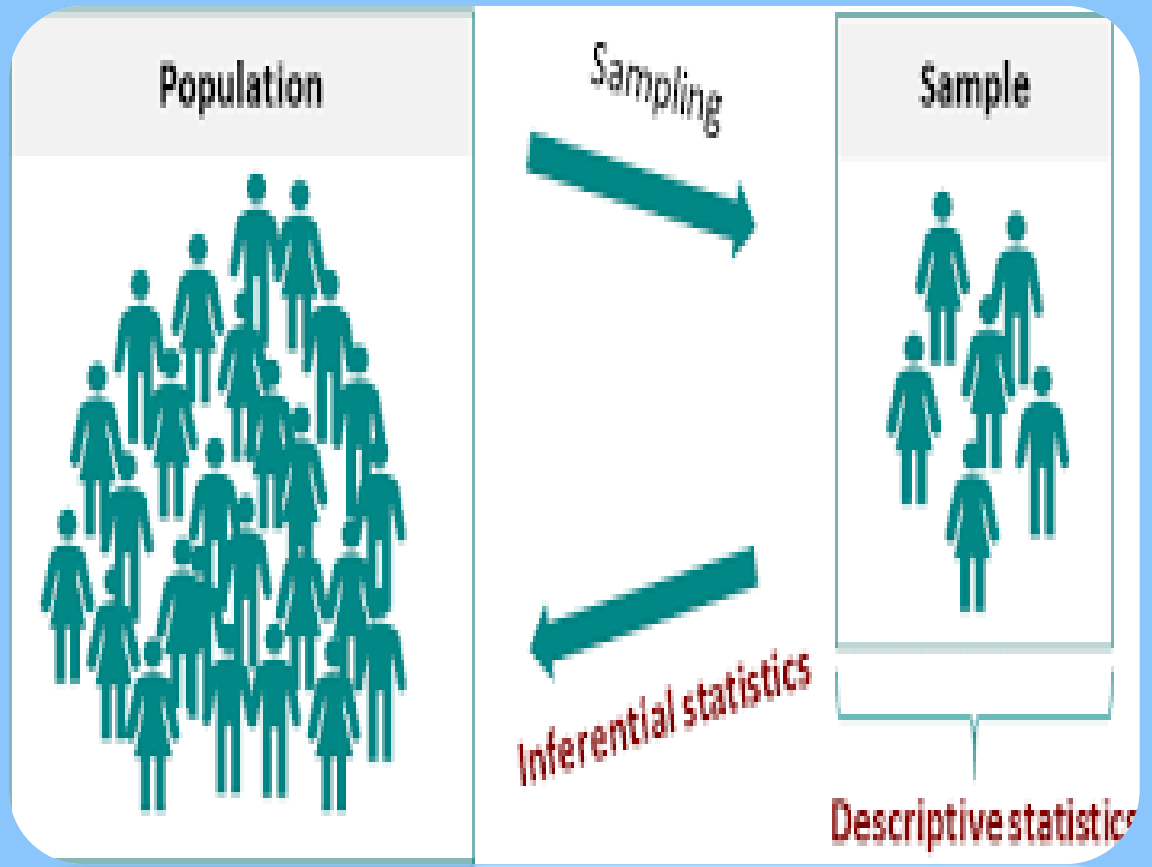
دورهٔ آموزشی «علم داده»

Data Science Course



جلسهٔ نهم:

آمار استنتاجی

Inferential Statistics

مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبدکاووس
پائیز ۱۳۹۹

About me

Mohammad Fozouni (Ph.D.)
Dep. of Math. & Stat.
Gonbad Kavous University

- fozouni@hotmail.com
- https://m-fozouni.ir
- http://profs.gonbad.ac.ir/fozouni/en

#data_science_fozouni

تاکنون چه آموختیم؟

1. مقدمه
2. احتمالات – ترکیب‌شناسی
3. احتمالات – قضیه بیز
4. توزیع‌های احتمال – گسسته و پیوسته
5. یک مثال عملی در خصوص کاربرد احتمالات – مثال فیفا ۲۰۱۹
6. چرا باید احتمالات بخوانیم؟ کاربرد در سرمایه‌گذاری، آمار و دیتا ساینس
7. آمار توصیفی و تحلیل داده‌ها

راجع به پایتون چه مواردی را بحث کردیم؟

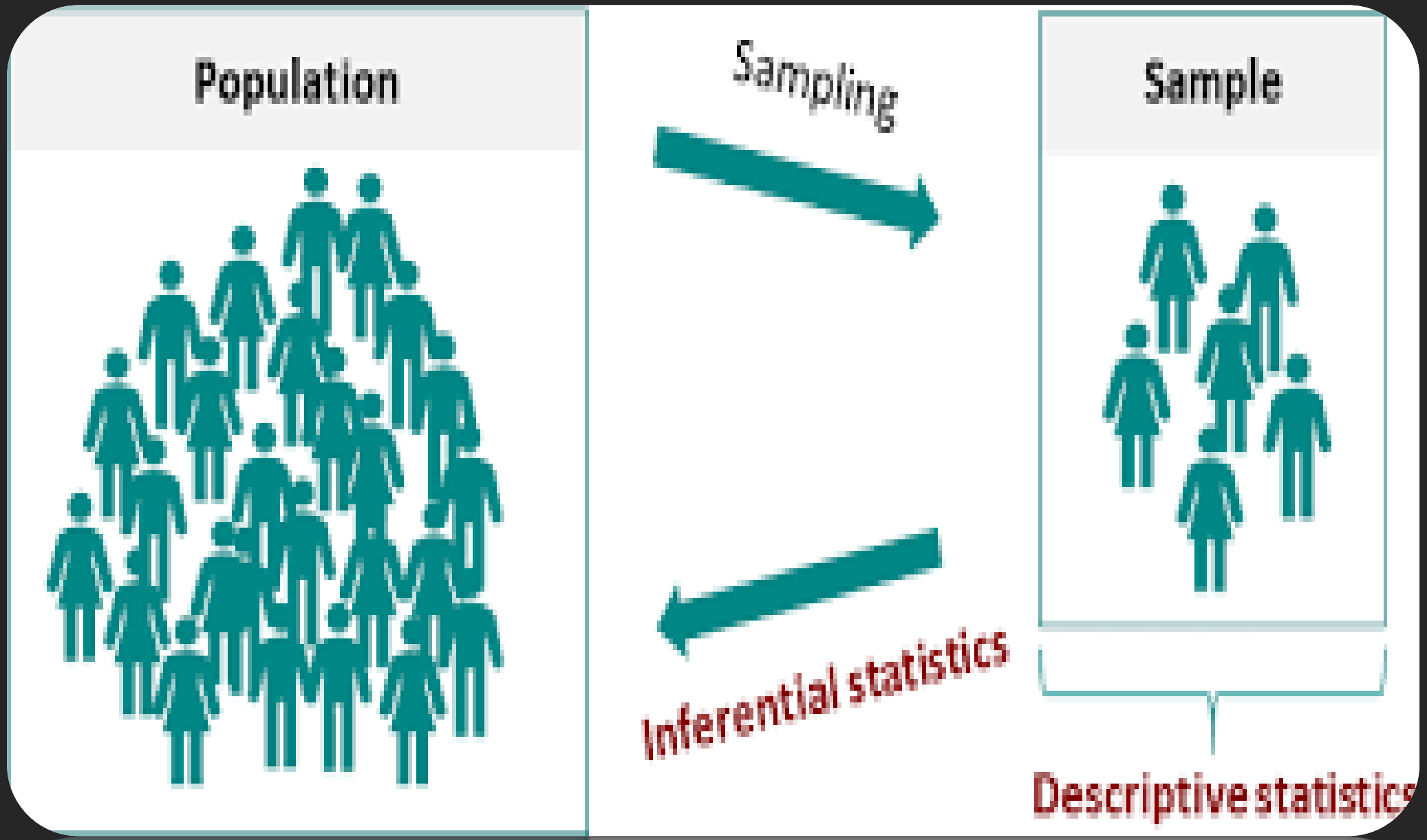
1. مقدمه و نصب ژوپیتر نوت‌بوک
2. متغیرها
3. عملگرها
4. شرطی‌ها
5. توابع
6. دنباله‌ها
7. حلقه‌ی تکرار
8. مباحث پیشرفته مثل OOP و ...

