Basics of Statistics

What is SciPy?

- SciPy is a free and open-source Python library used for scientific computing and technical computing.
- It is a collection of mathematical algorithms and convenience functions built on the NumPy extension of Python.
- It adds significant power to the interactive Python session by providing the user with high-level commands and classes for manipulating and visualizing data.
- Ref.c:/10-python/matplotlib_assignment.py

Install SciPy using pip

- pip install scipy
- conda install –c anaconda scipy

Sr.	Sub-Package	Description
1.	scipy.cluster	Cluster algorithms are used to vector quantization/ Kmeans.
2.	scipy.constants	It represents physical and mathematical constants.
3.	scipy.fftpack	It is used for Fourier transform.
4.	scipy.integrate	Integration routines
5.	scipy.interpolation	Interpolation
6.	scipy.linalg	It is used for linear algebra routine.
7.	scipy.io	It is used for data input and output.
8.	scipy.ndimage	It is used for the n-dimension image.
9.	scipy.odr	Orthogonal distance regression.
10.	scipy.optimize	It is used for optimization.
11.	scipy.signal	It is used in signal processing.
12.	scipy.sparse	Sparse matrices and associated routines.
13.	scipy.spatial	Spatial data structures and algorithms.
14.	scipy.special	Special Function.
15.	scipy.stats	Statistics.
16.	scipy.weaves	It is a tool for writing.

- **SciPy Cluster**
- Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. Each group which is formed from clustering is known as a cluster. There are two types of the cluster, which are:
- Central
- Hierarchy


2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.5
1.0
1.5
2.0
2.5

- **SciPy constants**
- There are a variety of constants that are included in the scipy.constant sub-package.
- These constants are used in the general scientific area. Let us see how these constant variables are imported and used.

```
#Import golden constant from the scipy
import scipy
print("sciPy -golden ratio  Value = %.18f"%scipy.constants.golden)
```

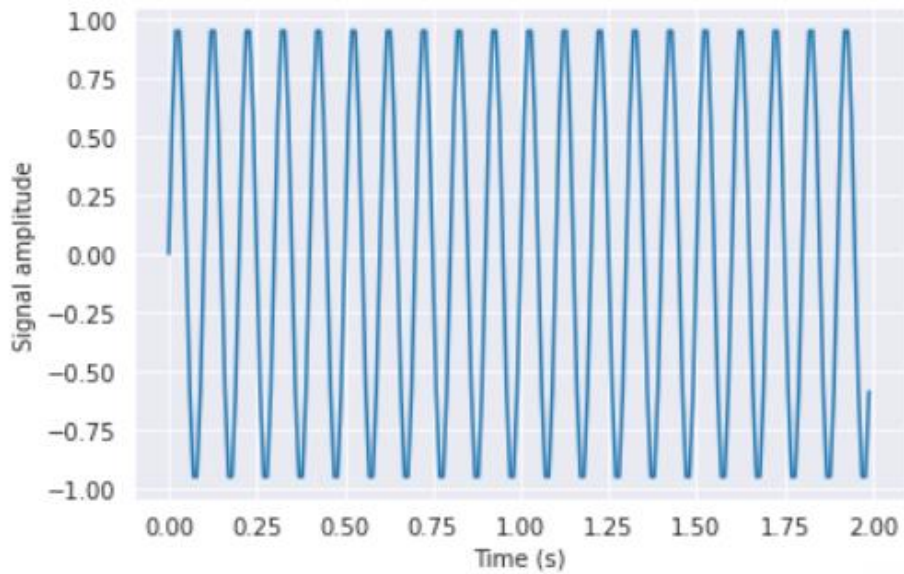
```
Output:
sciPy -golden ratio  Value = 1.6180339887498949 03
```

As you can see, we imported and printed the golden ratio constant using SciPy.The scipy.constant also provides the find() function, which returns a list of physical_constant keys containing a given string.

- **SciPy FFTpack**
- The FFT stands for Fast Fourier Transformation which is an algorithm for computing DFT.
- DFT is a mathematical technique which is used in converting spatial data into frequency data.
- SciPy provides the fftpack module, which is used to calculate Fourier transformation.
- In the example below, we will plot a simple periodic function of sin and see how the scipy.fft function will transform it.

```python
from matplotlib import pyplot as plt
import numpy as np
import seaborn as sns
sns.set_style("darkgrid")
#Frequency in terms of Hertz
fre  = 10
#Sample rate
fre_samp = 100
t = np.linspace(0, 2, 2 * fre_samp, endpoint = False )
a = np.sin(fre  * 2 * np.pi * t)
plt.plot(t, a)
plt.xlabel('Time (s)')
plt.ylabel('Signal amplitude')
plt.show()
```

Output:


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Signal amplitude
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Time (s)

- **SciPy integrate**
- The integrate sub package of the Scipy package contains a lot of functions that allow us to calculate the integral of some complex functions. If you use the help function, you will find all the different types of integrals you can calculate. Here is how:
- **Single Integrals:**
- We use the quad function to calculate the single integral of a function. Numerical integrate is sometimes called quadrature and hence the name quad. The function has three parameters:

Single Integrals:

We use the quad function to calculate the single integral of a function. Numerical integrate is sometimes called quadrature and hence the name quad. The function has three parameters:

```
scipy.integrate.quad(f,a,b)
 f – Function to be integrated.
 a–lower limit.
 b– upper limit.
```

- **SciPy Interpolation**
- Interpolation is the process of estimating unknown values that fall between known values.
- SciPy provides us with a sub-package scipy.
- interpolation which makes this task easy for us. Using this package, we can perform 1-D or univariate interpolation and Multivariate interpolation.
- Multivariate interpolation (spatial interpolation) is a kind interpolation on functions that consist of more than one variables.

- **SciPy linalg**
- SciPy has very fast linear algebra capabilities as it is built using the optimized ATLAS (Automatically Tuned Linear Algebra Software), LAPACK(Linear Algebra Package) and BLAS(Basic Linear Algebra Subprograms) libraries. All of these linear algebra routines can operate on an object that can be converted into a two-dimensional array and also returns the output as a two-dimensional array.

- **Finding a determinant of a square matrix**
- The determinant is a scalar value that can be computed from the elements of a square matrix and encodes certain properties of the linear transformation described by the matrix. The determinant of a matrix A is denoted det, det A, or |A|. In SciPy, this is computed using the det() function. It takes a matrix as input and returns a scalar value.

```python
#importing the scipy and numpy packages
from scipy import linalg
import numpy as np

#Declaring the numpy array
A = np.array([[1,2,9],[3,4,8],[7,8,4]])
#Passing the values to the det function
x = linalg.det(A)

#printing the result
print('Determinant of \n{} \n is {}'.format(A,x))
```

```
Output:
Determinant of
[[1 2 9]
 [3 4 8]
 [7 8 4]]
 is 3.999999999999986
```

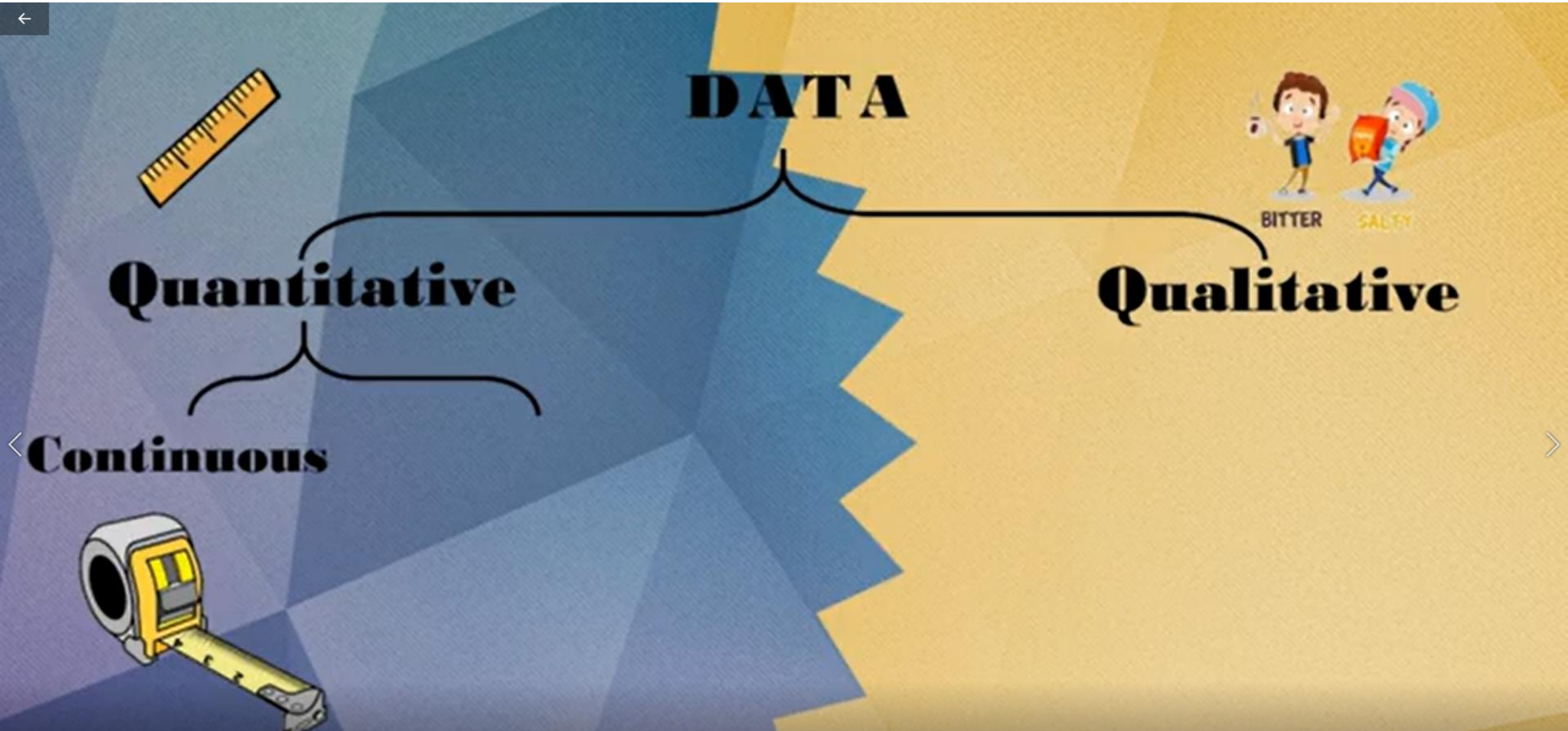
- **SciPy IO (Input & Output)**

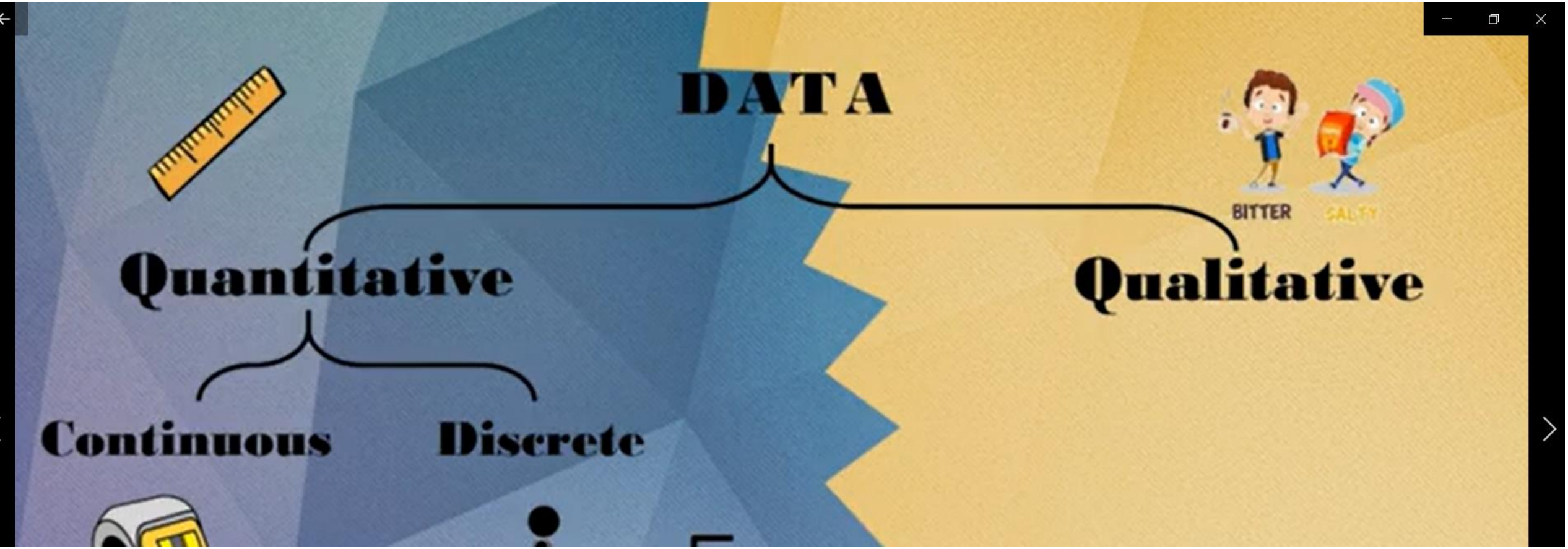
- The functions provided by the scipy.io package enables us to work around with different formats of files such as:
- Matlab
- IDL
- Matrix Market
- Wave

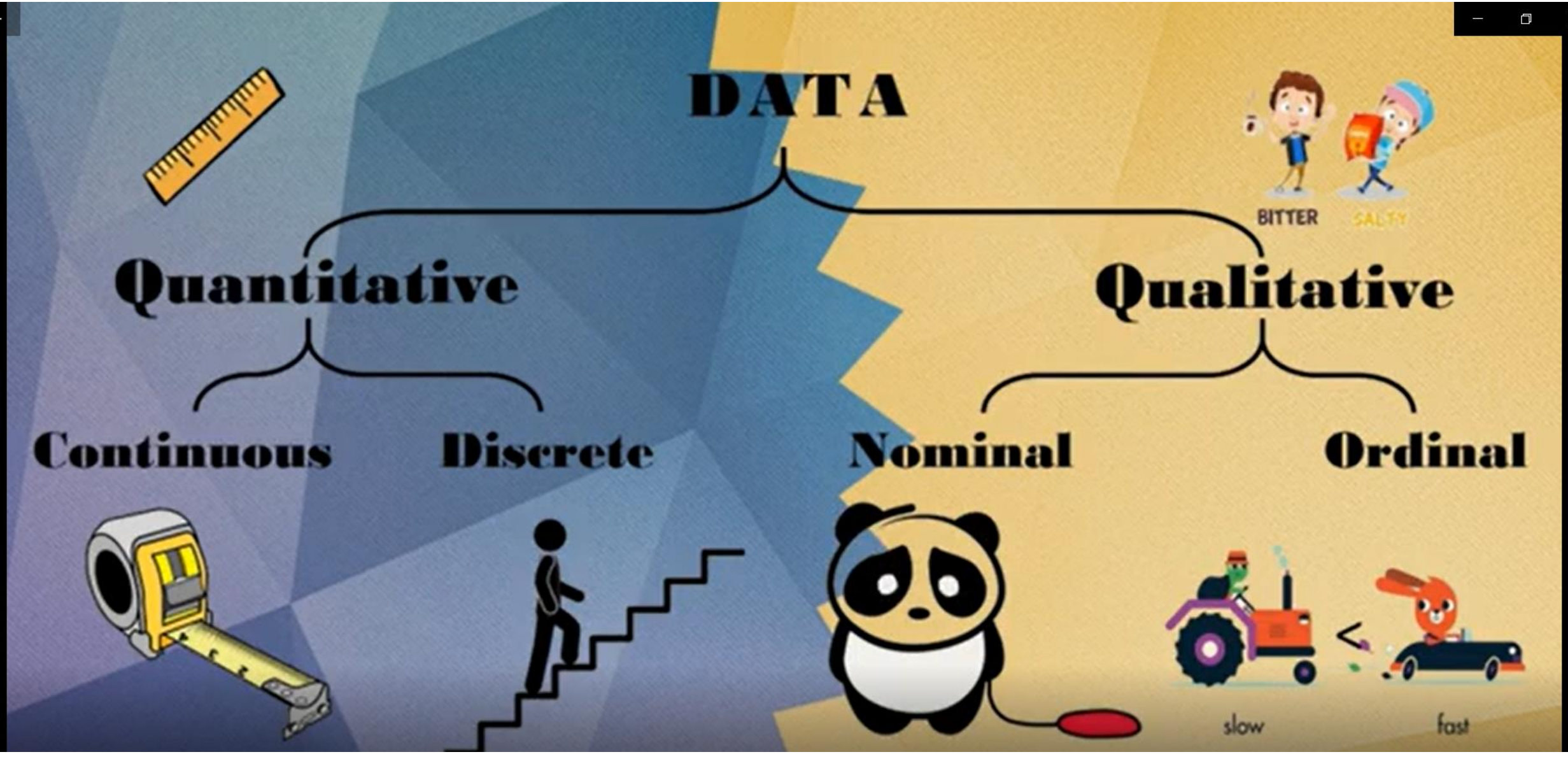
- **SciPy Ndimage**

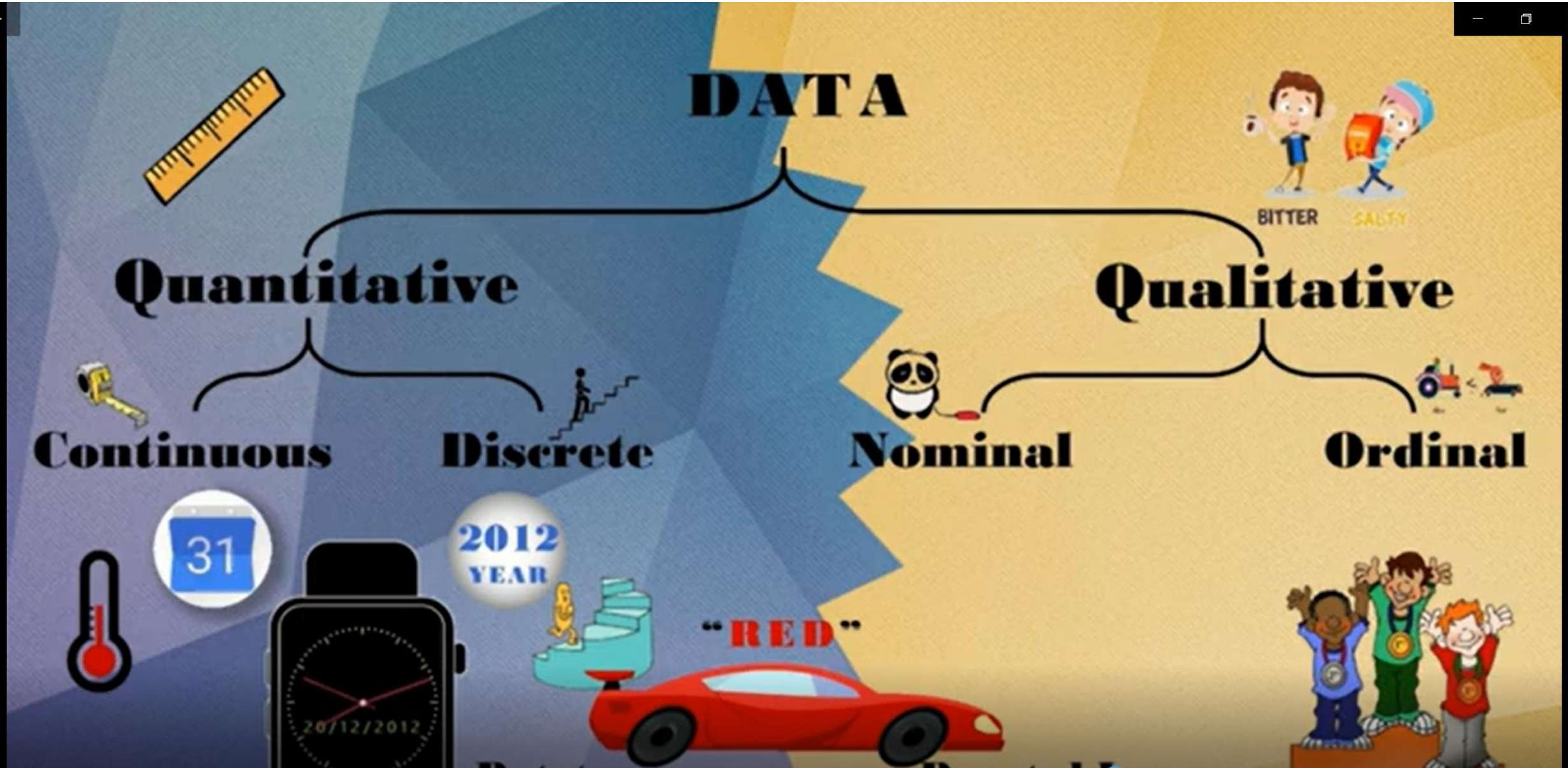
- The SciPy provides the ndimage (n-dimensional image) package, that contains the number of general image processing and analysis functions. Some of the most common tasks in image processing are as follows:
- Basic manipulations − Cropping, flipping, rotating, etc.
- Image filtering − Denoising, sharpening, etc.
- Image segmentation − Labeling pixels corresponding to different objects
- Classification
- Feature extraction

- import scipy.misc
- import matplotlib.pyplot as plt
- face = scipy.misc.face()#returns an image of raccoon
- #display image using matplotlib
- plt.imshow(face)
- plt.show()


DATA
Quantitative
Qualitative
Continuous
BITTER
SALTY


DATA
Quantitative
Qualitative
Continuous
Discrete
BITTER
SALTY


DATA
Quantitative
Qualitative
Continuous
Discrete
Nominal
Ordinal
BITTER
SALTY
slow
fast


DATA
Quantitative
Qualitative
Continuous
Discrete
Nominal
Ordinal
BITTER
SALTY
31
2012
YEAR
"RED"
20/12/2012

Mean

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000

Average	6250

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000
Average	6250

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000
Elon Musk	10 million
Average	1.43 million

Nina	Rob	Rafiq	Tao	Sofia	Mohan	Elon Musk
4000	5000	6000	7000	7500	8000	10 million

Median = 7000

Median


Nina
Rob
Rafiq
Tao
Sofia
Mohan
Elon Musk
4000
5000
6000
7000
7500
8000
10 million
Median = 7000
Nina
Rob
Rafiq
Tao
Prem
Sofia
Mohan
Elon Musk
4000
5000
6000
7000
8000
7500
8000
10 million
Median = 7500

Use case # 2: Handling missing values

Name	Monthly Income ($)	Credit Score	Approve Loan?
Rob	5000	650	No
Rafiq	6000	400	No
Nina	4000	780	Yes
Sofia	??	810	Yes
Mohan	8000	410	No
Tao	7000	850	Yes
Elon Musk	10 million	880	Yes

Name	Monthly Income ($)	Credit Score	Approve Loan?
Rob	5000	650	No
Rafiq	6000	400	No
Nina	4000	780	Yes
Sofia	1.6 million	810	Yes
Mohan	8000	410	No
Tao	7000	850	Yes
Elon Musk	10 million	880	Yes

Name	Monthly Income ($)	Credit Score	Approve Loan?
Rob	5000	650	No
Rafiq	6000	400	No
Nina	4000	780	Yes
Sofia	6500	810	Yes
Mohan	8000	410	No
Tao	7000	850	Yes
Elon Musk	10 million	880	Yes

4000 5000 6000 7000 8000 10 million

Name	Monthly Income ($)	Credit Score	Approve Loan?
Rob	5000	650	No
Rafiq	6000	400	No
Nina	4000	780	Yes
Sofia	6600	810	Yes
Mohan	8000	410	No
Tao	7000	850	Yes
Elon Musk	10 million	880	Yes

4000 5000 6000 7000 8000 10 million

How mean and median is used in data science?

Opening car show room

Descriptive Analysis

Loan approval model

Data Cleaning (Filling NA Values)

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000
Elon Musk	10 million

Outlier

Nina	Rob	Rafiq	Tao	Sofia	Mohan	Elon Musk
4000	5000	6000	7000	7500	8000	10 million

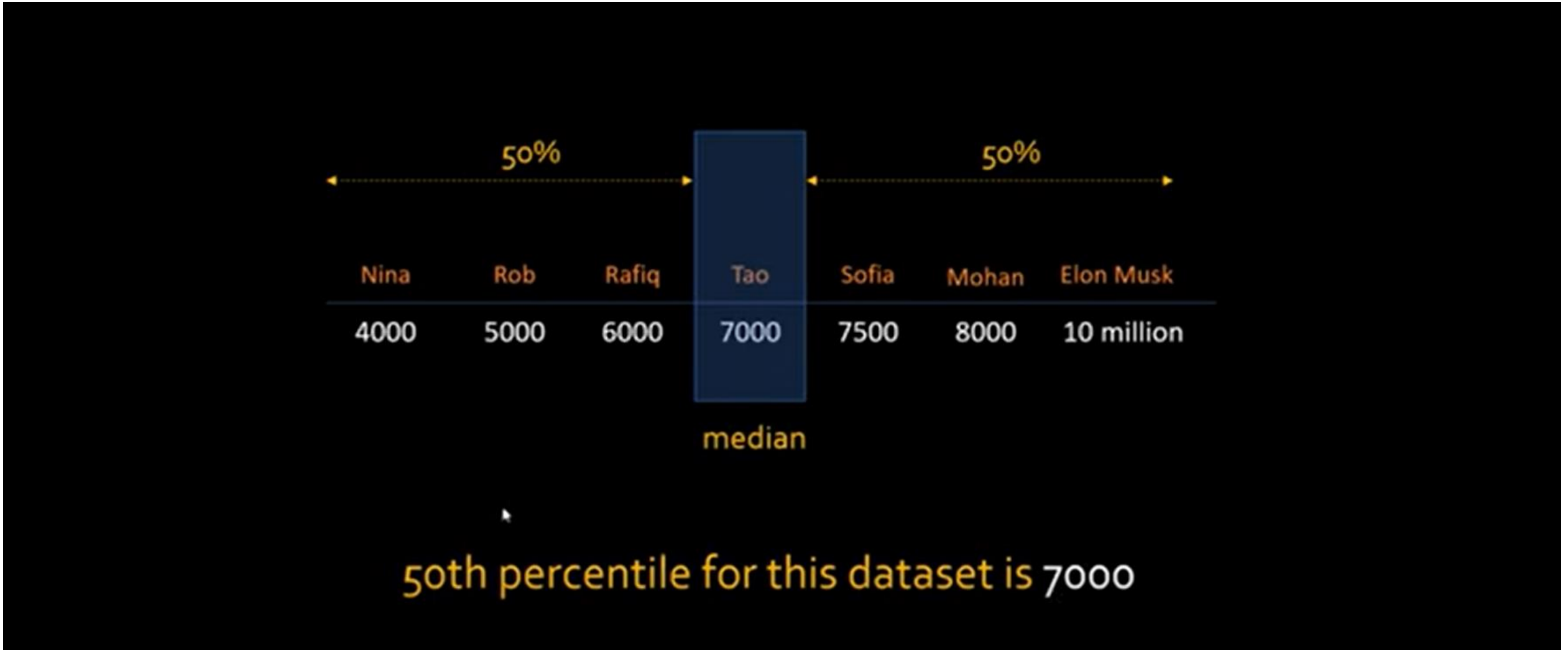
Percentile

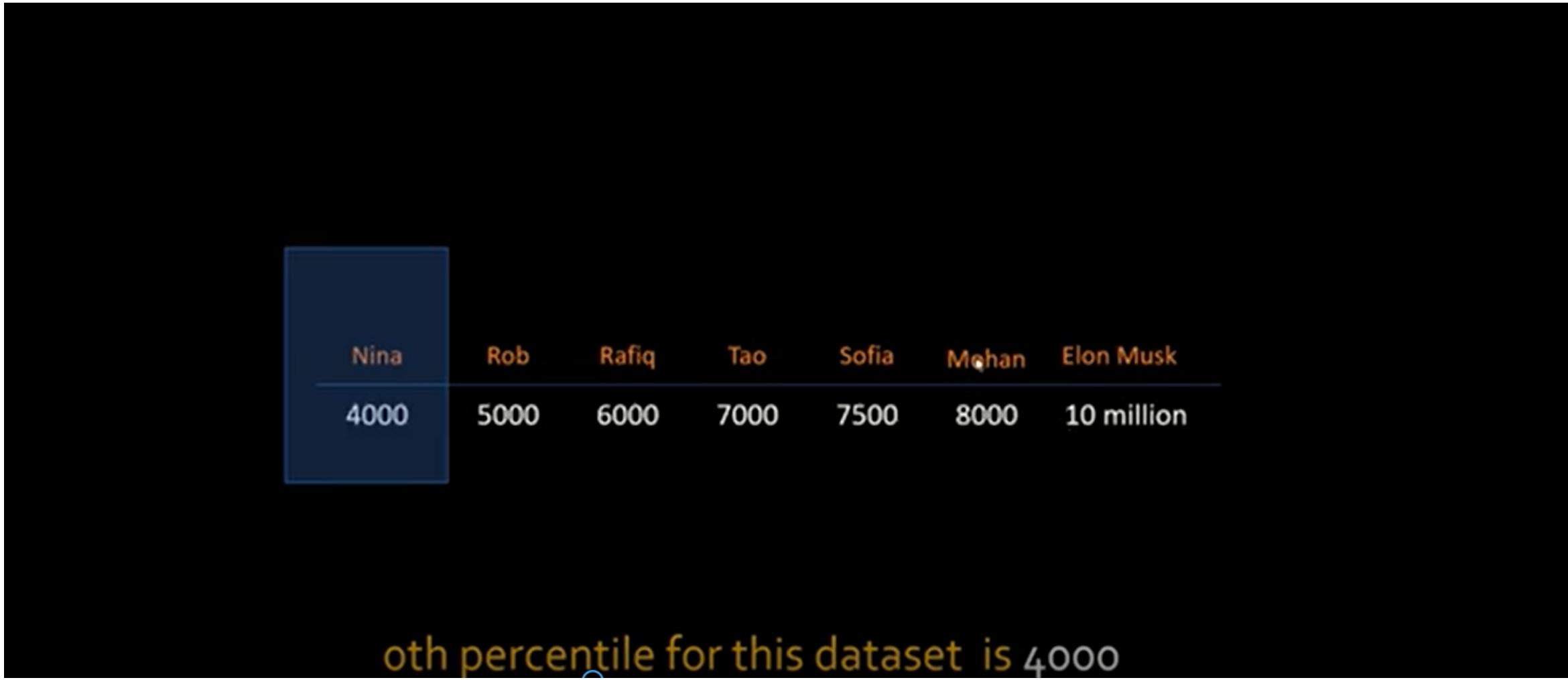
Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000
Elon Musk	10 million

Outlier

Nina	Rob	Rafiq	Tao	Sofia	Mohan	Elon Musk
4000	5000	6000	7000	7500	8000	10 million

Percentile


50%
50%
Nina
Rob
Rafiq
Tao
Sofia
Mohan
Elon Musk
4000
5000
6000
7000
7500
8000
10 million
median
50th percentile for this dataset is 7000


Nina
Rob
Rafiq
Tao
Sofia
Mohan
Elon Musk
4000
5000
6000
7000
7500
8000
10 million
oth percentile for this dataset is 4000

What is 25th percentile for this dataset ?

Total values =7

25% of 7 is 1.75 which is approximately 2 data points.

Nina	Rob	Rafiq	Tao	Sofia	Mohan	Elon Musk
4000	5000	6000	7000	7500	8000	10 million

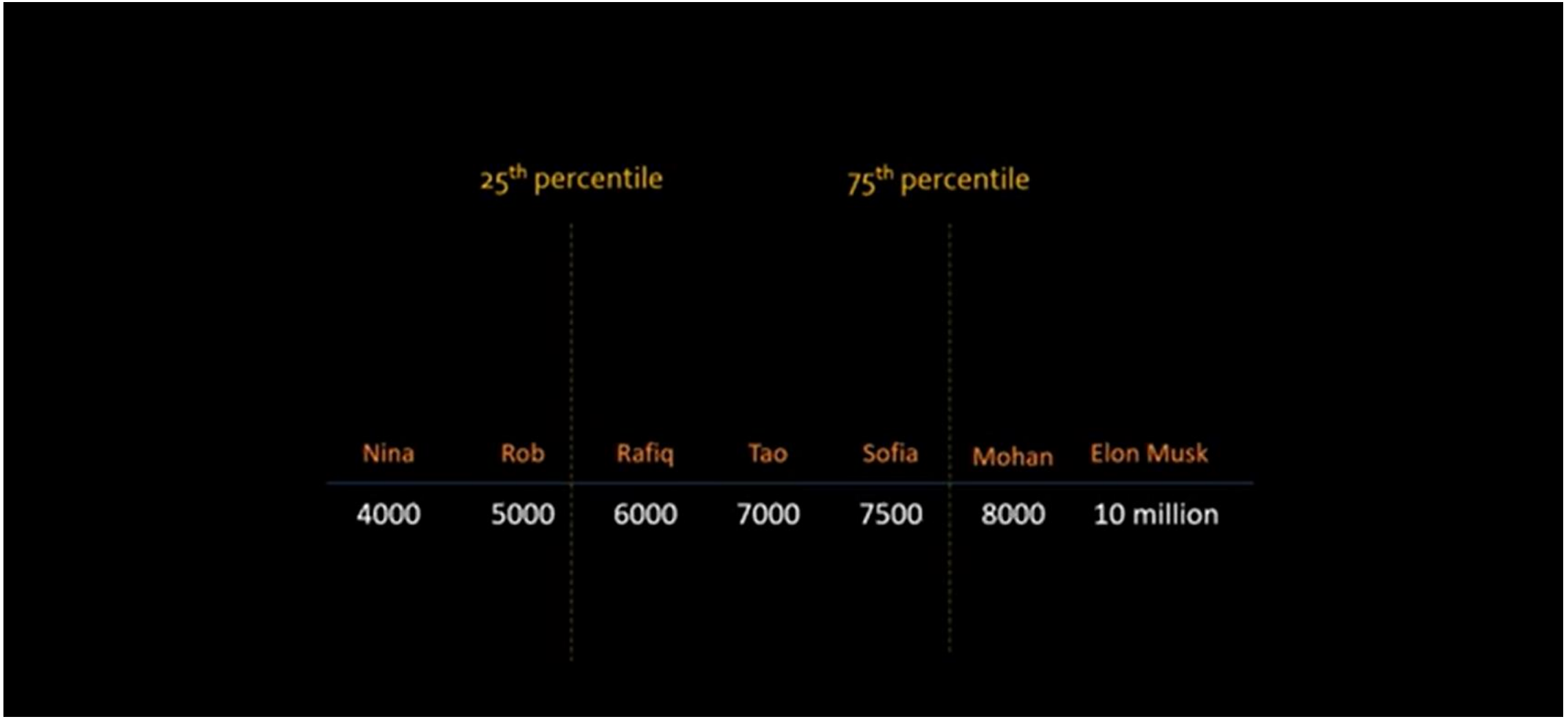
What is 75th percentile for this dataset ?

Total values =7

75% of 7 is 5.25 ~ 5 data points.

Nina	Rob	Rafiq	Tao	Sofia	Mohan	Elon Musk
4000	5000	6000	7000	7500	8000	10 million

75^{th} percentile is 7750


25th percentile
75th percentile
Nina
Rob
Rafiq
Tao
Sofia
Mohan
Elon Musk
4000
5000
6000
7000
7500
8000
10 million

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000
Elon Musk	10 million

Outlier

99%

Nina	Rob	Rafiq	Tao	Sofia	Mohan	Elon Musk
4000	5000	6000	7000	7500	8000	10 million

We can remove outlier using 99th percentile

What is Mode?

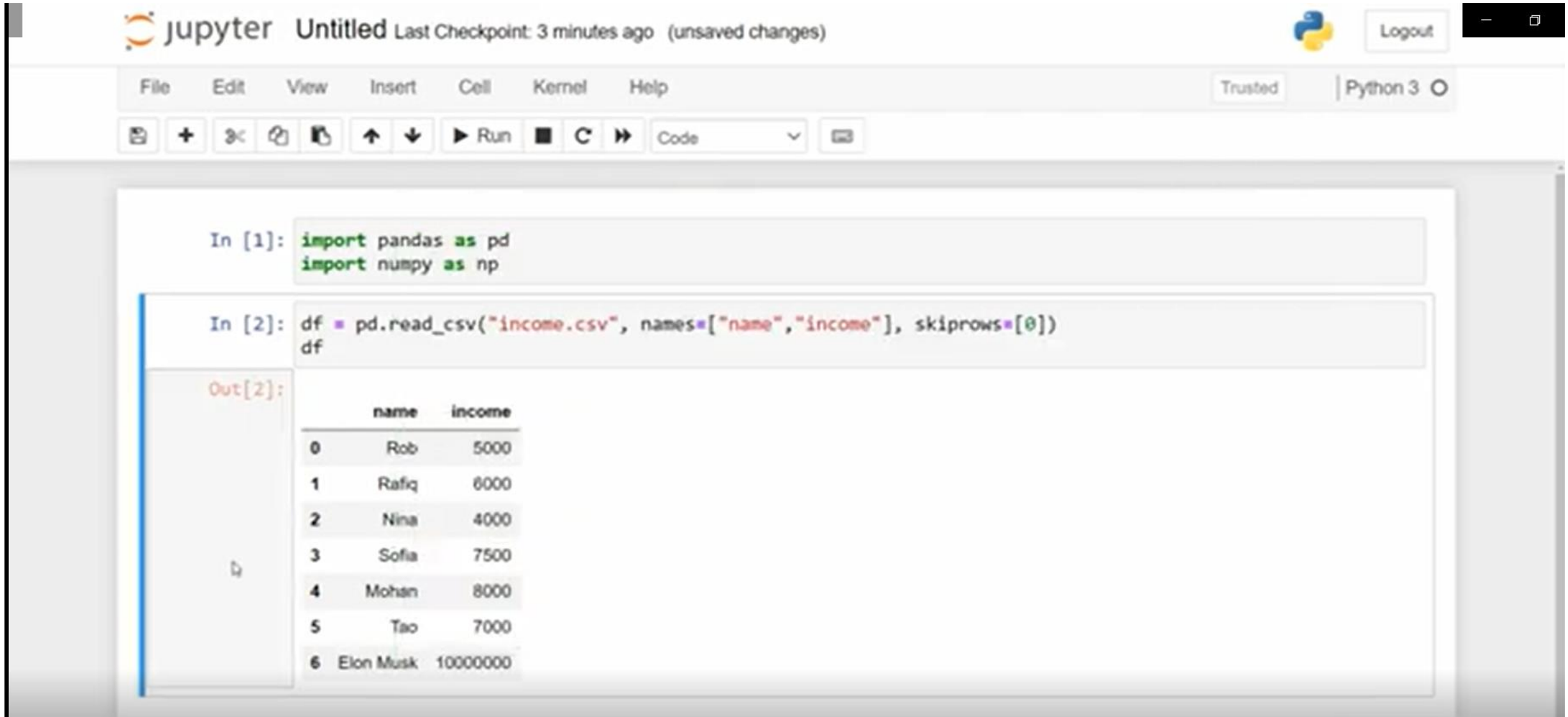
Name	Restaurant Vote
Rob	Mexican
Rafiq	Mexican
Nina	Italian
Sofia	Thai
Mohan	Italian
Tao	Mexican
Bantu	Indian

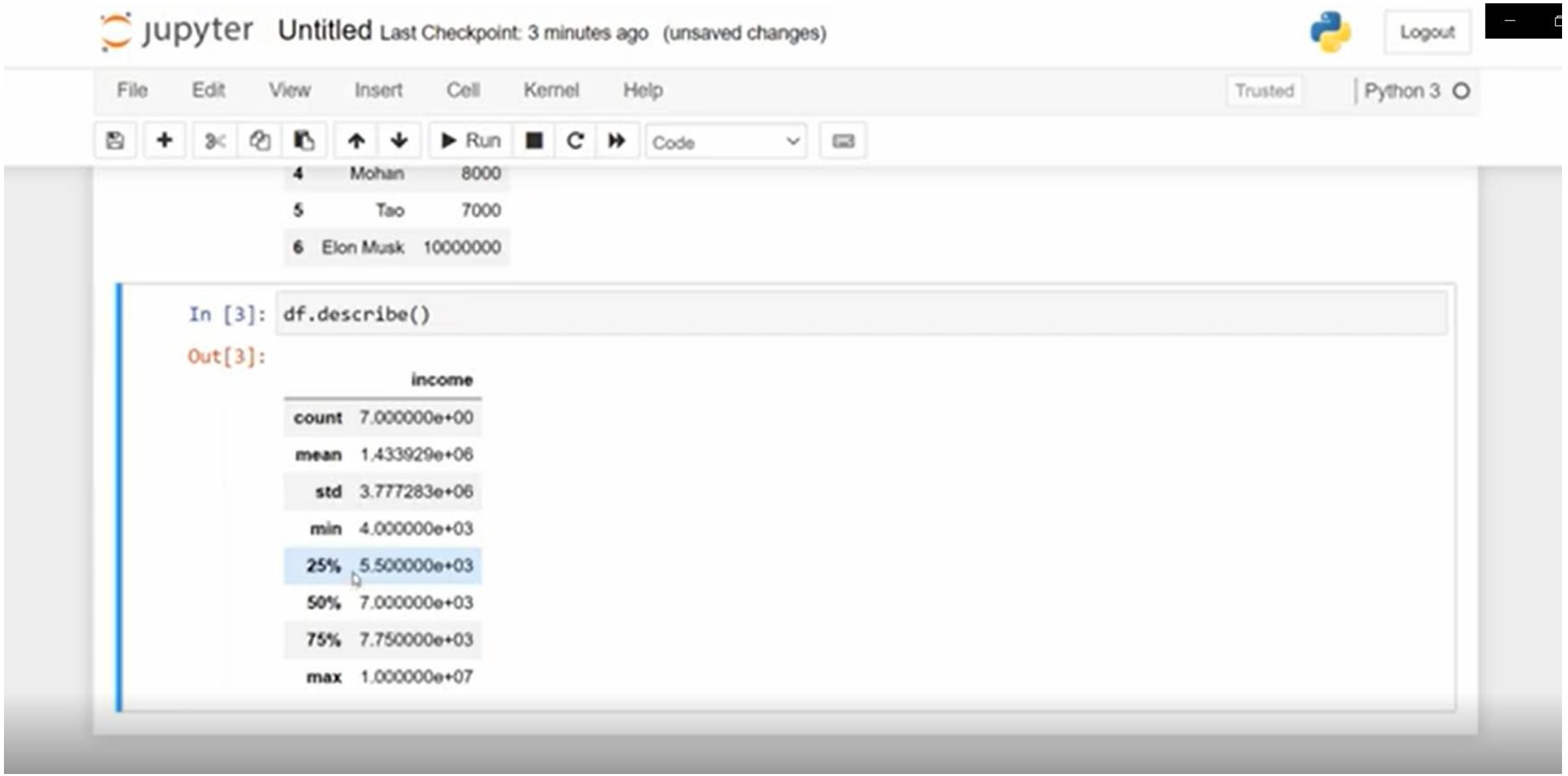
What is Mode?

Name	Restaurant Vote
Rob	Mexican
Rafiq	Mexican
Nina	Italian
Sofia	Thai
Mohan	Italian
Tao	Mexican
Bantu	Indian

Mode here is Mexican

Mode means most frequently occurring value in a dataset


jupyter Untitled Last Checkpoint: 3 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Code
In [1]:
import pandas as pd
import numpy as np
In [2]:
df = pd.read_csv("income.csv", names=["name","income"], skiprows=[0])
df
Out[2]:
name income
0 Rob 5000
1 Rafiq 6000
2 Nina 4000
3 Sofia 7500
4 Mohan 8000
5 Tao 7000
6 Elon Musk 10000000


jupyter Untitled Last Checkpoint: 3 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Code
4 Mohan 8000
5 Tao 7000
6 Elon Musk 10000000
In [3]: df.describe()
Out[3]:
income
count 7.000000e+00
mean 1.433929e+06
std 3.777283e+06
min 4.000000e+03
25% 5.500000e+03
50% 7.000000e+03
75% 7.750000e+03
max 1.000000e+07

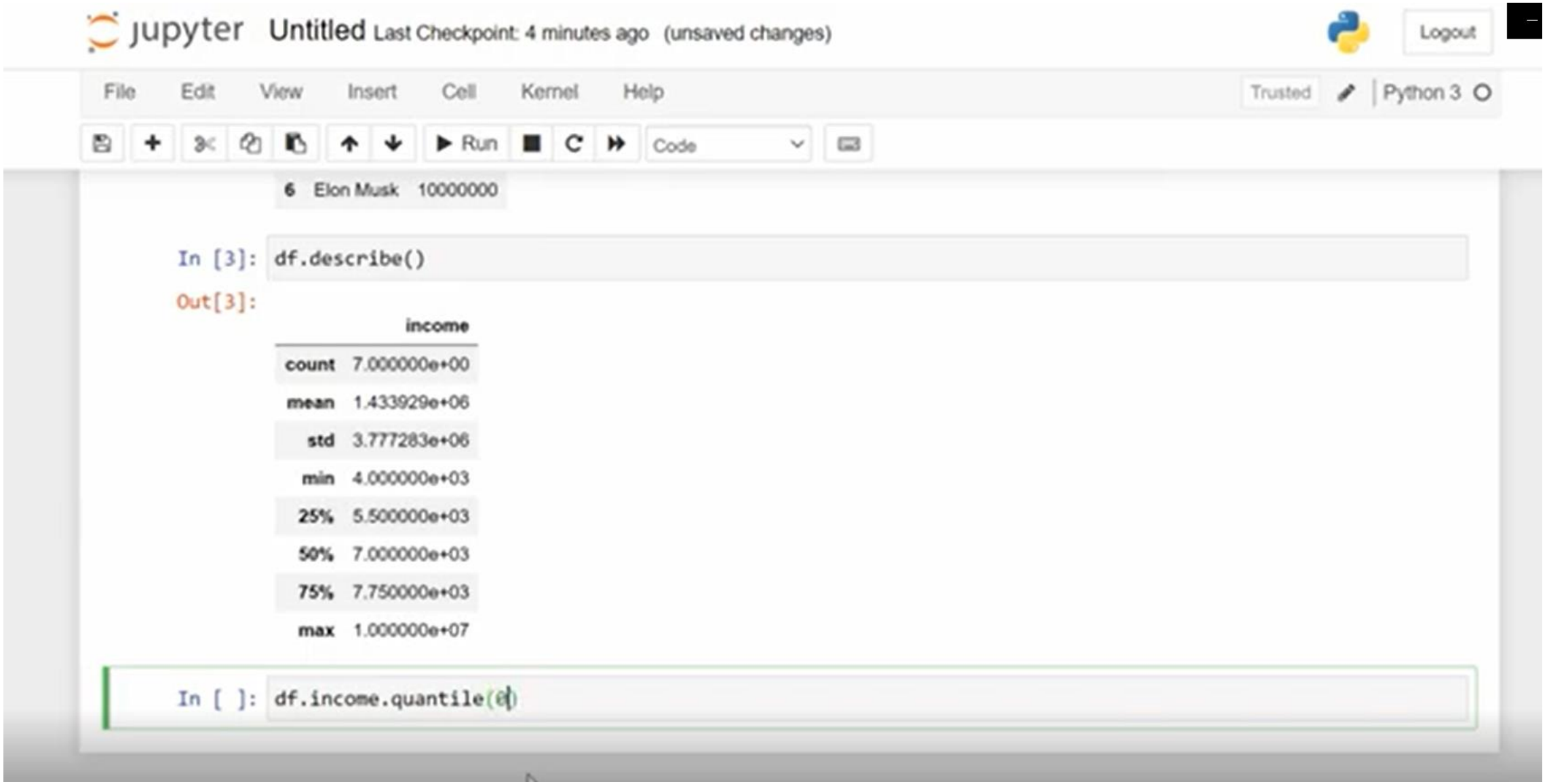
What is 25th percentile for this dataset ?

Total values =7

25% of 7 is 1.75 which is approximately 2 data points.

Nina	Rob	Rafiq	Tao	Sofia	Mohan	Elon Musk
4000	5000	6000	7000	7500	8000	10 million

25th percentile is 5500


jupyter Untitled Last Checkpoint: 4 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Code
6 Elon Musk 10000000
In [3]: df.describe()
Out[3]:
income
count 7.000000e+00
mean 1.433929e+06
std 3.777283e+06
min 4.000000e+03
25% 5.500000e+03
50% 7.000000e+03
75% 7.750000e+03
max 1.000000e+07
In [ ]: df.income.quantile(0)

jupyter Untitled Last Checkpoint: 4 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Help

Trusted Python 3

Run Code

6	Elon Musk	10000000

In [3]:
```
df.describe()
```

Out[3]:

	income
count	7.000000e+00
mean	1.433929e+06
std	3.777283e+06
min	4.000000e+03
25%	5.500000e+03
50%	7.000000e+03
75%	7.750000e+03
max	1.000000e+07

In [5]:
```
df.income.quantile(0.25)
```

Out[5]: 5500.0

jupyter Untitled Last Checkpoint: 5 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Help

Trusted Python 3

Run Code

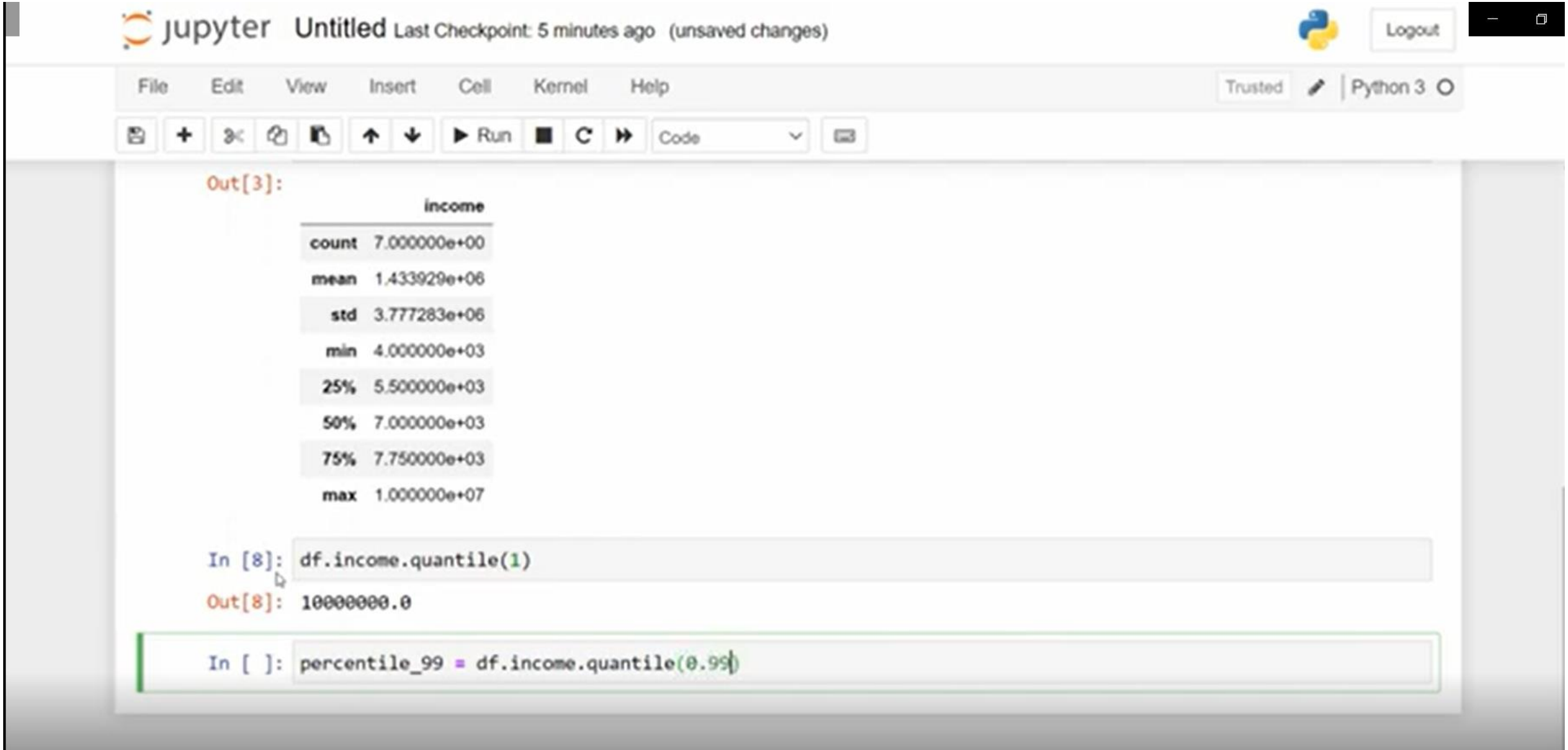
Out[3]:

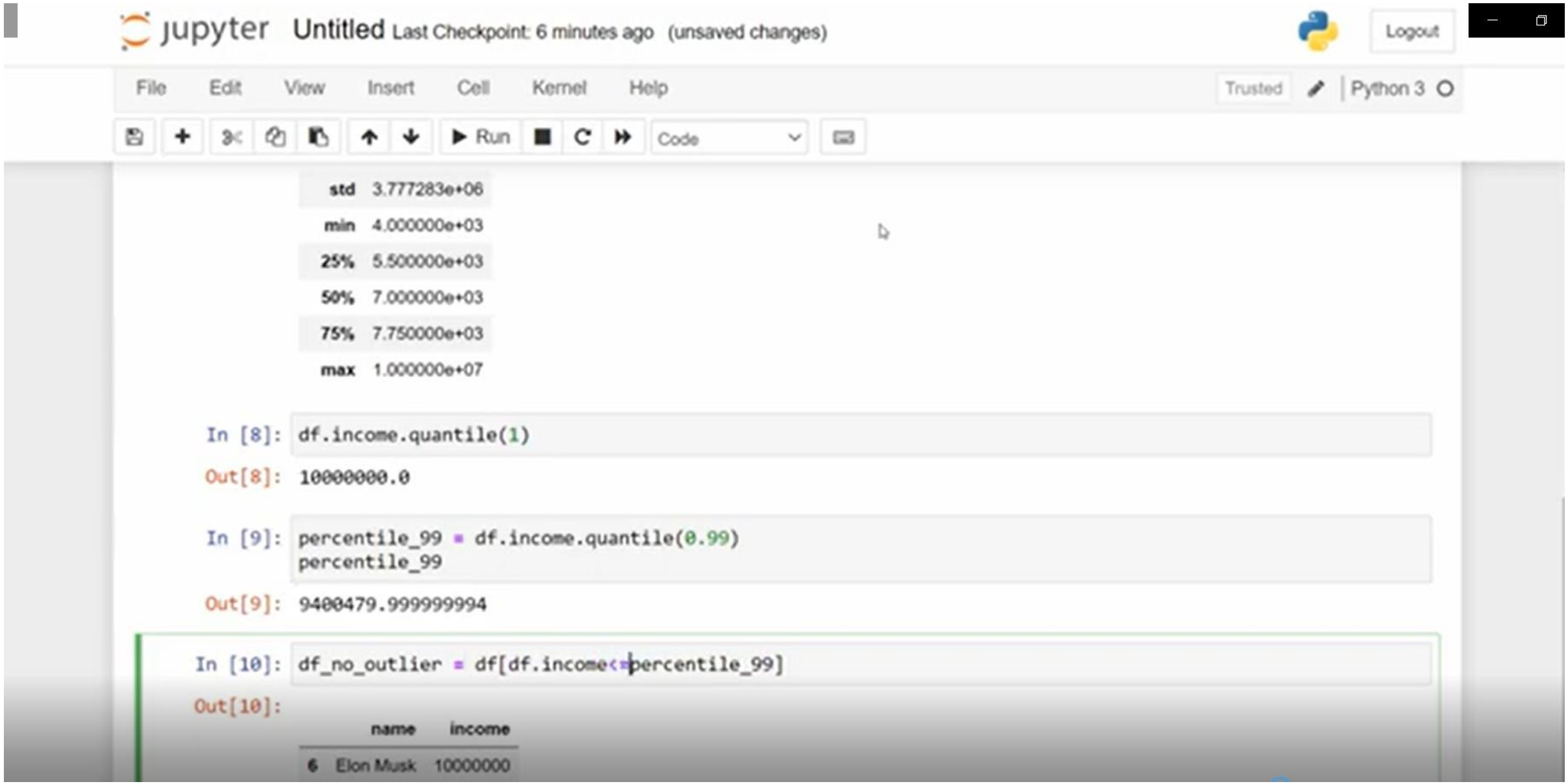
	income
count	7.000000e+00
mean	1.433929e+06
std	3.777283e+06
min	4.000000e+03
25%	5.500000e+03
50%	7.000000e+03
75%	7.750000e+03
max	1.000000e+07

In [8]:
```
df.income.quantile(1)
```

Out[8]: 10000000.0

In [ ]:


jupyter Untitled Last Checkpoint: 5 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Code
Out[3]:
income
count 7.000000e+00
mean 1.433929e+06
std 3.777283e+06
min 4.000000e+03
25% 5.500000e+03
50% 7.000000e+03
75% 7.750000e+03
max 1.000000e+07
In [8]: df.income.quantile(1)
Out[8]: 10000000.0
In [ ]: percentile_99 = df.income.quantile(0.99)


jupyter Untitled Last Checkpoint: 6 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Code
std 3.777283e+06
min 4.000000e+03
25% 5.500000e+03
50% 7.000000e+03
75% 7.750000e+03
max 1.000000e+07
In [8]: df.income.quantile(1)
Out[8]: 10000000.0
In [9]: percentile_99 = df.income.quantile(0.99)
percentile_99
Out[9]: 9400479.999999994
In [10]: df_no_outlier = df[df.income<=percentile_99]
Out[10]:
name income
6 Elon Musk 10000000

jupyter Untitled Last Checkpoint: 7 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Help

Trusted Python 3

Run Code

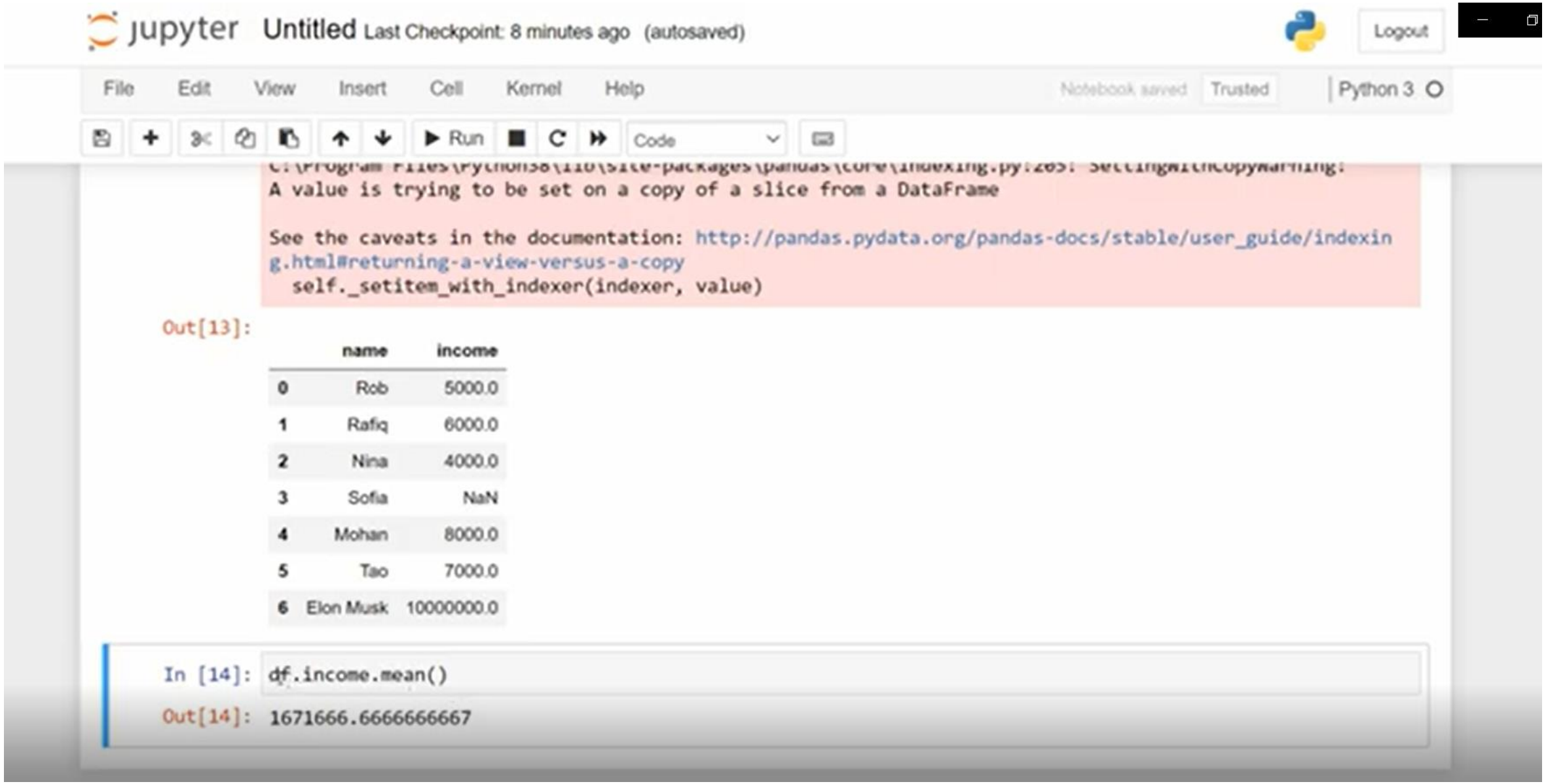
4	Mohan	8000
5	Tao	7000

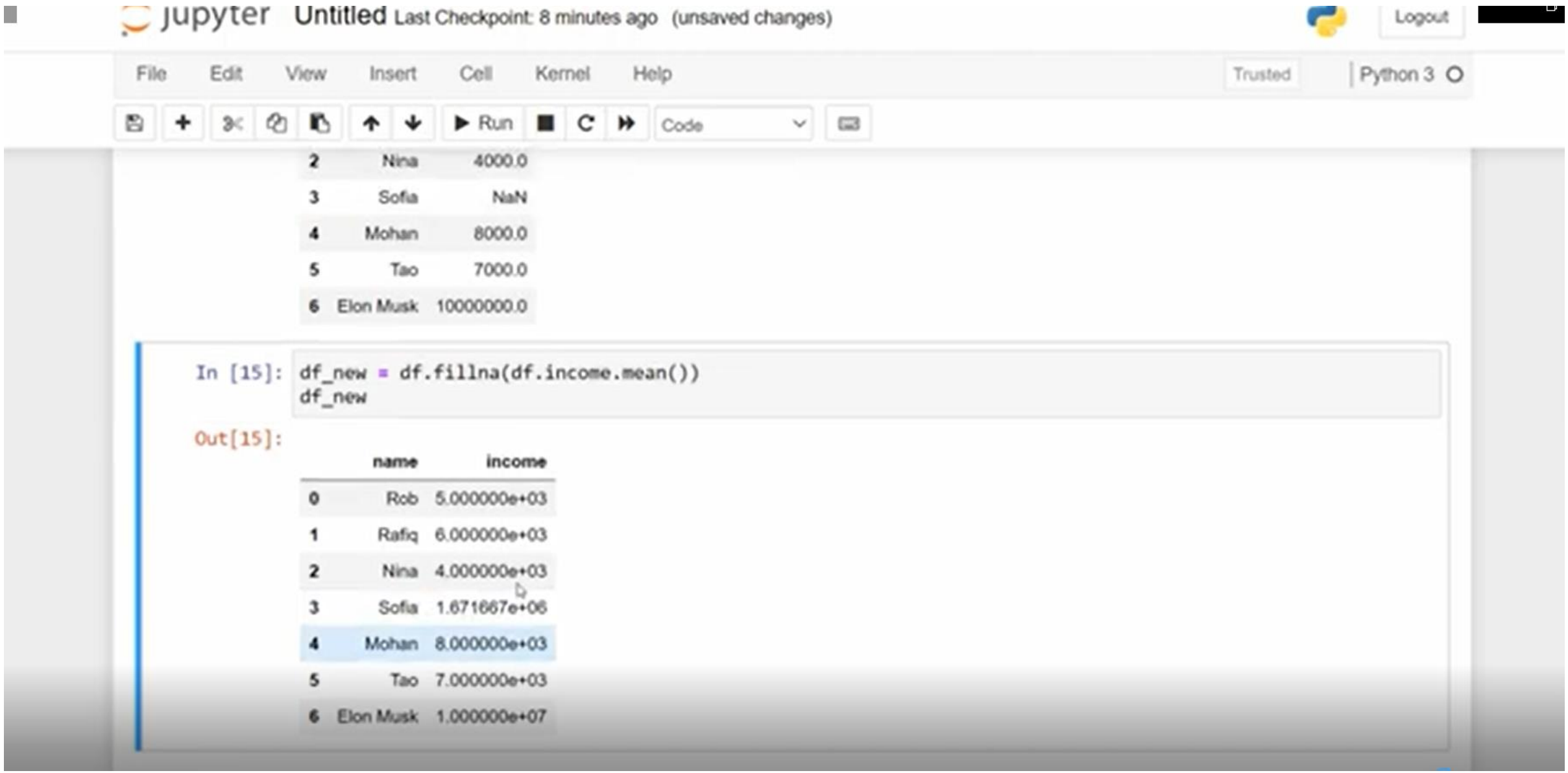
In [12]:
```
df
```

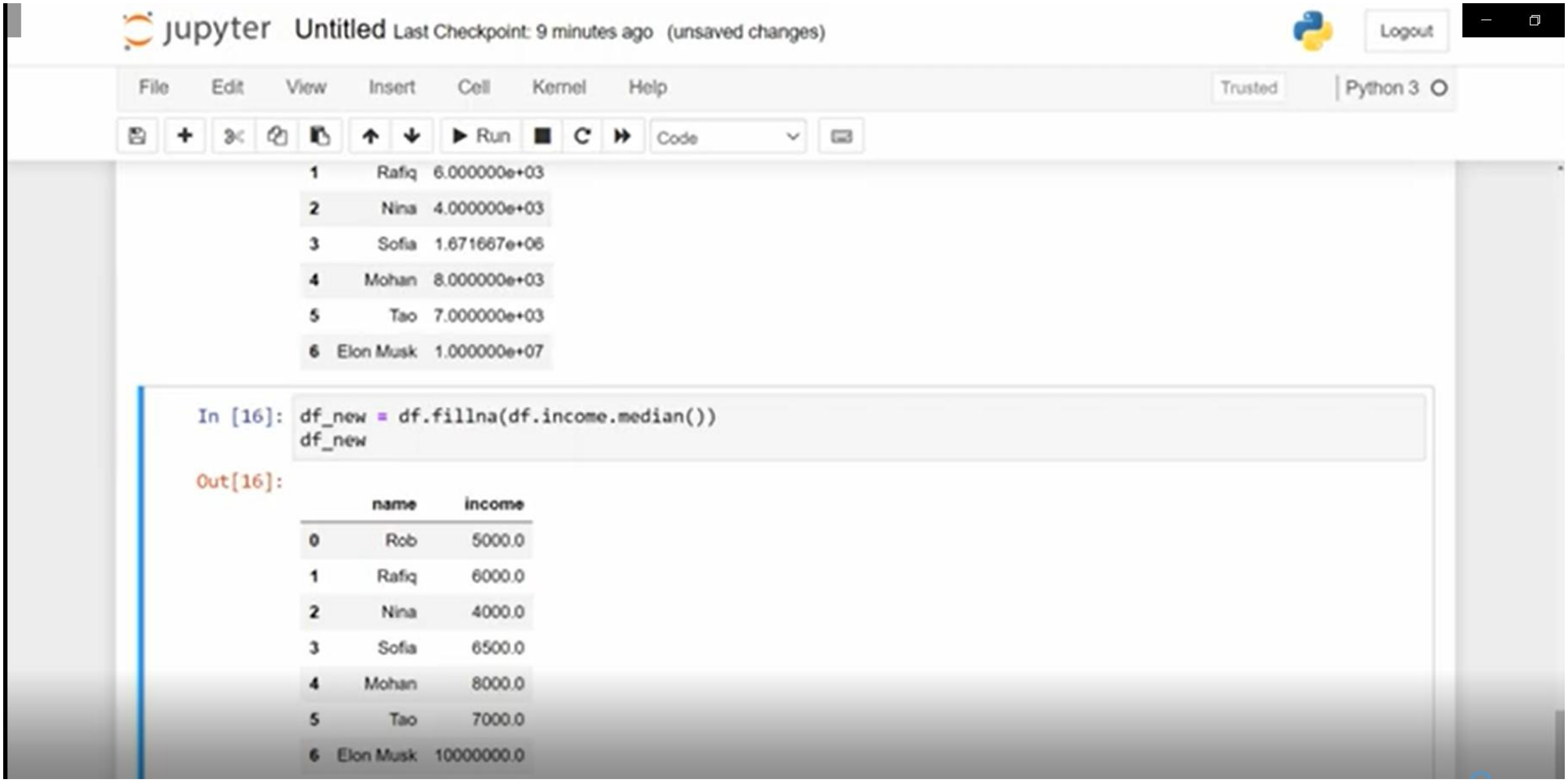
Out[12]:

	name	income
0	Rob	5000
1	Rafiq	6000
2	Nina	4000
3	Sofia	7500
4	Mohan	8000
5	Tao	7000
6	Elon Musk	10000000

