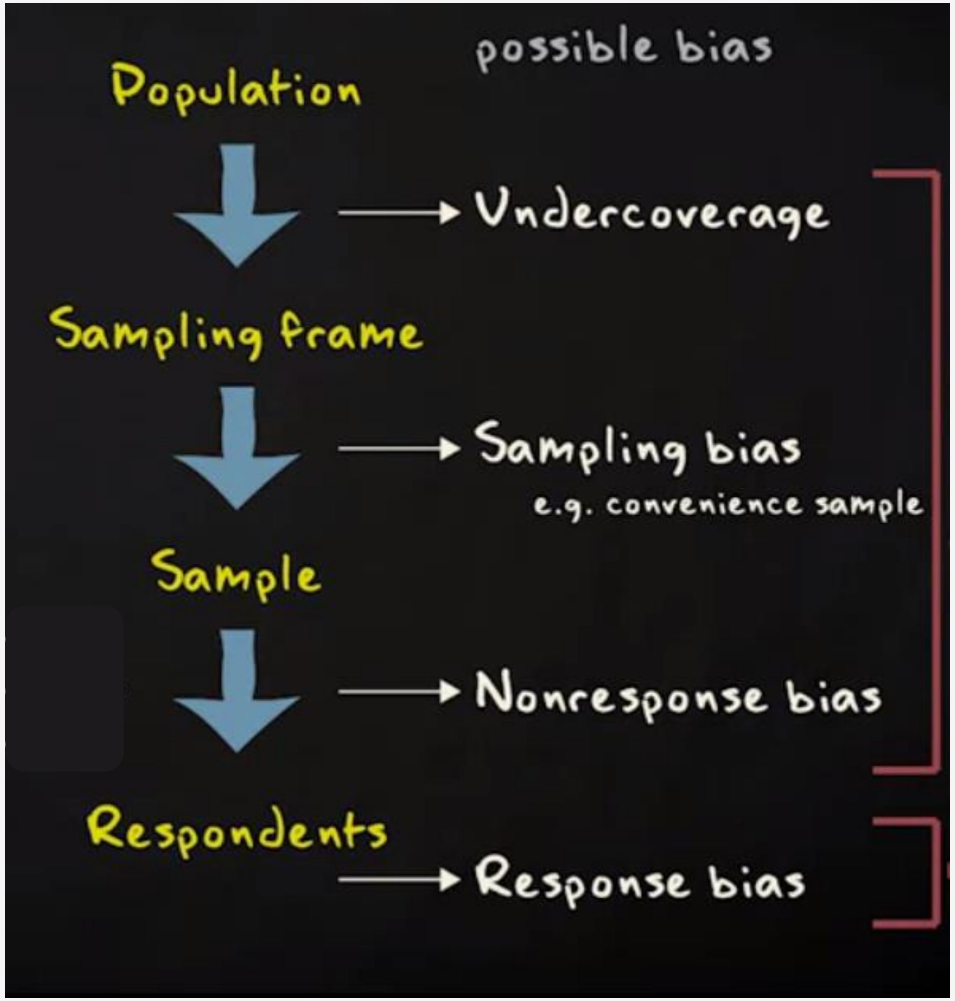


Sample and sampling

Along the way you will encounter various possible forms of bias.

The first one is **undercoverage**. This means that not everyone in the population is included in the sampling frame.

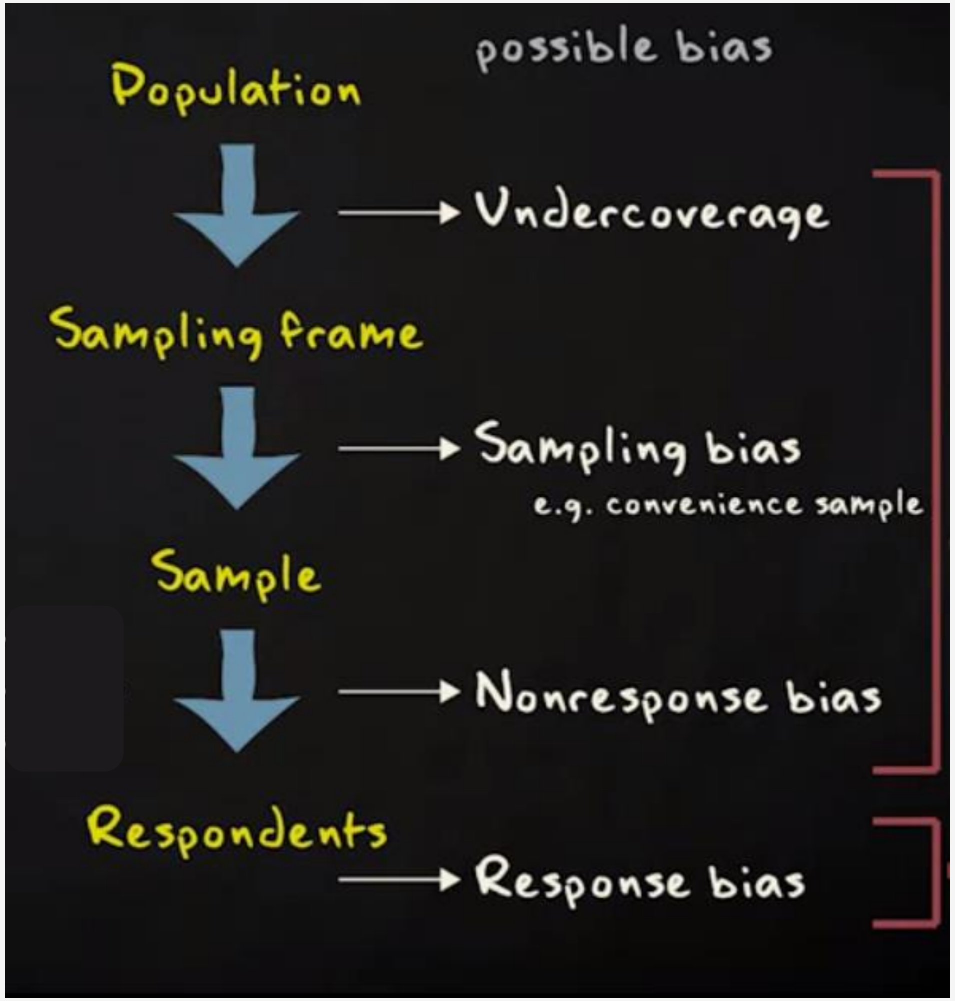
There can also be **sampling bias**. This means that not every person in your sampling frame is equally likely to be included in the sample. This is what happens if you fail to draw a random sample. This is the case if, for instance, you randomly approach people on the street. This is what we call a convenience sample.



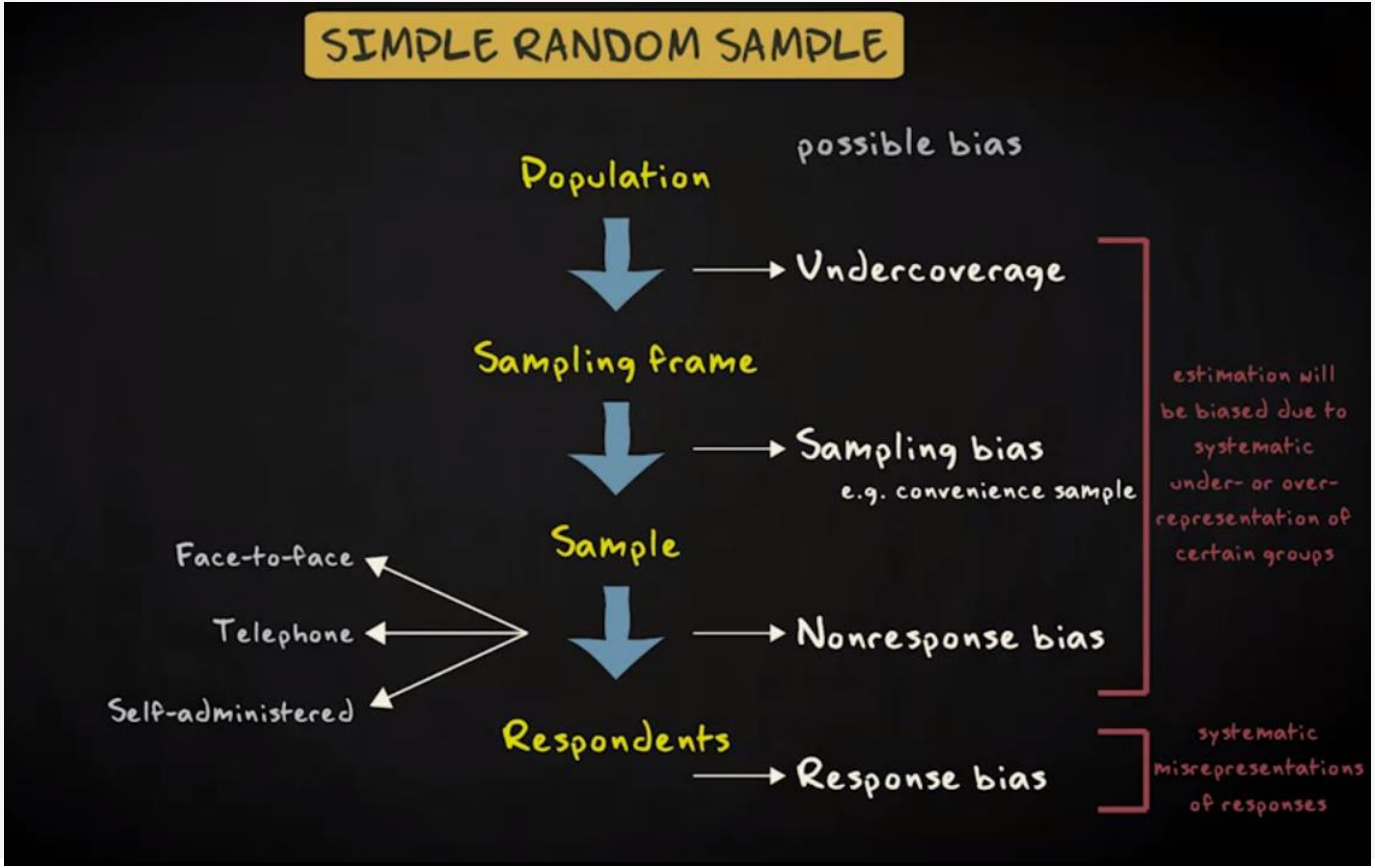
Sample and sampling

Thirdly, once you have your sample, another possible form of bias is **nonresponse bias**. Some selected subjects might refuse to participate, or they might simply be unreachable. Some respondents who have agreed to participate might not be willing to respond to particular questions.

Finally, there can be **response bias**. In this case the actual given responses are biased. This could for instance be due to an interviewer asking leading questions or because respondents think that some answers are socially unacceptable. For instance, a student might identify as a Hipster but tell an interviewer that she doesn't because she thinks that the interviewer doesn't like Hipsters. I



Sample and sampling

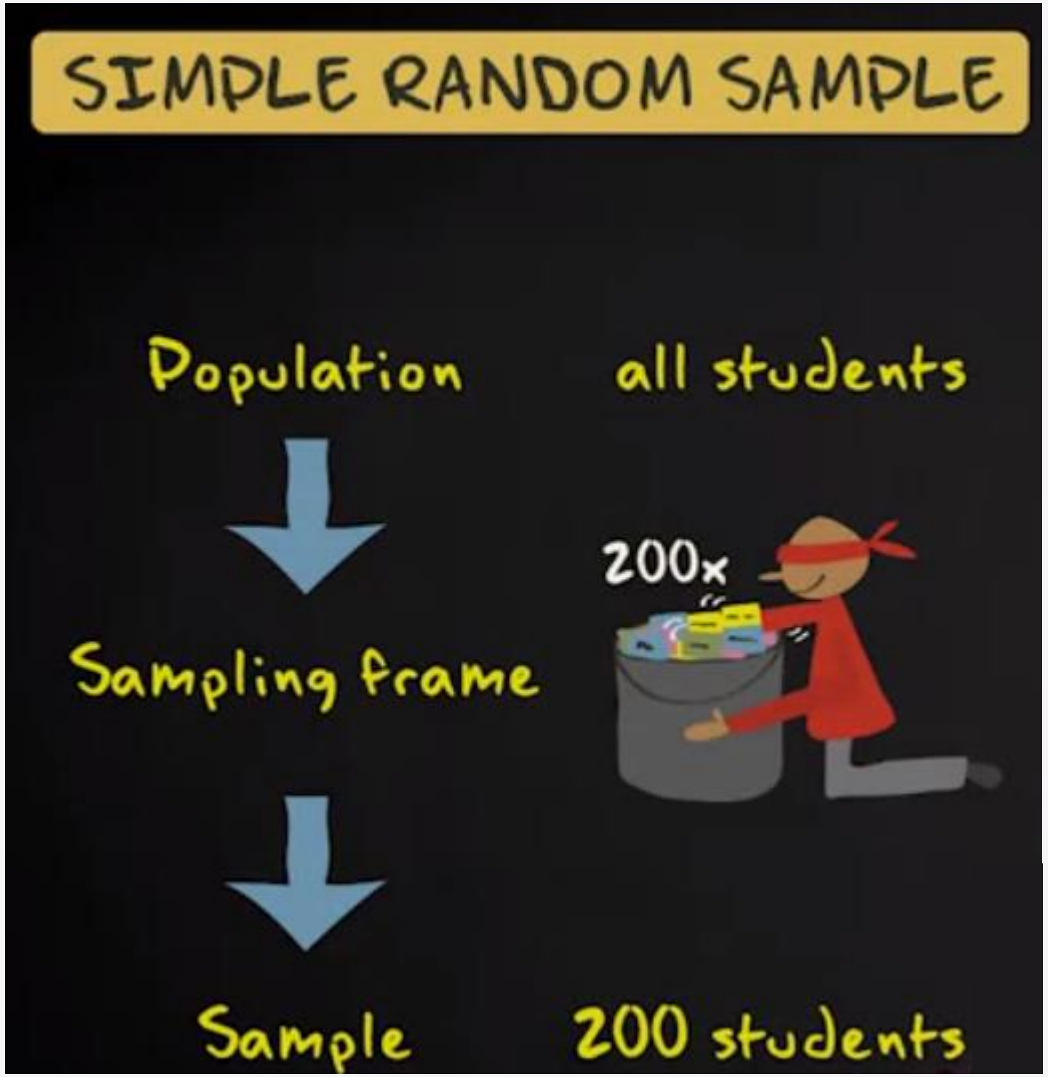


Sample and sampling

So, when drawing a sample, you should make sure that your sample is a simple random sample and that you keep these forms of bias to a minimum.

However, many times it will be almost impossible to draw a simple random sample. Luckily, there are **two other ways** of random sampling that are almost as good.

Sample and sampling



Sample and sampling

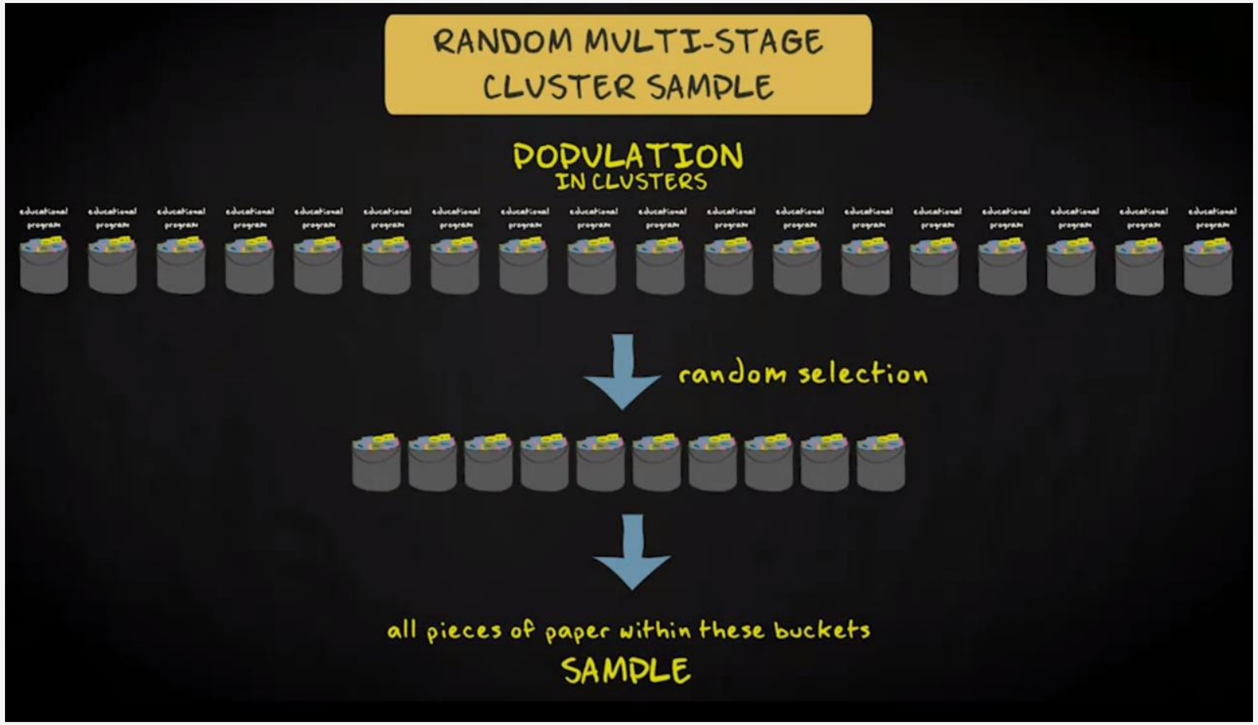
The first alternative is a **random multi-stage cluster sample**.

It works as follows.

First you identify a large number of clusters within your population; for instance, the various educational programs in London in which the students are enrolled. Every program is represented by a bucket and you put the pieces of paper with the names of students in the buckets of the programs in which they are enrolled.

Next, you randomly pick a number of buckets. Say… 10. Then, you select all pieces of paper within these buckets. That's your sample.

A random multi-stage cluster sample is a good choice if you don't have a good sampling frame or if drawing a simple random sample would be very expensive.

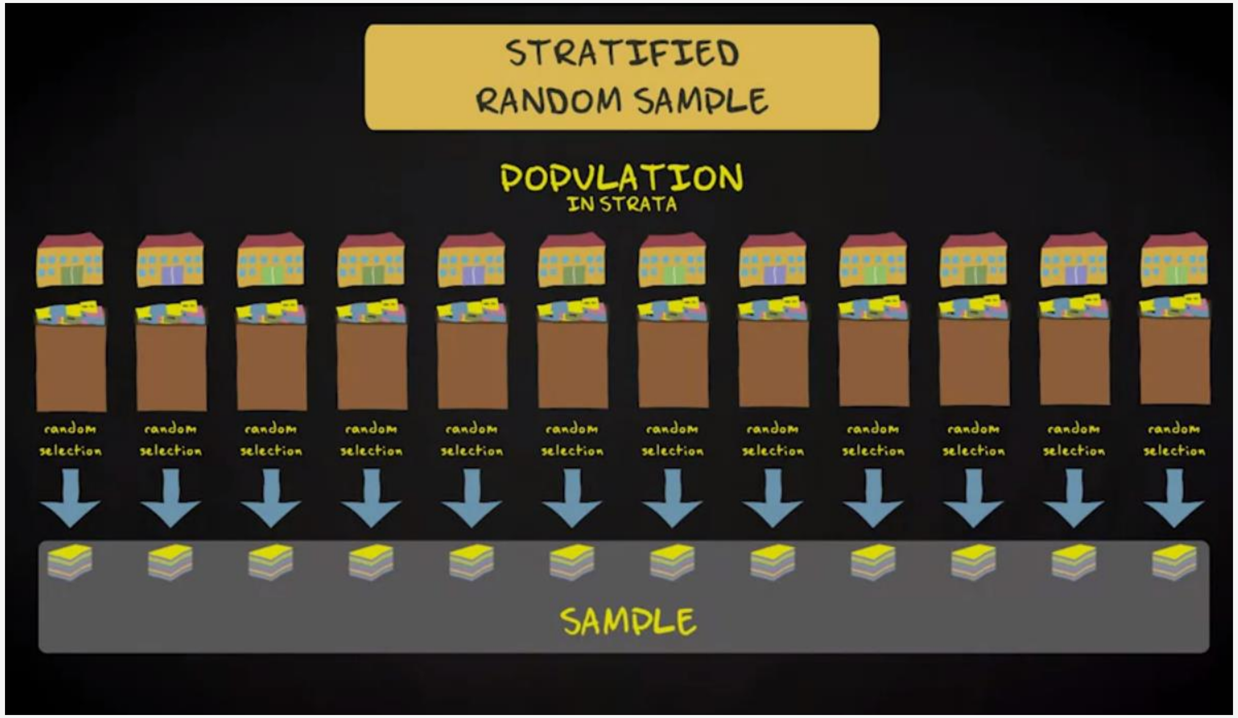

RANDOM MULTI-STAGE CLUSTER SAMPLE
POPULATION
IN CLUSTERS
educational program
random selection
all pieces of paper within these buckets
SAMPLE

Sample and sampling

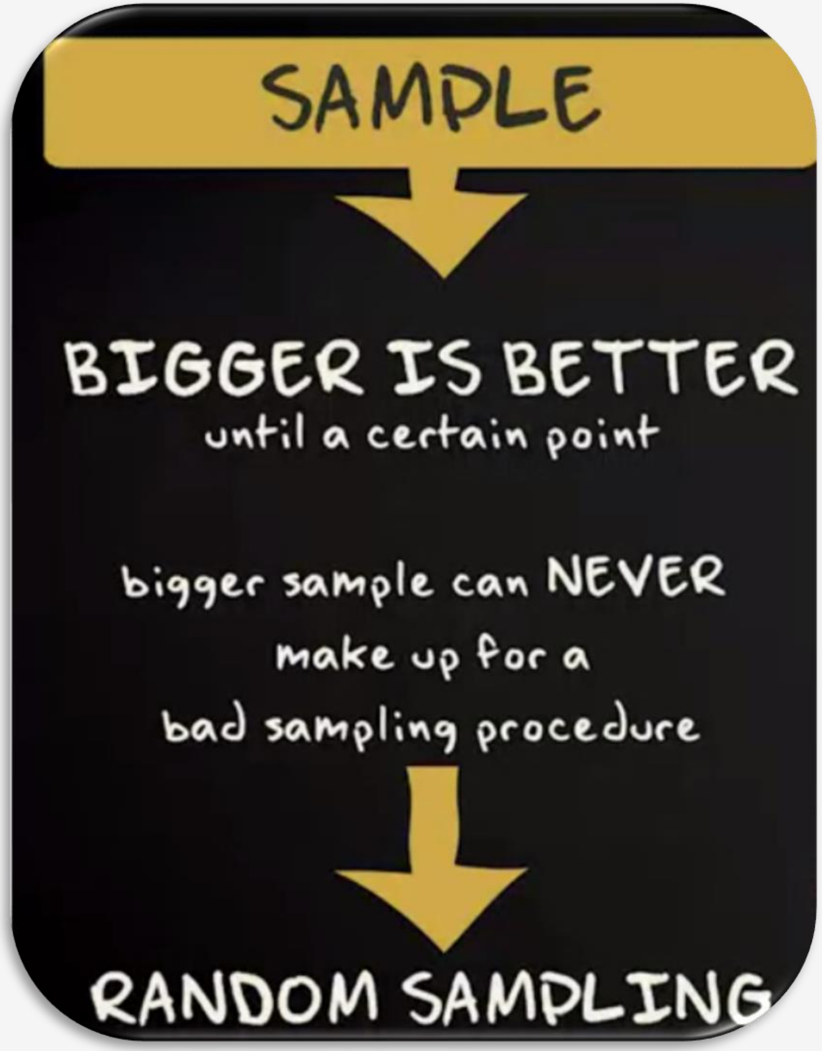
The second alternative is a stratified random sample.