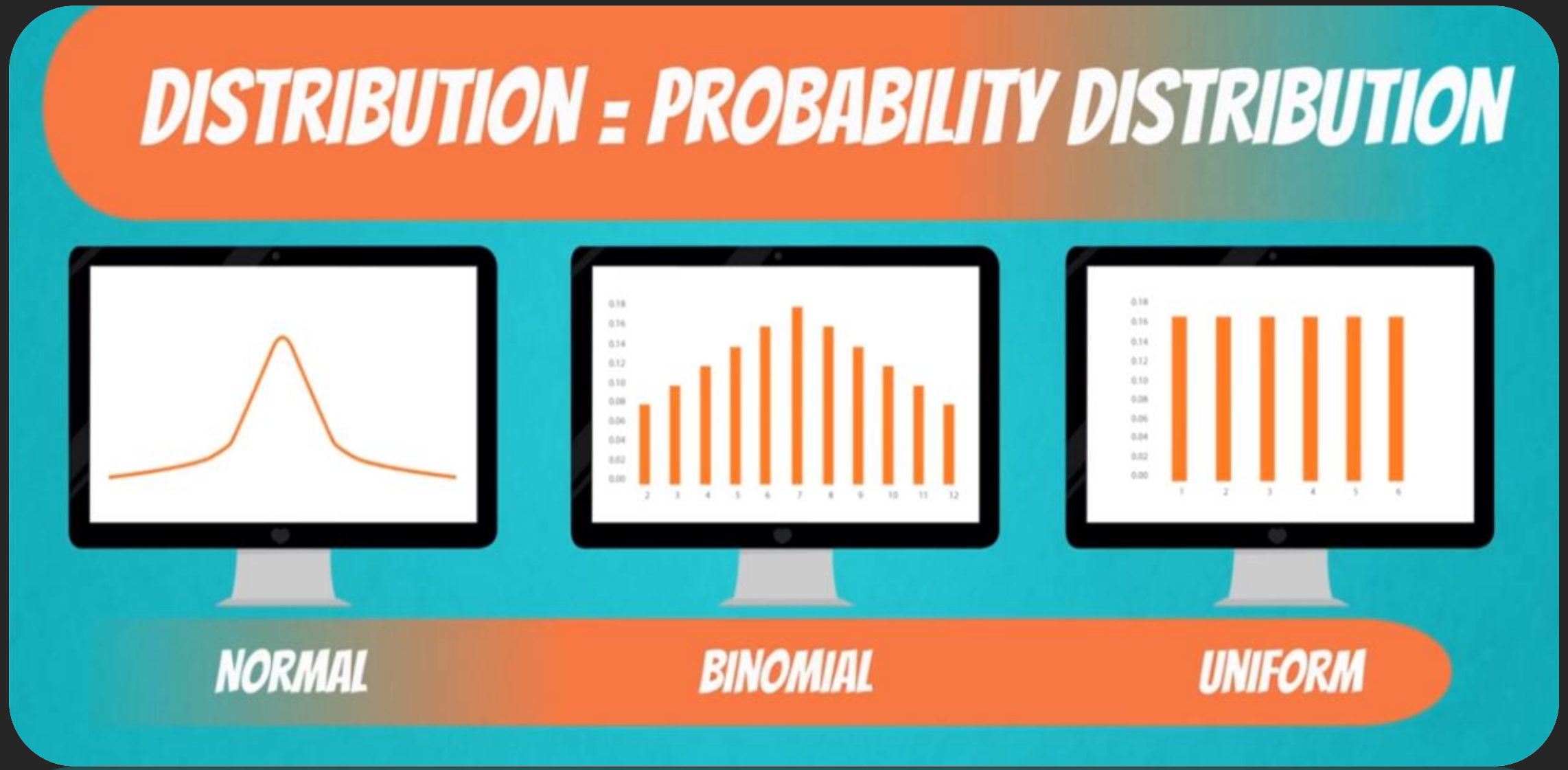
Probability theory

Distributions

INFERENTIAL STATISTICS


IN STATISTICS
DISTRIBUTION
=
PROBABILITY DISTRIBUTION

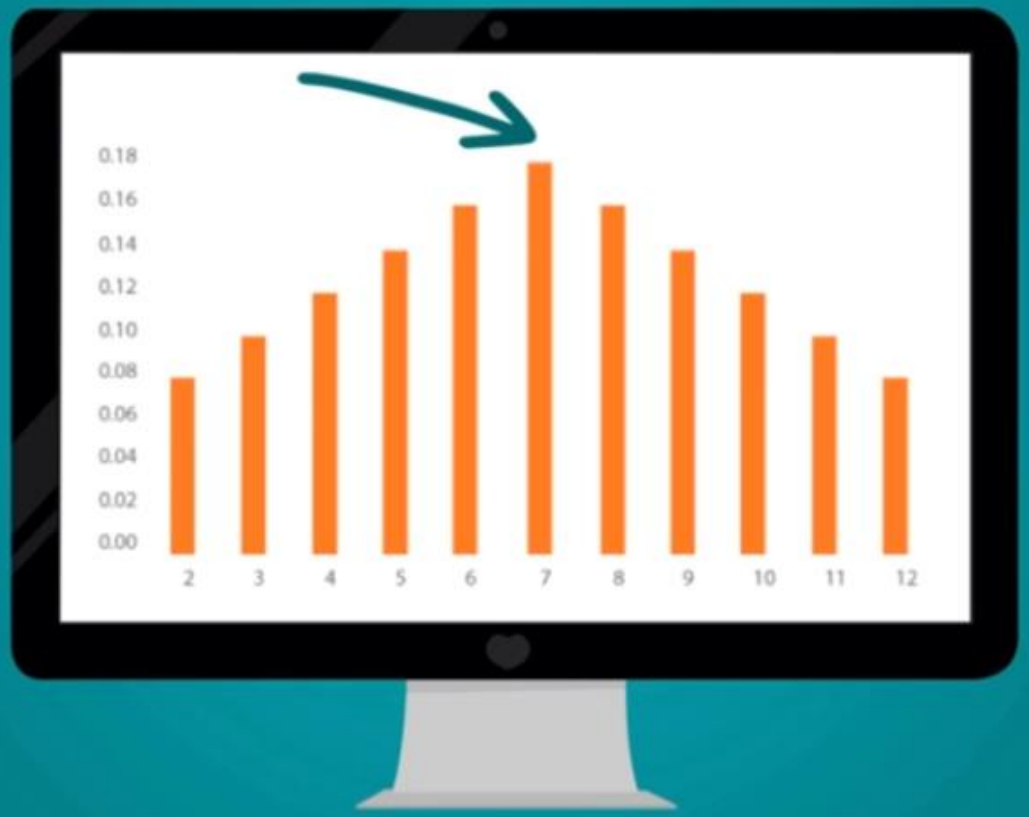

Population
Sampling
Sample
Inferential statistics
Descriptive statistics

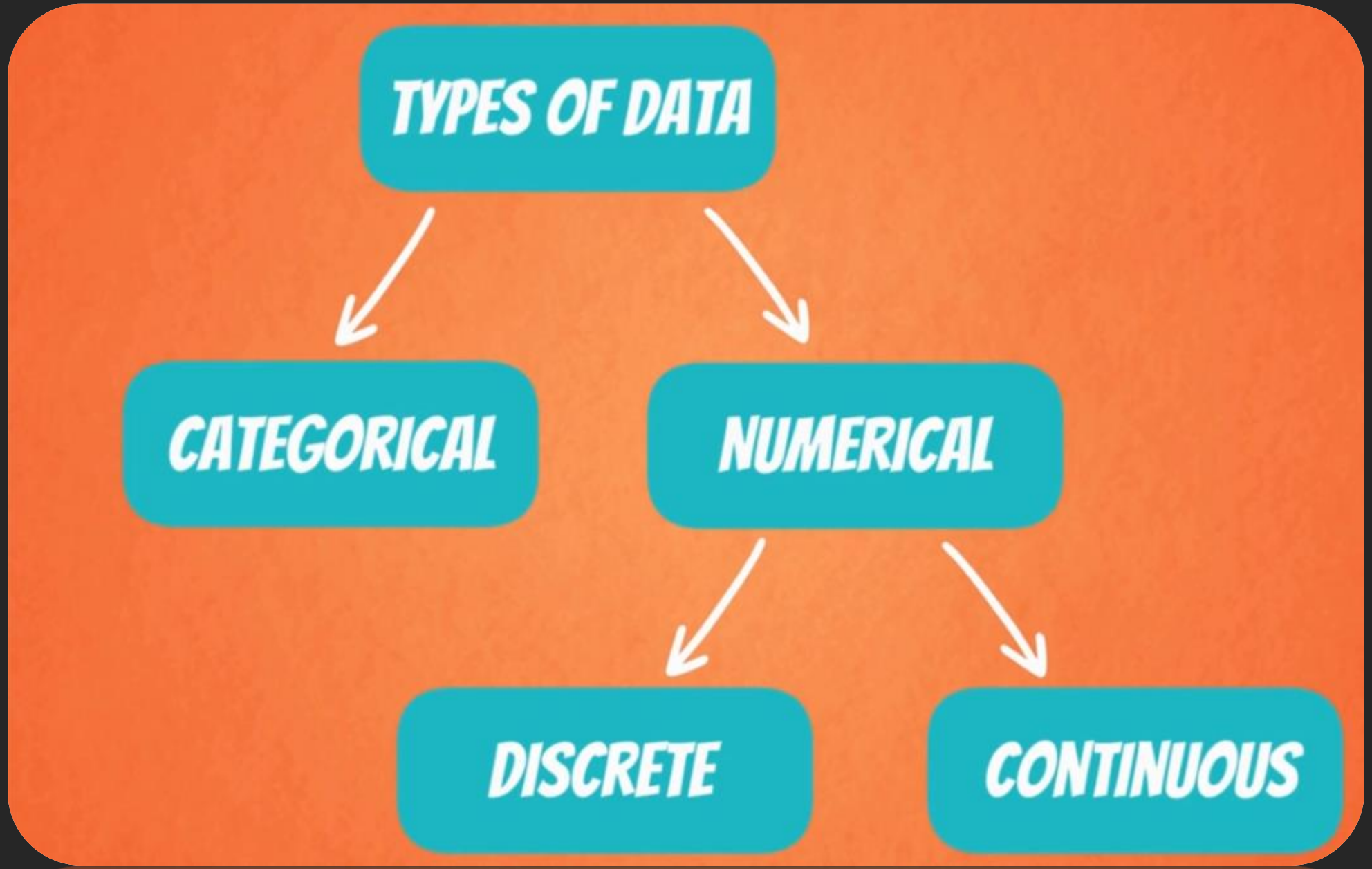

DISTRIBUTION = PROBABILITY DISTRIBUTION
NORMAL
BINOMIAL
UNIFORM

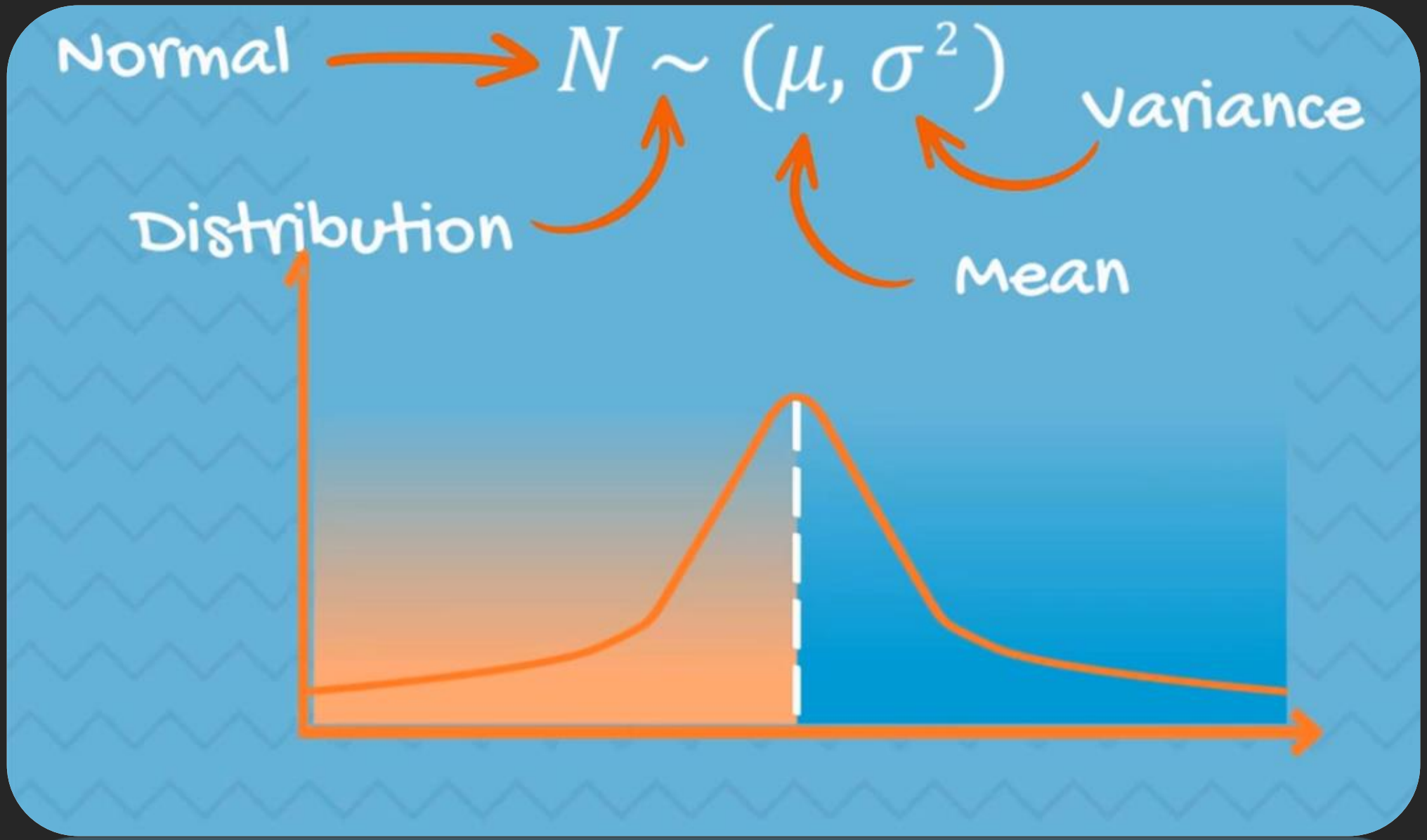
DEFINITION

A distribution is a function that shows the possible values for a variable and how often they occur.

PROBABILITY OF GETTING A 7 IS THE HIGHEST




TYPES OF DATA
CATEGORICAL
NUMERICAL
DISCRETE
CONTINUOUS


Normal
$N \sim (\mu, \sigma^2)$
Variance
Distribution
Mean

STANDARDIZATION

of a Normal distribution

$$\sim N(\mu, \sigma^2) \Rightarrow \sim N(0,1)$$

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

When we standardize a Normal distribution, the result is a Standard Normal distribution

Standard normal distribution

Standardization

Original dataset
1
2
2
3
3
3
4
4
5

Mean 3
St. dev 1.22

N~(3,1.49)

Subtract mean
-2
-1
-1
0
0
0
1
1
2

Mean 0
St. dev 1.22

N~(0,1.49)

Divide by std
-1.63
-0.82
-0.82
0.00
0.00
0.00
0.82
0.82
1.63

Mean 0.00
St. dev 1.00

N~(0,1)

