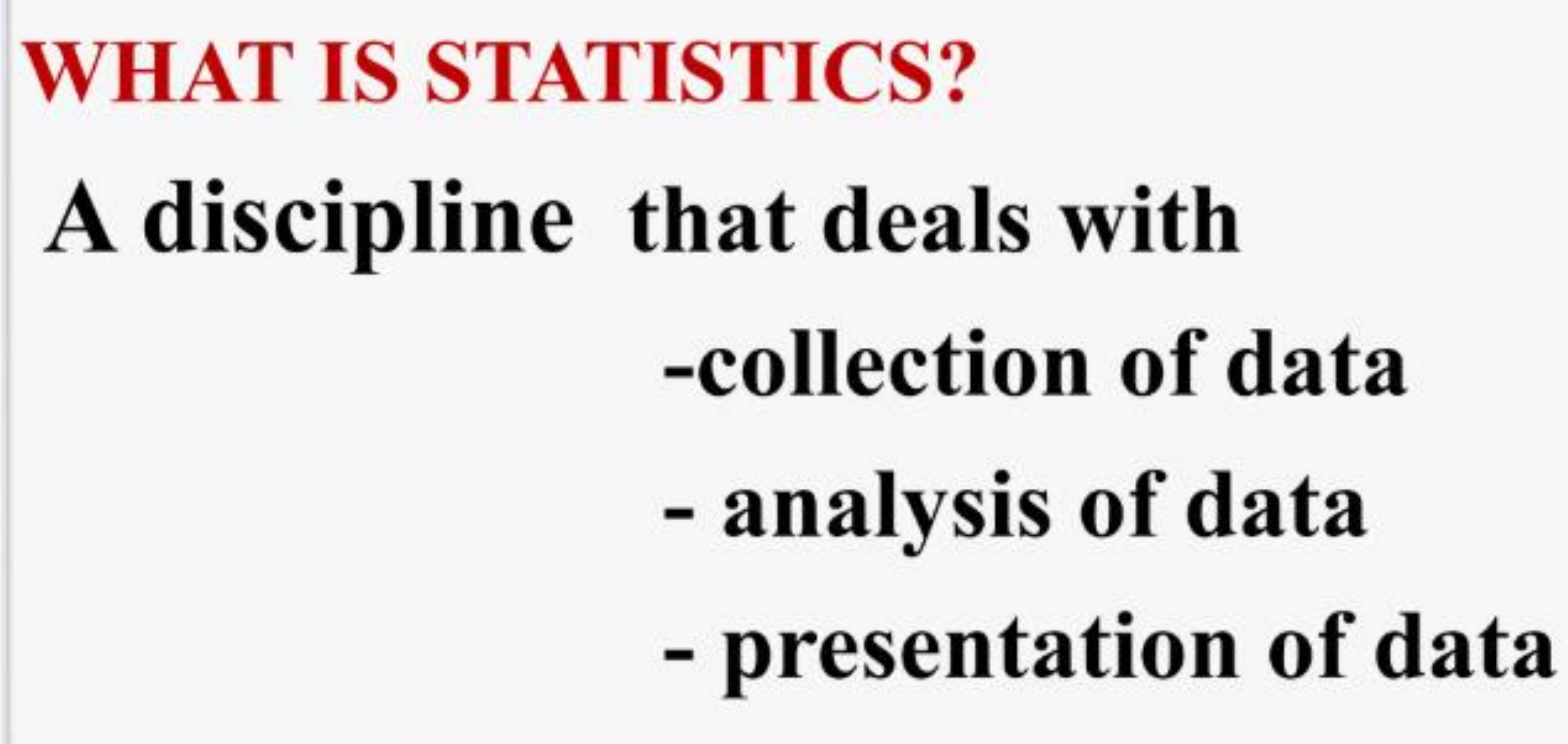
Inferential Statistics

ANISHA JOSEPH



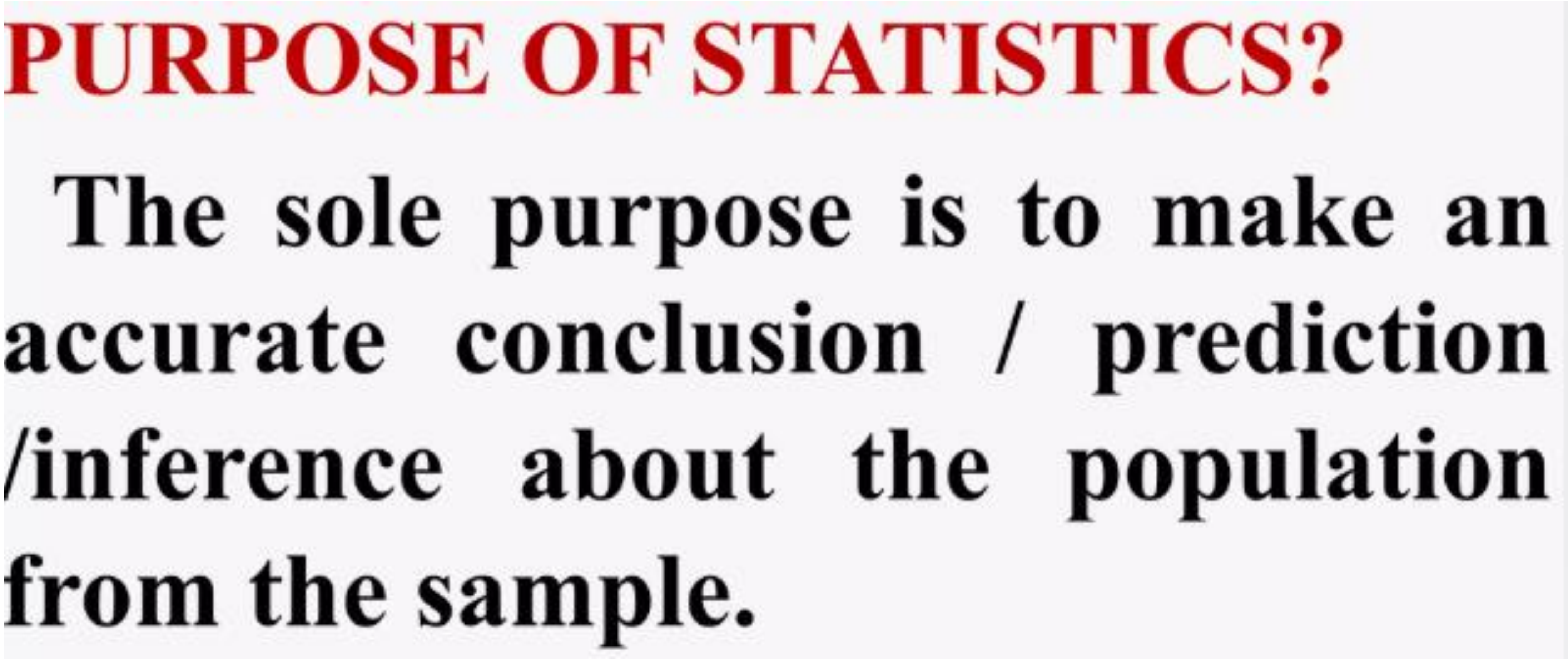
WHAT IS STATISTICS?

A discipline that deals with

-collection of data

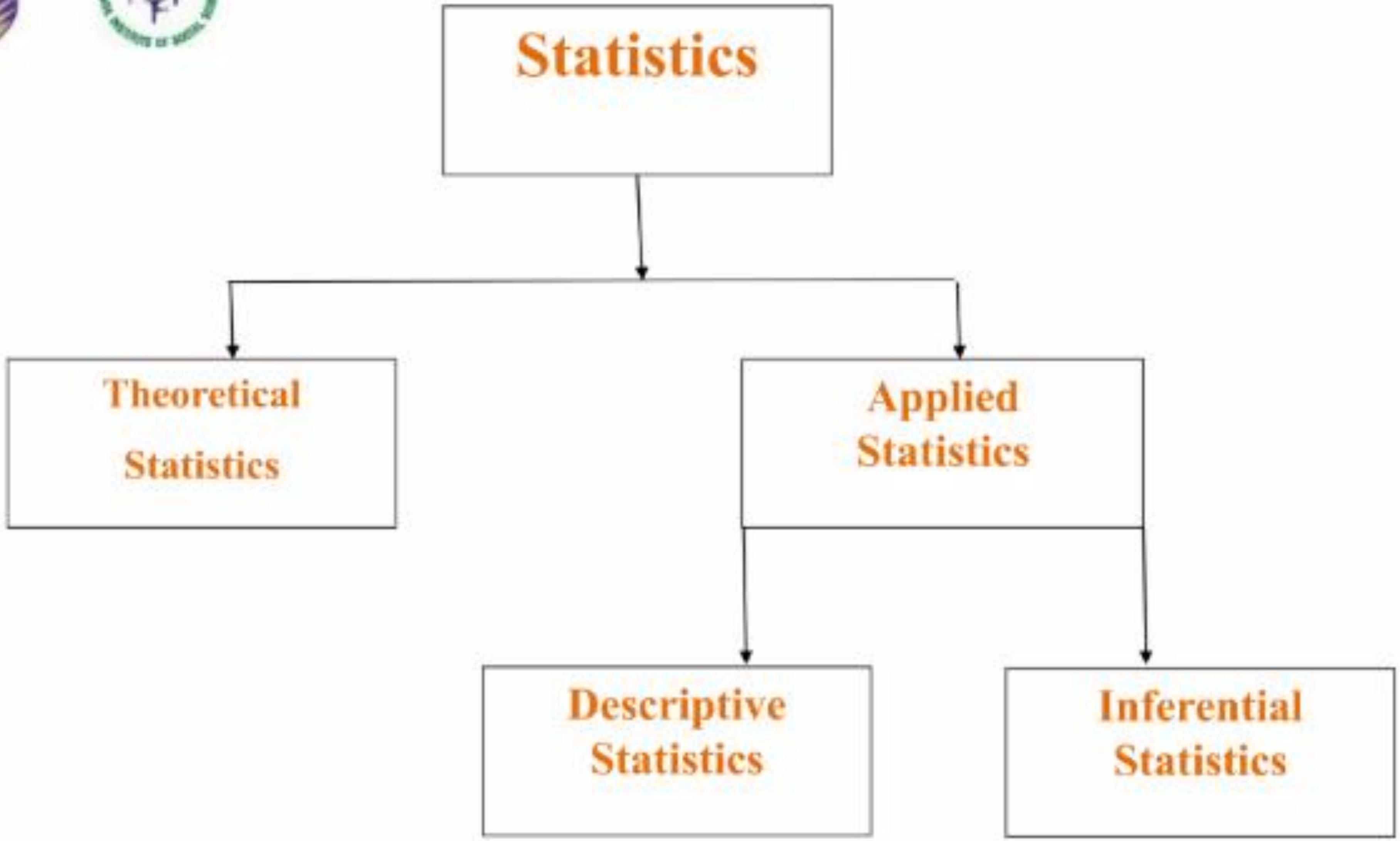
- analysis of data

- presentation of data



PURPOSE OF STATISTICS?

The sole purpose is to make an accurate conclusion / prediction /inference about the population from the sample.