In [ ]:
```
df['income'][3]=np.NaN
```


jupyter Untitled Last Checkpoint: 8 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Help
Notebook saved
Trusted
Python 3
Run
Code
A value is trying to be set on a copy of a slice from a DataFrame
See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
self._setitem_with_indexer(indexer, value)
Out[13]:
name income
0 Rob 5000.0
1 Rafiq 6000.0
2 Nina 4000.0
3 Sofia NaN
4 Mohan 8000.0
5 Tao 7000.0
6 Elon Musk 10000000.0
In [14]: df.income.mean()
Out[14]: 1671666.6666666667


jupyter Untitled Last Checkpoint: 8 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Code
2 Nina 4000.0
3 Sofia NaN
4 Mohan 8000.0
5 Tao 7000.0
6 Elon Musk 10000000.0
In [15]:
df_new = df.fillna(df.income.mean())
df_new
Out[15]:
name income
0 Rob 5.000000e+03
1 Rafiq 6.000000e+03
2 Nina 4.000000e+03
3 Sofia 1.671667e+06
4 Mohan 8.000000e+03
5 Tao 7.000000e+03
6 Elon Musk 1.000000e+07


jupyter Untitled Last Checkpoint: 9 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Code
1 Rafiq 6.000000e+03
2 Nina 4.000000e+03
3 Sofia 1.671667e+06
4 Mohan 8.000000e+03
5 Tao 7.000000e+03
6 Elon Musk 1.000000e+07
In [16]:
df_new = df.fillna(df.income.median())
df_new
Out[16]:
name income
0 Rob 5000.0
1 Rafiq 6000.0
2 Nina 4000.0
3 Sofia 6500.0
4 Mohan 8000.0
5 Tao 7000.0
6 Elon Musk 10000000.0

History Test

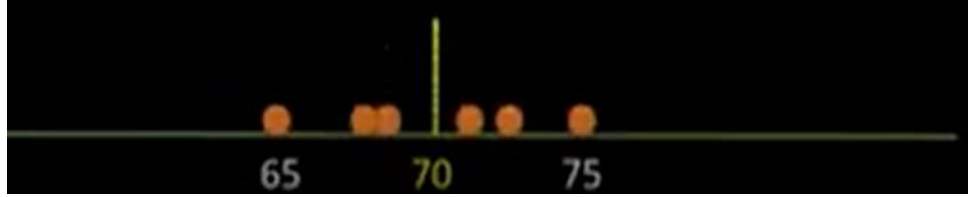
Name	Score
Mohan	75
Andrea	72
Sofia	68
Joe	65
Virat	67
Abdul	73

Average = 70

History Test

Name	Score
Mohan	75
Andrea	72
Sofia	68
Joe	65
Virat	67
Abdul	73

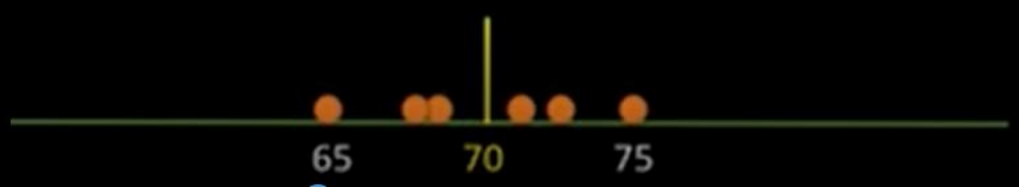
Average = 70



History Test

Name	Score
Mohan	75
Andrea	72
Sofia	68
Joe	65
Virat	67
Abdul	73

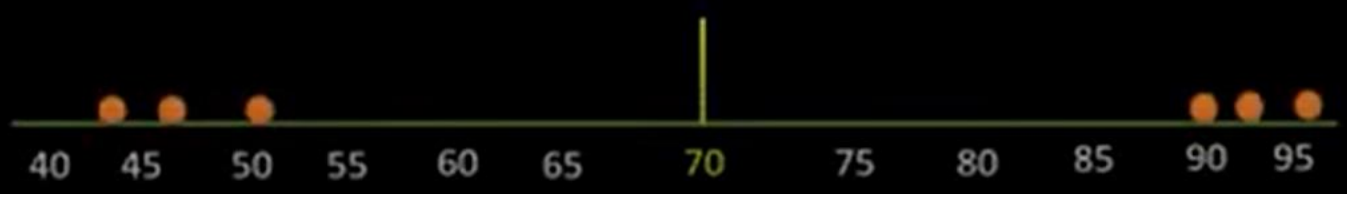
Average = 70



Math Test

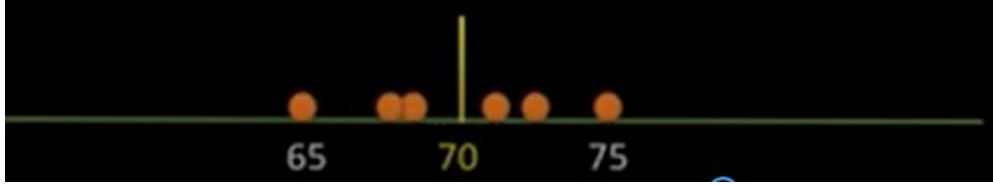
Name	Score
Mohan	93
Andrea	96
Sofia	43
Joe	47
Virat	51
Abdul	90

Average = 70



History Test

Name	Score	Abs (Score – Avg)
Mohan	75	5
Andrea	72	2
Sofia	68	2
Joe	65	5
Virat	67	3
Abdul	73	2
	Mean	**3.16**

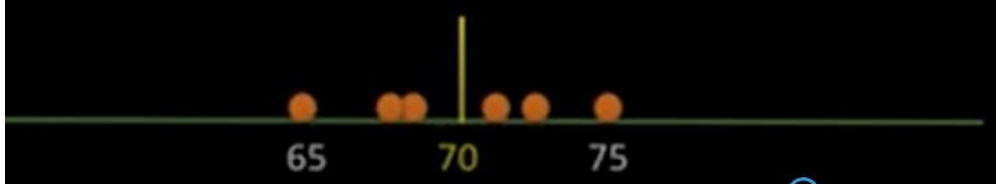


Math Test

Name	Score	Abs (Score – Avg)
Mohan	93	23
Andrea	96	26
Sofia	43	27
Joe	47	23
Virat	51	19
Abdul	90	20

History Test

Name	Score	Abs (Score – Avg)
Mohan	75	5
Andrea	72	2
Sofia	68	2
Joe	65	5
Virat	67	3
Abdul	73	2
	Mean	**3.16**



Math Test

Name	Score	Abs (Score – Avg)
Mohan	93	23
Andrea	96	26
Sofia	43	27
Joe	47	23
Virat	51	19
Abdul	90	20
	Mean	**23**

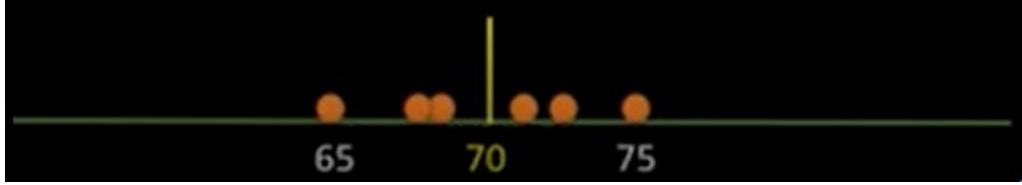
History Test

Name	Score	Abs (Score – Avg)
Mohan	75	5
Andrea	72	2
Sofia	68	2
Joe	65	5
Virat	67	3
Abdul	73	2
	MAD	3.16

Math Test

Name	Score	Abs (Score – Avg)
Mohan	93	23
Andrea	96	26
Sofia	43	27
Joe	47	23
Virat	51	19
Abdul	90	20
	MAD	23

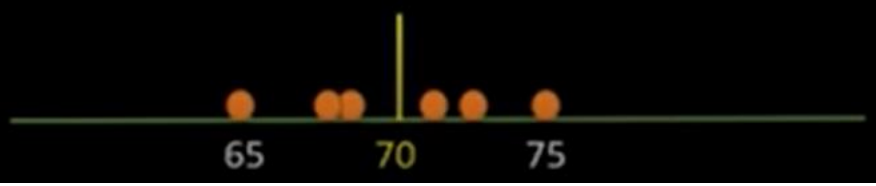
Mean Absolute Deviation

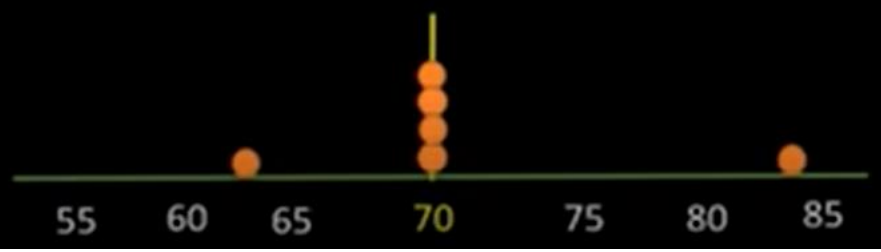


Name	Score	Abs (Score – Avg)
Mohan	75	5
Andrea	72	2
Sofia	68	2
Joe	65	5
Virat	67	3
Abdul	73	3
	MAD	3·33

History Test

Name	Score	Abs (Score – Avg)
Mohan	83	13
Andrea	70	0
Sofia	70	0
Joe	63	7
Virat	70	0
Abdul	70	0
	MAD	3·33





Average = 70

Name	Score	Abs (Score – Avg)	Abs (Score – Avg)^2
Mohan	75	5	25
Andrea	72	2	4
Sofia	68	2	4
Joe	65	5	25
Virat	67	3	9
Abdul	73	3	9
		Avg	12.66
		$\sqrt{Avg}$	3.55

Average = 70

Name	Score	Abs (Score – Avg)	Abs (Score – Avg)^2
Mohan	83	13	169
Andrea	70	0	0
Sofia	70	0	0
Joe	63	7	49
Virat	70	0	0
Abdul	70	0	0
		Avg	36.33
		$\sqrt{Avg}$	6.02

Formula

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

σ = population standard deviation

N = the size of the population

x_i = each value from the population

μ = the population mean

From the web

To find the **standard deviation**, we take the square root of the variance. From learning that **SD** = 13.31, we can say that each score deviates from the mean by 13.31 points on average. Sep 17, 2020

https://www.scribbr.com › statistics › standard-deviation

Standard Deviation | A Step by Step Guide with Formulas

9

Both measure the dispersion of your data by computing the distance of the data to its mean.

1. the **mean absolute deviation** is using norm L1 (it is also called Manhattan distance or rectilinear distance)
2. the **standard deviation** is using norm L2 (also called Euclidean distance)

The difference between the two norms is that the **standard deviation** is calculating the square of the difference whereas the **mean absolute deviation** is only looking at the absolute difference. Hence large outliers will create a higher dispersion when using the standard deviation instead of the other method. The Euclidean distance is indeed also more often used. The main reason is that the **standard deviation** have nice properties when the data is normally distributed. So under this assumption, it is recommended to use it. However people often do this assumption for data which is actually not normally distributed which creates issues. If your data is not normally distributed, you can still use the standard deviation, but you should be careful with the interpretation of the results.

Finally you should know that both measures of dispersion are particular cases of the Minkowski distance, for p=1 and p=2. You can increase p to get other measures of the dispersion of your data.

Share Cite Improve this answer Follow

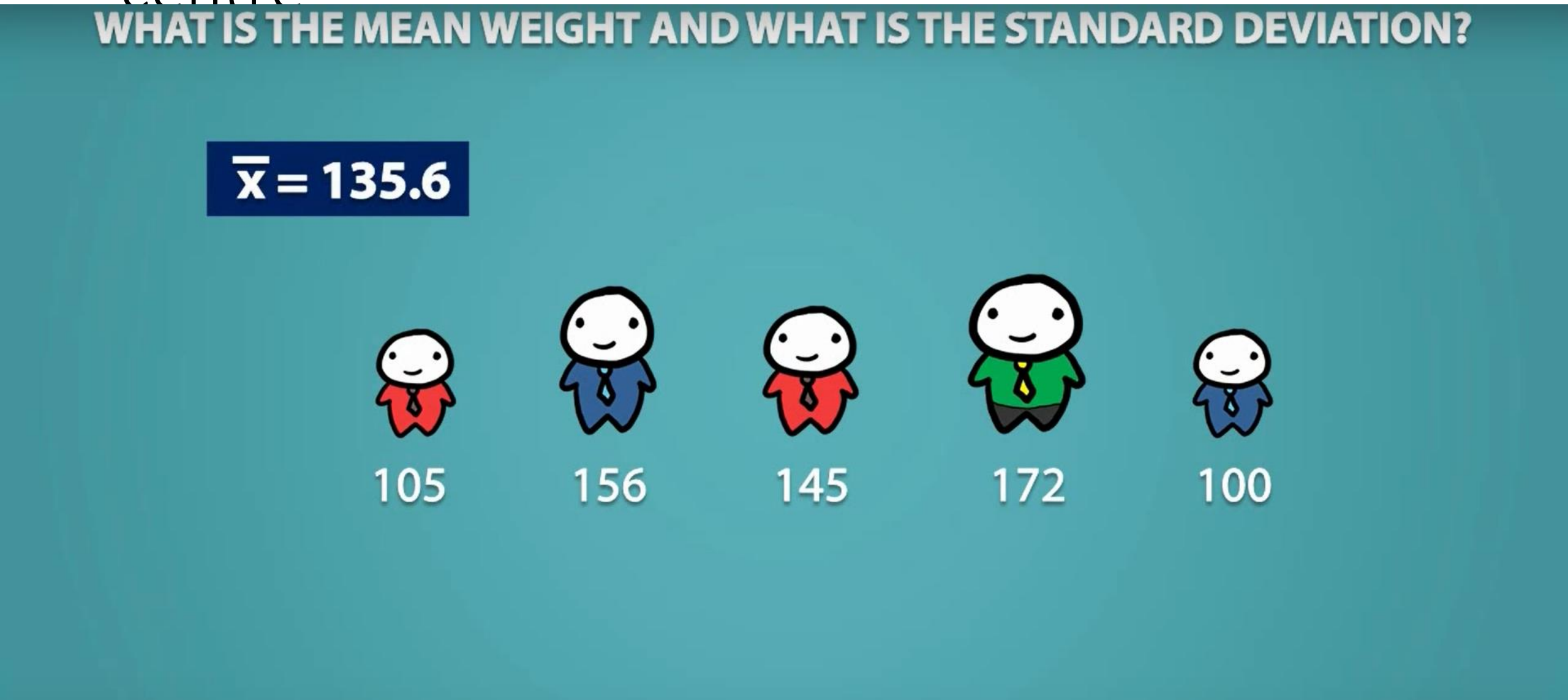
edited Mar 5 '14 at 3:04

answered Mar 5 '14 at 2:51

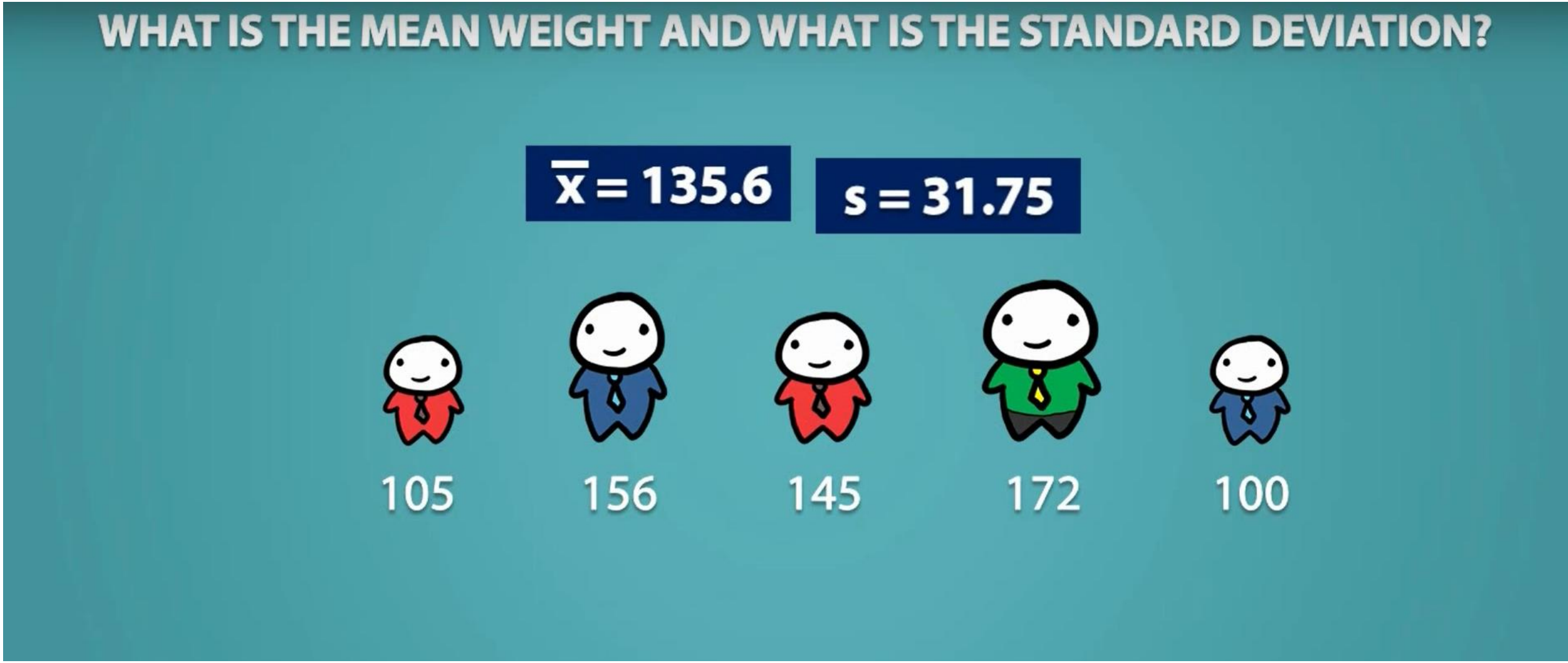
RockScience

2,613 4 26 44

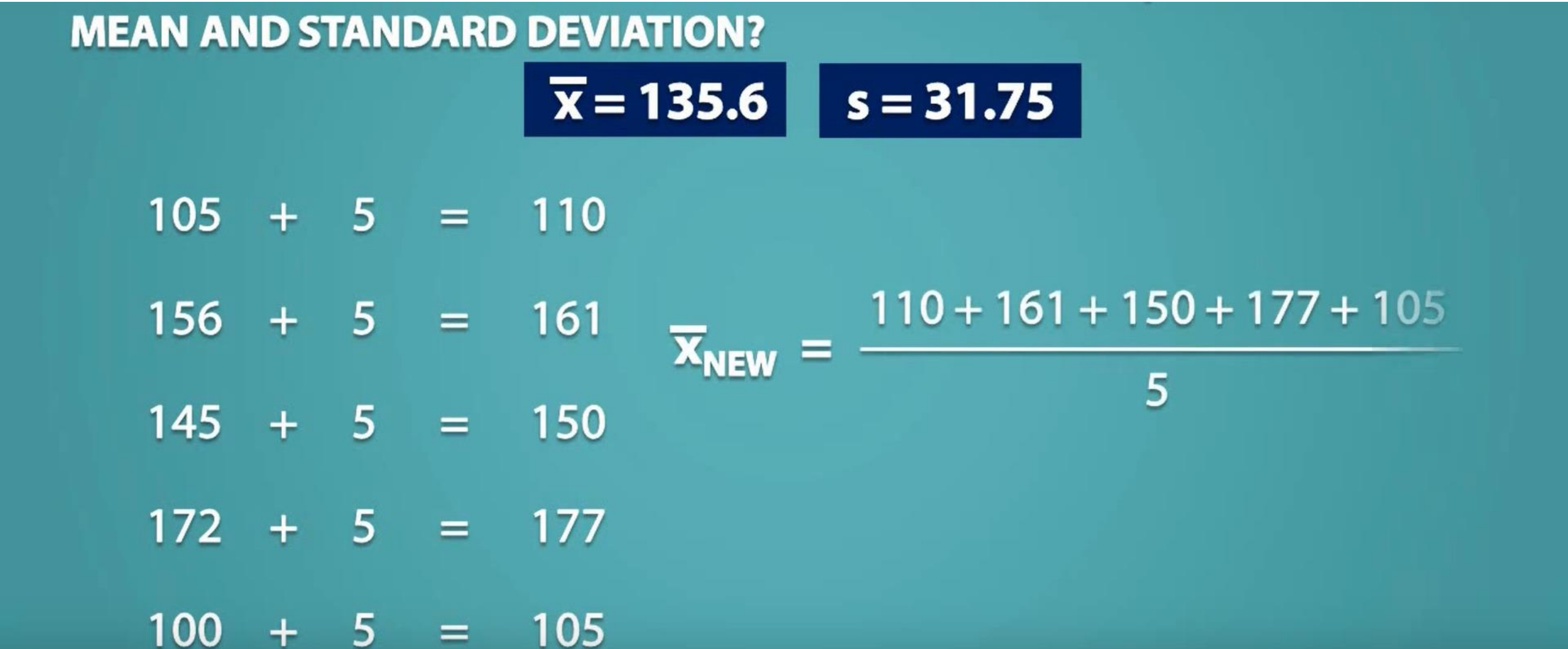
Effect of transforming data on spread and centre



Effect of transforming data on spread and centre



Suppose during winter each person wears 5 pound cloths then mean and standard deviation is-



MEAN AND STANDARD DEVIATION?

$\bar{x} = 135.6$ $s = 31.75$

105

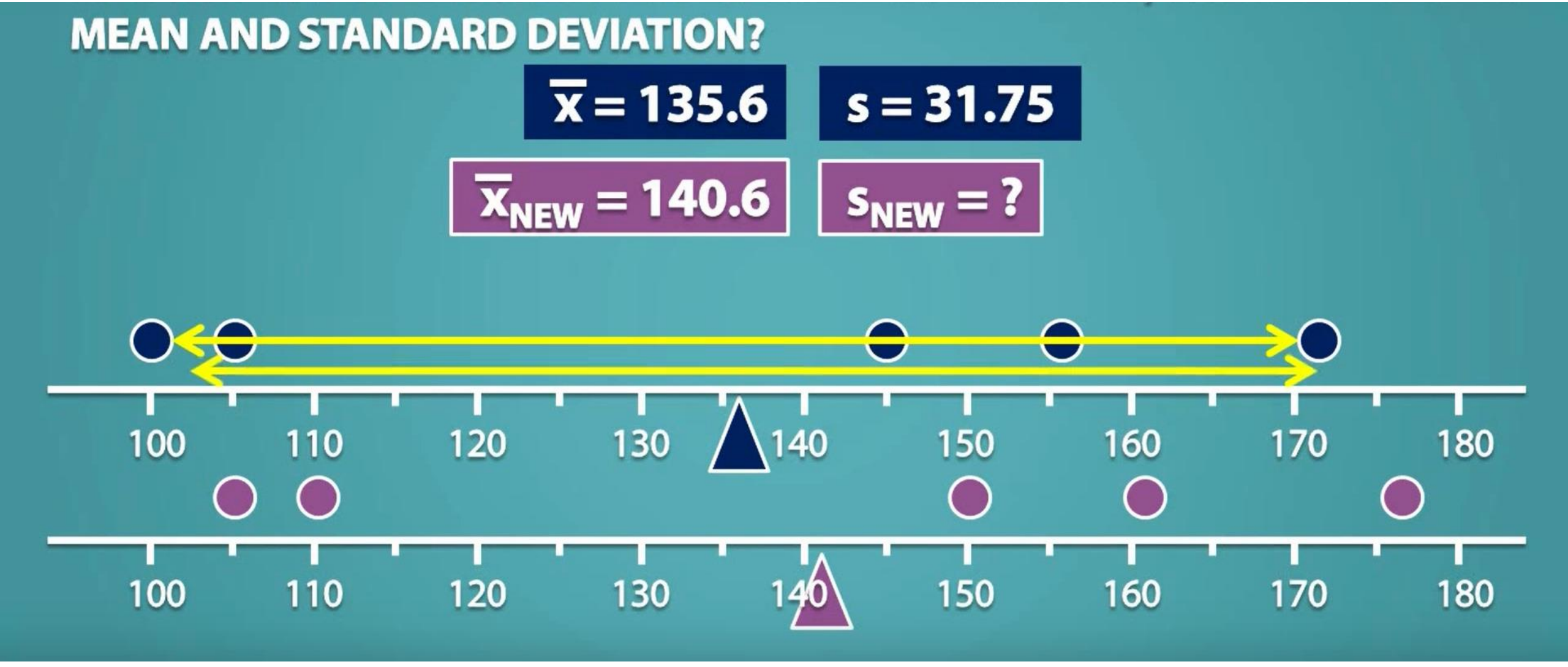
156

145

172

100

$$\bar{x}_{NEW} = 135.6 + 5 = 140.6$$


MEAN AND STANDARD DEVIATION?
$\bar{x} = 135.6$
$s = 31.75$
$\bar{x}_{NEW} = 140.6$
$s_{NEW} = ?$
100
110
120
130
140
150
160
170
180
100
110
120
130
140
150
160
170
180

MEAN AND STANDARD DEVIATION?

$\bar{x} = 135.6$ $s = 31.75$

$\bar{x}_{NEW} = 140.6$ $s_{NEW} = 31.75$

TRANSFORMING DATA

GUIDELINES

MEASURES OF CENTRE

AFFECTED BY:

MODE, MEDIAN, MEAN

MEASURES OF SPREAD

AFFECTED BY:

RANGE, STANDARD DEVIATION

Suppose each student drink

2.5mL OF WATER FOR EVERY POUND THEY WEIGH; PLUS 750mL OF WATER A DAY. WHAT IS THE MEAN AND STANDARD DEVIATION FOR THE AMOUNT OF WATER CONSUMED EVERY DAY?

105 156 145 172 100

2.5mL OF WATER FOR EVERY POUND THEY WEIGH; PLUS 750mL OF WATER A DAY. WHAT IS THE MEAN AND STANDARD DEVIATION FOR THE AMOUNT OF WATER CONSUMED EVERY DAY?

105

156

145

172

100

$\bar{x} = 135.6$

$s = 31.75$

2.5mL OF WATER FOR EVERY POUND THEY WEIGH; PLUS 750mL OF WATER A DAY. WHAT IS THE MEAN AND STANDARD DEVIATION FOR THE AMOUNT OF WATER CONSUMED EVERY DAY?

105 $\bar{x} = 135.6$ x 2.5 + 750

156 $\bar{x}_{NEW} = (135.6)(2.5) + 750 = 1089$

145 $s = 31.75$ x 2.5

172 $s_{NEW} = (31.75)(2.5) = 79.38$

100

$\bar{x} = 135.6$ x 2.5 + 750

MEASURES OF CENTRE

$CENTRE_{NEW} = (CENTRE_{OLD})(X) + B$

$\bar{x}_{NEW} = (135.6)(2.5) + 750 = 1089$

$s = 31.75$ x 2.5

$s_{NEW} = (31.75)(2.5) = 79.38$

MEASURES OF CENTRE	$\text{CENTRE}_{\text{NEW}} = (\text{CENTRE}_{\text{OLD}})(X) + B$
MEASURES OF SPREAD	$\text{SPREAD}_{\text{NEW}} = (\text{SPREAD}_{\text{OLD}})(X)$

OUTLIER

A DATA VALUE THAT IS NUMERICALLY DISTANT FROM A DATA SET

OUTLIER

A DATA VALUE THAT IS NUMERICALLY DISTANT FROM A DATA SET

EXAMPLES


FREQUENCY
8
7
6
5
4
3
2
1
5
10
15
20
25
30
35
40
45
50
55
60
RANDOM VARIABLE


OUTLIER
21
20
22
25
21
9000
23

TEMPERATURE OF WINNIPEG ON JULY 1ST

YEAR	TEMPERATURE
2015	26.0 ℃
2014	15.0 ℃
2013	20.5 ℃
2012	31.0 ℃
2011	-350.0 ℃ OUTLIER
2010	31.0 ℃
2009	30.5 ℃

$\overline{x} = -28\,℃$

26.0 ℃

15.0 ℃

20.5 ℃

31.0 ℃

-350.0 ℃ OUTLIER

31.0 ℃

30.5 ℃

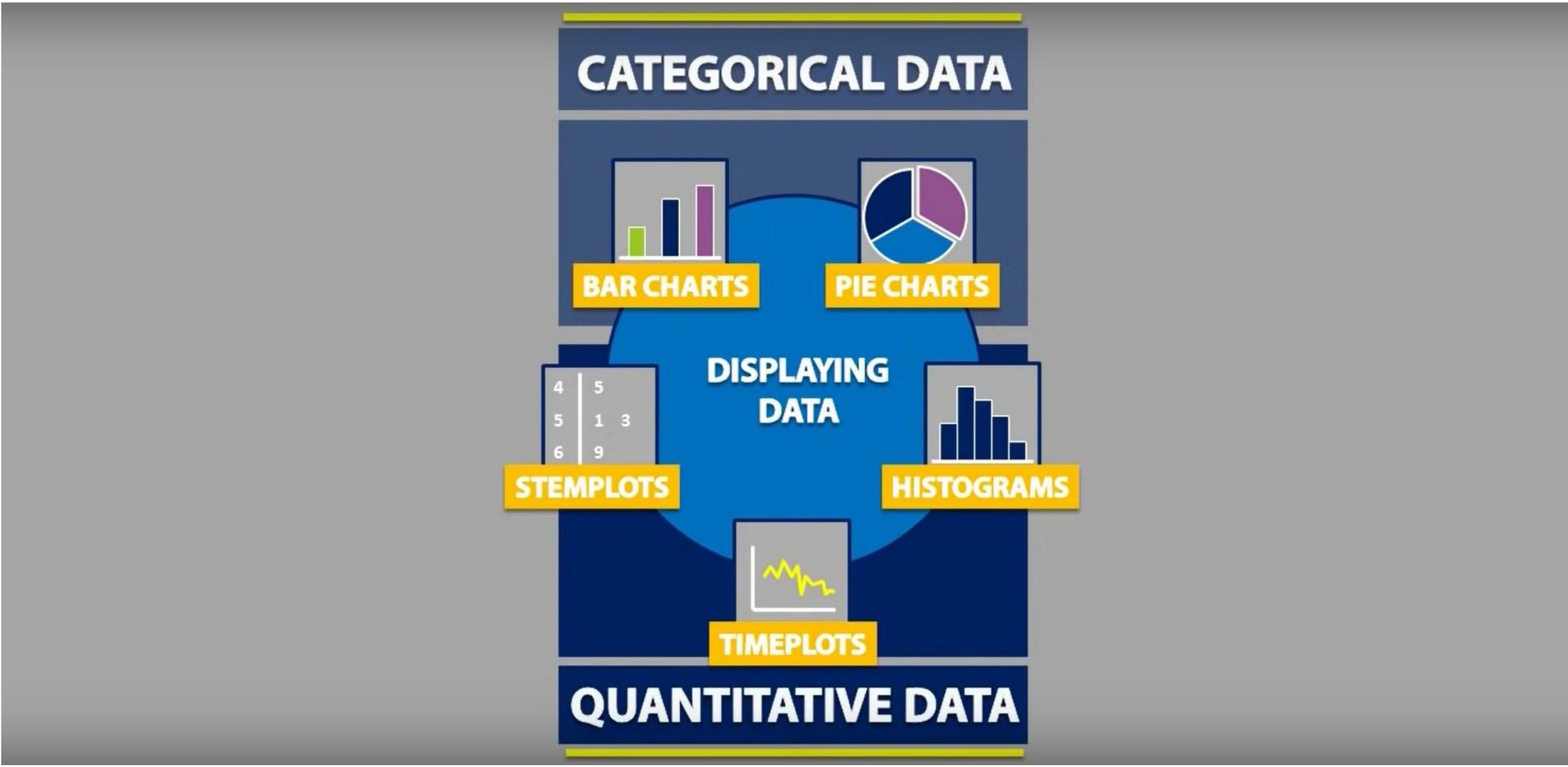
DATA SET	
OUTLIER	-350.0 ℃
	15.0 ℃
	20.5 ℃
	26.0 ℃
	30.5 ℃
	31.0 ℃
	31.0 ℃

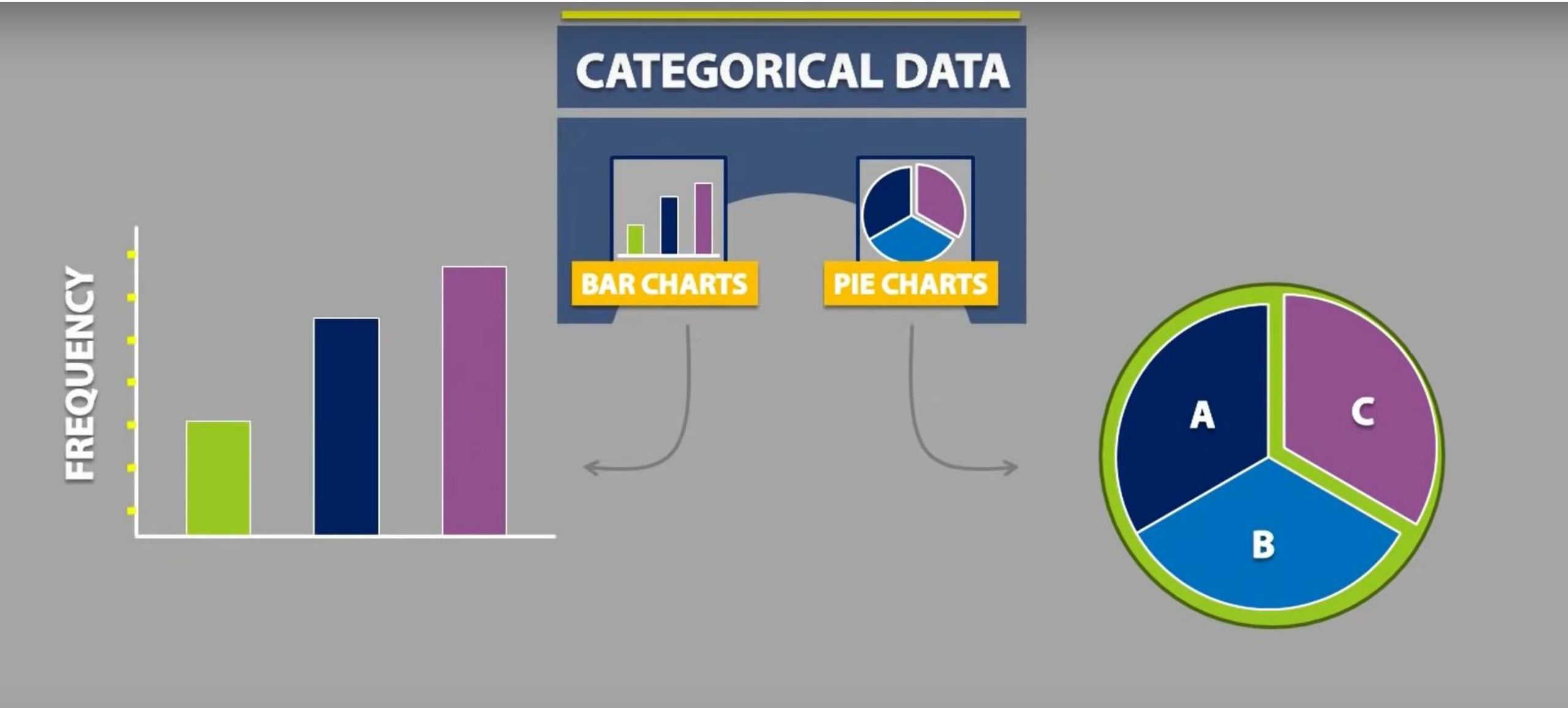
CALCULATIONS

MEASURE	WITH OUTLIER	WITHOUT OUTLIER
MEAN	– 28	25.667
MEDIAN	26	28.25
MODE	31	31
RANGE	381	16

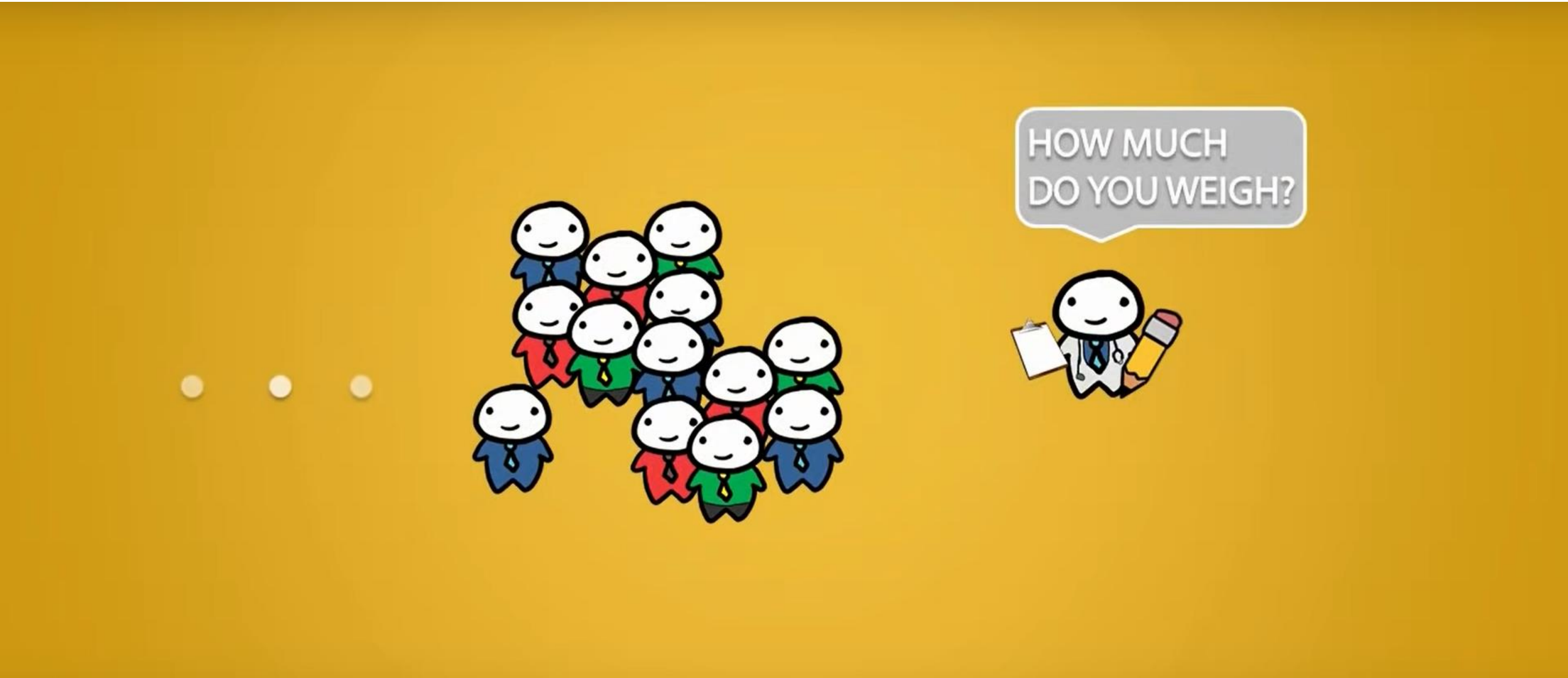
RESPONSE TO AN OUTLIER	MEASURE	WITH OUTLIER	WITHOUT OUTLIER
AFFECTED	MEAN	– 28	25.667
RESISTANT	MEDIAN	26	28.25
RESISTANT	MODE	31	31
AFFECTED	RANGE	381 ⟷	16

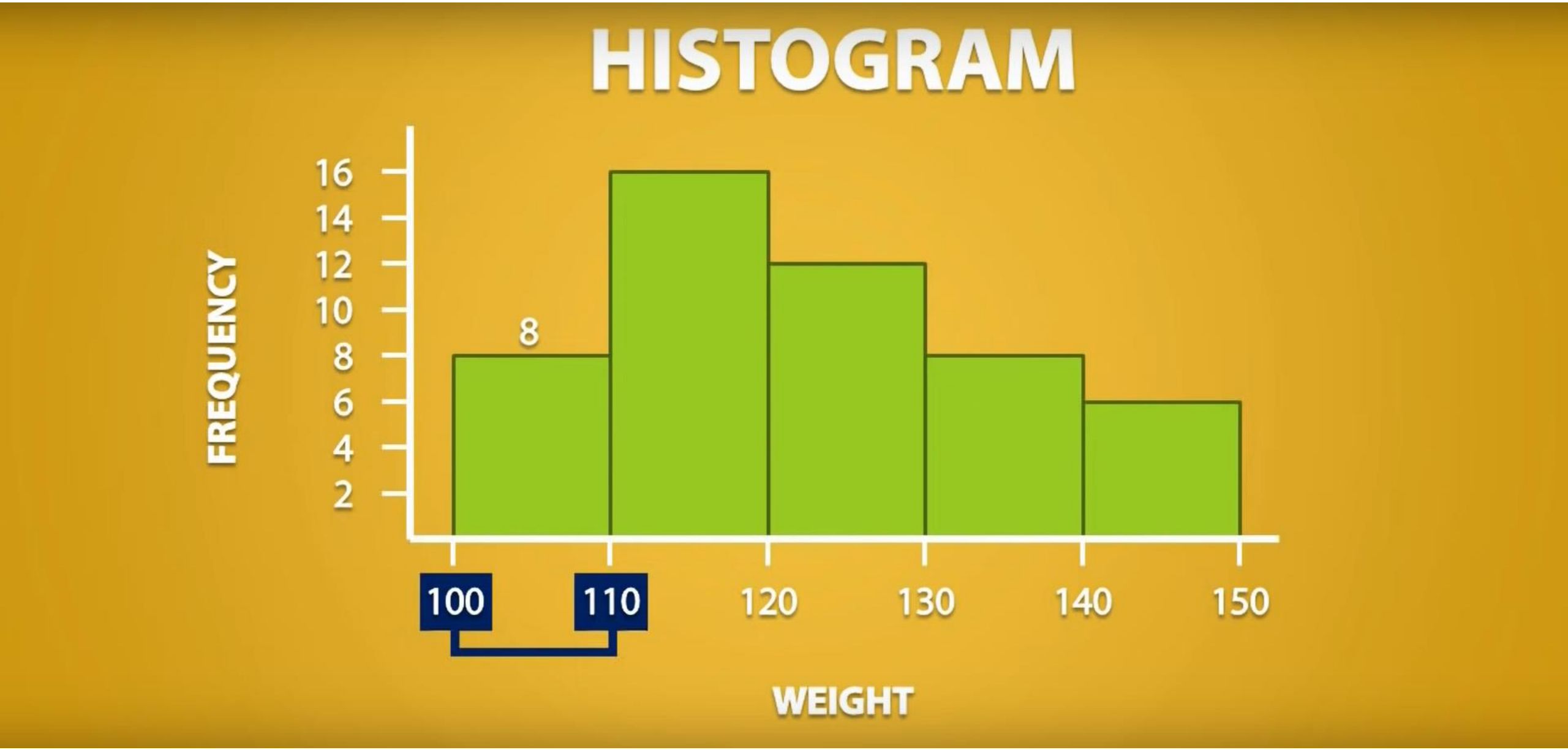
R = MAXIMUM – MINIMUM

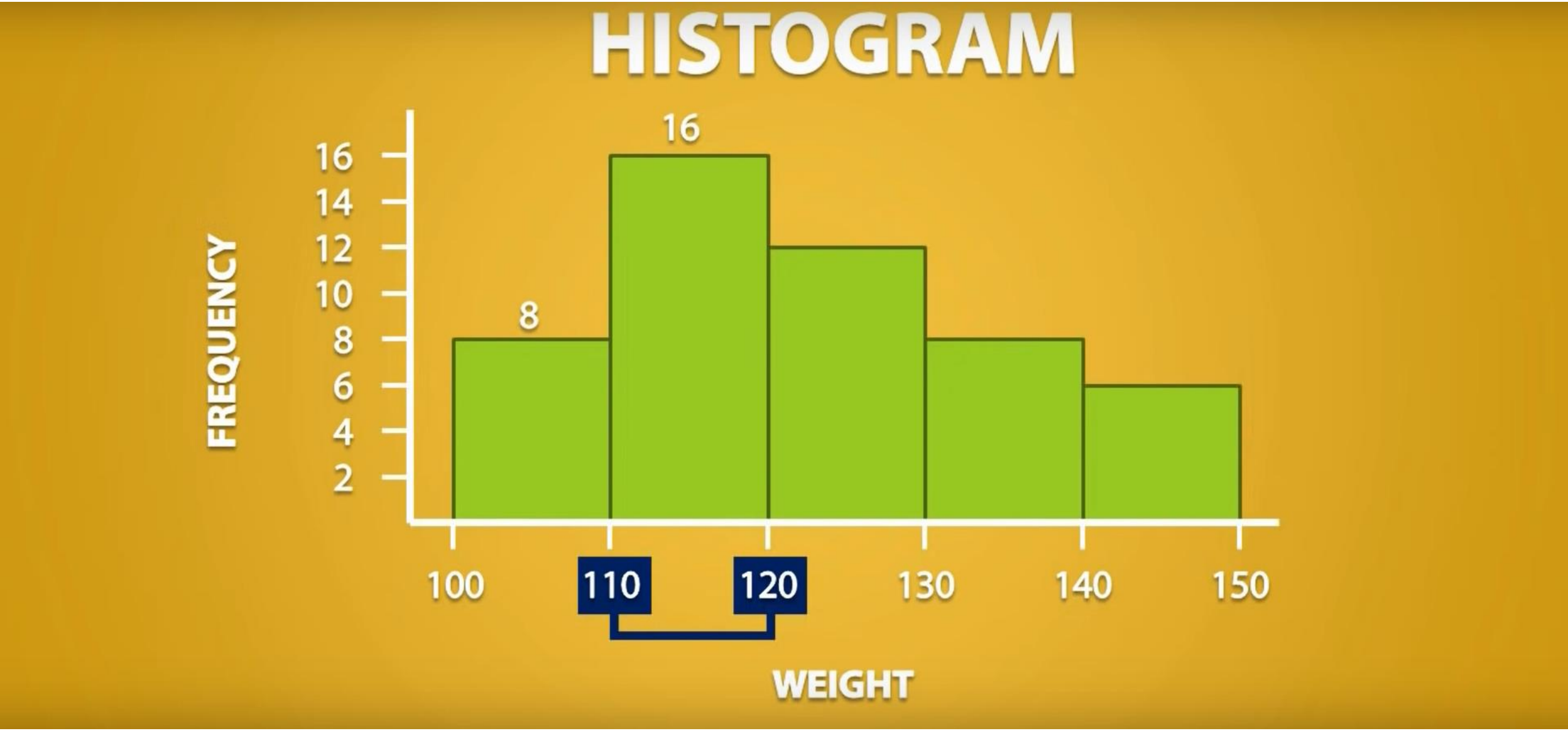

CATEGORICAL DATA
BAR CHARTS
PIE CHARTS
DISPLAYING DATA
4 5
5 1 3
6 9
STEMPLOTS
HISTOGRAMS
TIMEPLOTS
QUANTITATIVE DATA


CATEGORICAL DATA
BAR CHARTS
PIE CHARTS
FREQUENCY
A
B
C

Quantative data




HISTOGRAM
FREQUENCY
16
14
12
10
8
6
4
2
8
100
110
120
130
140
150
WEIGHT


HISTOGRAM
FREQUENCY
16
14
12
10
8
6
4
2
16
8
100
110
120
130
140
150
WEIGHT

FREQUENCY DISTRIBUTION

WEIGHT		FREQUENCY
100 – 110	←	8
110 – 120	←	16
120 – 130	←	12
130 – 140	←	8
140 – 150	←	6

Now suppose we have 120 data point?where should we include?

FREQUENCY DISTRIBUTION

WEIGHT	FREQUENCY
100 – 110	8
110 – 120	16
120 – 130	12
130 – 140	8
140 – 150	6

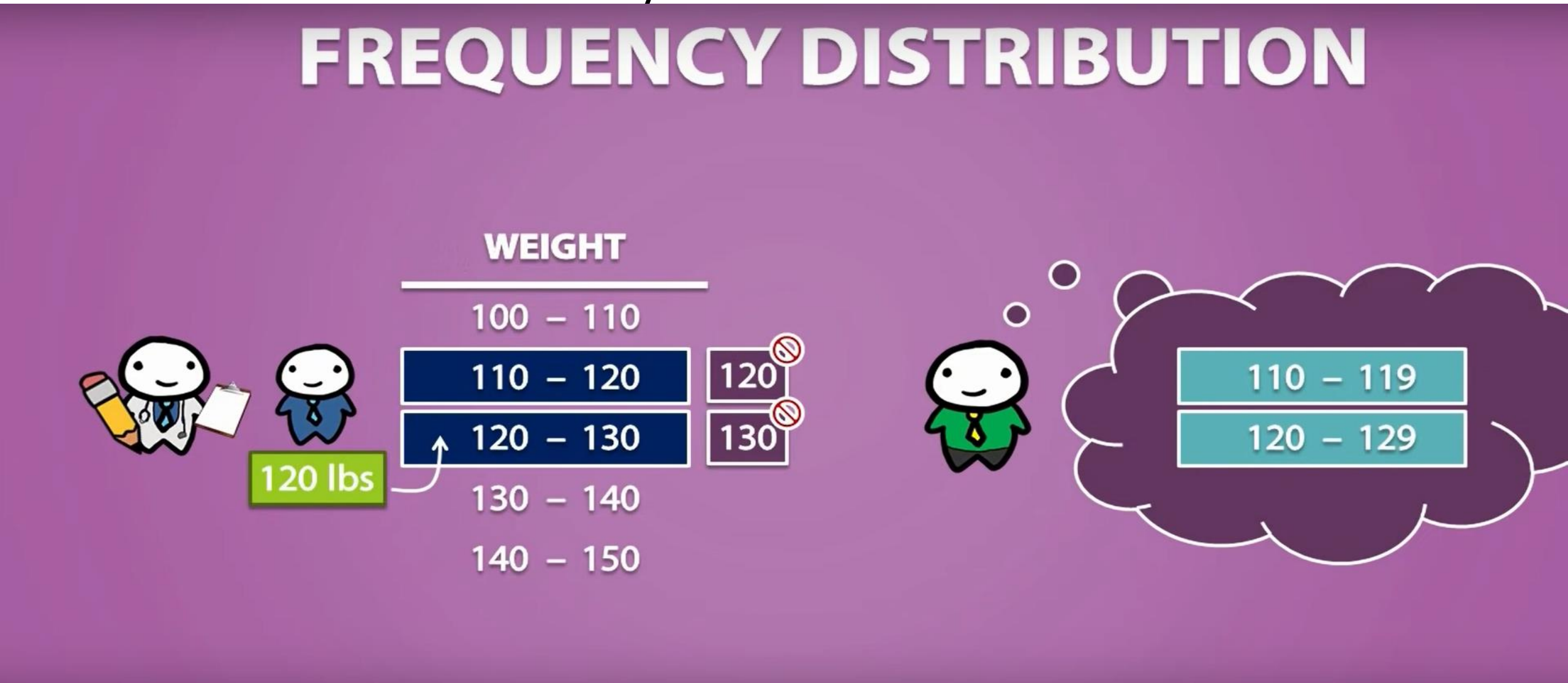
FREQUENCY DISTRIBUTION

WEIGHT
100 – 110
110 – 120
120 – 130
130 – 140
140 – 150

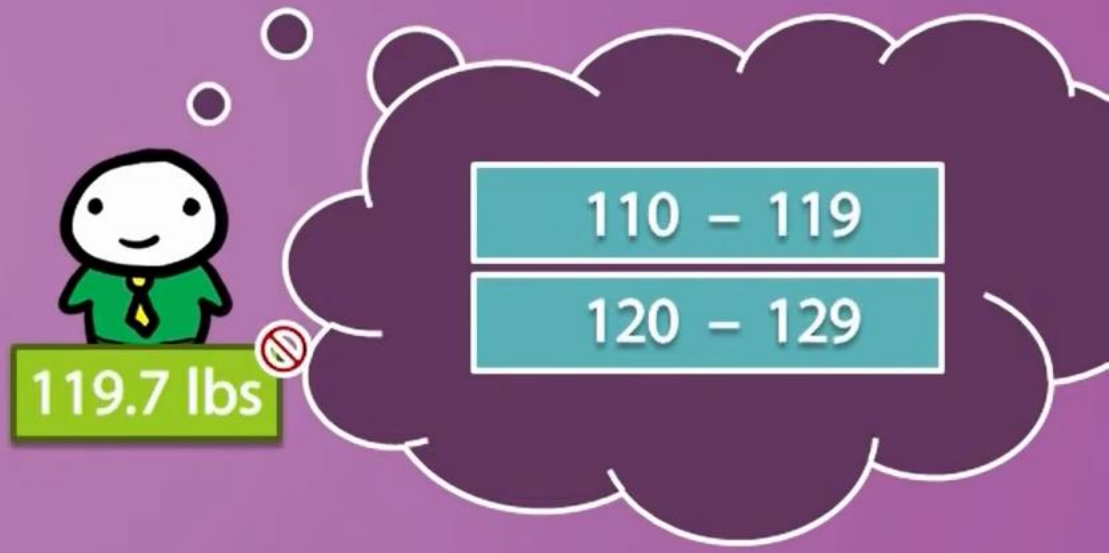
120

BY CONVENTION, WE SAY THAT EACH INTERVAL DOES NOT INCLUDE THE RIGHT END POINT

Can we have range 110-119…but there will not data continuity



FREQUENCY DISTRIBUTION



Frequency distribution can be converted to relative frequency distribution

FREQUENCY DISTRIBUTION

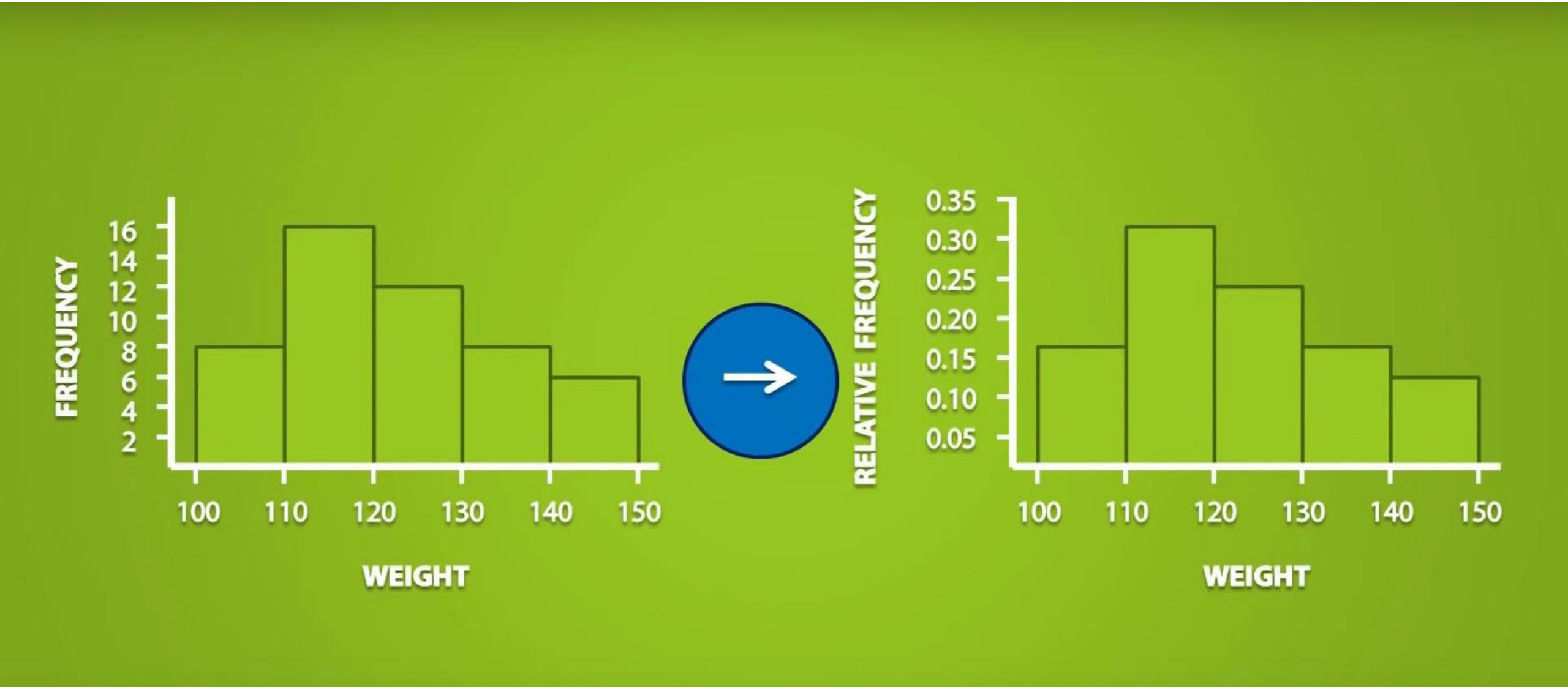
WEIGHT	FREQUENCY
100 – 110	8
110 – 120	16
120 – 130	12
130 – 140	8
140 – 150	6

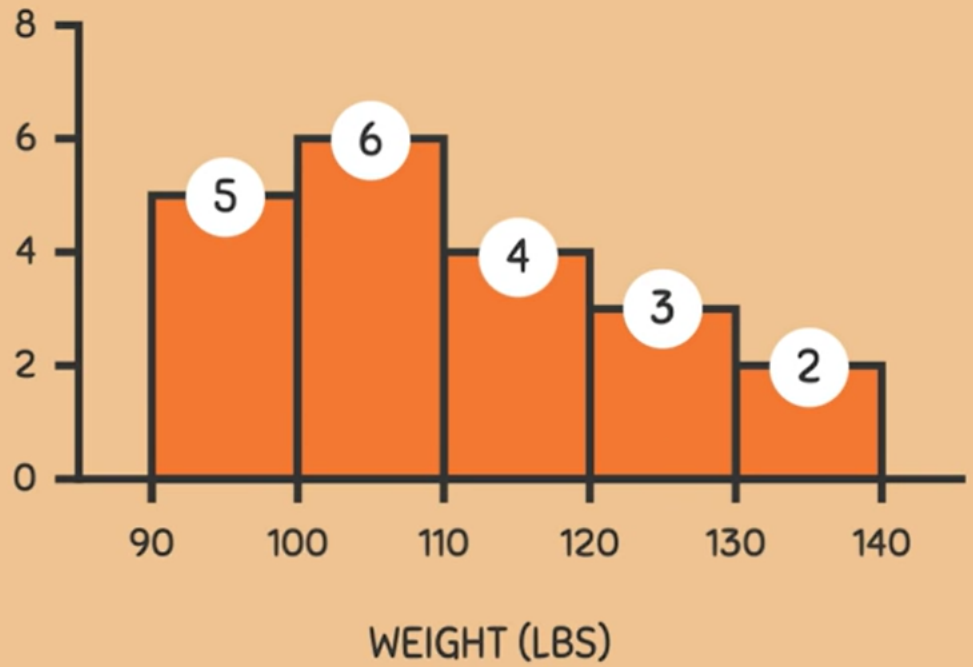
RELATIVE FREQUENCY DISTRIBUTION

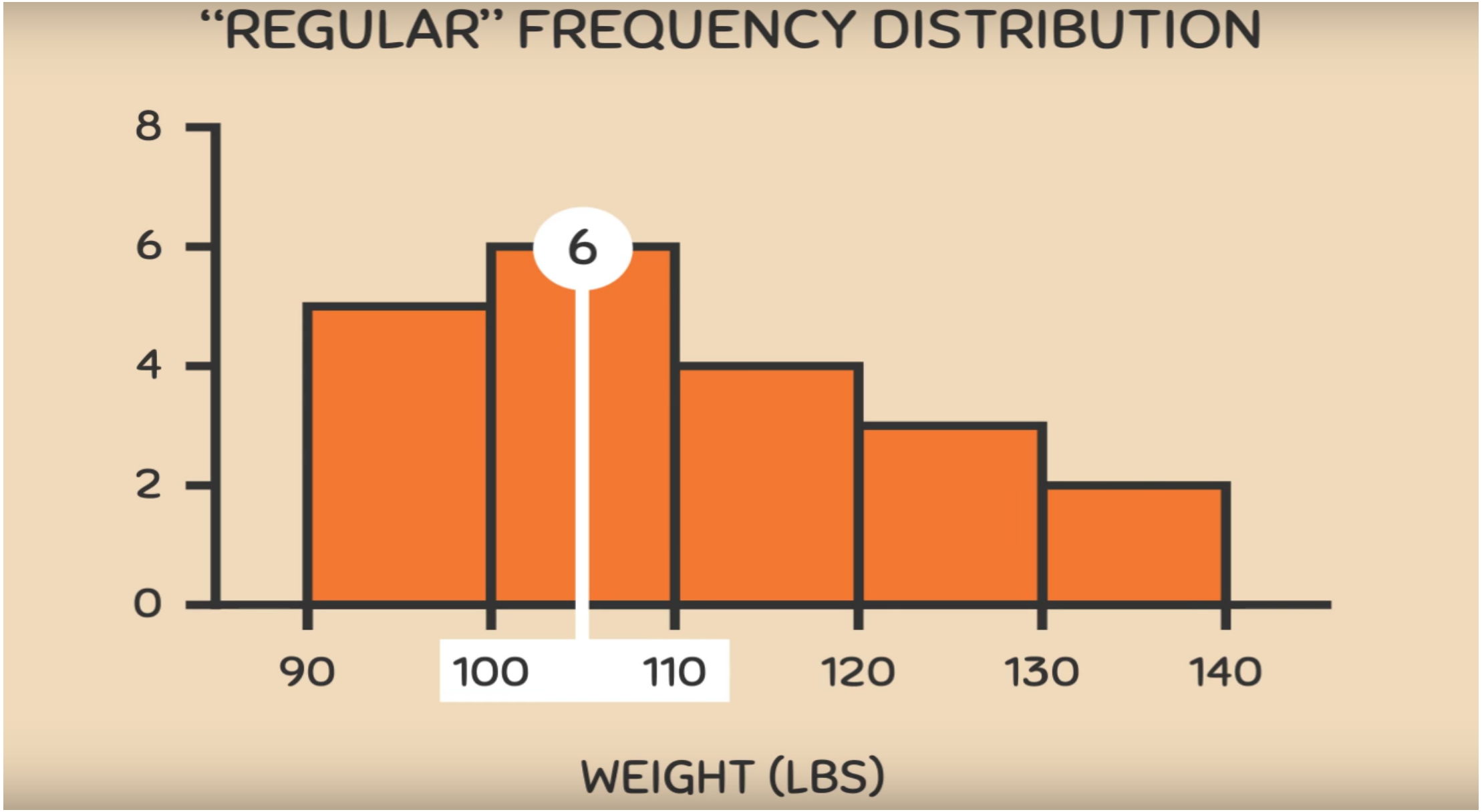
WEIGHT	RELATIVE FREQUENCY
100 – 110	0.16
110 – 120	0.32
120 – 130	0.24
130 – 140	0.16
140 – 150	0.12