$$x$$

$$x - \mu$$

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

CLT

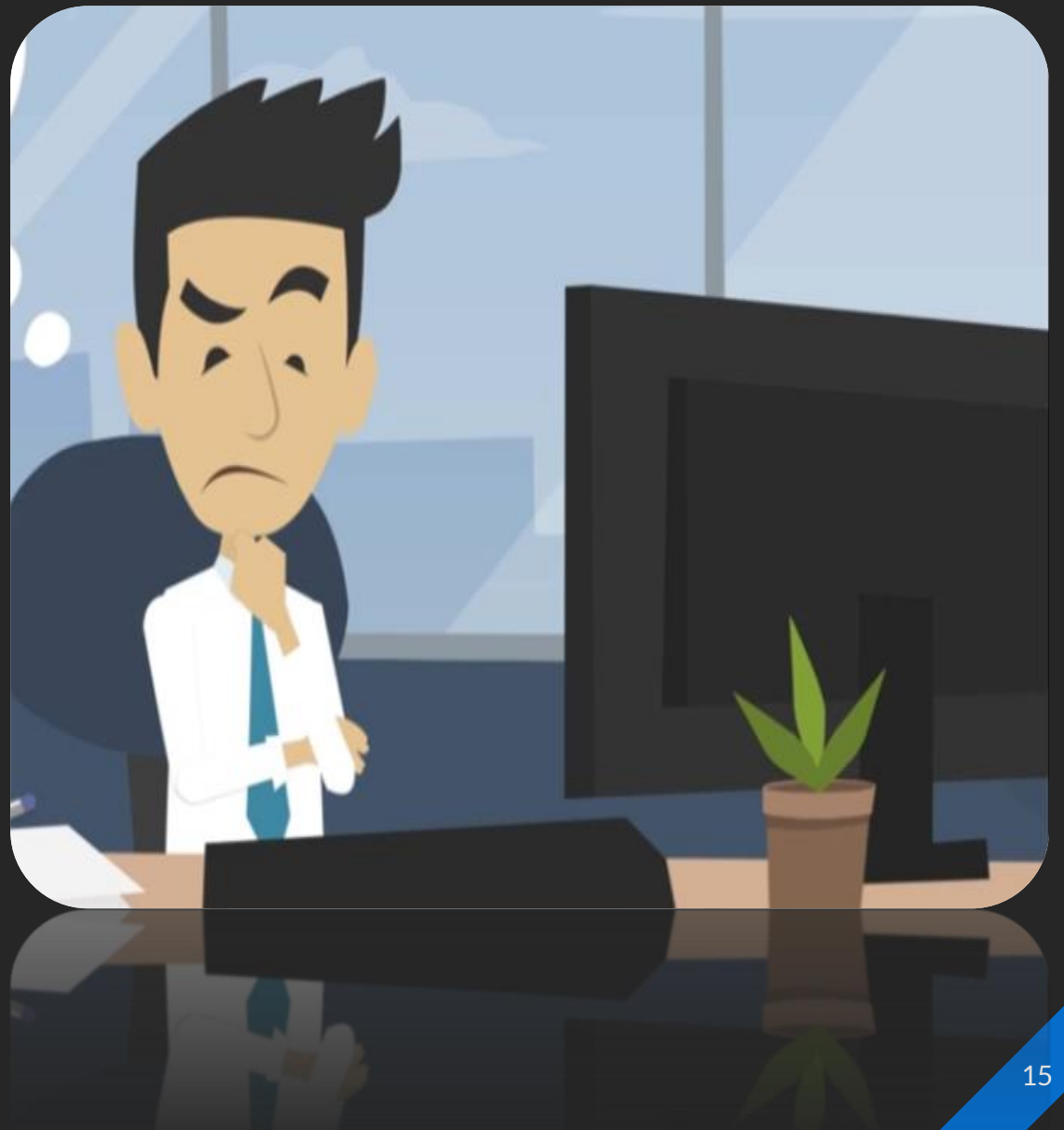
or

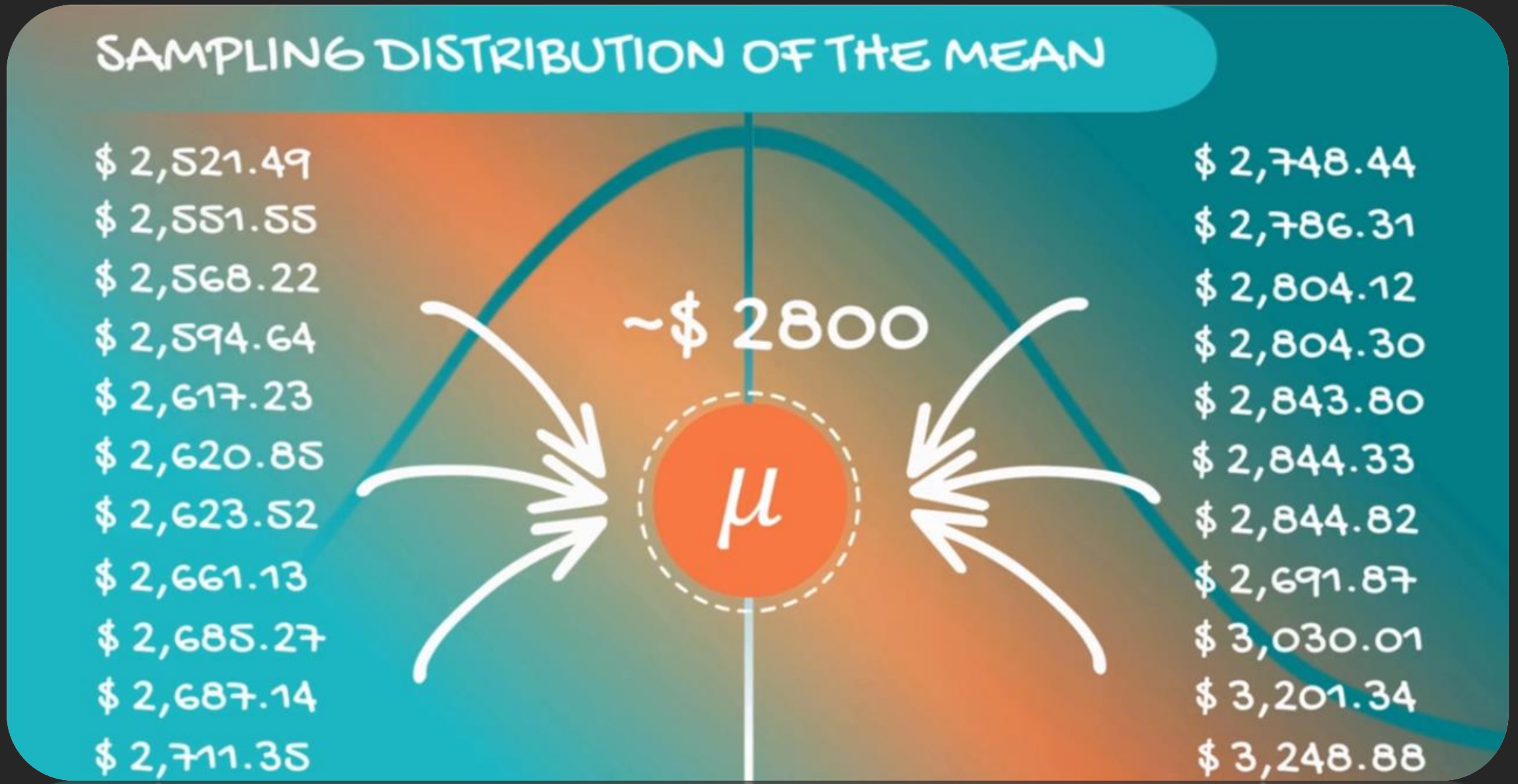
Central Limit Theorem

Is one of the building blocks of Statistics

BUT

WHY???




SAMPLING DISTRIBUTION OF THE MEAN
$ 2,521.49
$ 2,551.55
$ 2,568.22
$ 2,594.64
$ 2,617.23
$ 2,620.85
$ 2,623.52
$ 2,661.13
$ 2,685.27
$ 2,687.14
$ 2,711.35
~$ 2800
μ
$ 2,748.44
$ 2,786.31
$ 2,804.12
$ 2,804.30
$ 2,843.80
$ 2,844.33
$ 2,844.82
$ 2,691.87
$ 3,030.01
$ 3,201.34
$ 3,248.88