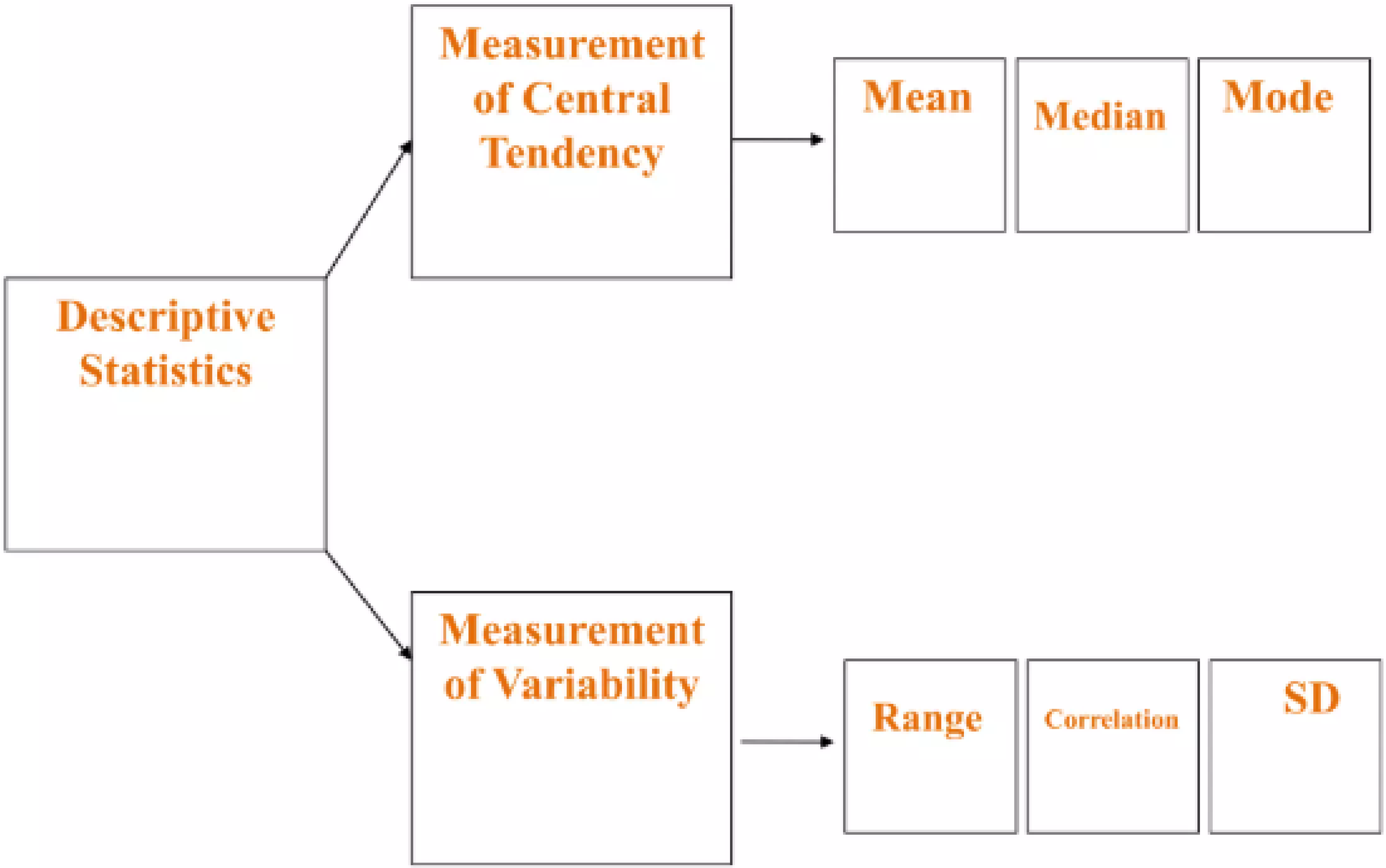
Statistics

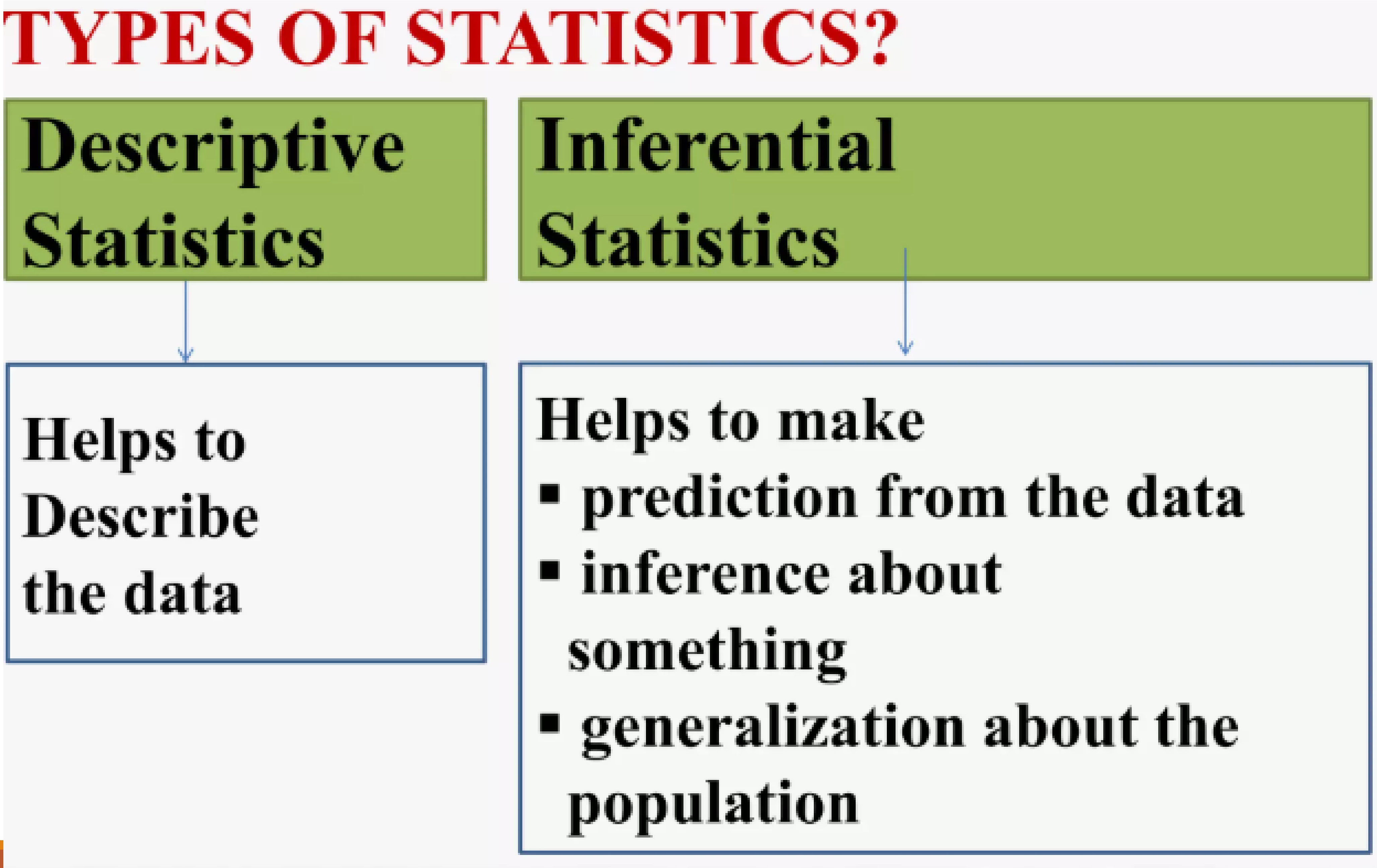
Theoretical Statistics

Applied Statistics

Descriptive Statistics

Inferential Statistics


Descriptive Statistics
Measurement of Central Tendency
Mean
Median
Mode
Measurement of Variability
Range
Correlation
SD



TYPES OF STATISTICS?

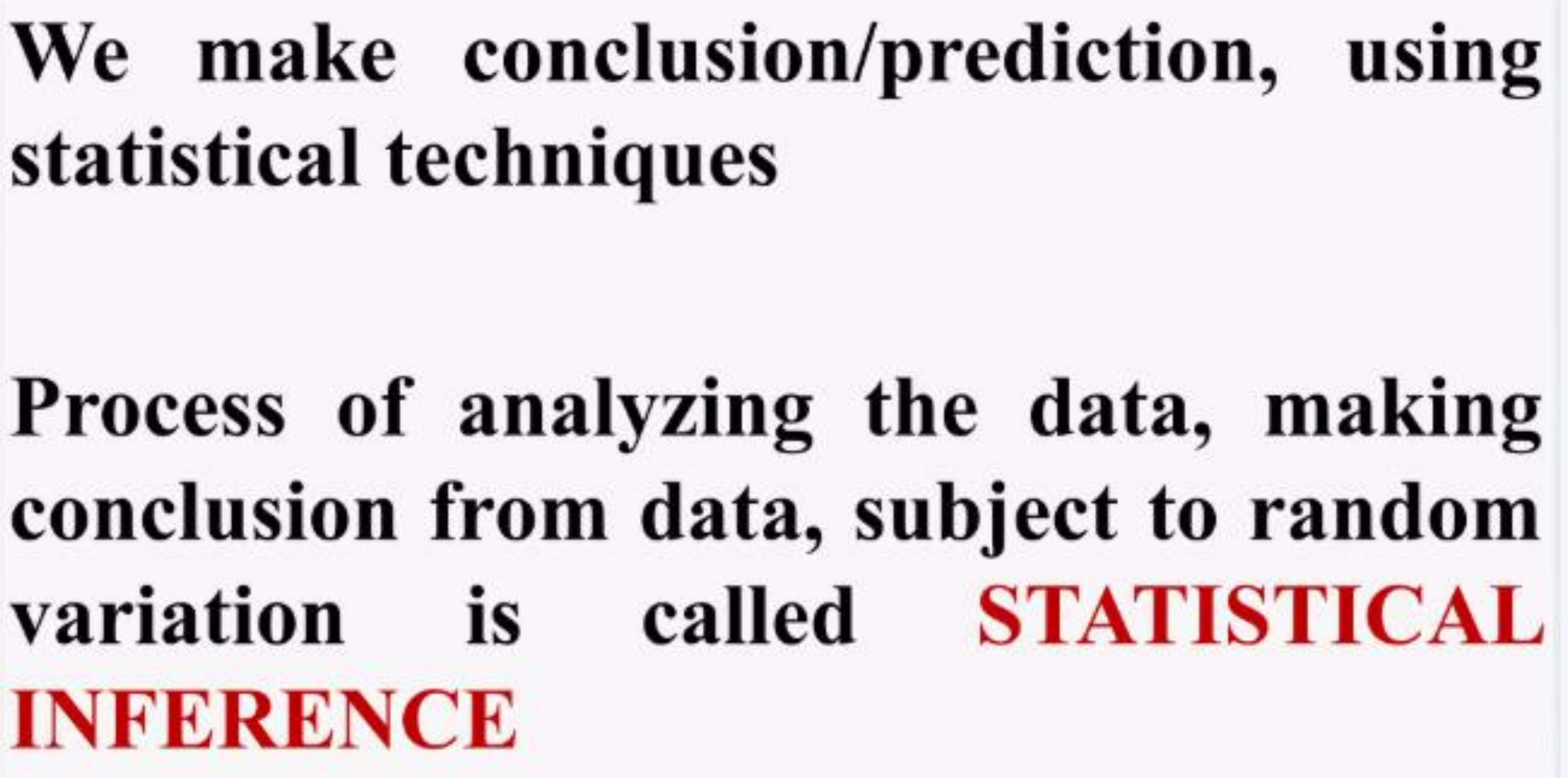
Descriptive Statistics

Helps to Describe the data

Inferential Statistics

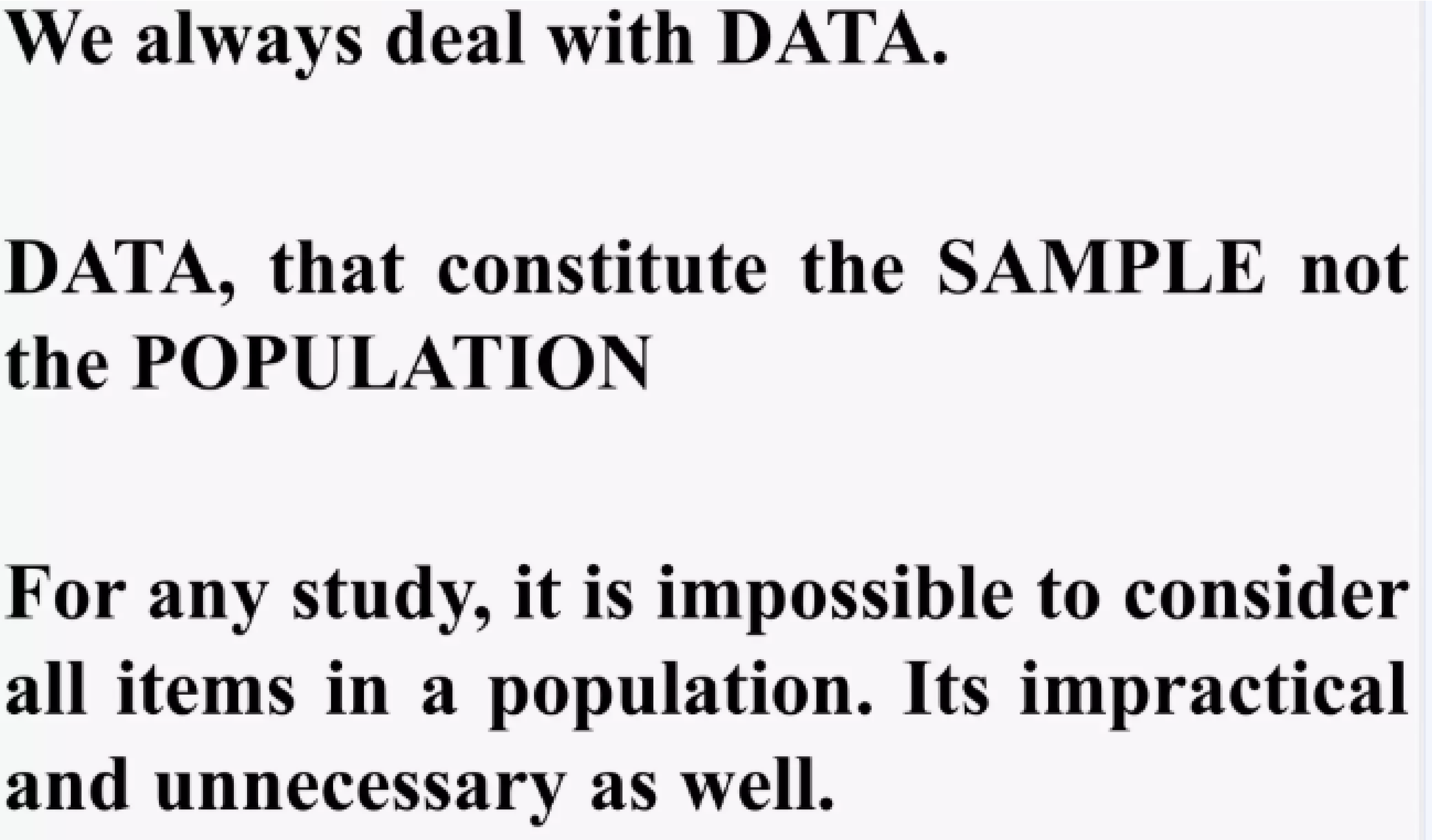
Helps to make

- prediction from the data
- inference about something
- generalization about the population



We make conclusion/prediction, using statistical techniques

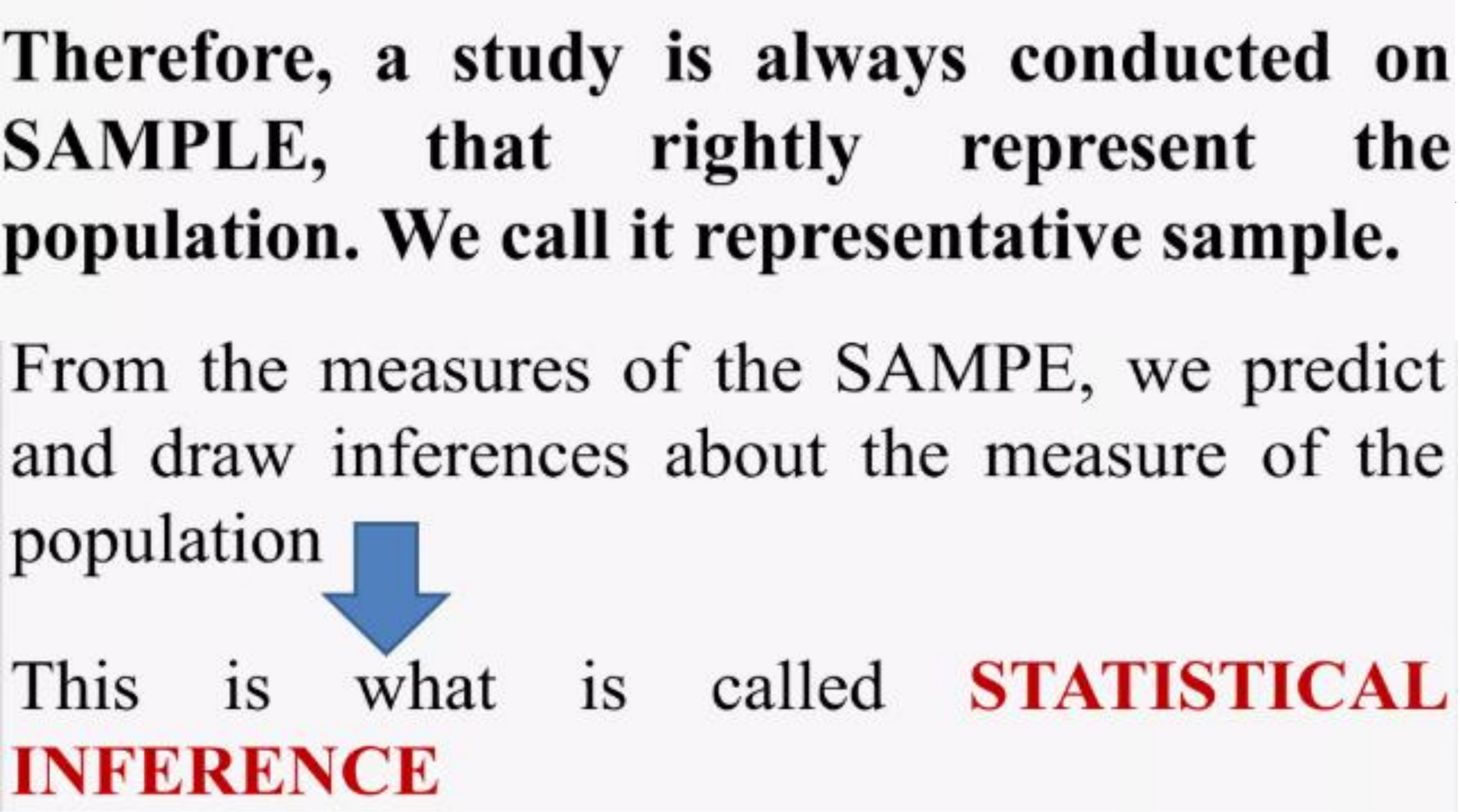
Process of analyzing the data, making conclusion from data, subject to random variation is called STATISTICAL INFERENCE



We always deal with DATA.