Now you divide the population in separate groups which you call “strata”. For instance, the various universities in London. Every university is represented by a box and you put the pieces of papers with the names of the students in the box of the university where they are registered. Next, you select a simple random sample of pieces of paper from each box. All these pieces of paper together form your sample.

An advantage of this method is that you can make sure that you have enough subjects from every stratum in your sample. A disadvantage is that you need a sampling frame and that you need to know to which stratum each respondent belongs.



Sample and sampling



SAMPLE

BIGGER IS BETTER
until a certain point

bigger sample can NEVER
make up for a
bad sampling procedure

RANDOM SAMPLING

Sampling distribution of sample mean and central limit theorem

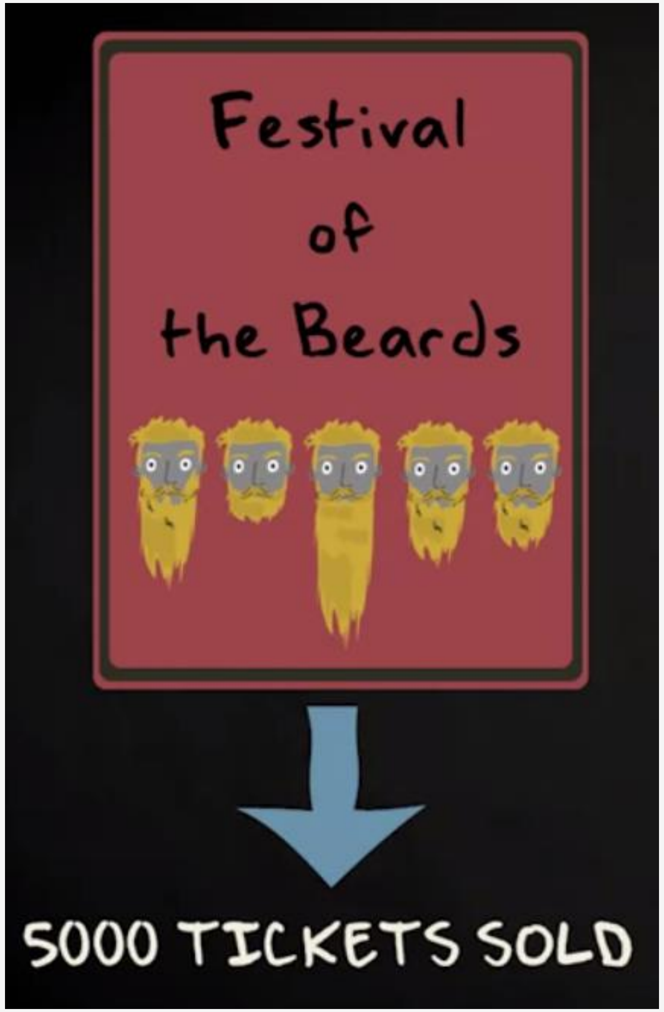
Sampling distribution of sample mean and central limit theorem

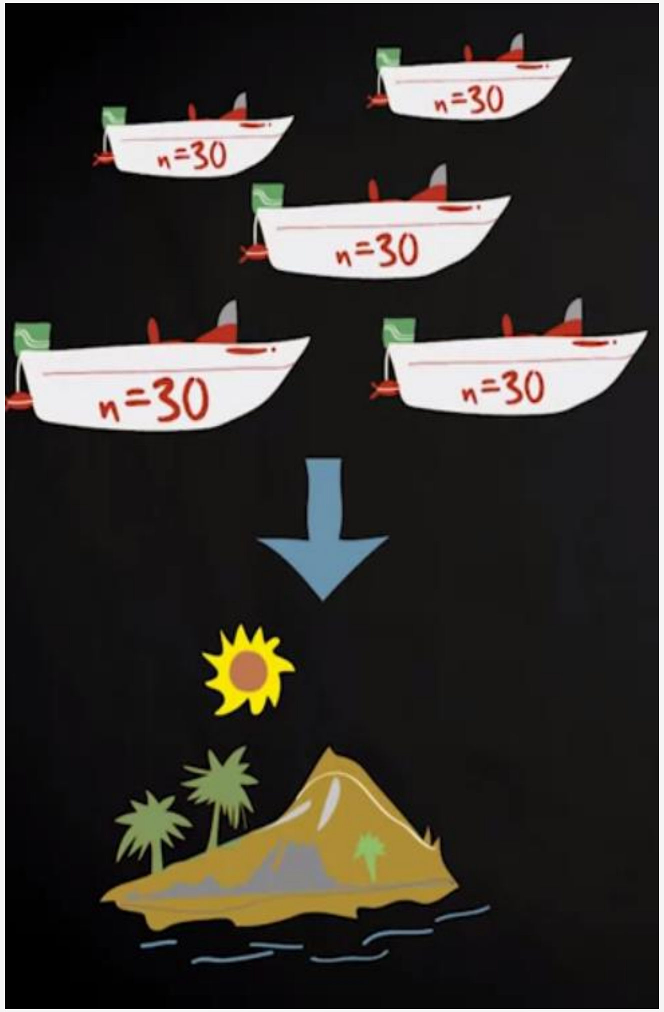
We've seen that researchers often use a sample to draw conclusions about the population their sample is from. To do so, they make use of a probability distribution that is very important in the world of statistics: **the sampling distribution.**

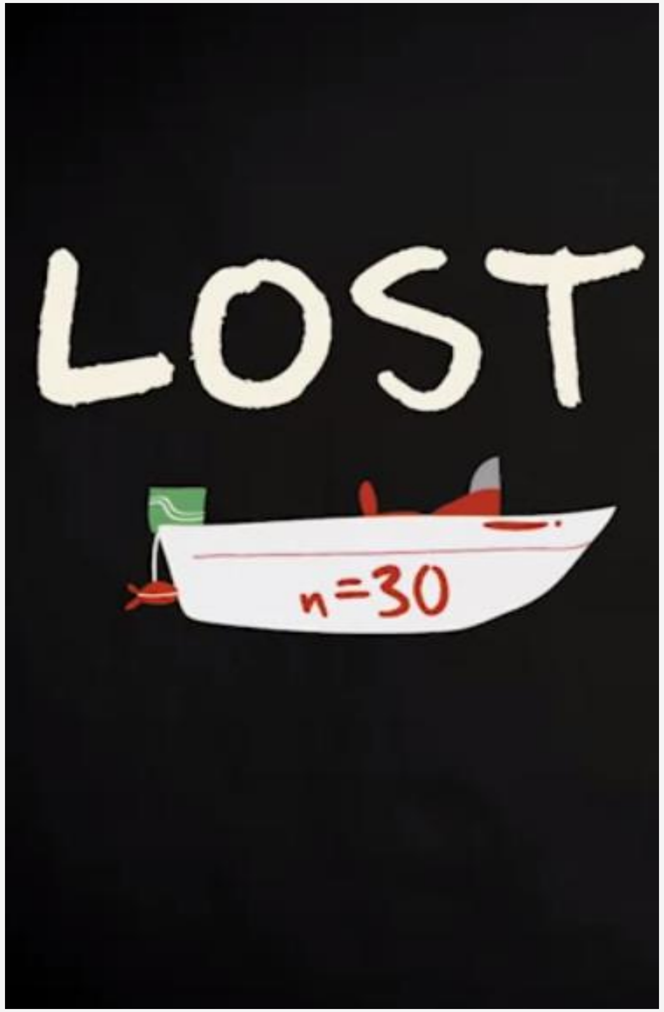
The sampling distribution of the sample mean is the distribution that you get if you draw an infinite number of samples from your population and compute the mean of all the collected sample means.

The **central limit theorem** says that, provided that the sample size is sufficiently large, the sampling distribution of the sample mean has an approximately normal distribution. The mean of the sampling distribution equals the population mean, and the standard deviation of the sampling distribution equals the standard deviation in the population divided by the square root of the sample size.