WEIGHT	FREQUENCY	CALCULATIONS	RELATIVE FREQUENCY
100 – 110	8	8 ÷ 50 =	0.16
110 – 120	16	16 ÷ 50 =	0.32
120 – 130	12	12 ÷ 50 =	0.24
130 – 140	8	8 ÷ 50 =	0.16
140 – 150	+ 6	6 ÷ 50 =	+ 0.12
	SUM = 50		SUM = 1

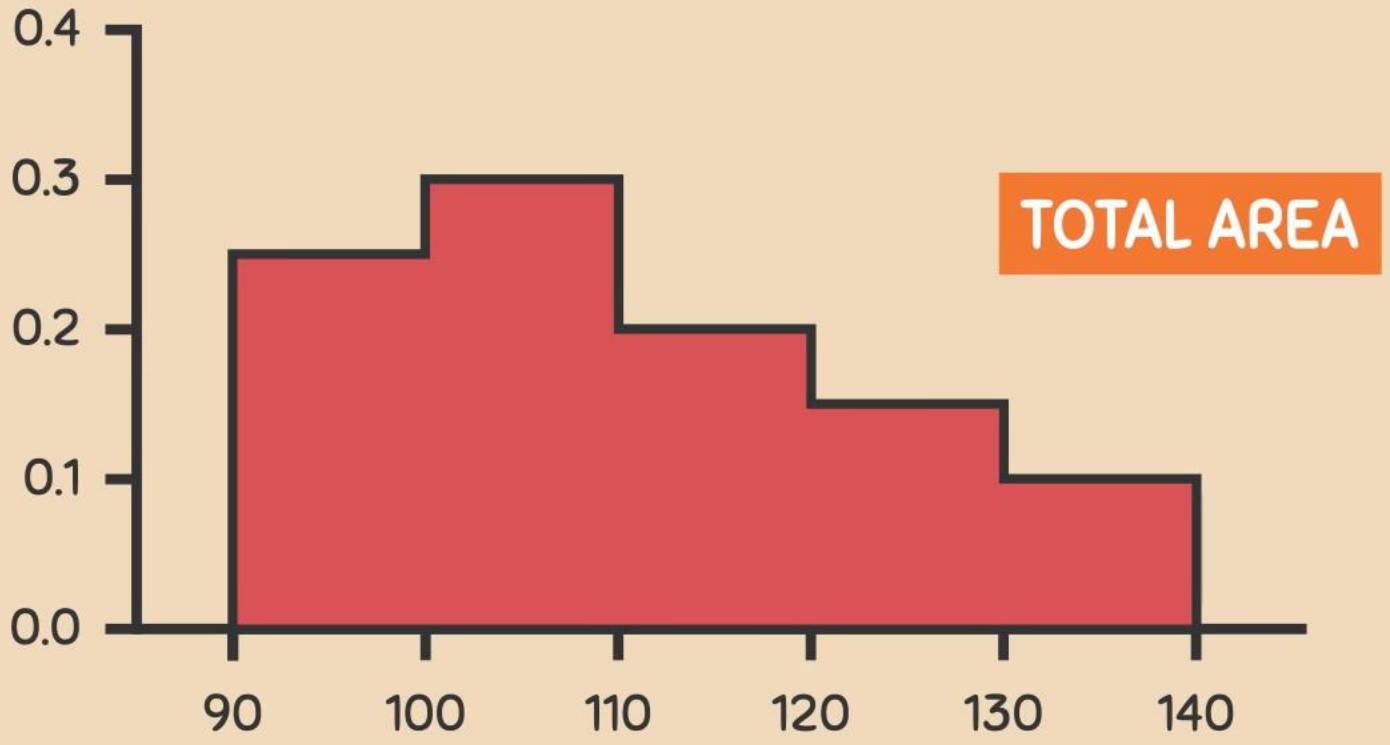
WEIGHT	FREQUENCY	CALCULATIONS	RELATIVE FREQUENCY
100 – 110	8	8 ÷ 50 =	16%
110 – 120	16	16 ÷ 50 =	32%
120 – 130	12	12 ÷ 50 =	24%
130 – 140	8	8 ÷ 50 =	16%
140 – 150	+ 6	6 ÷ 50 =	+ 12%
	SUM = 50		SUM = 100%


FREQUENCY
16
14
12
10
8
6
4
2
100
110
120
130
140
150
WEIGHT
RELATIVE FREQUENCY
0.35
0.30
0.25
0.20
0.15
0.10
0.05
100
110
120
130
140
150
WEIGHT

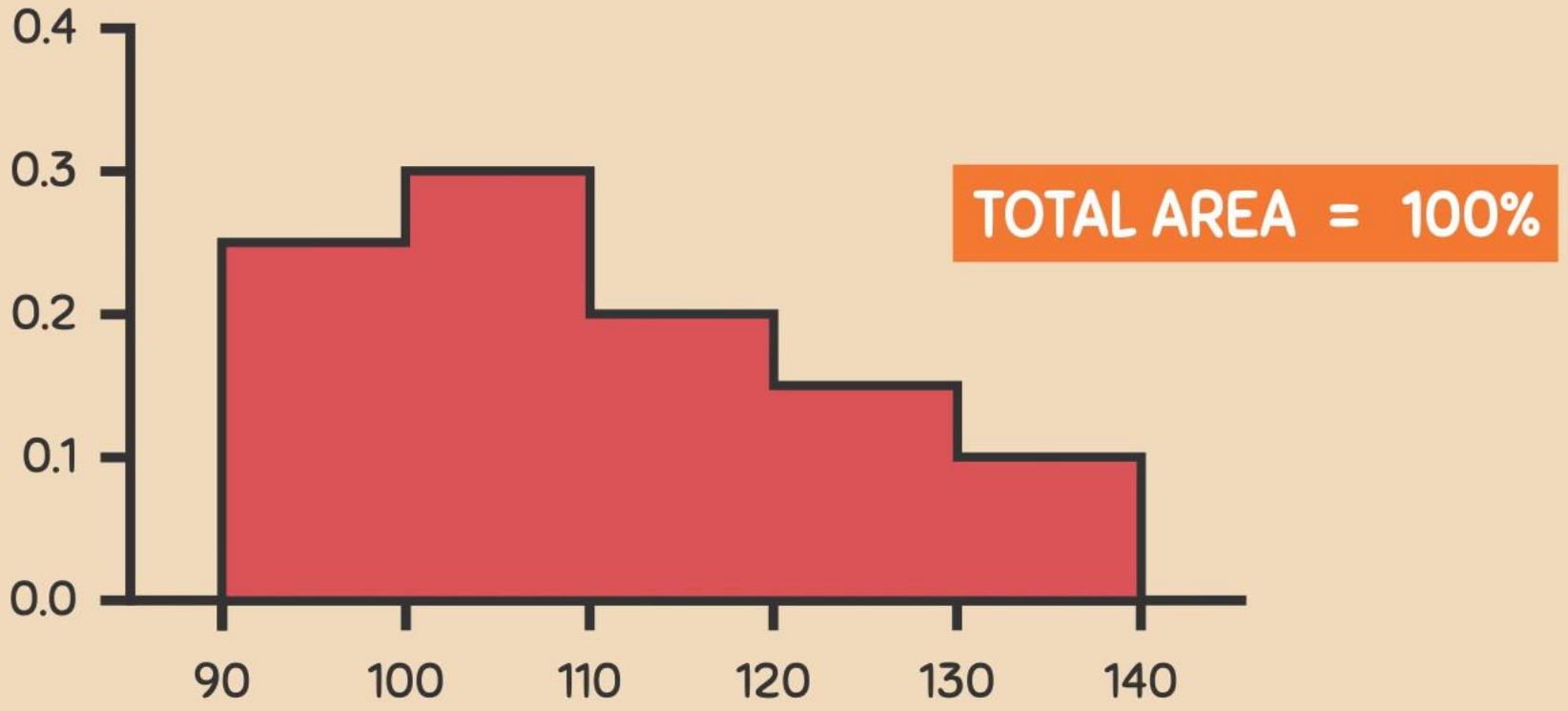

"REGULAR" FREQUENCY DISTRIBUTION
8
6
4
2
0
5
6
4
3
2
90
100
110
120
130
140
WEIGHT (LBS)


"REGULAR" FREQUENCY DISTRIBUTION
8
6
4
2
0
6
90
100
110
120
130
140
WEIGHT (LBS)

RELATIVE FREQUENCY DISTRIBUTION

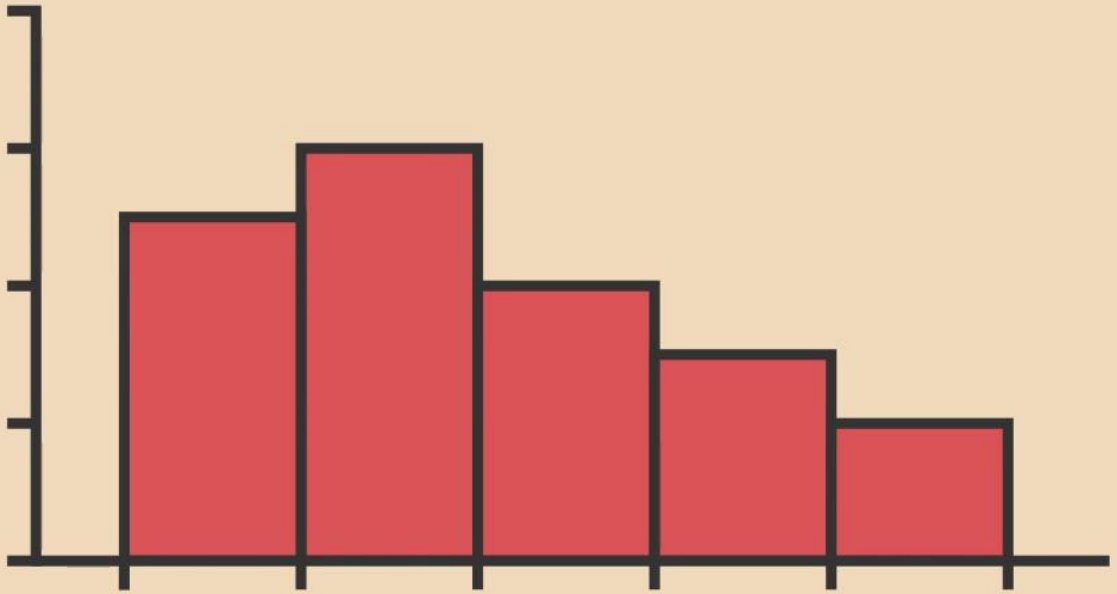


RELATIVE FREQUENCY DISTRIBUTION

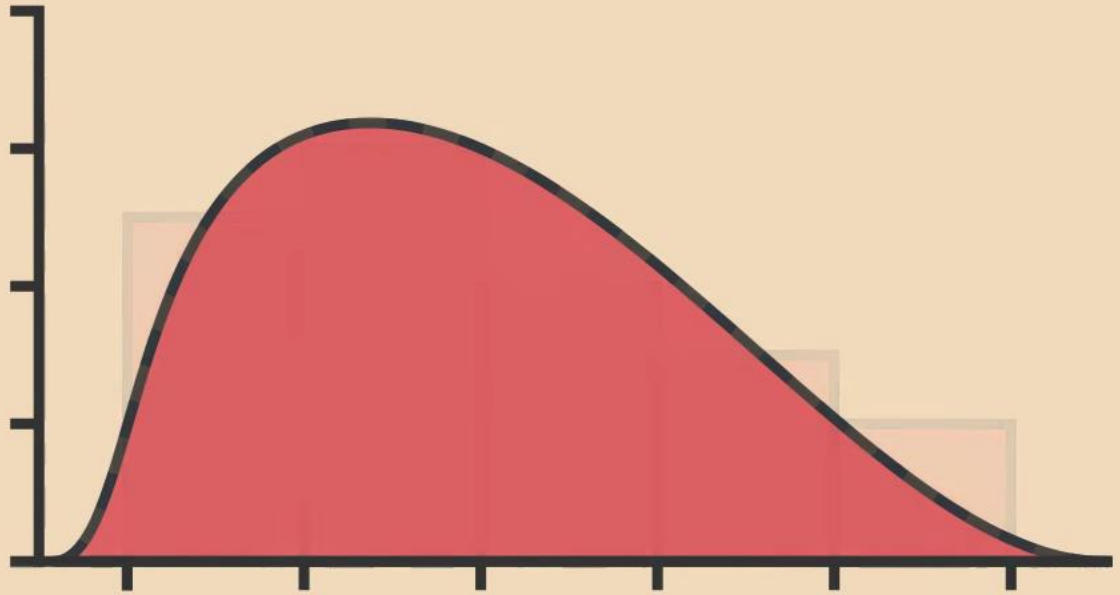


TOTAL AREA = 100%

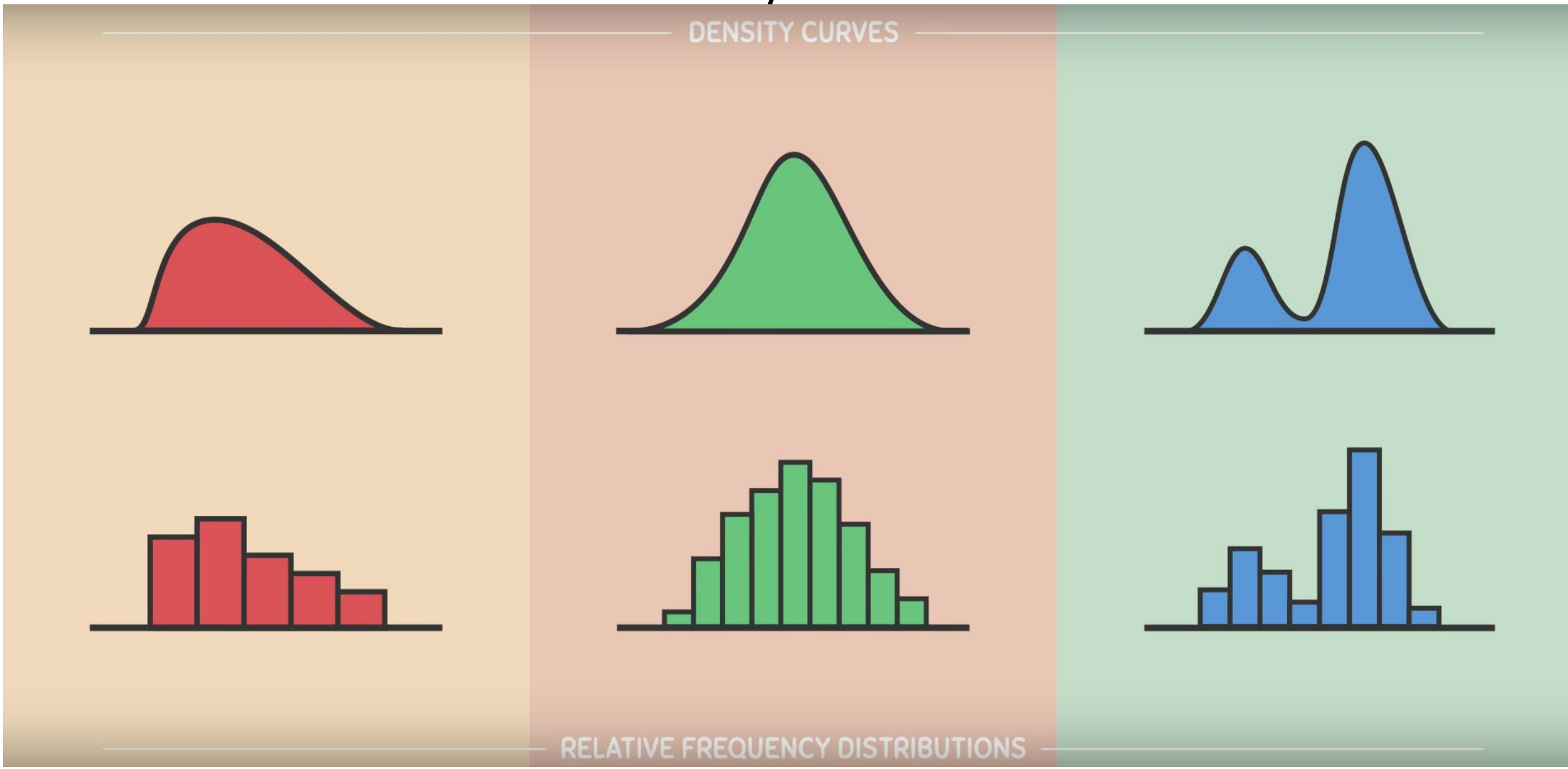
How histogram is converted to density curve



The equivalent density curve

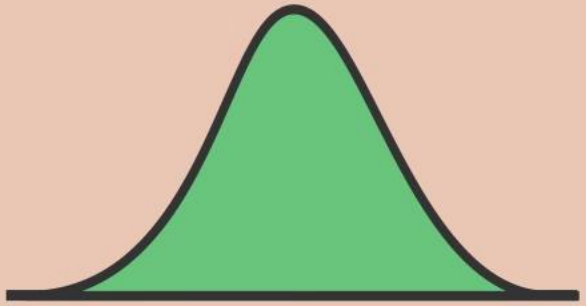


Various form of density curves

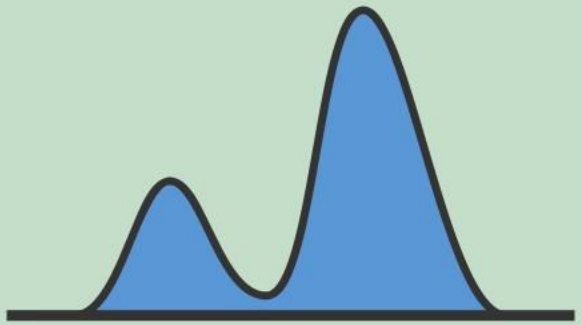


DENSITY CURVES

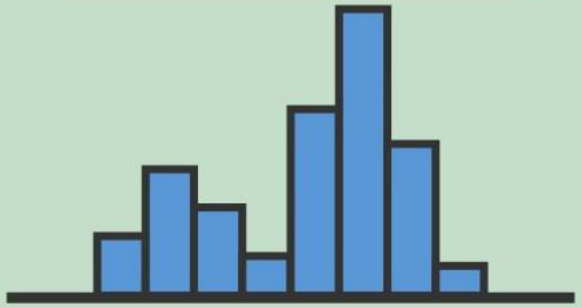
TOTAL AREA = 1



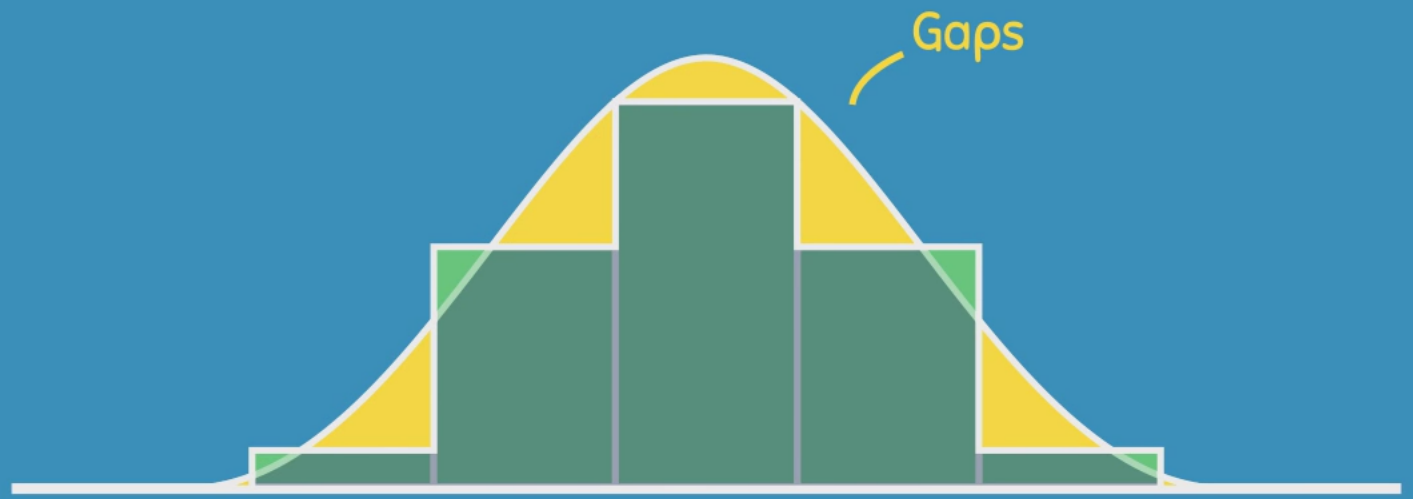
TOTAL AREA = 1



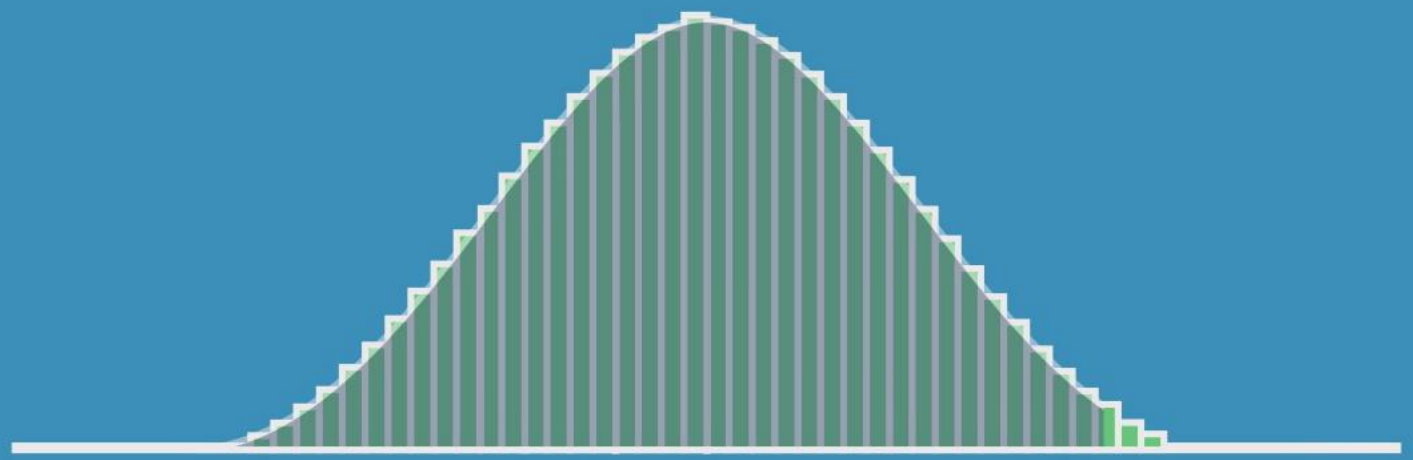
TOTAL AREA = 1



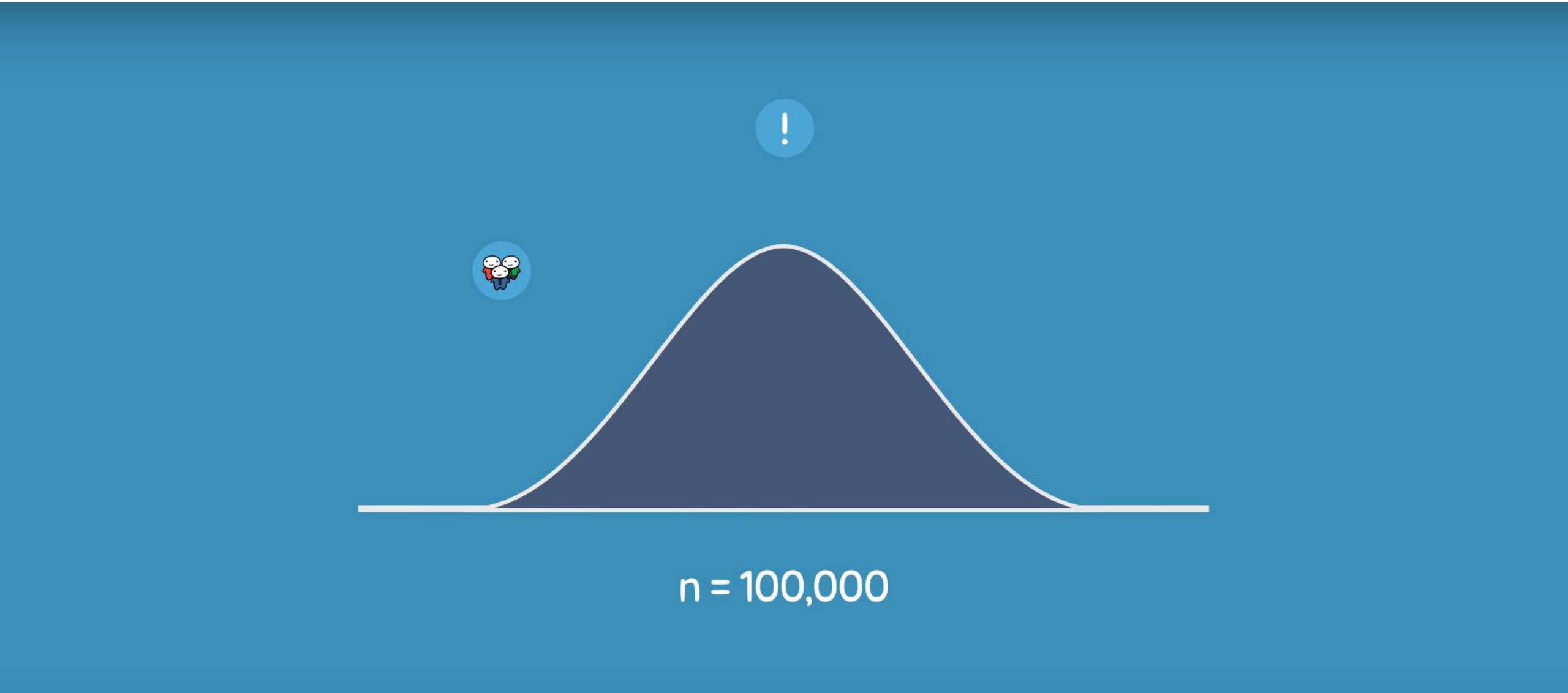
RELATIVE FREQUENCY DISTRIBUTIONS



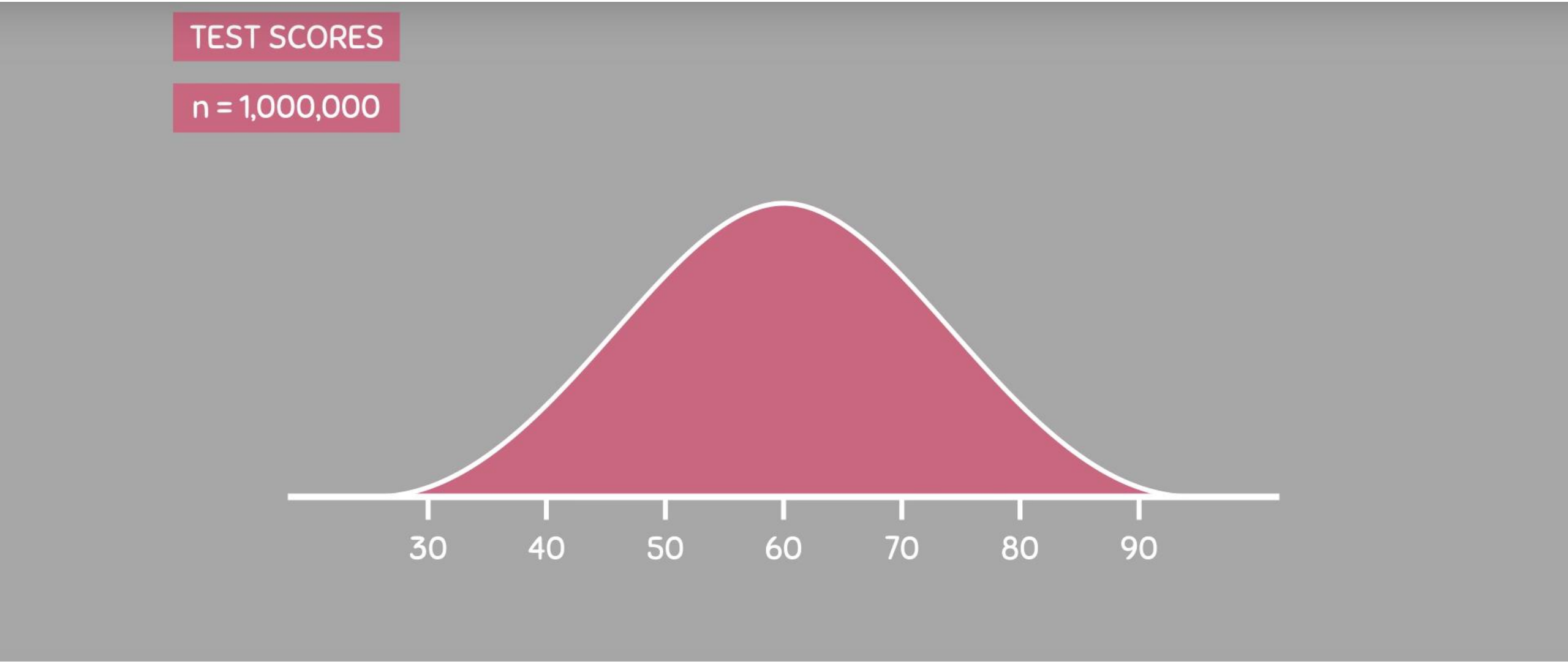
n = 50



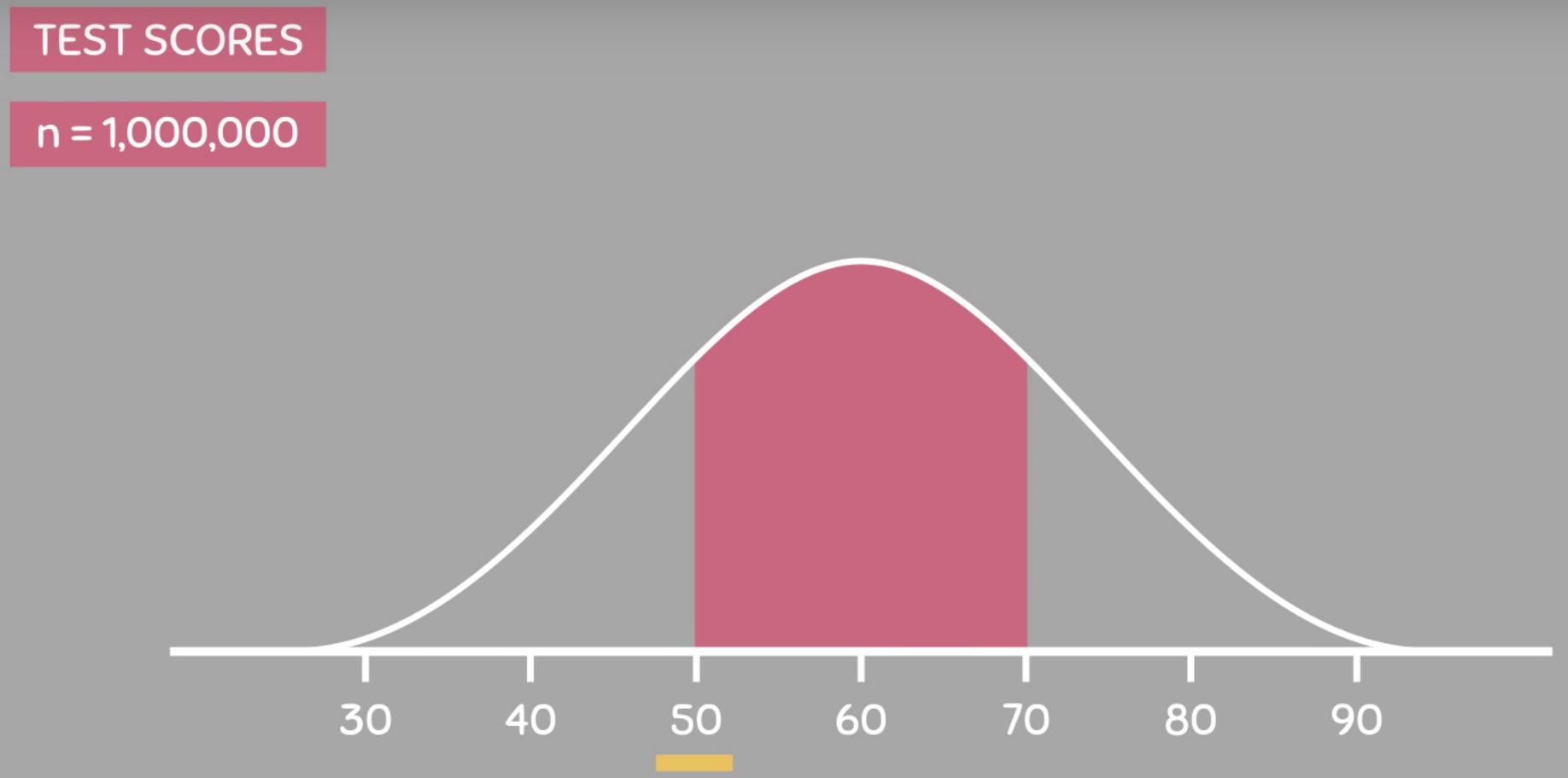
n = 100,000


!
n = 100,000

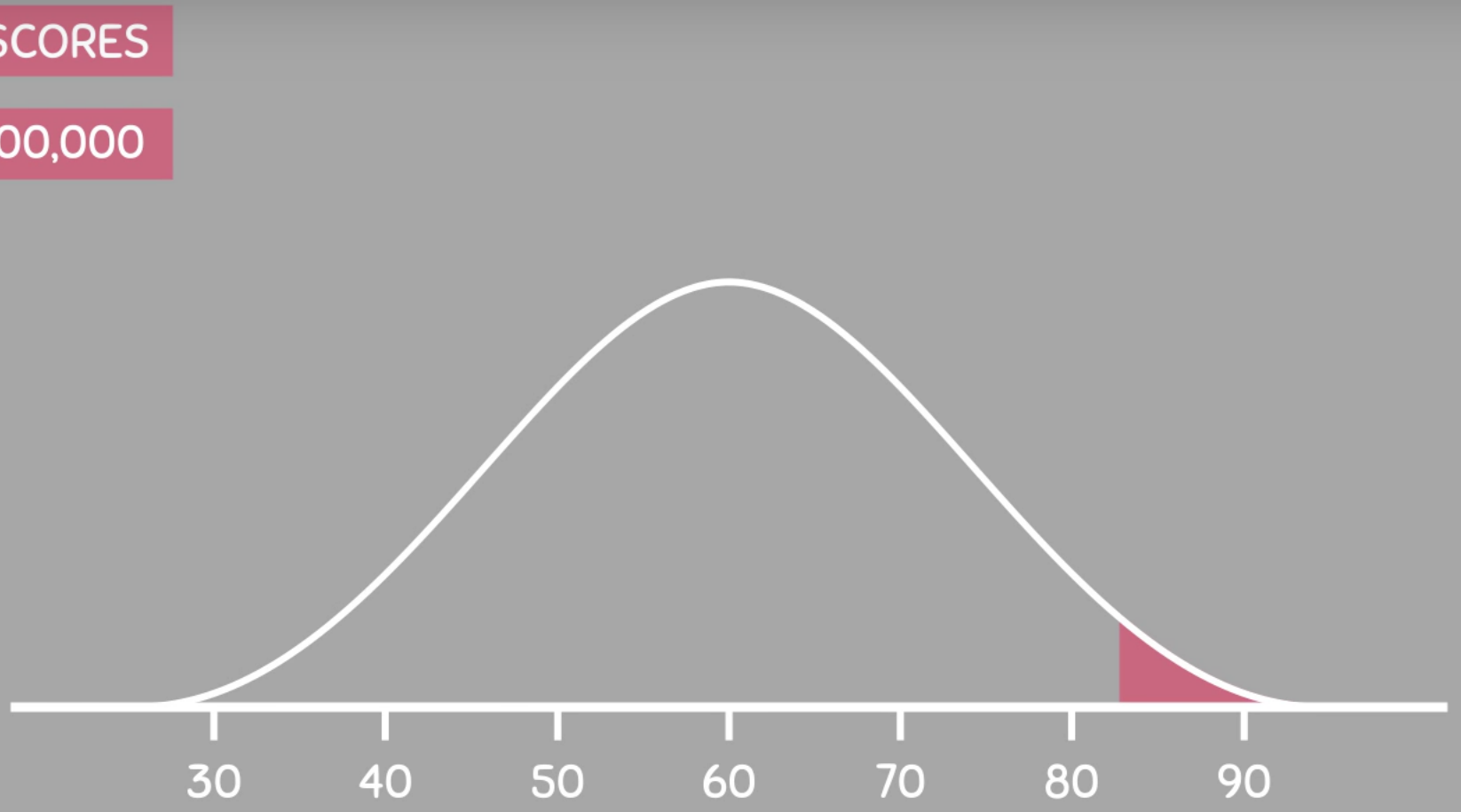
How to read density curves



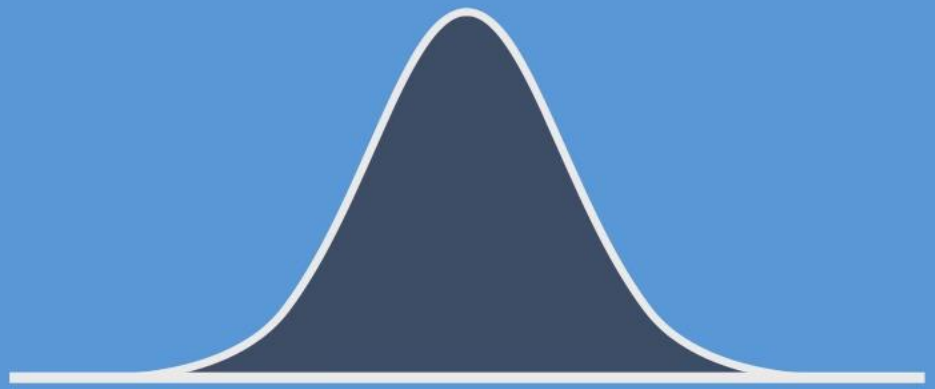
Large majority of students scored score between 50 to 70



Very few students scored score between 85to 90



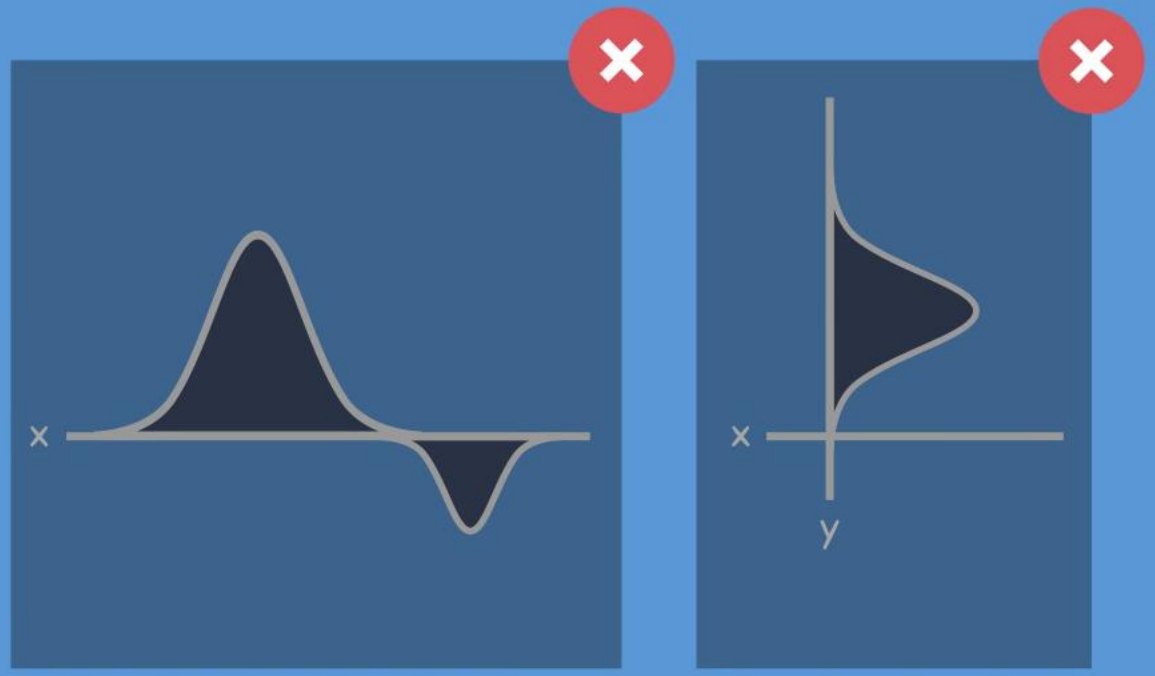
DENSITY CURVE



PROPERTIES OF DENSITY CURVES

1. A density curve must lie on or above the horizontal axis

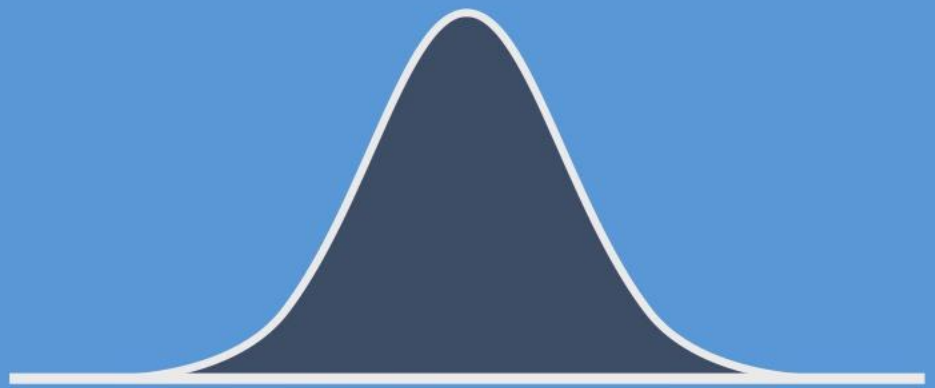
DENSITY CURVE



PROPERTIES OF DENSITY CURVES

1. A density curve must lie on or above the horizontal axis

DENSITY CURVE

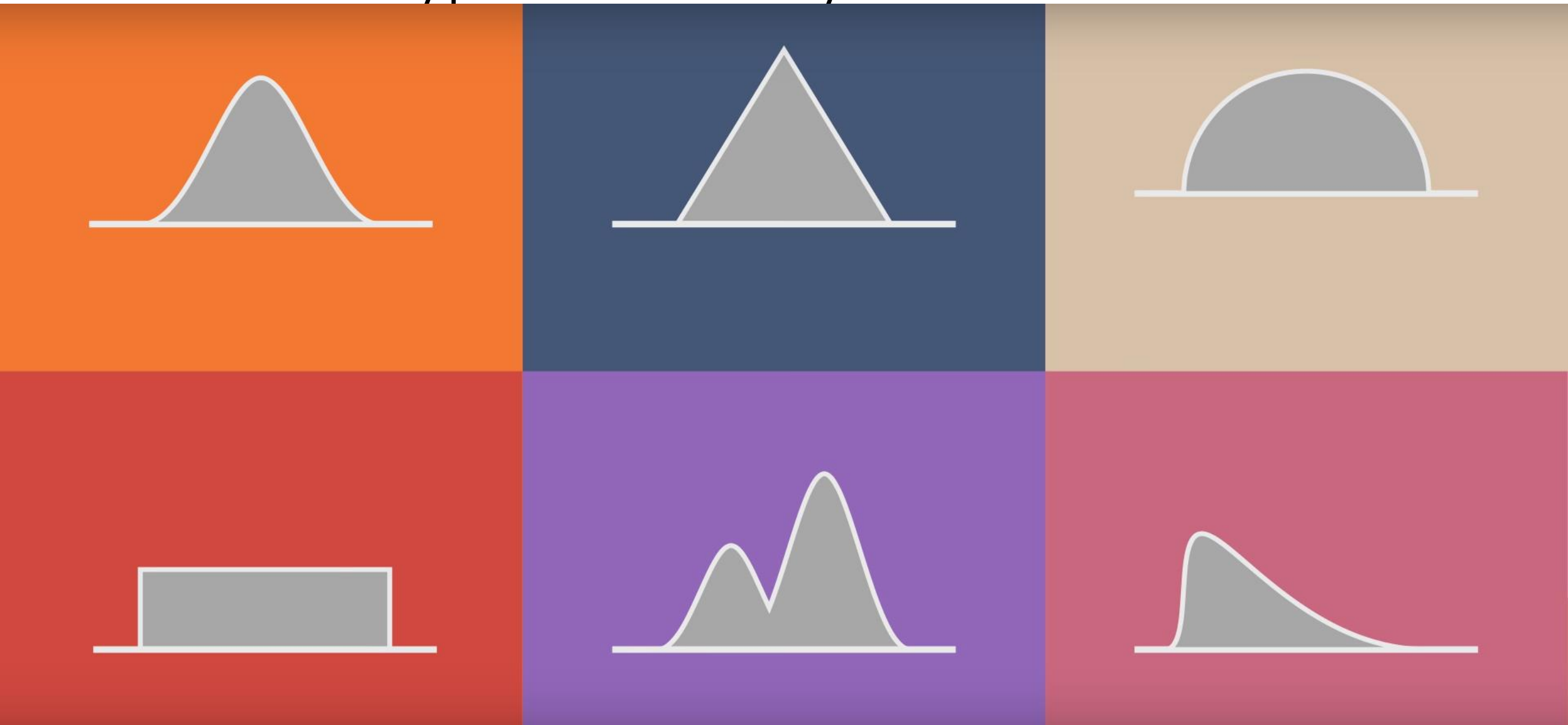


PROPERTIES OF DENSITY CURVES

1. A density curve must lie on or above the horizontal axis

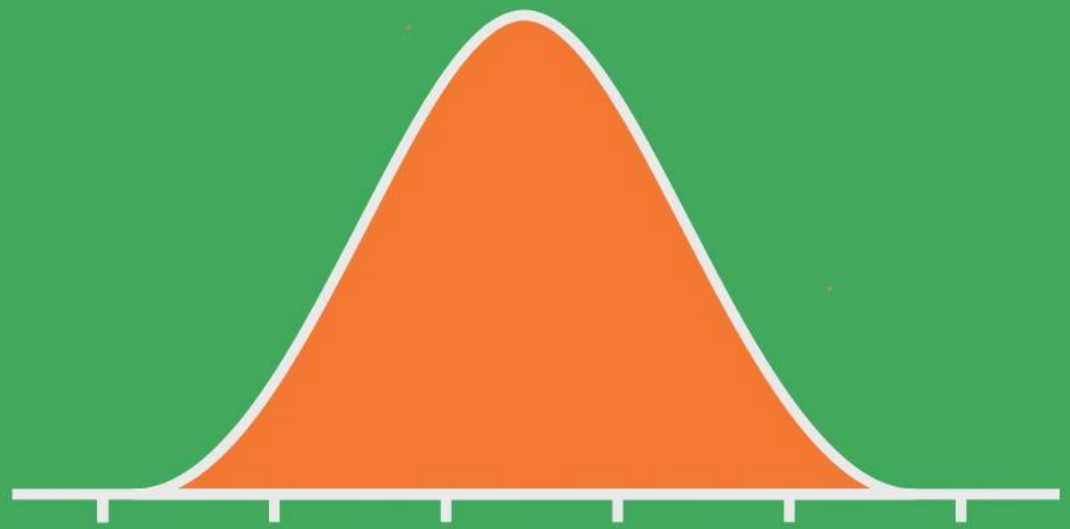
2. The total area under the curve is always equal to 1

Different types of density curves



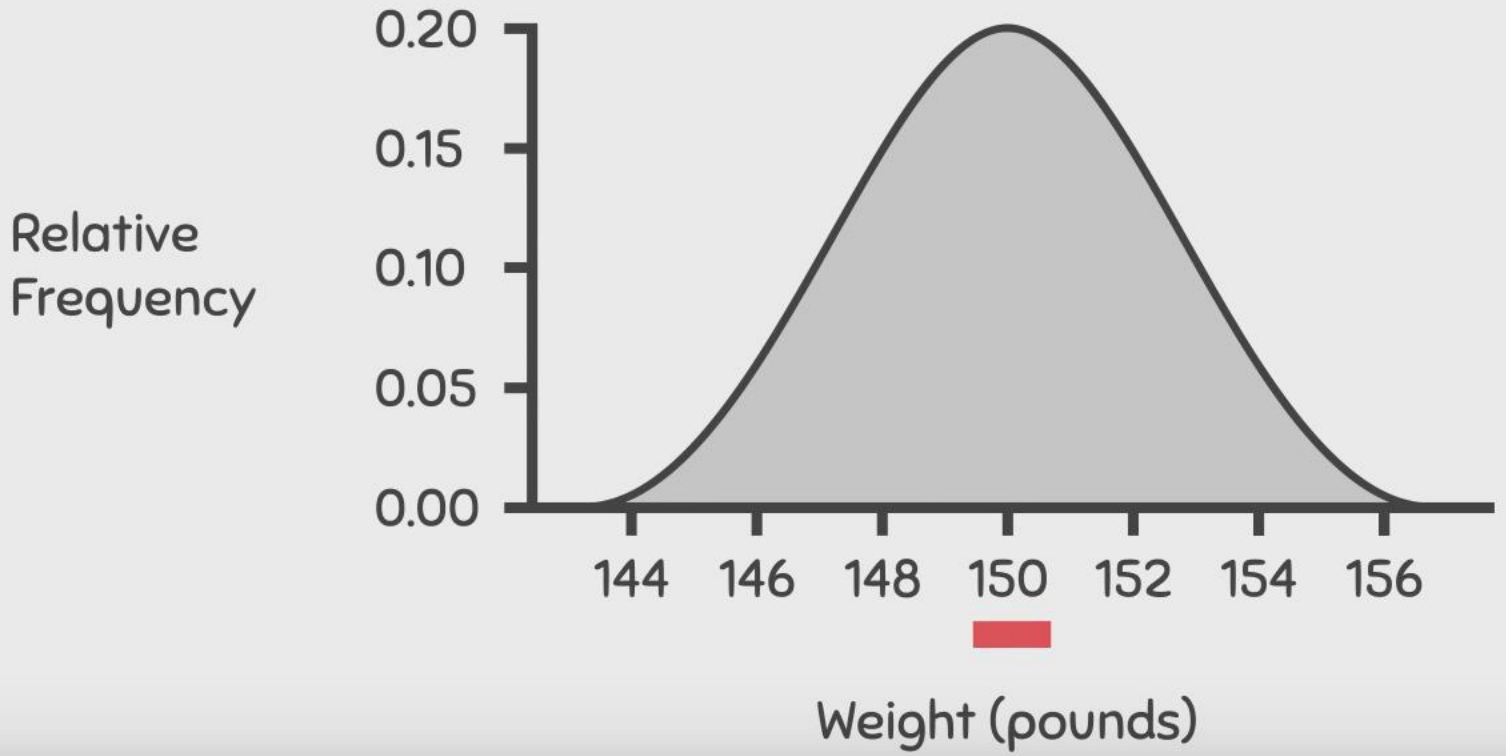
UNIFORM DISTRIBUTION

NORMAL DISTRIBUTION



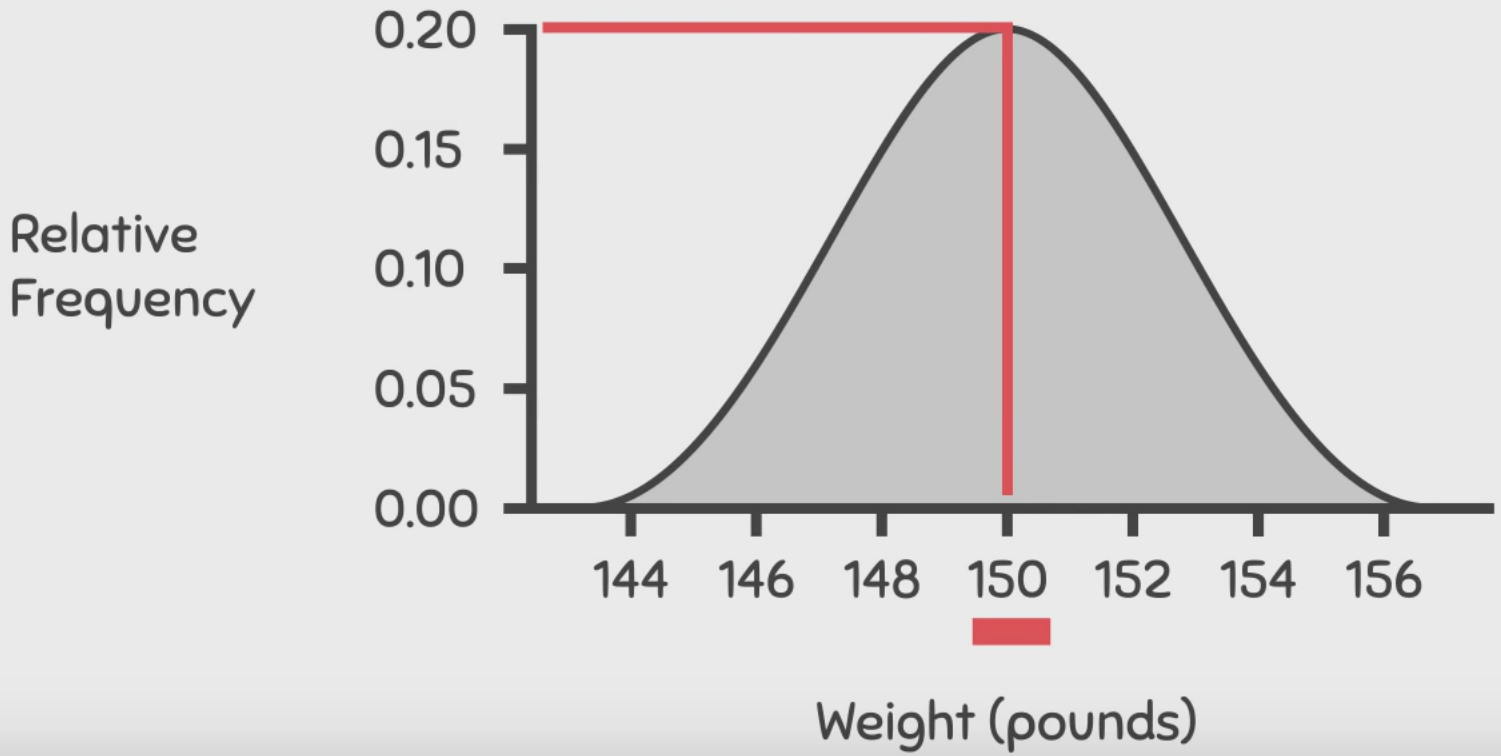
PRACTICE QUESTIONS

1. For the density curve below, approximately what percentage of people weigh exactly 150 pounds?



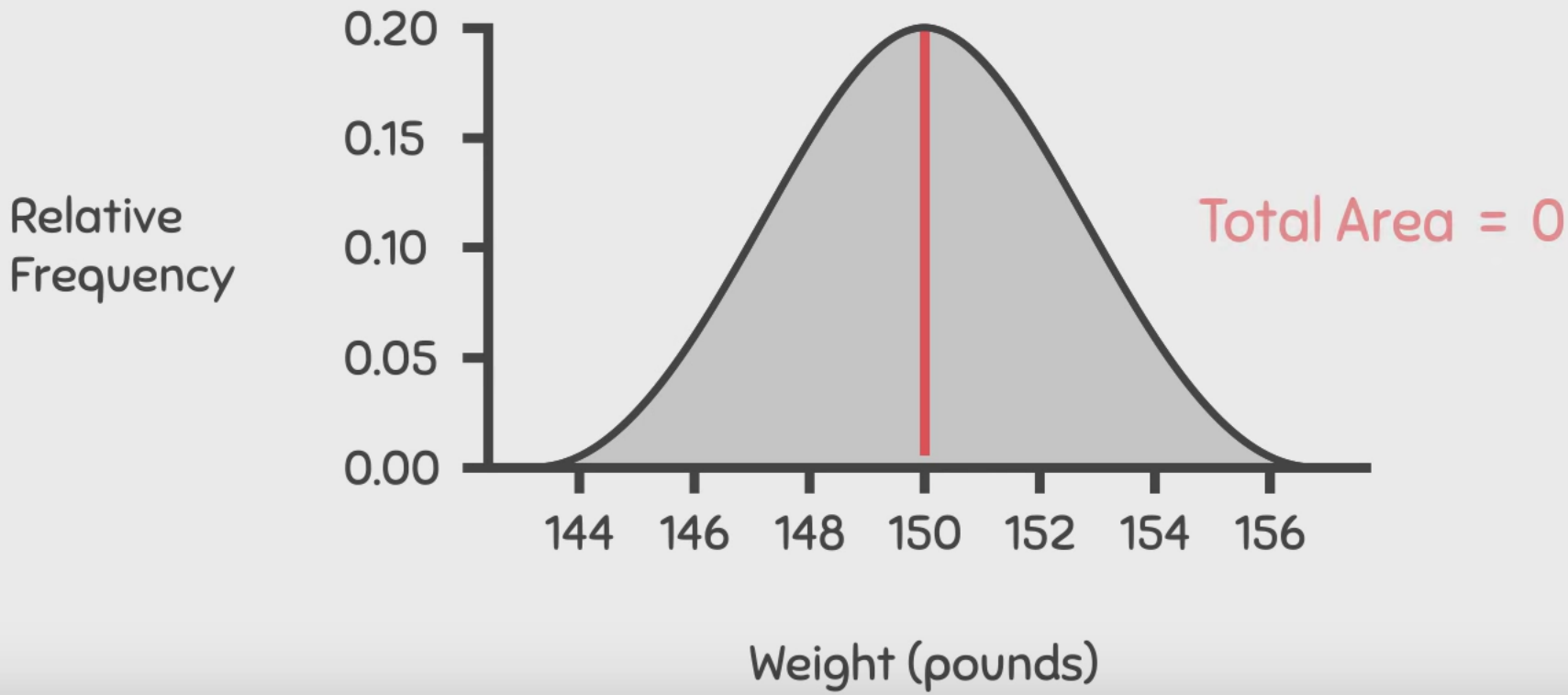
Wrong answer

❶ For the density curve below, approximately what percentage of people weigh exactly 150 pounds?

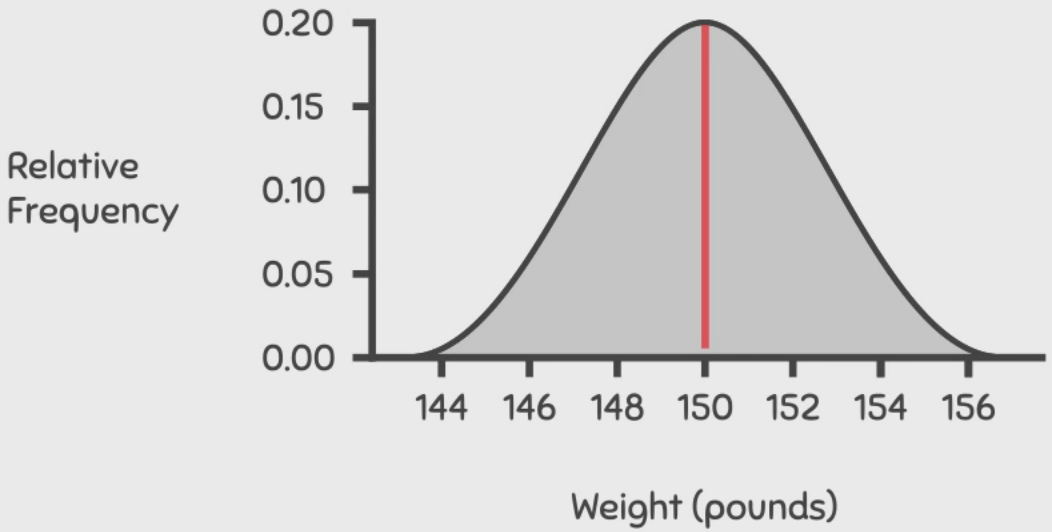


❶ For the density curve below, approximately what percentage of people weigh exactly 150 pounds?

❶ For the density curve below, approximately what percentage of people weigh exactly 150 pounds?



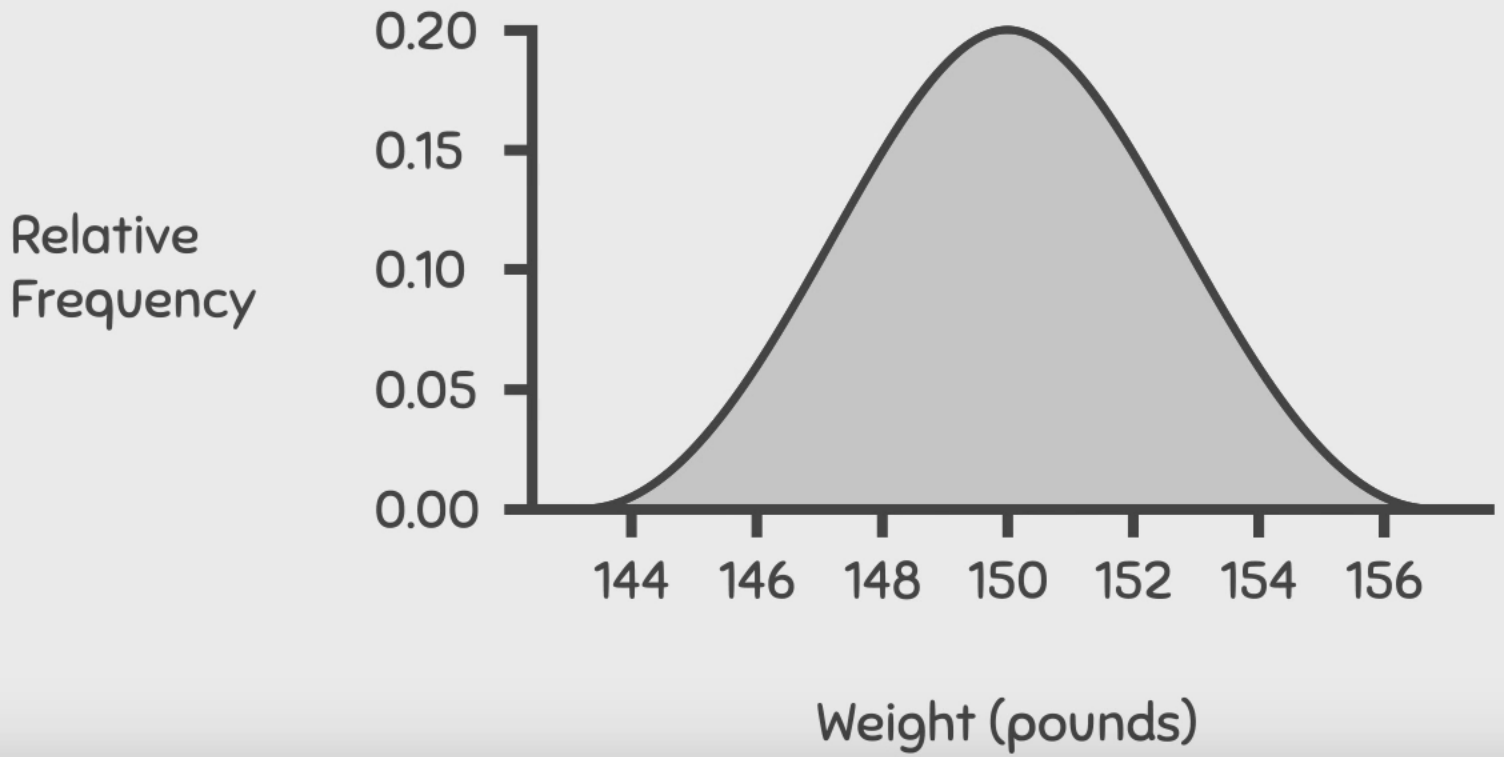
❶ For the density curve below, approximately what percentage of people weigh exactly 150 pounds?