DATA, that constitute the SAMPLE not the POPULATION

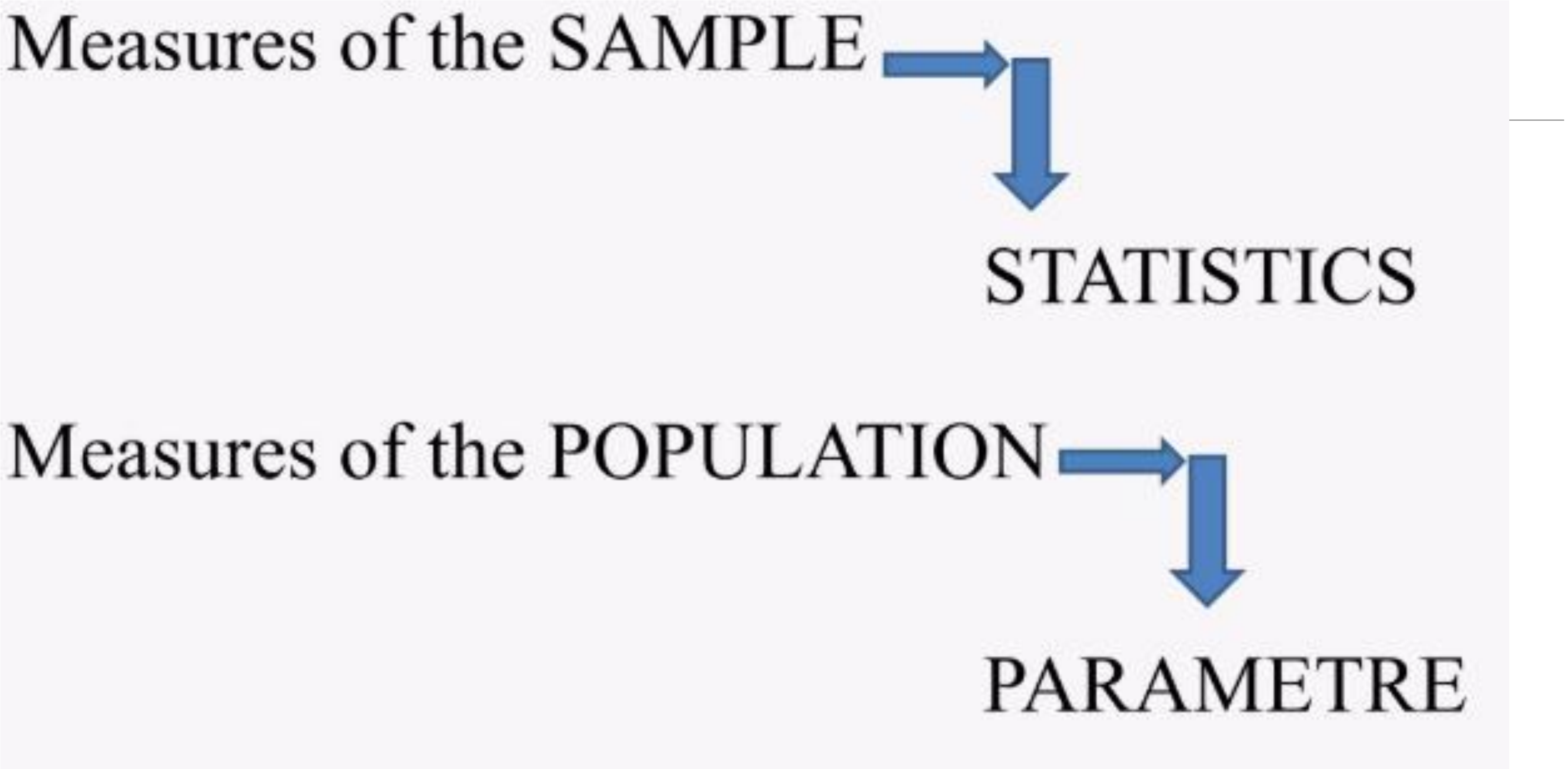
For any study, it is impossible to consider all items in a population. Its impractical and unnecessary as well.



Therefore, a study is always conducted on SAMPLE, that rightly represent the population. We call it representative sample.

From the measures of the SAMPE, we predict and draw inferences about the measure of the population

This is what is called **STATISTICAL INFERENCE**

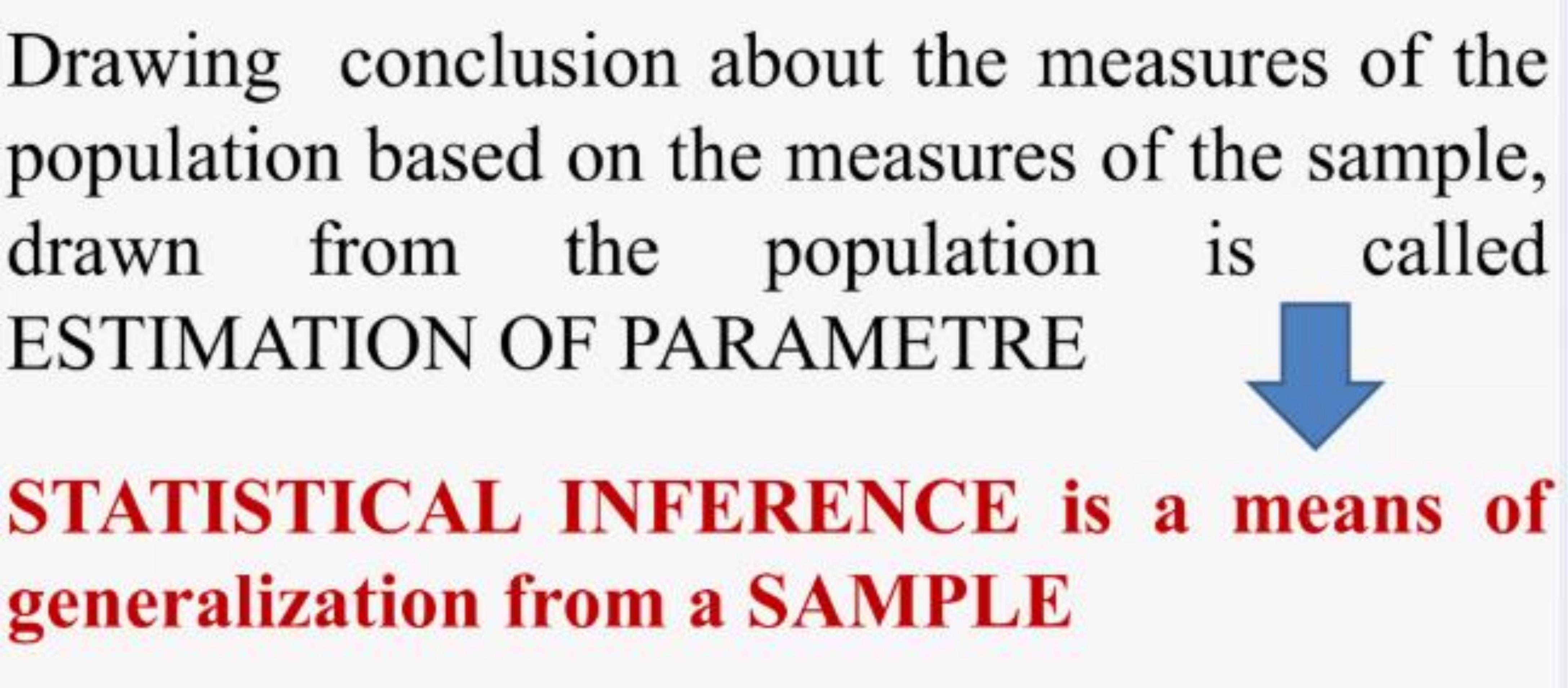


Measures of the SAMPLE

STATISTICS

Measures of the POPULATION

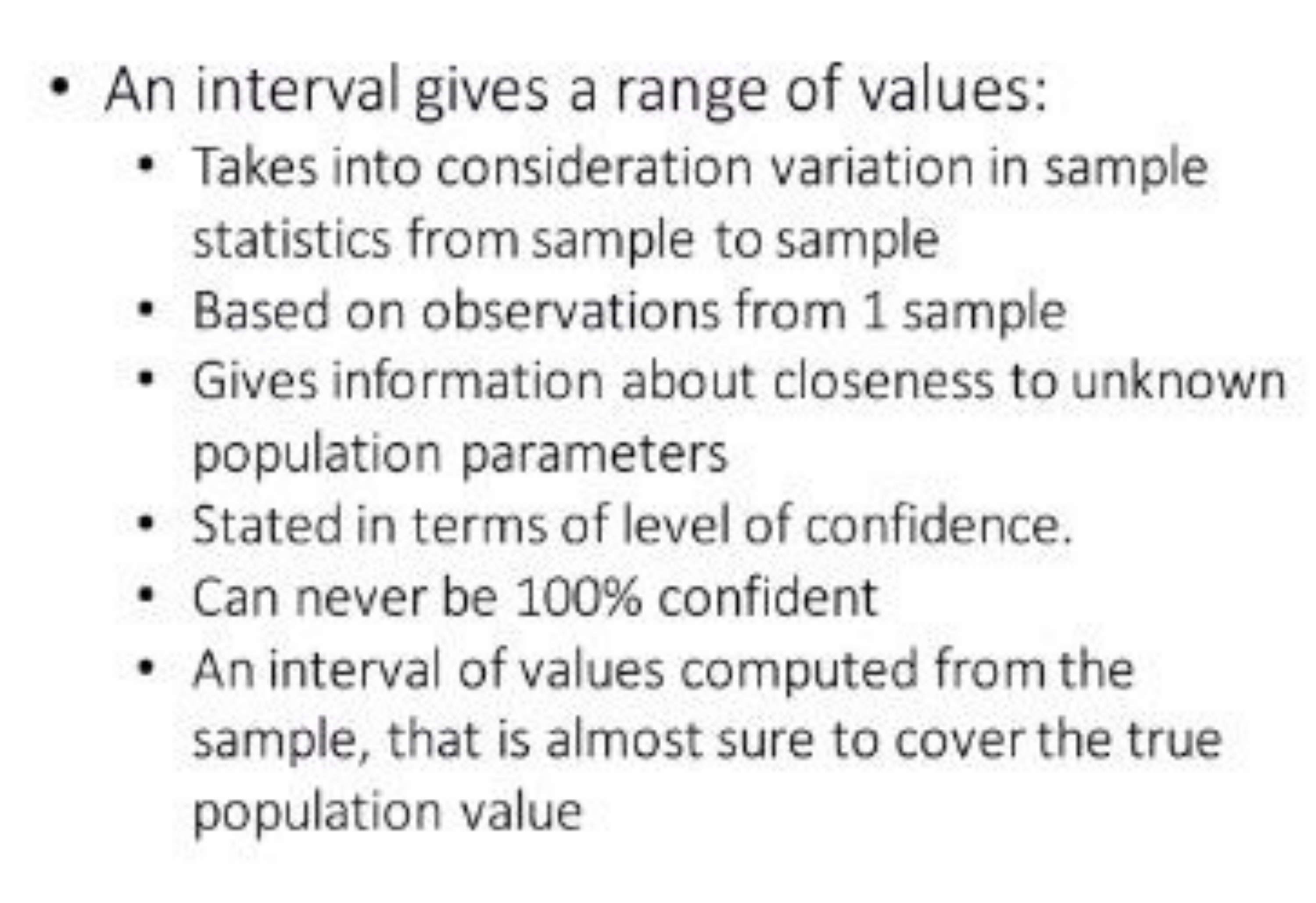
PARAMETRE



Drawing conclusion about the measures of the population based on the measures of the sample, drawn from the population is called ESTIMATION OF PARAMETRE

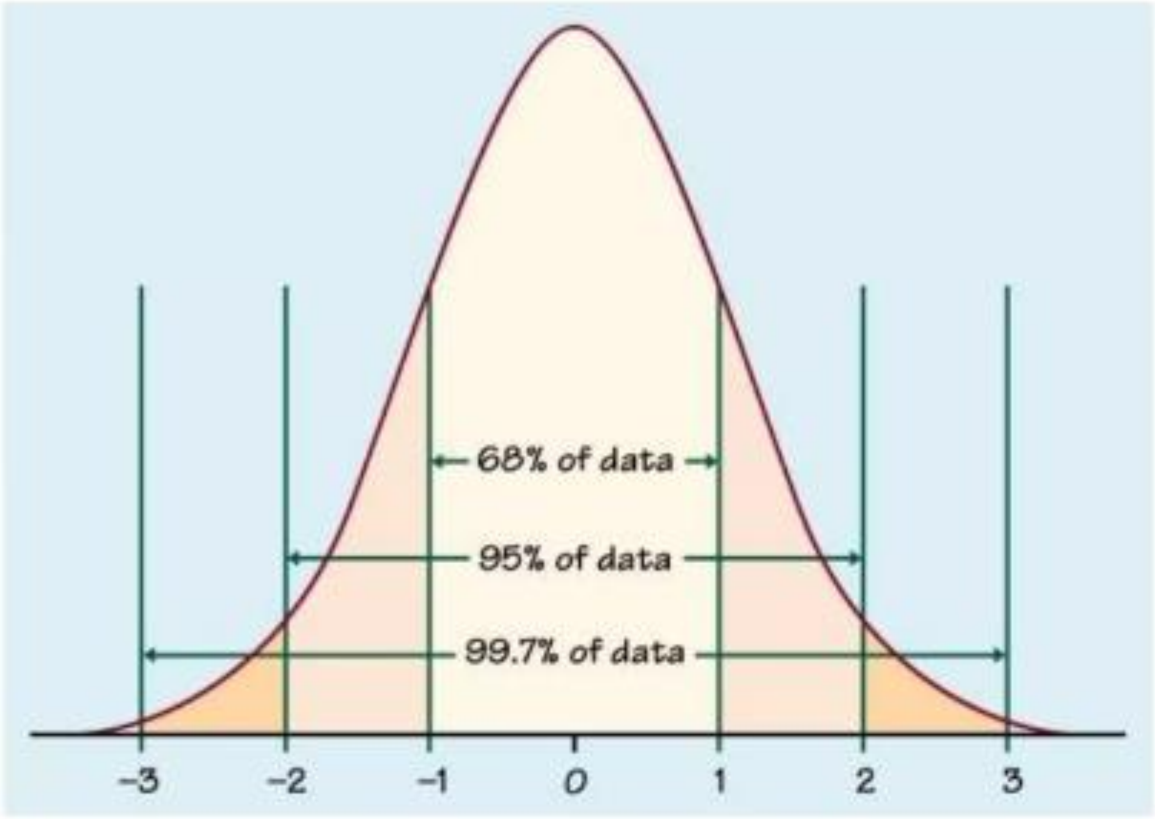
STATISTICAL INFERENCE is a means of generalization from a SAMPLE