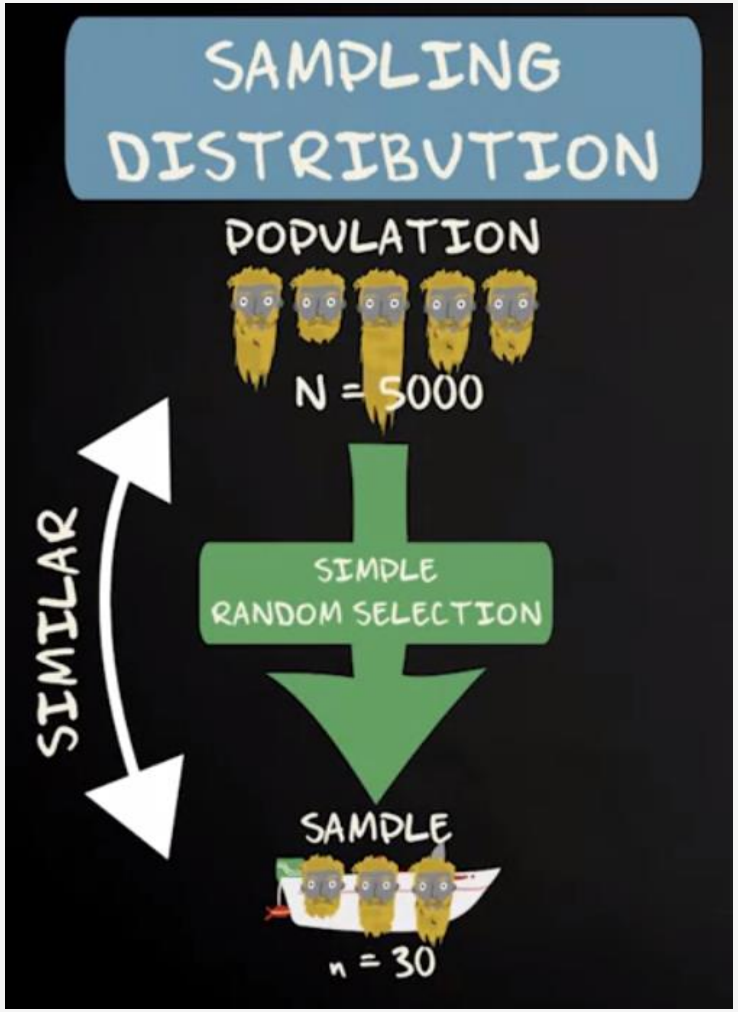
Sample and sampling





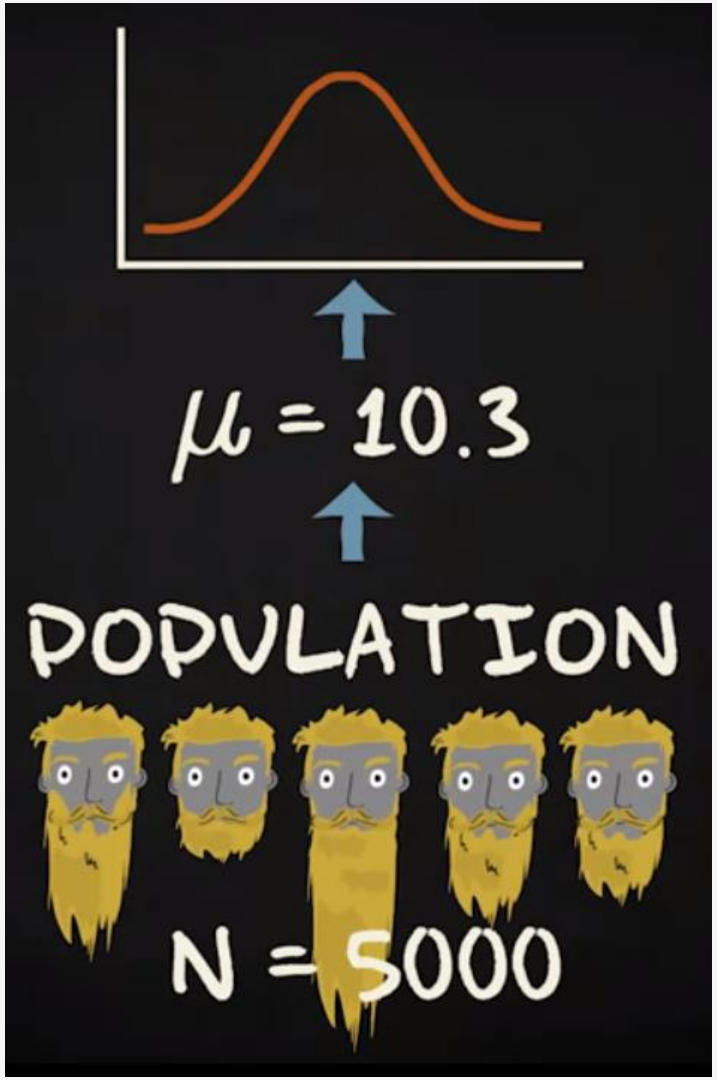


Sample and sampling

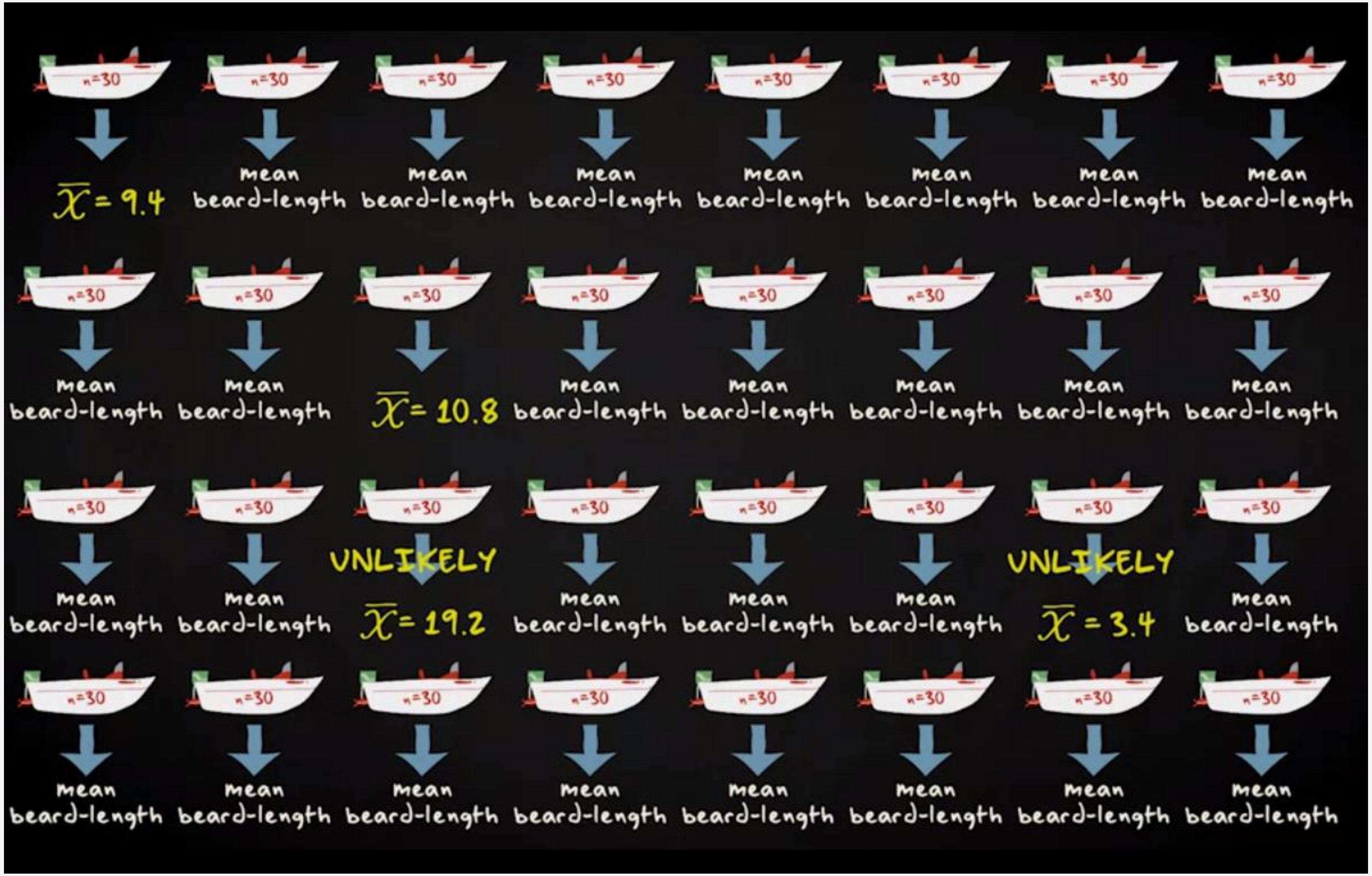


VERY UNLIKELY
SIMPLE RANDOM SAMPLE
Festival of the Beards

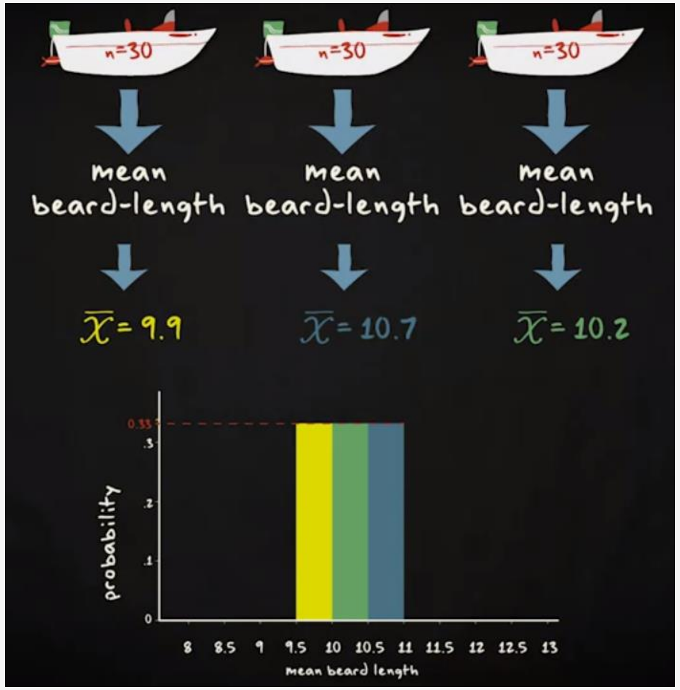
Sample and sampling

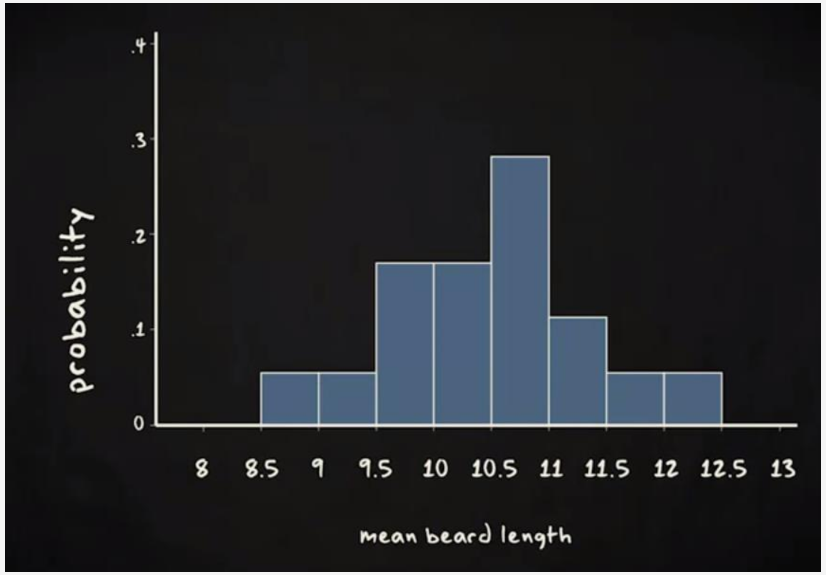


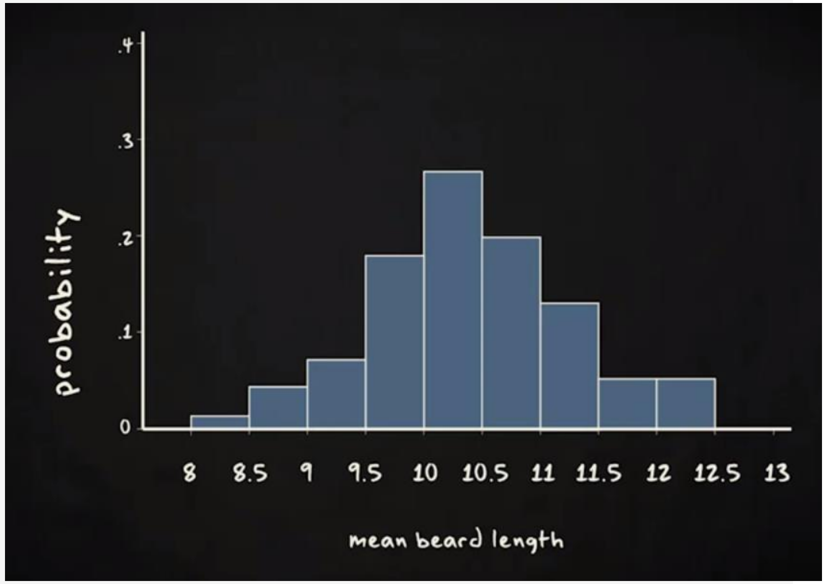
Sample and sampling



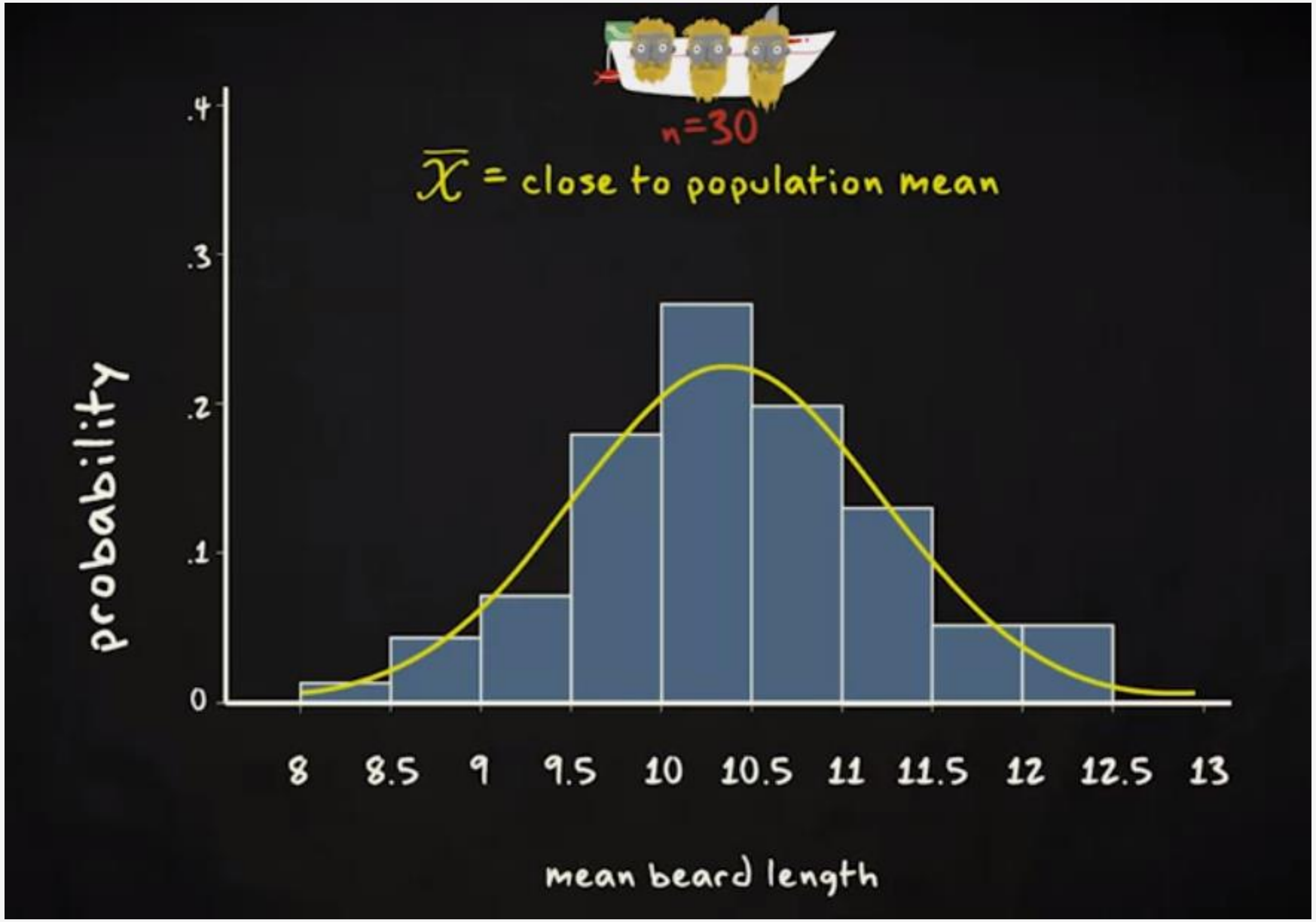
Sample and sampling



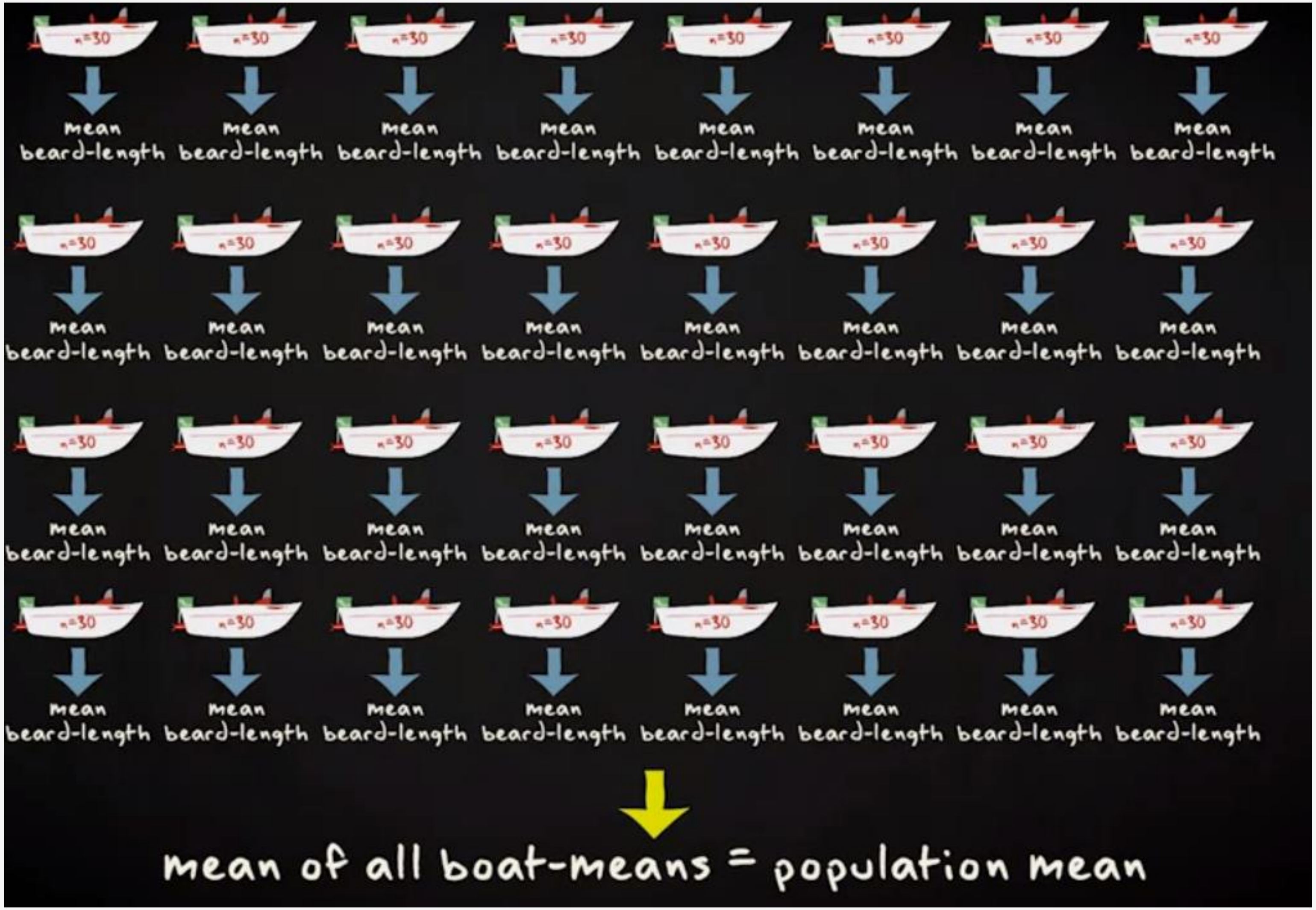




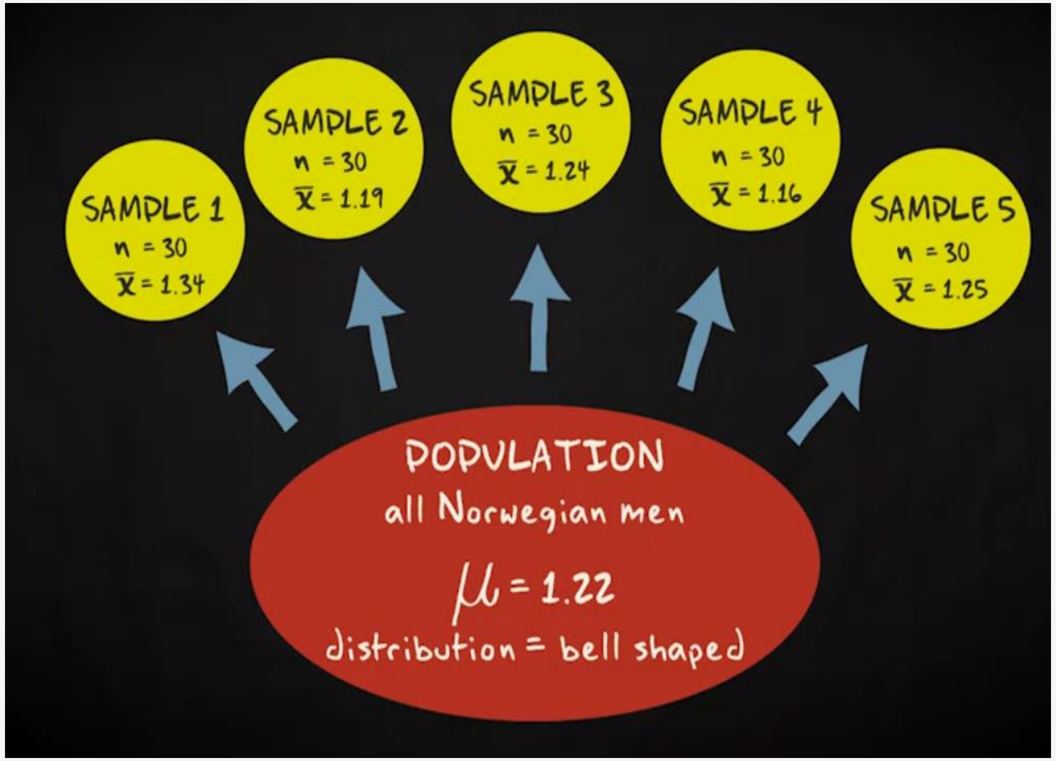
Sample and sampling

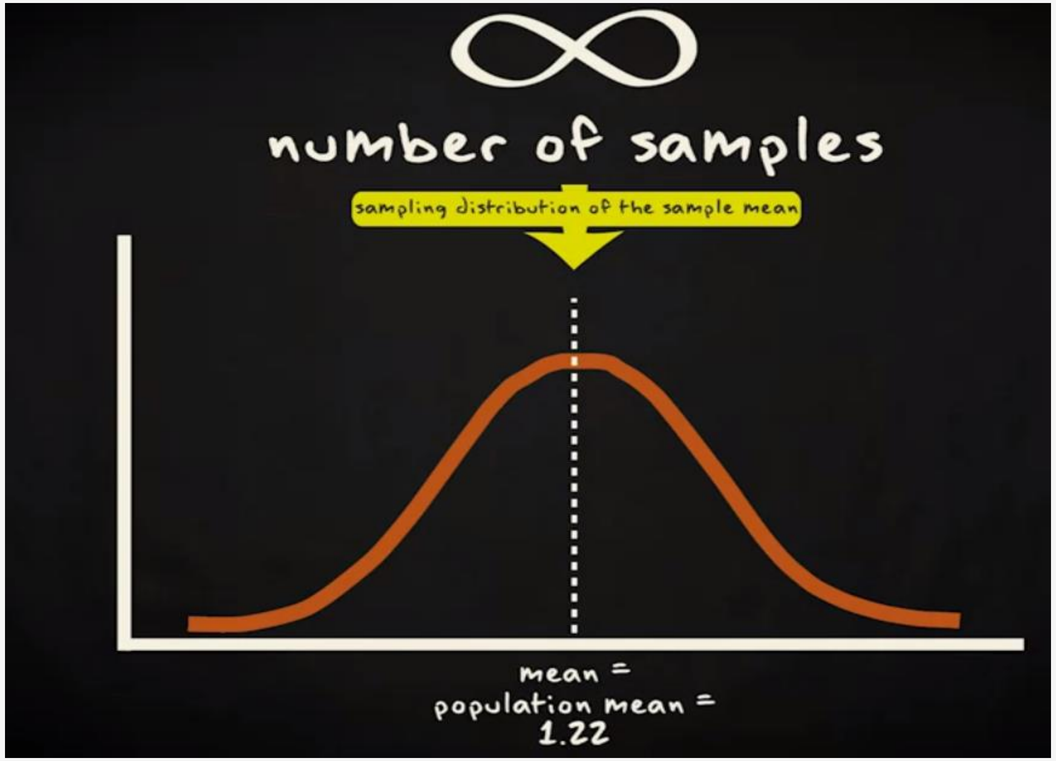


Sample and sampling

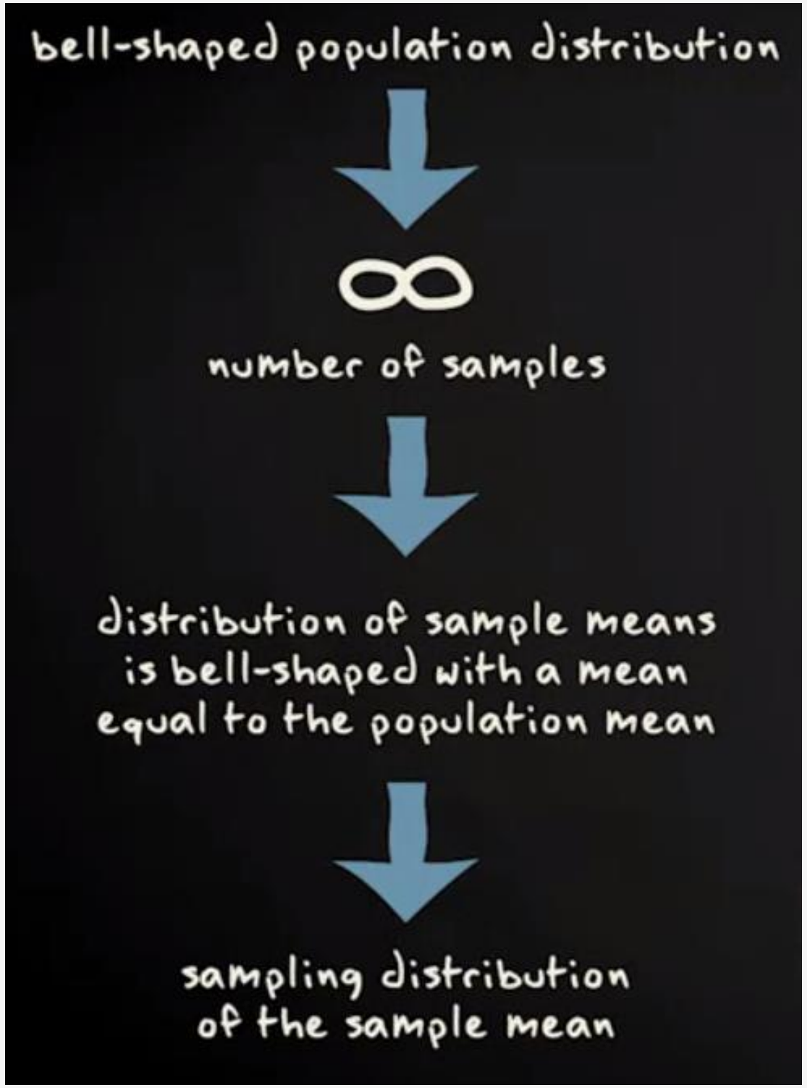


Sample and sampling





Sample and sampling



bell-shaped population distribution

∞

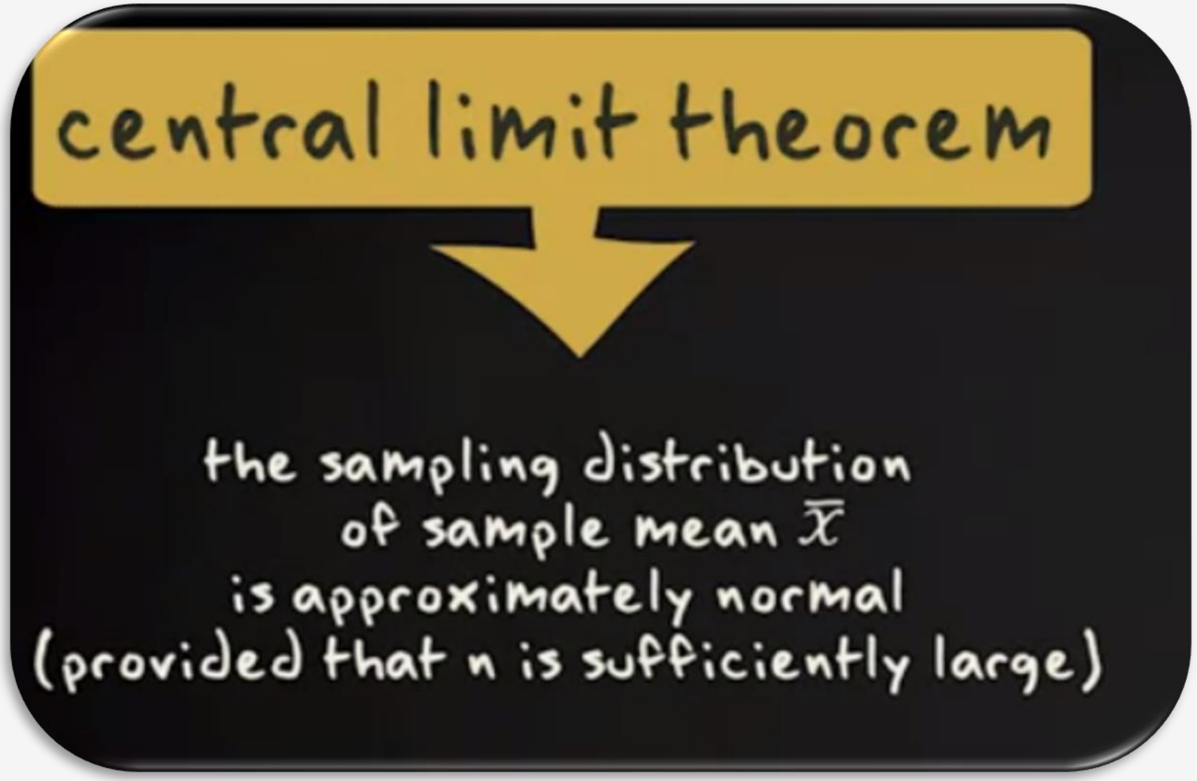
number of samples

distribution of sample means is bell-shaped with a mean equal to the population mean

sampling distribution of the sample mean

Sample and sampling

The central limit theorem says that provided that the sample size is sufficiently large, the sampling distribution of sample mean x bar has an approximately normal distribution, even if the variable of interest is not normally distributed in the population.

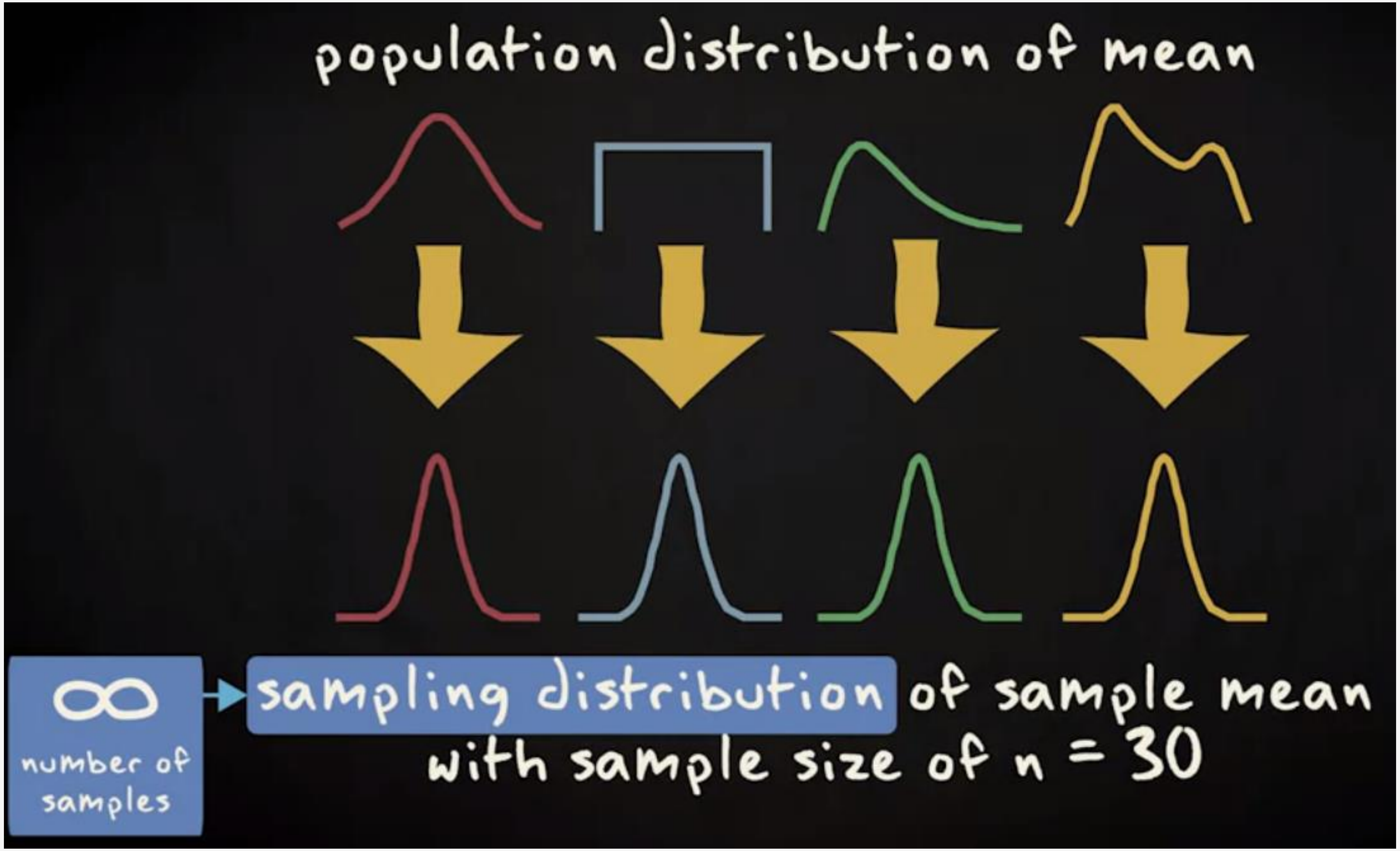


central limit theorem

the sampling distribution
of sample mean $\overline{x}$
is approximately normal
(provided that n is sufficiently large)

Isn't that cool? No matter how a variable is distributed in the population, the sampling distribution of the sample mean is always, always approximately normal, as long as the sample size is large enough.