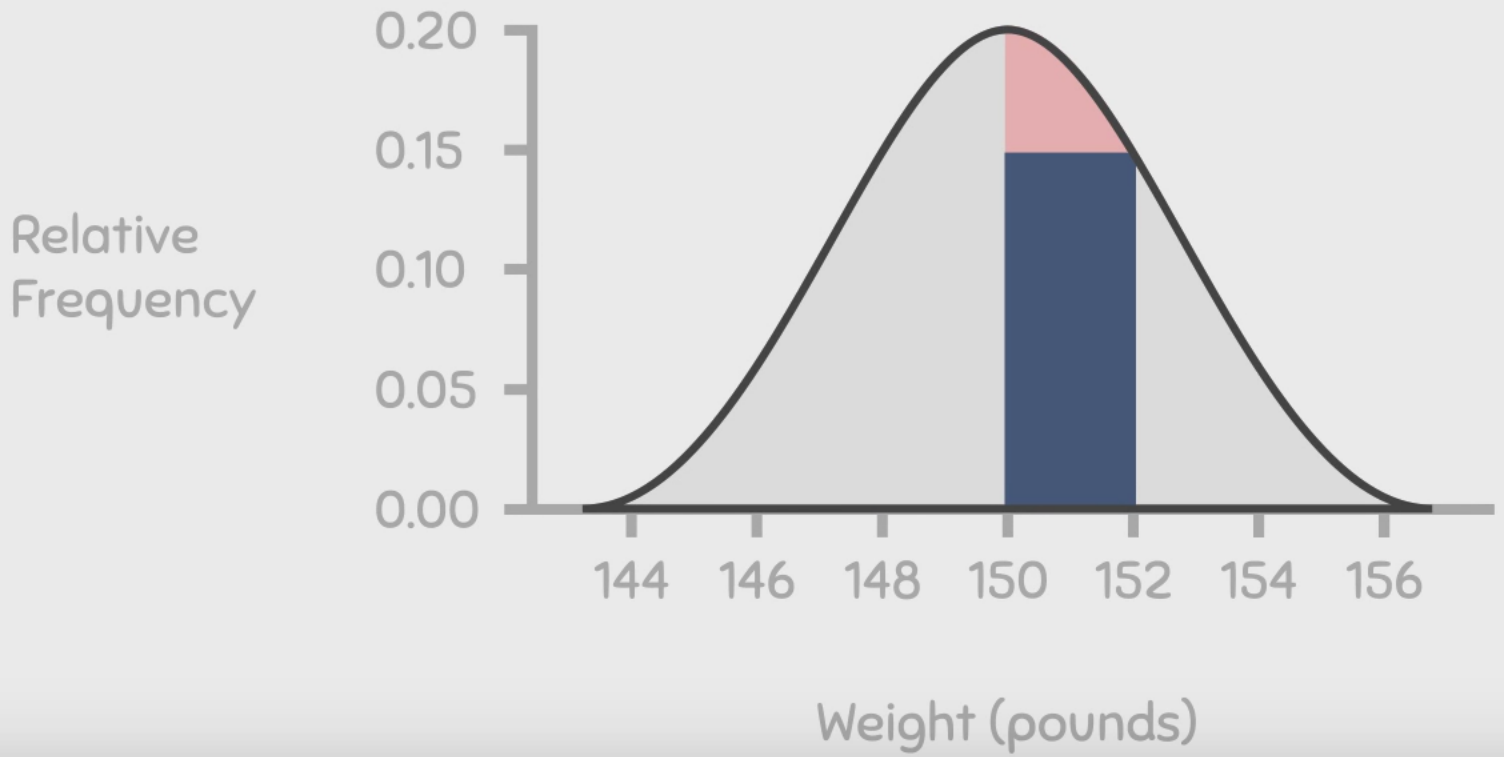
ANSWER = 0

150.000 lbs

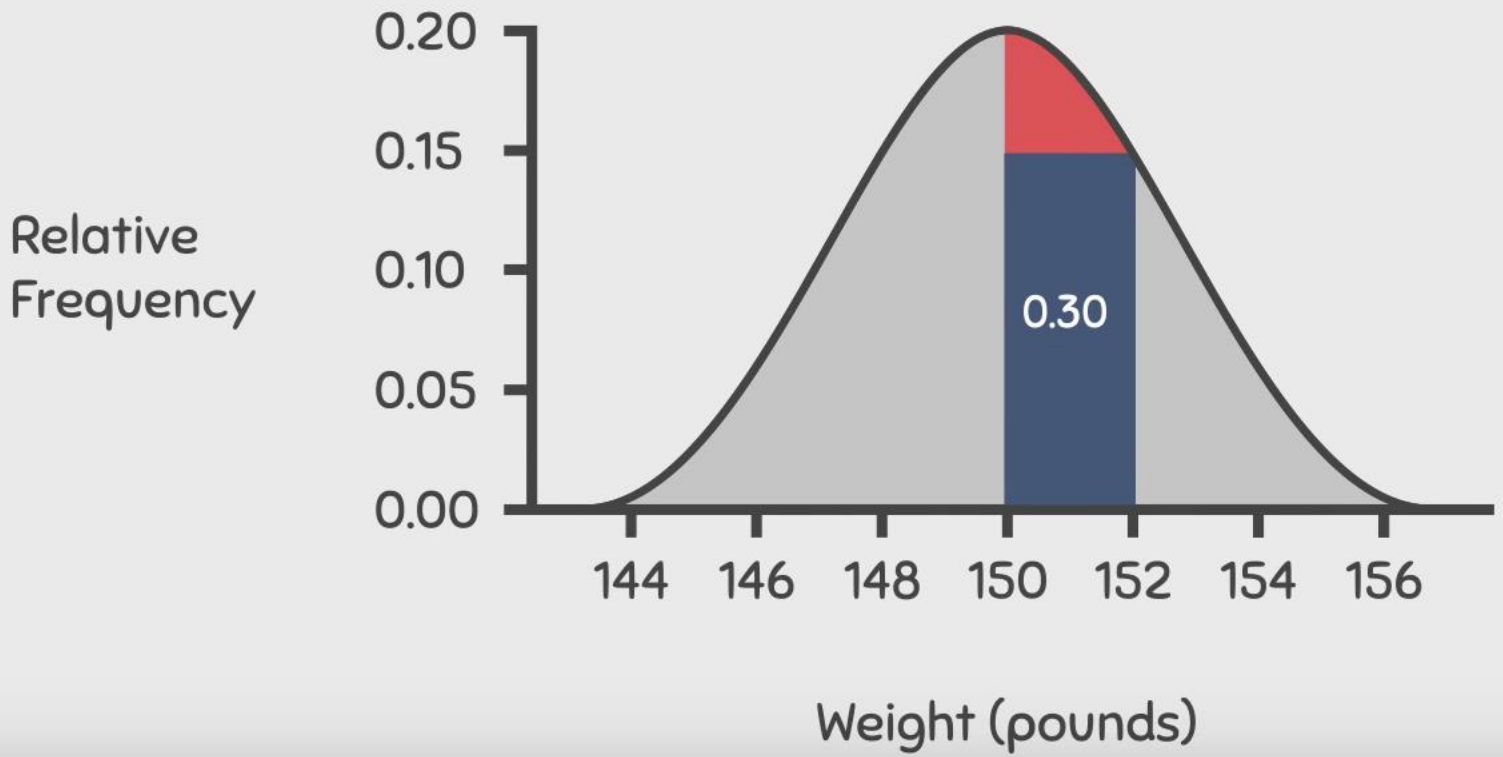
What percentage of people weigh between 150 and 152 pounds?



What percentage of people weigh between 150 and 152 pounds?



What percentage of people weigh between 150 and 152 pounds?

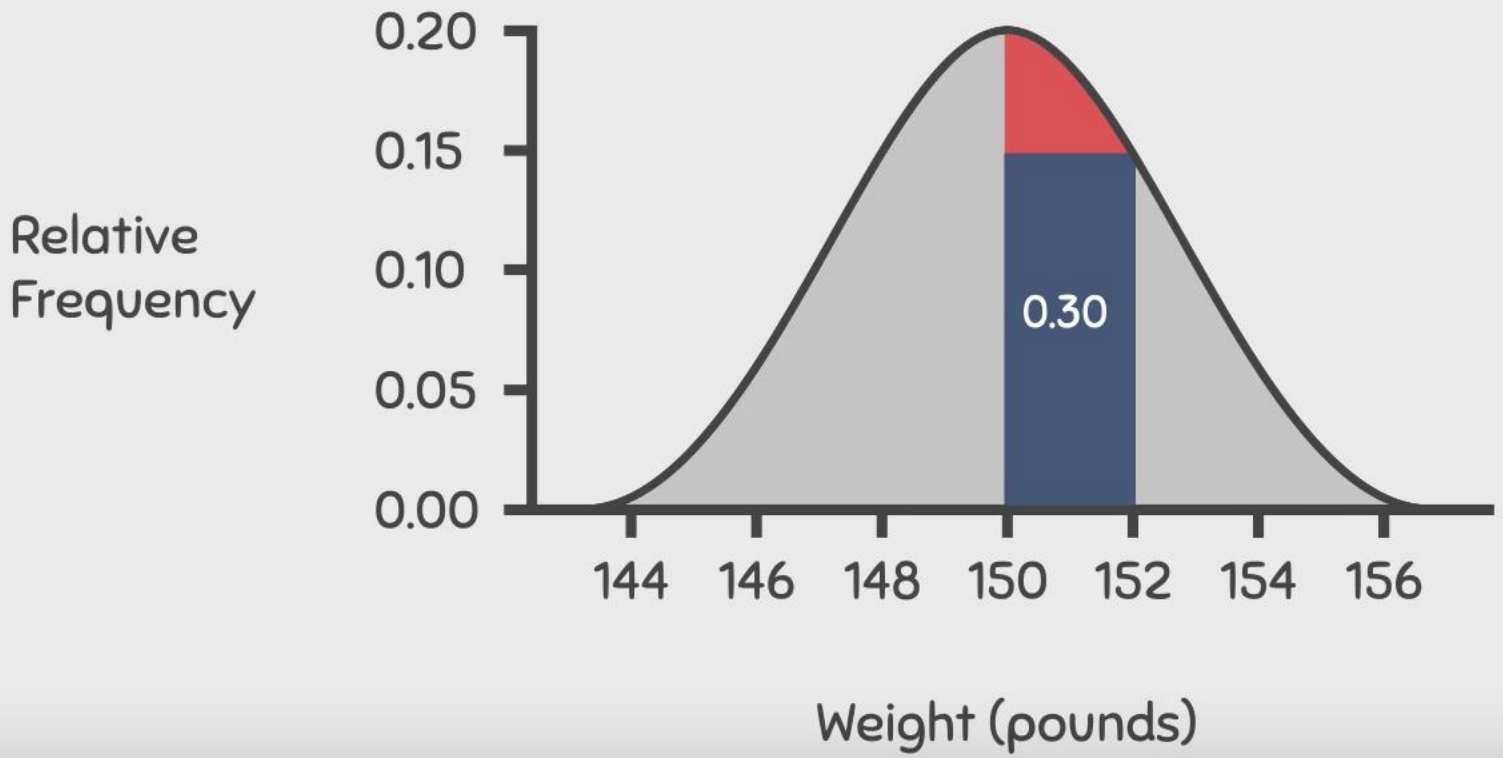


Area = L x W

= 0.15 x 2

= 0.30

What percentage of people weigh between 150 and 152 pounds?



Area = L x W

= 0.15 x 2

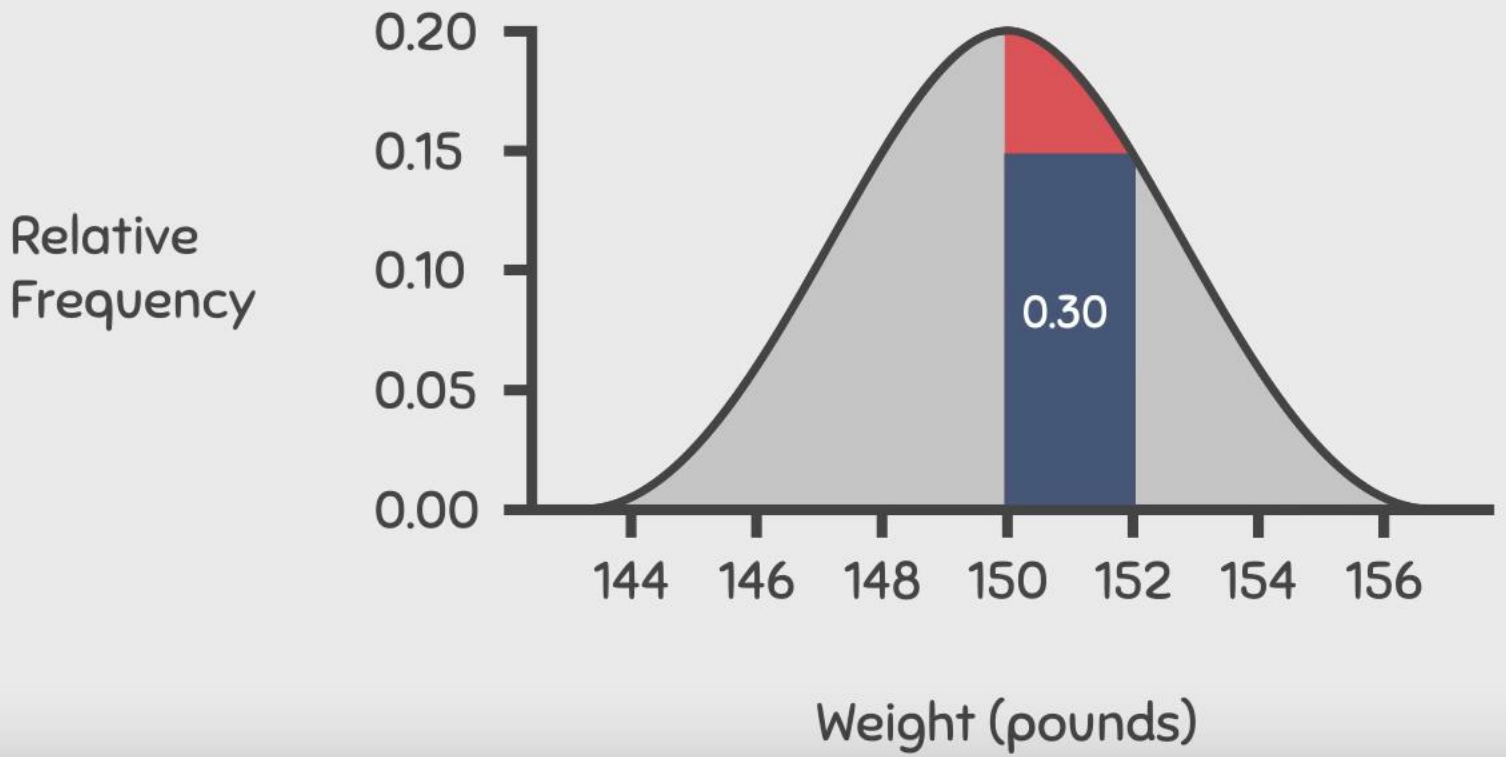
= 0.30

Area = L x W

2

0.05

What percentage of people weigh between 150 and 152 pounds?



Area = L x W

= 0.15 x 2

= 0.30

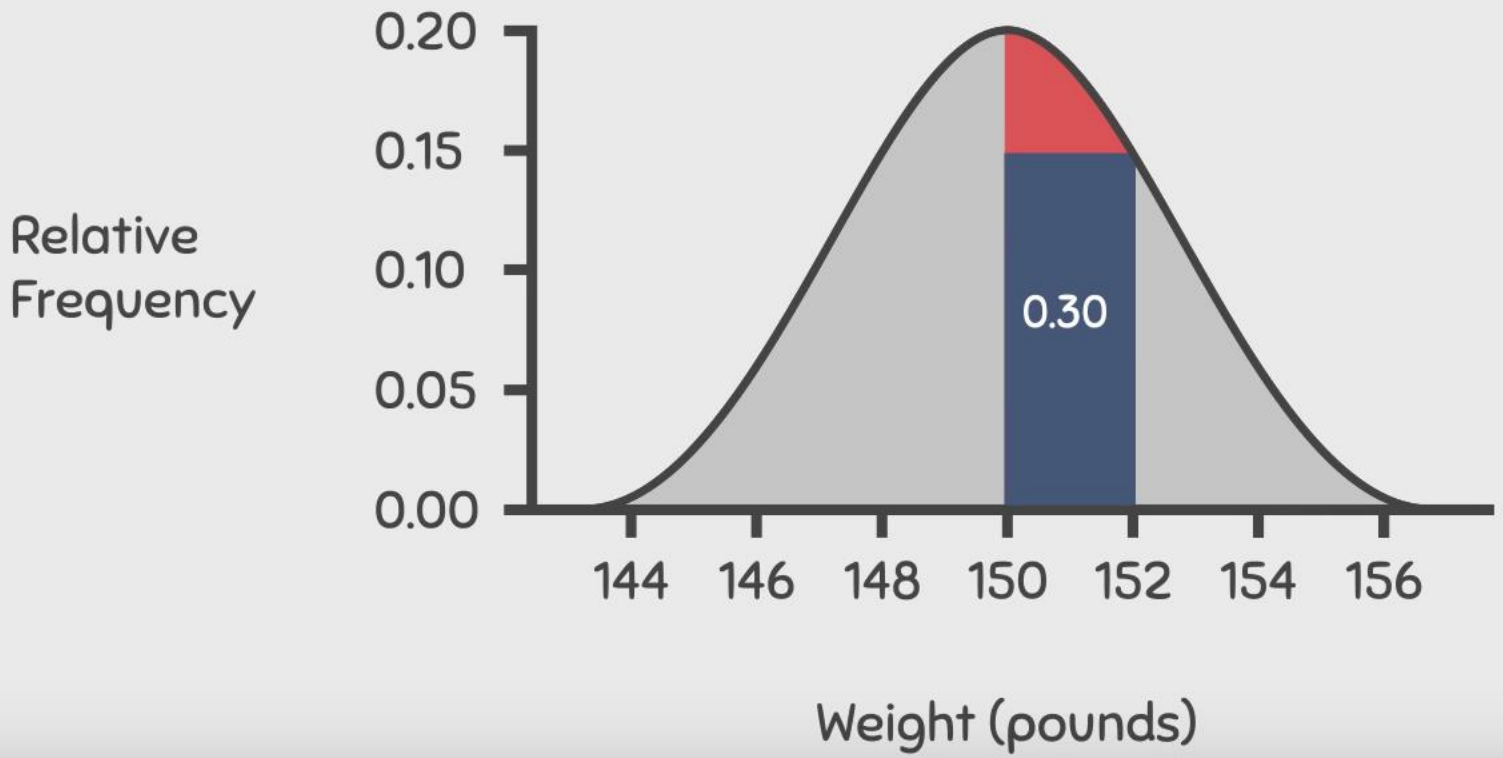
Area = L x W

= 0.05 x 2

= 0.1

0.1

What percentage of people weigh between 150 and 152 pounds?



Area = L x W

= 0.15 x 2

= 0.30

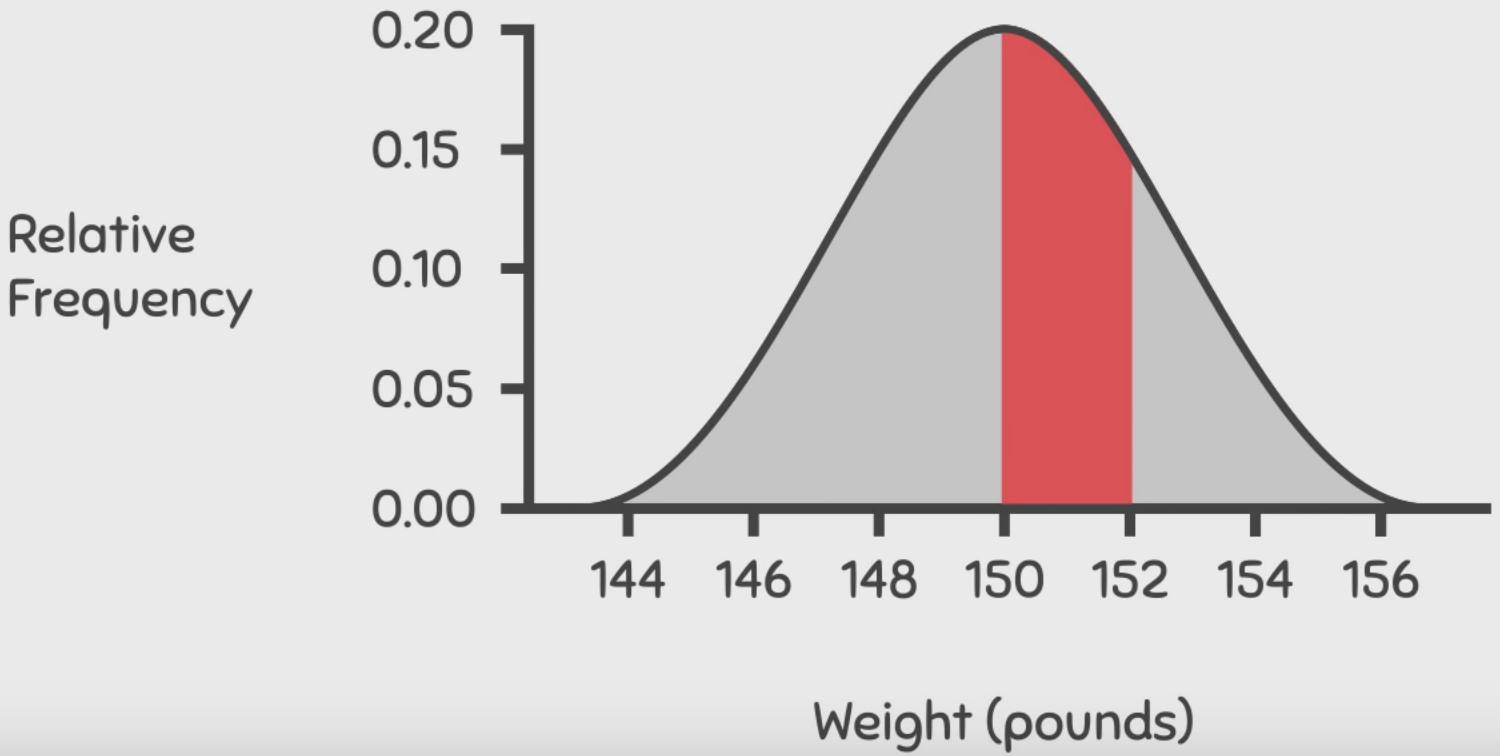
Area = L x W

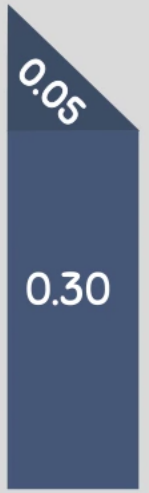
= 0.05 x 2

= 0.1

0.1 ÷ 2 = 0.05

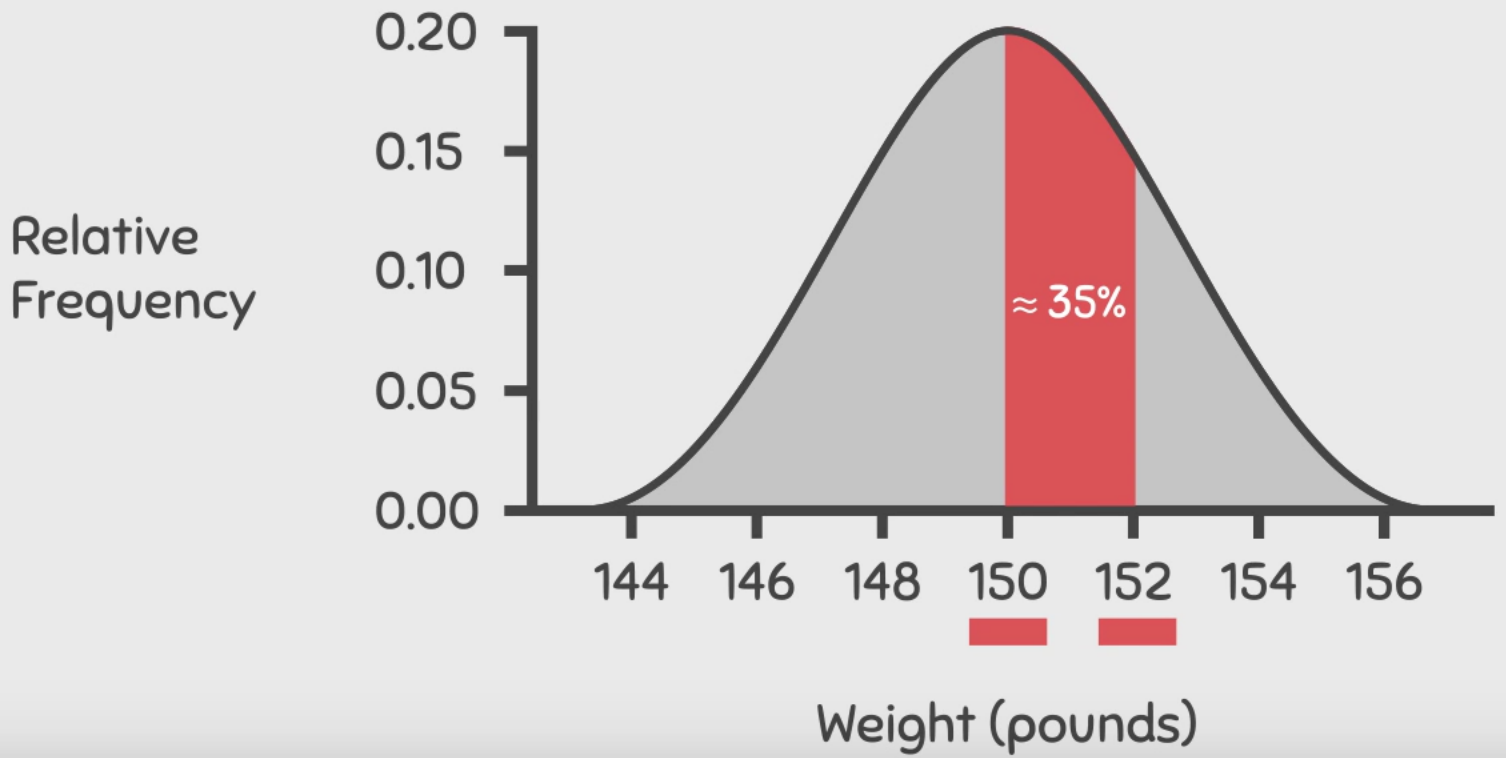
What percentage of people weigh between 150 and 152 pounds?



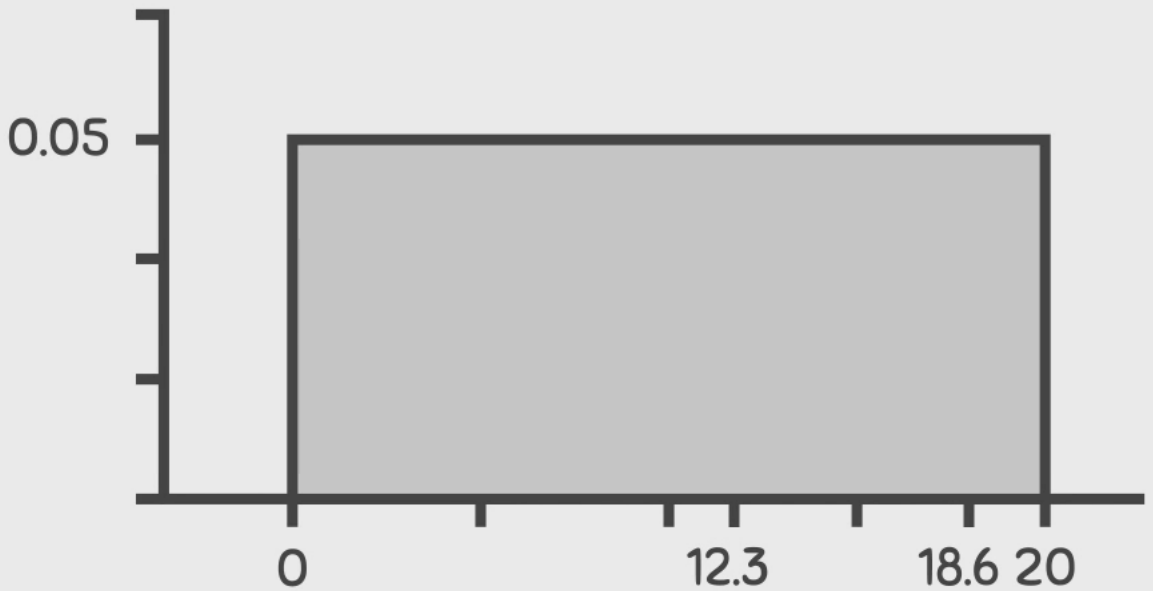


Total Area = 0.05 + 0.30

What percentage of people weigh between 150 and 152 pounds?

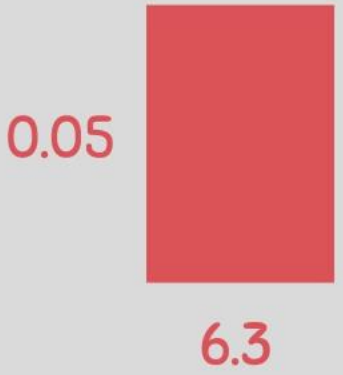


For the uniform distribution below, what proportion of values are located between 12.3 and 18.6?

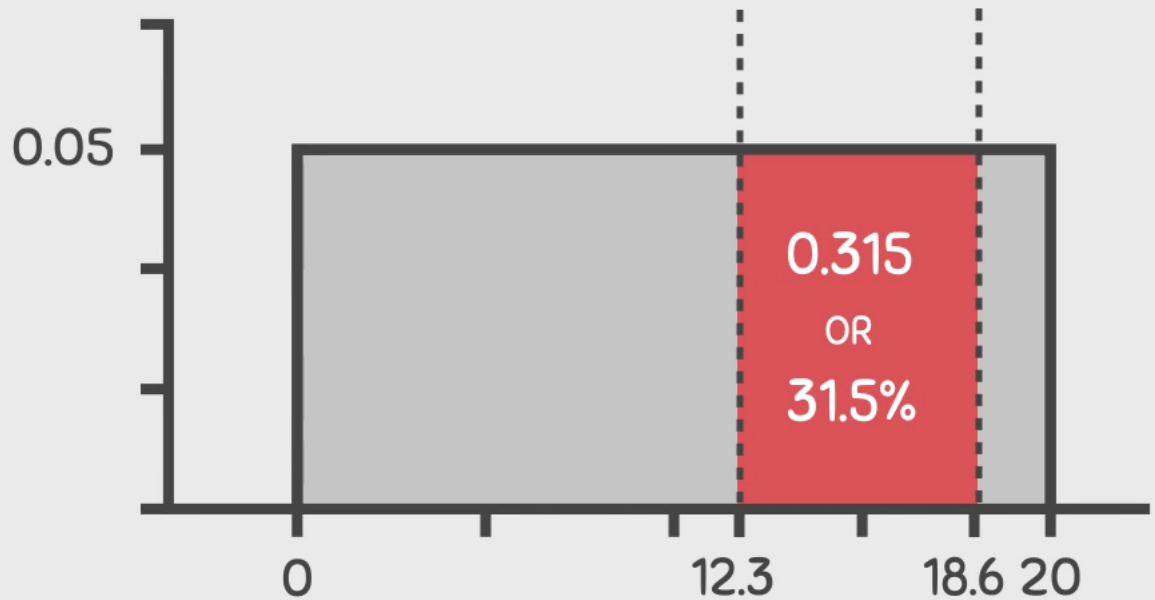


❸ For the uniform distribution below, what proportion of values are located between 12.3 and 18.6?

Area = L x W



For the uniform distribution below, what proportion of values are located between 12.3 and 18.6?



Normal Distribution

PARAMETER

A number that describes the data from a population

STATISTIC

A number that describes the data from a sample

	SAMPLE	POPULATION
MEAN	$\bar{x}$	μ
STANDARD DEVIATION	s	σ
	STATISTIC	PARAMETER

BELL CURVE

NORMAL CURVE

NORMAL DISTRIBUTION

WEIGHT

HEIGHT

μ

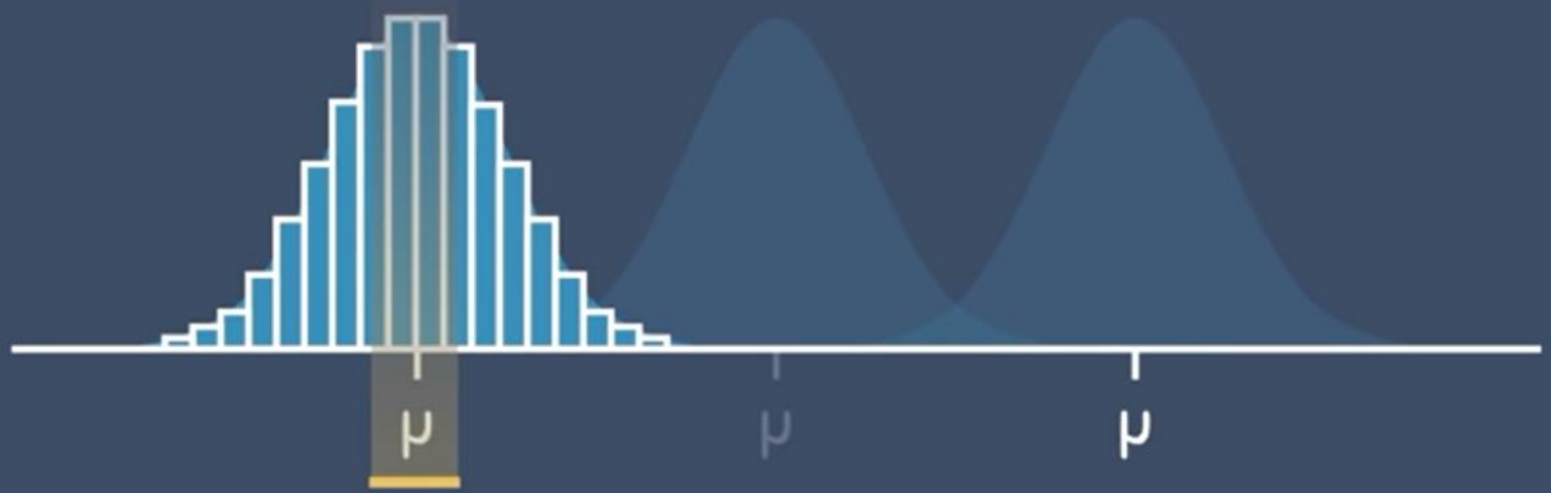
POPULATION MEAN

CHARACTERIZES THE POSITION OF THE NORMAL DISTRIBUTION

μ

POPULATION MEAN

CHARACTERIZES THE POSITION OF THE NORMAL DISTRIBUTION



σ

POPULATION
STANDARD DEVIATION

CHARACTERIZES THE SPREAD OF THE NORMAL DISTRIBUTION

σ = small

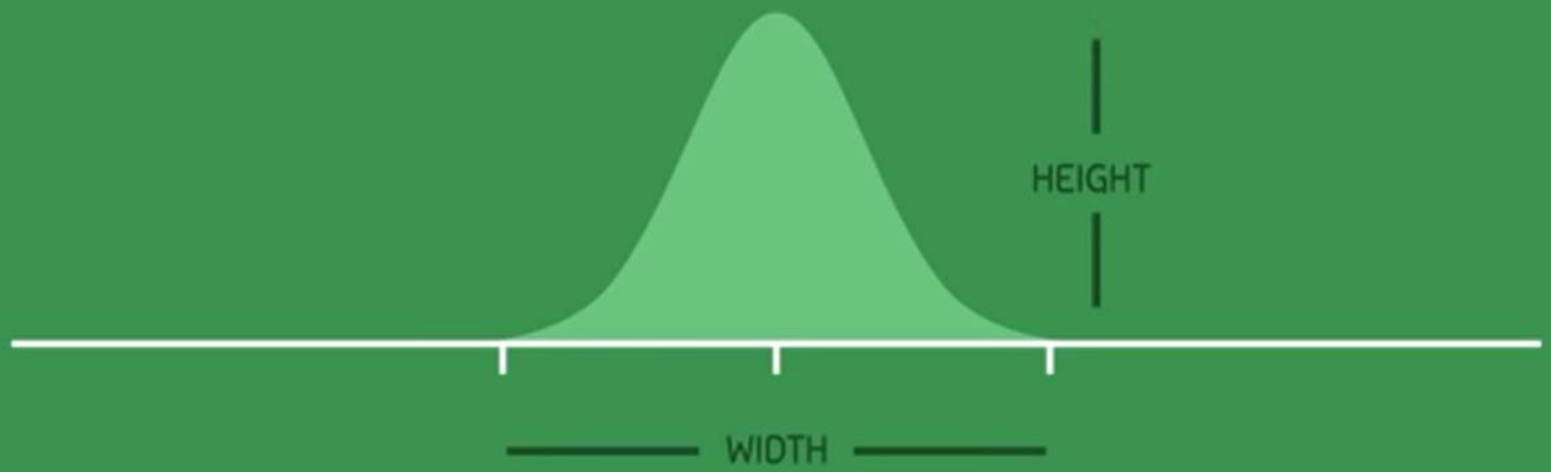
σ

POPULATION
STANDARD DEVIATION

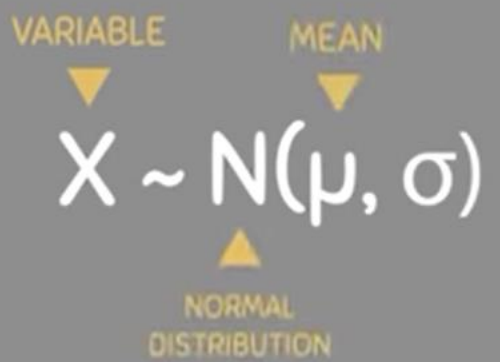
CHARACTERIZES THE SPREAD OF THE NORMAL DISTRIBUTION

DENSITY CURVE

TOTAL AREA = 100%



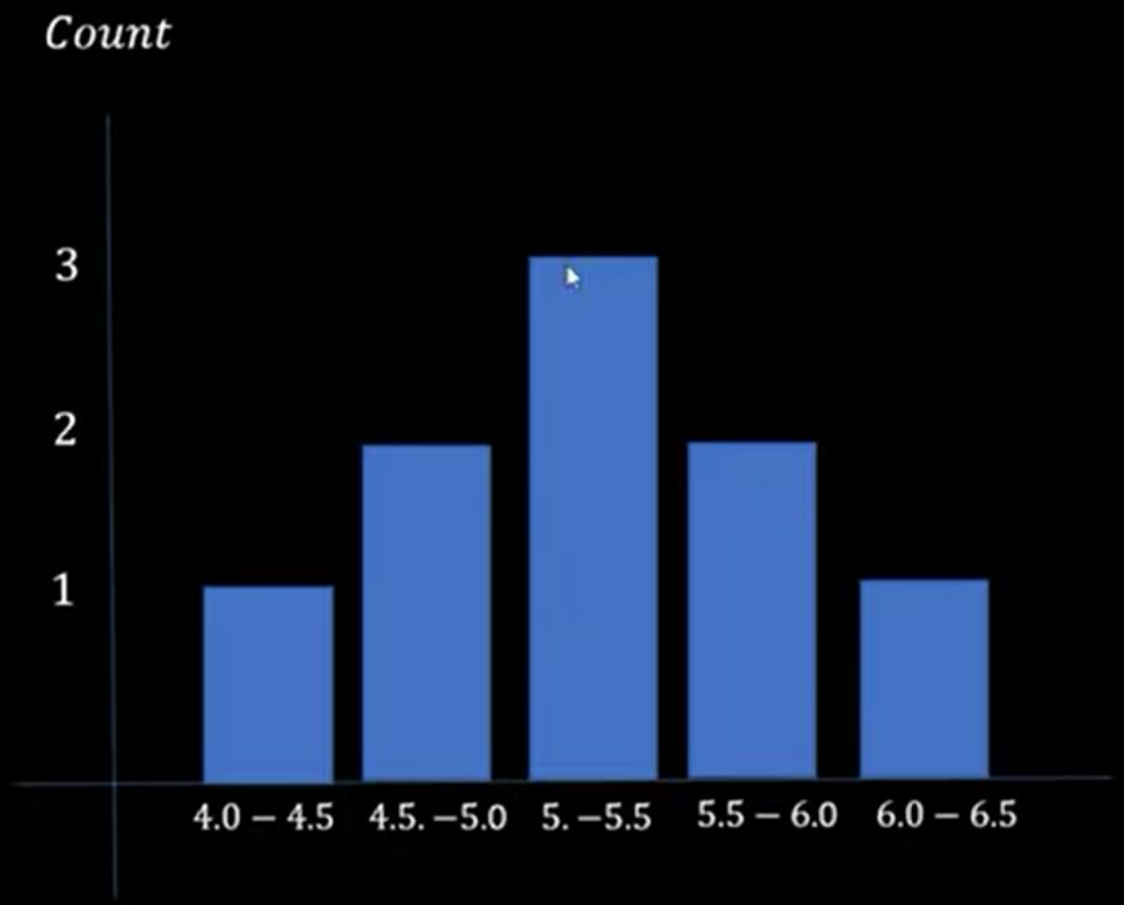
1. The normal distribution is unimodal
2. The normal curve is symmetric about its mean
3. The parameters μ and σ completely characterize the normal distribution
4. $X \sim N(\mu, \sigma)$



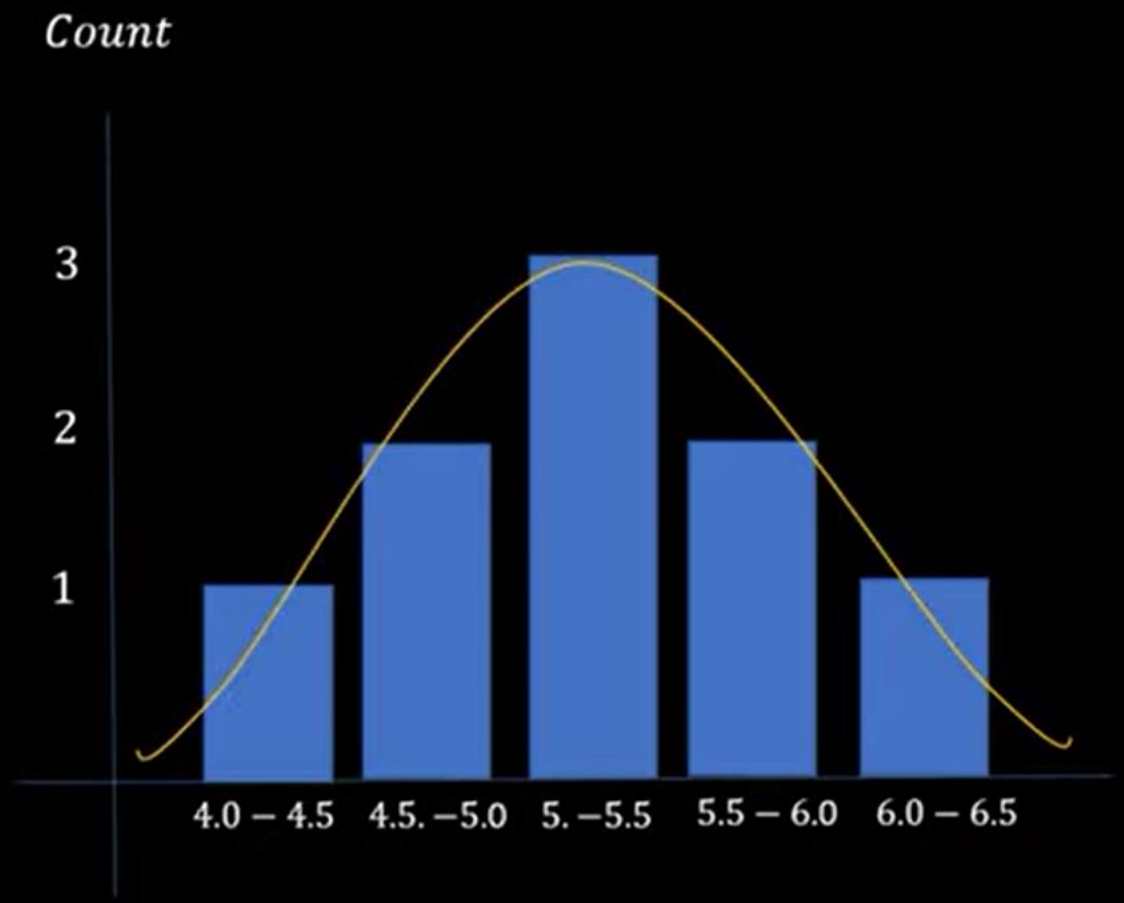
Regular normal distribution

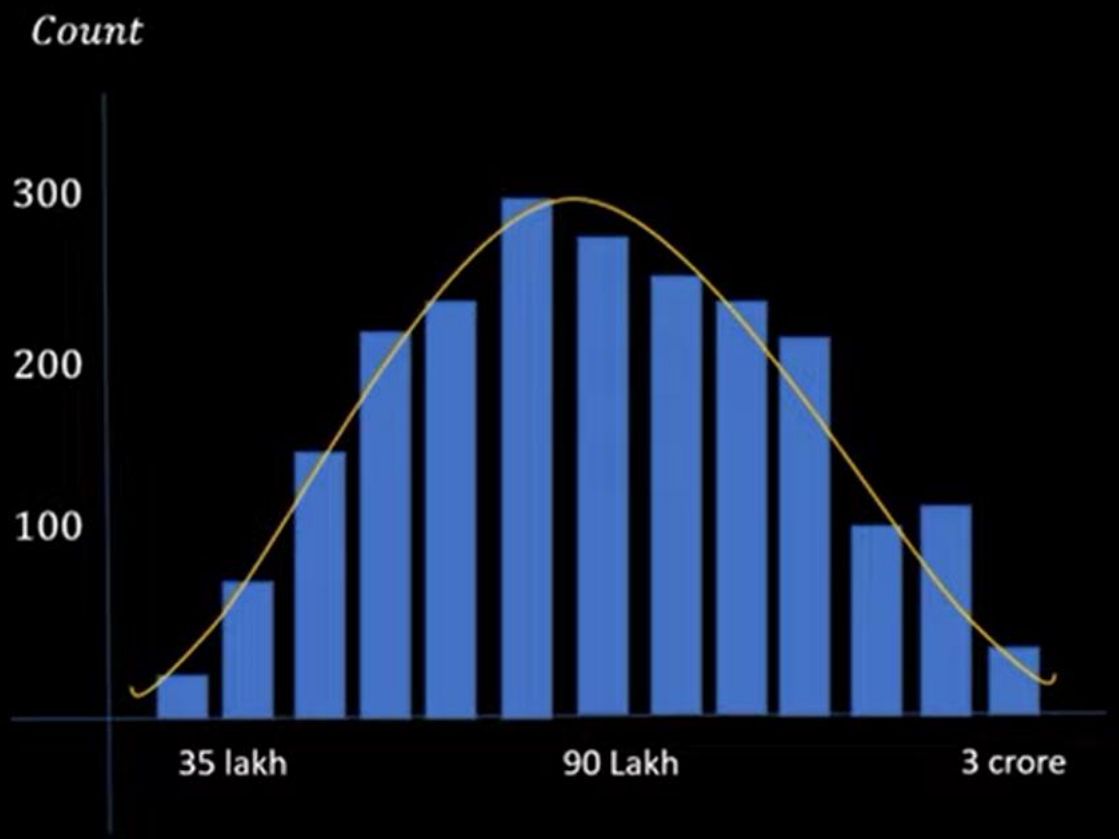
Name	Height (ft)
Rob	6.2
Thomas	5.7
Nina	4.6
Mittal	5.4
Sofia	5.9
Mohan	4.3
Tao	5.1
Deepika	5.2
Rafiq	4.9

Name	Height (ft)
Rob	6.2
Thomas	5.7
Nina	4.6
Mittal	5.4
Sofia	5.9
Mohan	4.3
Tao	5.1
Deepika	5.2
Rafiq	4.9

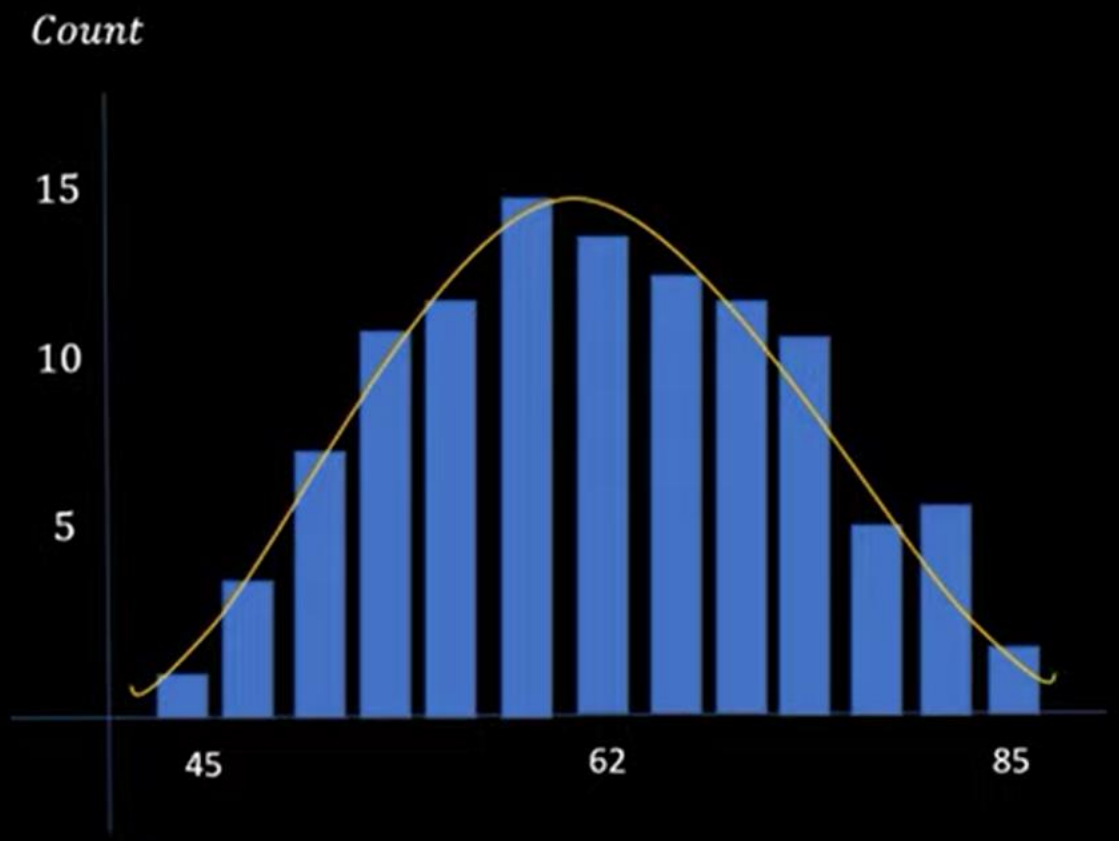


Name	Height (ft)
Rob	6.2
Thomas	5.7
Nina	4.6
Mittal	5.4
Sofia	5.9
Mohan	4.3
Tao	5.1
Deepika	5.2
Rafiq	4.9

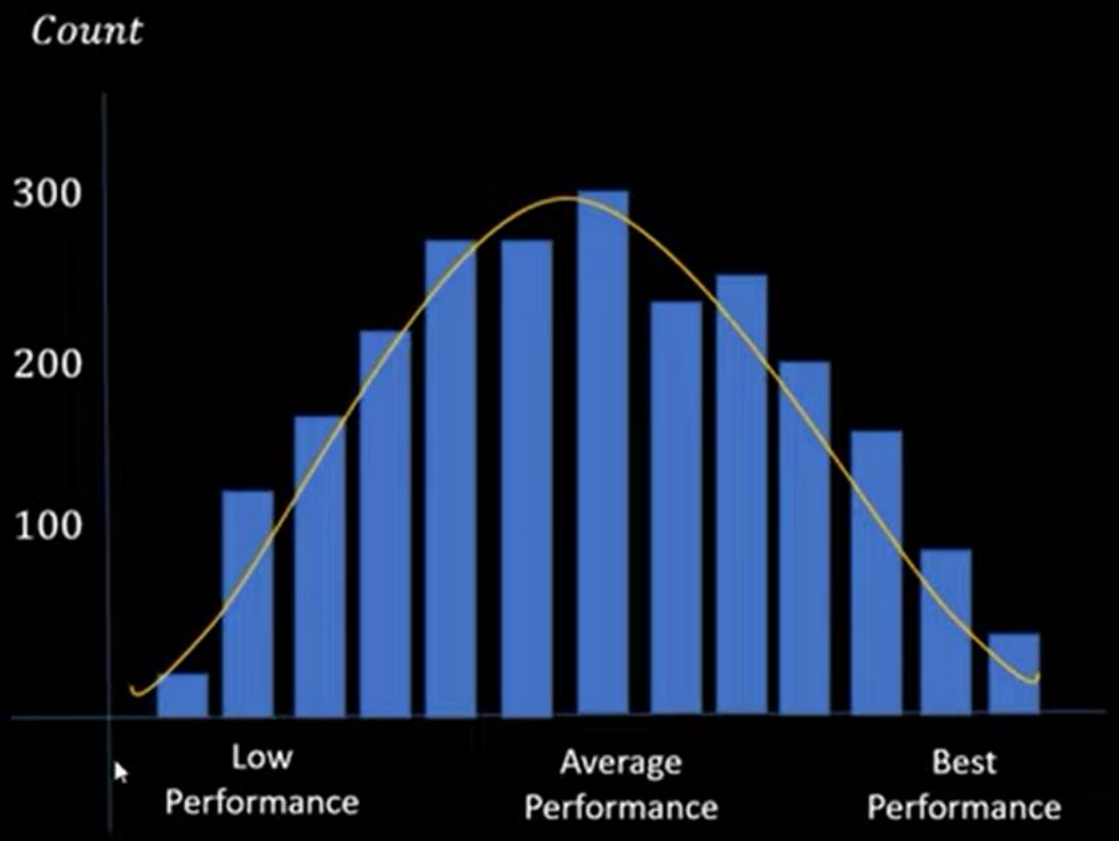




2 bedroom apartment price in Bangalore city



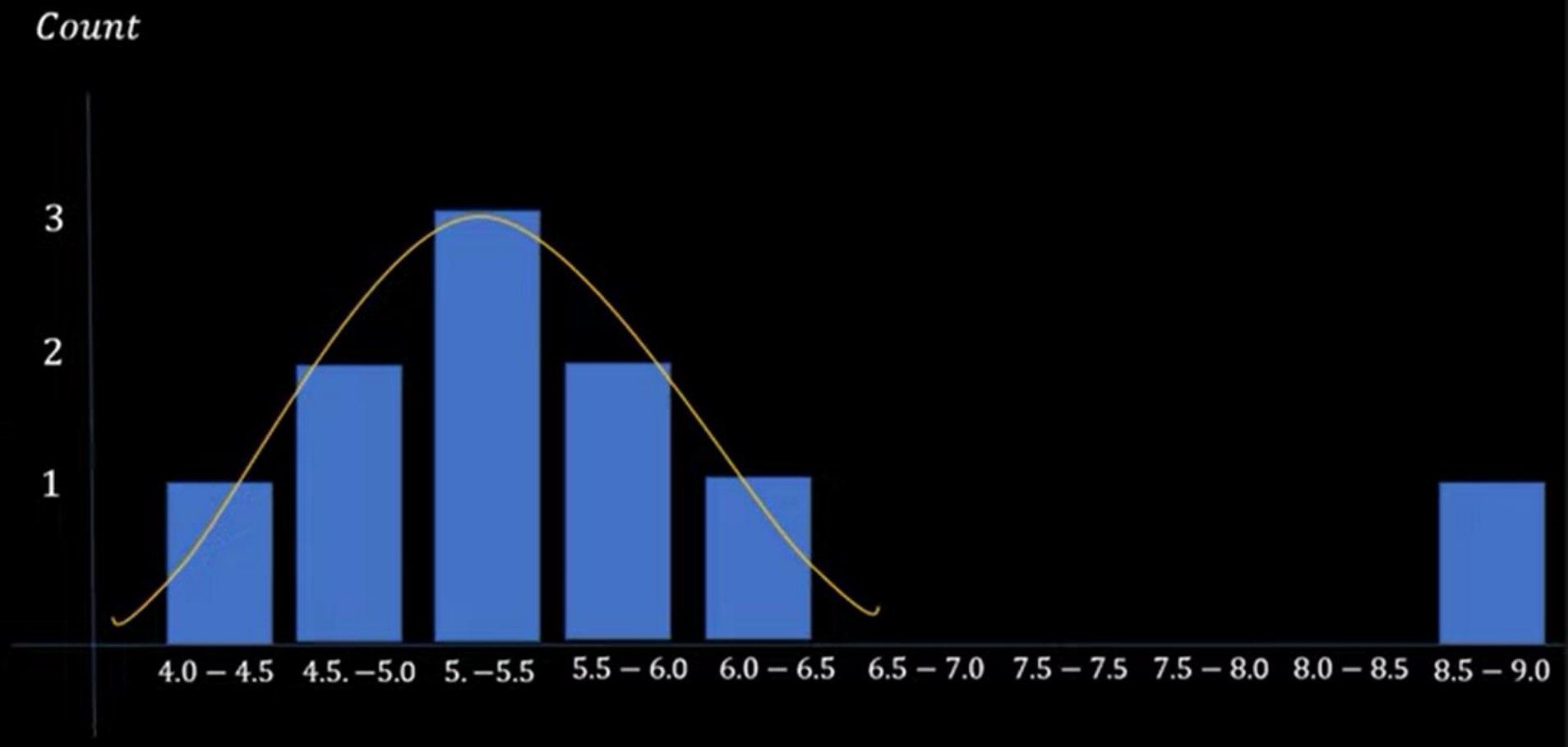
Test score out of 100



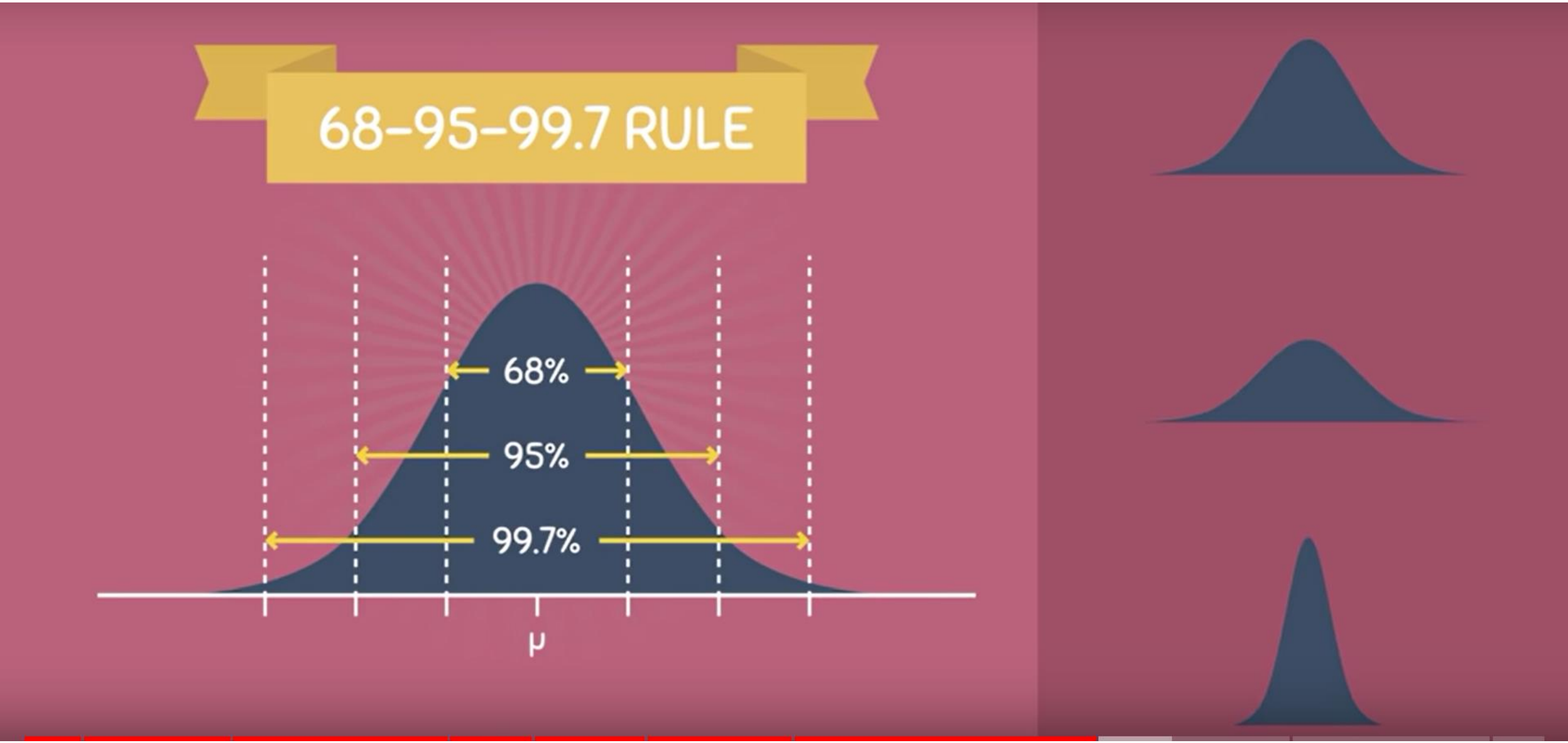
Employee Performance

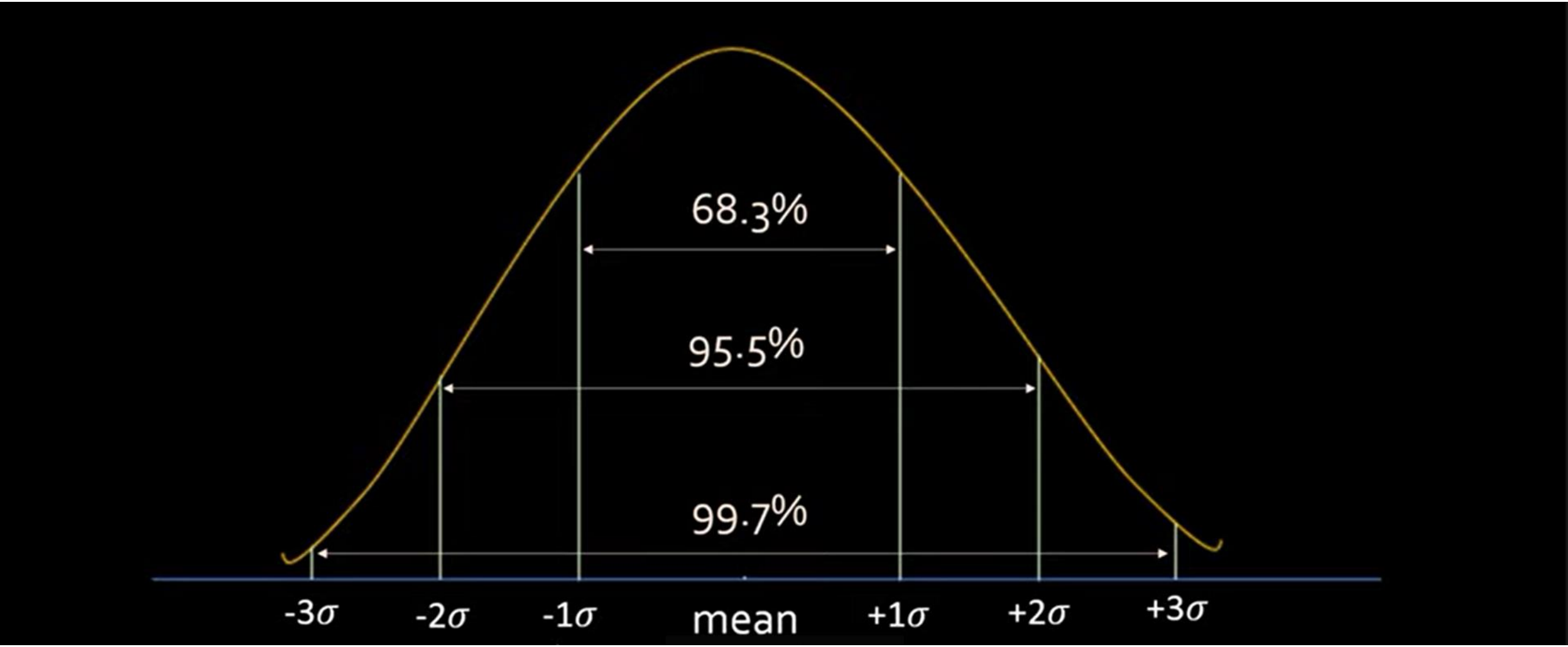
If there is an outlier

Name	Height (ft)
Rob	6.2
Thomas	5.7
Nina	4.6
Mittal	5.4
Sofia	5.9
Mohan	4.3
Tao	5.1
Deepika	5.2
Rafiq	4.9
Smith	9.0

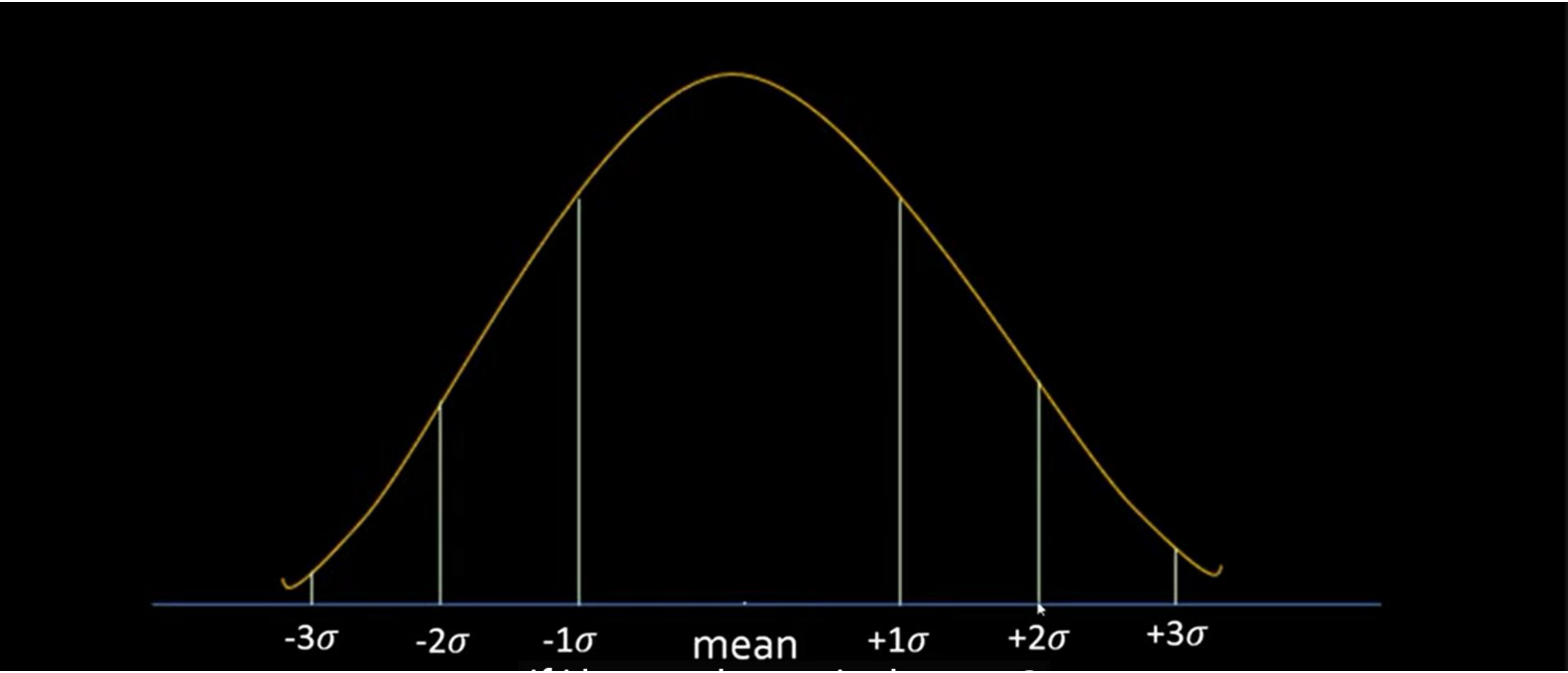


What formula do we use to remove outliers?