Confidence Interval

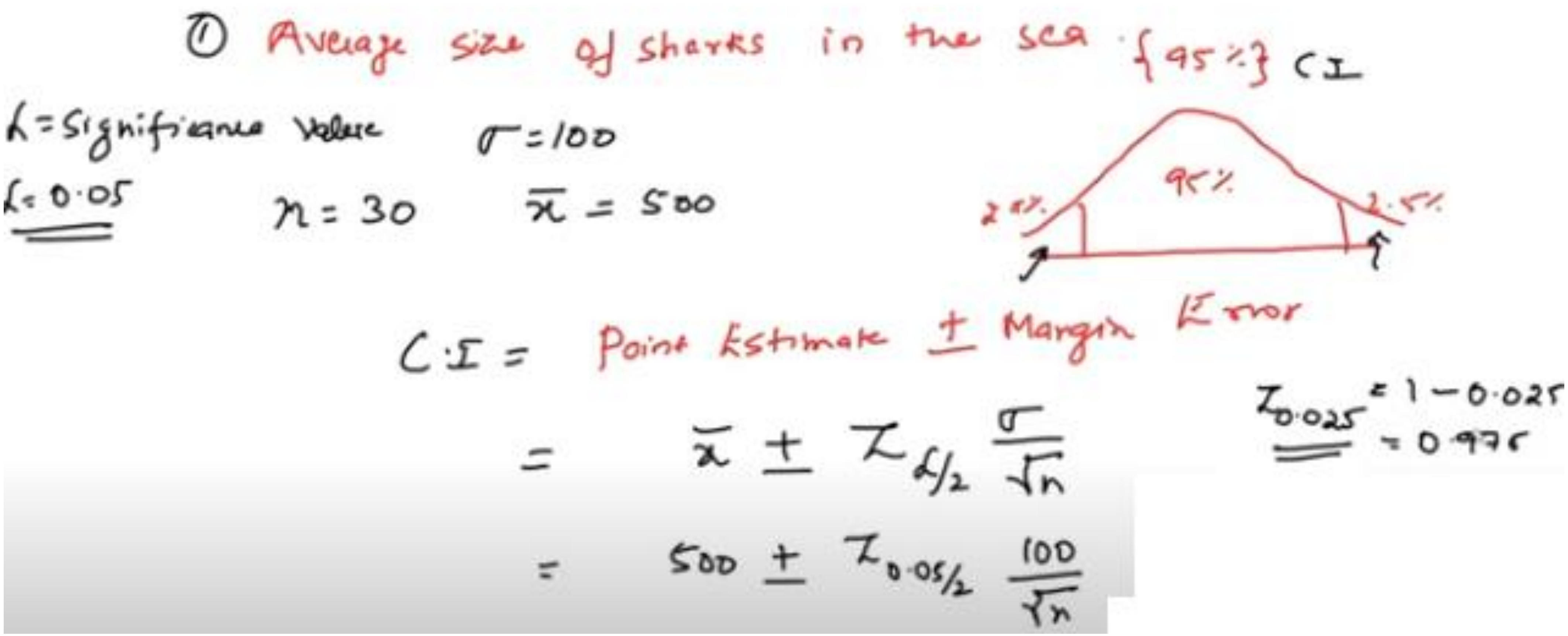


- An interval gives a range of values:
  - Takes into consideration variation in sample statistics from sample to sample
  - Based on observations from 1 sample
  - Gives information about closeness to unknown population parameters
  - Stated in terms of level of confidence.
  - Can never be 100% confident
  - An interval of values computed from the sample, that is almost sure to cover the true population value

NORMAL DISTRIBUTION CURVE



Confidence Interval -Example



① Average size of sharks in the sea. {95%} CI

α = Significance value $\quad \sigma = 100$

$\alpha = 0.05 \quad n = 30 \quad \bar{x} = 500$

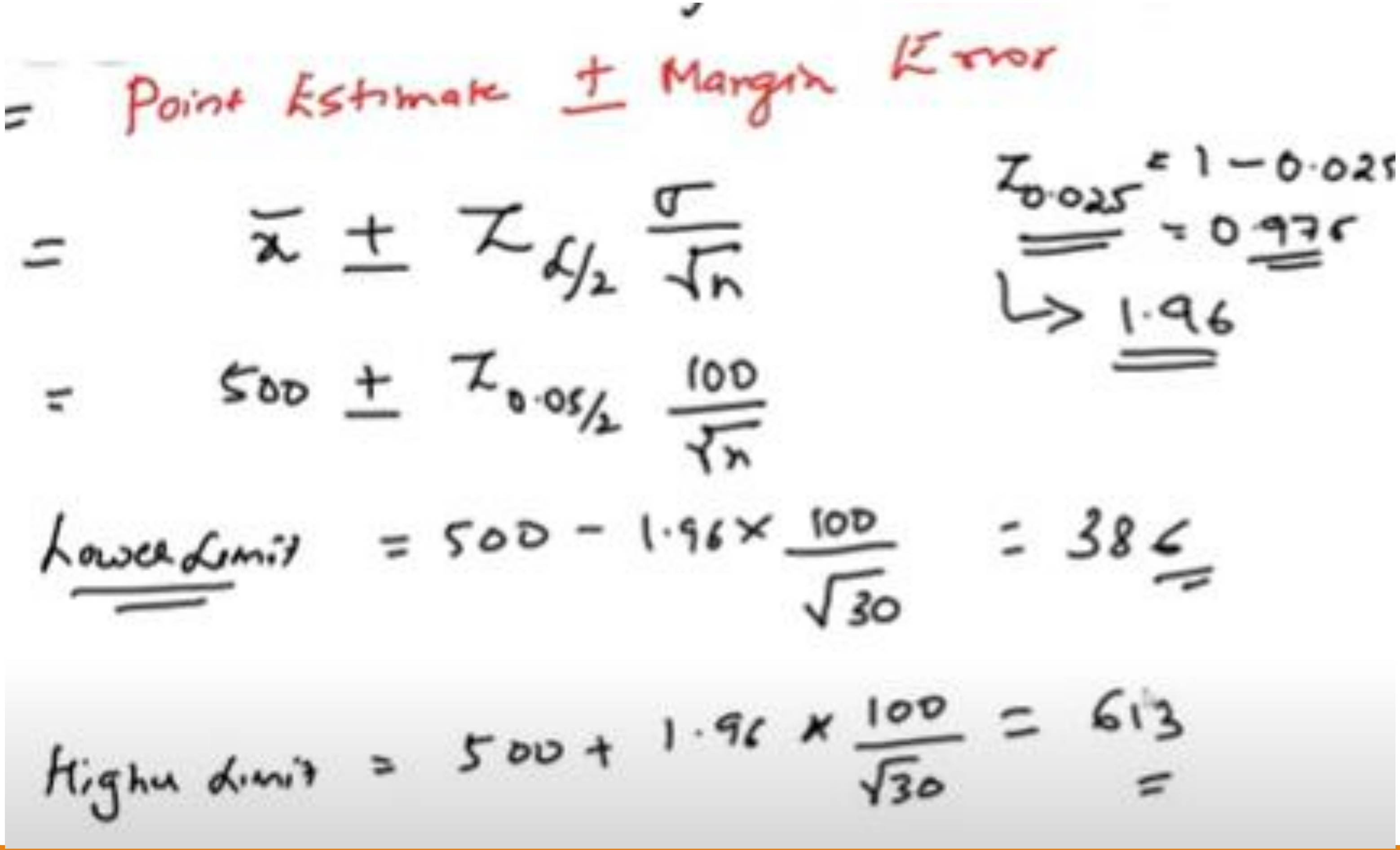
C.I = Point Estimate ± Margin Error

$$= \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$= 500 \pm Z_{0.05/2} \frac{100}{\sqrt{n}}$$

$$Z_{0.025} = 1 - 0.025 = 0.975$$

Confidence Interval -Example



= Point Estimate ± Margin Error

$$= \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$= 500 \pm Z_{0.05/2} \frac{100}{\sqrt{n}}$$

$$\underline{\underline{Z_{0.025}}} = 1 - 0.025$$

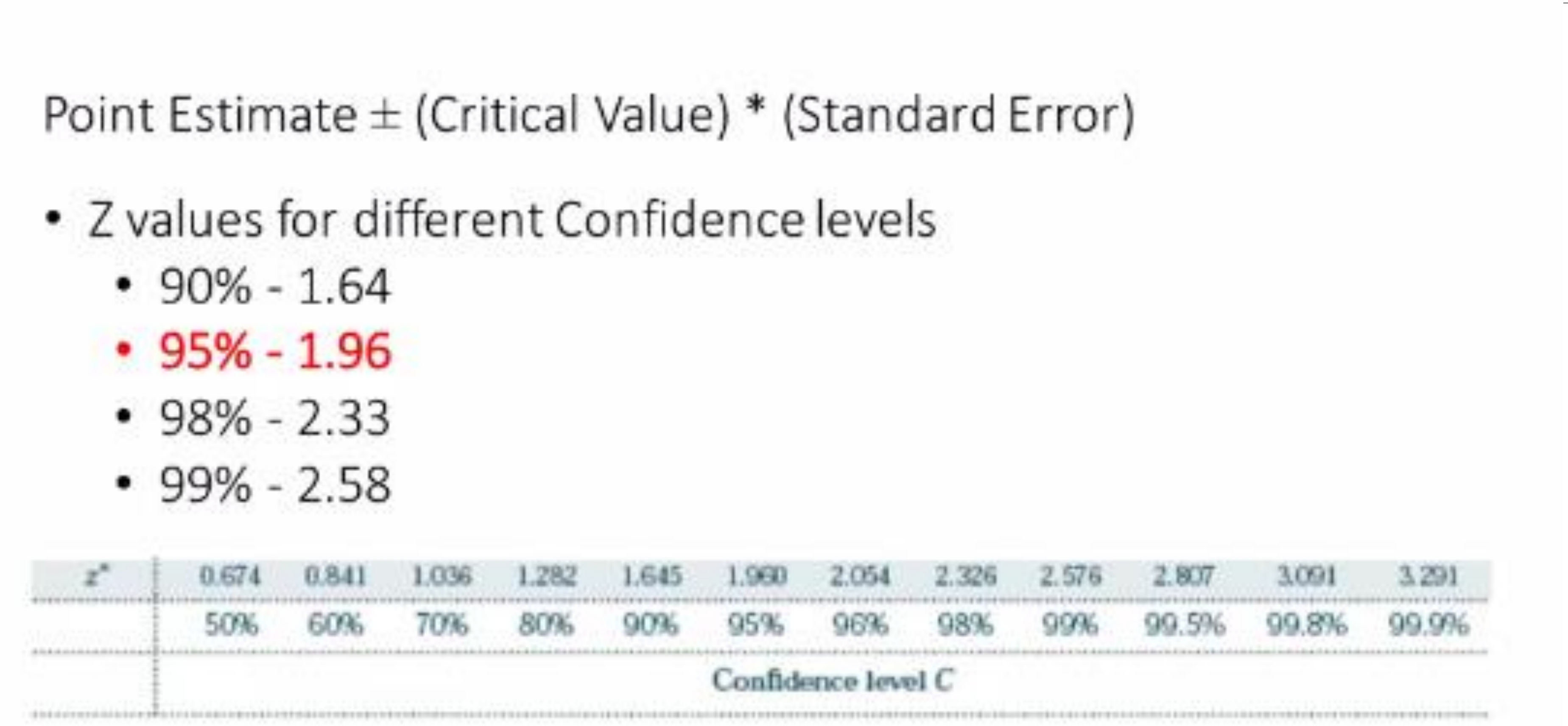
$$= 0.\underline{\underline{975}}$$

$$\hookrightarrow \underline{\underline{1.96}}$$

$\underline{\underline{\text{Lower Limit}}} = 500 - 1.96 \times \frac{100}{\sqrt{30}} = \underline{\underline{385}}$

$\text{Higher Limit} = 500 + 1.96 \times \frac{100}{\sqrt{30}} = \underline{\underline{613}}$

Confidence Interval



Point Estimate ± (Critical Value) * (Standard Error)

- Z values for different Confidence levels
    - 90% - 1.64
    - 95% - 1.96
    - 98% - 2.33
    - 99% - 2.58

z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

- Descriptive Statistics

Mean:

- ❑ The total of all the values divided by the size of the data set.
- ❑ It is the most commonly used statistic of position.
- ❑ It is easy to understand and calculate.
- ❑ It works well when the distribution is symmetric and there are no outliers.
- ❑ The mean of a sample is denoted by ‘**x-bar**’.
- ❑ The mean of a population is denoted by ‘**μ**’.

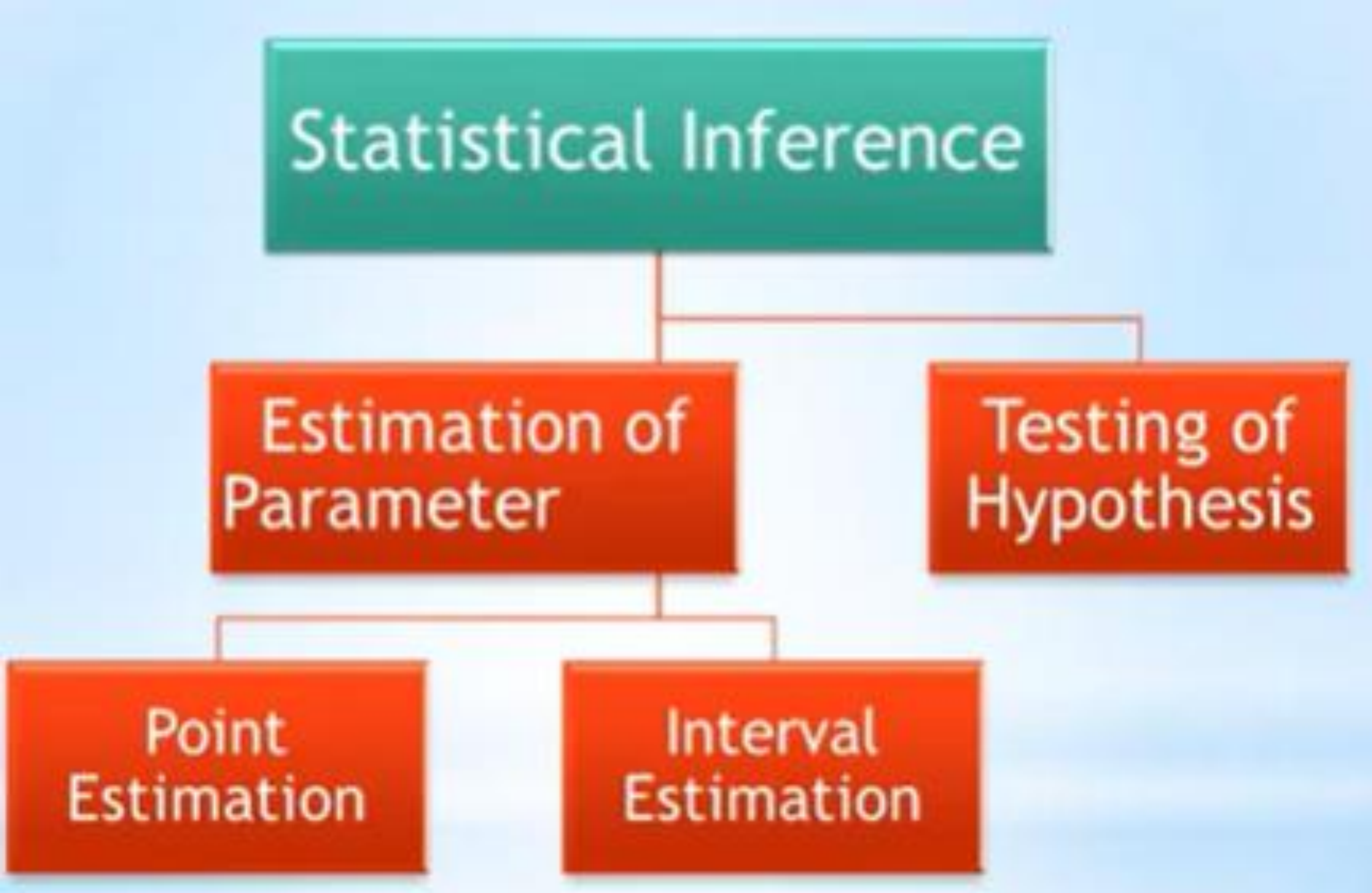
$$\text{Mean} = \bar{x} = \frac{\sum_{i}^{n} x_i}{n}$$

Standard Deviation and variance

- The best-known estimates for variability are the variance and the standard deviation, which are based on squared deviations.
- The **variance** is an average of the squared deviations, and the **standard deviation** is the square root of the variance.