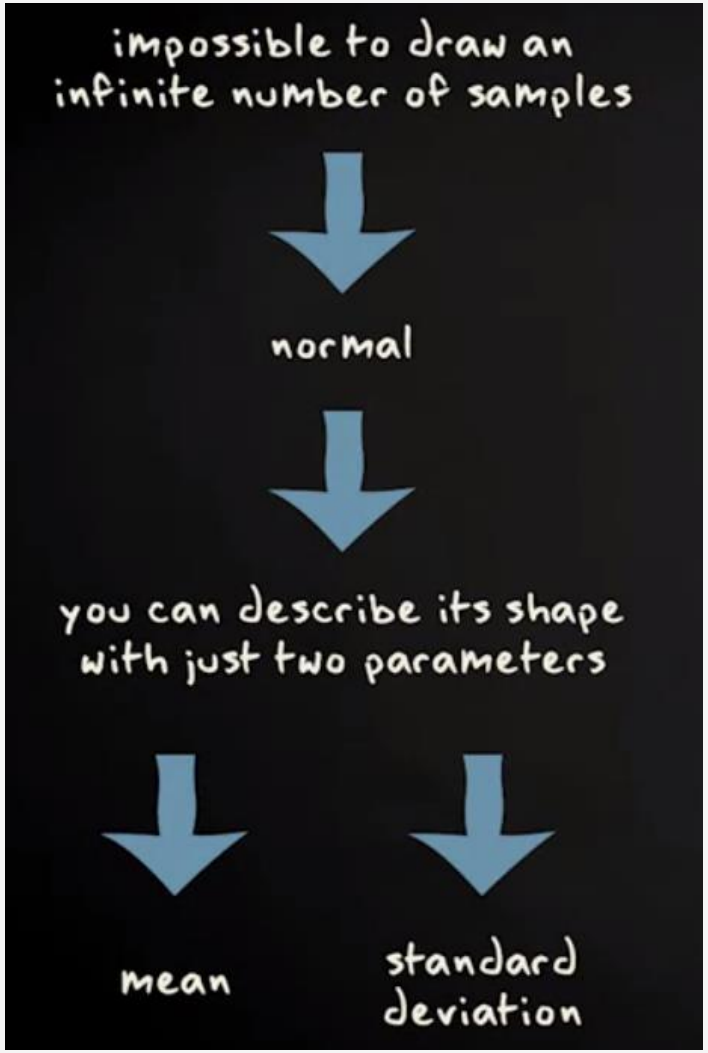
Sample and sampling


population distribution of mean
∞
number of samples
sampling distribution of sample mean
with sample size of n = 30

Sample and sampling

In practice, it is impossible to draw an infinite number of samples, but then, the good news is that drawing multiple samples is not required at all, to determine the shape of the sampling distribution.

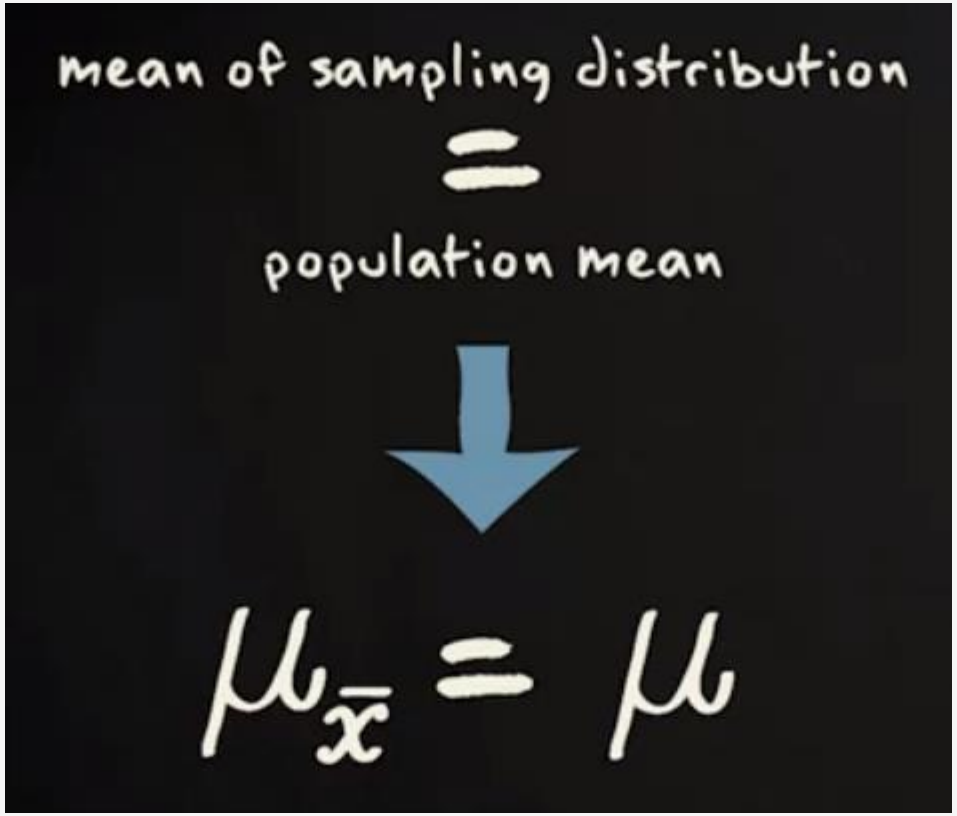
Because if it's normal, you can describe its shape by just two parameters, mean and standard deviation. So it is sufficient to estimate these two parameters.


impossible to draw an infinite number of samples
normal
you can describe its shape with just two parameters
mean
standard deviation

Sample and sampling

The mean of the sampling distribution is equal to the mean of the population distribution.

We can display that as follows, mu x bar is equal to mu. Mu stands for the population mean, and mu x bar stands for the mean of the sampling distribution of the sample mean.

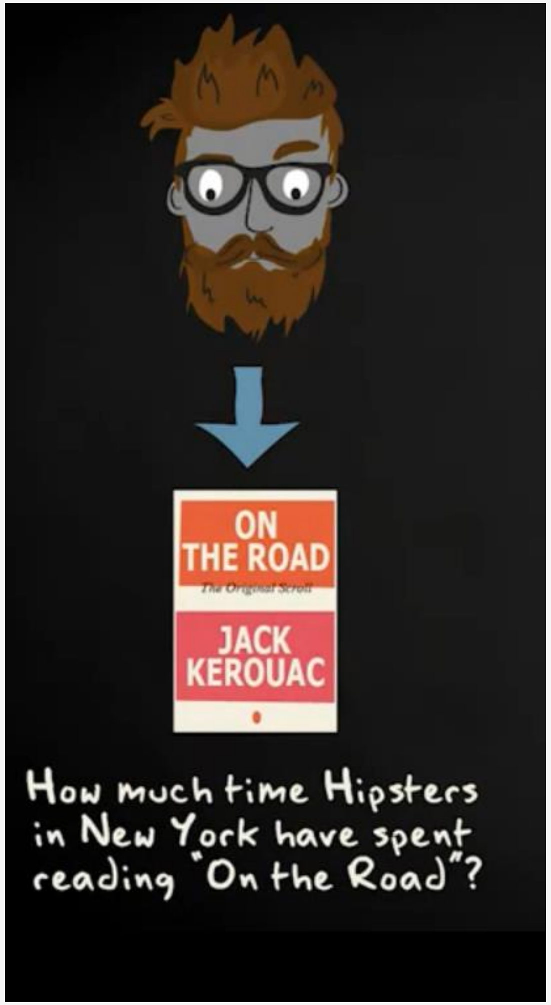


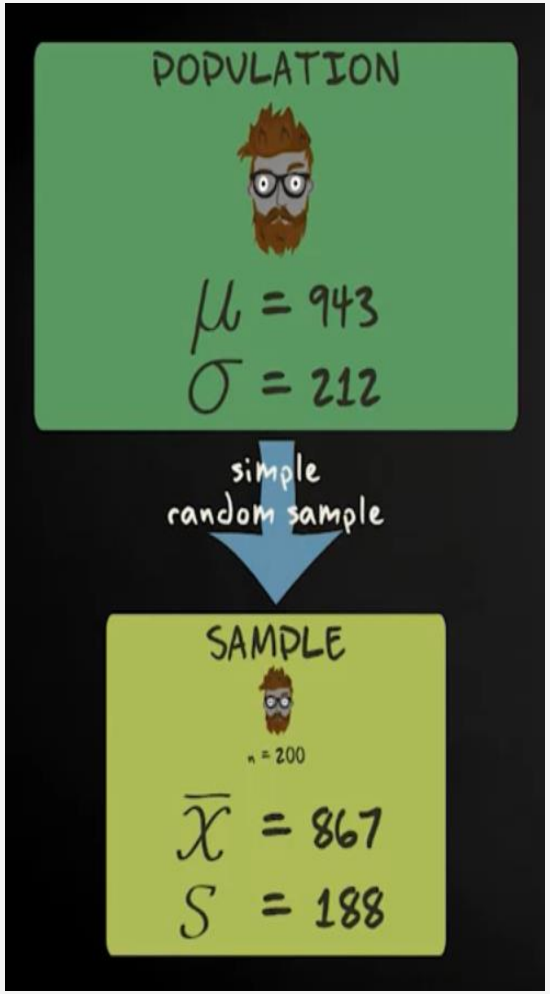
Sample and sampling

We are interested in the question **how much time Hipsters in New York have spent reading "On the Road".**

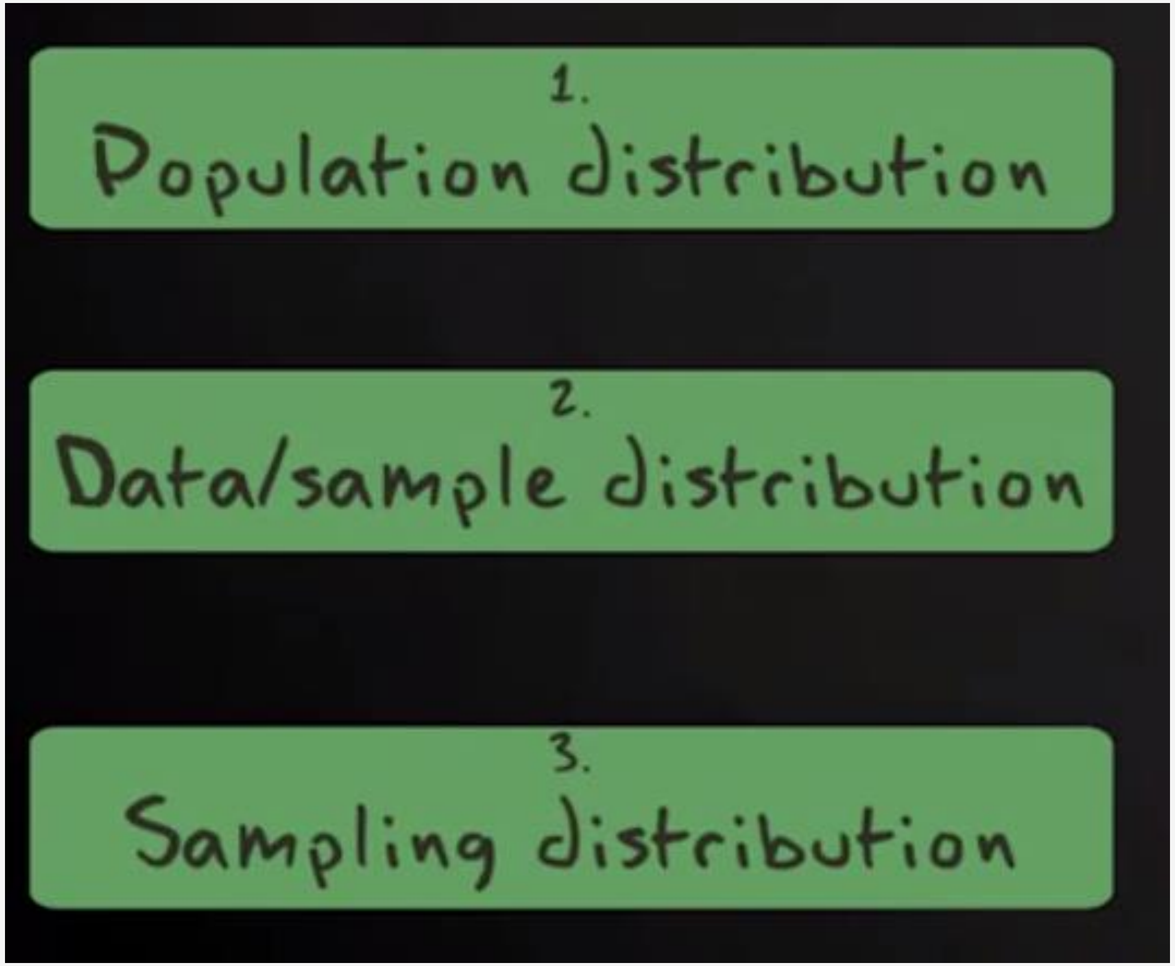
Here we will see about three distributions that are of importance to our research project: **the population distribution**, the data or **sample distribution** and the **sampling distribution**.

We will also show you how you can **compute the probabilities of selecting individuals with particular scores** and samples with particular sample means from this population


ON THE ROAD
The Original Scroll
JACK KEROUAC
How much time Hipsters in New York have spent reading "On the Road"?