68-95-99.7 RULE
68%
95%
99.7%
μ


68.3%
95.5%
99.7%
-3σ
-2σ
-1σ
mean
+1σ
+2σ
+3σ

Z Score


-3σ
-2σ
-1σ
mean
+1σ
+2σ
+3σ

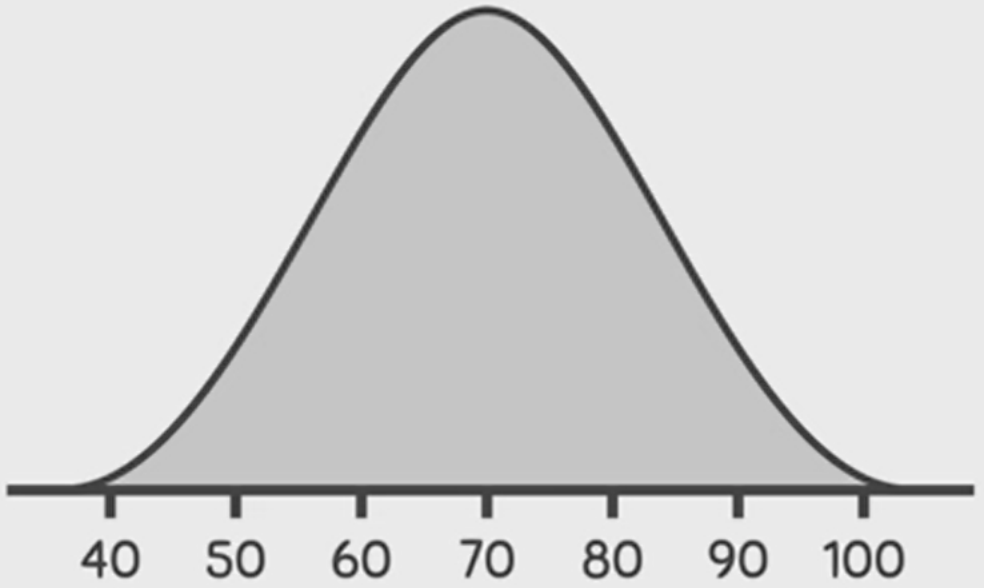
Z Score: How many standard deviation away a datapoint is from mean

Standard deviation=0.61

Name	Height (ft)	Z Score
Rob	6.2	(6.2-5.25) / 0.61 = 1.53
Thomas	5.7	(5.7-5.25) / 0.61 = 0.72
Nina	4.6	(4.6 – 5.25) / 0.61 = -1.06
Mittal	5.4	(5.4 – 5.25) / 0.61 = 0.23
Sofia	5.9	(5.9 – 5.25) / 0.61 = 1.04
Mohan	4.3	(4.3 – 5.25) / 0.61 = -1.55
Tao	5.1	(5.1 – 5.25) / 0.61 = -0.25
Deepika	5.2	(5.2 – 5.25) / 0.61 = -0.09
Rafiq	4.9	(4.9 – 5.25) / 0.61 = -0.58

$Average = 5.25$

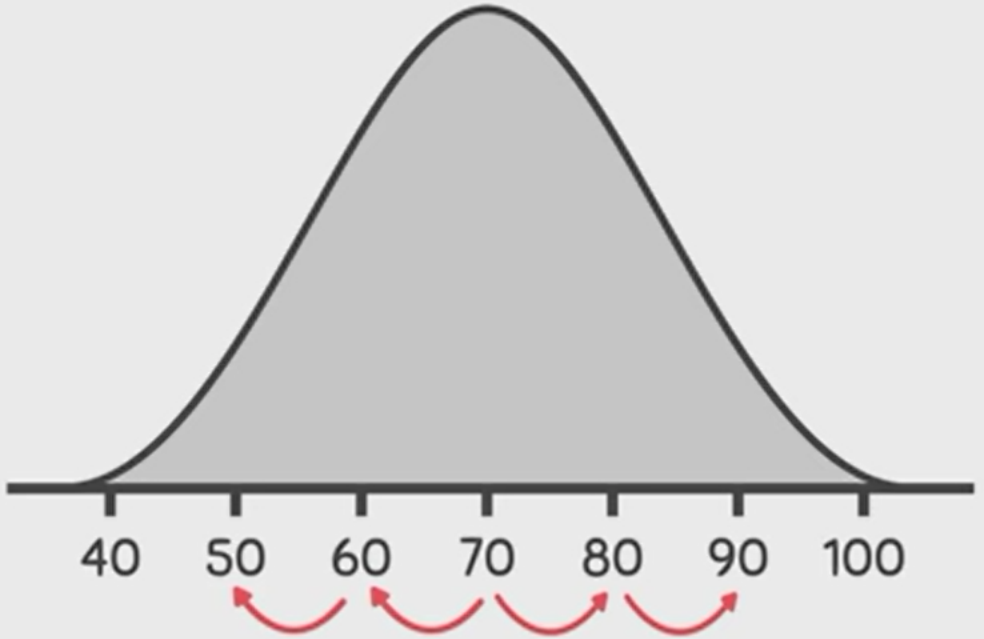
❶ The normal distribution below has a standard deviation of 10. Approximately what area is contained between 70 and 90?

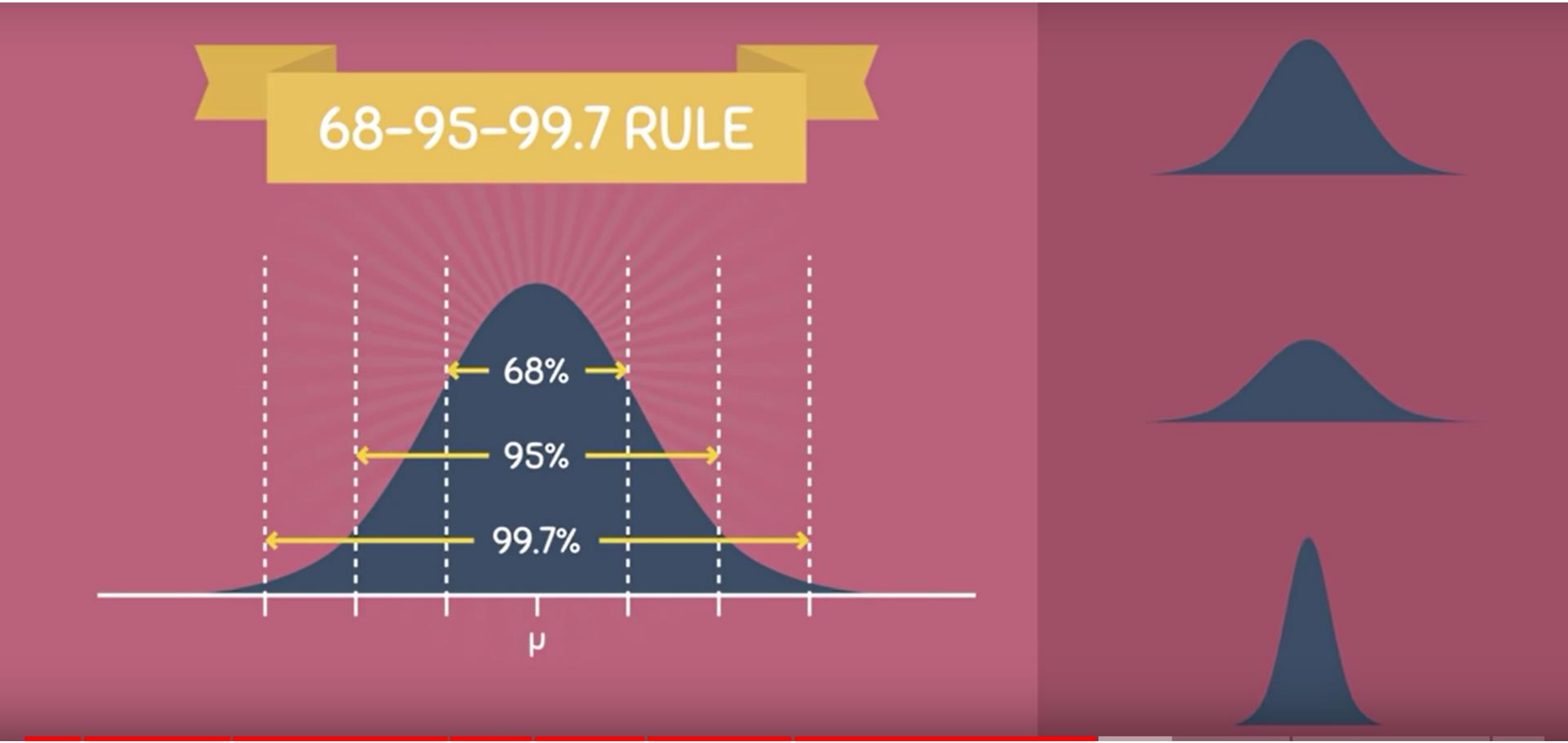


❶ The normal distribution below has a standard deviation of 10. Approximately what area is contained between 70 and 90?

$\mu = 70$

$\sigma = 10$

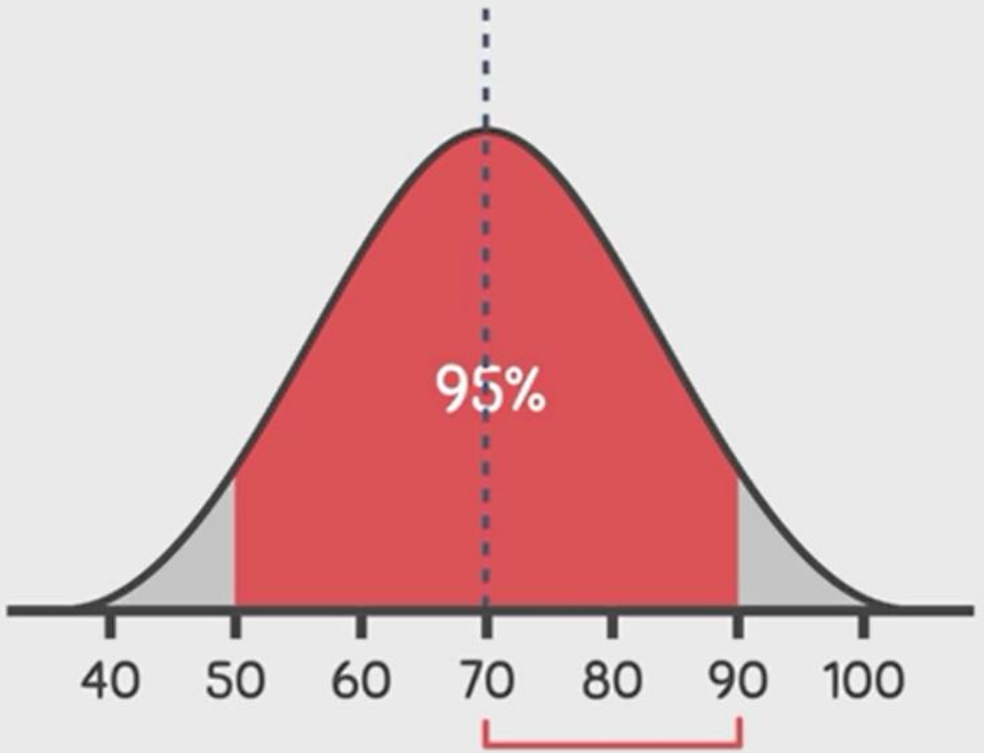



68-95-99.7 RULE
68%
95%
99.7%
μ

❶ The normal distribution below has a standard deviation of 10. Approximately what area is contained between 70 and 90?

$\mu = 70$

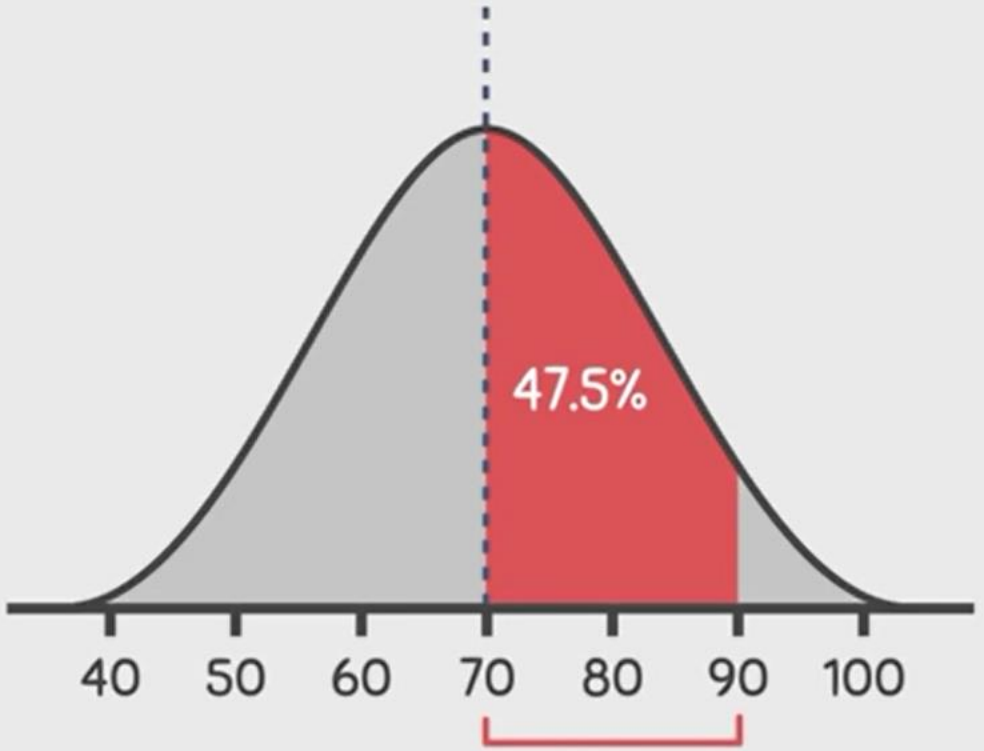
$\sigma = 10$



❶ The normal distribution below has a standard deviation of 10. Approximately what area is contained between 70 and 90?

$\mu = 70$

$\sigma = 10$

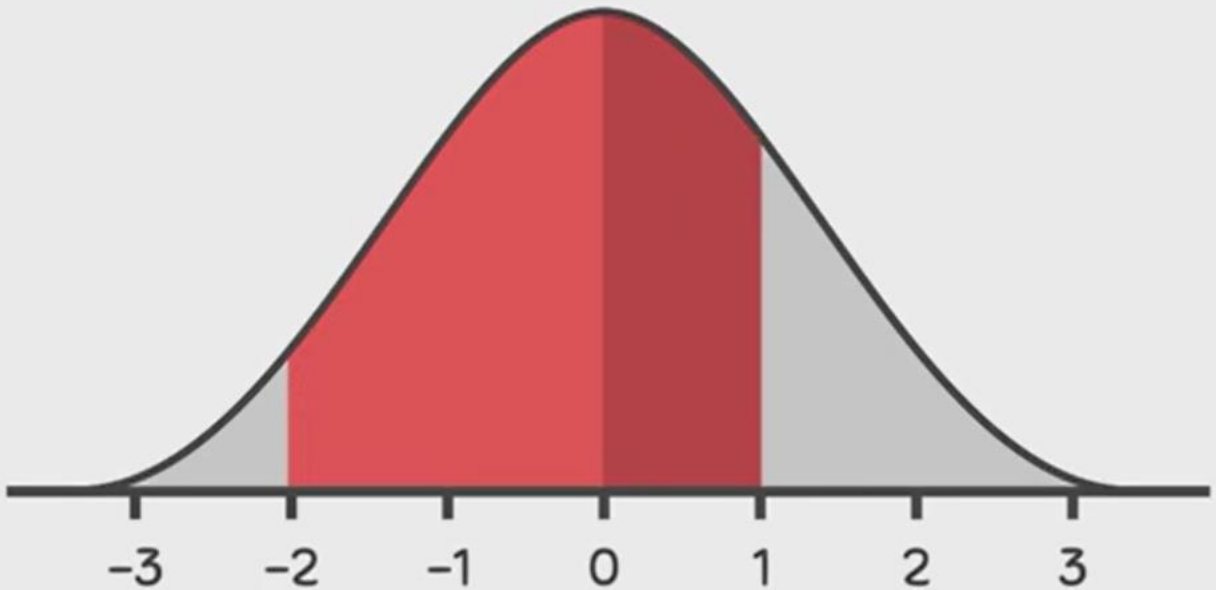


❷ For the normal distribution below, approximately what area is contained between −2 and 1?

❷ For the normal distribution below, approximately what area is contained between −2 and 1?

$\mu = 0$

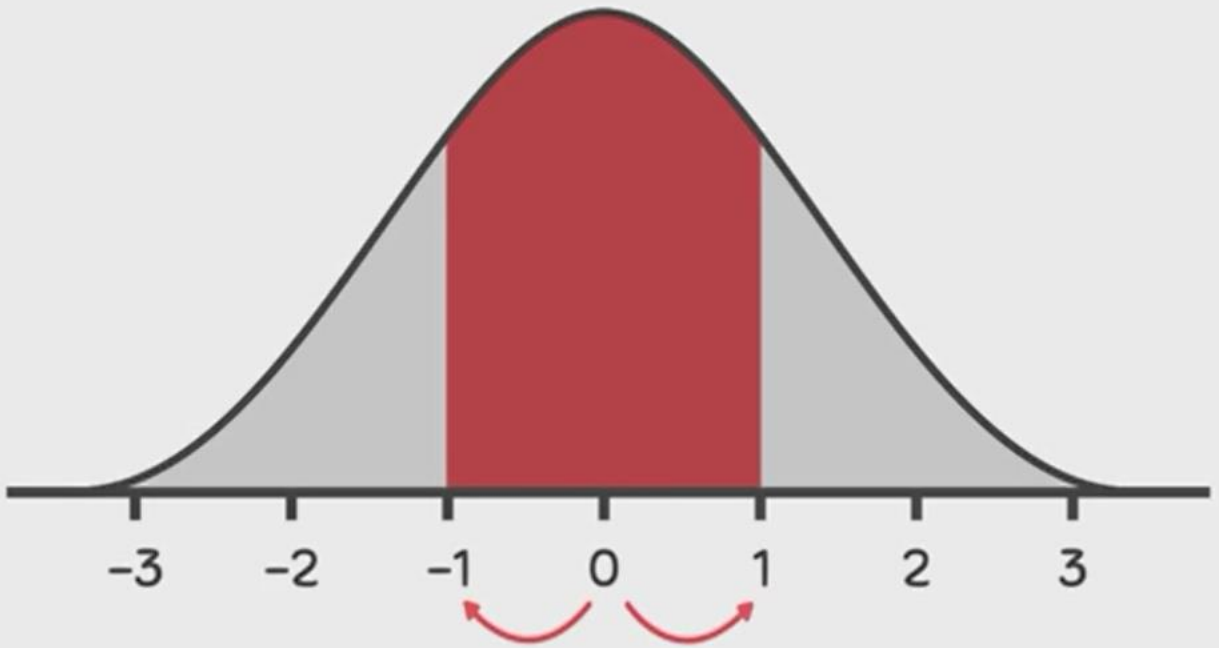
$\sigma = 1$



❷ For the normal distribution below, approximately what area is contained between −2 and 1?

$\mu = 0$

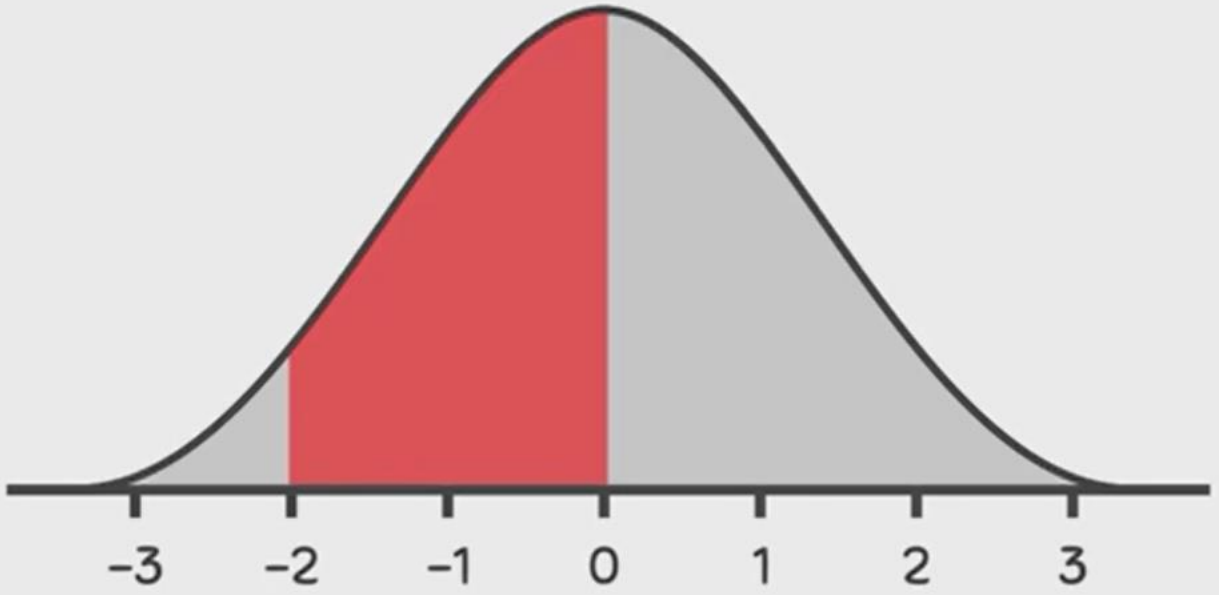
$\sigma = 1$



❷ For the normal distribution below, approximately what area is contained between −2 and 1?

$\mu = 0$

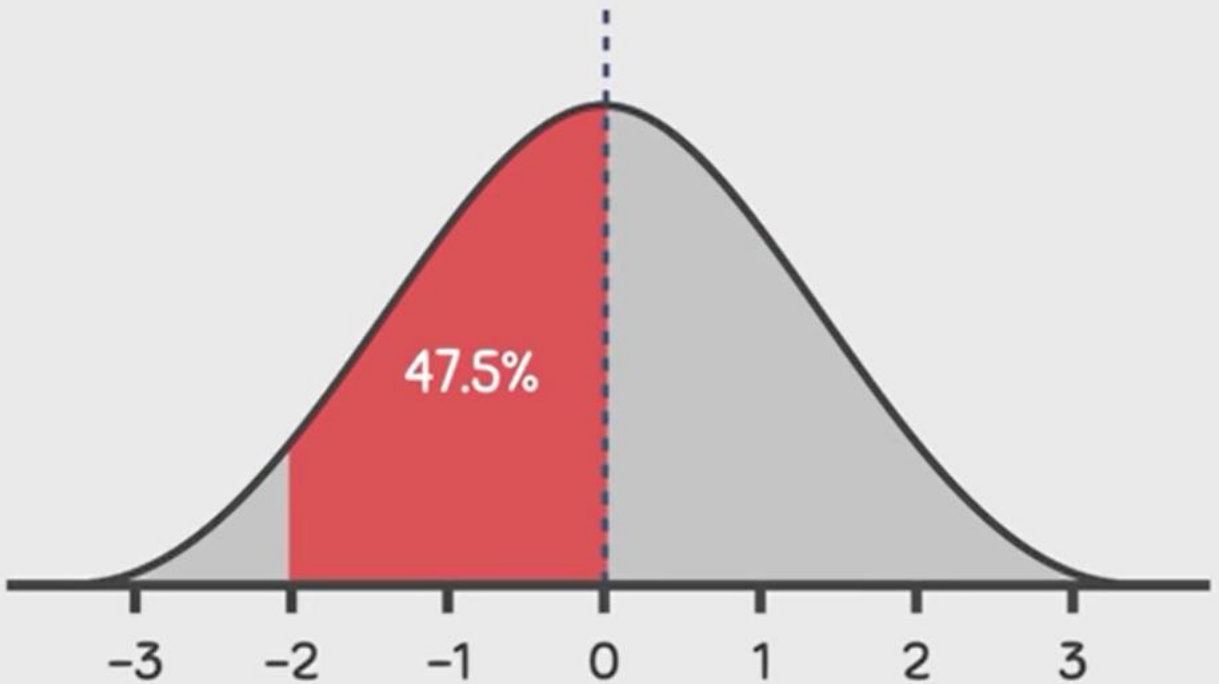
$\sigma = 1$



❷ For the normal distribution below, approximately what area is contained between −2 and 1?

$\mu = 0$

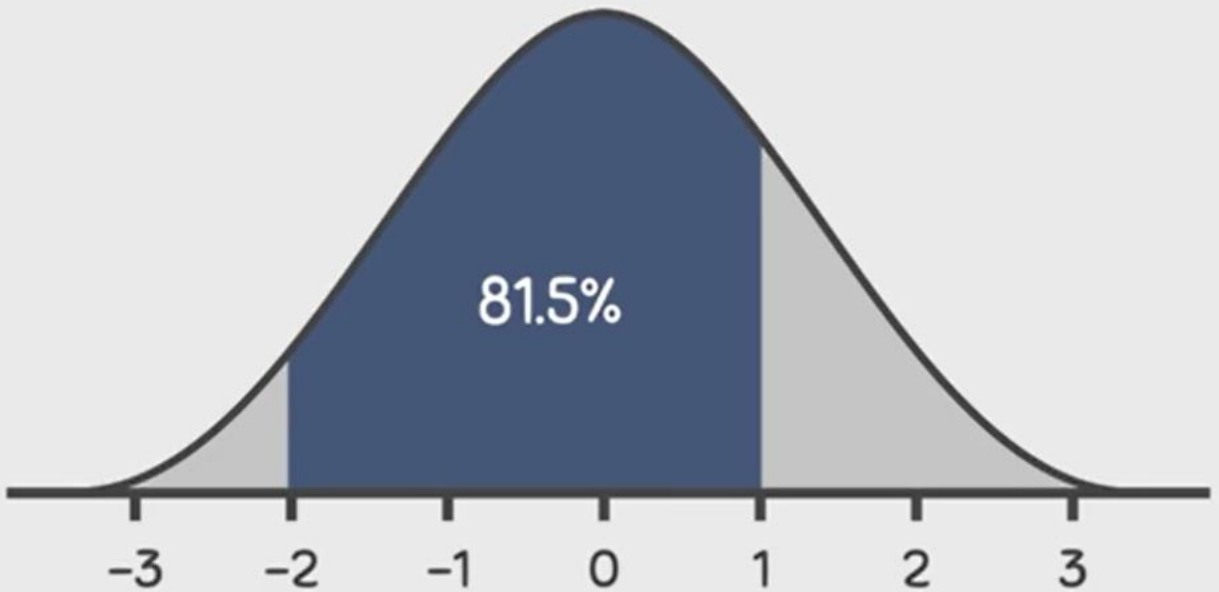
$\sigma = 1$



❷ For the normal distribution below, approximately what area is contained between −2 and 1?

$\mu = 0$

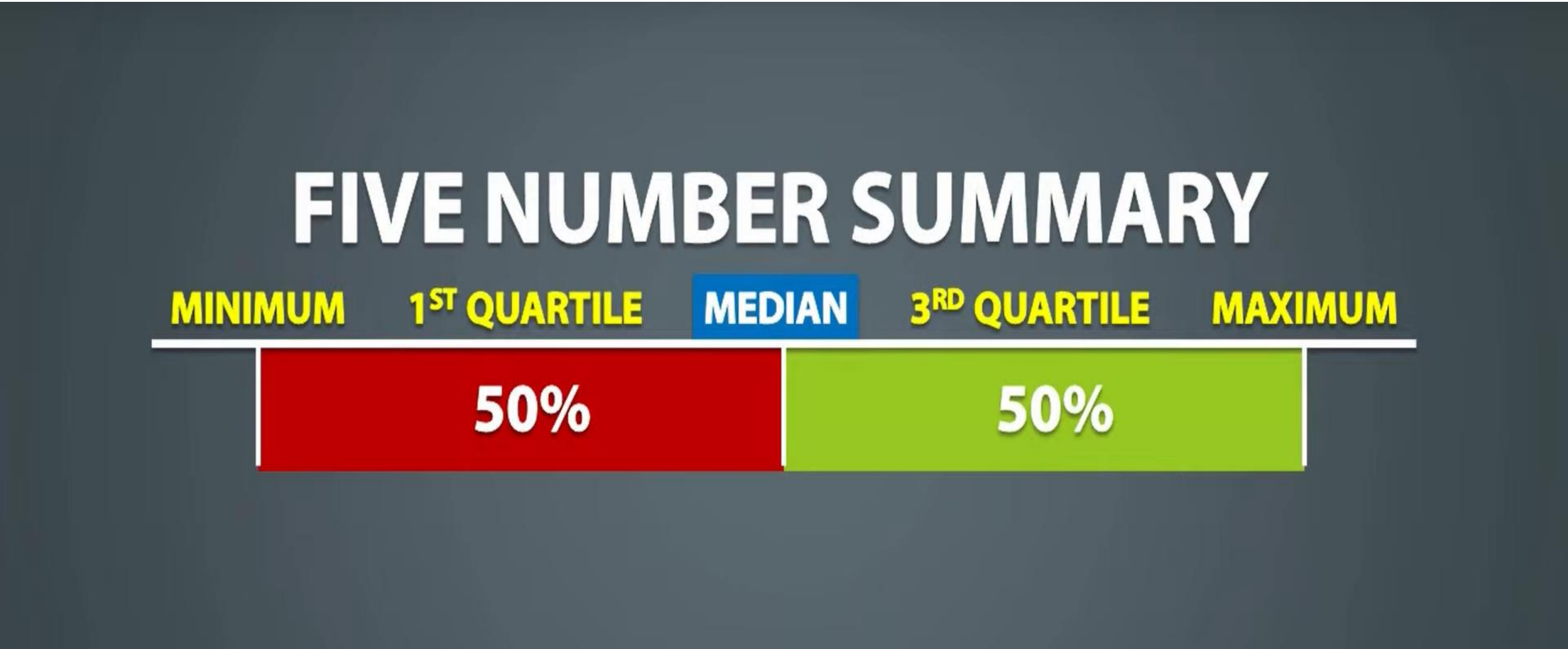
$\sigma = 1$



FIVE NUMBER SUMMARY

GIVES US A WAY TO DESCRIBE A DISTRIBUTION
USING ONLY FIVE NUMBERS


MINIMUM
1ST QUARTILE
MEDIAN
3RD QUARTILE
MAXIMUM
FIVE NUMBER SUMMARY
GIVES US A WAY TO DESCRIBE A DISTRIBUTION
USING ONLY FIVE NUMBERS


FIVE NUMBER SUMMARY
MINIMUM
1ST QUARTILE
MEDIAN
3RD QUARTILE
MAXIMUM
50%
50%

FIVE NUMBER SUMMARY

MINIMUM | 1ST QUARTILE | MEDIAN | 3RD QUARTILE | MAXIMUM

33

8

10 11 12 25 25 27 31 **33** 34 34 35 36 43 50 59

POSITION OF MEDIAN = 8

FIVE NUMBER SUMMARY

MINIMUM	1ST QUARTILE	MEDIAN	3RD QUARTILE	MAXIMUM
		33		

10 11 12 25 25 27 31 **33** 34 34 35 36 43 50 59

$$\text{POSITION OF Q1} = \frac{n+1}{2}$$

"n" refers to the number of data values BELOW the median

FIVE NUMBER SUMMARY

MINIMUM | 1ST QUARTILE | MEDIAN | 3RD QUARTILE | MAXIMUM

33

10 11 12 25 25 27 31 **33** 34 34 35 36 43 50 59

$$\text{POSITION OF Q1} = \frac{7 + 1}{2}$$

"n" refers to the number of data values BELOW the median

FIVE NUMBER SUMMARY

MINIMUM | 1ST QUARTILE | MEDIAN | 3RD QUARTILE | MAXIMUM

33

4

10 11 12 25 25 27 31 33 34 34 35 36 43 50 59

POSITION OF Q1 = 4

FIVE NUMBER SUMMARY

MINIMUM	1ST QUARTILE	MEDIAN	3RD QUARTILE	MAXIMUM
	25	33		

10 11 12 **25** 25 27 31 **33** 34 34 35 36 43 50 59

$$\text{POSITION OF Q3} = \frac{n+1}{2}$$

"n" refers to the number of data values that are ABOVE the median

FIVE NUMBER SUMMARY

MINIMUM | 1ST QUARTILE | MEDIAN | 3RD QUARTILE | MAXIMUM

25 | 33

4

10 11 12 **25** 25 27 31 **33** 34 34 35 **36** 43 50 59

POSITION OF Q3 = 4

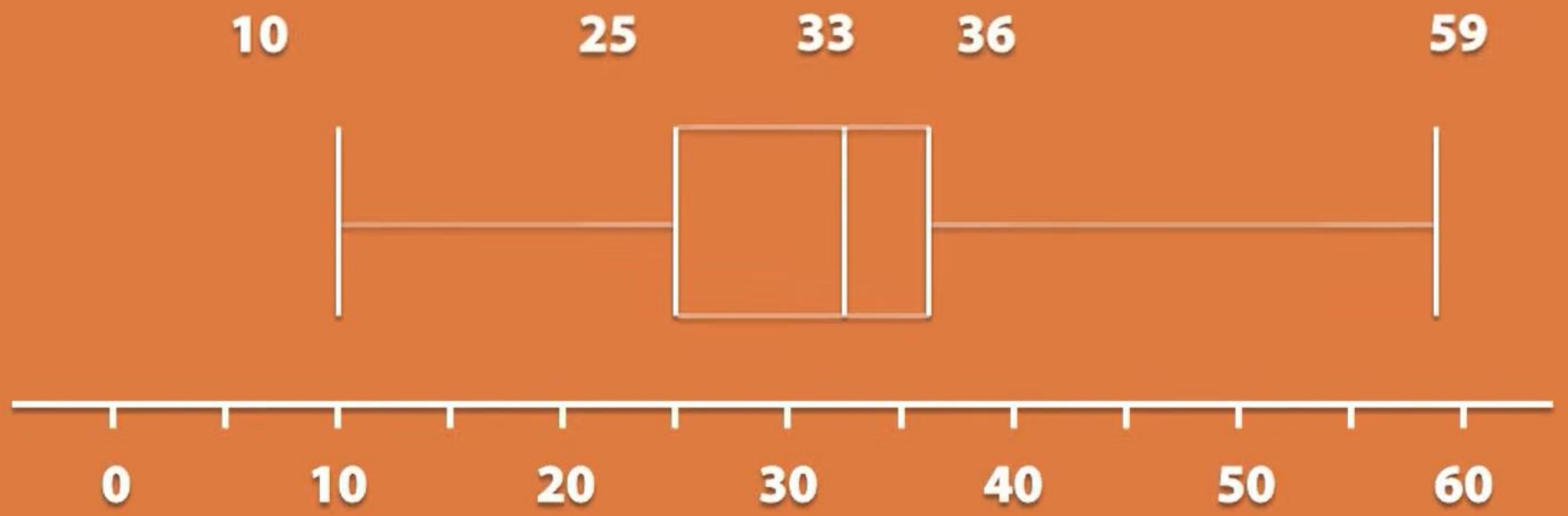
FIVE NUMBER SUMMARY

MINIMUM	1ST QUARTILE	MEDIAN	3RD QUARTILE	MAXIMUM
	25	33	36	

10 11 12 **25** 25 27 31 **33** 34 34 35 **36** 43 50 **59**

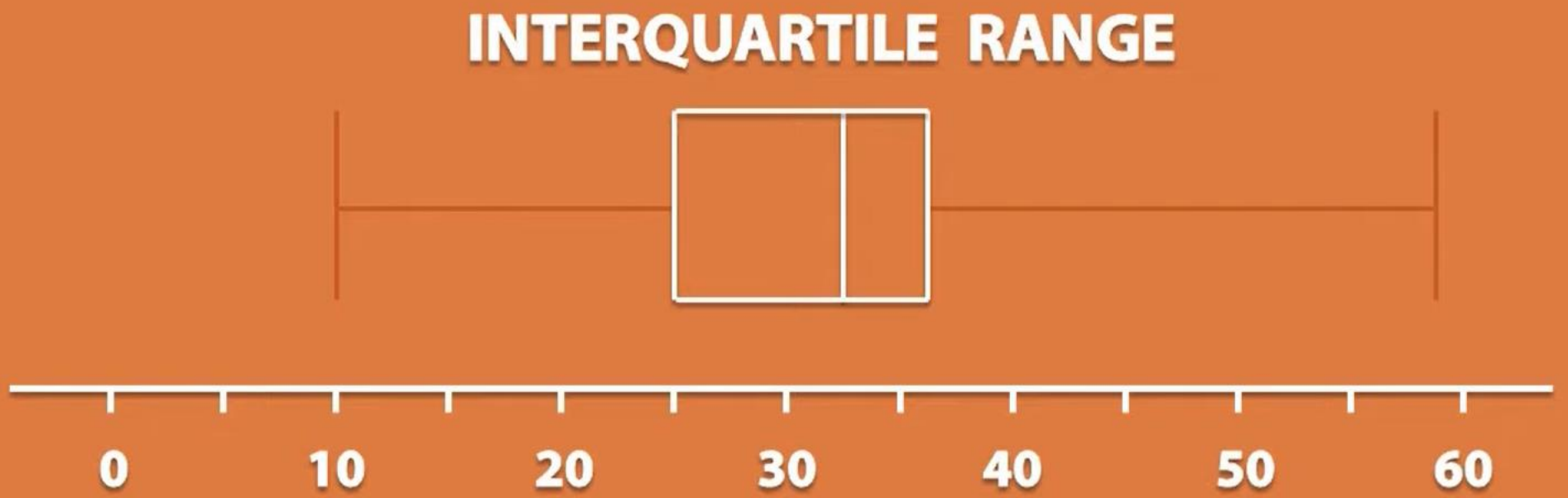
BOXPLOT

GIVES US A VISUAL REPRESENTATION OF THE FIVE NUMBER SUMMARY

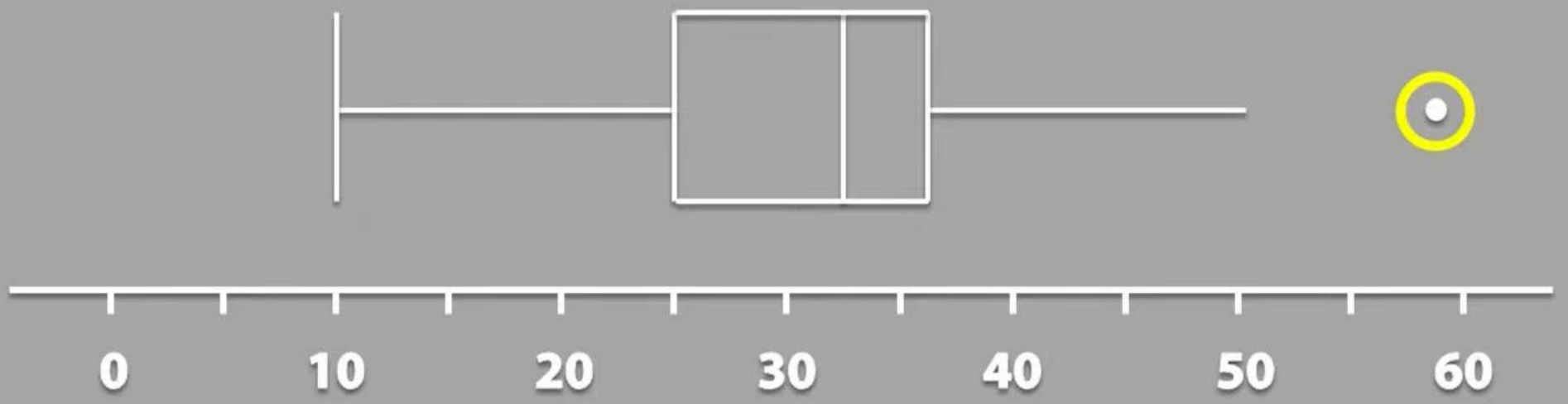


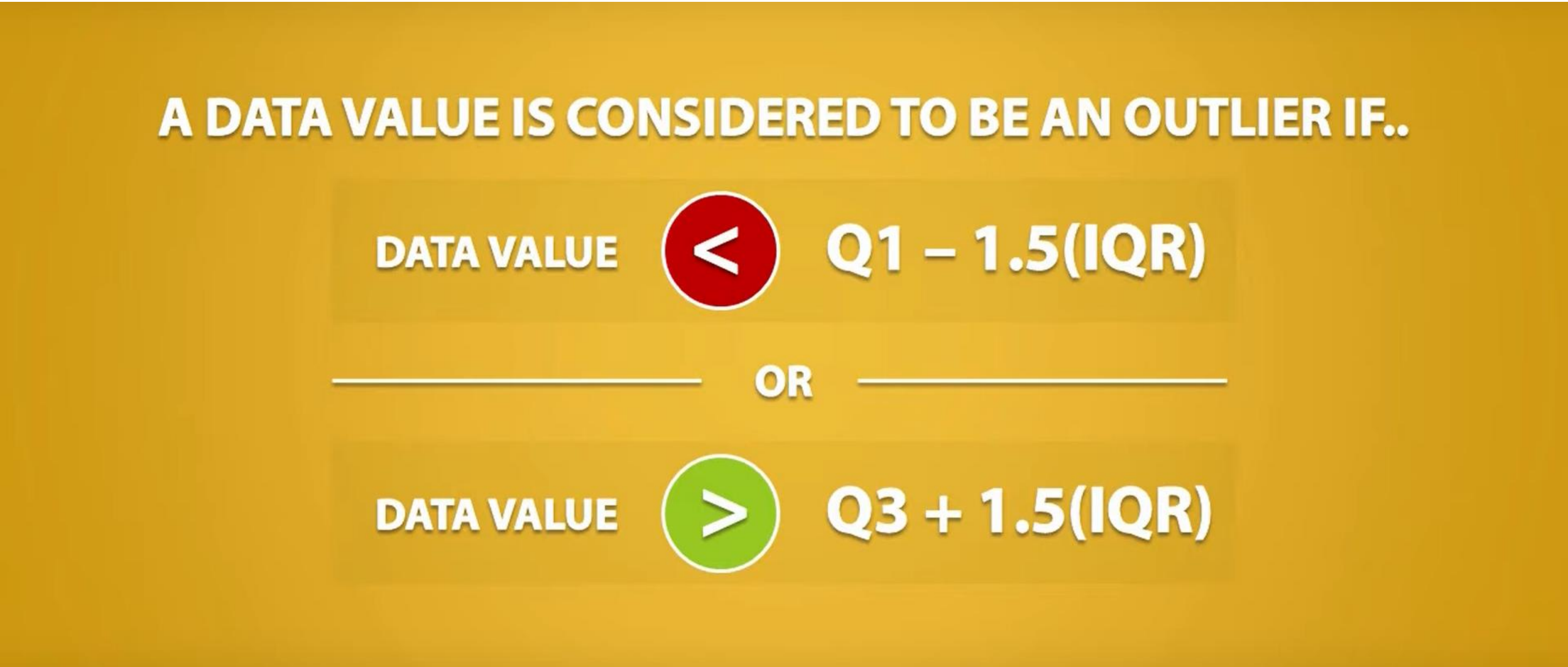
BOXPLOT

GIVES US A VISUAL REPRESENTATION OF THE FIVE NUMBER SUMMARY

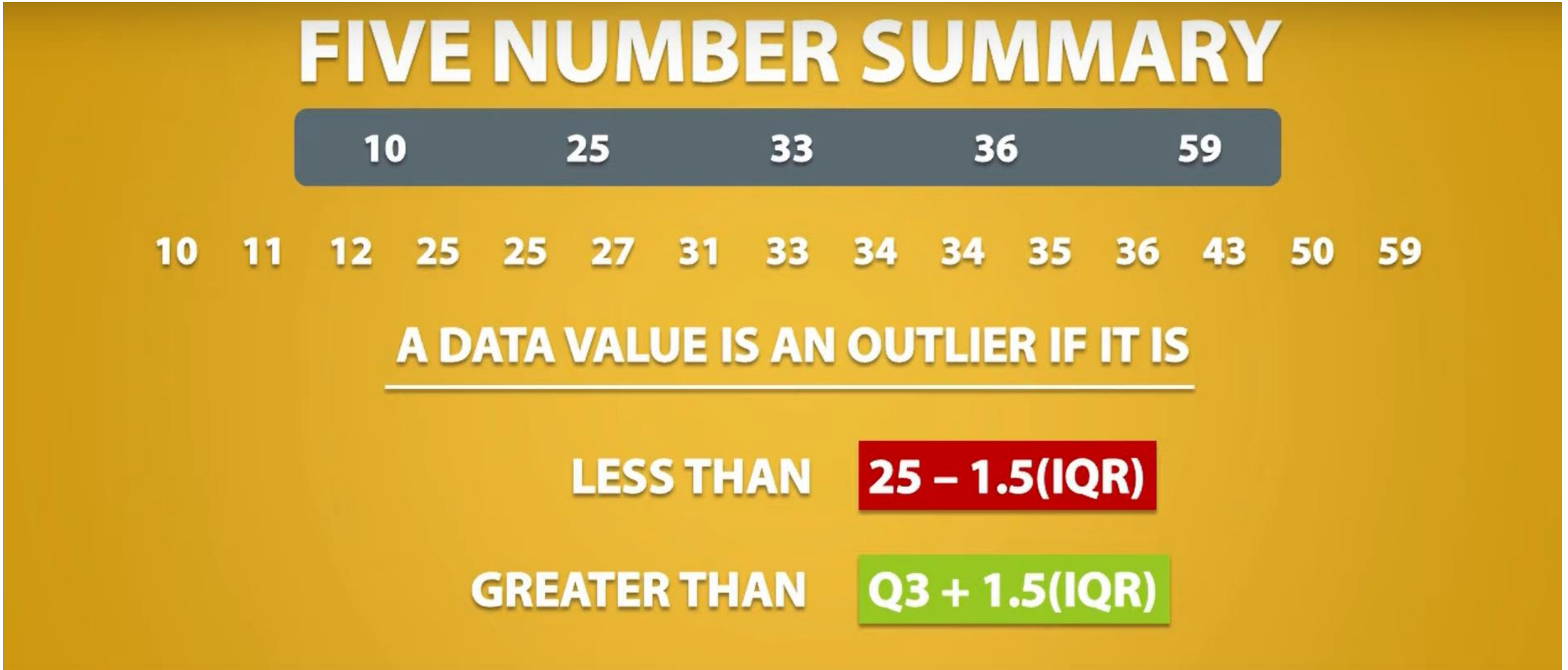


MODIFIED BOXPLOT




A DATA VALUE IS CONSIDERED TO BE AN OUTLIER IF..
DATA VALUE < Q1 – 1.5(IQR)
OR
DATA VALUE > Q3 + 1.5(IQR)

IQR=11


FIVE NUMBER SUMMARY
10
25
33
36
59
10 11 12 25 25 27 31 33 34 34 35 36 43 50 59
A DATA VALUE IS AN OUTLIER IF IT IS
LESS THAN
25 – 1.5(IQR)
GREATER THAN
Q3 + 1.5(IQR)

FIVE NUMBER SUMMARY

10	25	33	36	59

10 11 12 25 25 27 31 33 34 34 35 36 43 50 59

A DATA VALUE IS AN OUTLIER IF IT IS

LESS THAN $25 - 1.5(11)$

GREATER THAN $36 + 1.5(11)$

FIVE NUMBER SUMMARY

10 25 33 36 59

10 11 12 25 25 27 31 33 34 34 35 36 43 50 59

A DATA VALUE IS AN OUTLIER IF IT IS

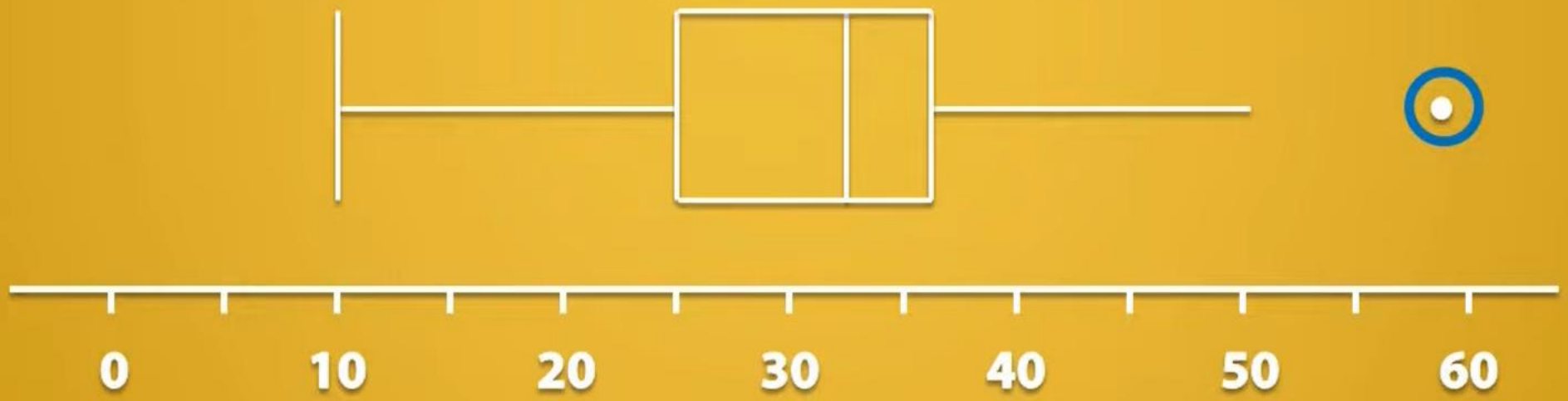
LESS THAN **8.5**

GREATER THAN **52.5**

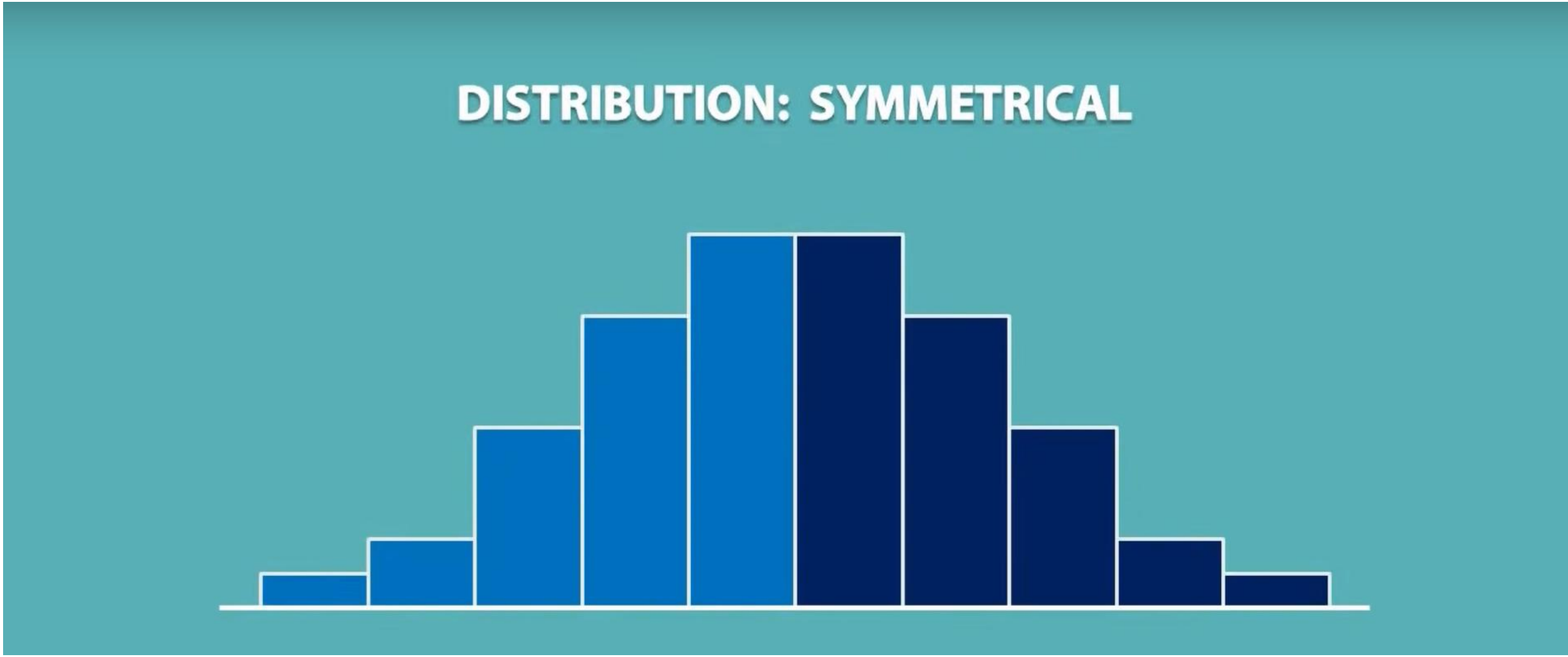
FIVE NUMBER SUMMARY

10 25 33 36 59

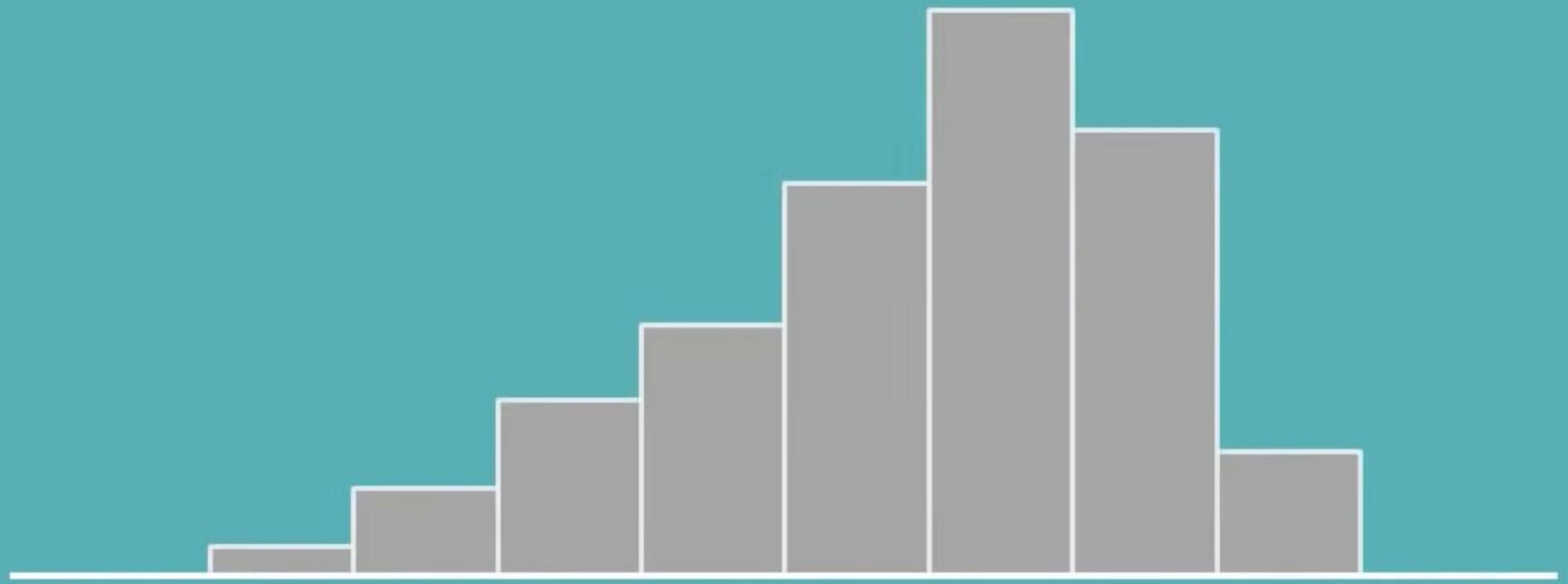
10 11 12 25 25 27 31 33 34 34 35 36 43 50 59