$$\text{Variance} \;=\; s^2 = \frac{\sum (x-\overline{x})^2}{n-1}$$

$$\text{Standard deviation} \;=\; s = \sqrt{\text{Variance}}$$



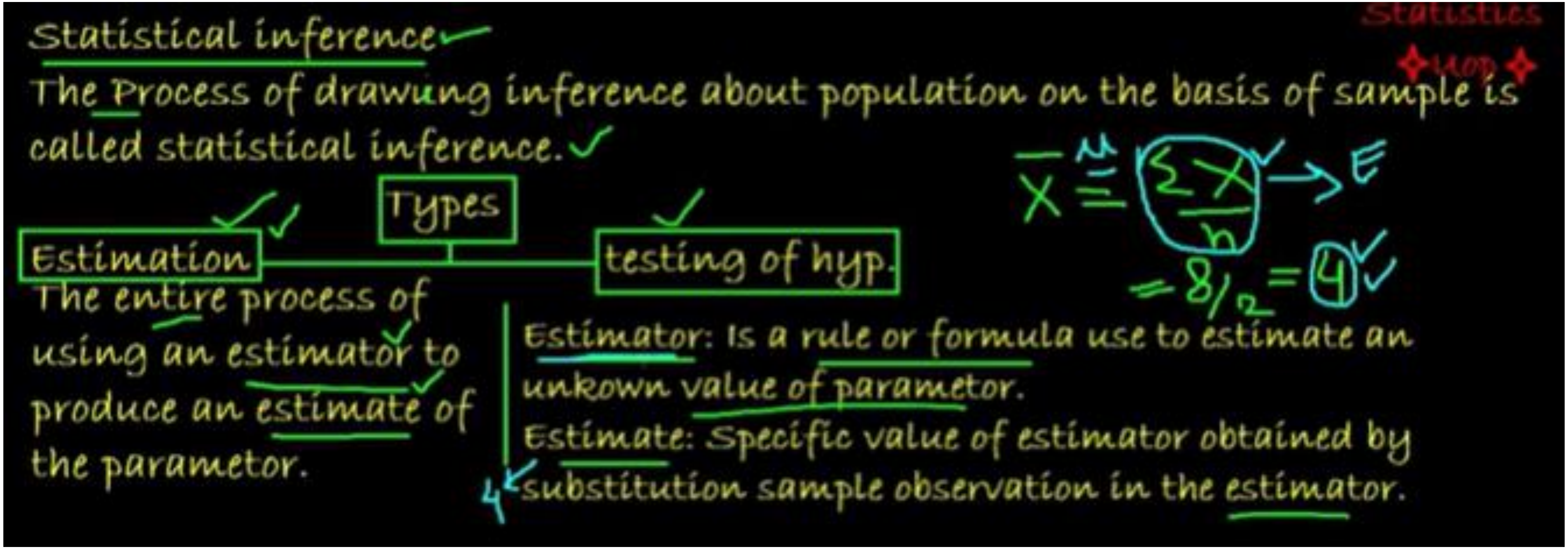
Statistical Inference

Estimation of Parameter

Testing of Hypothesis

Point Estimation

Interval Estimation



Statistical inference ✓

The Process of drawung inference about population on the basis of sample is called statistical inference. ✓

Types

Estimation ✓ ✓

testing of hyp. ✓

$$\bar{X} \overset{\mu}{=} \frac{\sum X}{n} \to E$$

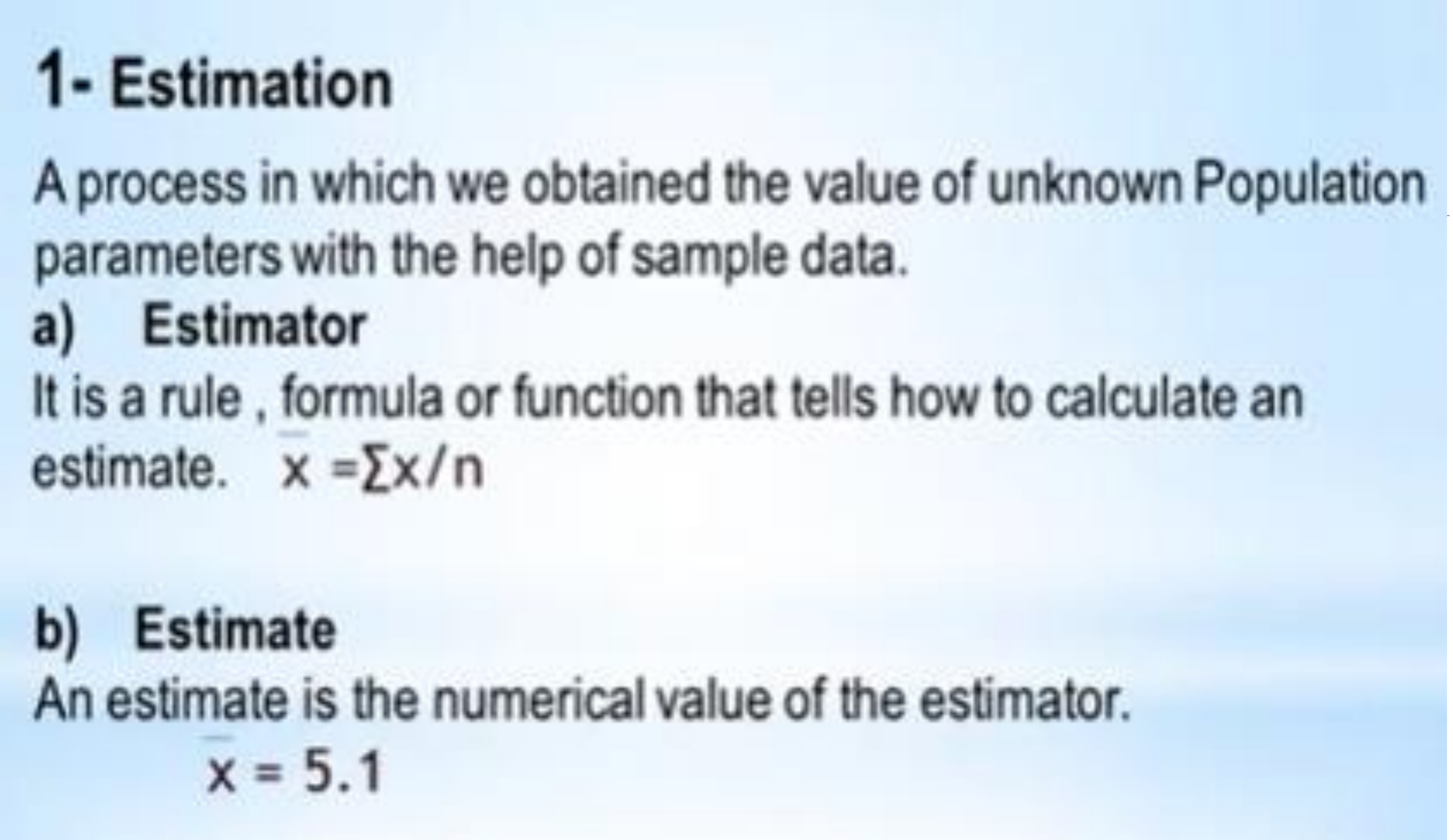
$$= 8/2 = \boxed{4} ✓ ✓$$

The entire process of using an estimator ✓ to produce an estimate ✓ of the parameter.

Estimator: Is a rule or formula use to estimate an unkown value of parameter.

Estimate: Specific value of estimator obtained by

4 ← substitution sample observation in the estimator.



1- Estimation

A process in which we obtained the value of unknown Population parameters with the help of sample data.

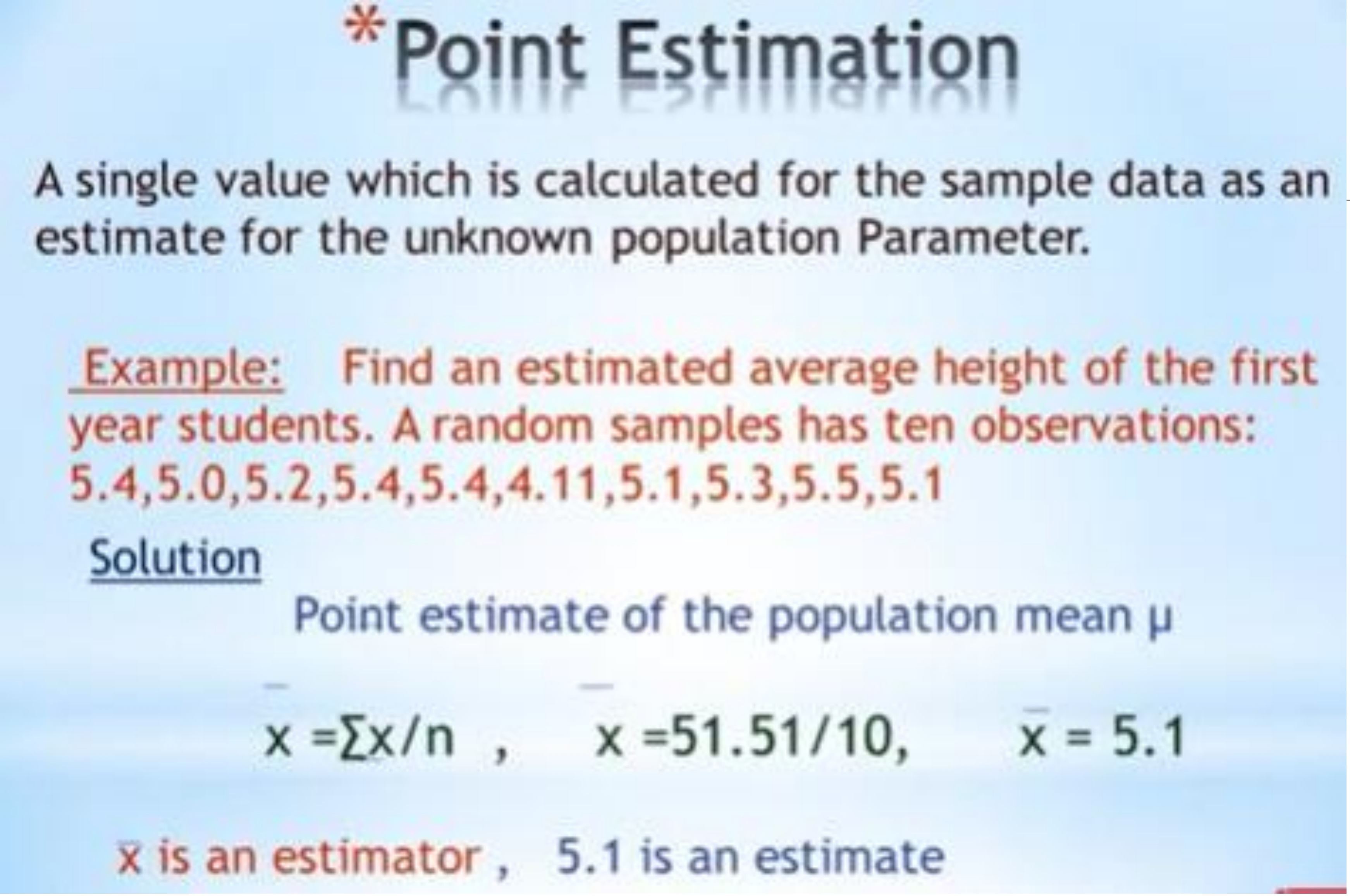
a) Estimator

It is a rule , formula or function that tells how to calculate an estimate. $\bar{x} = \sum x / n$

b) Estimate

An estimate is the numerical value of the estimator.

$$\bar{x} = 5.1$$



*Point Estimation

A single value which is calculated for the sample data as an estimate for the unknown population Parameter.

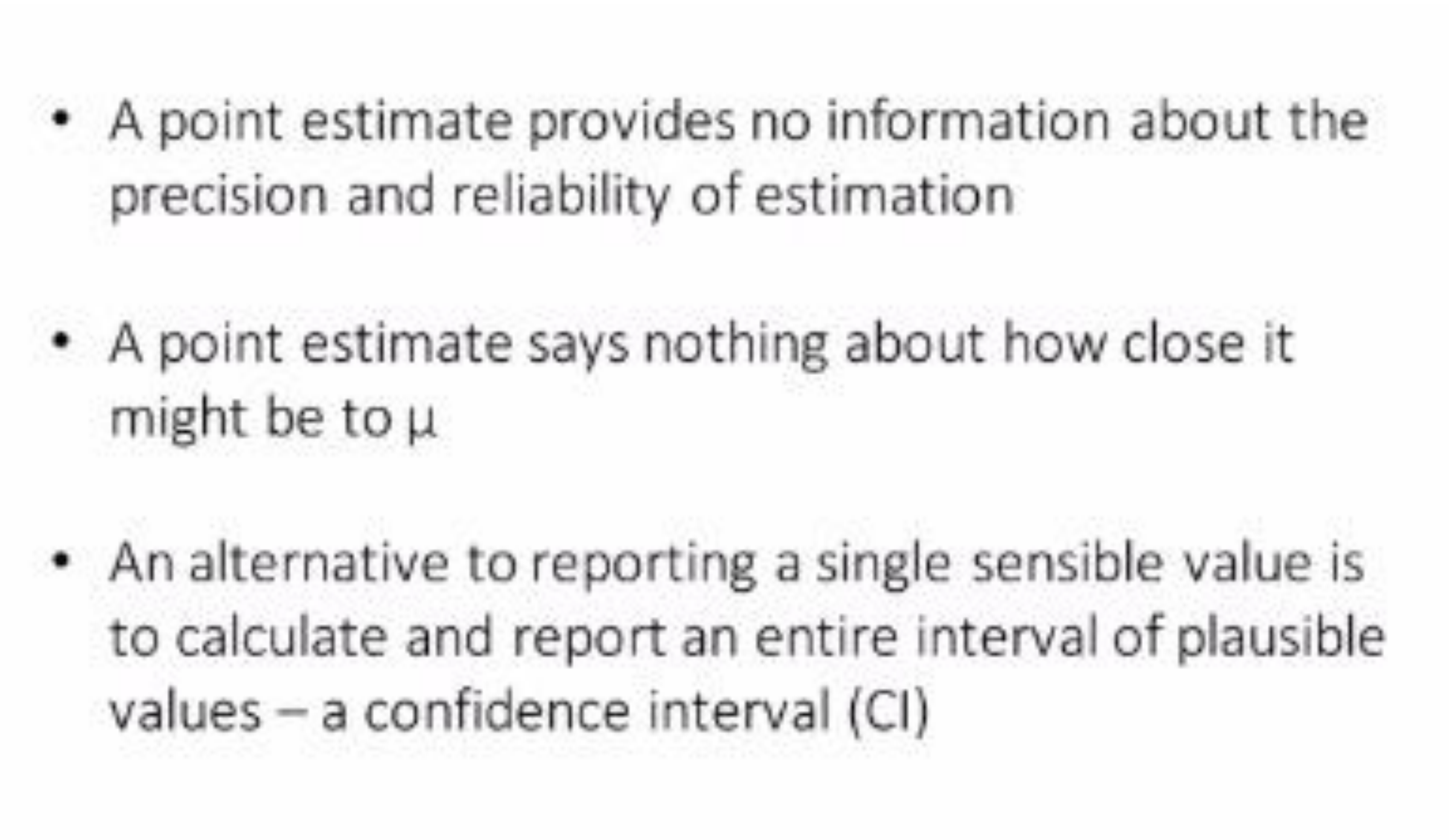
Example: Find an estimated average height of the first year students. A random samples has ten observations: 5.4,5.0,5.2,5.4,5.4,4.11,5.1,5.3,5.5,5.1