Sample and sampling

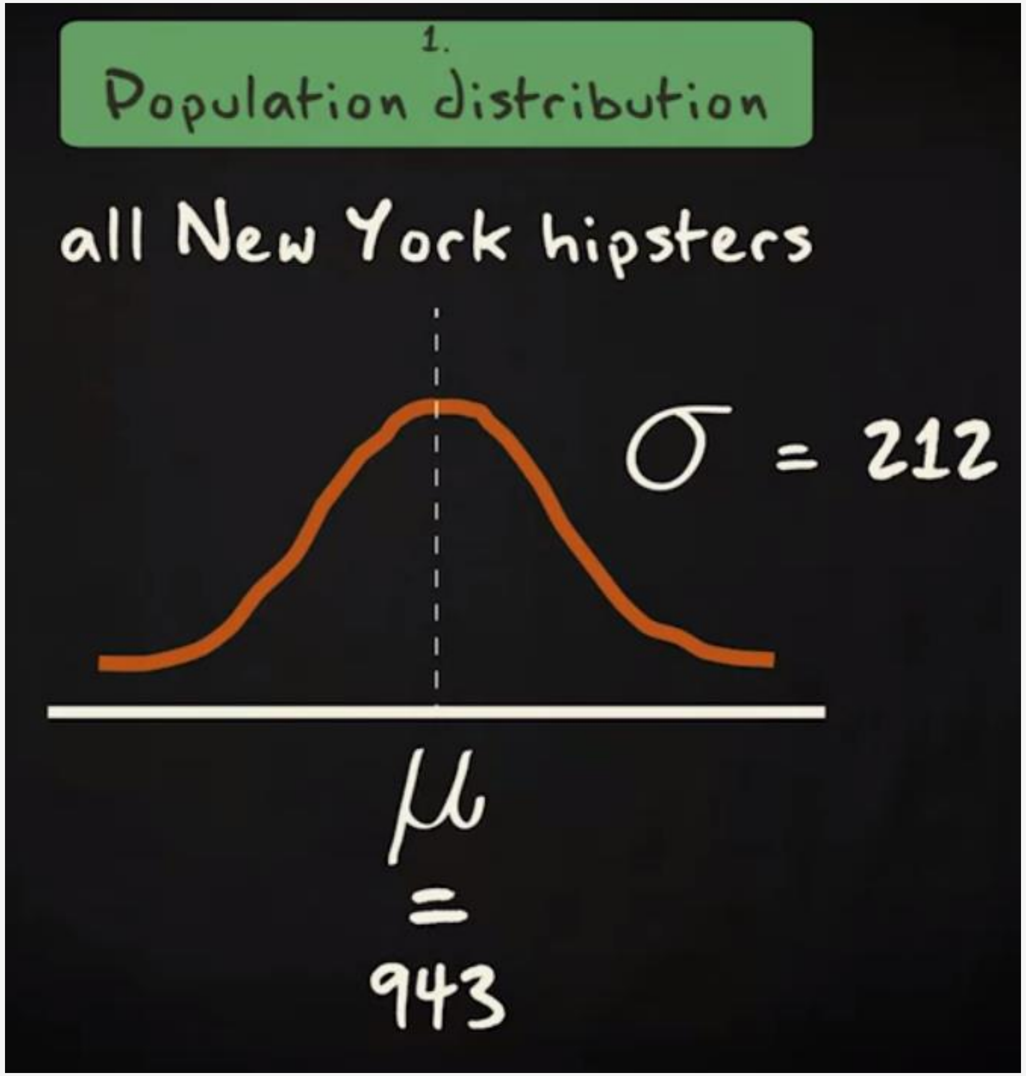


1. Population distribution

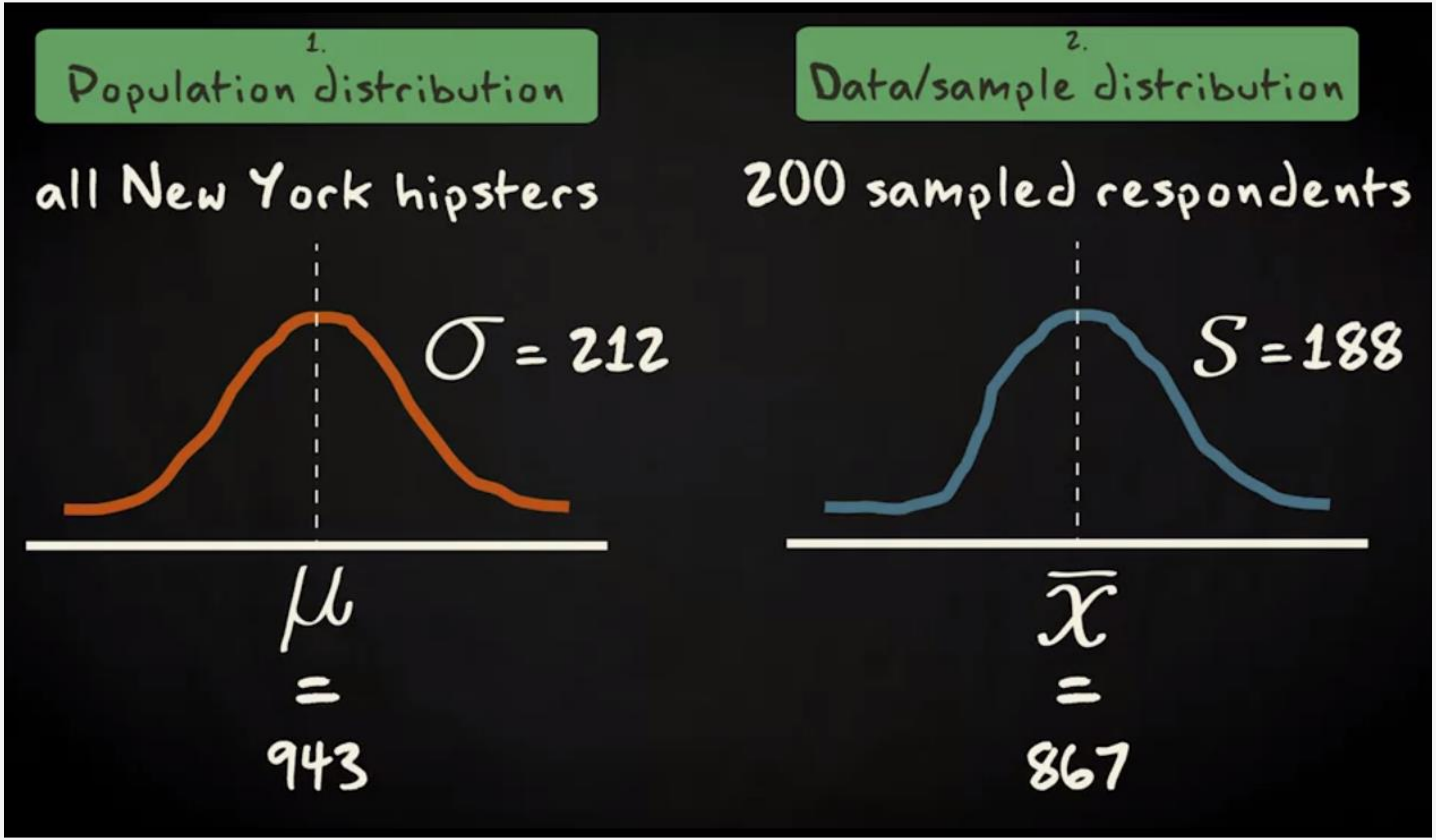
2. Data/sample distribution

3. Sampling distribution

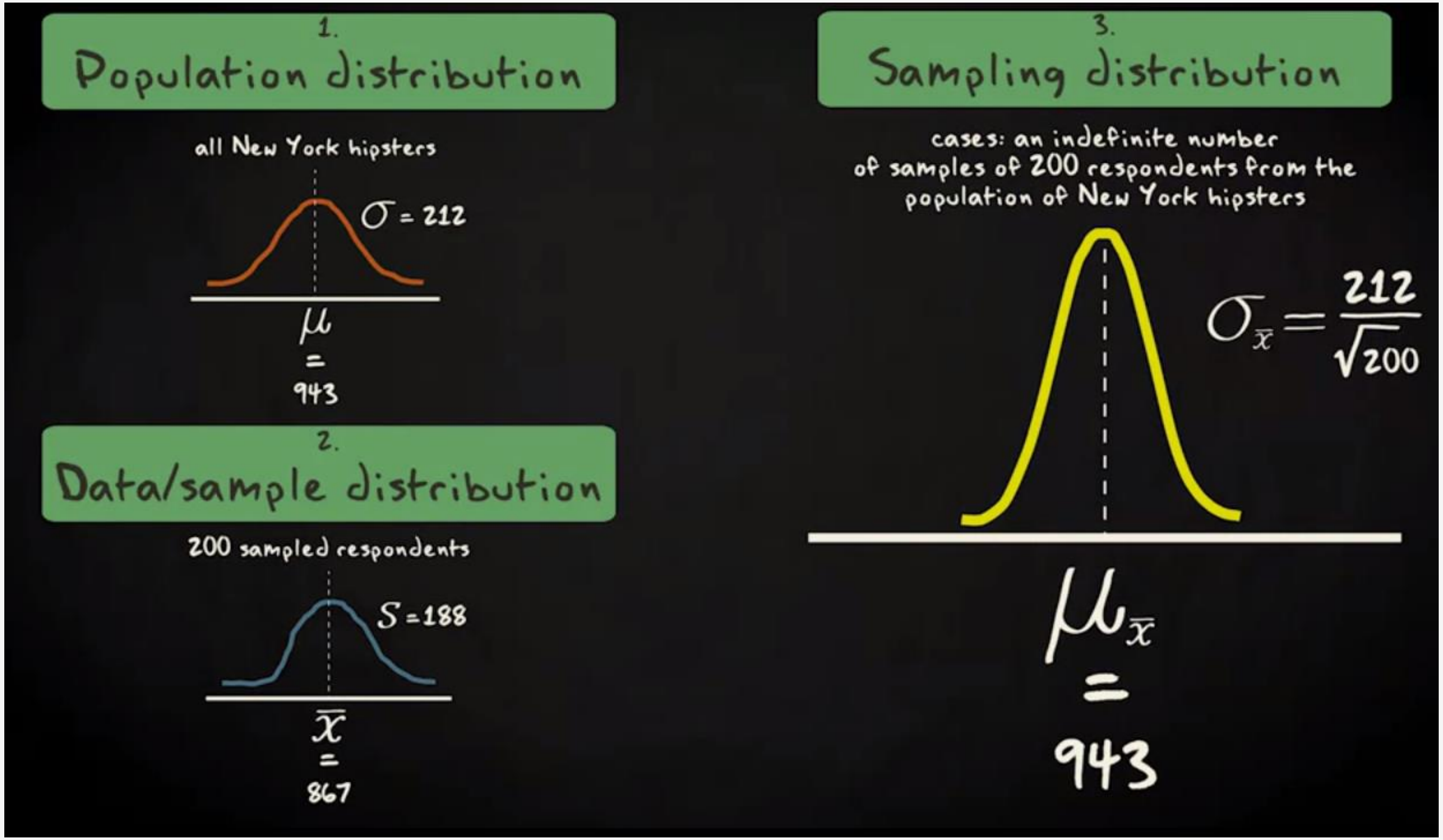
Sample and sampling



Sample and sampling



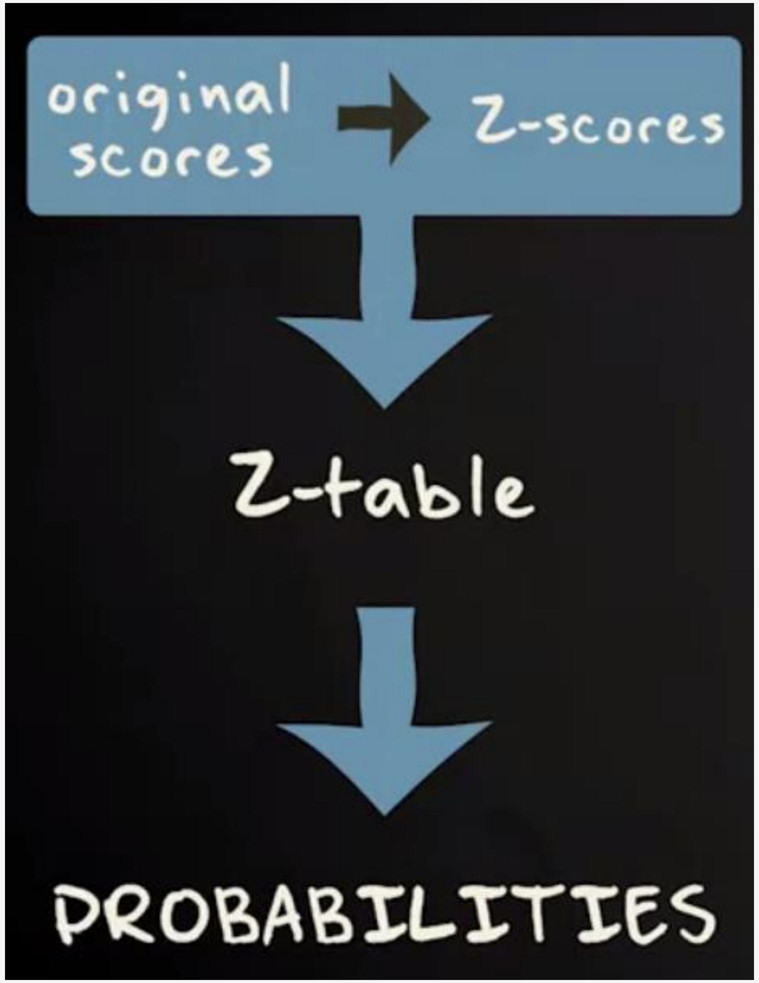
Sample and sampling



Sample and sampling

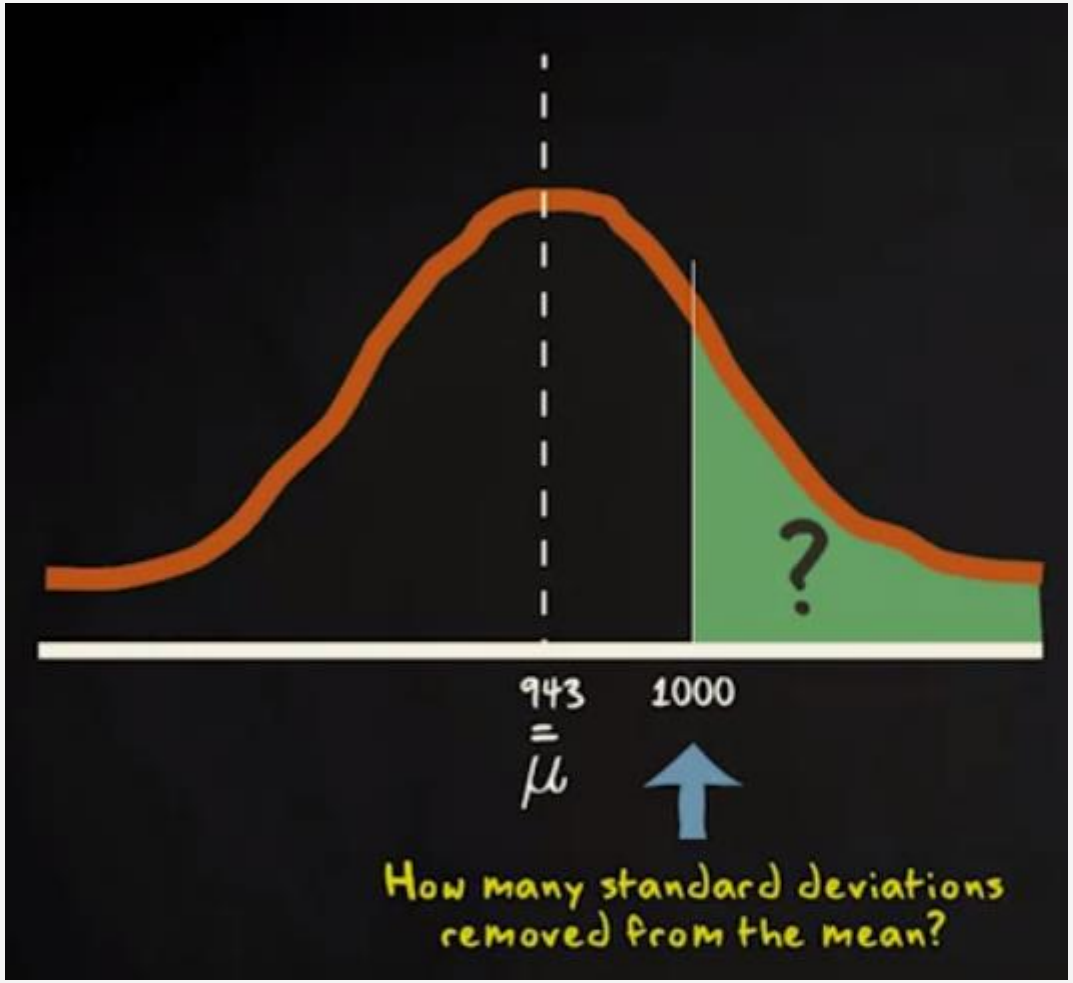
The very nice thing about normal distributions, is that we can find probabilities by changing original scores into z scores. And by then employing the z table.

Now, if we would like to know what the probabilities are, of selecting random samples or subjects from a population, we can apply this logic to sampling distributions and as long as they are normally distributed to population distributions.

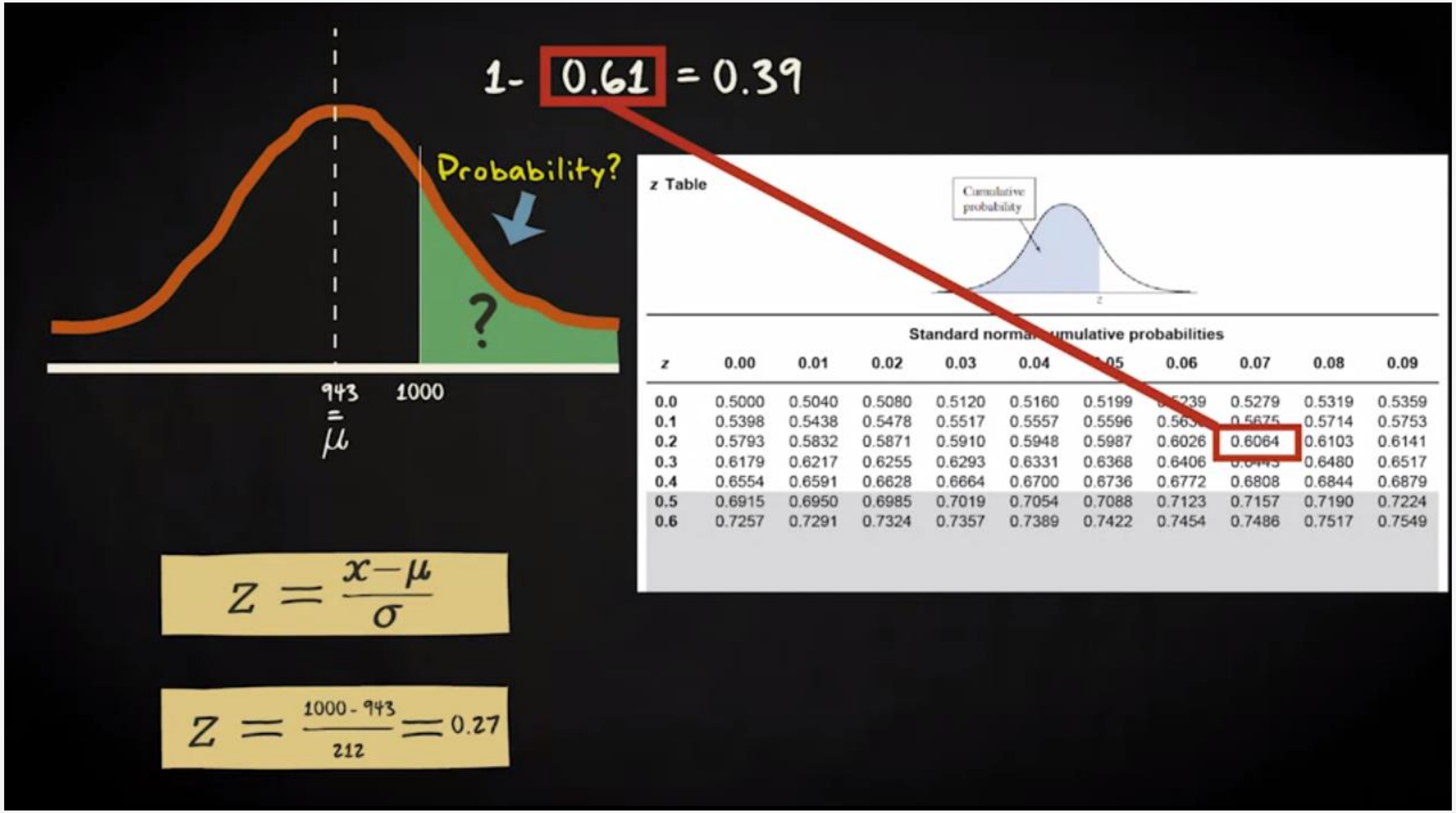


Sample and sampling

Now image you select a random hipster from the population. What is the probability that this hipster has a reading time of 1,000 minutes or more?



Sample and sampling



$$1-\boxed{0.61}=0.39$$

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

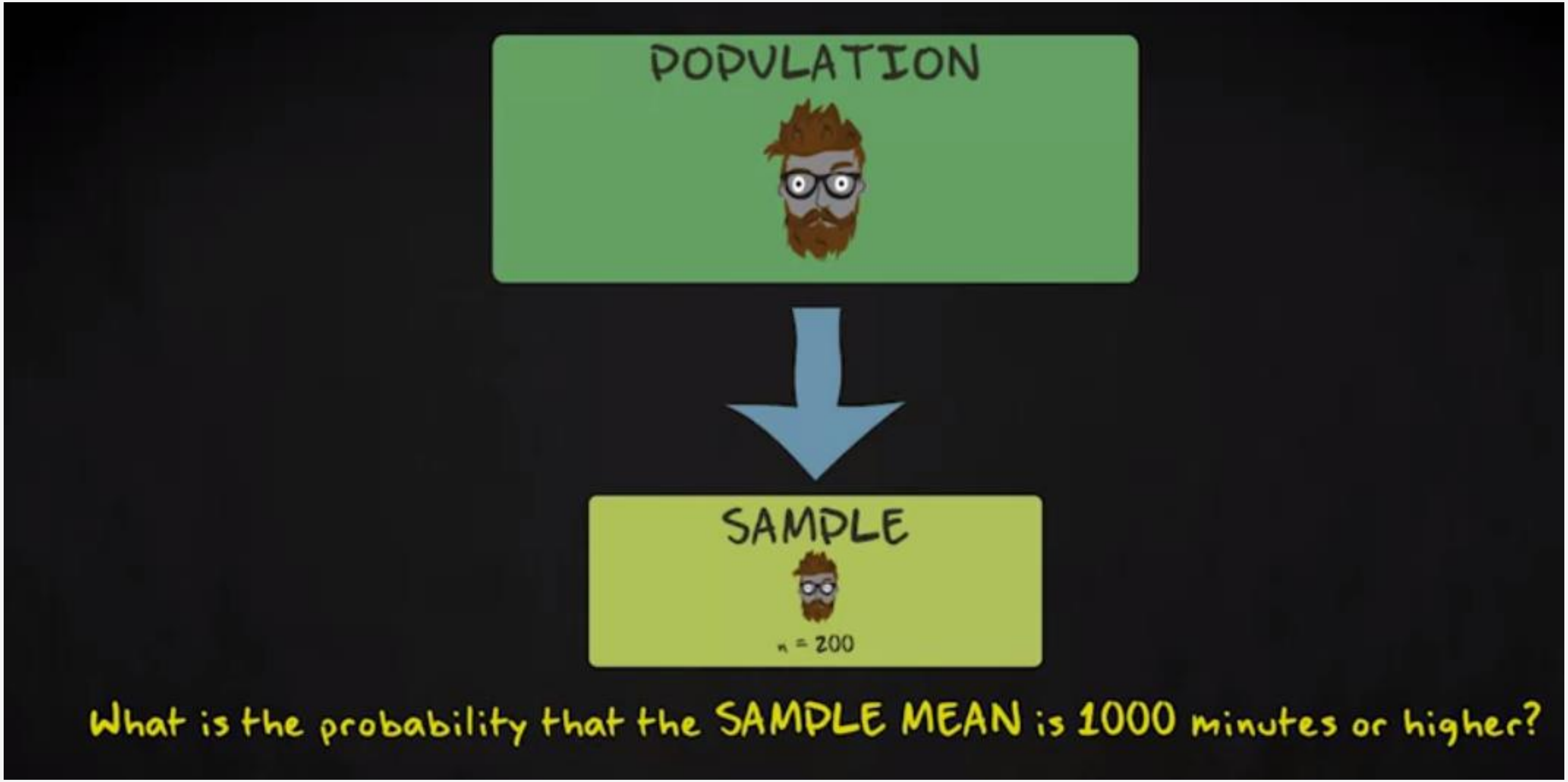
$$Z=\frac{1000-943}{212}=0.27$$

z Table

Standard normal cumulative probabilities

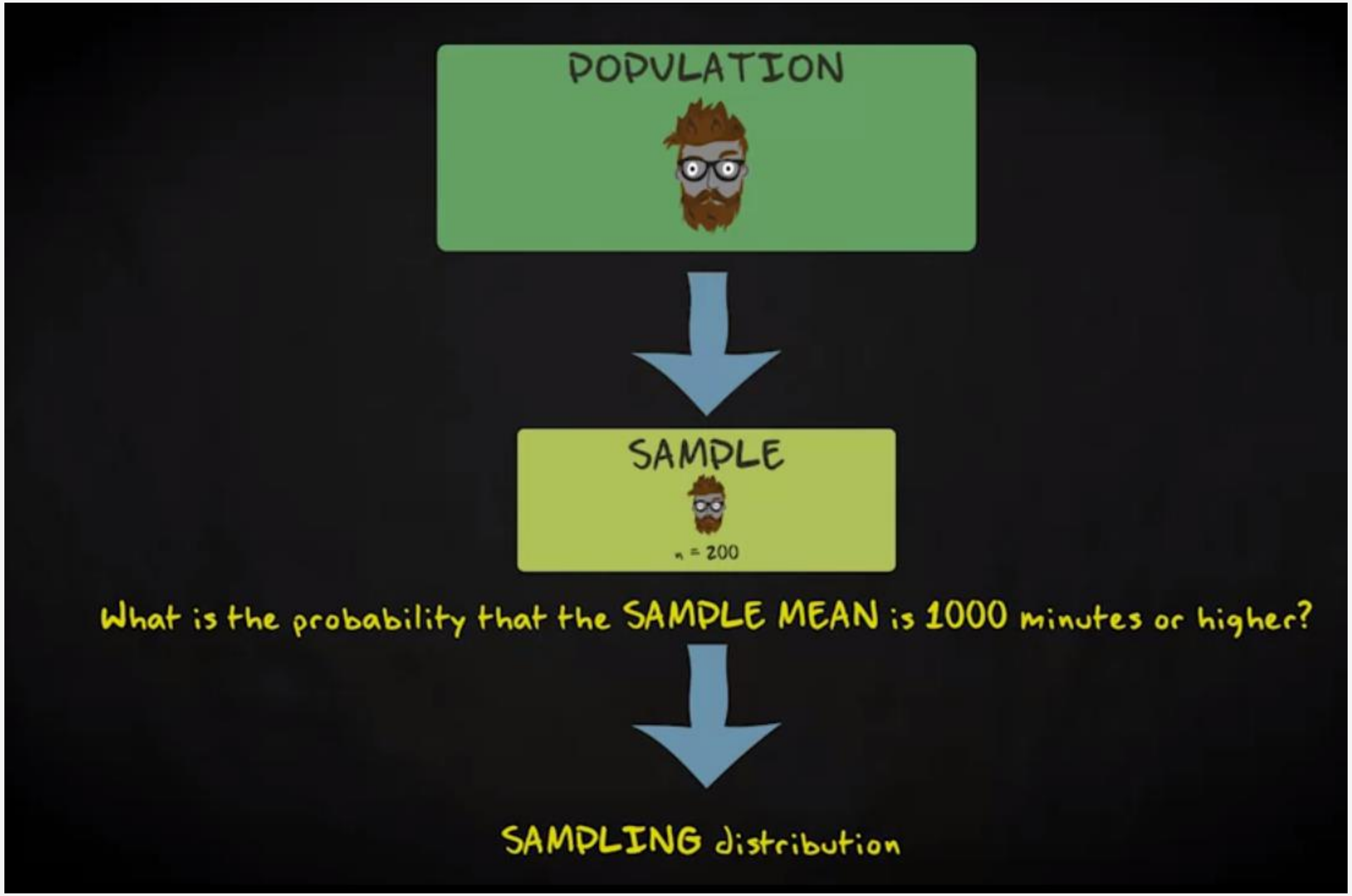
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	[illegible]	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	$\boxed{0.6064}$	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	[illegible]	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549

Sample and sampling



What is the probability that the SAMPLE MEAN is 1000 minutes or higher?

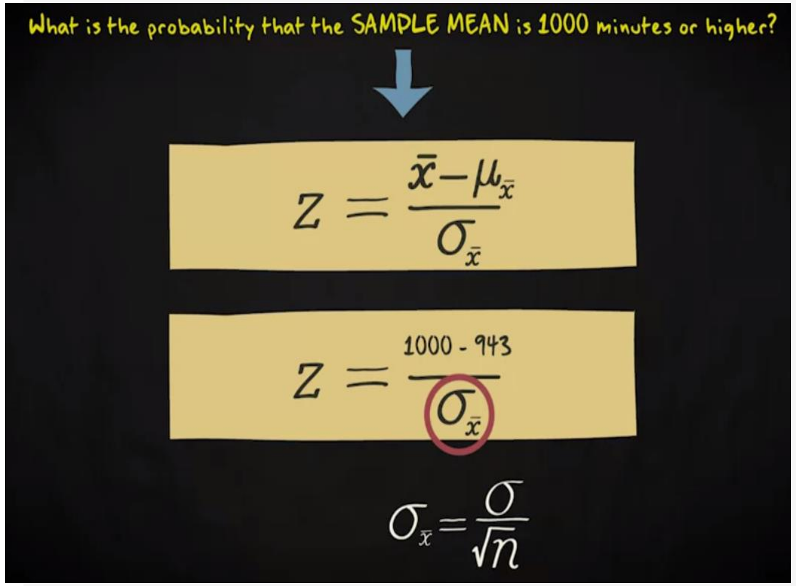
Sample and sampling



What is the probability that the SAMPLE MEAN is 1000 minutes or higher?