CENTRAL LIMIT THEOREM

Original distribution

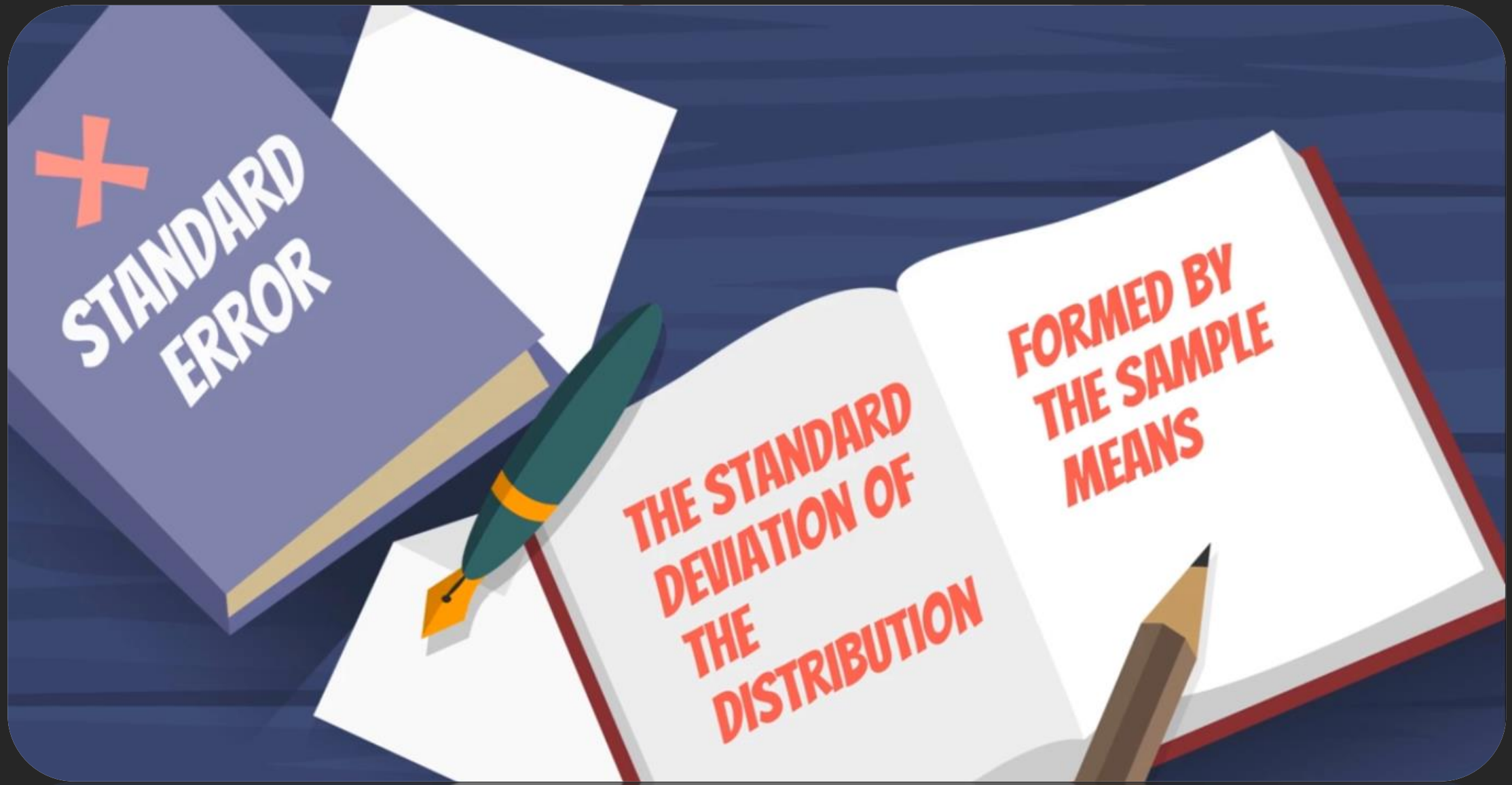
$\mu \quad \sigma^2$

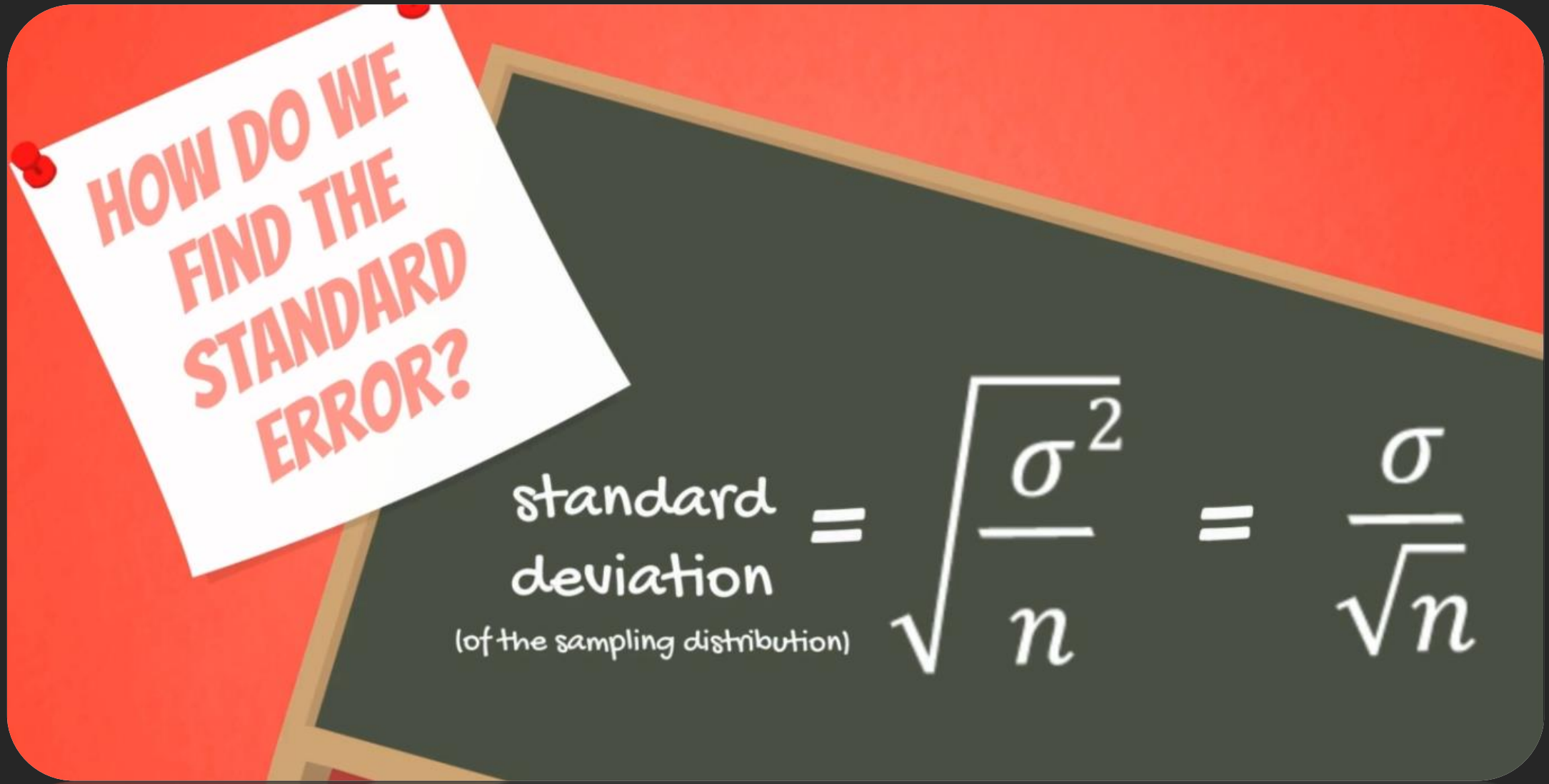
Sampling distribution

$N\left(\mu, \frac{\sigma^2}{n}\right)$

No matter the underlying distribution,
the sampling distribution approximates a Normal