Solution

Point estimate of the population mean μ

$\bar{x} = \sum x/n$, $\bar{x}$ =51.51/10, $\bar{x}$ = 5.1

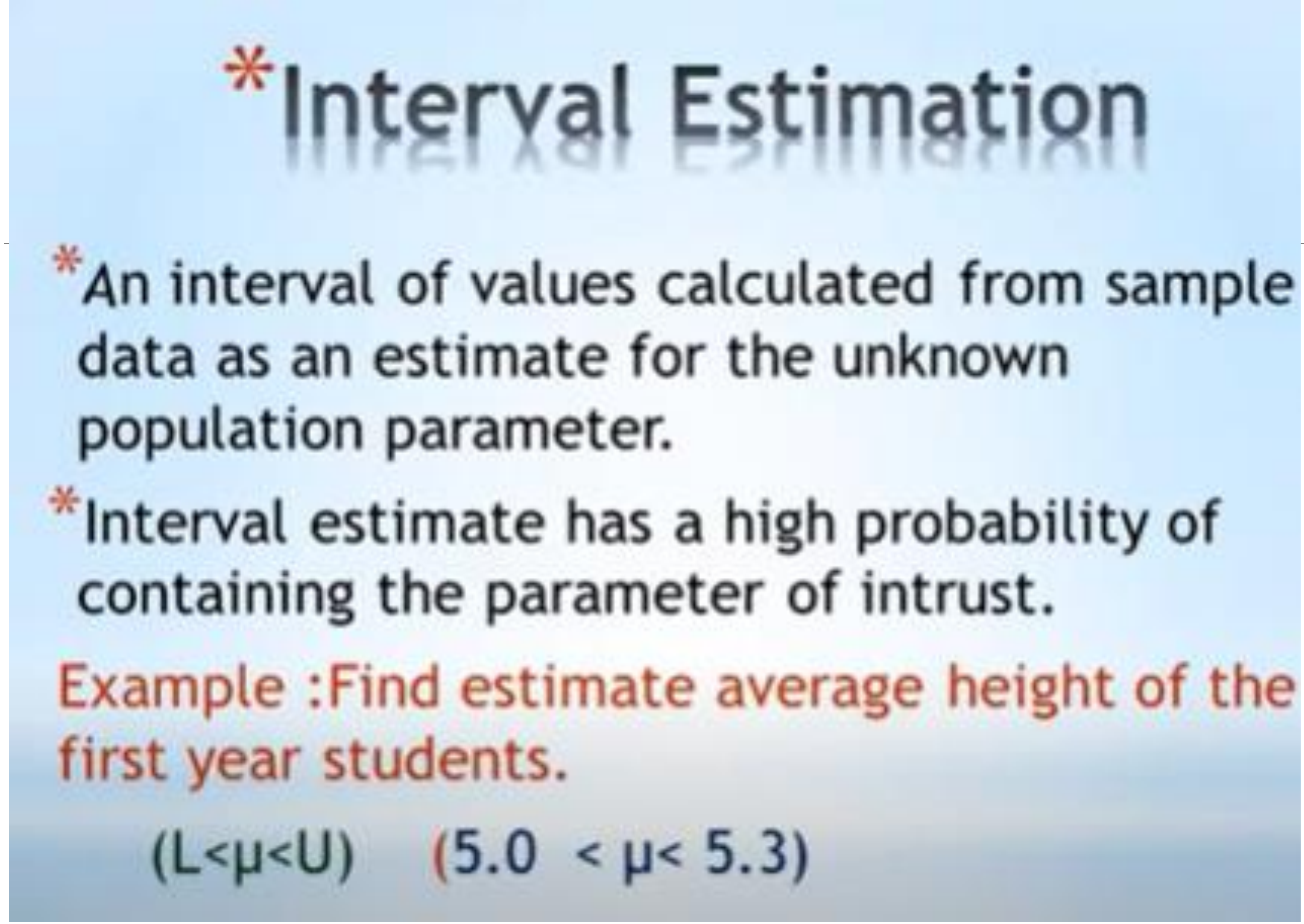
$\bar{x}$ is an estimator , 5.1 is an estimate



- A point estimate provides no information about the precision and reliability of estimation

- A point estimate says nothing about how close it might be to μ

- An alternative to reporting a single sensible value is to calculate and report an entire interval of plausible values – a confidence interval (CI)



*Interval Estimation

*An interval of values calculated from sample data as an estimate for the unknown population parameter.

*Interval estimate has a high probability of containing the parameter of intrust.

Example :Find estimate average height of the first year students.

$(L<\mu<U)$ $(5.0 < \mu < 5.3)$

Confidence Interval (CI) for the Mean μ

- Draw an SRS of size n from N(μ, σ). A level C confidence interval for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}.$$

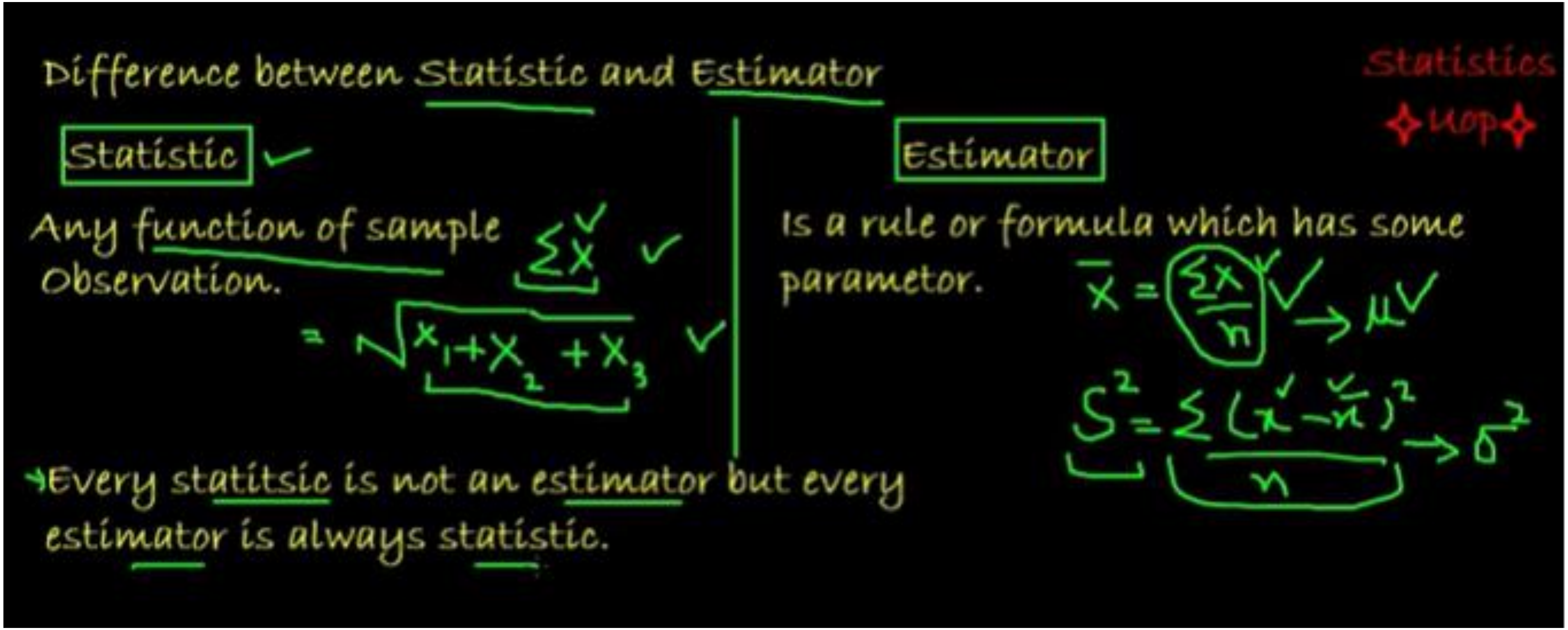
- Here z^* is called a critical value (CV), which, along with – z^*, marks the middle (100C)% of all values from N(0, 1). Note that the CI is written as **est ± (cv)(std).**
- There is a trade-off between the confidence level and the margin of error: *to obtain a small margin of error from the same data, you must be willing to accept lower confidence.*
- Increasing the sample size reduces the margin of error for any fixed confidence level.

Example: (Plasma Aldosterone in Dogs) Aldosterone is a hormone involved in maintaining fluid balance in the body. In a veterinary study, 8 dogs with heart failure were treated with the drug Captopril, and plasma concentrations of aldosterone were measured before and after the treatment.

Dogs	1	2	3	4	5	6	7	8
Before	210	219	222	224	310	345	360	400
After	202	211	210	211	289	310	323	339

Suppose that the before-after change (before – after) in concentration has a normal distribution with standard deviation 15.

1. Display all values on a single graph, with paired values connected by a line.
2. Find the 95% CI for the mean change. Answer: -24.4 ± 12.95



Difference between Statistic and Estimator

Statistic ✓

Any function of sample Observation.

$$\underbrace{\sum X^{\checkmark}} \checkmark$$

$$= \sqrt{\underbrace{x_1 + X_2 + X_3}} \checkmark$$

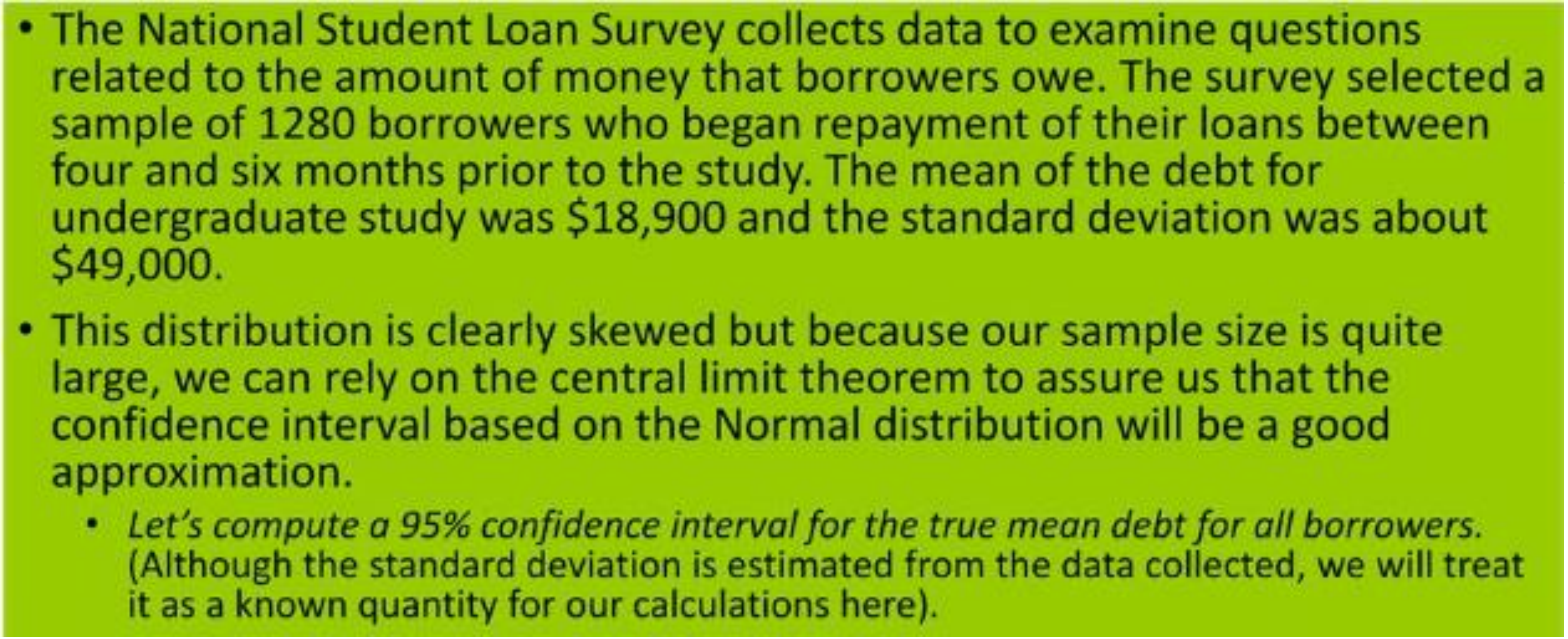
Estimator

Is a rule or formula which has some parameter.

$$\bar{X} = \frac{\sum X}{n} \checkmark \rightarrow \mu \checkmark$$

$$\underbrace{S^2} = \underbrace{\frac{\sum (x - \bar{x})^2}{n}} \rightarrow \sigma^2$$

→Every statitsic is not an estimator but every estimator is always statistic.



- The National Student Loan Survey collects data to examine questions related to the amount of money that borrowers owe. The survey selected a sample of 1280 borrowers who began repayment of their loans between four and six months prior to the study. The mean of the debt for undergraduate study was $18,900 and the standard deviation was about $49,000.
- This distribution is clearly skewed but because our sample size is quite large, we can rely on the central limit theorem to assure us that the confidence interval based on the Normal distribution will be a good approximation.
  - *Let's compute a 95% confidence interval for the true mean debt for all borrowers.* (Although the standard deviation is estimated from the data collected, we will treat it as a known quantity for our calculations here).

Example

- Calculations:

$$\begin{aligned} m &= z^* \frac{\sigma}{\sqrt{n}} \\ &= 1.960 \frac{49{,}000}{\sqrt{1280}} \\ &= 2684 \end{aligned}$$

- We’ll round 2684 to 2700 for the purposes of this example.

$$\begin{aligned} \bar{x} \pm m &= 18{,}900 \pm 2700 \\ &= (16{,}200,\ 21{,}600) \end{aligned}$$

Example

- Suppose the researchers who designed the National Student Loan Survey had used a different sample size.
  - How would this affect the confidence interval?
- We can answer this question by changing the sample size in our calculations and assuming that the mean and standard deviation are the same.
- Let’s assume that the sample mean of the debt for undergraduate study is $18,900 and the standard deviation is about $49,000, as in the previous example. But suppose that the sample size is only 320.
- The margin of error for 95% confidence is?

Example

$$\begin{aligned} m &= z^* \frac{\sigma}{\sqrt{n}} \\ &= 1.960 \frac{49{,}000}{\sqrt{320}} \\ &= 5400 \end{aligned}$$

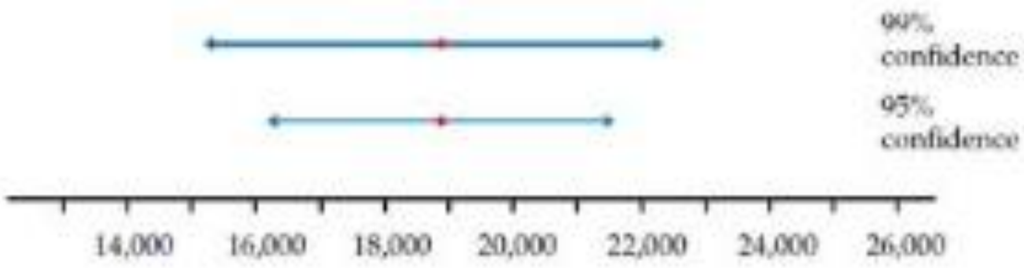
$$\begin{aligned} \bar{x} \pm m &= 18{,}900 \pm 5400 \\ &= (13{,}500,\ 24{,}300) \end{aligned}$$

5400 Vs 2700

- Notice that the margin of error for this example is twice as large as the margin of error that we just computed.
- The only change that we made was to assume that the sample size is 320 rather than 1280.
- This sample size is exactly one-fourth of the original 1280.
- Thus, we approximately double the margin of error when we reduce the sample size to one-fourth of the original value.