SAMPLING distribution

Sample and sampling

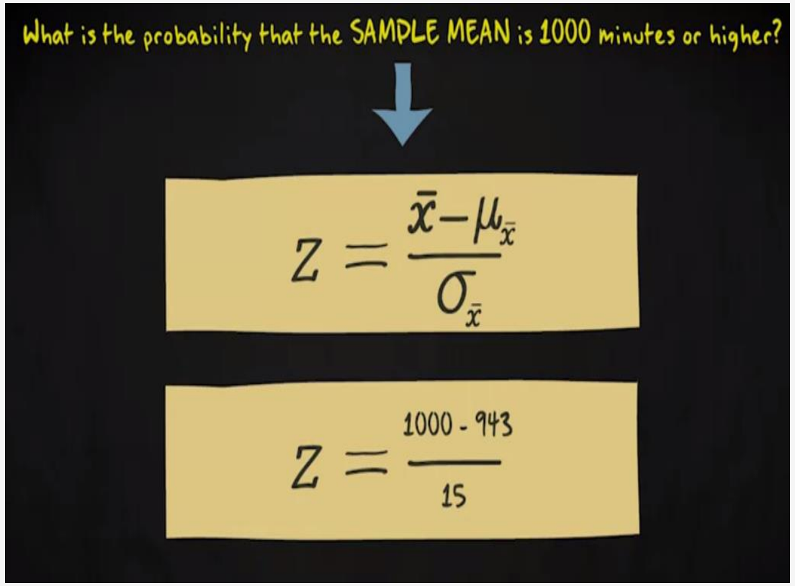


What is the probability that the SAMPLE MEAN is 1000 minutes or higher?

$$Z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

$$Z = \frac{1000 - 943}{\sigma_{\bar{x}}}$$

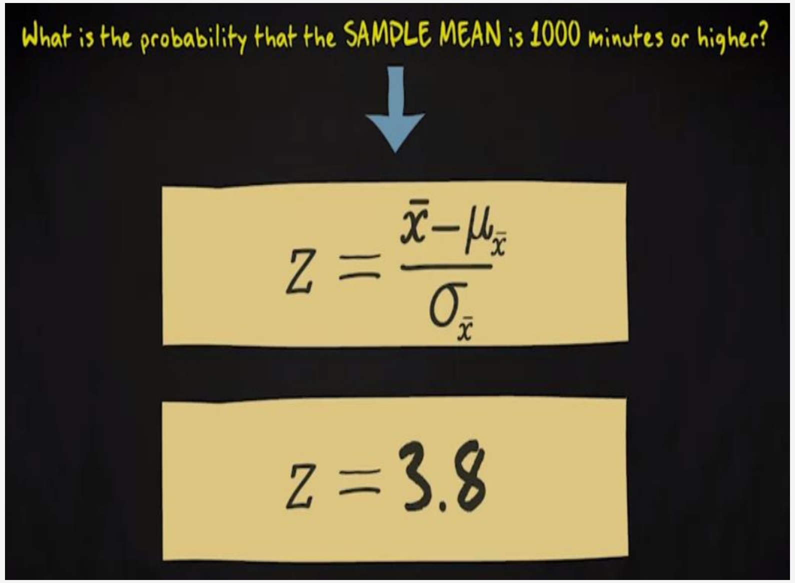
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



What is the probability that the SAMPLE MEAN is 1000 minutes or higher?

$$Z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

$$Z = \frac{1000 - 943}{15}$$



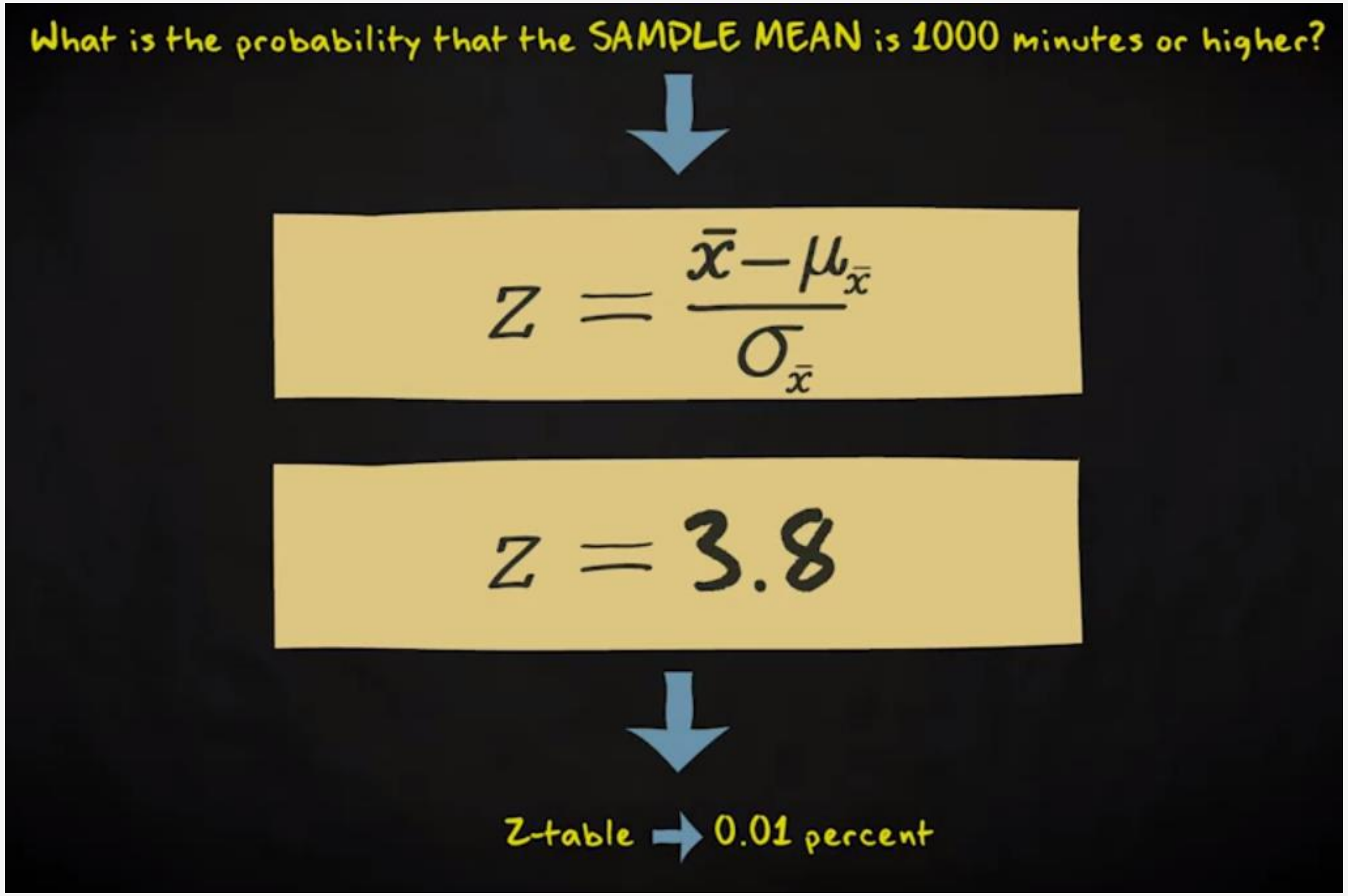
What is the probability that the SAMPLE MEAN is 1000 minutes or higher?

$$Z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

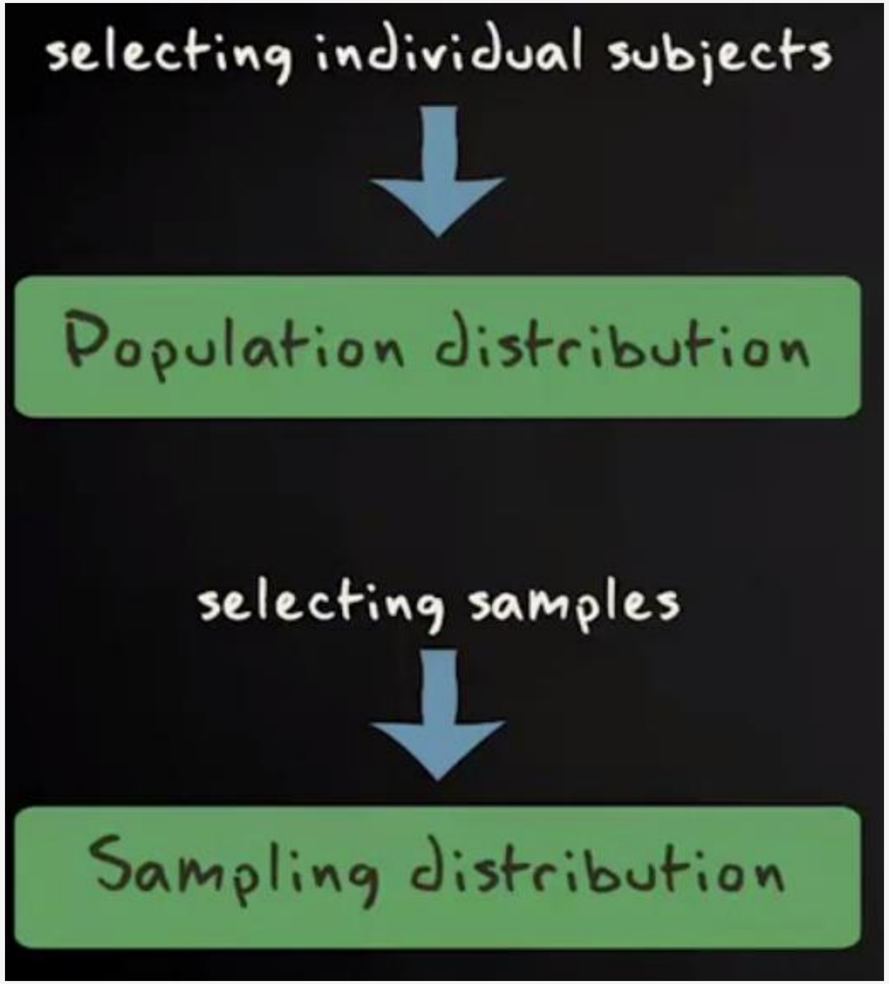
$$Z = 3.8$$

Sample and sampling

That gives a z score of 3.8. If we look it up in our z table, we find that the chance of drawing a sample with the mean reading time of one thousand minutes or more is 0.01 percent.


What is the probability that the SAMPLE MEAN is 1000 minutes or higher?
$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$
$z = 3.8$
Z-table → 0.01 percent

Sample and sampling

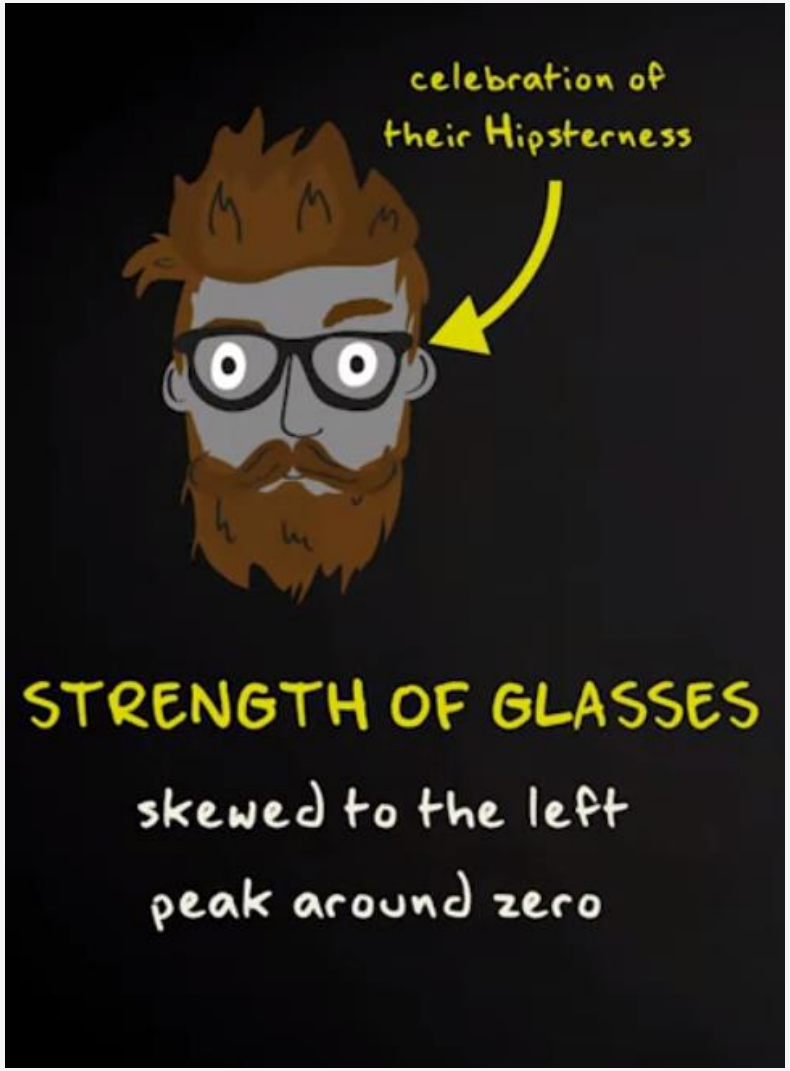


Exercise

As you might have noticed, hipsters often wear oversized glasses, yet many hipsters don't really need glasses. The glasses only serve as a celebration as their hipsterness.

Suppose you studied the strengths of the glasses of the hipsters in Italy. You know that the variable strength of glasses has a population distribution that is **skewed to the left**.

The peak will be around 0 because most glasses are fake glasses. But, of course, not all glasses are fake because hipsters are generally relatively young and **young people are more often short sighted** than long sighted, there will be more hipsters who are short sighted than hipsters who are long sighted.



Exercise

Those who are short sighted have a negative score on strengths of glasses. And those who are long sighted have a positive score.

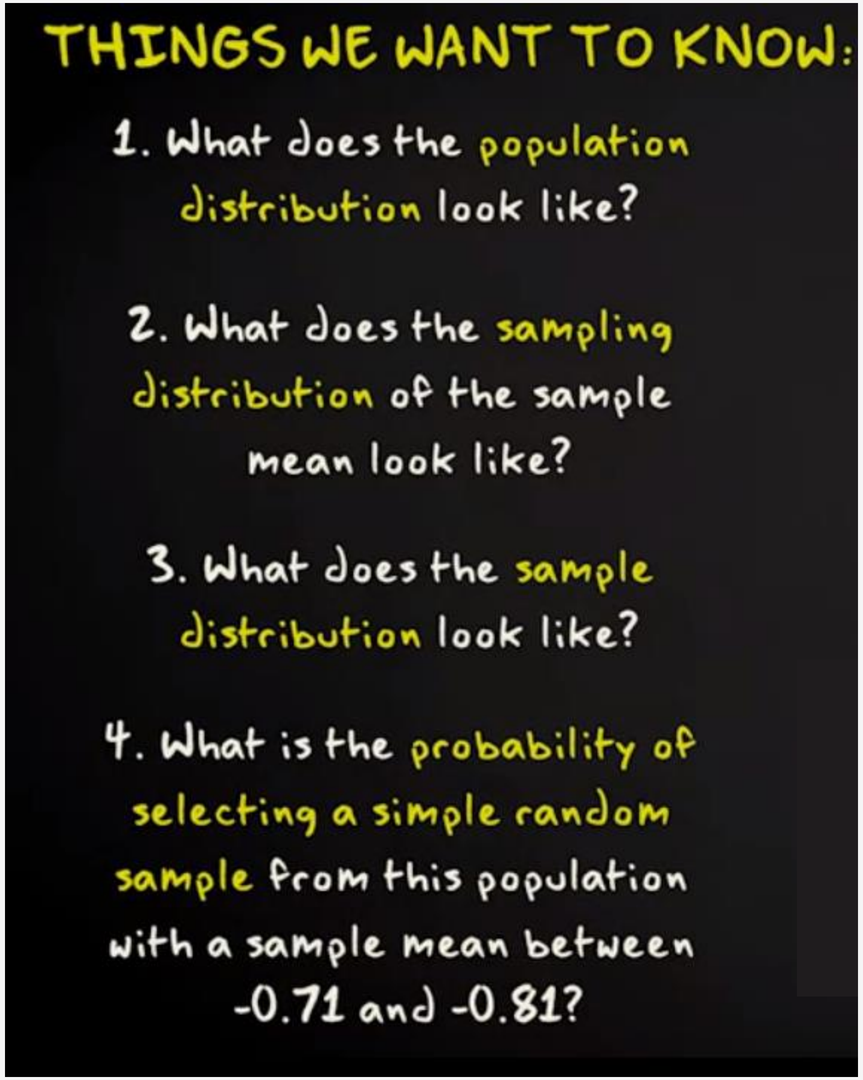
The mean in the population is -0.75 and that the standard deviation is 2.89. We would like to know four things.

First, what does the **population distribution** look like? We would like to see shape, mean and standard deviation.

Second, what does the **sampling distribution of the sample mean** look like based on a sample size of n equals 3000?

Third, what does the **sample or data distribution** look like if you draw a simple random sample of 3000 cases from this population?

And **fourth**, what is the **probability of selecting such a sample from this population** with a sample mean between -0.71 and -0.81?



THINGS WE WANT TO KNOW:

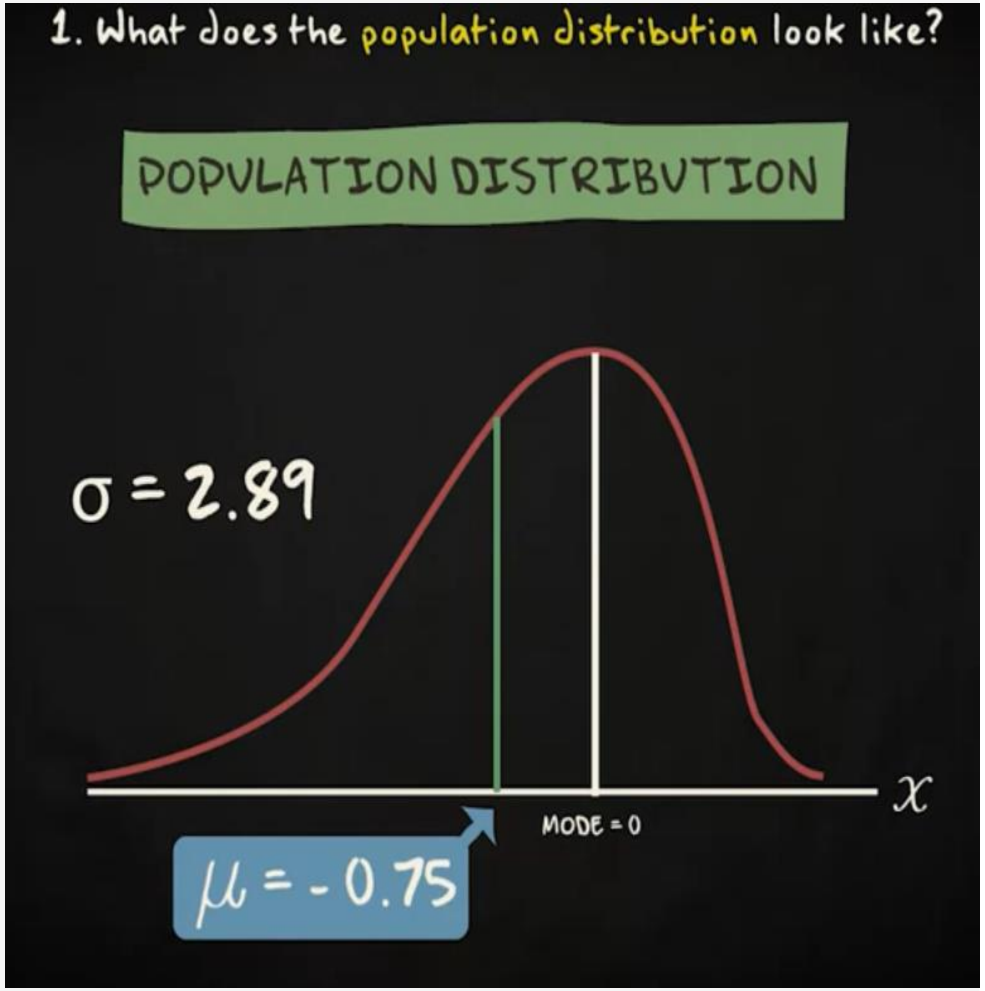
1. What does the population distribution look like?
2. What does the sampling distribution of the sample mean look like?
3. What does the sample distribution look like?
4. What is the probability of selecting a simple random sample from this population with a sample mean between -0.71 and -0.81?

Exercise

Let's start with the first question. The distribution would look something like this.

The peak is around 0 and the distribution is skewed to the left.

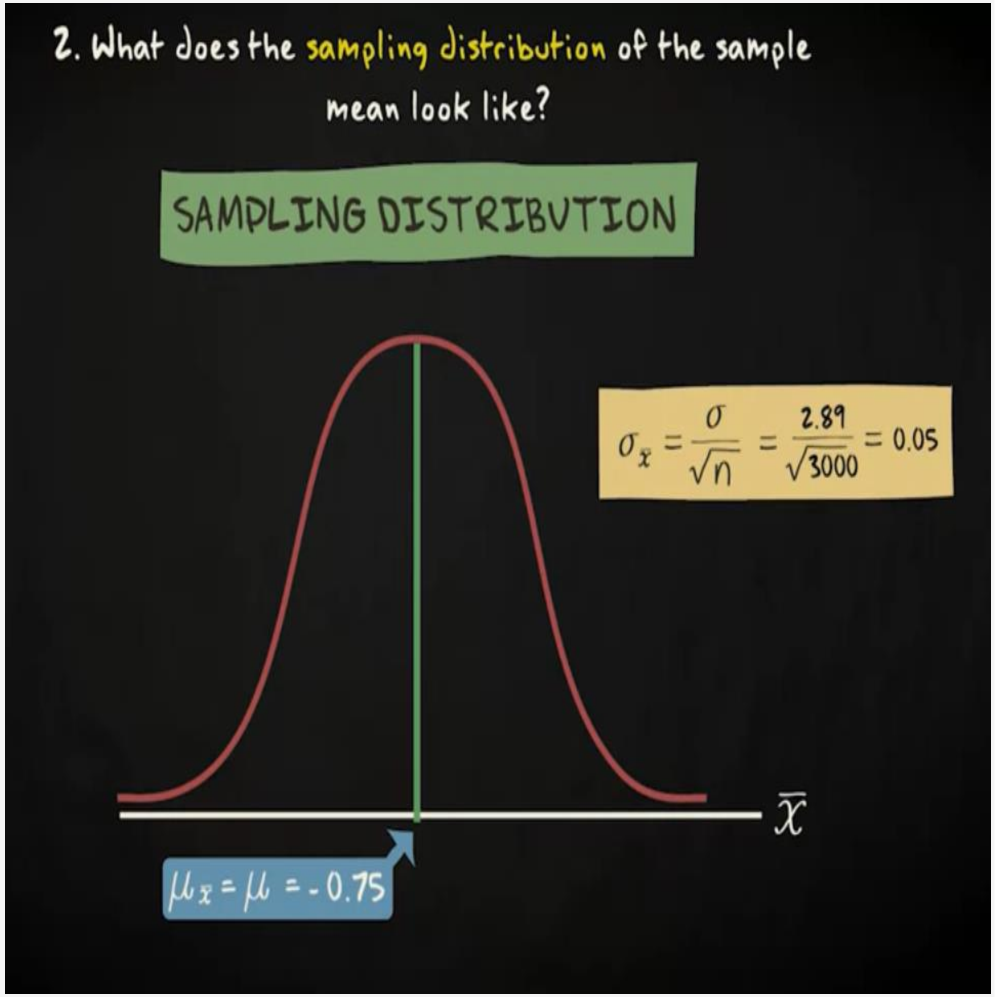
The population mean is left of the population mode. We know that the score of this parameter, symbolized by Mu, is -0.75. The population standard deviation, symbolized by Sigma, is 2.89.