Sampling distribution ~ $N\left(\mu, \frac{\sigma^2}{n}\right)$, $n > 30$

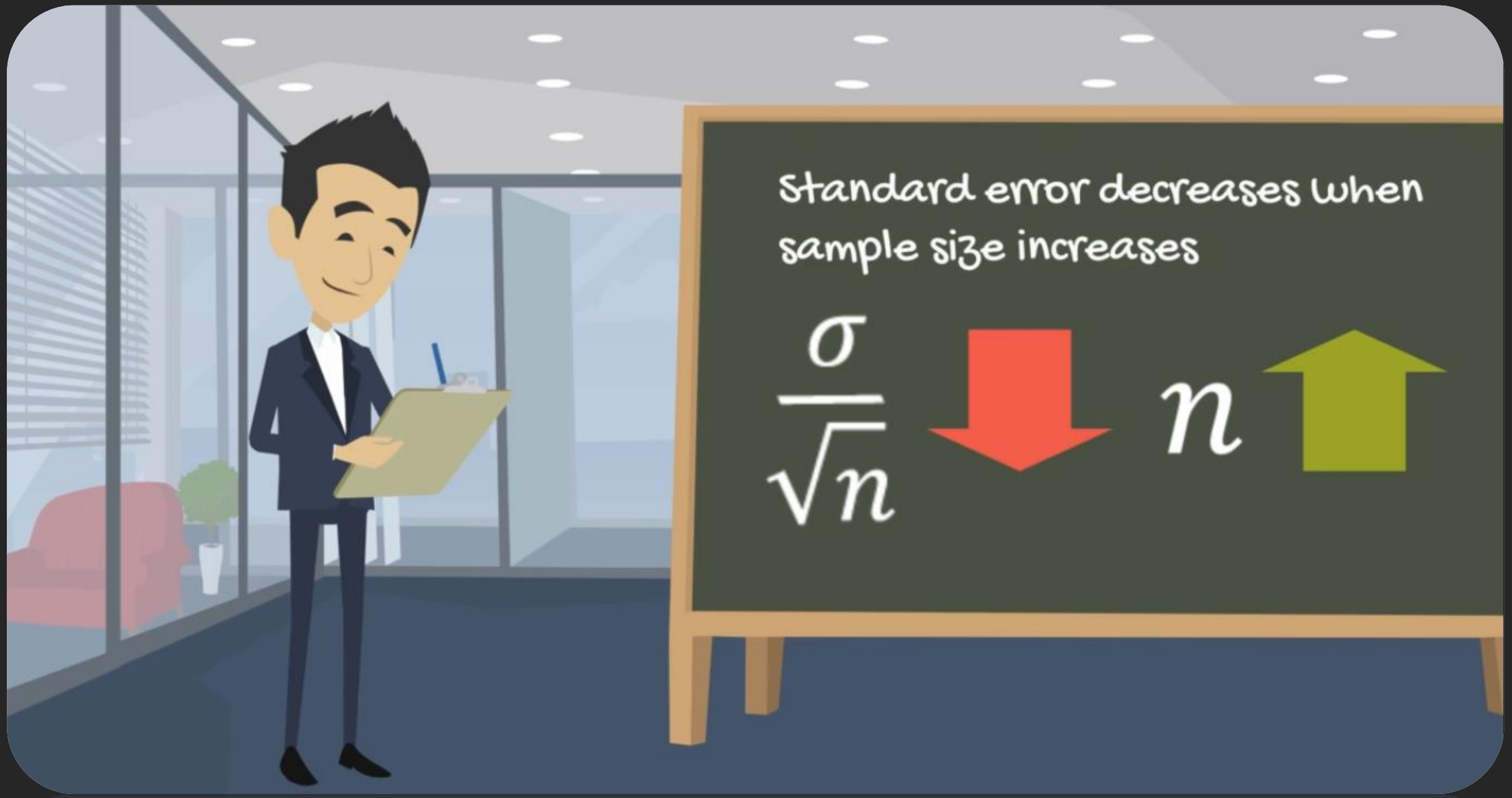

STANDARD ERROR
THE STANDARD DEVIATION OF THE DISTRIBUTION
FORMED BY THE SAMPLE MEANS


HOW DO WE FIND THE STANDARD ERROR?
standard deviation
(of the sampling distribution)
$= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

WHY IS IT IMPORTANT?

Used in most statistical tests

Because it shows how well you approximated the true mean


Standard error decreases when sample size increases
$\frac{\sigma}{\sqrt{n}}$
n


POINT ESTIMATES
CONFIDENCE INTERVALS

Estimators and Estimates

Estimators

Broadly, an estimator is a mathematical function that approximates a population parameter depending only on sample information.

Examples of estimators and the corresponding parameters:

Term	Estimator	Parameter
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Correlation	r	ρ

Estimators have two important properties:

- Bias

The expected value of an unbiased estimator is the population parameter. The bias in this case is 0. If the expected value of an estimator is (parameter + b), then the bias is b.

- Efficiency

The most efficient estimator is the one with the smallest variance.

Estimates

An estimate is the output that you get from the estimator (when you apply the formula). There are two types of estimates: point estimates and confidence interval estimates.

Point estimates

A single value.

Examples:

- 1
- 5
- 122.67
- 0.32

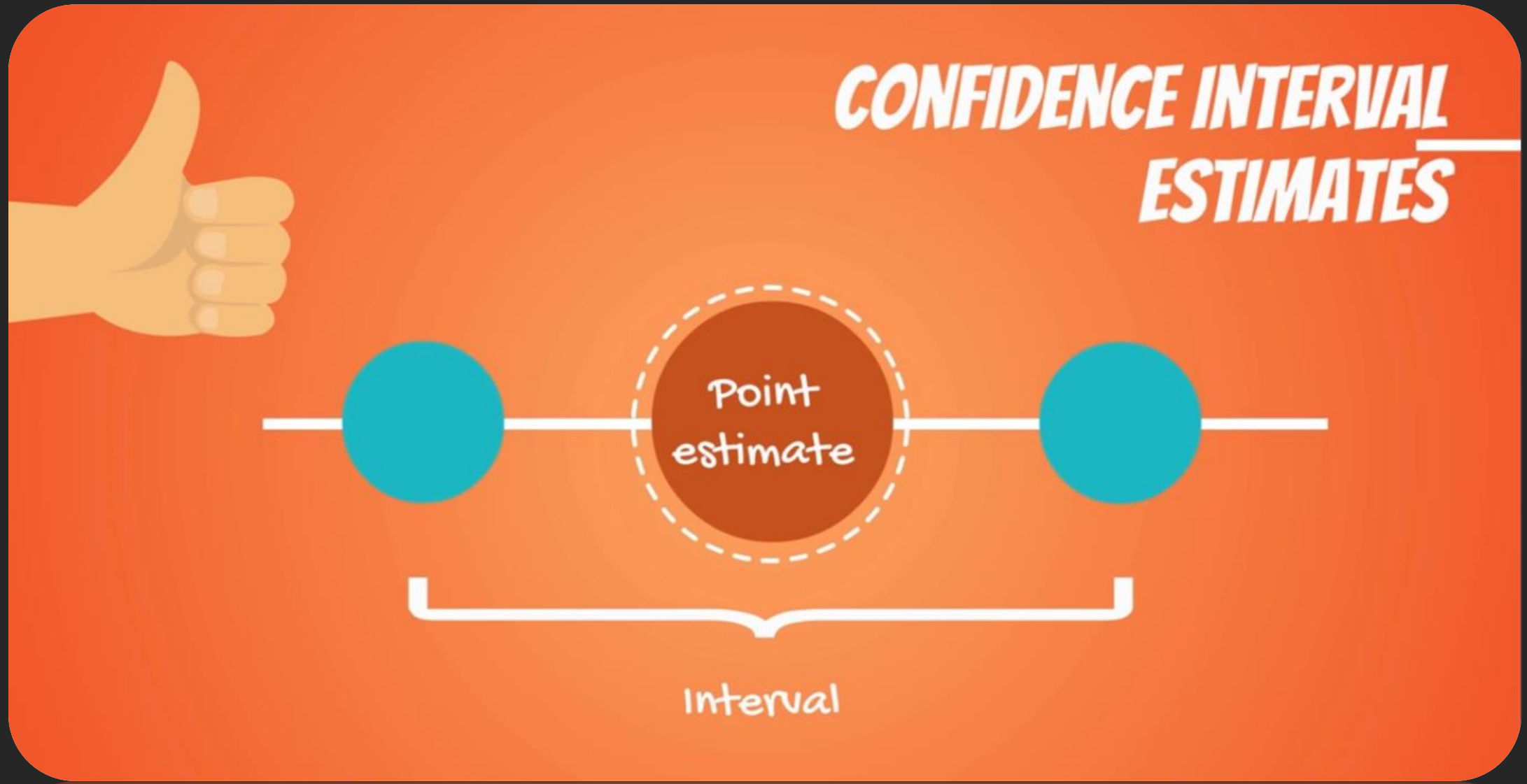
Confidence intervals

An interval.

Examples:

- (1 , 5)
- (12 , 33)
- (221.78 , 745.66)
- (- 0.71 , 0.11)

Confidence intervals are much more precise than point estimates. That is why they are preferred when making inferences.


CONFIDENCE INTERVAL ESTIMATES
Point estimate
Interval

POINT ESTIMATORS AND ESTIMATES

Estimator /how to estimate/		Parameter /what to estimate/	Estimate /concrete result/
$\bar{x}$	of	μ	52.22
s^2	of	σ^2	1724.93

Estimators have two important properties:

- Bias

The expected value of an unbiased estimator is the population parameter. The bias in this case is 0. If the expected value of an estimator is (parameter + b), then the bias is b.