Symmetry and skewness



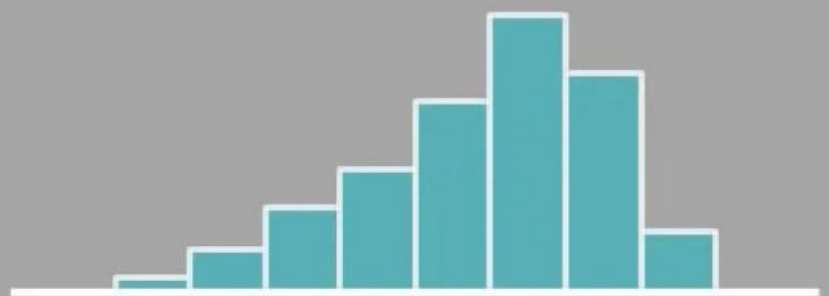
DISTRIBUTION: SKEWED



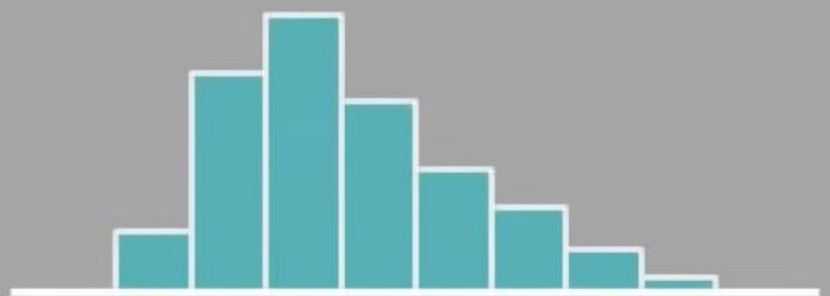
SKEWNESS

REFERS TO ASYMMETRY

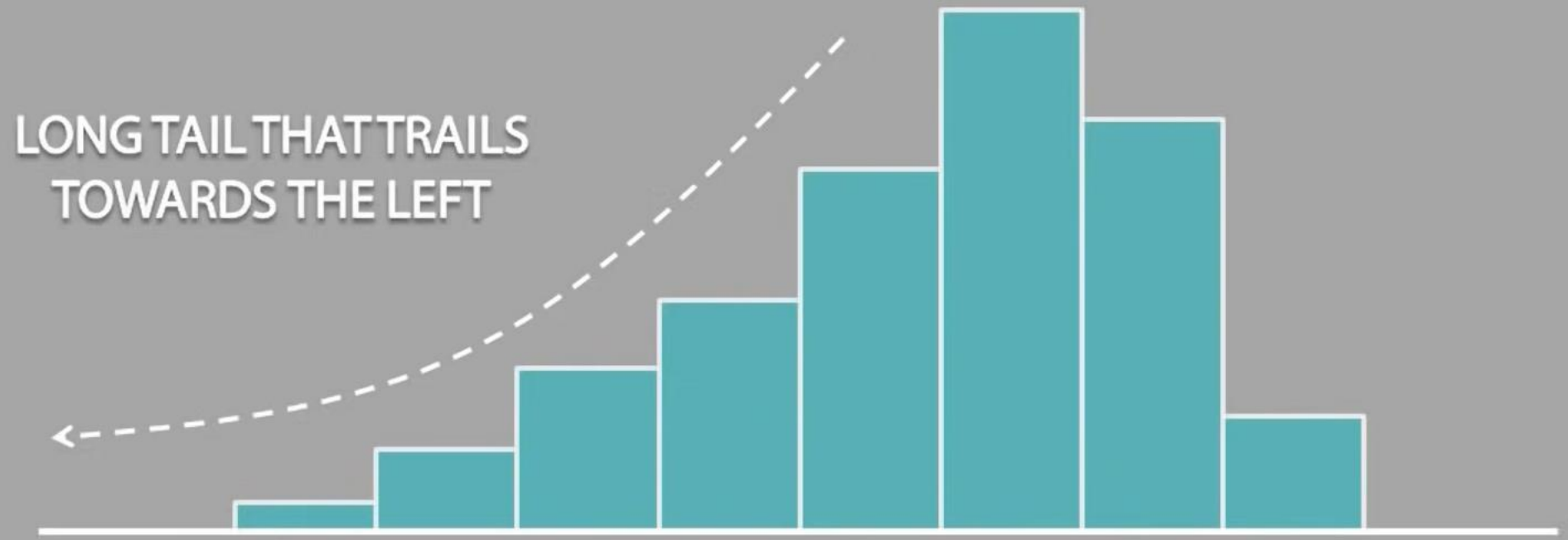
SKEWED TO THE LEFT



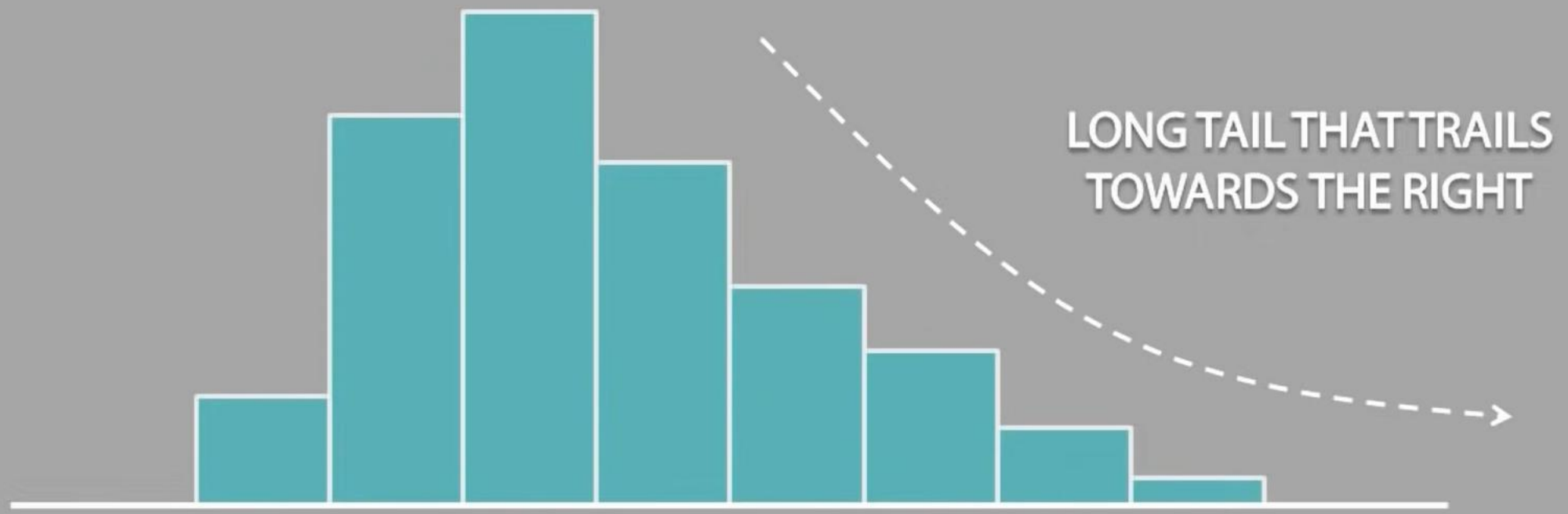
SKEWED TO THE RIGHT



SKEWED TO THE LEFT

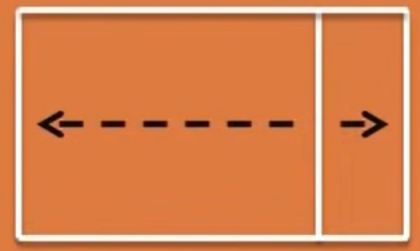


SKEWED TO THE RIGHT



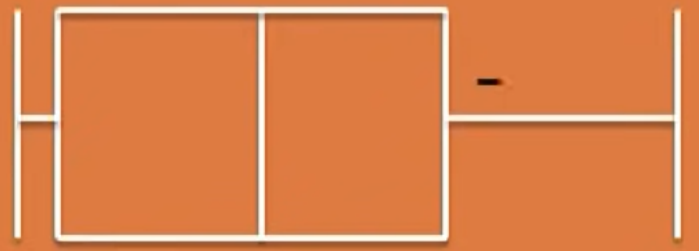
STRATEGIES FOR DETERMINING THE SKEWNESS FOR A BOXPLOT

UNEQUAL BOXES

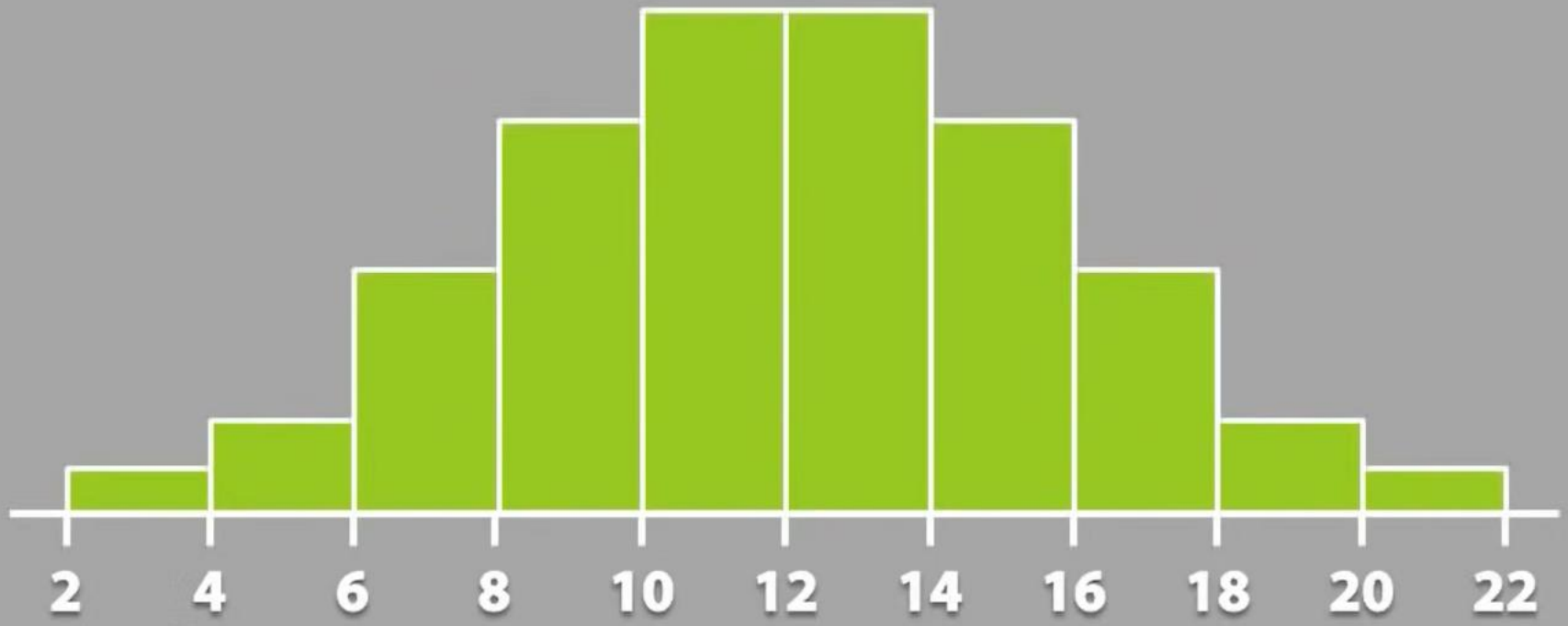


SKEWED TO THE LEFT

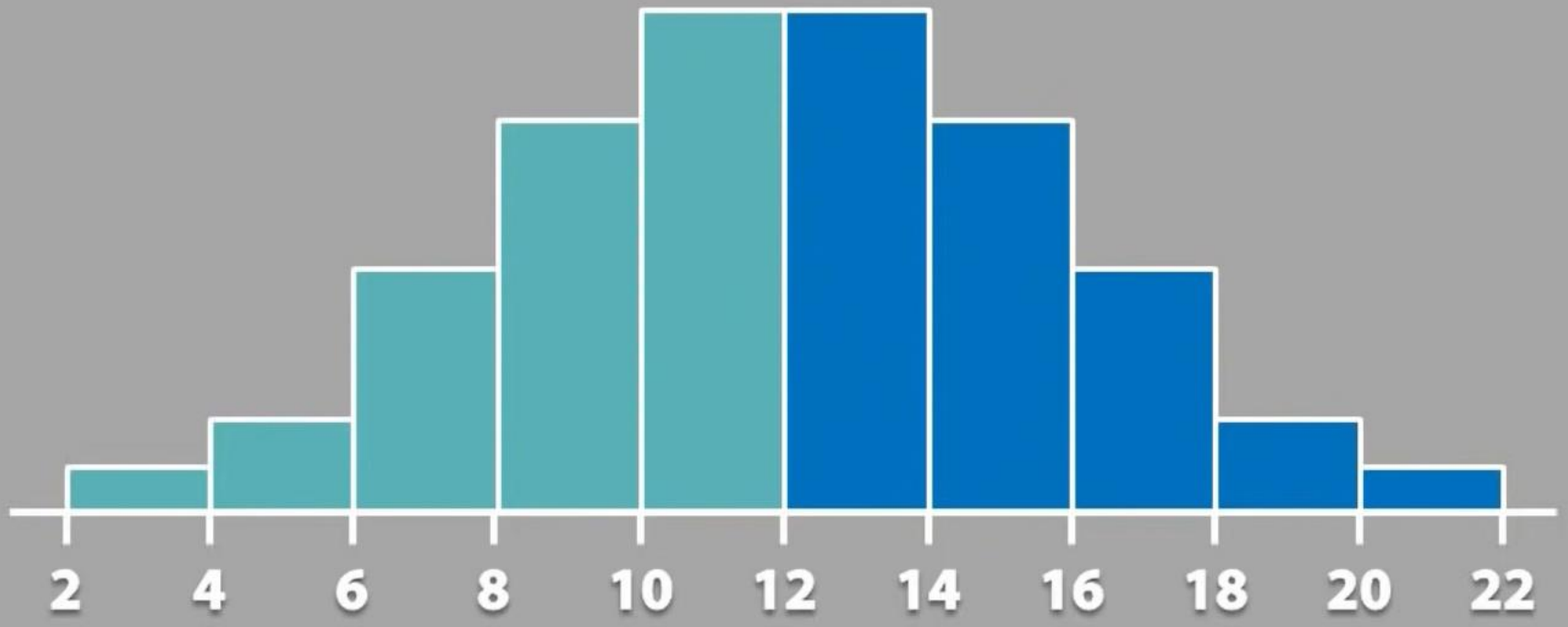
EQUAL BOXES



DISTRIBUTION: SYMMETRICAL

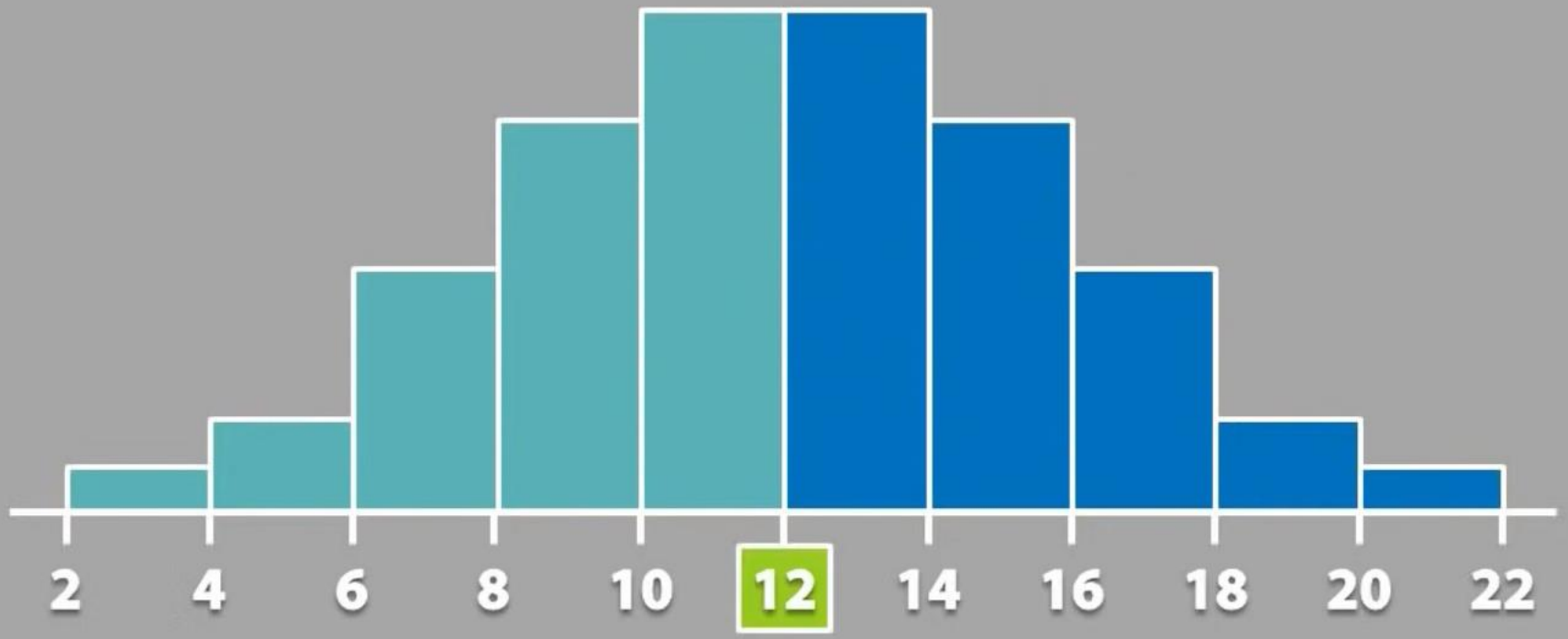


DISTRIBUTION: SYMMETRICAL

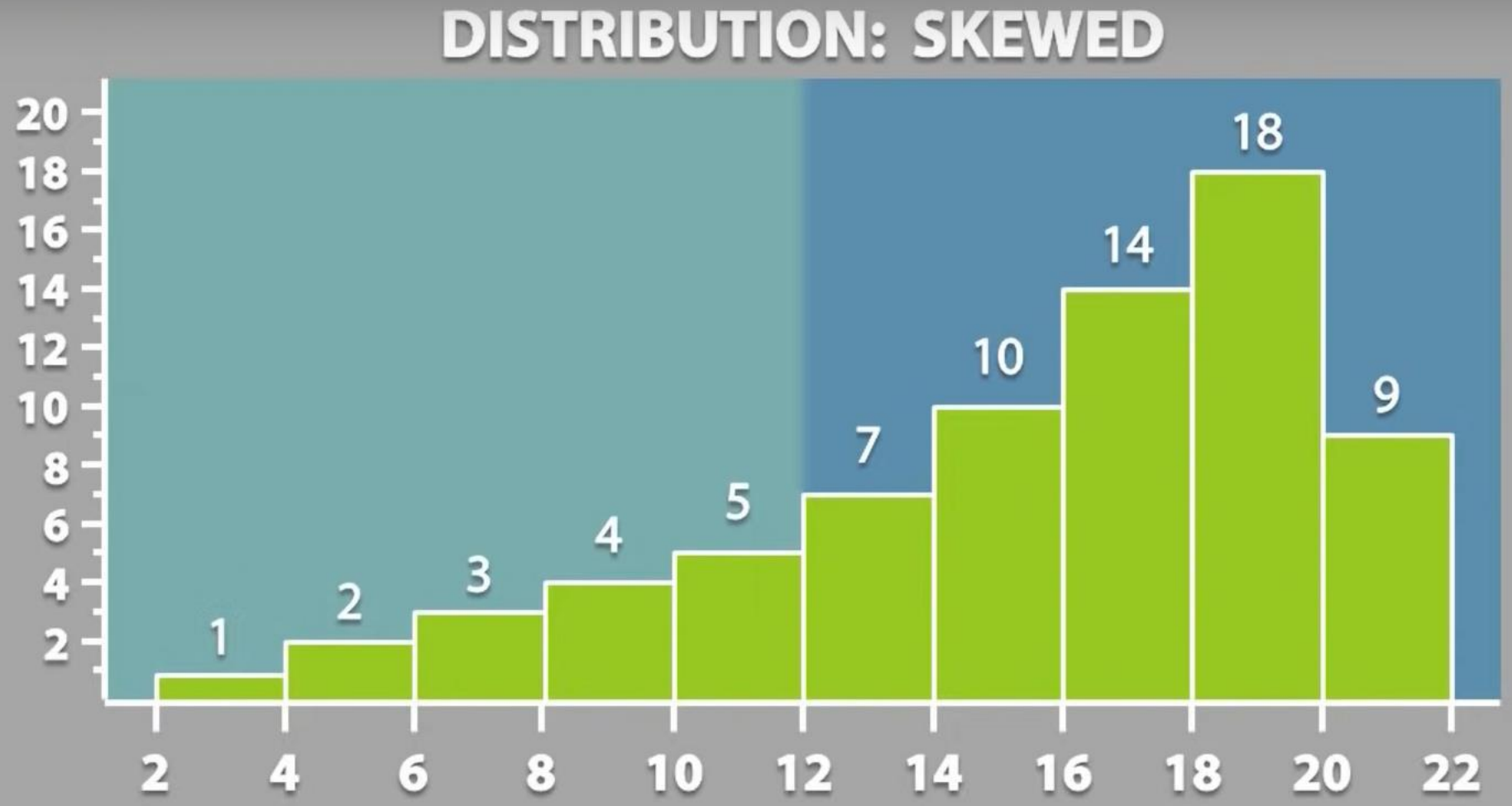


DISTRIBUTION: SYMMETRICAL

MEAN = MEDIAN

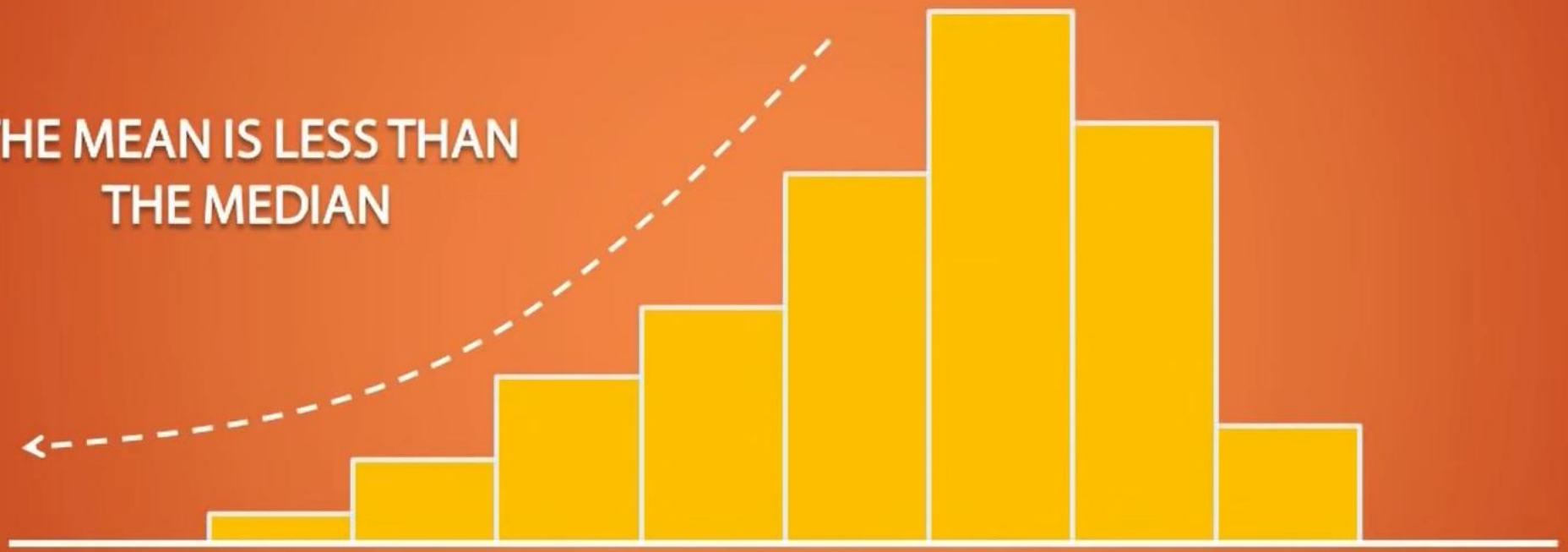


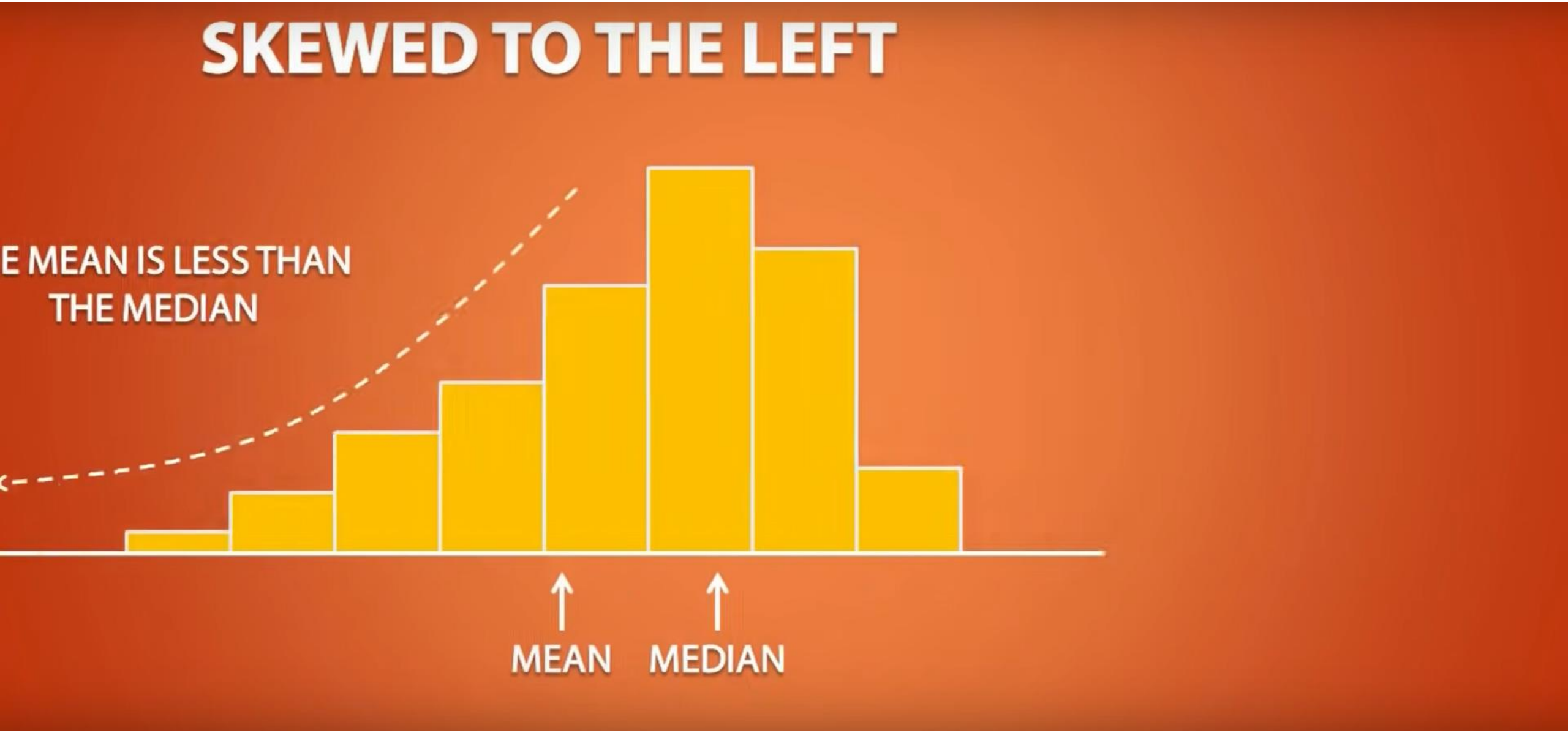
If skewed-Right hand side we have more data values than left hand



SKEWED TO THE LEFT

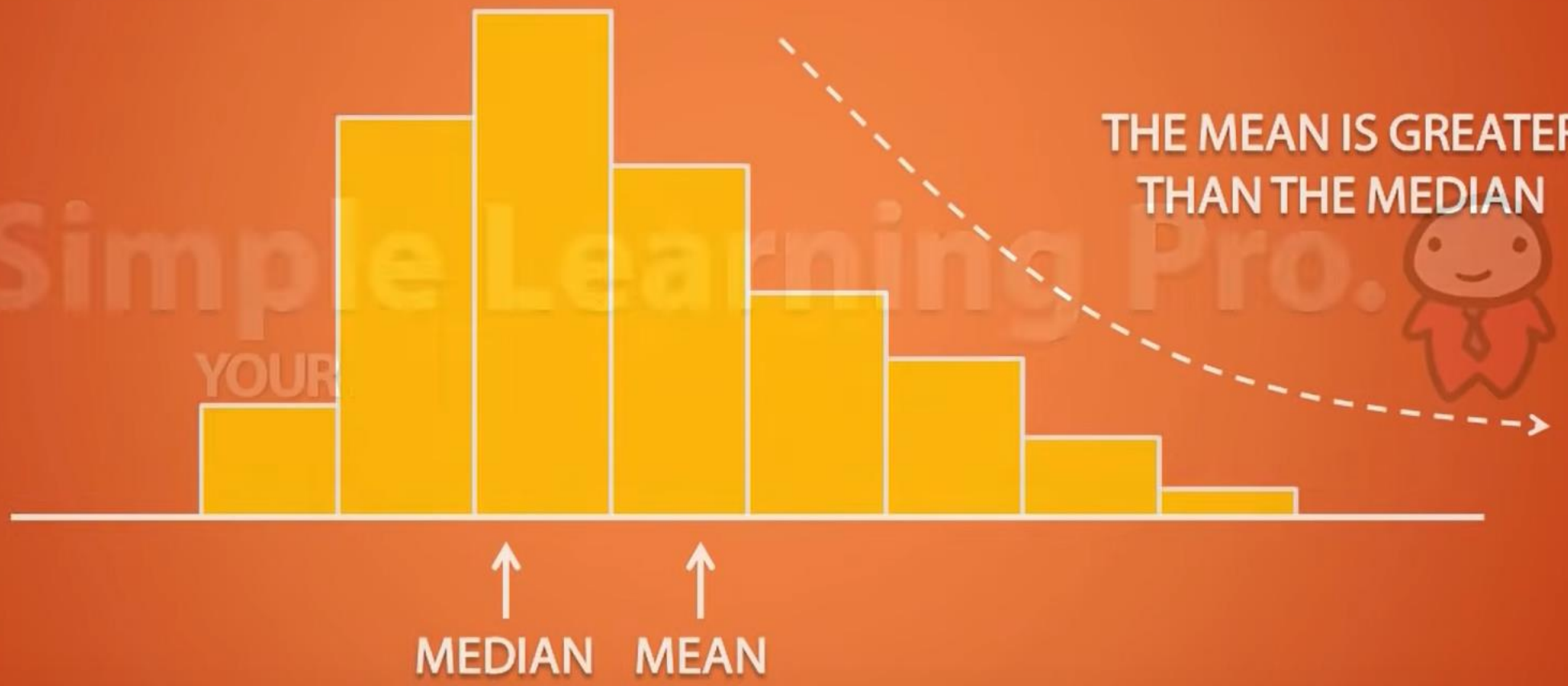
THE MEAN IS LESS THAN
THE MEDIAN

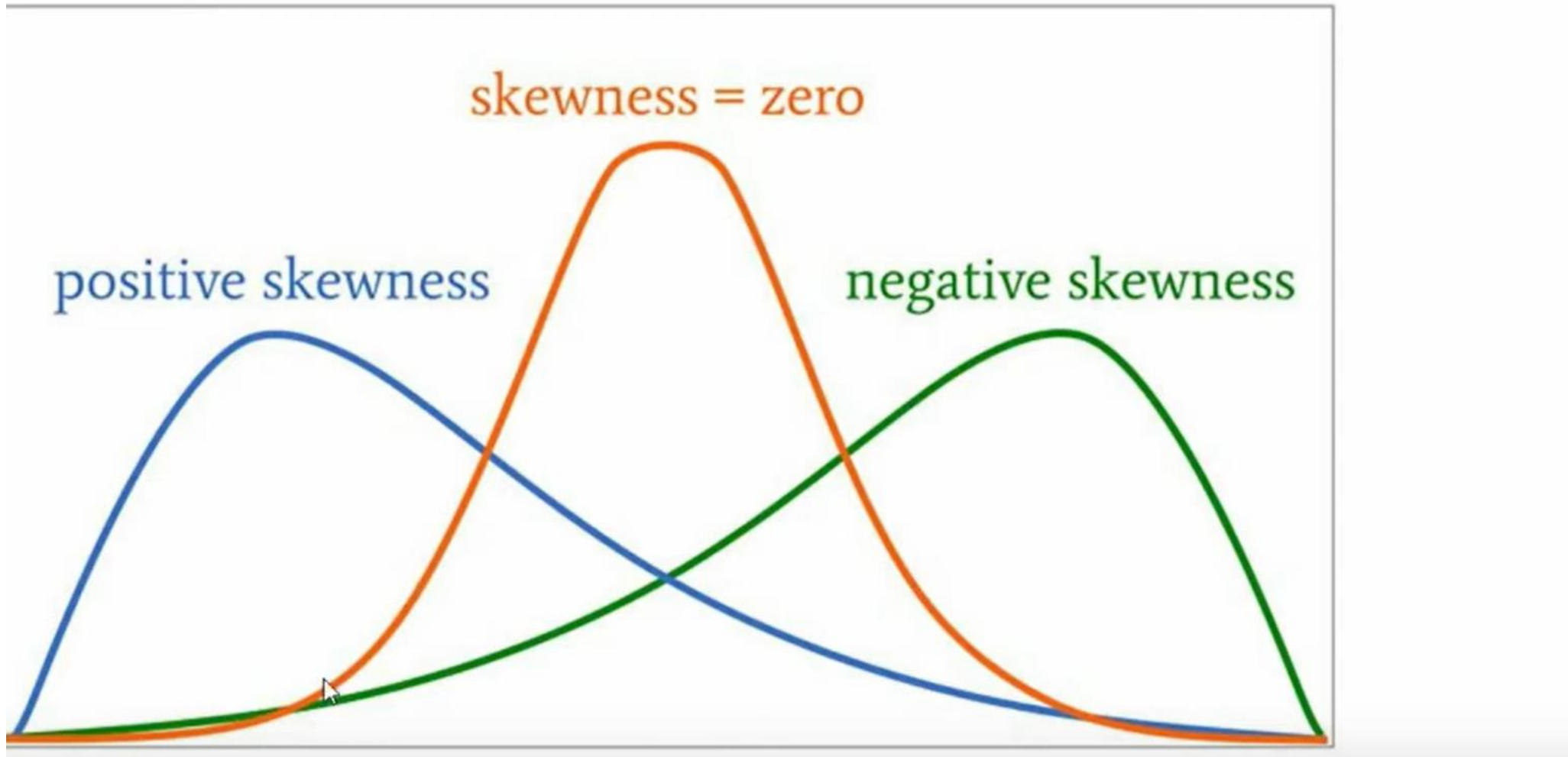


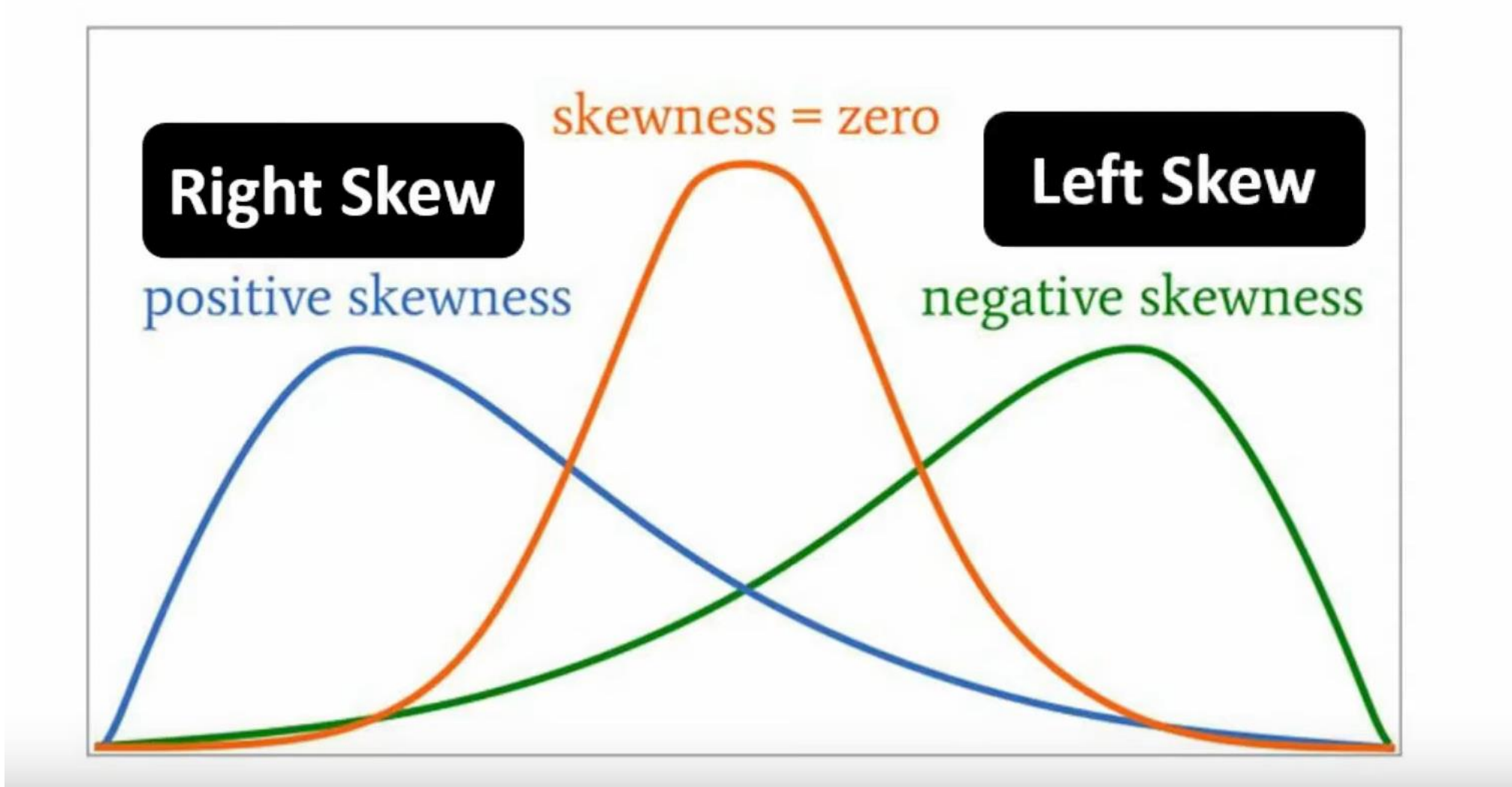

SKEWED TO THE LEFT
E MEAN IS LESS THAN
THE MEDIAN
MEAN
MEDIAN

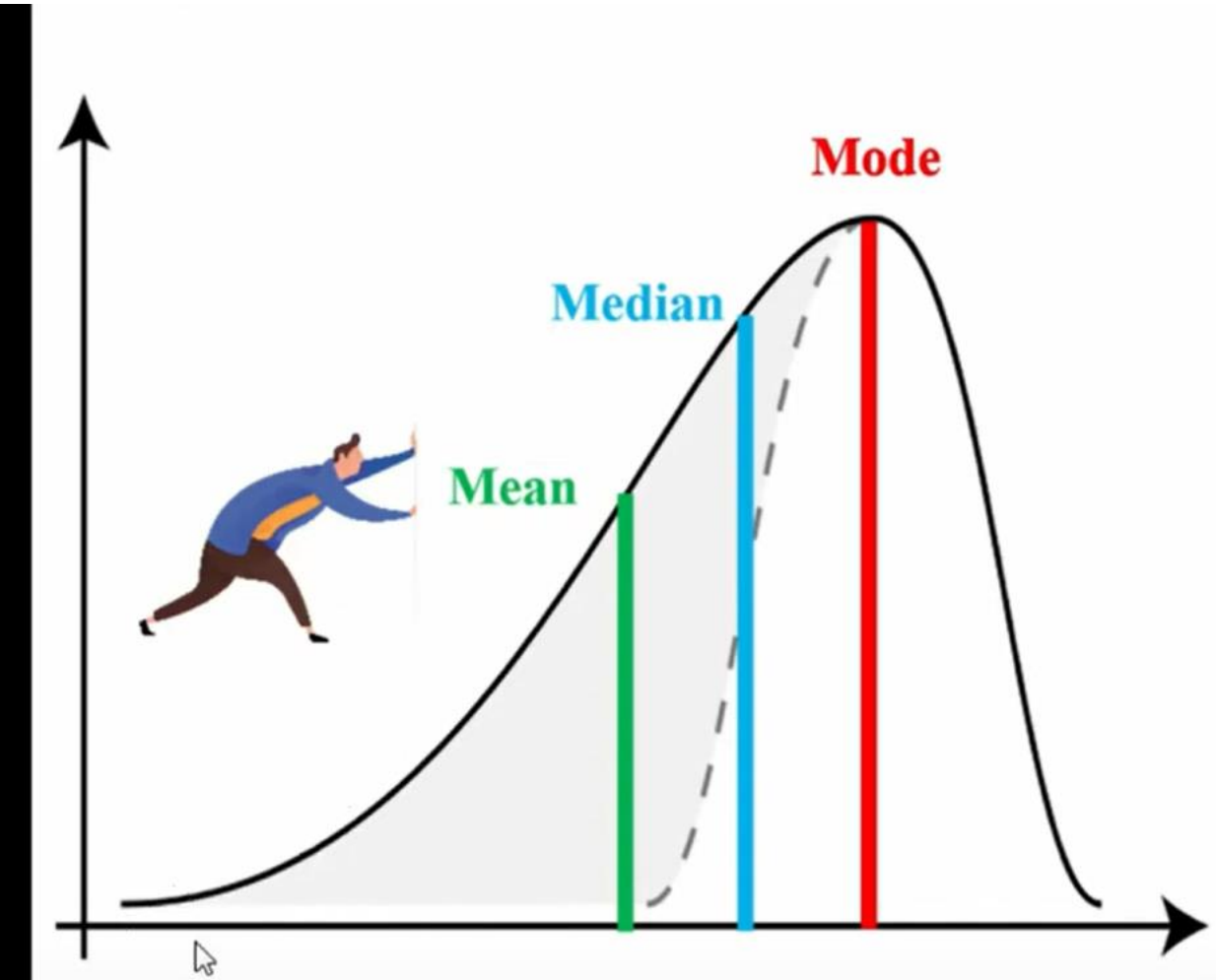
SKEWED TO THE RIGHT

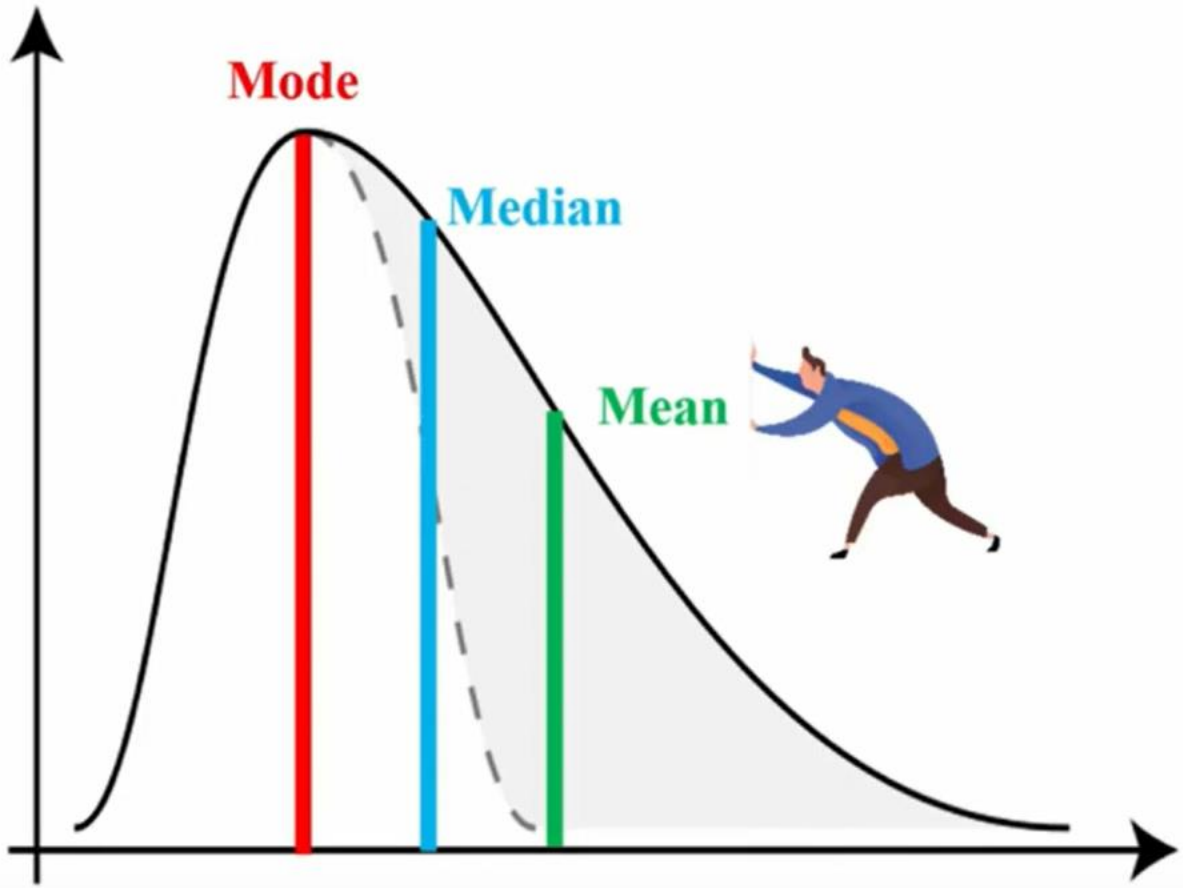
SKEWED TO THE RIGHT



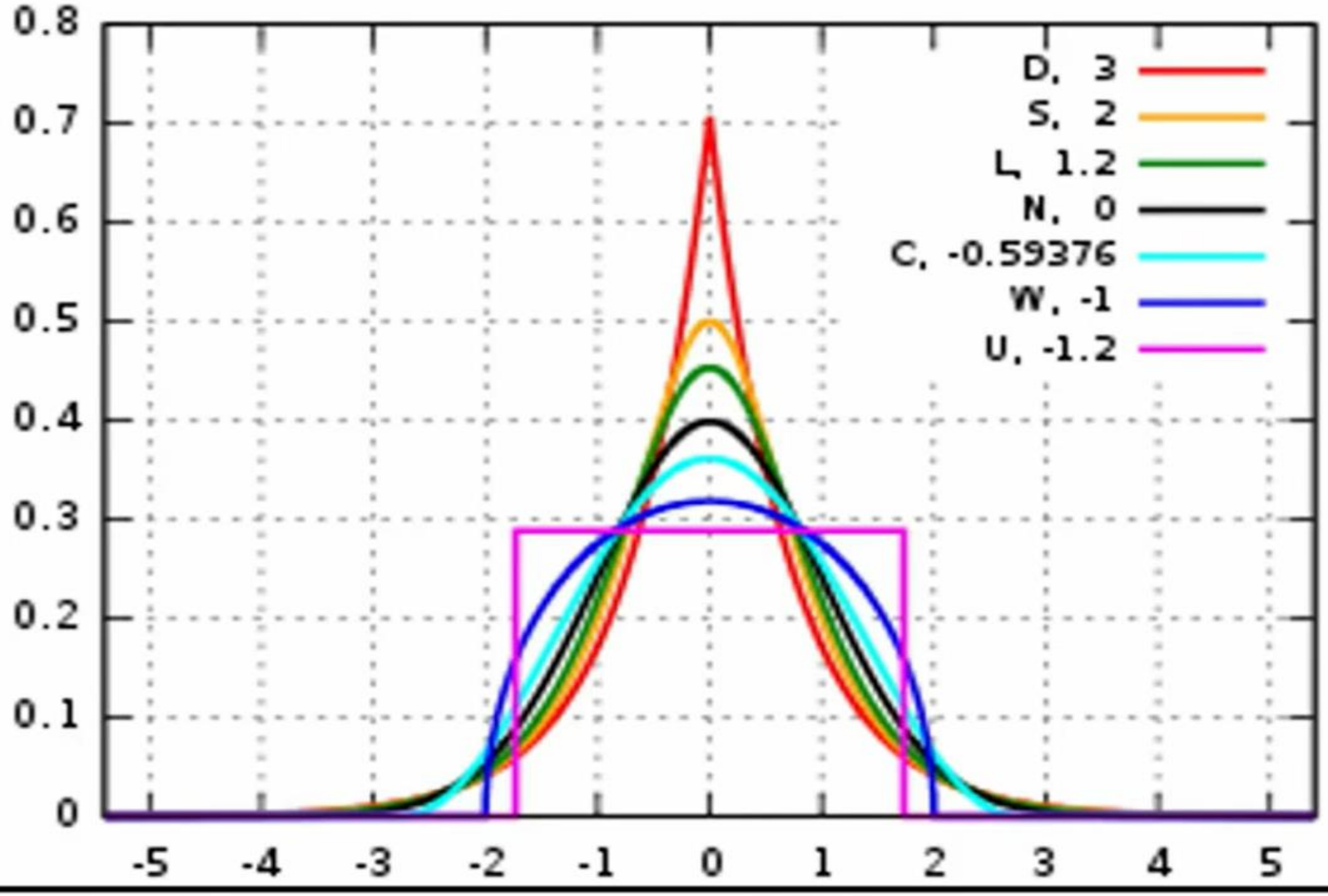

skewness = zero
positive skewness
negative skewness


skewness = zero
Right Skew
Left Skew
positive skewness
negative skewness


Mode
Median
Mean


Mode
Median
Mean

Kurtosis



Interpretation of Kurtosis

- Kurtosis can be understood with the help of Standard Deviation.
- **Smaller the Standard Deviation, Steeper the Distribution whereas**
- **Higher the Standard Deviation, Flatter the distribution**.

Q-Q Plot

Quantile-Quantile Plot

Q-Q Plot

Quantile-Quantile Plot

Is Given Data is Gaussian Distributed or Normal Distributed?

Q-Q Plot

Step 1 Sort the Data Feature Values and Make Percentiles

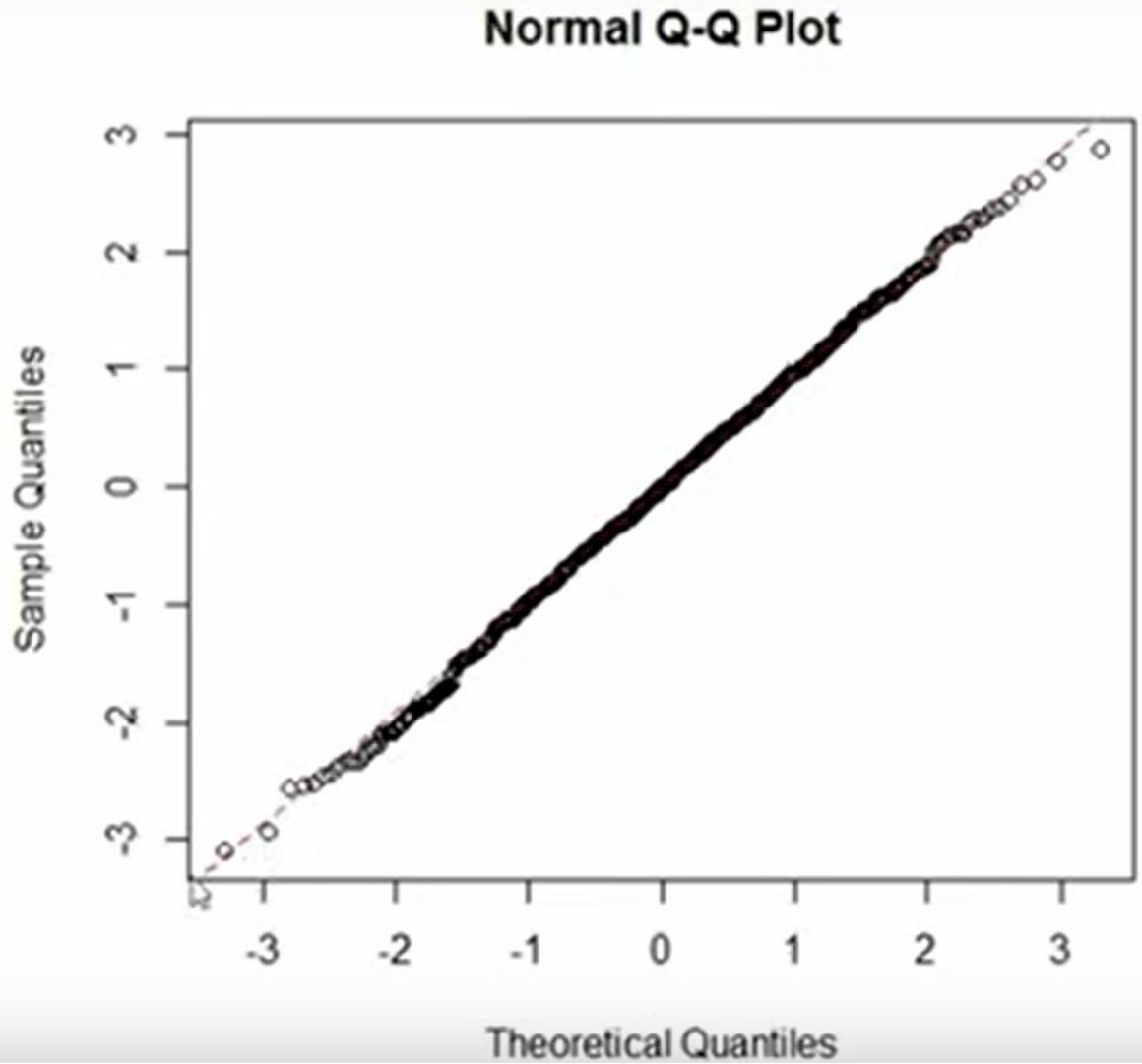
1,2,3,4,5,……,500 → 5,10,15,……,500

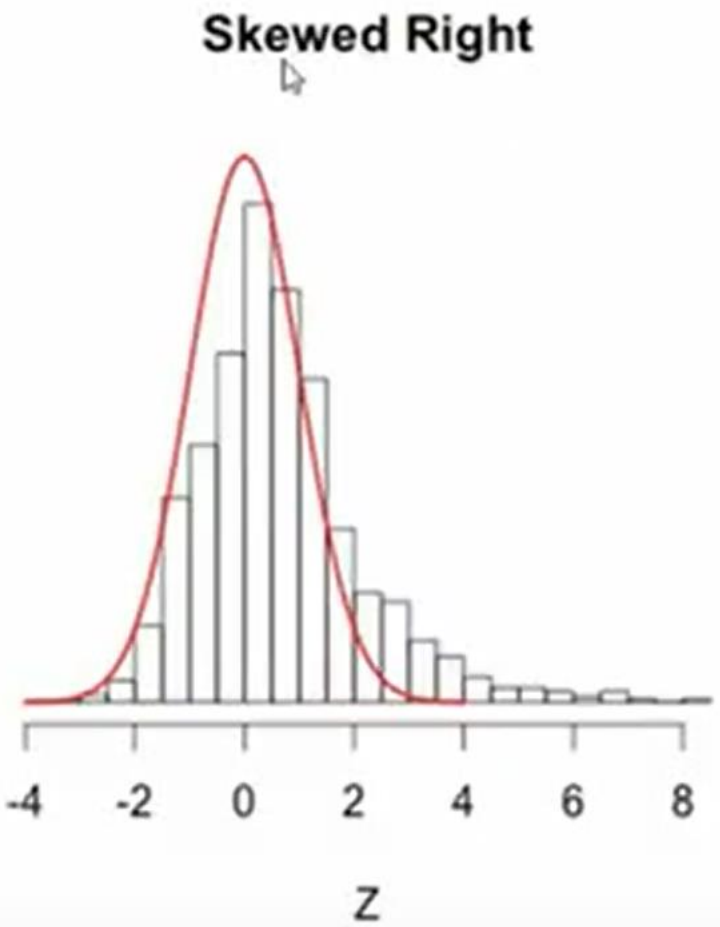
Step 2 Y ~ N(0,1) → y1,y2,y3,y4,y5,……,y1000

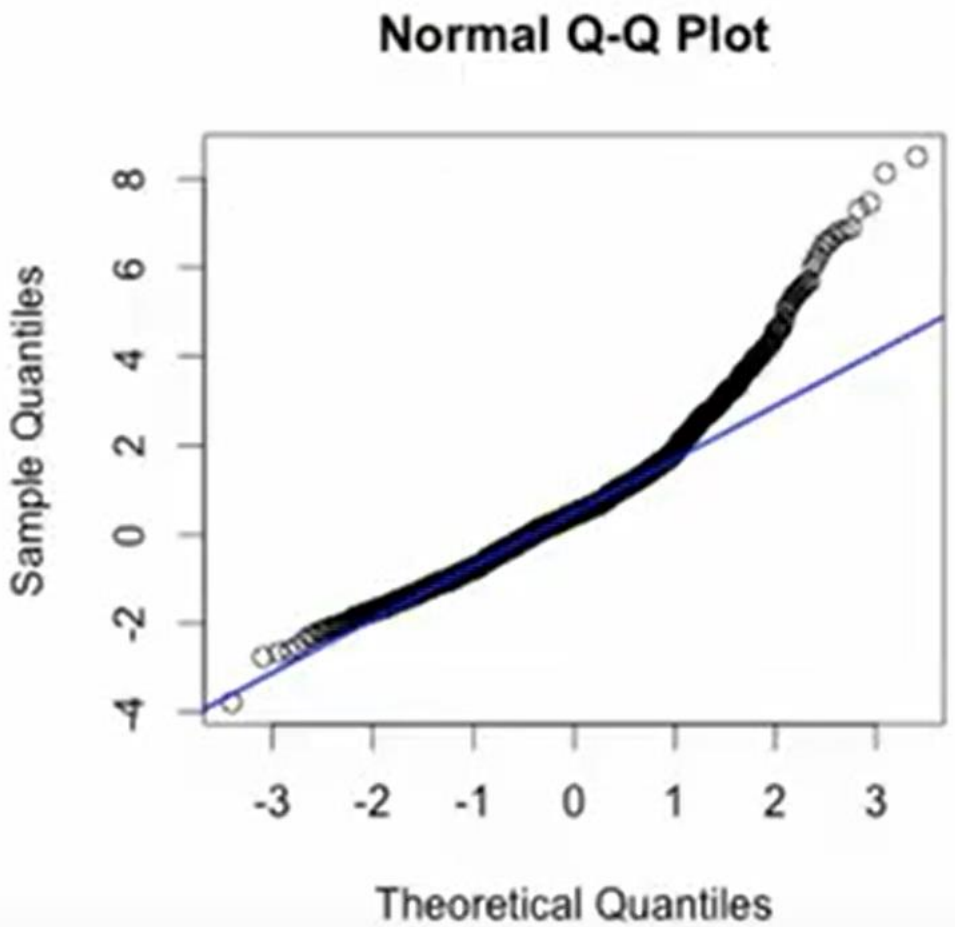
y10,y20,y30,……,y1000

Step 3 Plot it on Graph

Q-Q Plot

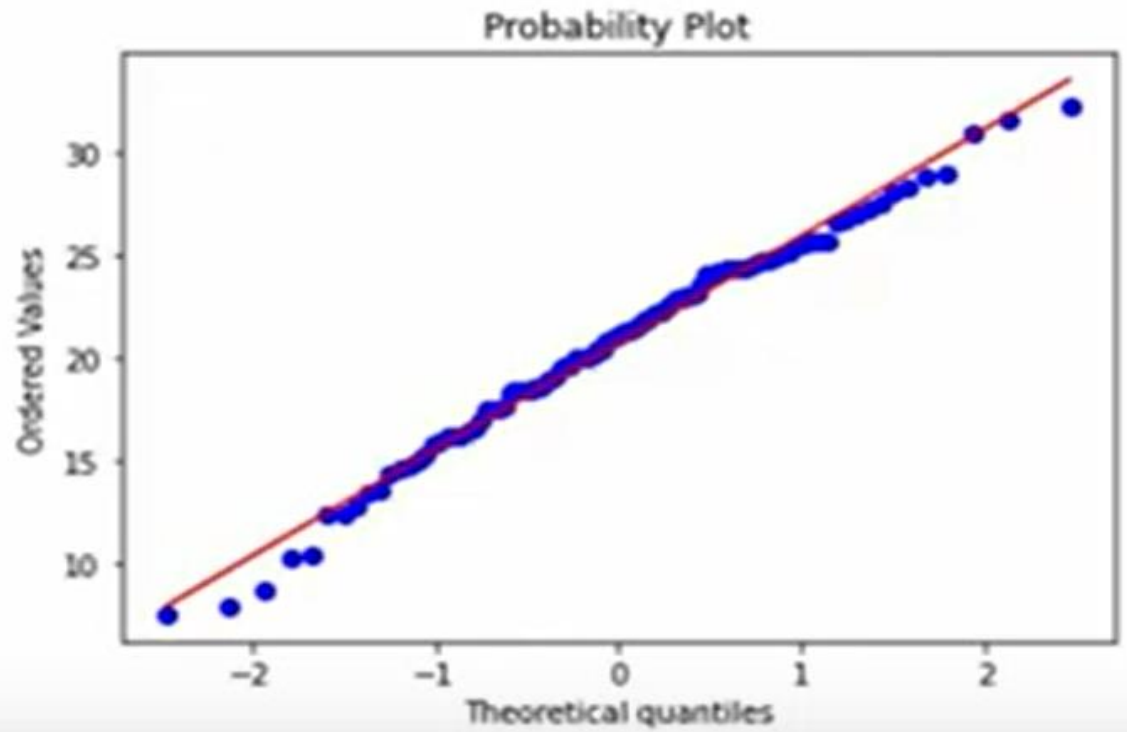

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-3
-2
-1
0
1
2
3


Skewed Right
-4
-2
0
2
4
6
8
Z


Normal Q-Q Plot
Sample Quantiles
-4
-2
0
2
4
6
8
-3
-2
-1
0
1
2
3
Theoretical Quantiles

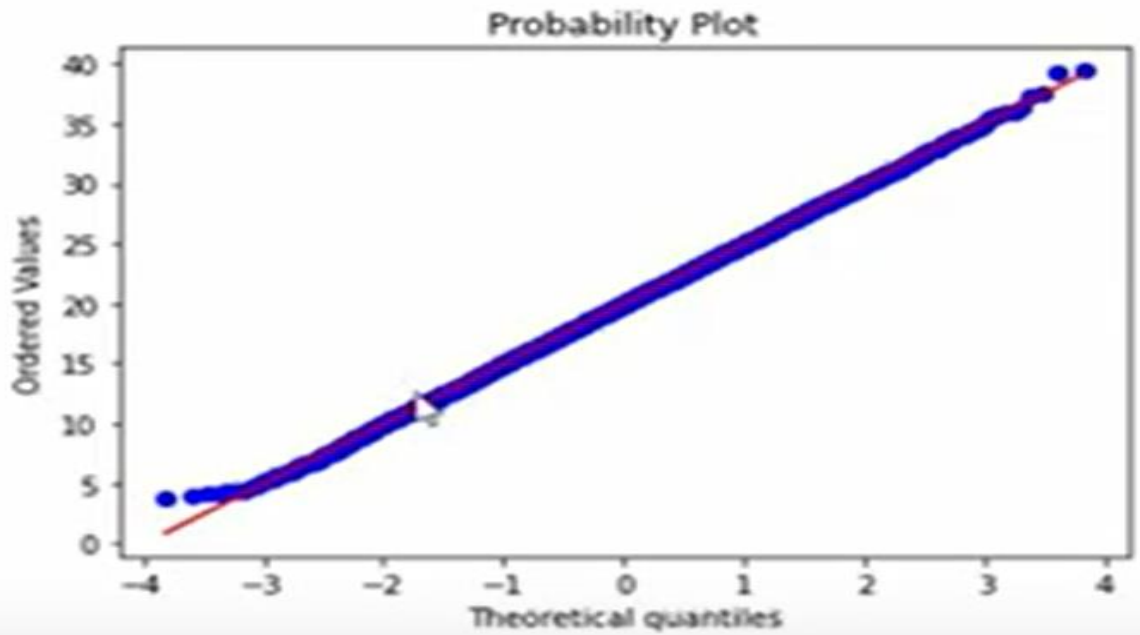
Q-Q Plot

```
import numpy as np
import pylab
import scipy.stats as stats

measurements = np.random.normal(loc = 20, scale = 5, size=100)
stats.probplot(measurements, dist="norm", plot=pylab)
pylab.show()
```



Q-Q Plot

```
[3]  import numpy as np
     import pylab
     import scipy.stats as stats

     measurements = np.random.normal(loc = 20, scale = 5, size=10000)
     stats.probplot(measurements, dist="norm", plot=pylab)
     pylab.show()
```



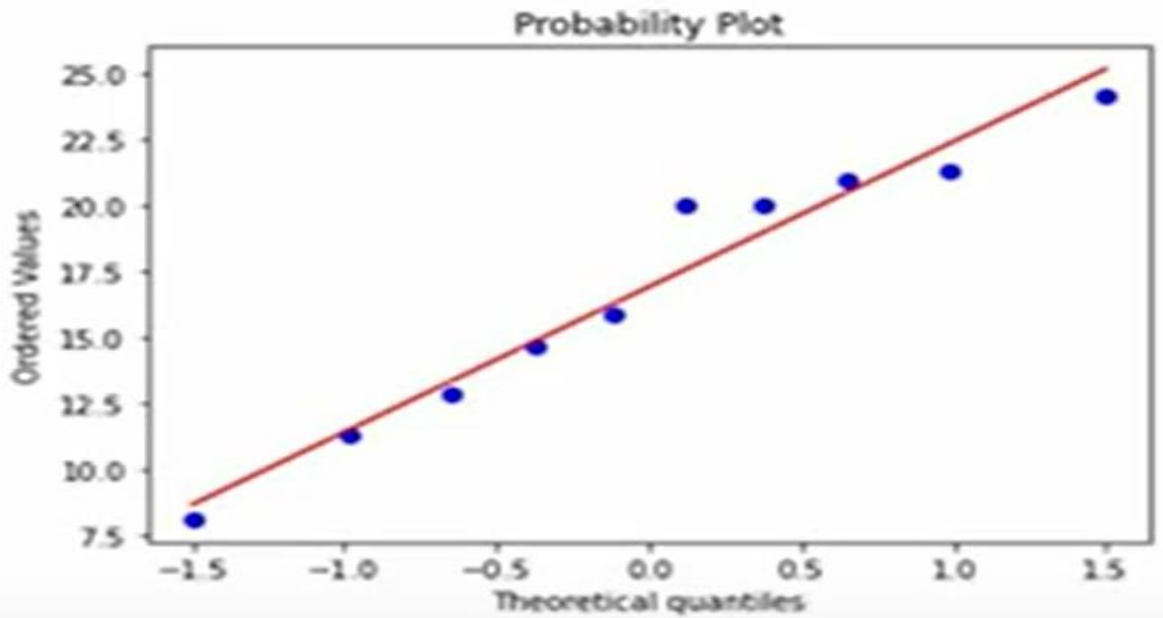
Q-Q Plot

Limitations

IF number of sample are small
then it hard to interpret Q-Q Plot

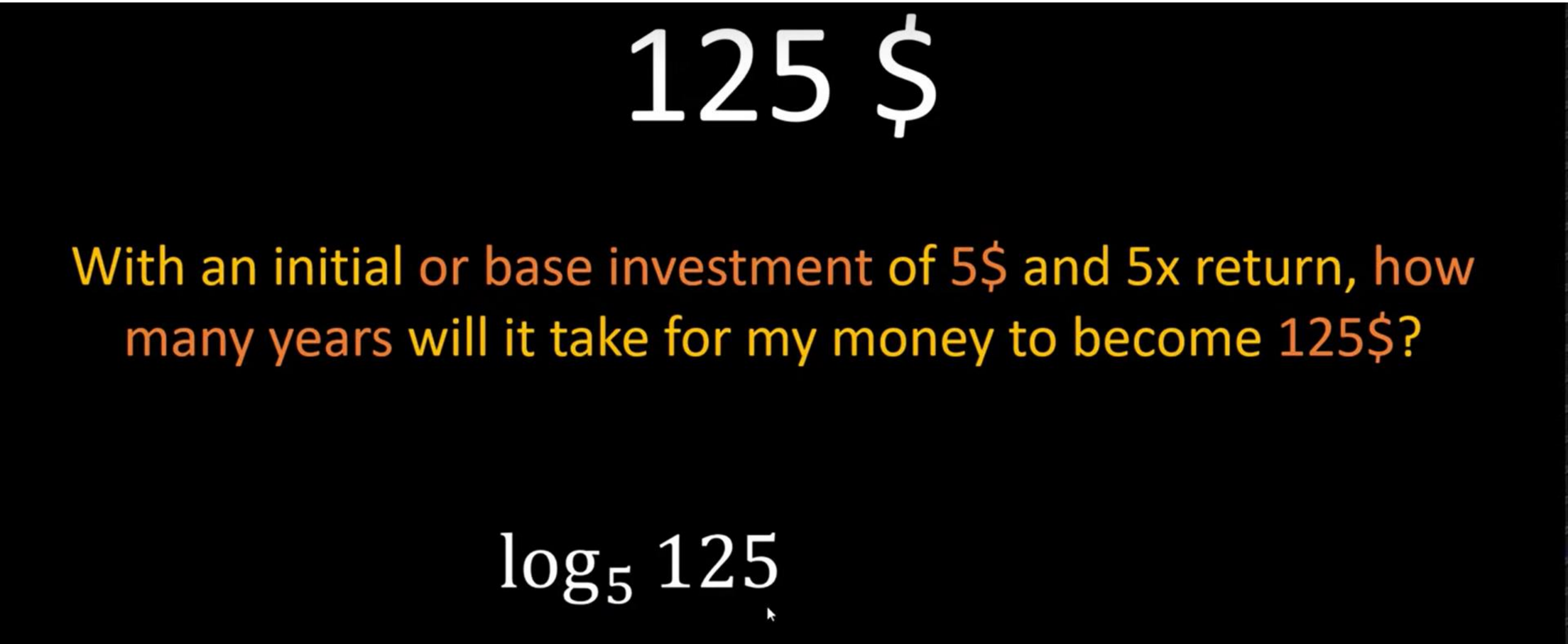
Q-Q Plot

```
import numpy as np
import pylab
import scipy.stats as stats

measurements = np.random.normal(loc = 20, scale = 5, size=10)
stats.probplot(measurements, dist="norm", plot=pylab)
pylab.show()
```



Role of log in Machine Learning models

Let us take reverse case,that you have 125$



Logarithm is an inverse of an exponent

125 $

With an initial or base investment of 5$ and 5x return, how many years will it take for my money to become 125$?

$$\log_5 125 \rightarrow 3$$

$\log_{10} 10 \rightarrow 1$

$\log_{10} 100 \rightarrow \log_{10} 10^2 \rightarrow 2 \log_{10} 10 \rightarrow 2$

$\log_{10} 1000 \rightarrow \log_{10} 10^3 \rightarrow 3 \log_{10} 10 \rightarrow 3$

Ref logarithm_in_data_analysis.ipnb on Jupyter notebook

Log transform in machine learning

person name	credit score	income	age	loan approved?
Rob	750	80000	32	Y
Tom	310	32000	45	N
Xi	475	77000	33	Y
Mohan	600	65000	51	N
Pooja	820	**550000**	35	Y
Sofiya	780	75000	31	Y

If we will apply this data,ML model will produce vague output

Log transform in machine learning

person name	credit score	income	age	loan approved?
Rob	750	80000	32	Y
Tom	310	32000	45	N
Xi	475	77000	33	Y
Mohan	600	65000	51	N
Pooja	820	**550000**	35	Y
Sofiya	780	75000	31	Y