Exercise

We know that when the sample size is sufficiently large (which is, with an n of 3000, clearly the case in this example), the sampling distribution is bell-shaped with a **mean that equals the population mean**.

The standard deviation of the sampling distribution, symbolized by **Sigma-X-bar**, can easily be computed. It is the standard deviation in the population divided by the square root of n.

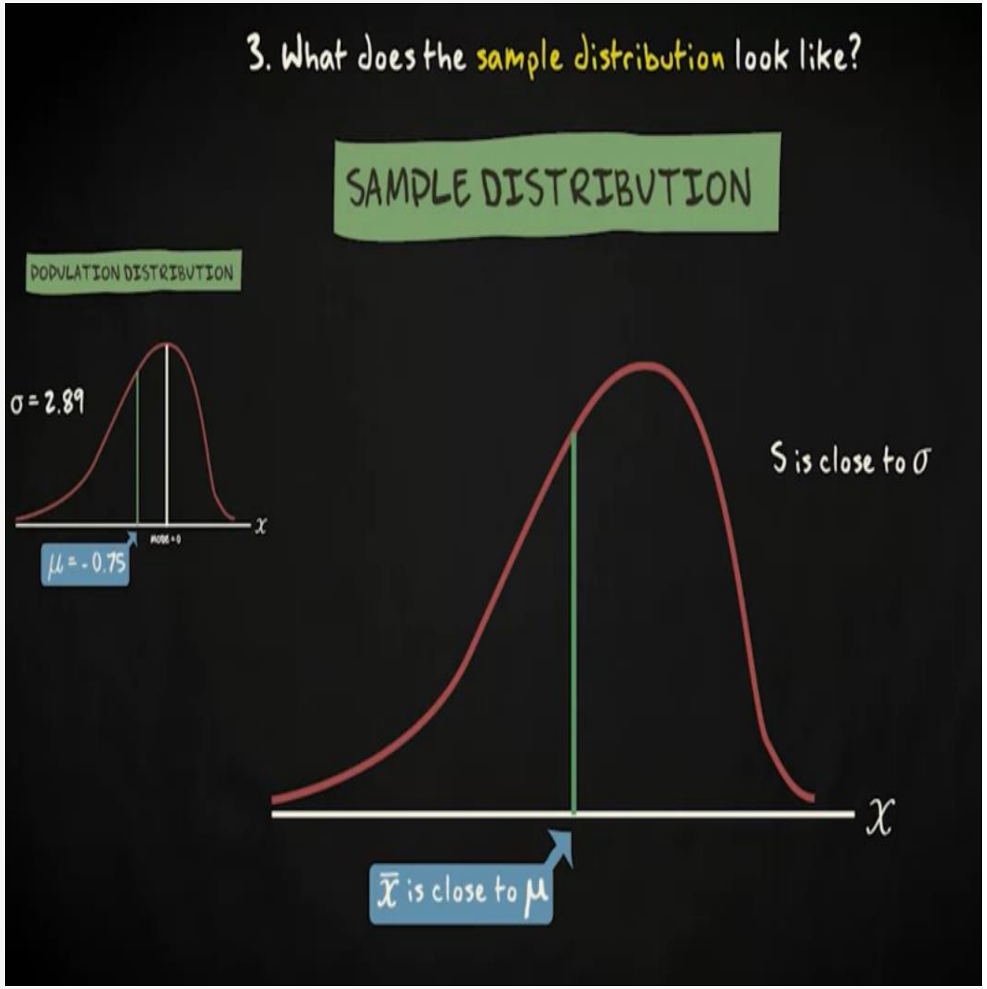

2. What does the sampling distribution of the sample mean look like?
SAMPLING DISTRIBUTION
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.89}{\sqrt{3000}} = 0.05$
$\bar{x}$
$\mu_{\bar{x}} = \mu = -0.75$

Exercise

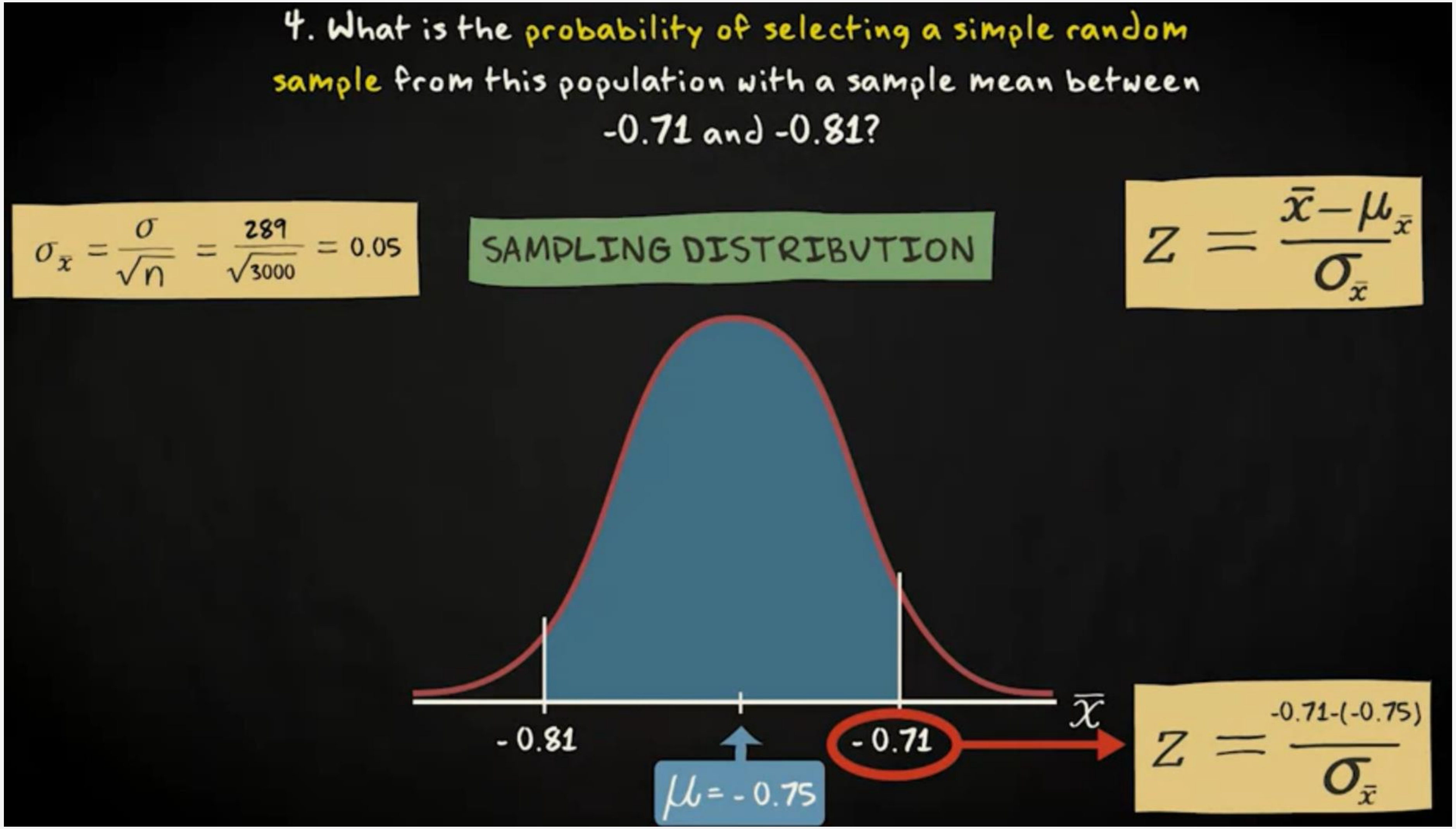
Because we are dealing with a simple random sample with a fairly large sample size, we can be pretty confident that the sample resembles the population.

The shape of its distribution will be very similar to the shape of the population distribution, and the sample mean, symbolized by X-bar, will be close to the population mean of - 0.75.

The sample standard deviation, symbolized by s, will be close to the population standard deviation of 2.89.



Exercise



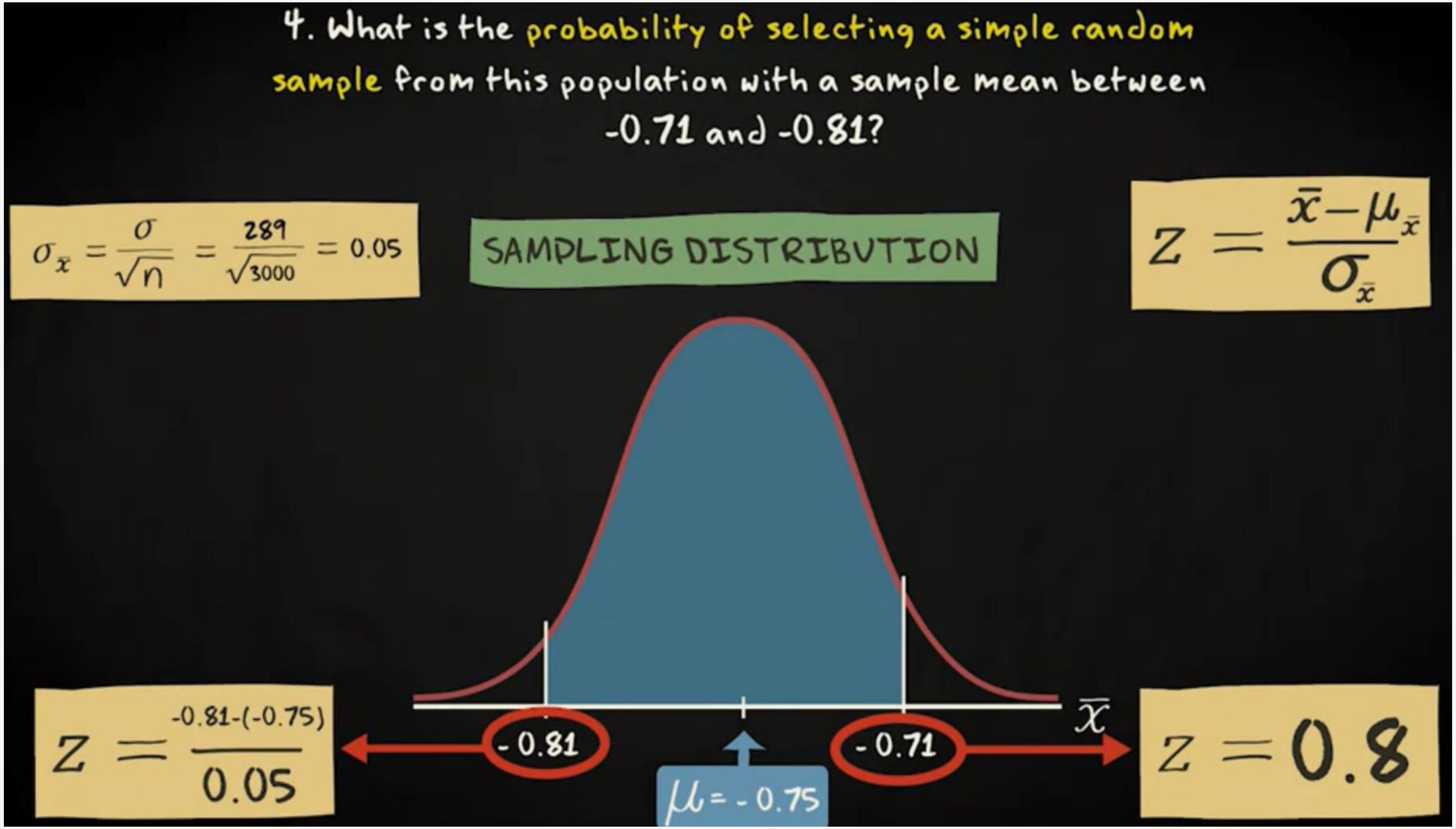
4. What is the probability of selecting a simple random sample from this population with a sample mean between -0.71 and -0.81?

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{289}{\sqrt{3000}} = 0.05$$

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

$$Z = \frac{-0.71-(-0.75)}{\sigma_{\bar{x}}}$$

Exercise



4. What is the probability of selecting a simple random sample from this population with a sample mean between -0.71 and -0.81?

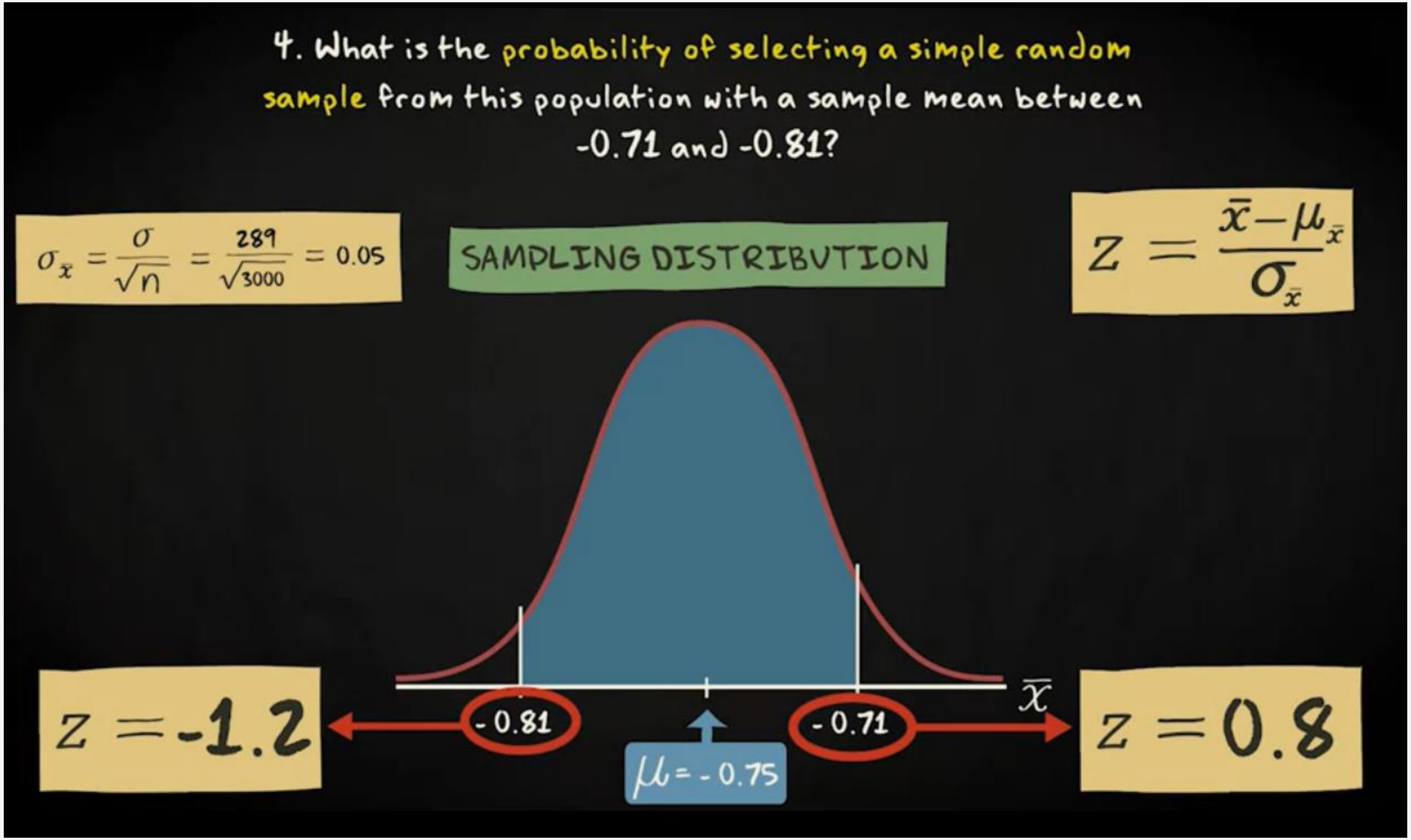
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{289}{\sqrt{3000}} = 0.05$$

$$Z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

$$Z = \frac{-0.81-(-0.75)}{0.05}$$

$$z = 0.8$$

Exercise



4. What is the probability of selecting a simple random sample from this population with a sample mean between -0.71 and -0.81?

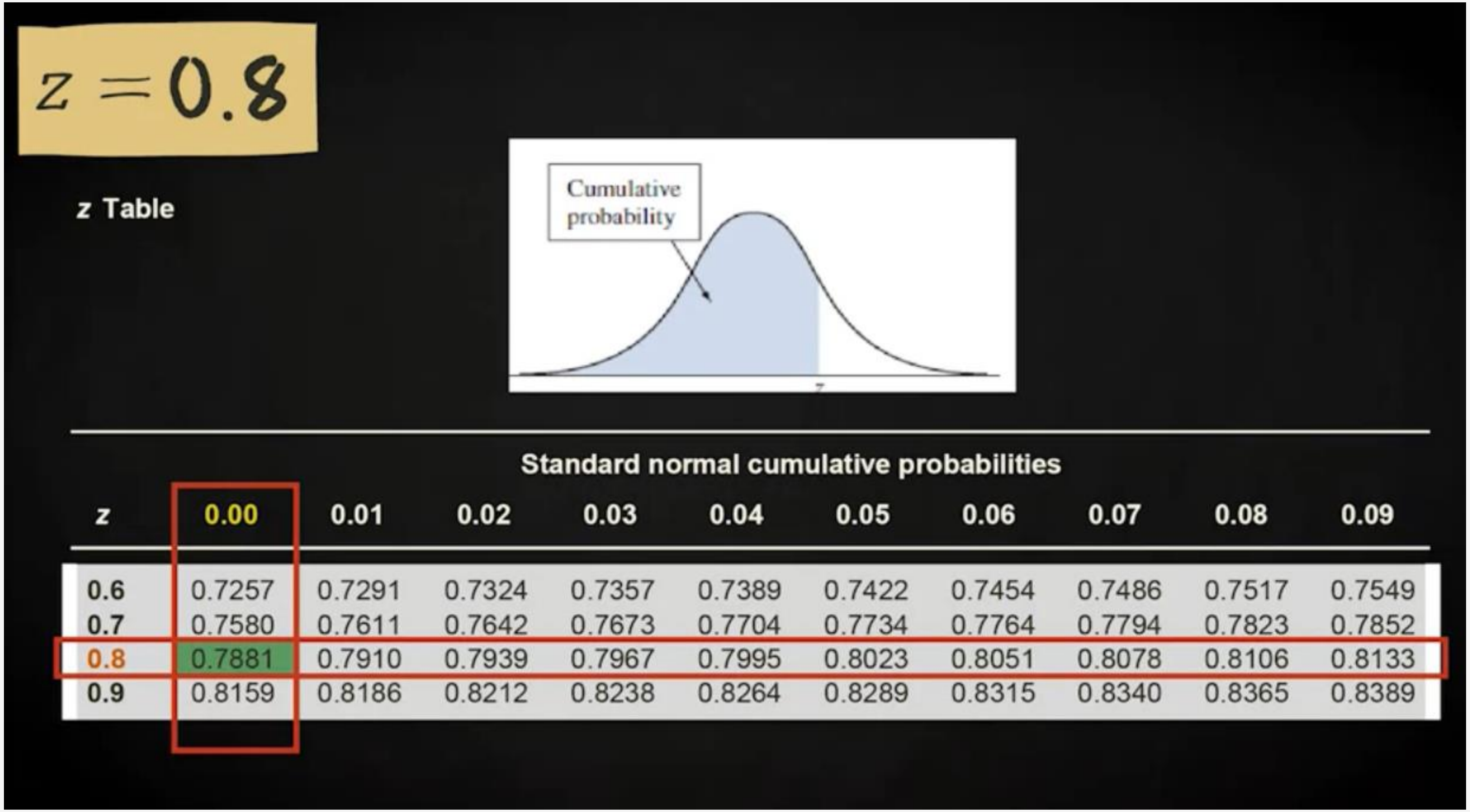
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{289}{\sqrt{3000}} = 0.05$$