- Efficiency

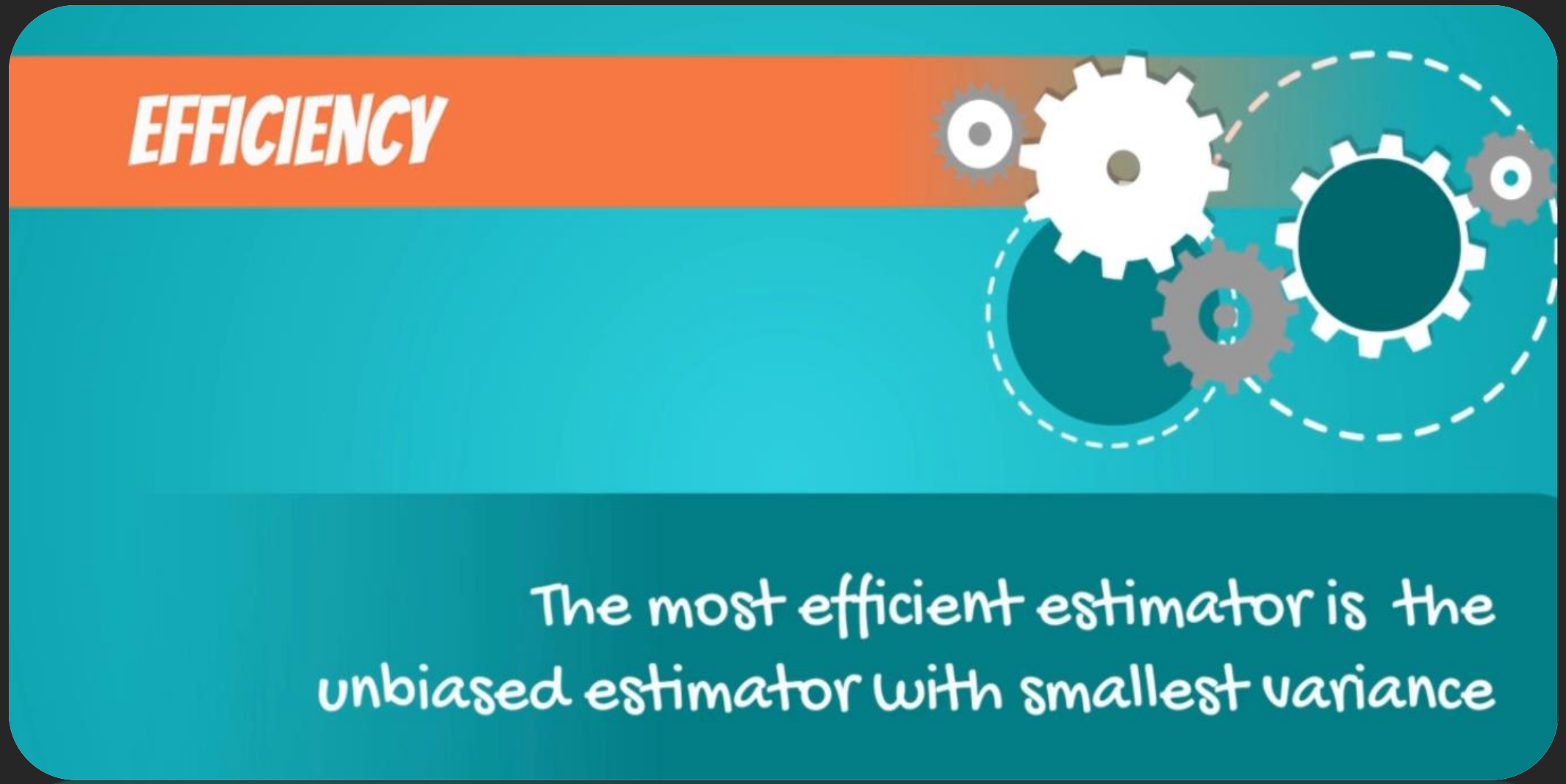
The most efficient estimator is the one with the smallest variance.

UNBIASED ESTIMATOR

expected value = population parameter

e.g. $\bar{x}$ has an expected value of μ


BIAS
$\bar{x}$ estimates μ with NO BIAS
$\bar{x}$ +1FT estimates μ with a bias of +1FT


EFFICIENCY
The most efficient estimator is the
unbiased estimator with smallest variance

Why we should read English?
Is it possible to get a ahead start in research only by reading in Persian or just in our mother tongue?

Thanks for watching AMIGOS