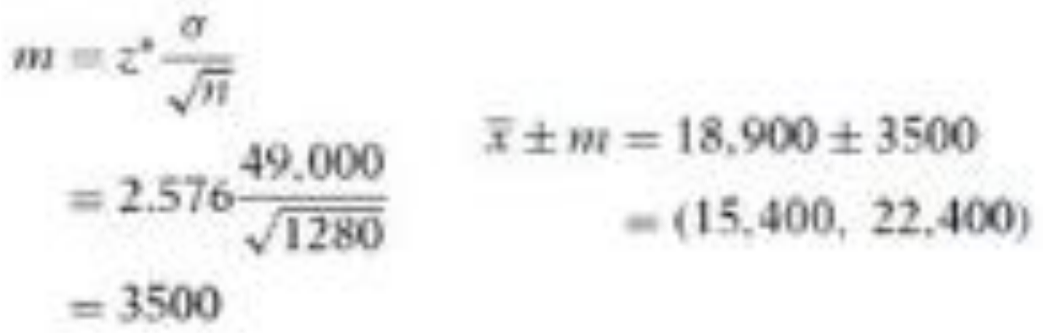
Confidence Interval

- One thing to note that by increasing the confidence interval from 95% to 99%. We make the interval bigger not smaller!
- This may seem strange at first but this diagram shows why:



- Suppose that for the student loan data in our example we wanted 99% confidence. For 99% confidence, z∗ = 2.576. The margin of error for 99% confidence based on 1280 observations is:

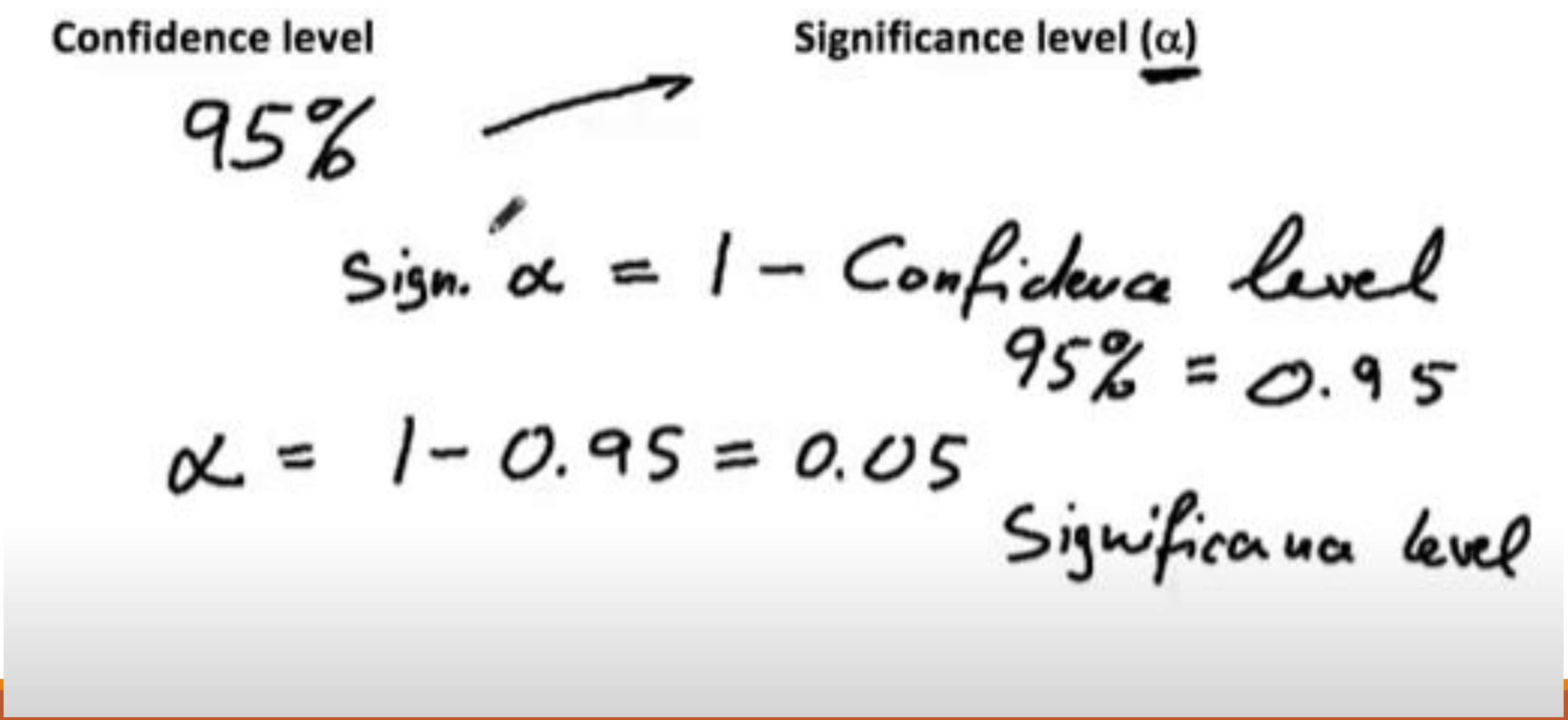
Confidence Interval



$$
\begin{aligned}
m &= z^* \frac{\sigma}{\sqrt{n}} \\
&= 2.576 \frac{49{,}000}{\sqrt{1280}} \\
&= 3500
\end{aligned}
$$

$$
\begin{aligned}
\bar{x} \pm m &= 18{,}900 \pm 3500 \\
&= (15{,}400,\ 22{,}400)
\end{aligned}
$$

Level of significance



Confidence level $\longrightarrow$ **Significance level ($\underline{\alpha}$)**

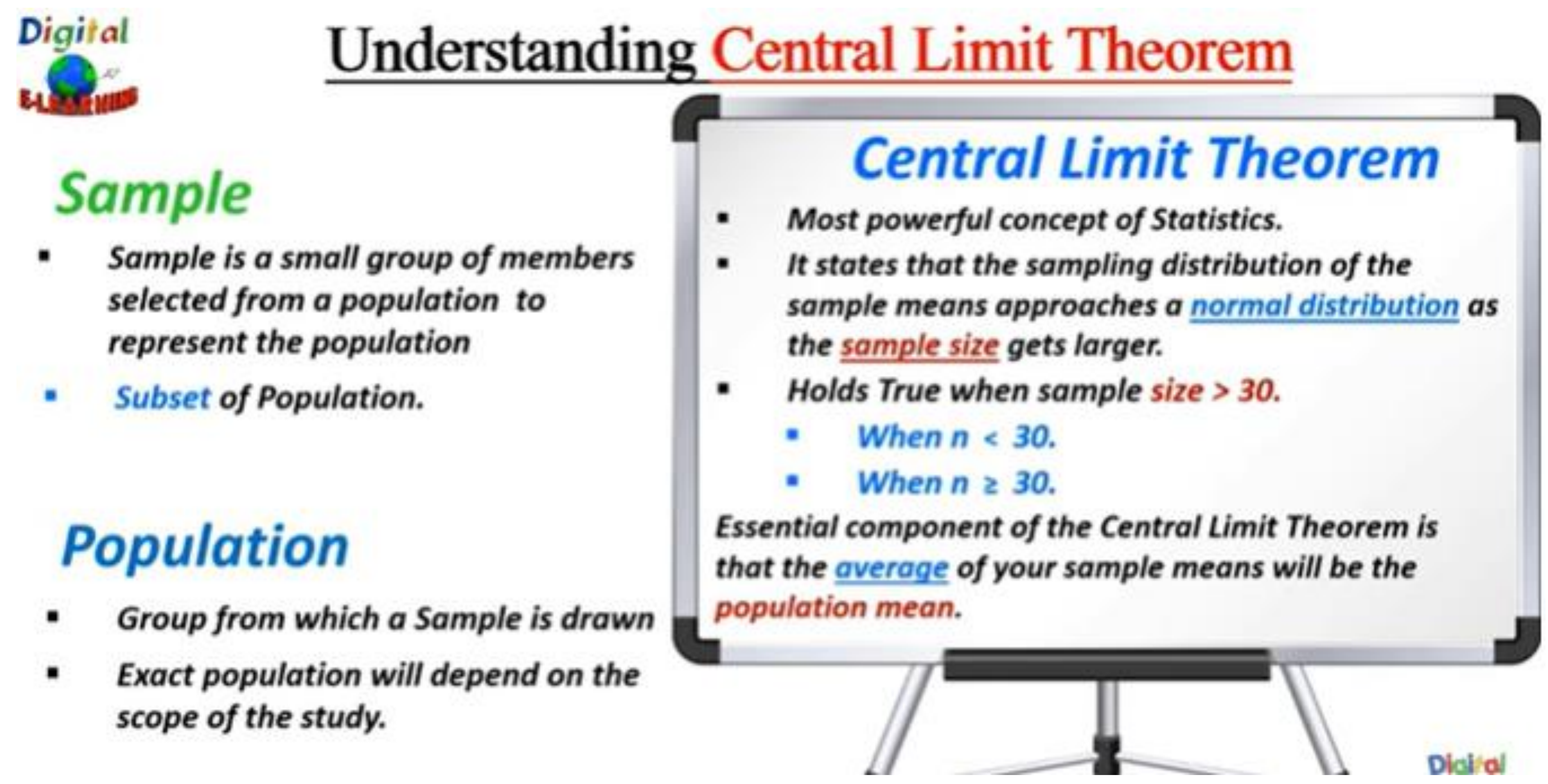
95%

Sign. $\alpha = 1 -$ Confidence level

$95\% = 0.95$

$\alpha = 1 - 0.95 = 0.05$

Significance level



Understanding Central Limit Theorem

Sample

- *Sample is a small group of members selected from a population to represent the population*
- *Subset of Population.*

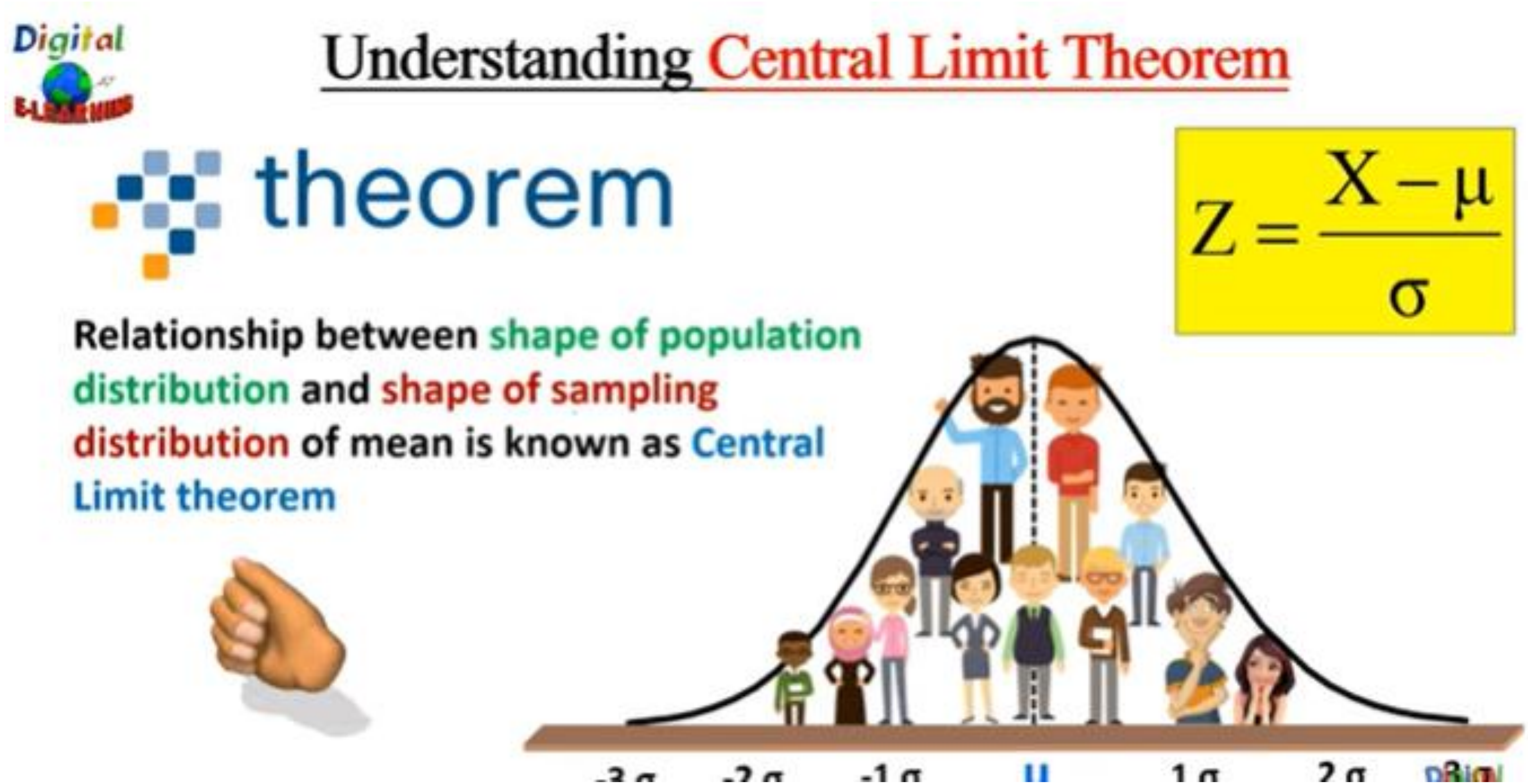
Population

- *Group from which a Sample is drawn*
- *Exact population will depend on the scope of the study.*

Central Limit Theorem

- *Most powerful concept of Statistics.*
- *It states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger.*
- *Holds True when sample size > 30.*
  - *When n < 30.*
  - *When n ≥ 30.*

Essential component of the Central Limit Theorem is that the average of your sample means will be the population mean.



Understanding Central Limit Theorem

theorem

Relationship between shape of population distribution and shape of sampling distribution of mean is known as Central Limit theorem

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

-3 σ
-2 σ
-1 σ
μ
1 σ
2 σ
3σ

Implement

Take a data set

Plot population distribution

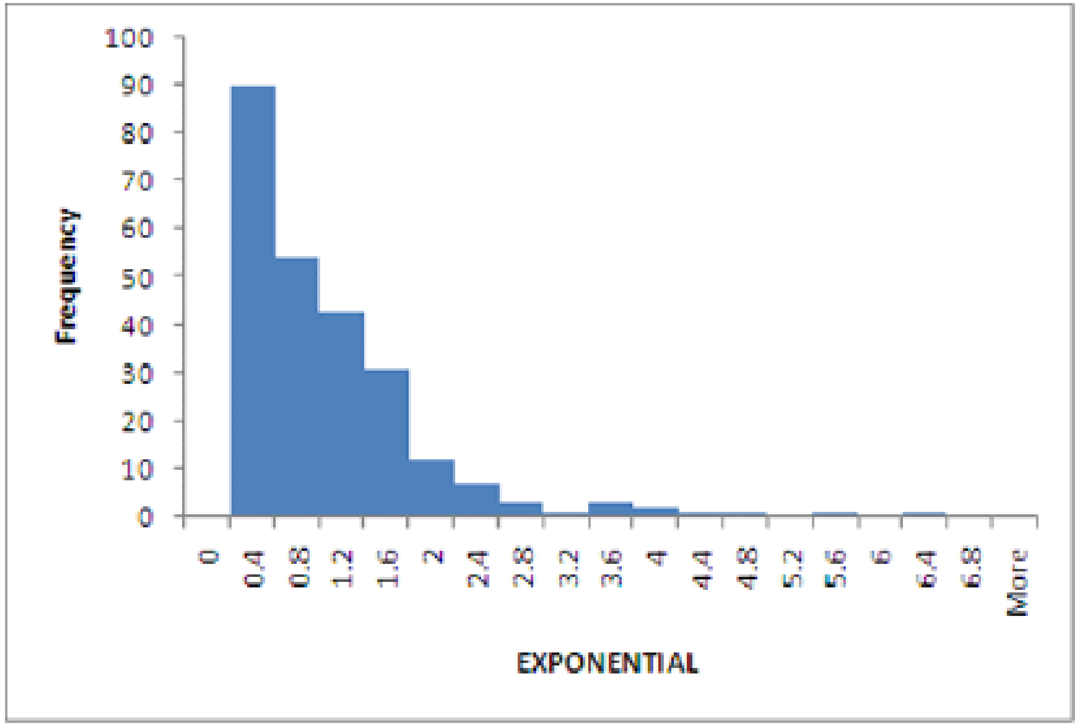
Plot sampling distribution of mean with 20 samples