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

$$z = -1.2$$

$$z = 0.8$$

Exercise



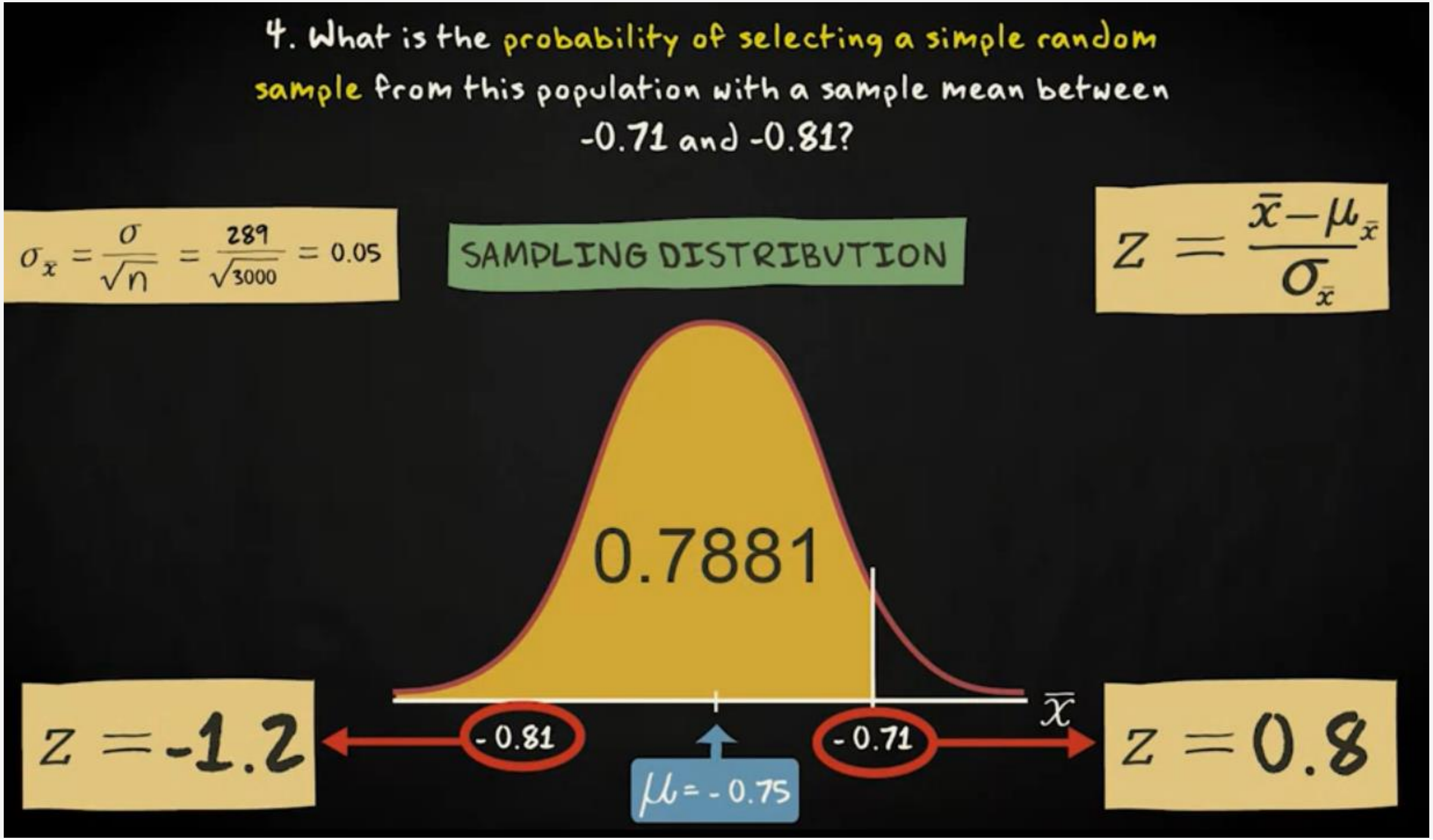
$z = 0.8$

z Table

Standard normal cumulative probabilities

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389

Exercise



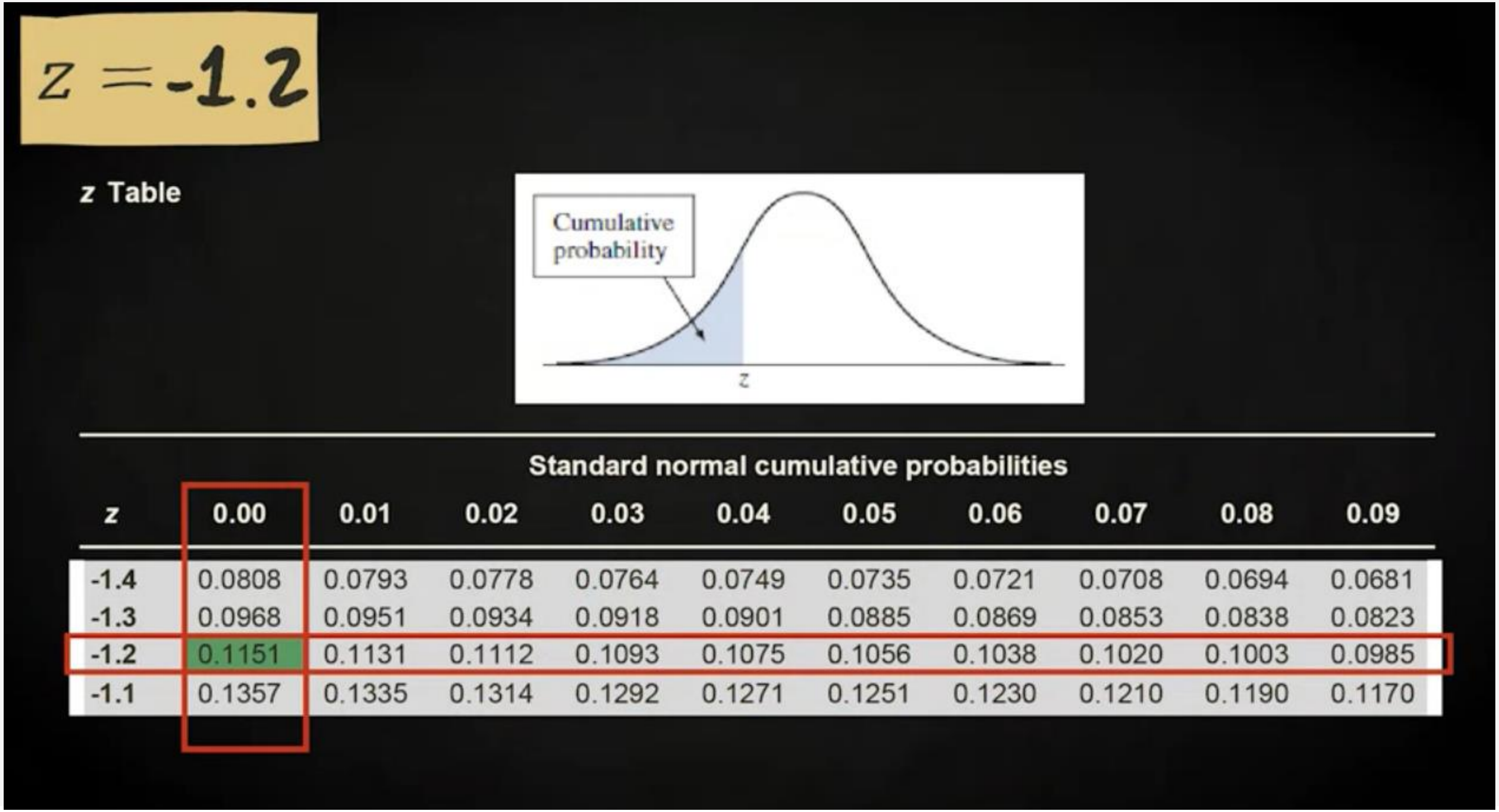
4. What is the probability of selecting a simple random sample from this population with a sample mean between -0.71 and -0.81?

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{289}{\sqrt{3000}} = 0.05$$

SAMPLING DISTRIBUTION

$$z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Exercise



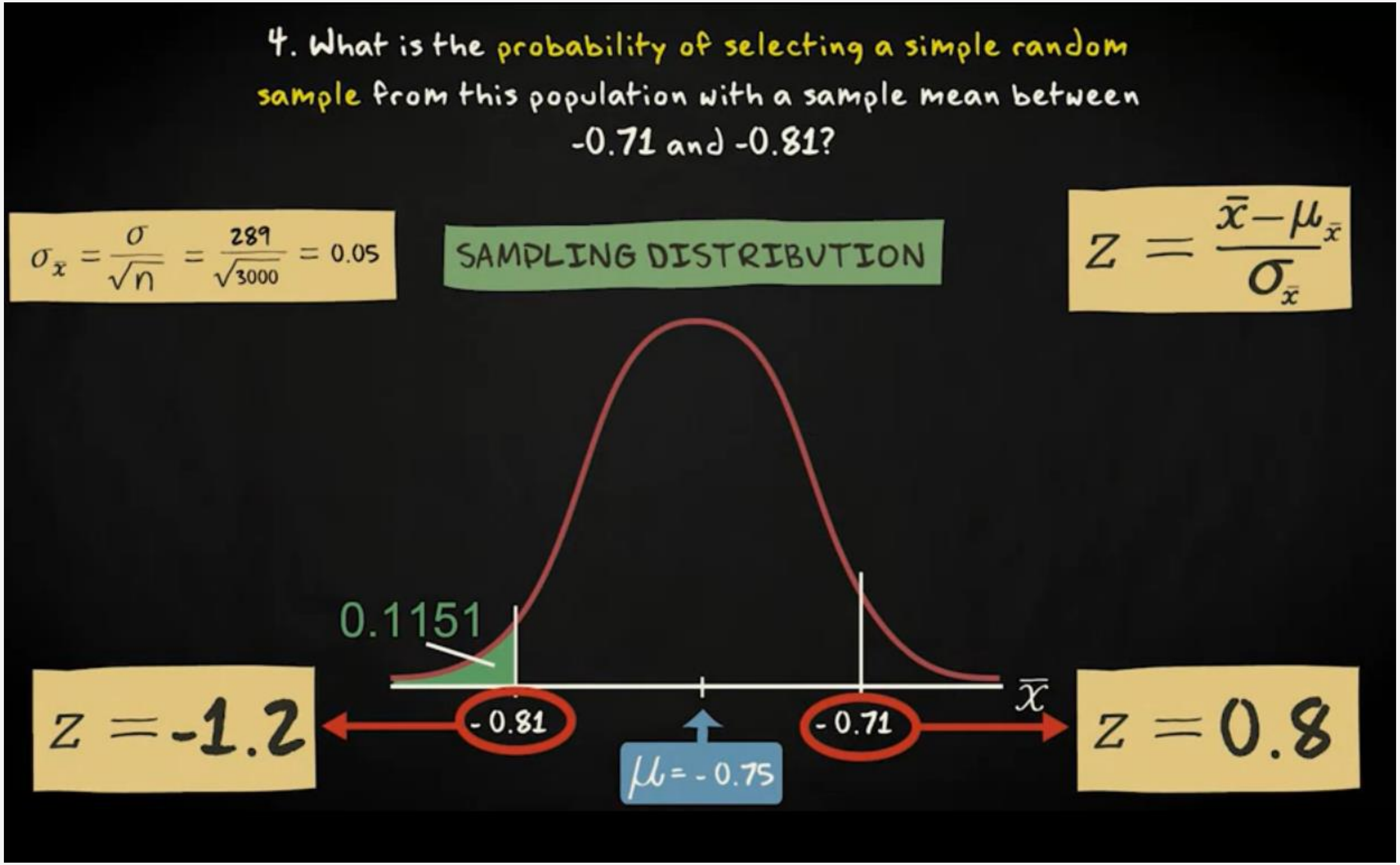
$z = -1.2$

z Table

Standard normal cumulative probabilities

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170

Exercise

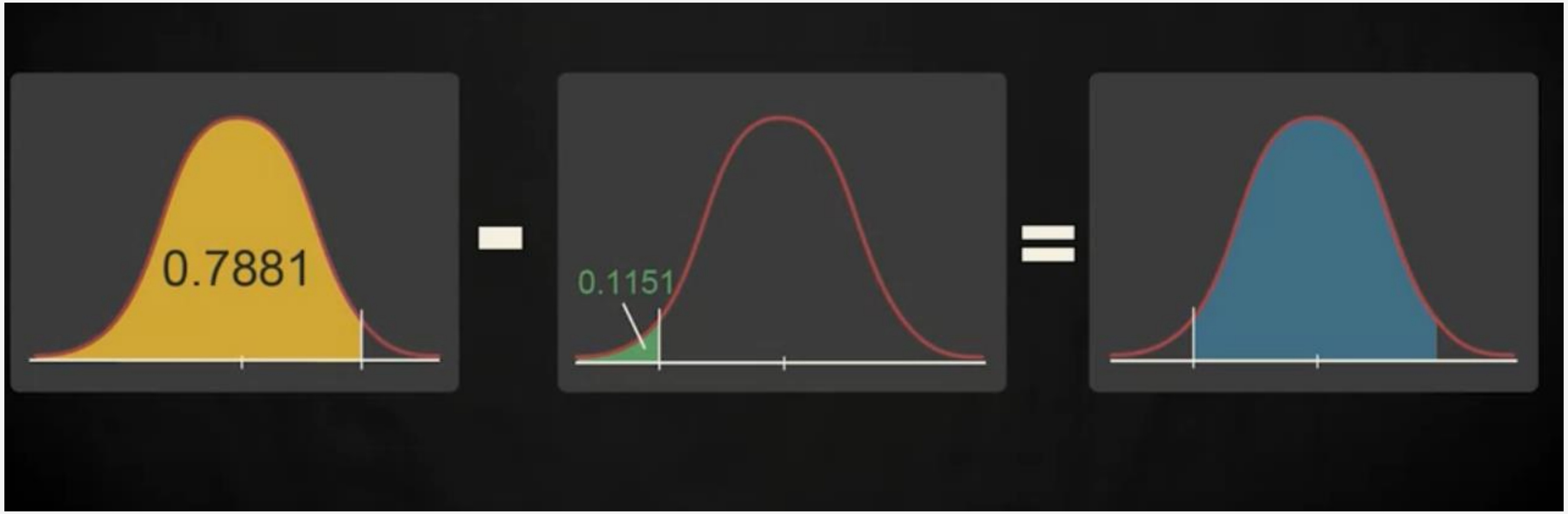


4. What is the probability of selecting a simple random sample from this population with a sample mean between -0.71 and -0.81?

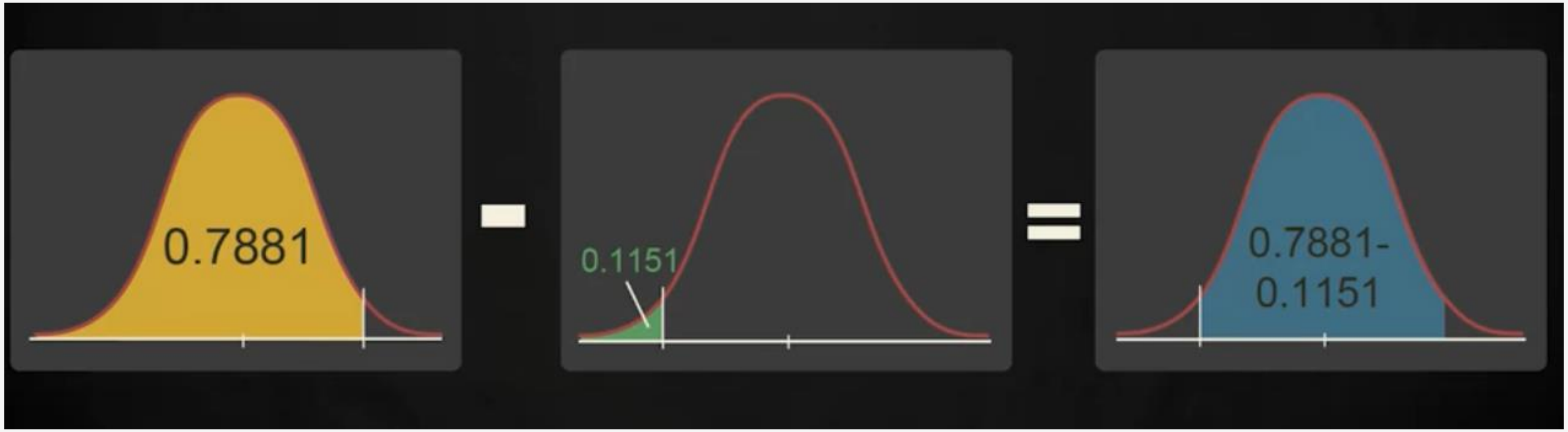
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{289}{\sqrt{3000}} = 0.05$$

$$Z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

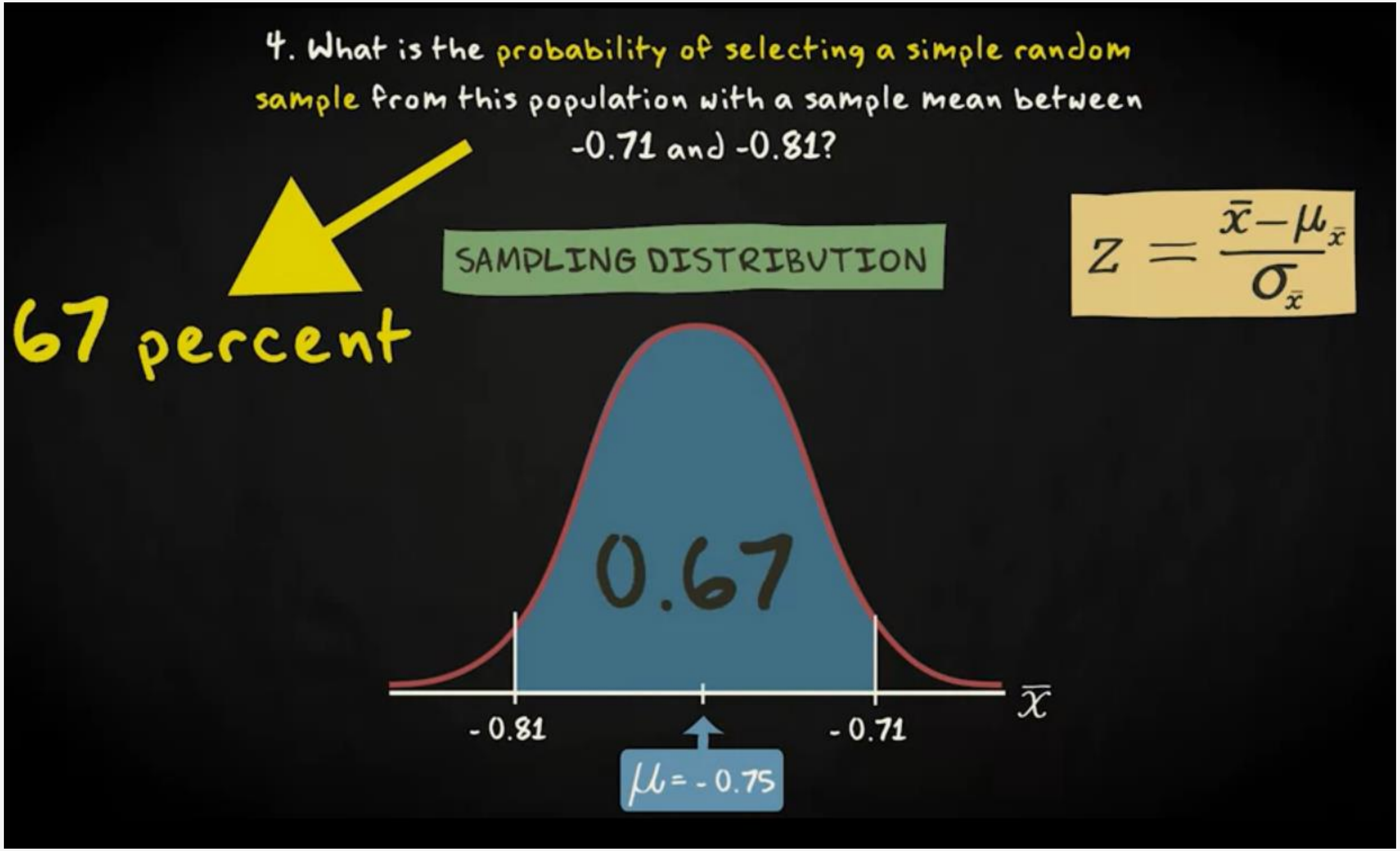
Exercise



Exercise



Exercise



4. What is the probability of selecting a simple random sample from this population with a sample mean between -0.71 and -0.81?

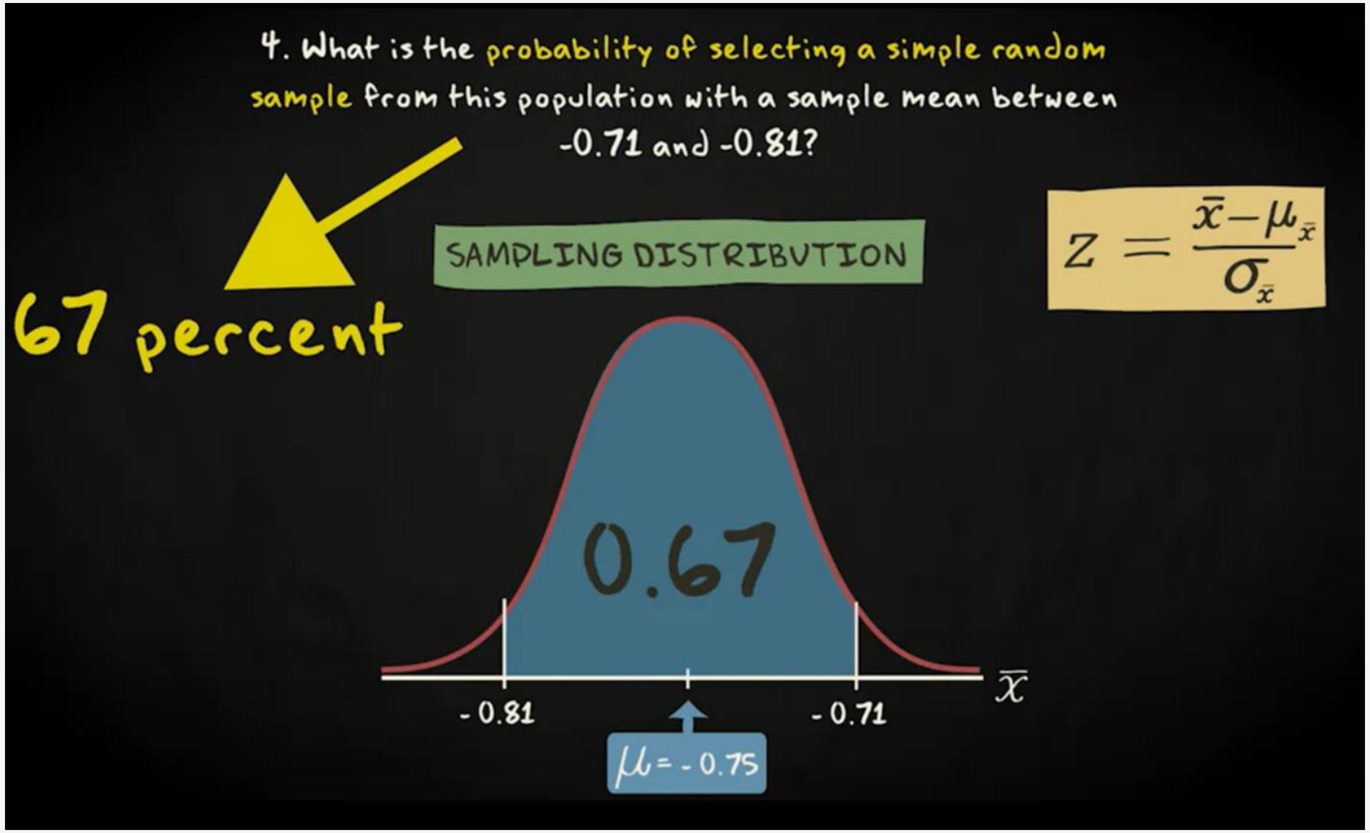
67 percent

SAMPLING DISTRIBUTION

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Exercise

Conclusion, **the probability of selecting a sample of n equals 3,000 with a sample mean between -0.71 and -0.81 is 67%.**



4. What is the probability of selecting a simple random sample from this population with a sample mean between -0.71 and -0.81?