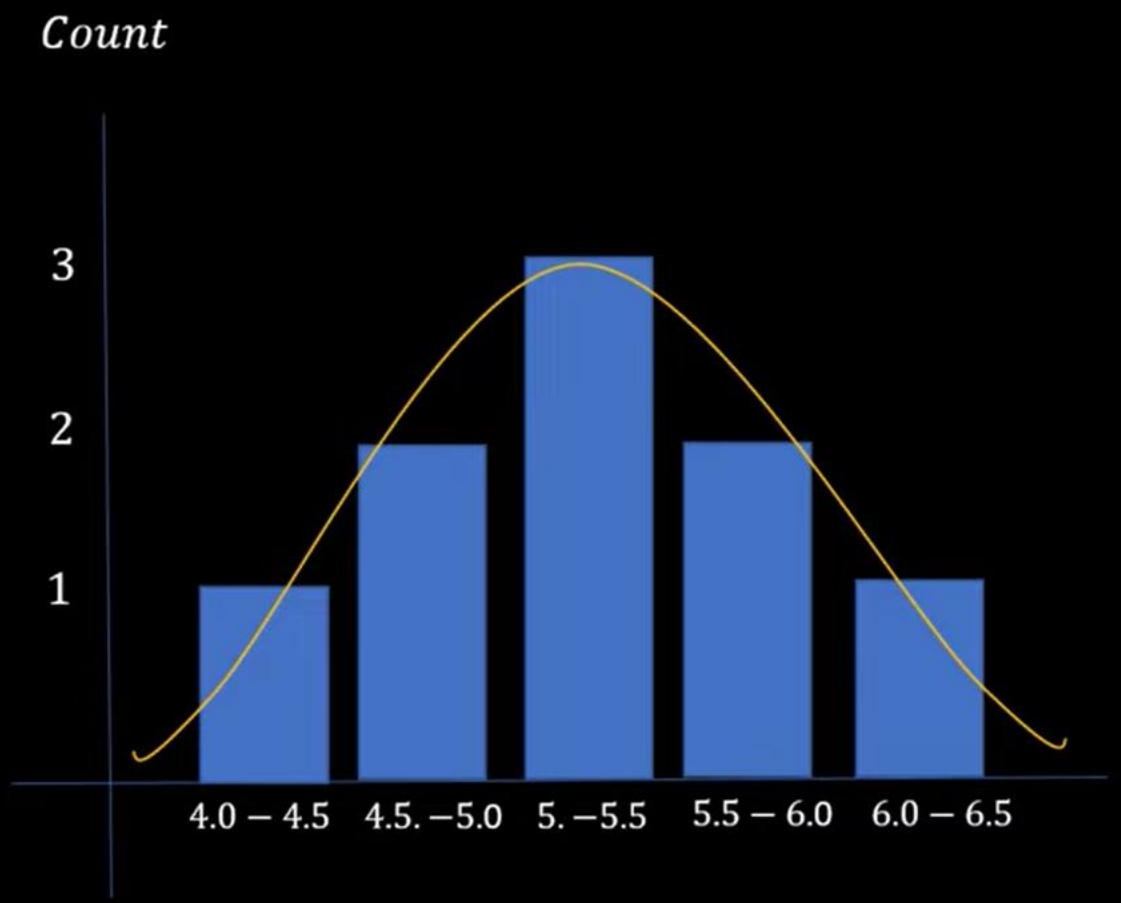
person name	credit score	income	age	loan approved?	log income
Rob	750	80000	32	Y	4.903089987
Tom	310	32000	45	N	4.505149978
Xi	475	77000	33	Y	4.886490725
Mohan	600	65000	51	N	4.812913357
Pooja	820	**550000**	35	Y	5.740362689
Sofiya	780	75000	31	Y	4.875061263

person name	credit score	income	age	loan approved?	log income
Rob	750	80000	32	Y	4.903089987
Tom	310	32000	45	N	4.505149978
Xi	475	77000	33	Y	4.886490725
Mohan	600	65000	51	N	4.812913357
Pooja	820	**550000**	35	Y	5.740362689
Sofiya	780	75000	31	Y	4.875061263

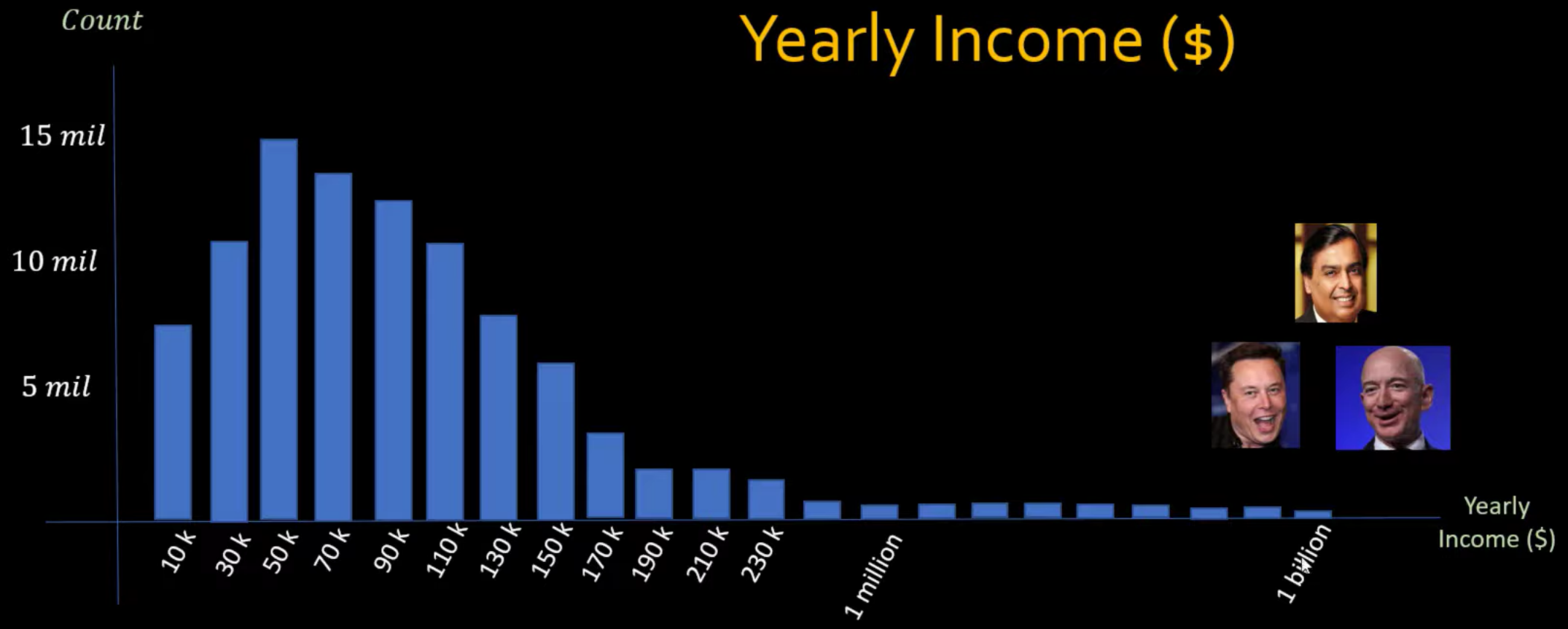
Earthquake → Ritcher Scale

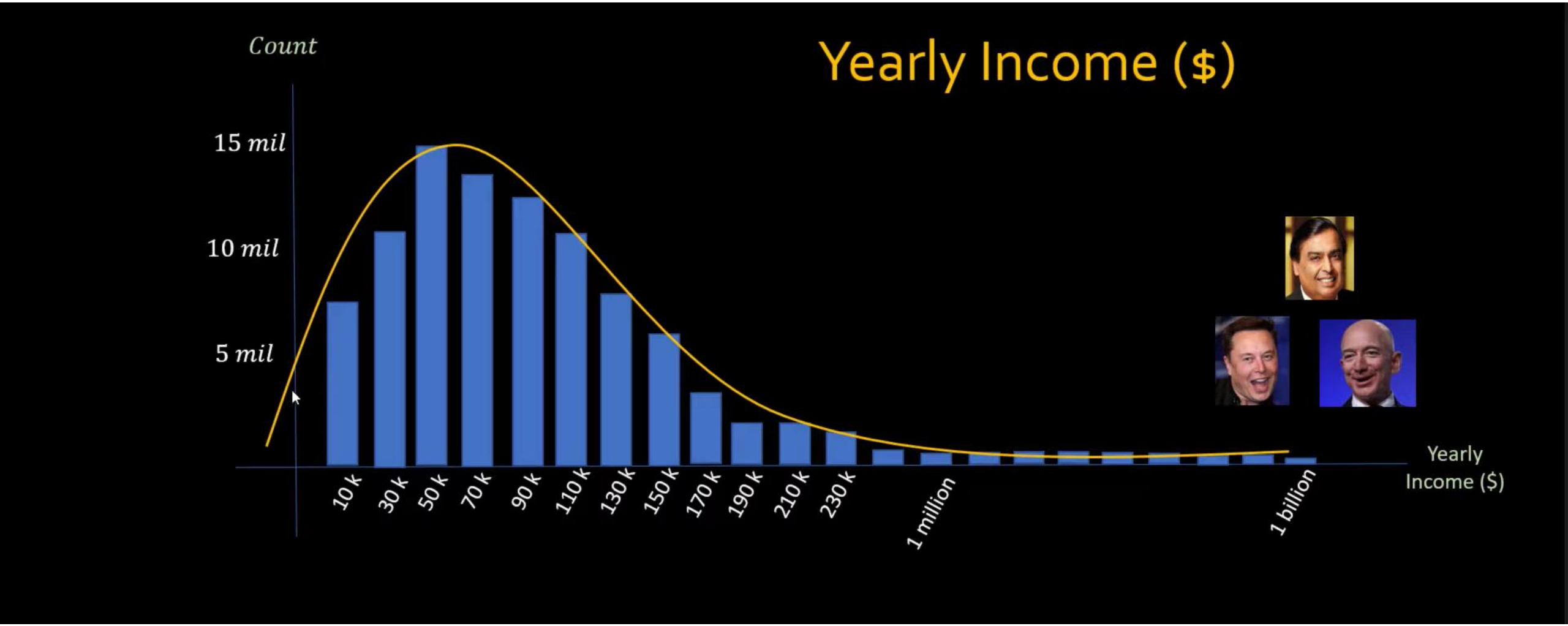
Earthquake of scale 5 is 10 times more powerful than scale 4

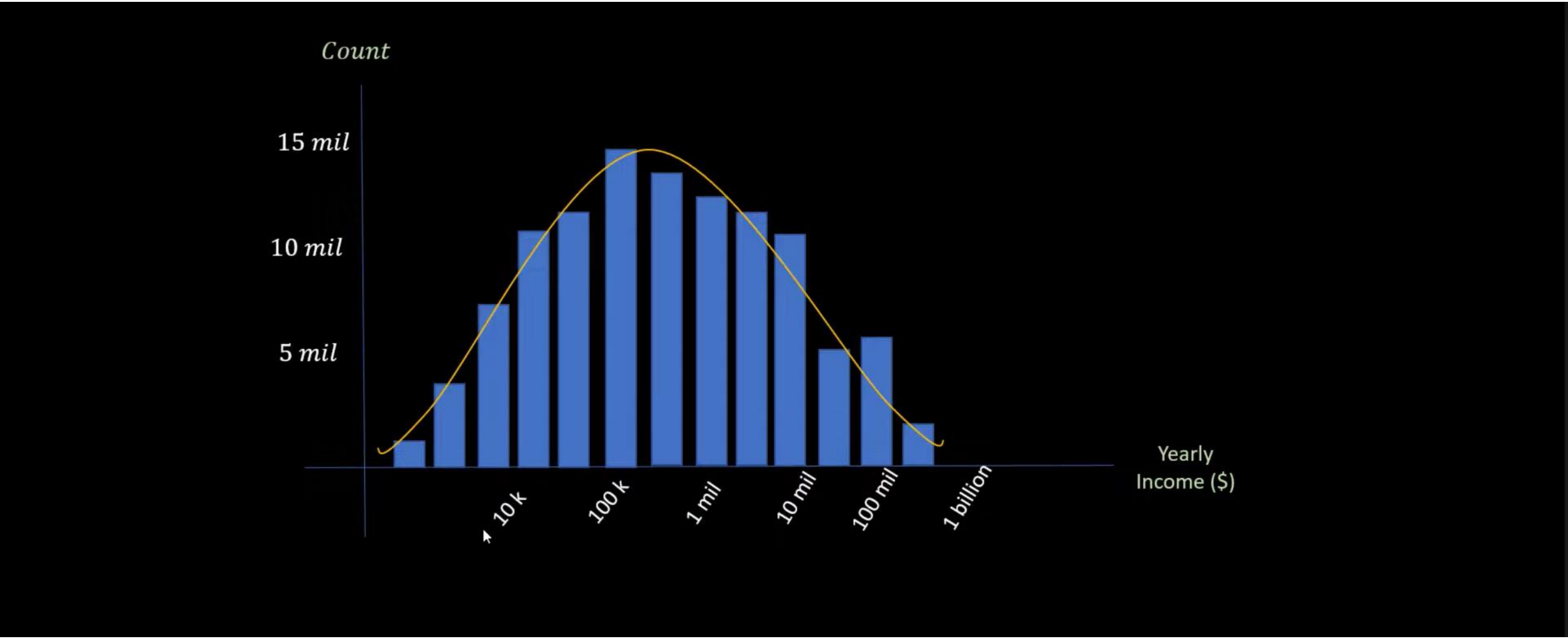
Name	Height (ft)
Rob	6.2
Thomas	5.7
Nina	4.6
Mittal	5.4
Sofia	5.9
Mohan	4.3
Tao	5.1
Deepika	5.2
Rafiq	4.9

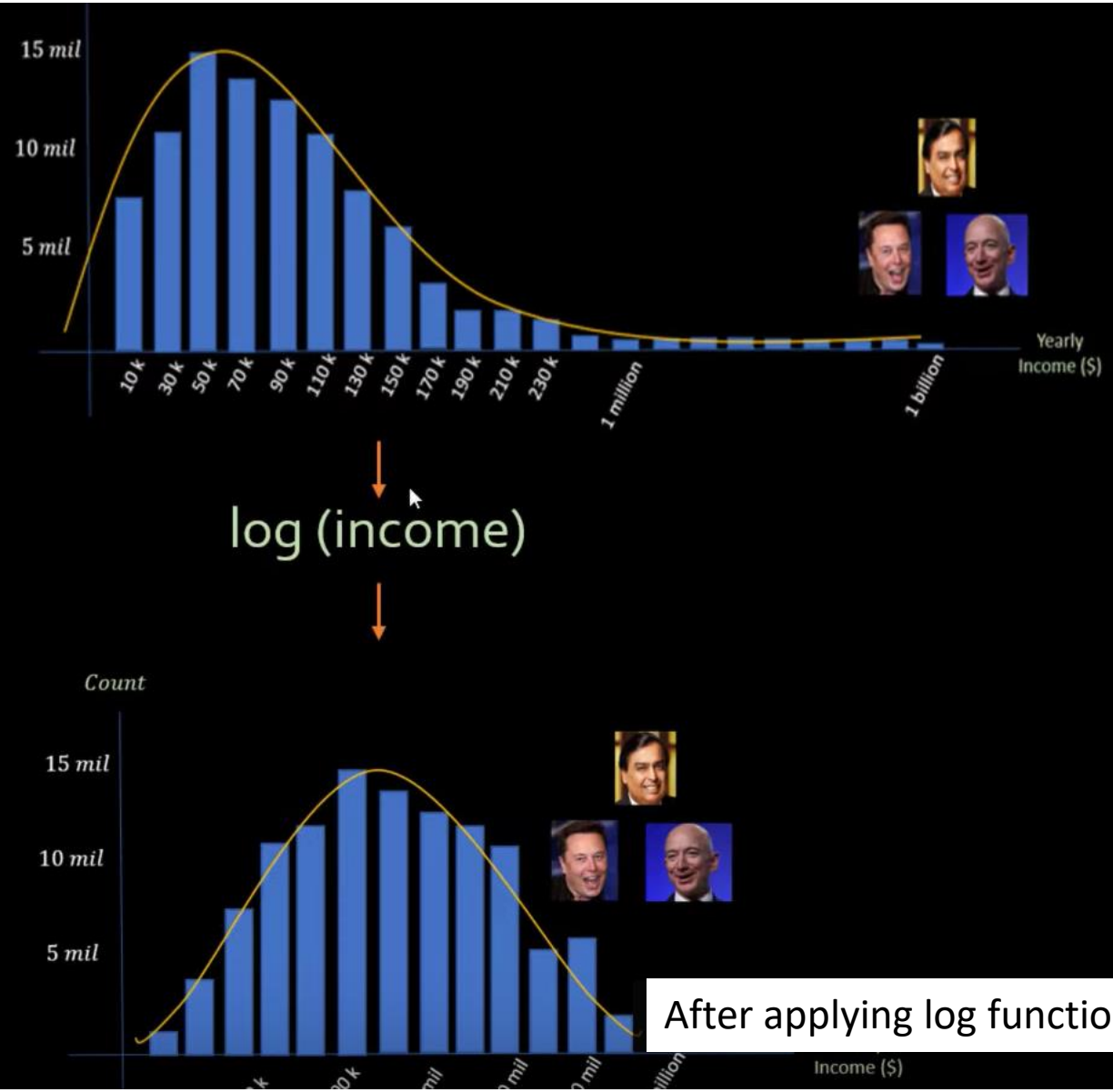


Yearly Income ($)



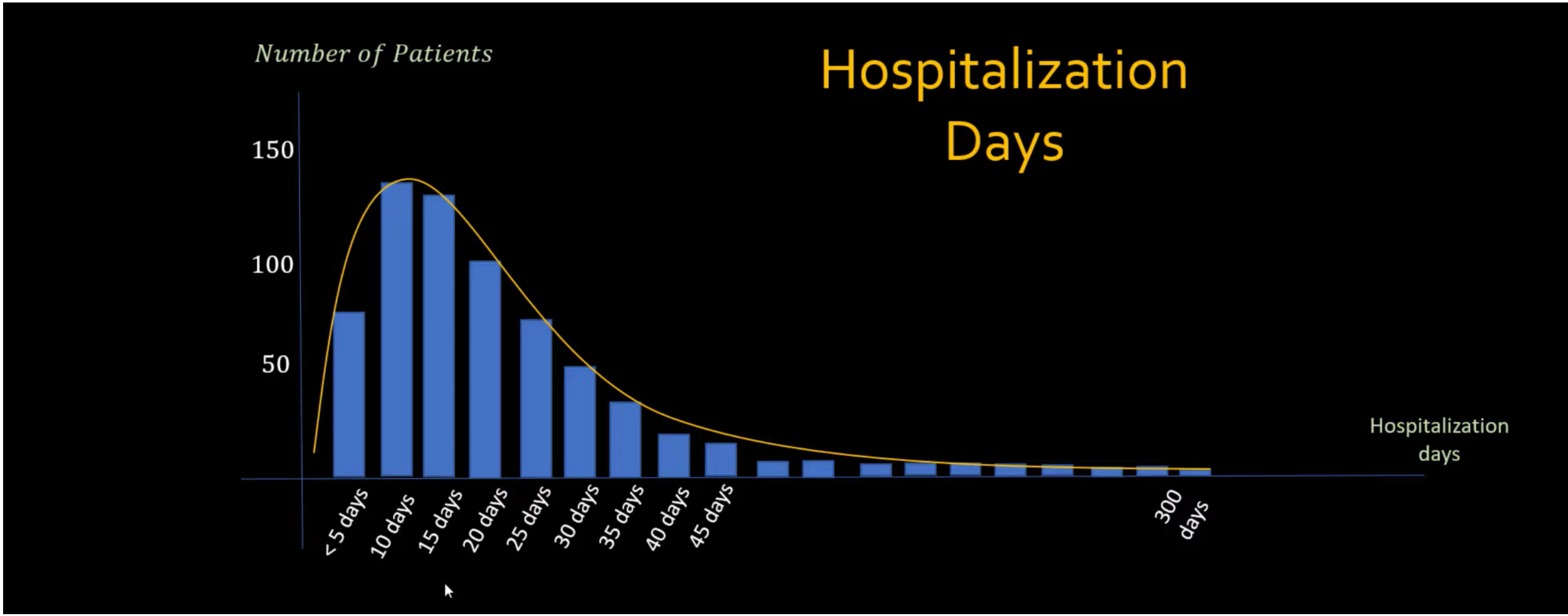

Count
15 mil
10 mil
5 mil
Yearly Income ($)
10 k
30 k
50 k
70 k
90 k
110 k
130 k
150 k
170 k
190 k
210 k
230 k
1 million
1 billion
Yearly Income ($)

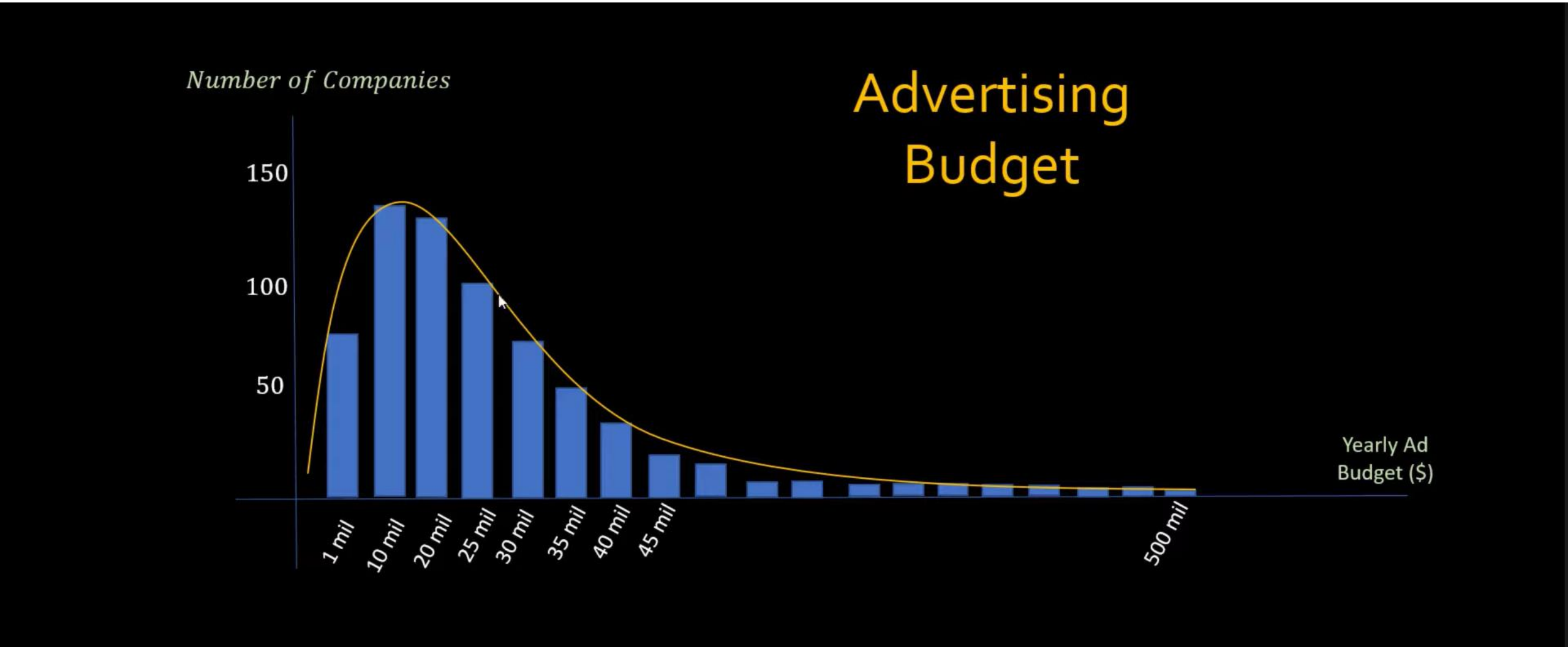

Count
15 mil
10 mil
5 mil
10 k
100 k
1 mil
10 mil
100 mil
1 billion
Yearly
Income ($)



If you get a normal distribution by applying a log function to a dataset then dataset is log normally distributed

After applying log function


Number of Patients
Hospitalization Days
150
100
50
Hospitalization days
< 5 days
10 days
15 days
20 days
25 days
30 days
35 days
40 days
45 days
300 days


Number of Companies
Advertising Budget
150
100
50
Yearly Ad Budget ($)
1 mil
10 mil
20 mil
25 mil
30 mil
35 mil
40 mil
45 mil
500 mil

How it is used in data science?

person name	credit score	income	age	loan approved?
Rob	750	80000	32	Y
Tom	310	32000	45	N
Xi	475	77000	33	Y
Mohan	600	65000	51	N
Pooja	820	**550000**	35	Y
Sofiya	780	75000	31	Y

Let us take another example

B15 | 6097

Table HINC-06. Income Distribution to $250,000 or More for Households: 2017

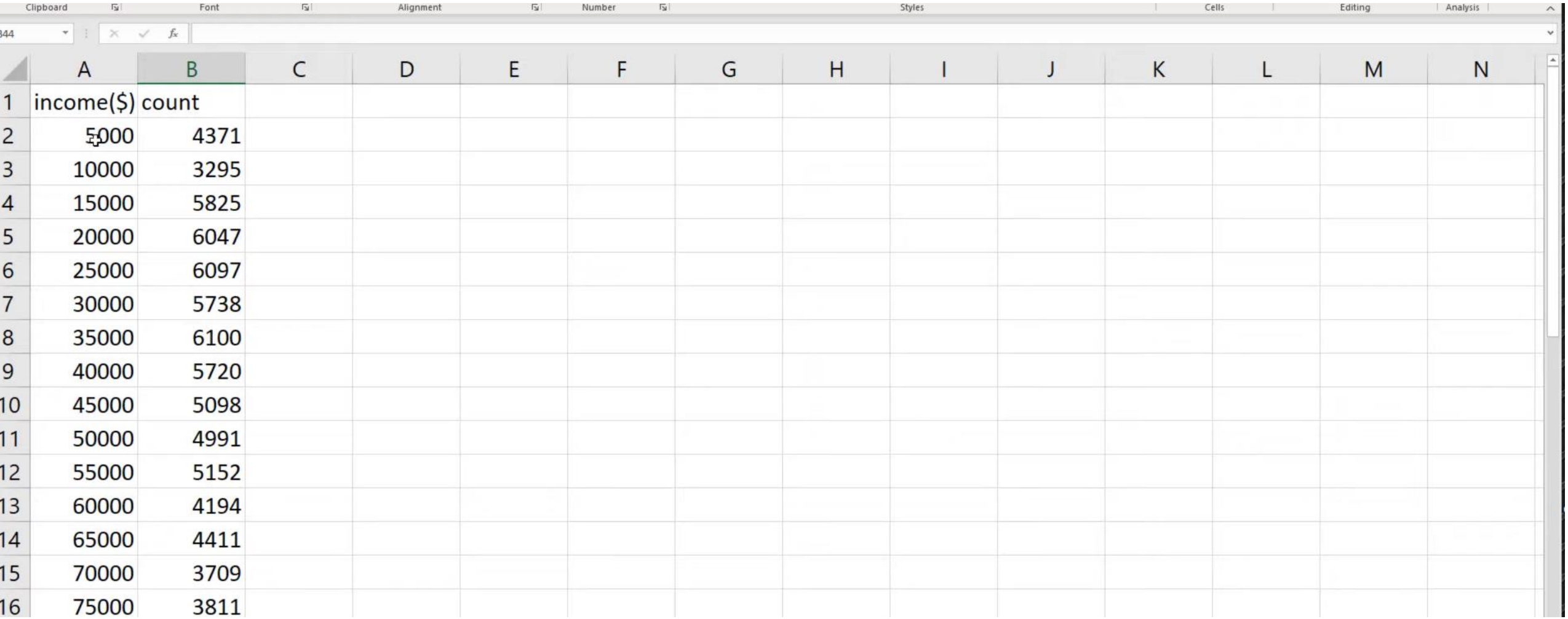
Data reflect the implementation of an updated processing system that incorporates content from earlier questionnaire redesigns related to income, health insurance, and demographics.

For information on confidentiality protection, sampling error, nonsampling error, and definitions, see <www2.census.gov/programs-surveys/cps/techdocs/cpsmar18.pdf>.

Source: U.S. Census Bureau, Current Population Survey, 2018 Annual Social and Economic Supplement.

(Numbers in thousands. Households as of March of the following year. A.O.I.C. stands for alone or in combination. Standard errors calculated using replicate weights)

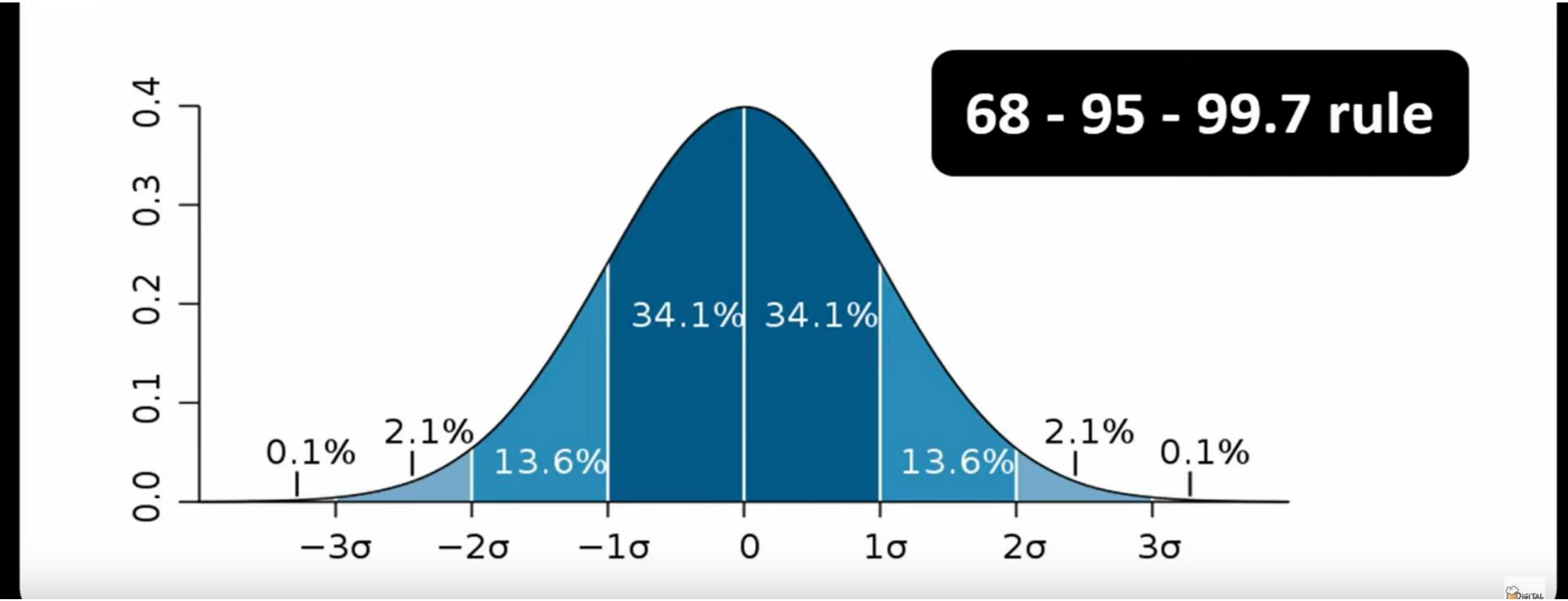
Income of Household	All Races			White A.O.I.C.			White alone (1)			White alone, not His	
	Number	Mean Income		Number	Mean Income		Number	Mean Income		Number	Mean
		Dollars	Standard Error		Dollars	Standard Error		Dollars	Standard Error		Dollars
Total	127,669	87,643	570	101,974	91,246	631	100,113	91,519	642	84,706	95,759
Under $5,000	4,371	1,128	46	2,815	1,173	61	2,739	1,147	61	2,154	1,155
$5,000 to $9,999	3,295	7,933	45	2,196	7,882	56	2,154	7,881	57	1,673	7,897
$10,000 to $14,999	5,825	12,319	36	4,254	12,338	40	4,142	12,341	40	3,367	12,390
$15,000 to $19,999	6,047	17,260	33	4,574	17,282	40	4,494	17,286	41	3,675	17,302
$20,000 to $24,999	6,097	22,224	32	4,682	22,281	37	4,579	22,279	37	3,682	22,341
$25,000 to $29,999	5,738	27,038	34	4,447	27,073	39	4,333	27,082	39	3,464	27,155
$30,000 to $34,999	6,100	32,018	34	4,676	32,052	40	4,575	32,052	40	3,691	32,100
$35,000 to $39,999	5,720	37,060	36	4,429	37,140	40	4,354	37,142	41	3,529	37,181
$40,000 to $44,999	5,098	42,019	33	4,042	42,021	38	3,947	42,035	38	3,210	42,098
$45,000 to $49,999	4,991	47,037	34	4,044	47,054	39	3,975	47,054	40	3,209	47,064
$50,000 to $54,999	5,152	51,933	37	4,130	51,949	40	4,047	51,950	41	3,356	51,998



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	income($)	count												
2	5000	4371												
3	10000	3295												
4	15000	5825												
5	20000	6047												
6	25000	6097												
7	30000	5738												
8	35000	6100												
9	40000	5720												
10	45000	5098												
11	50000	4991												
12	55000	5152												
13	60000	4194												
14	65000	4411												
15	70000	3709												
16	75000	3811												

Ref lognormal_dist.ipynb

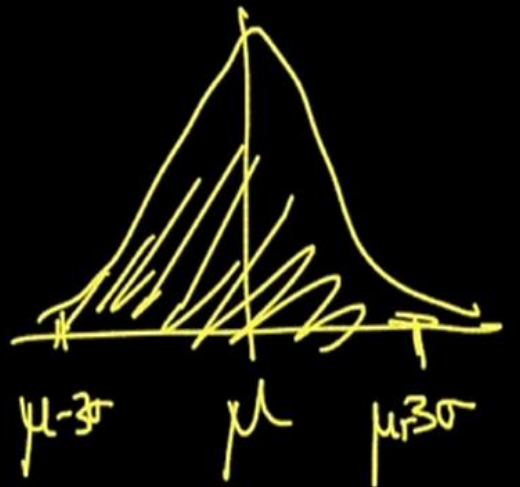
C


68 - 95 - 99.7 rule
0.0
0.1
0.2
0.3
0.4
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ

Chebyshev's Inequality

$X \rightarrow$ Gaussian distb $\rightarrow$ 68 - 95 - 99.7 rule

1σ, 2σ, 3σ



$X \leftarrow$ heights of students

$X \sim N(\mu, \sigma)$ $\rightarrow$ 1 cm

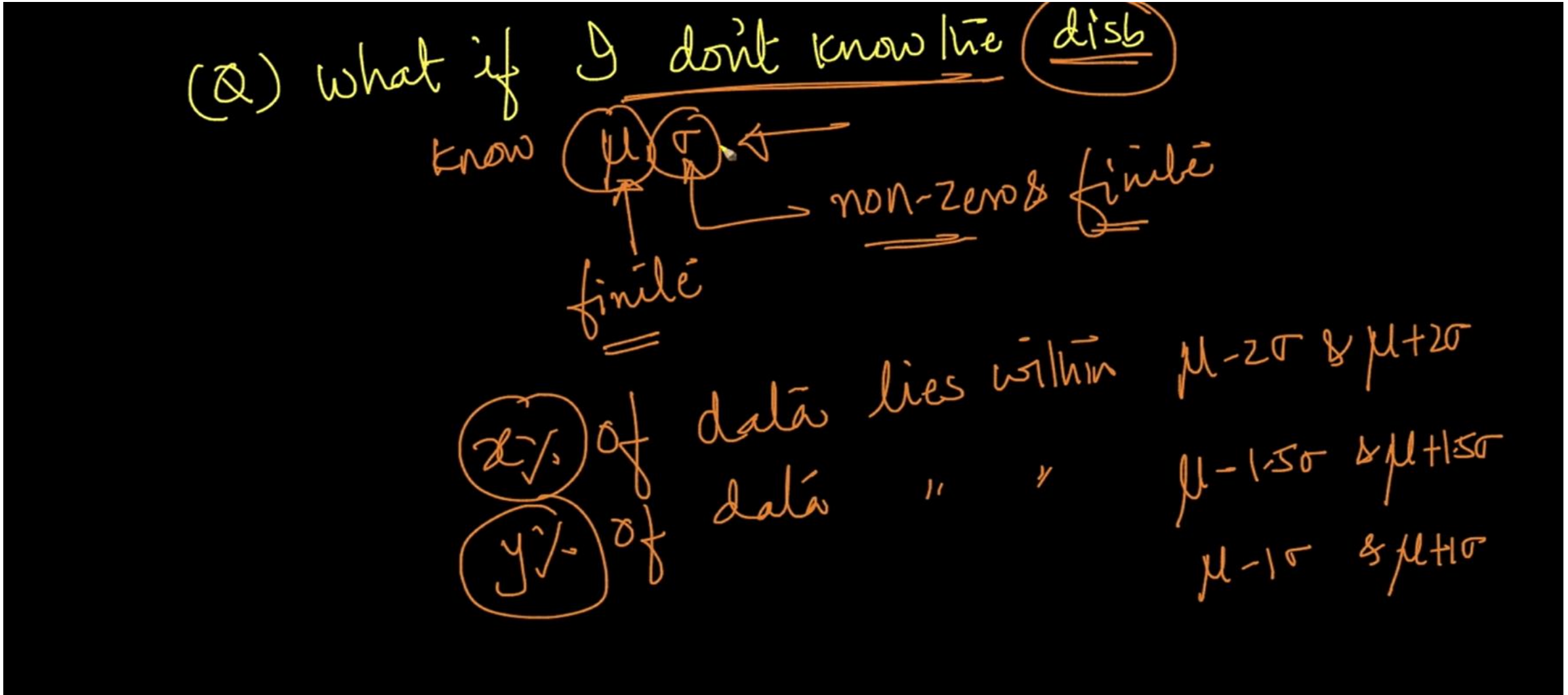
1 — 2 — 3

$\rightarrow$ 150 cm

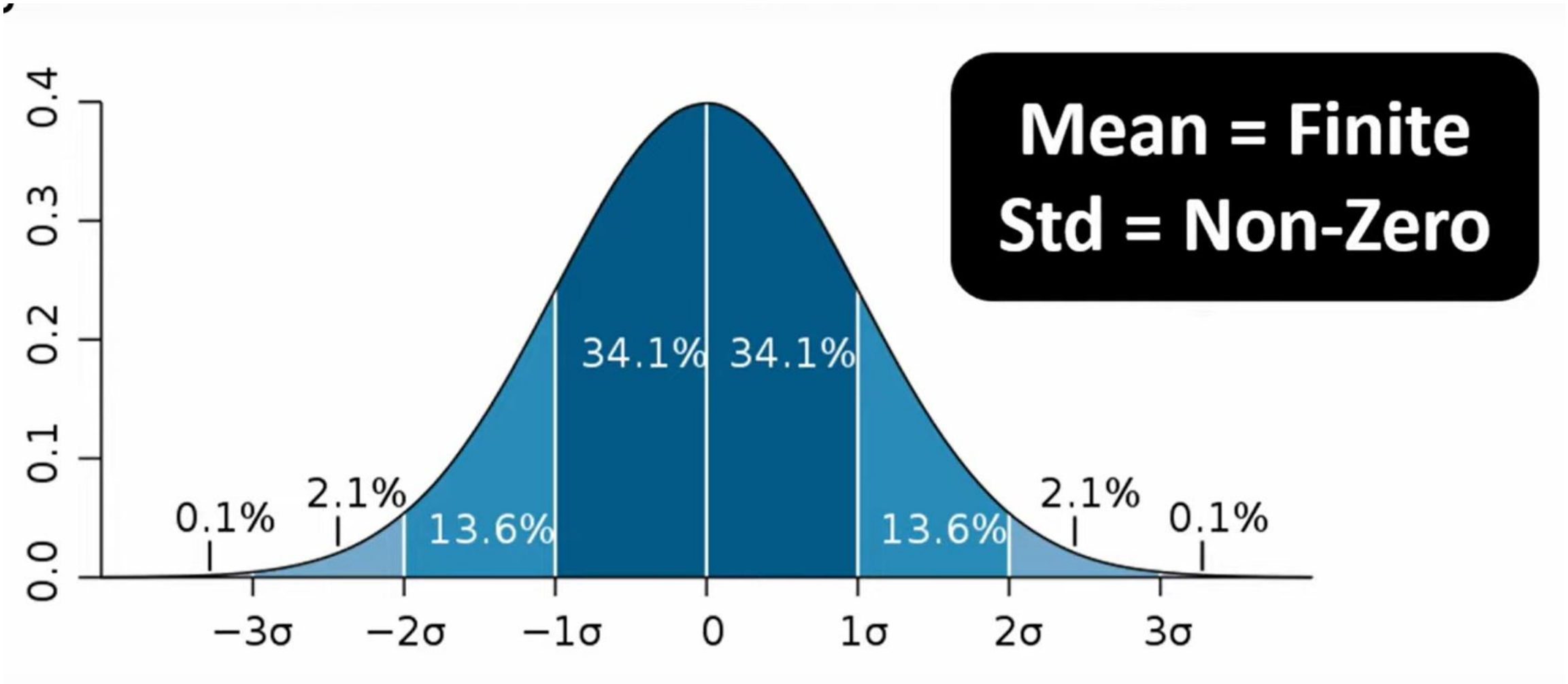
$Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95\%$

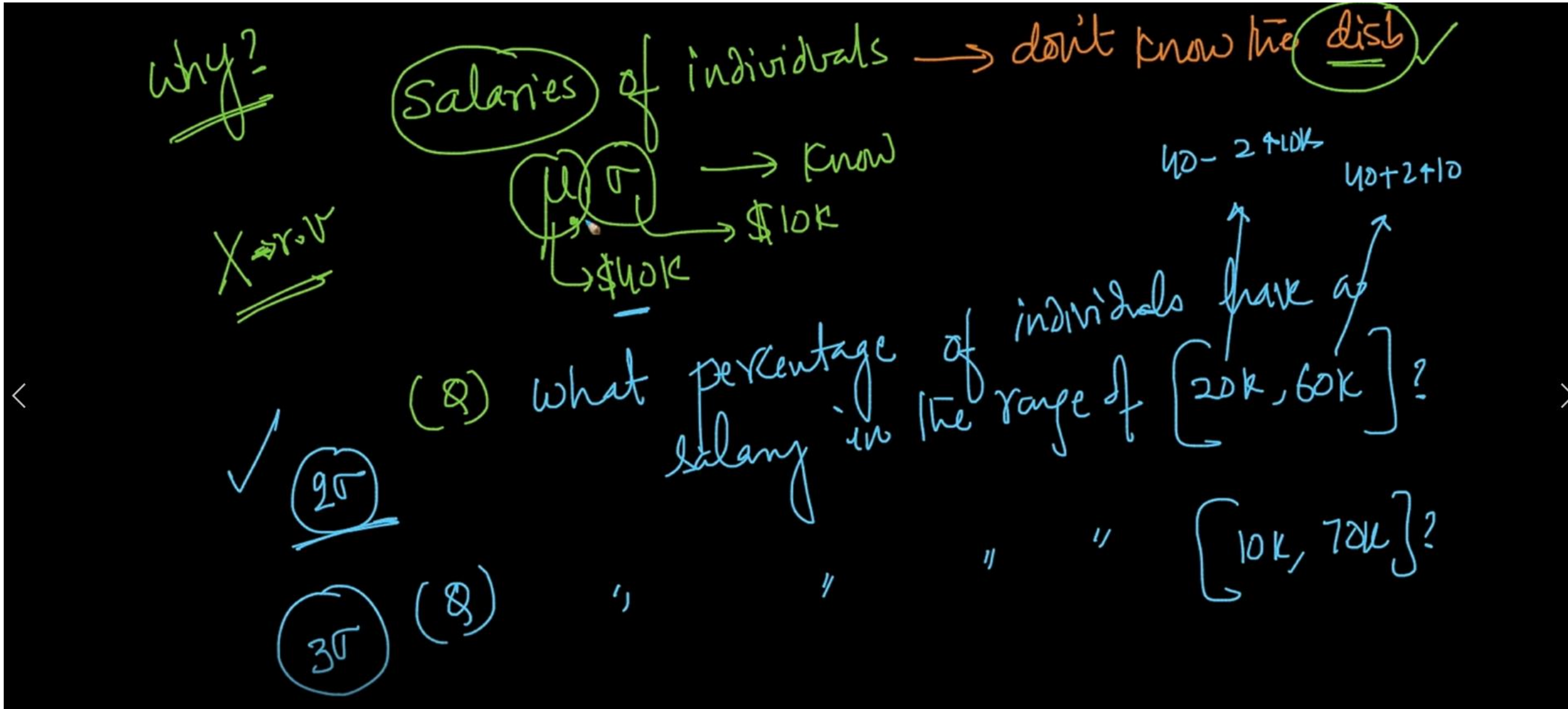
95% of students heights lie in the $[130, 170]$

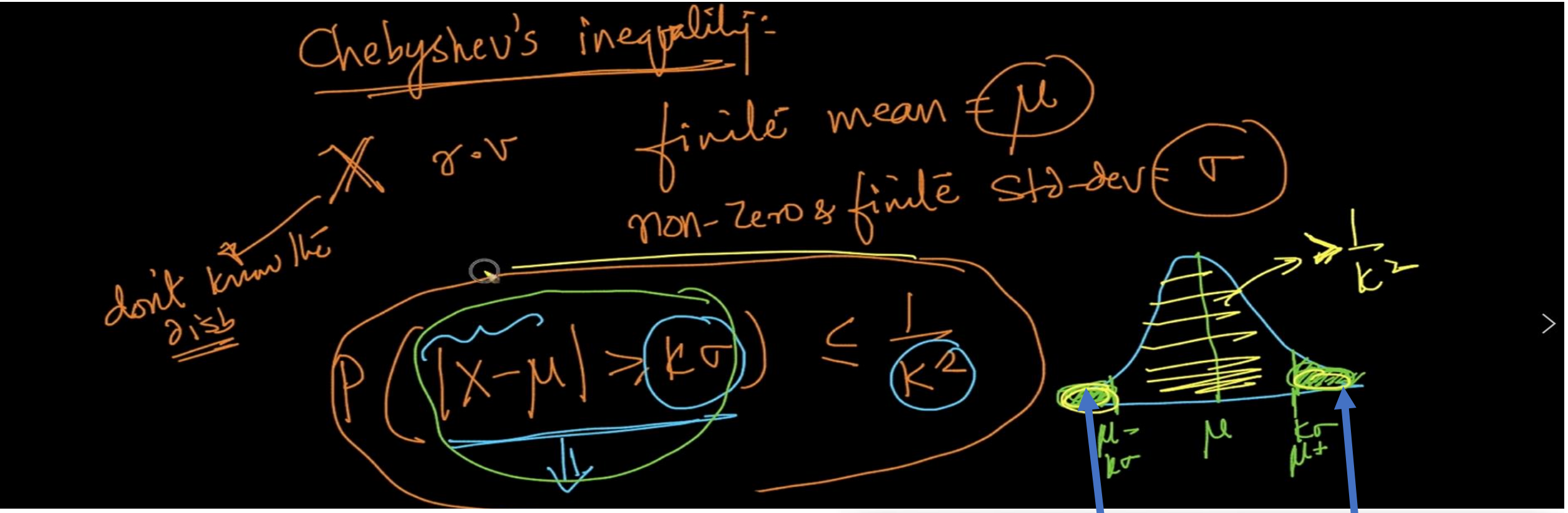
Can we find the % of data between the range,if we do not know the distribution?--NO

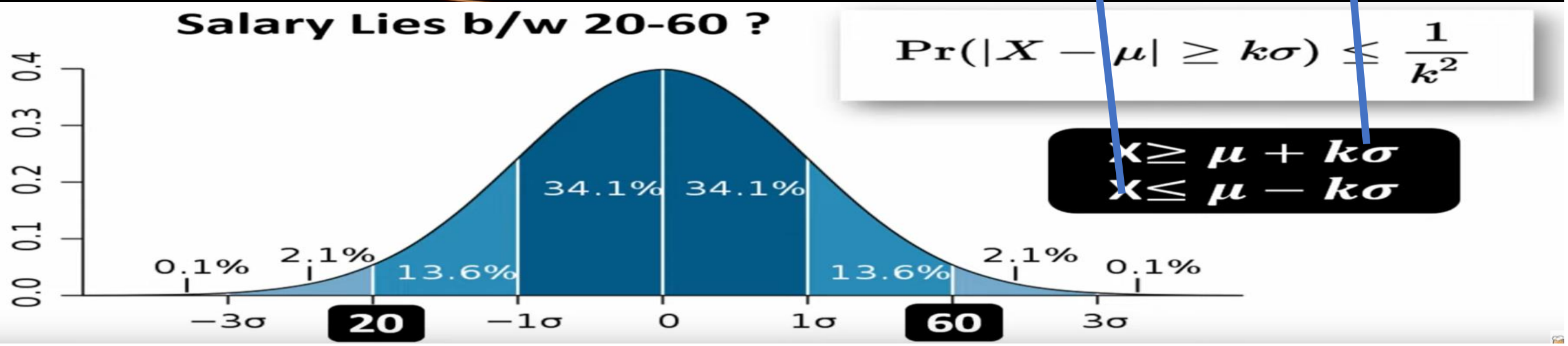


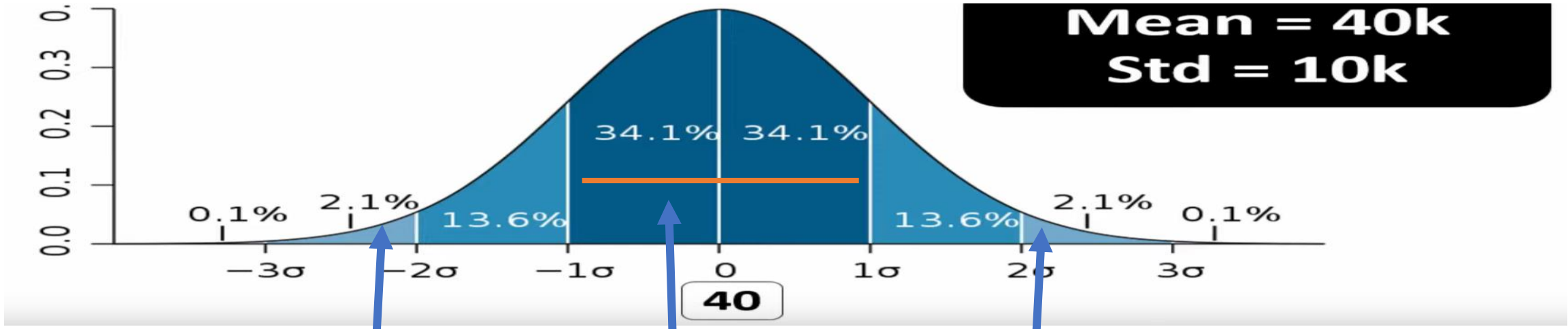
Chebyshev’s condition




why?
Salaries of individuals → don't know the dist
→ know
$10K
$40K
X ~ N(μ, σ²)
(Q) what percentage of individuals have a salary in the range of [20K, 60K]?
μ - 2*10K
μ + 2*10
2σ
3σ
(Q) " " " [10K, 70K]?


Chebyshev's inequality:
X r.v
finite mean = μ
non-zero & finite std-dev = σ
don't know the dist
P(|X−μ| ≥ kσ) ≤ 1/k²
1/k²
μ−kσ
μ
μ+kσ

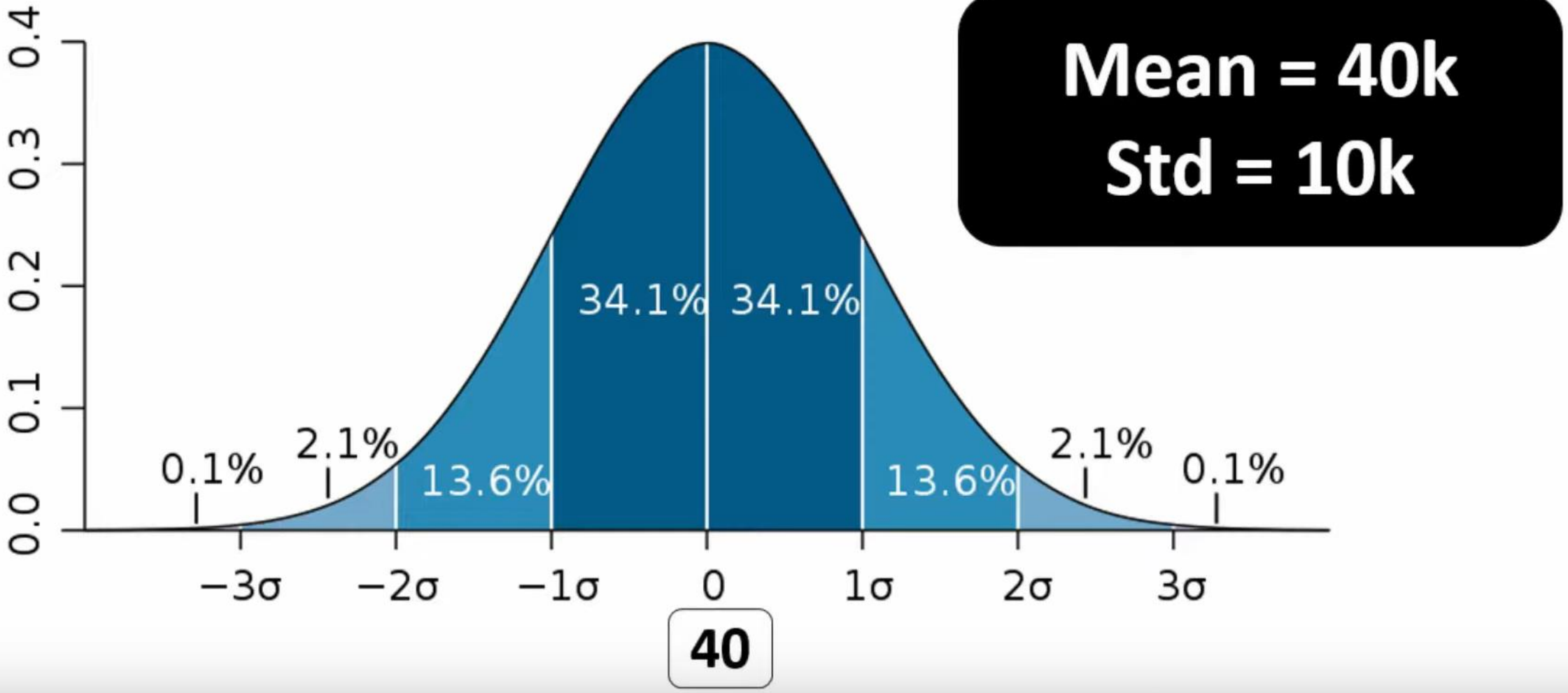

Salary Lies b/w 20-60 ?
Pr(|X − μ| ≥ kσ) ≤ 1/k²
X≥ μ + kσ
X≤ μ − kσ
0.0
0.1
0.2
0.3
0.4
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
−3σ
20
−1σ
0
1σ
60
3σ

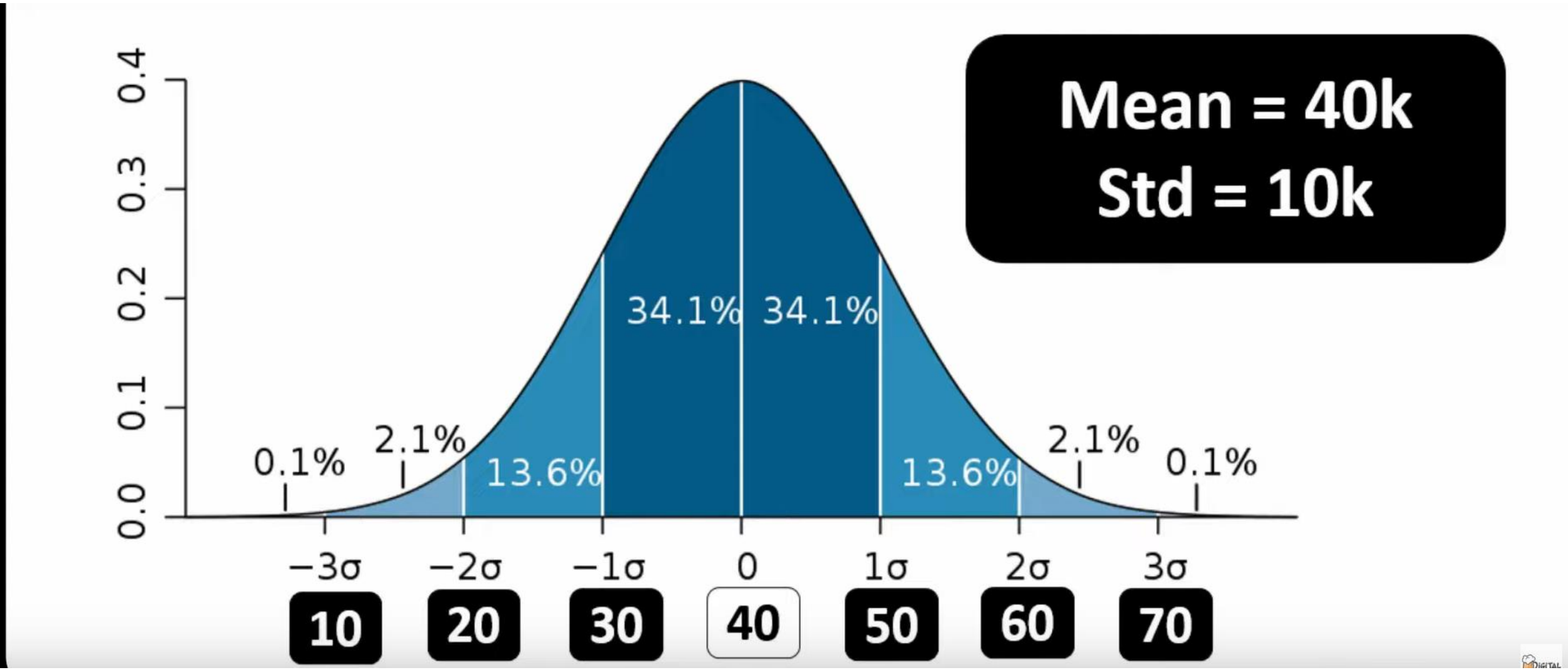


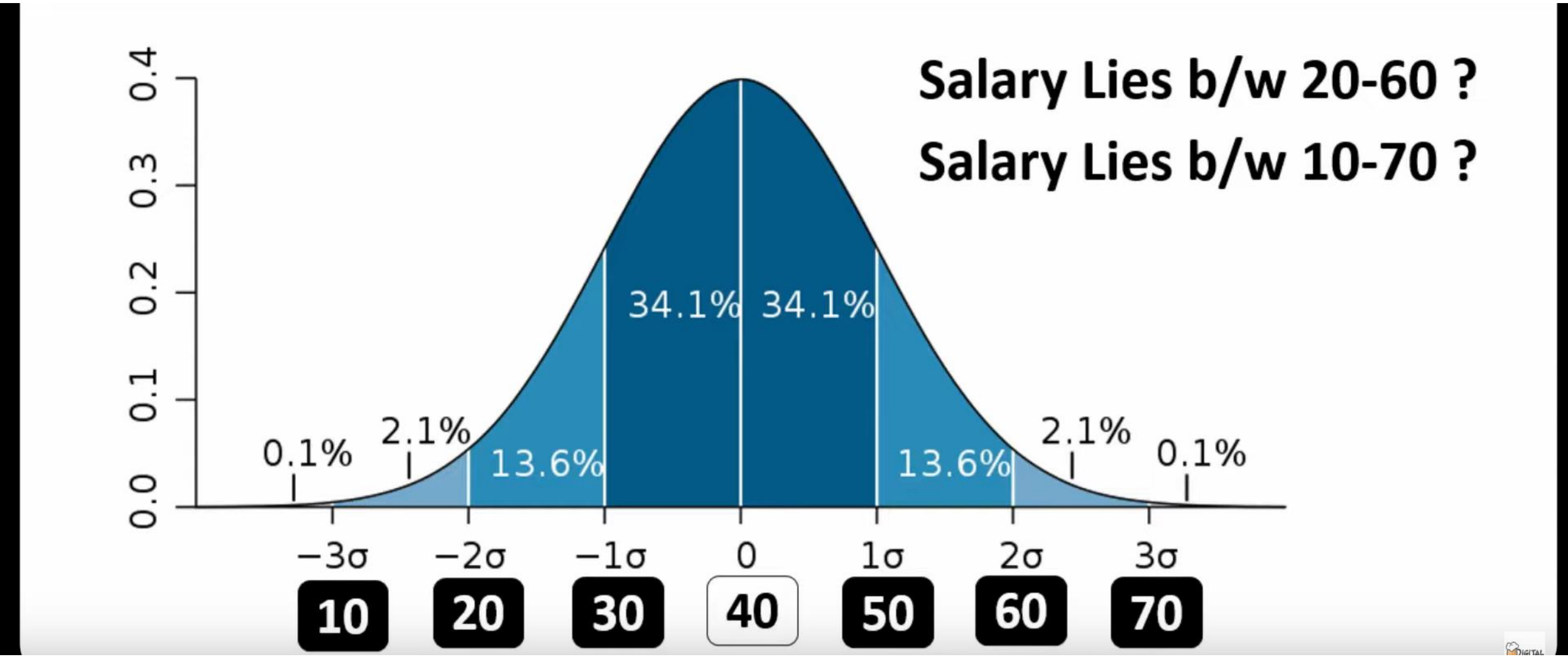
$$P\left(\begin{array}{l} X \geq \mu + k\sigma \\ X \leq \mu - k\sigma \end{array}\right) \leq \frac{1}{k^2}$$

(or)

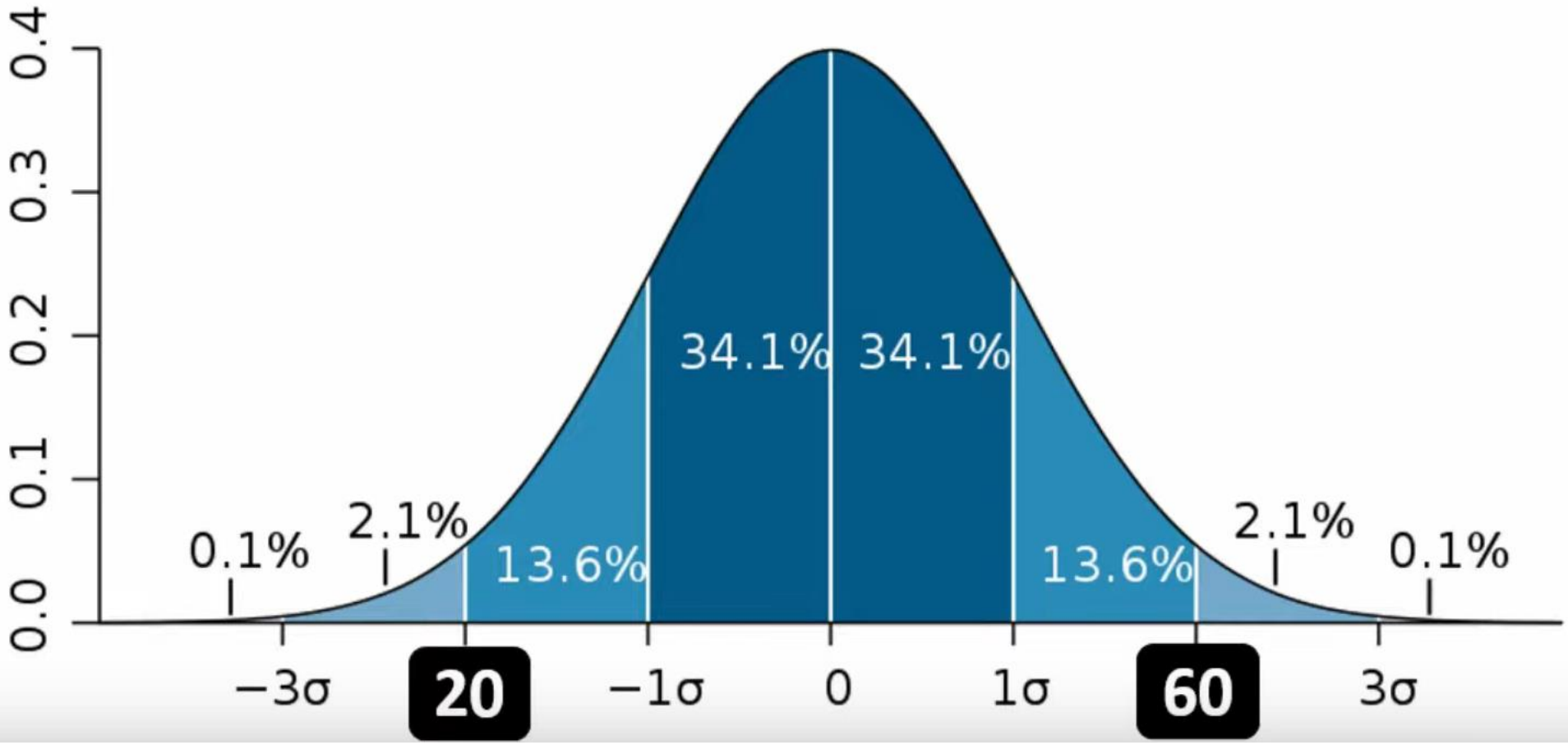
$$P(\mu - k\sigma < X < \mu + k\sigma) > 1 - \frac{1}{k^2}$$


Mean = 40k
Std = 10k
0.0
0.1
0.2
0.3
0.4
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ
40


Mean = 40k
Std = 10k
0.0
0.1
0.2
0.3
0.4
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ
10
20
30
40
50
60
70

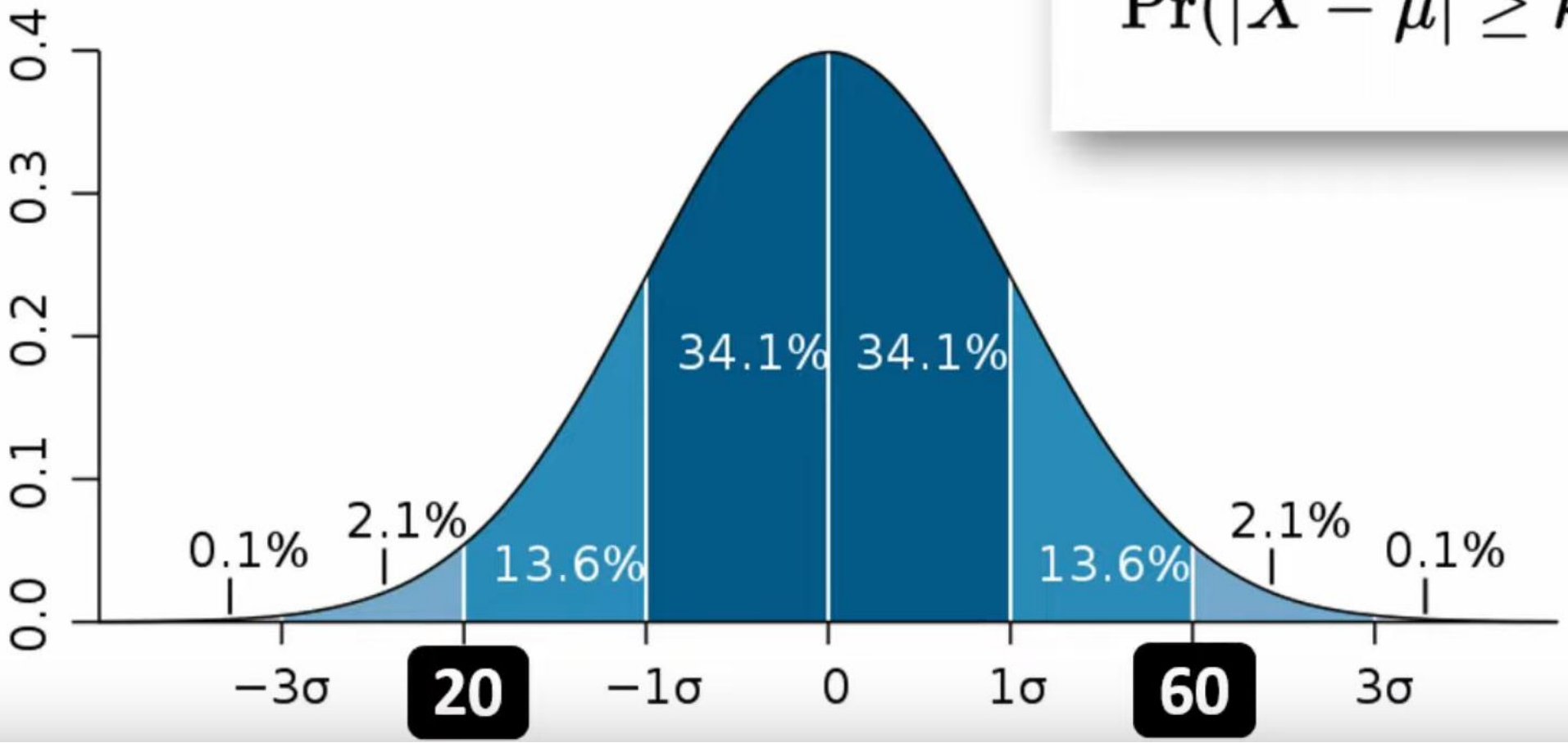

Salary Lies b/w 20-60 ?
Salary Lies b/w 10-70 ?
0.0
0.1
0.2
0.3
0.4
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ
10
20
30
40
50
60
70

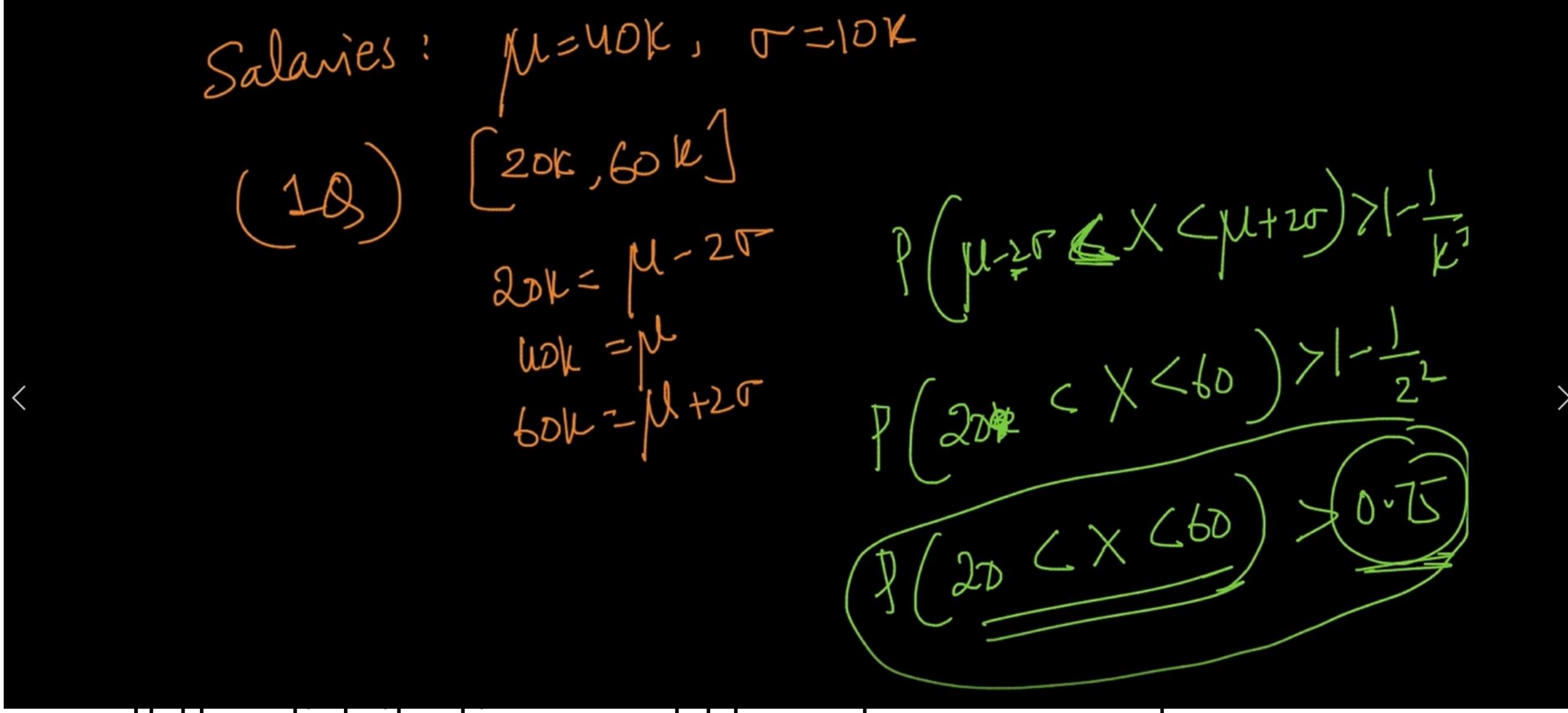
Salary Lies b/w 20-60 ?



Salary Lies b/w 20-60 ?

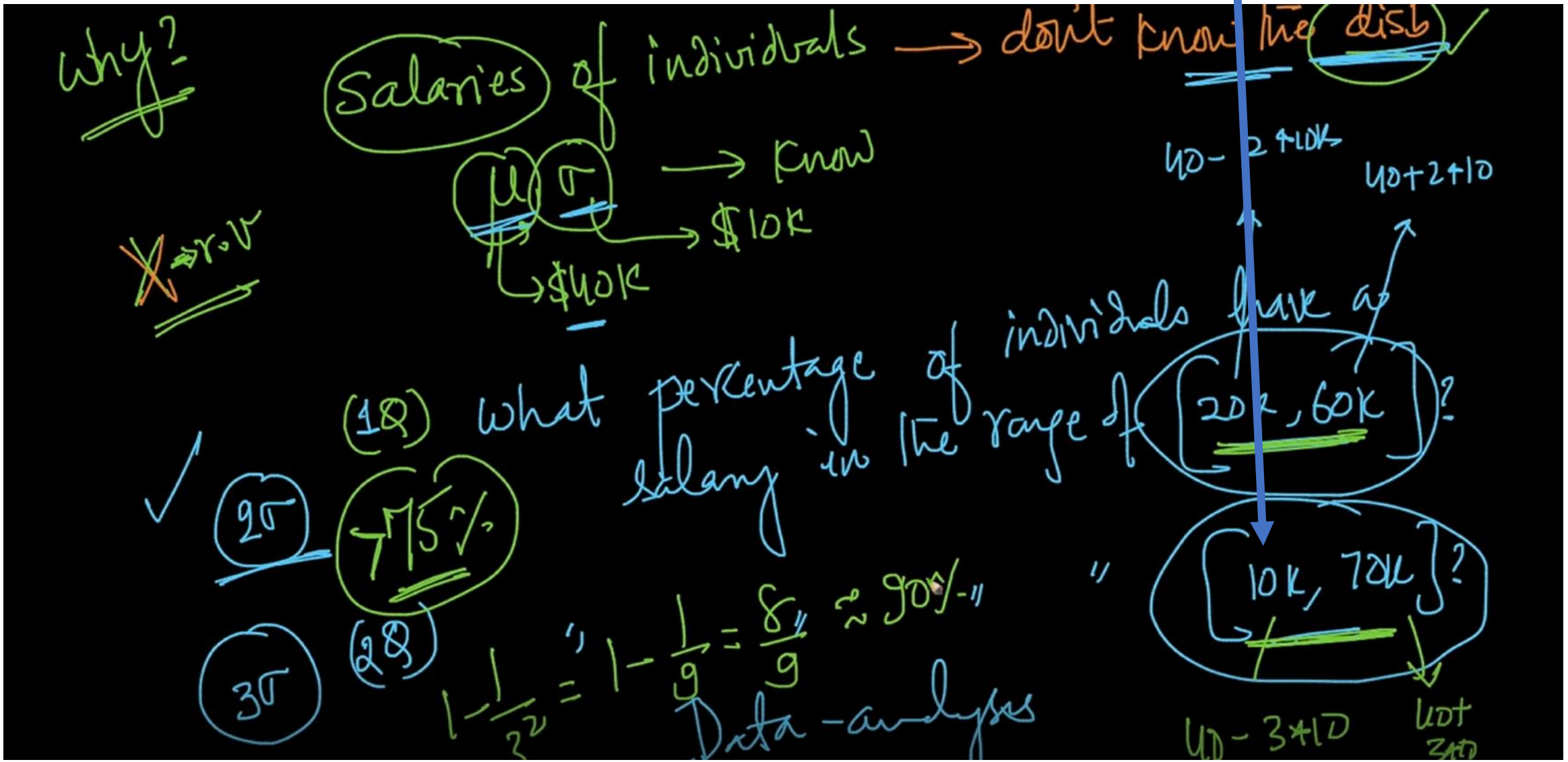
$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

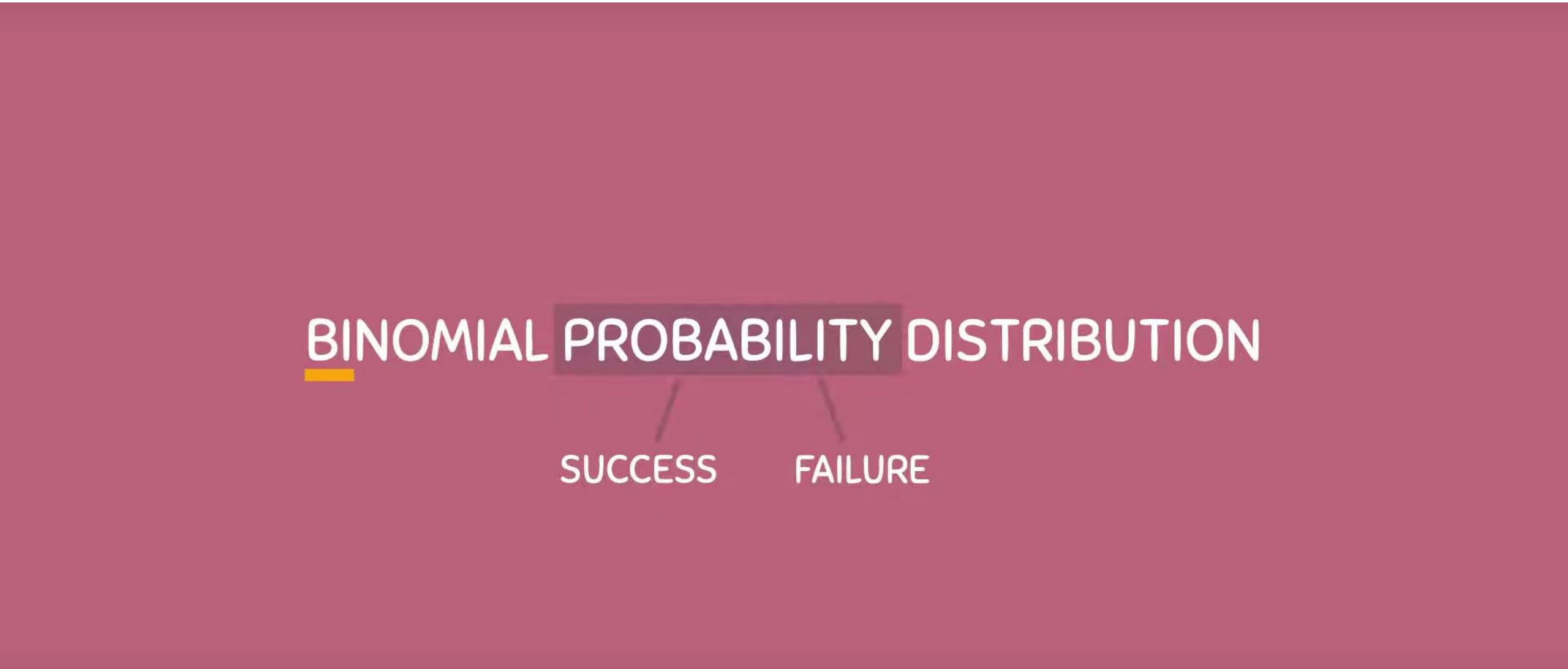




If the distribution would have been normal distribution, then salary would have been 95% but now it is at least 75%.

If we want salary between 10k to 70k




BINOMIAL PROBABILITY DISTRIBUTION
SUCCESS
FAILURE

BINOMIAL PROBABILITY DISTRIBUTION

Two

SUCCESS

BINOMIAL SETTING

1. The number of trials (n), must be fixed
2. There are only two possible outcomes for each trial
    - Success
    - Failure
3. The probability (p) of success must be constant for every trial
4. Each trial must be independent

EXAMPLE

If you flip a regular coin three times, what is the probability of getting exactly one head, and is this a binomial experiment?

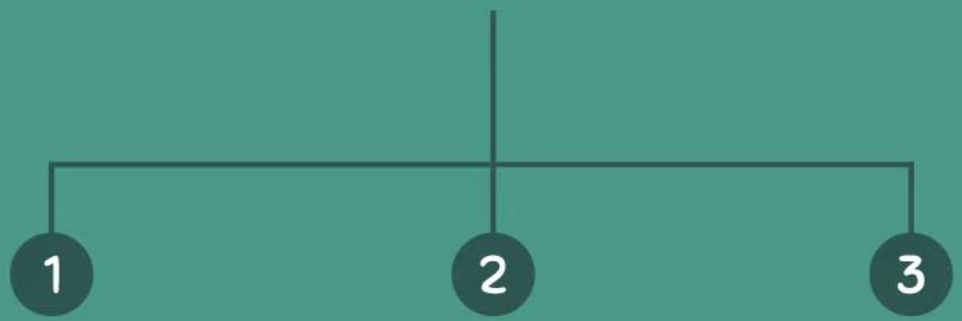
EXAMPLE

If you flip a regular coin three times, what is the probability of getting exactly one head, and is this a binomial experiment?

P(Heads) = 0.5

P(Tails) = 0.5

Three possible ways



P(HTT) = 0.125 P(THT) = 0.125 P(TTH) = 0.125

P(exactly one head) = 0.125 + 0.125 + 0.125 = 0.375

1. The number of trials (n), must be fixed
   - n = 3
2. There are only two possible outcomes for each trial
3. The probability (p) of success must be constant for every trial
4. Each trial must be independent

1 The number of trials (n), must be fixed

- n = 3

2 There are only two possible outcomes for each trial

- Success = Heads
- Failure = Not getting heads

3 The probability (p) of success must be constant for every trial

4 Each trial must be independent

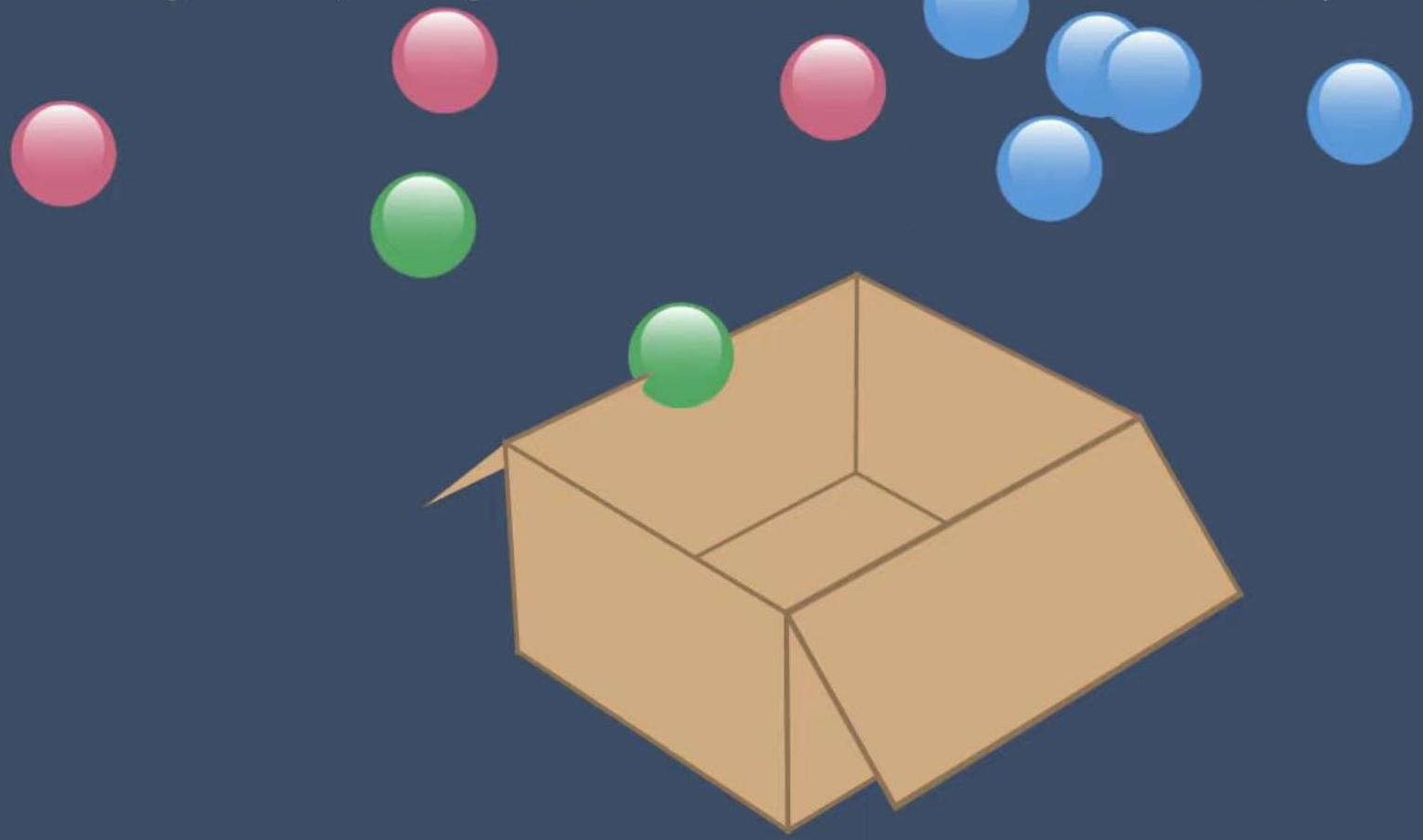
1. The number of trials (n), must be fixed
   - n = 3
2. There are only two possible outcomes for each trial
   - Success = Heads
   - Failure = Tails
3. The probability (ρ) of success must be constant for every trial
   - ρ = 0.5
4. Each trial must be independent

1. The number of trials (n), must be fixed
   - n = 3
2. There are only two possible outcomes for each trial
   - Success = Heads
   - Failure = Tails
3. The probability (p) of success must be constant for every trial
   - p = 0.5
4. Each trial must be independent

- ✓ The number of trials (n), must be fixed
  - $n = 3$
- ✓ There are only two possible outcomes for each trial
  - Success = Heads
  - Failure = Tails
- ✓ The probability (p) of success must be constant for every trial
  - $p = 0.5$
- ✓ Each trial must be independent
- ✓ Binomial Experiment

EXAMPLE

Suppose we have 10 marbles in a box. We have 3 pink marbles, 2 green marbles, and 5 blue marbles. If we pick out five marbles with replacement, what is the probability of drawing exactly two green marbles, and is this a binomial experiment?



1. Is there a fixed number of trials? Yes.
   - $n = 5$

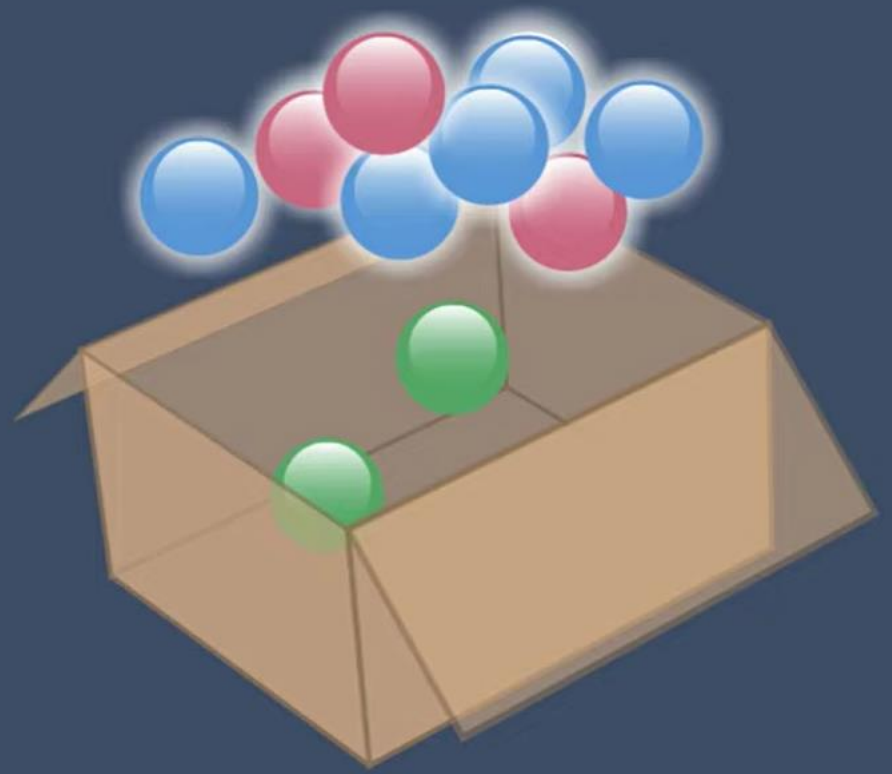
2. Are the two possible outcomes a success and a failure? Yes.
   - Success = Green marble
   - Failure = Not getting a green marble

3. Is the probability of success constant for each trial? Yes.
   - $p = 0.2$

4. Are trials independent of each other? Yes.

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
Outcome 1	G	G	–	–	–
Outcome 2	G	–	G	–	–
Outcome 3	G	–	–	G	–
Outcome 4	G	–	–	–	G
Outcome 5	–	G	G	–	–
Outcome 6	–	G	–	G	–
Outcome 7	–	G	–	–	G
Outcome 8	–	–	G	G	–
Outcome 9	–	–	G	–	G
Outcome 10	–	–	–	G	G

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
Outcome 1	S	S	F	F	F
Outcome 2	S	F	S	F	F
Outcome 3	S	F	F	S	F
Outcome 4	S	F	F	F	S
Outcome 5	F	S	S	F	F
Outcome 6	F	S	F	S	F
Outcome 7	F	S	F	F	S
Outcome 8	F	F	S	S	F
Outcome 9	F	F	S	F	S
Outcome 10	F	F	F	S	S



P(Success) = P(Green) = 0.2
P(Success) = 0.2

P(Failure) = P(Not Green) = 0.8

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Probability
Outcome 1	S	S	F	F	F	

$$0.2 \times 0.2 \times 0.8 \times 0.8 \times 0.8 = 0.02048$$

P(Success) = P(Green) = 0.2
P(Success) = 0.2

P(Failure) = P(Not Green) = 0.8
P(Failure) = 0.8

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Probability
Outcome 1	S	S	F	F	F	0.02048
Outcome 2	S	F	S	F	F	0.02048
Outcome 3	S	F	F	S	F	0.02048
Outcome 4	S	F	F	F	S	0.02048
Outcome 5	F	S	S	F	F	0.02048
Outcome 6	F	S	F	S	F	0.02048
Outcome 7	F	S	F	F	S	0.02048
Outcome 8	F	F	S	S	F	0.02048
Outcome 9	F	F	S	F	S	0.02048
Outcome 10	F	F	F	S	S	0.02048

	Probability
Outcome 1	0.02048
Outcome 2	0.02048
Outcome 3	0.02048
Outcome 4	0.02048
Outcome 5	0.02048
Outcome 6	0.02048
Outcome 7	0.02048
Outcome 8	0.02048
Outcome 9	0.02048
Outcome 10	0.02048

Sum

P(Exactly two Green marbles) = 0.2048

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

BINOMIAL FORMULA

Binomial Distribution Formula

Binomial Distribution $P(k) = {}^{n}C_{k} \cdot p^{k} (1 - p)^{n-k}$

Or, $P(r) = [n!/r!(n-r)!] \cdot p^{r} (1 - p)^{n-r}$

k = number of successes
n = number of trials
p = probability of success

$$\binom{n}{k}$$

Combination Formula

Binomial Distribution Formula

Binomial Distribution $P(k) = {}^{n}C_{k} \cdot p^{k} (1 - p)^{n-k}$

Or, $P(r) = [n!/r!(n-r)!] \cdot p^{r} (1 - p)^{n-r}$

EXAMPLE

If we pick out five marbles with replacement, what is the probability of drawing exactly two green marbles

- $n = 5$
- $k = 2$
- $p = 0.2$

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$P(2) = \binom{5}{2} 0.2^2 (1-0.2)^{5-2}$$

$$= 0.2048$$

k = number of successes
n = number of trials
p = probability of success

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

BINOMIAL FORMULA

Flipping a coin two times..

Success = Heads

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- $k = 0, 1$ or 2
- $n = 2$
- $p = 0.5$

Success = Heads

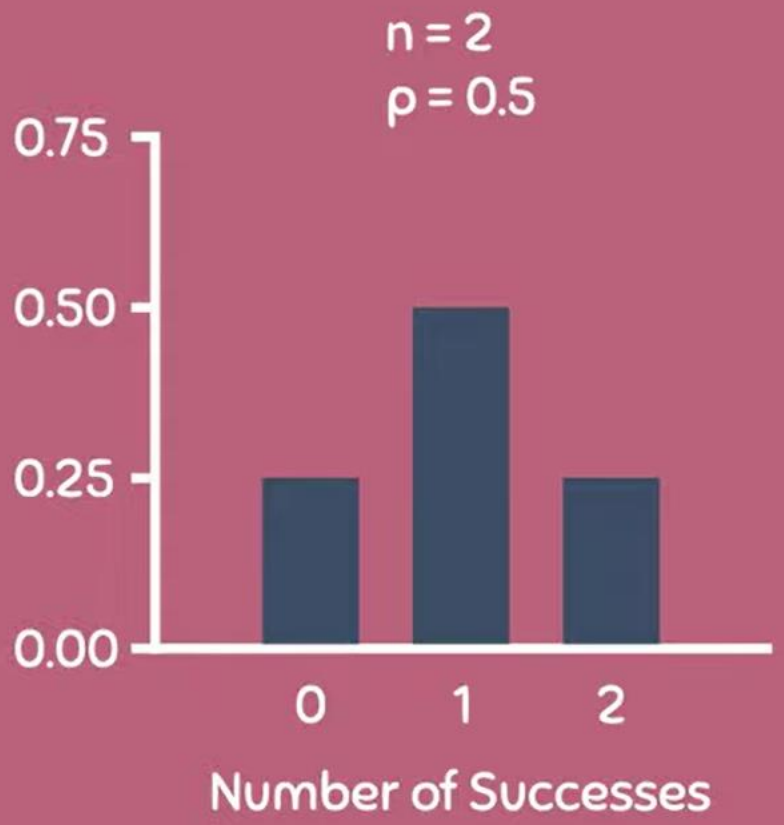
$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

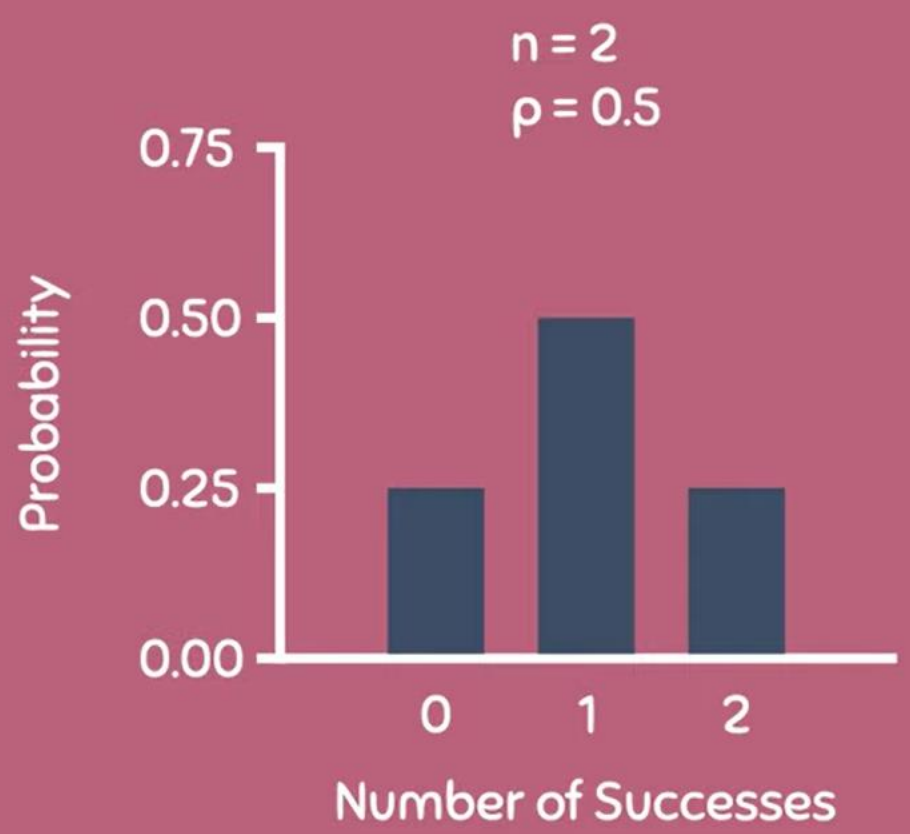
- $k = 0, 1$ or 2
- $n = 2$
- $p = 0.5$

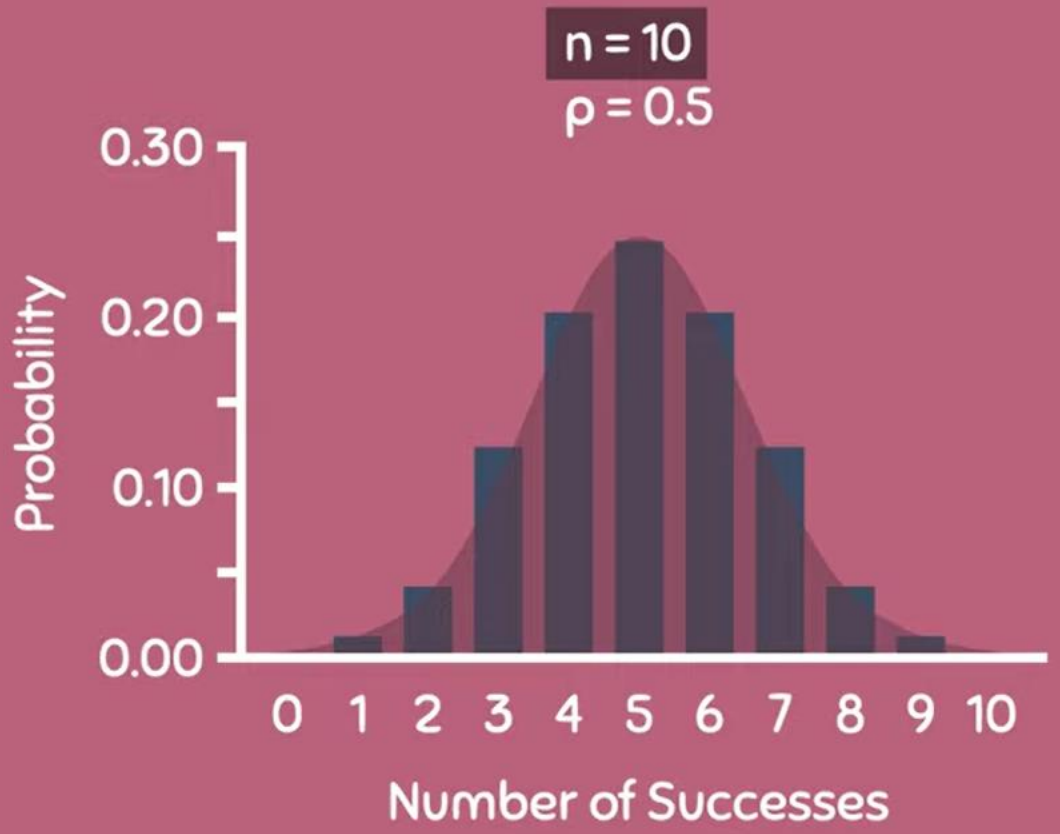
$$P(0) = \binom{2}{0} 0.5^0 (1-0.5)^{2-0} = 0.25$$

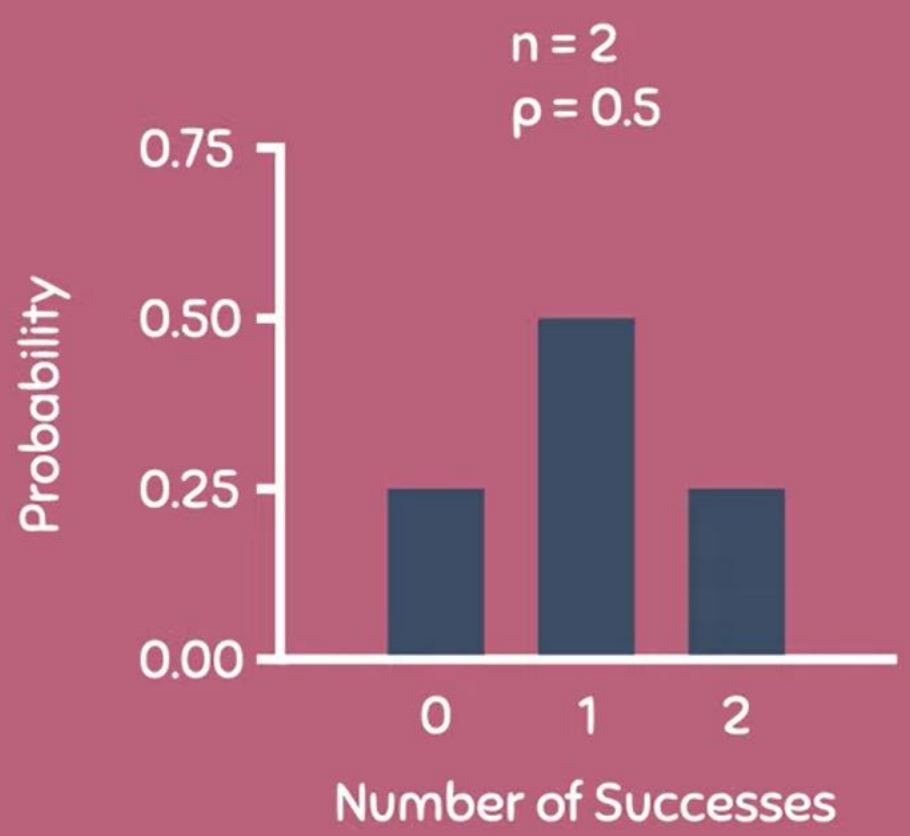
$$P(1) = \binom{2}{1} 0.5^1 (1-0.5)^{2-1} = 0.50$$

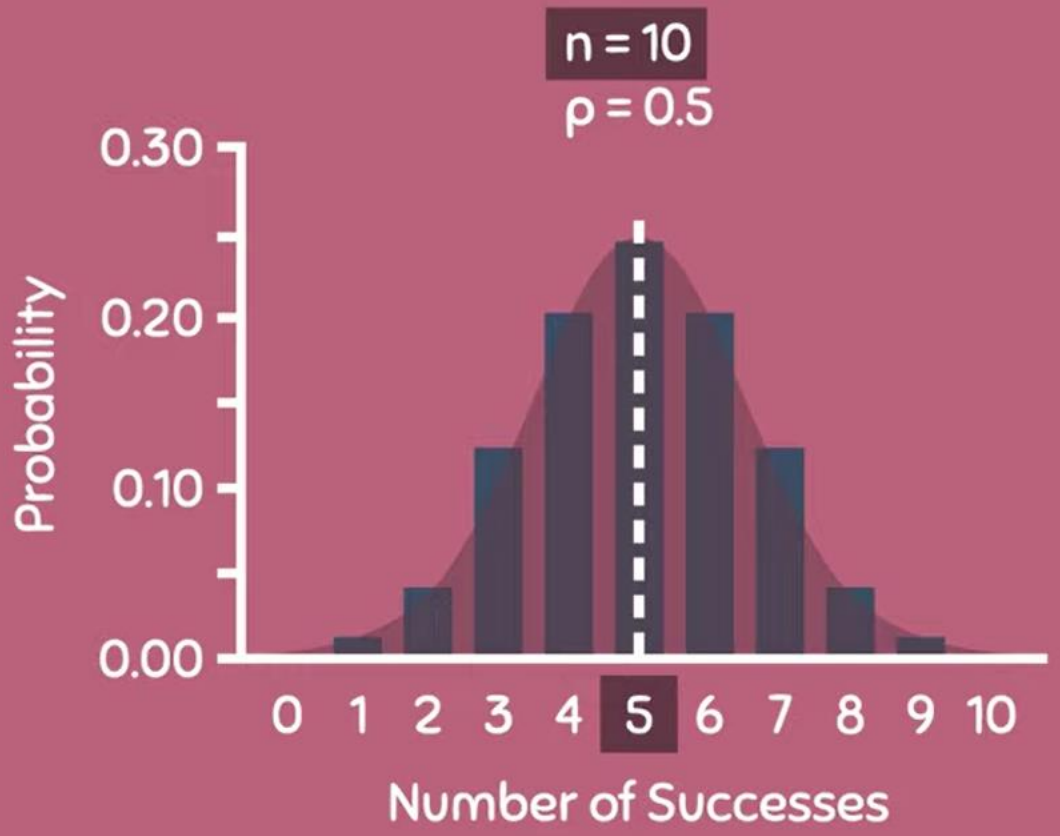
$$P(2) = \binom{2}{2} 0.5^2 (1-0.5)^{2-2} = 0.25$$


n = 2
p = 0.5
0.75
0.50
0.25
0.00
0
1
2
Number of Successes


n = 2
p = 0.5
0.75
0.50
0.25
0.00
Probability
0
1
2
Number of Successes

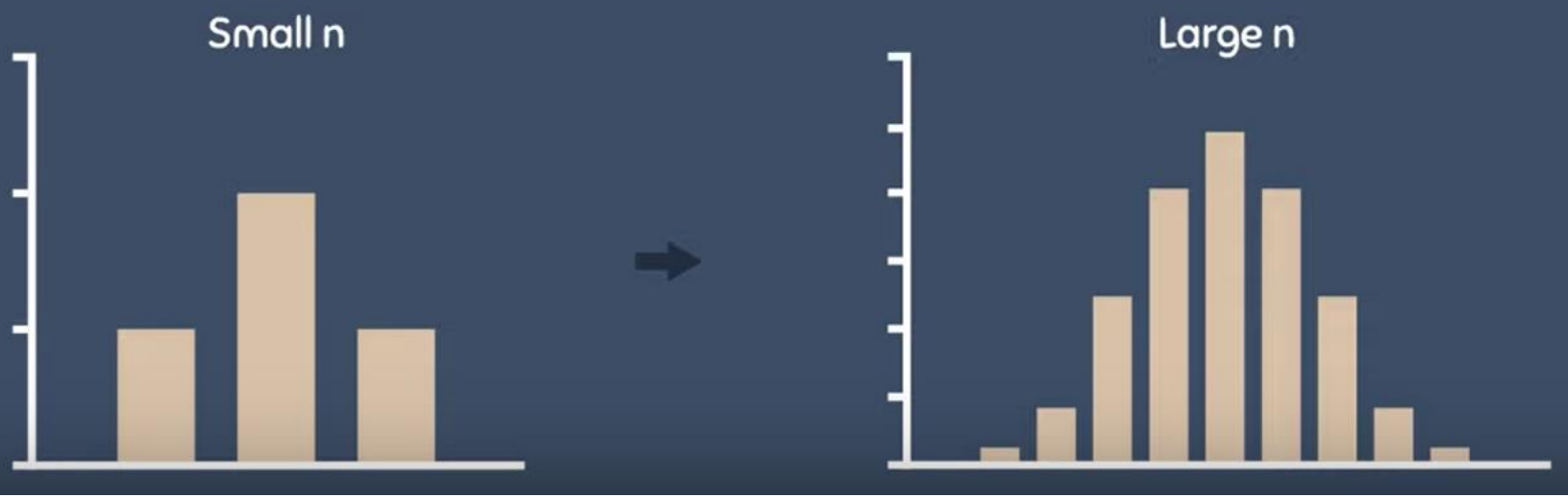

n = 10
p = 0.5
0.30
0.20
0.10
0.00
Probability
0 1 2 3 4 5 6 7 8 9 10
Number of Successes

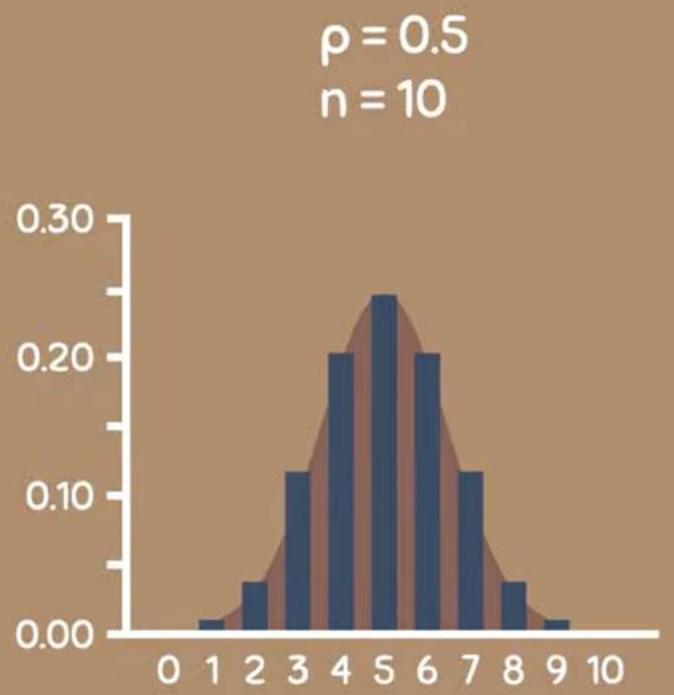

n = 2
p = 0.5
0.75
0.50
0.25
0.00
Probability
0
1
2
Number of Successes

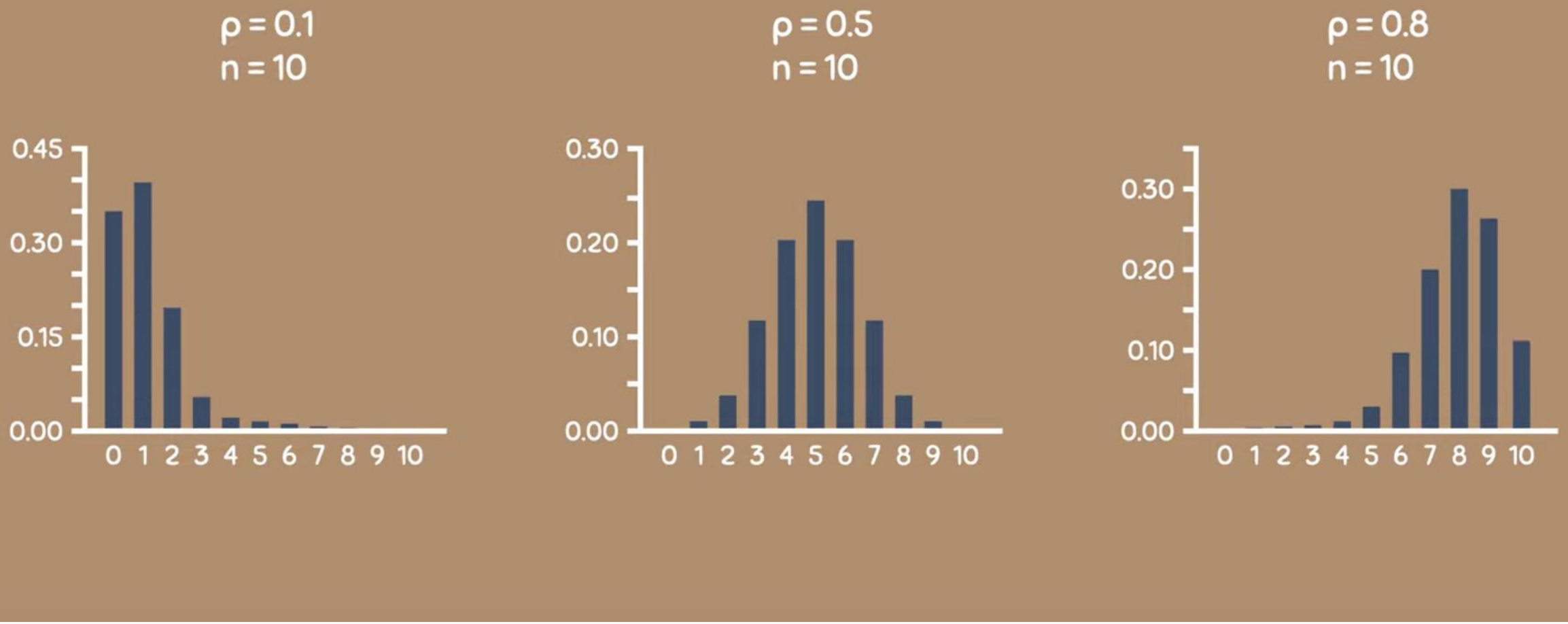

n = 10
p = 0.5
0.30
0.20
0.10
0.00
Probability
0 1 2 3 4 5 6 7 8 9 10
Number of Successes

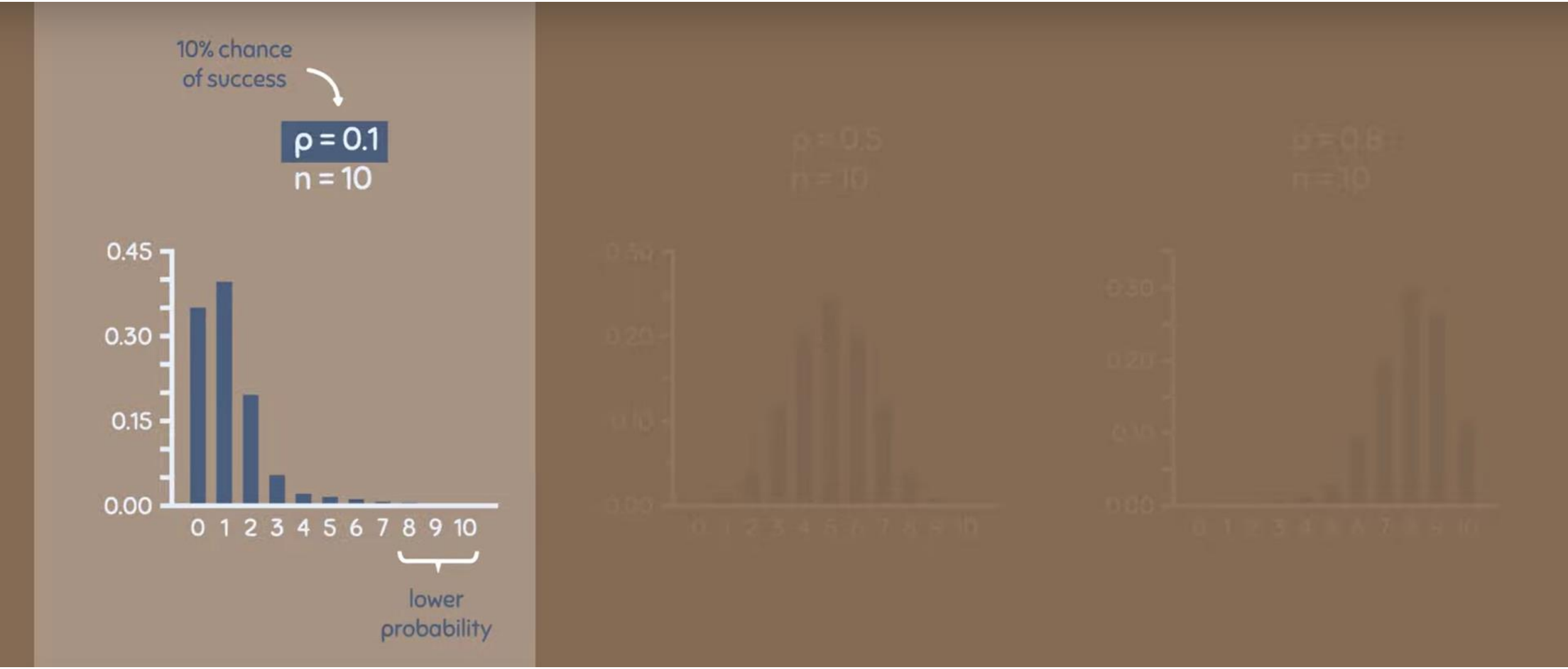
If a variable X follows a Binomial Distribution

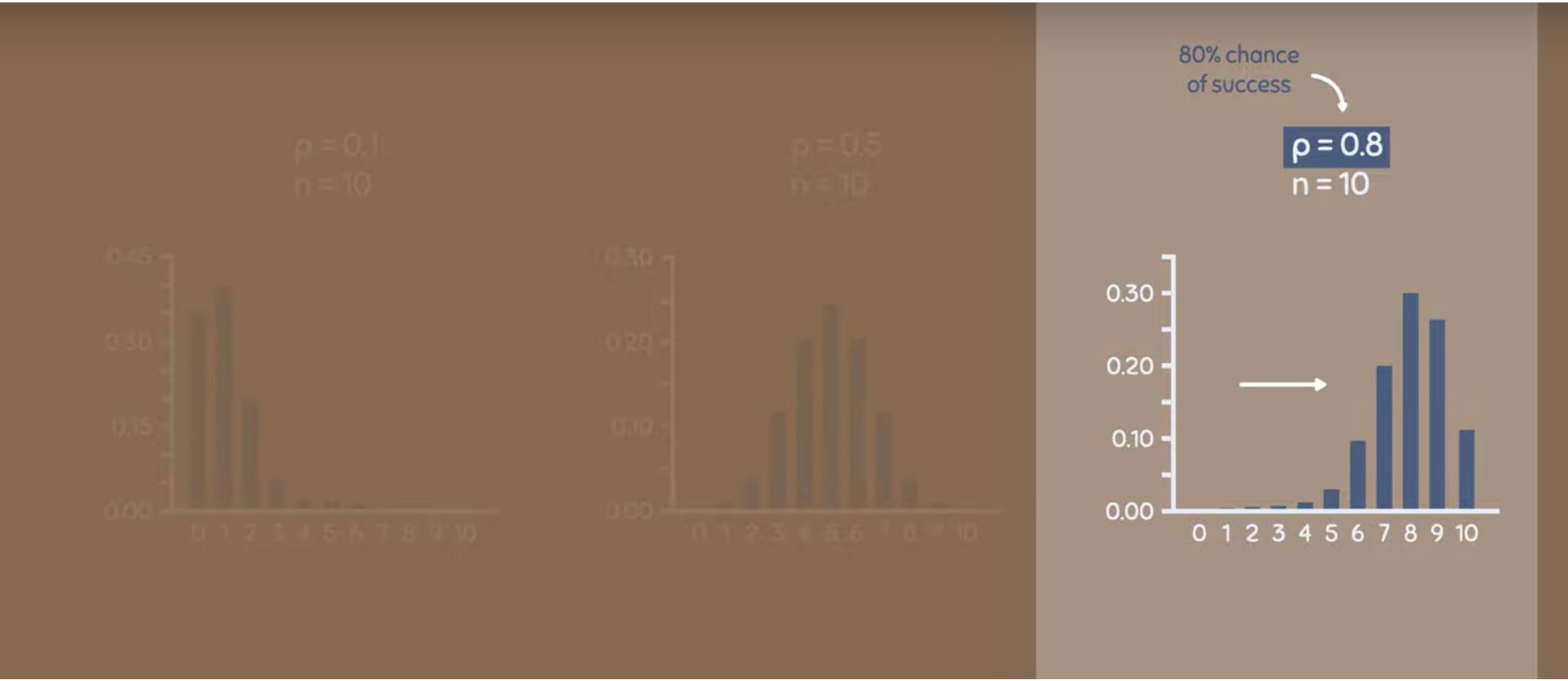
Mean	$\mu = np$
Variance	$\sigma^2 = np(1-p)$
Standard Deviation	$\sigma = \sqrt{np(1-p)}$



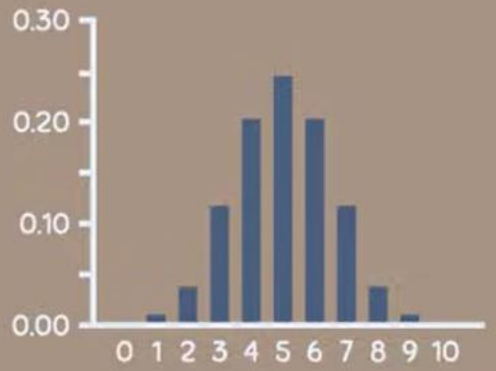

p = 0.5
n = 10
0.30
0.20
0.10
0.00
0 1 2 3 4 5 6 7 8 9 10


ρ = 0.1
n = 10
0.45
0.30
0.15
0.00
0 1 2 3 4 5 6 7 8 9 10
ρ = 0.5
n = 10
0.30
0.20
0.10
0.00
0 1 2 3 4 5 6 7 8 9 10
ρ = 0.8
n = 10
0.30
0.20
0.10
0.00
0 1 2 3 4 5 6 7 8 9 10

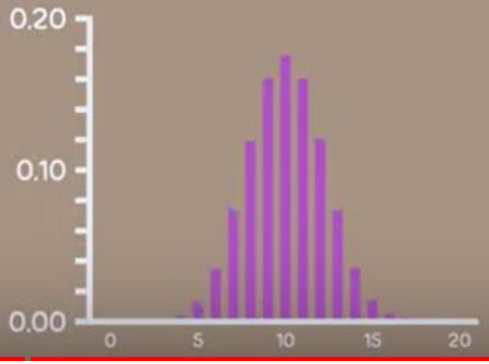

10% chance
of success
p = 0.1
n = 10
0.45
0.30
0.15
0.00
0 1 2 3 4 5 6 7 8 9 10
lower
probability


80% chance
of success
ρ = 0.8
n = 10
0.30
0.20
0.10
0.00
0 1 2 3 4 5 6 7 8 9 10

$p = 0.5$
$n = 10$



$p = 0.5$
$n = 20$



INDEPENDENT EVENTS

DEPENDENT EVENTS

INDEPENDENT EVENTS

Refer to the occurrence of one event
not affecting the probability of another event

ROLLING A DIE

FLIPPING A COIN

INDEPENDENT EVENTS

TWO INDEPENDENT EVENTS:

$$P(A \cap B) = P(A) \times P(B)$$

Probability of A and B

Probability of event A

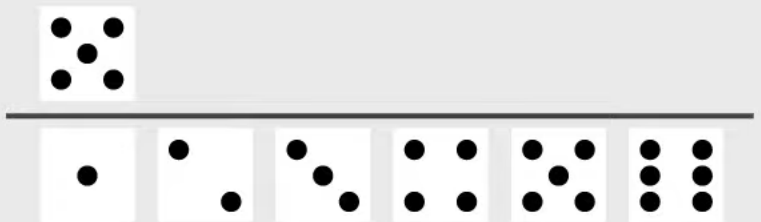
Probability of event B

EXAMPLE

If you roll a six-sided die and flip a coin, what is the probability of rolling a five and getting heads?

$$P(\text{Event}) = \frac{\text{total \# of favourable outcomes}}{\text{total \# of possible outcomes}}$$

$$P(A \cap B) = P(A) \times P(B)$$

P(Rolling a 5) 

EXAMPLE | If you roll a six-sided die and flip a coin, what is the probability of rolling a five and getting heads?

$$P(\text{Event}) = \frac{\text{total \# of favourable outcomes}}{\text{total \# of possible outcomes}}$$

$$P(\text{Rolling a 5}) = \frac{1}{6}$$

$$P(\text{Getting heads}) = \frac{1}{2}$$

$$P(A \cap B) = P(A) \times P(B)$$

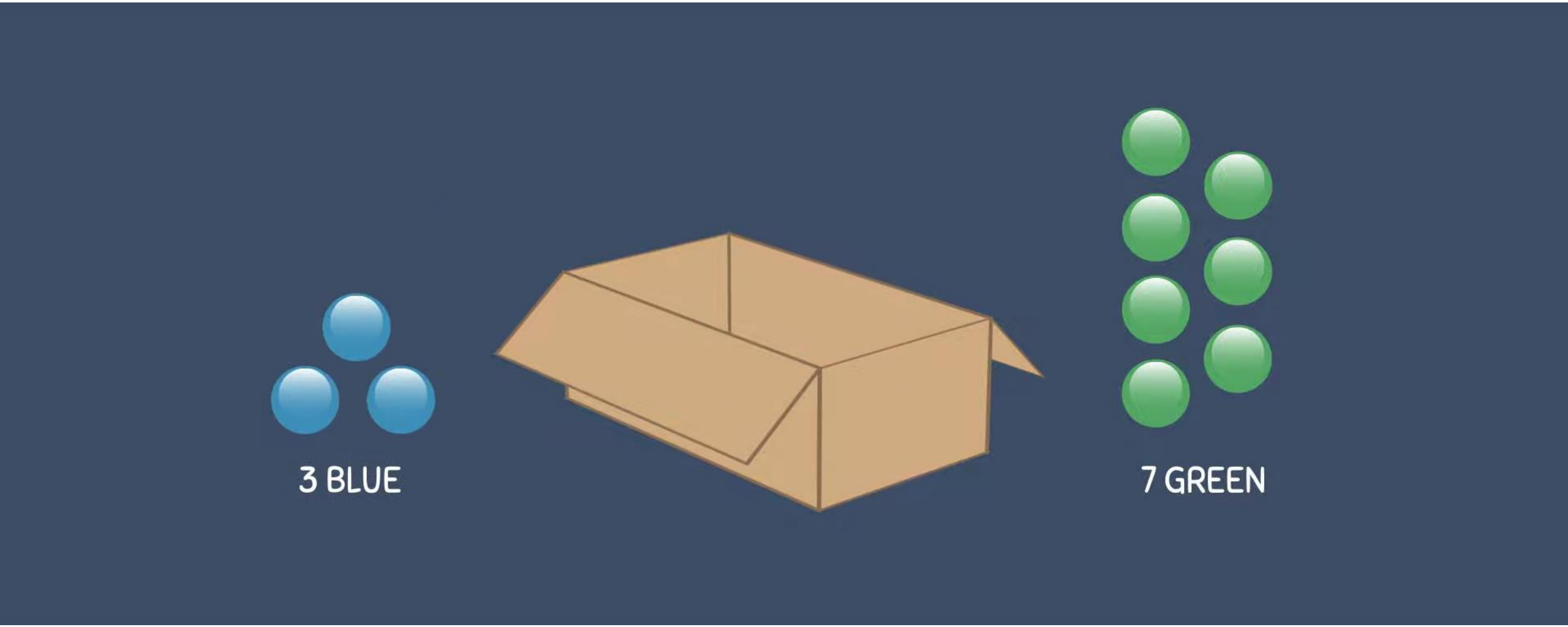
$$= \frac{1}{6} \times \frac{1}{2}$$

$$= \frac{1}{12}$$

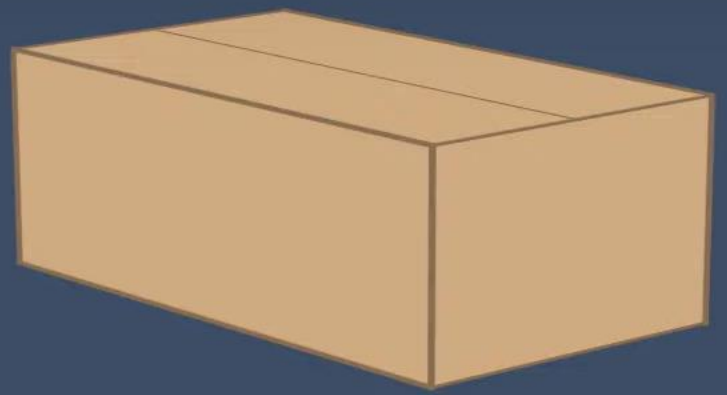
$$P(\text{Rolling a 5 and Getting heads}) = \frac{1}{12}$$

$$= 0.0833$$

DEPENDENT EVENTS


3 BLUE
7 GREEN

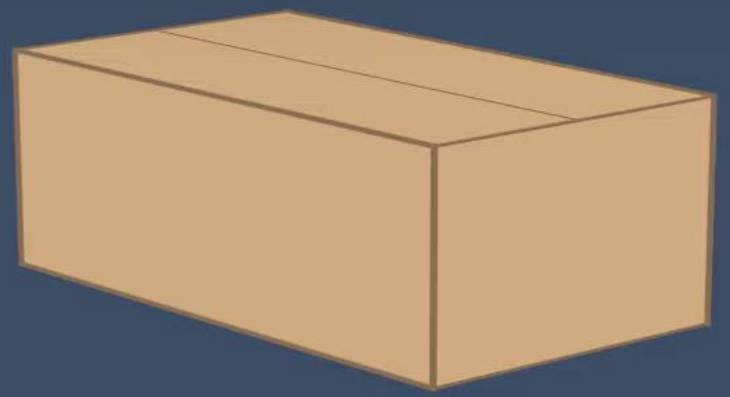
P(Blue) = 3/10



P(Green) = 7/10

If we randomly select two marbles from this box, what is the probability of drawing a green marble and then a blue marble, without replacement?

P(Blue) = 3/10



P(Green) = 7/10

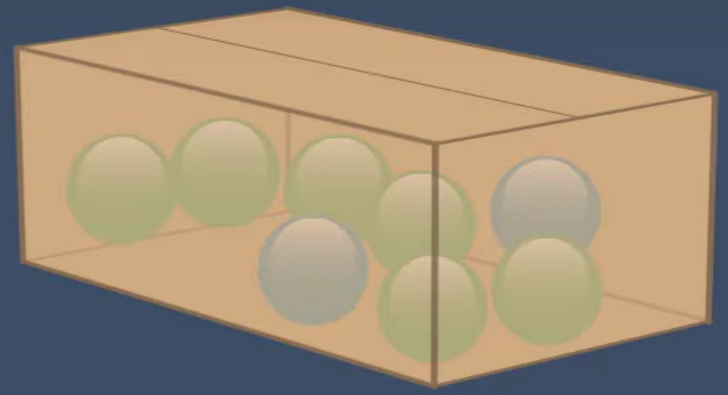
DEPENDENT EVENT

If we randomly select two marbles from this box, what is the probability of drawing a green marble and then a blue marble, without replacement?

$$P(A \cap B) = P(A) \times P(B)$$

$$= 7/10 \times 3/10$$

P(Blue) = 3/10



P(Green) = 7/10

DEPENDENT EVENT

If we randomly select two marbles from this box, what is the probability of drawing a green marble and then a blue marble, without replacement?

FIRST EVENT

P(GREEN) = 7/10

= 0.7

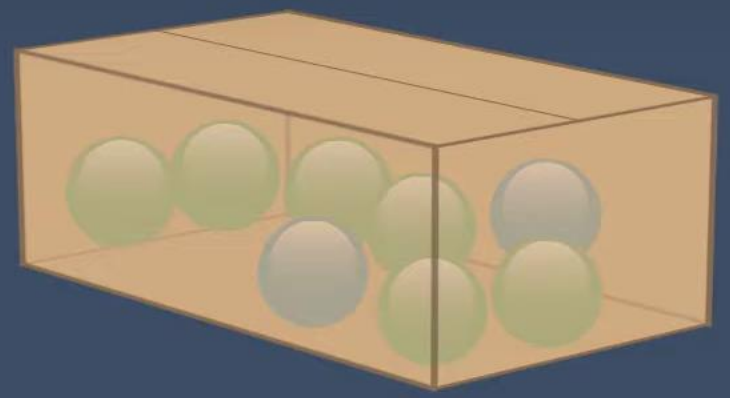
SECOND EVENT

P(BLUE) = 3/9

= 0.33

increased

P(Blue) = 3/10



P(Green) = 7/10

DEPENDENT EVENT

If we randomly select two marbles from this box, what is the probability of drawing a green marble and then a blue marble, without replacement?

$$P(A \cap B) = P(A) \times P(B)$$

$$= 7/10 \times 3/9$$

$$= 7/30$$

$$= 0.233$$

P(BLUE) = 3/9

P(GREEN) = 7/10