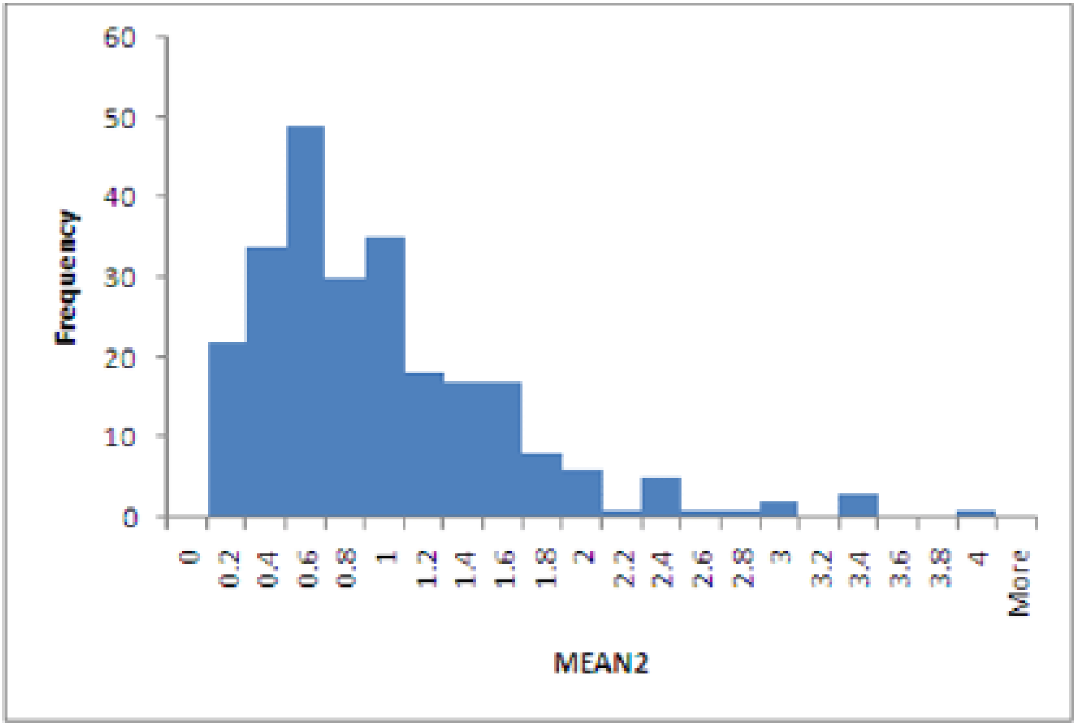
Plot sampling distribution of mean with 30 samples

Plot sampling distribution of mean with 50 samples

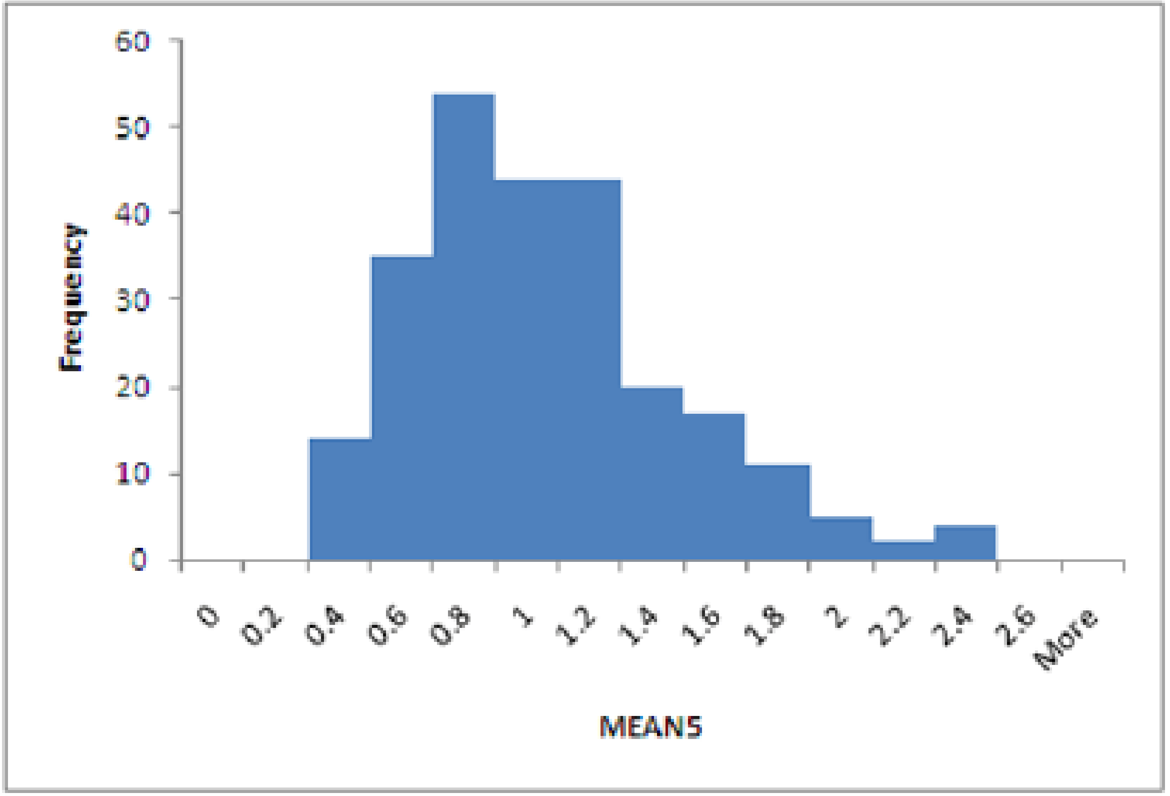
Plot sampling distribution of mean with 100 samples

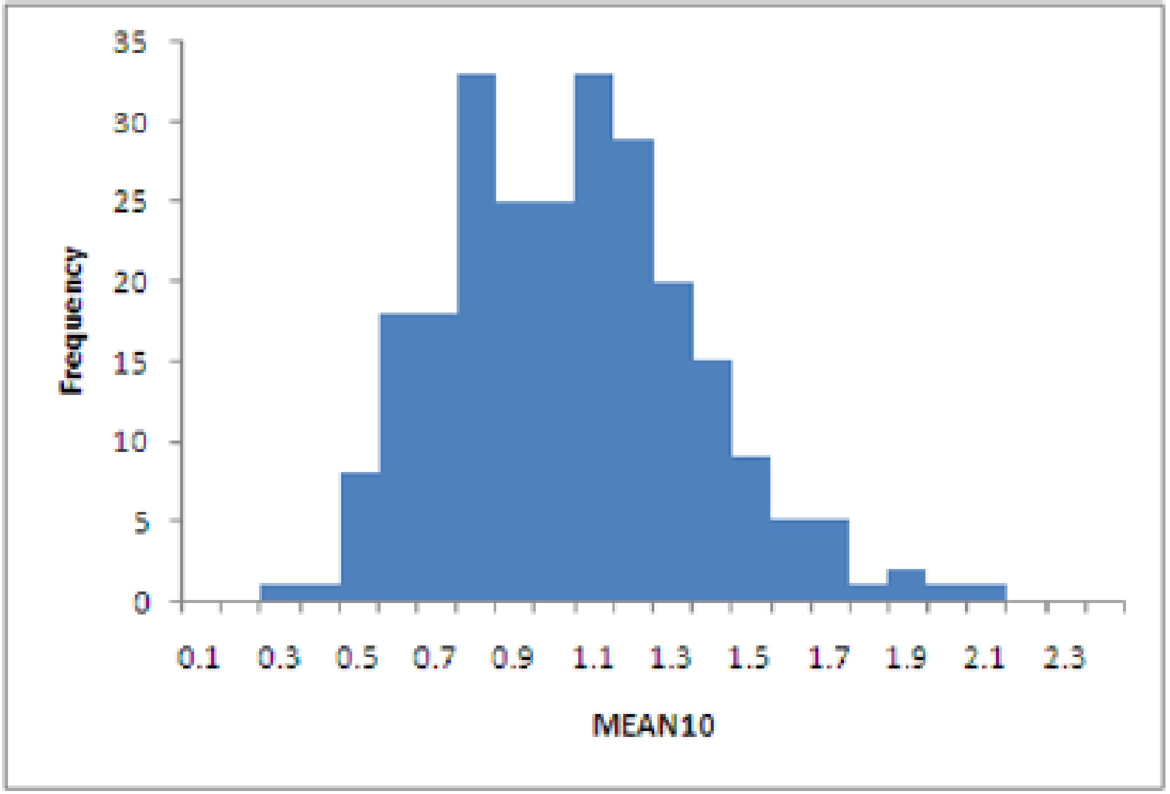
Sampling distributions - Example



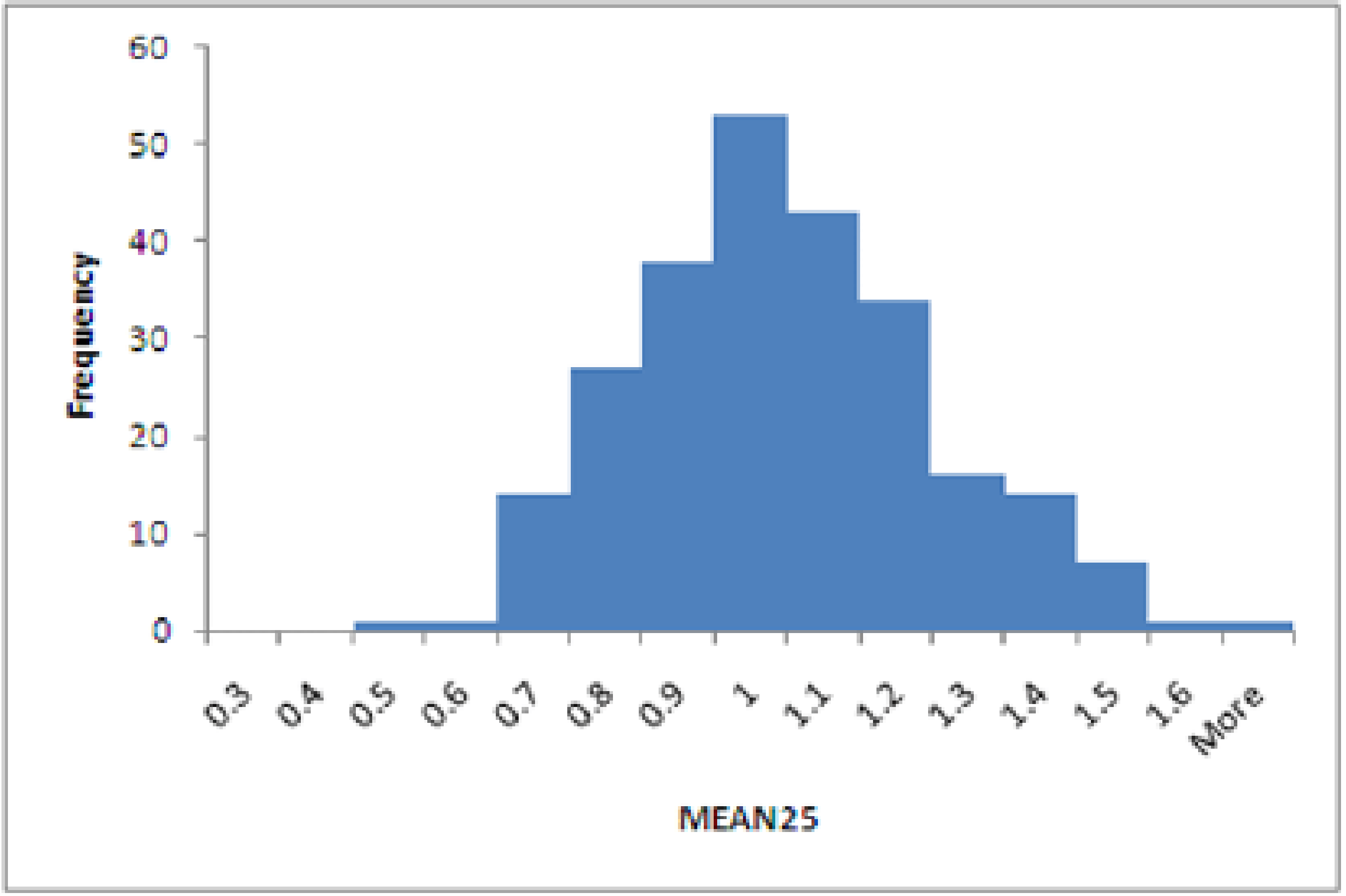


Sampling distributions - Example





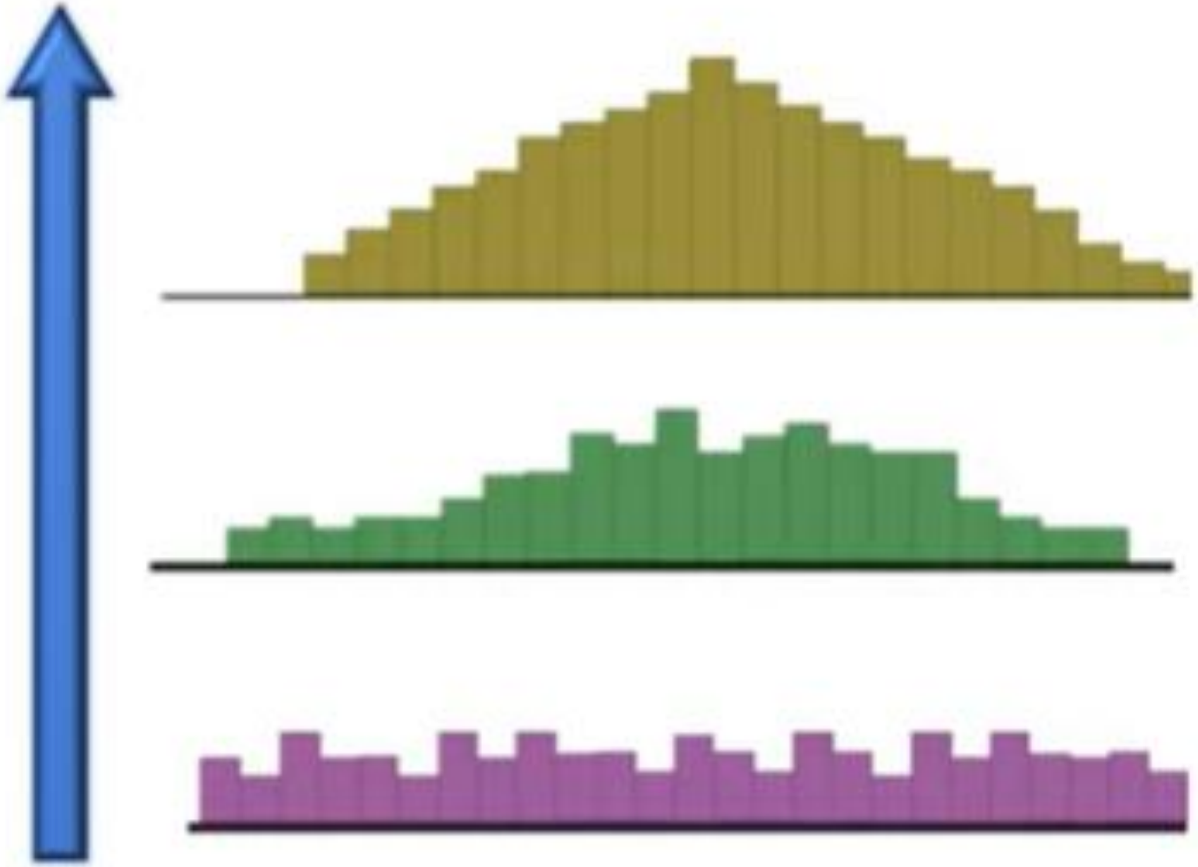
Sampling distributions - Example



Understanding Central Limit Theorem

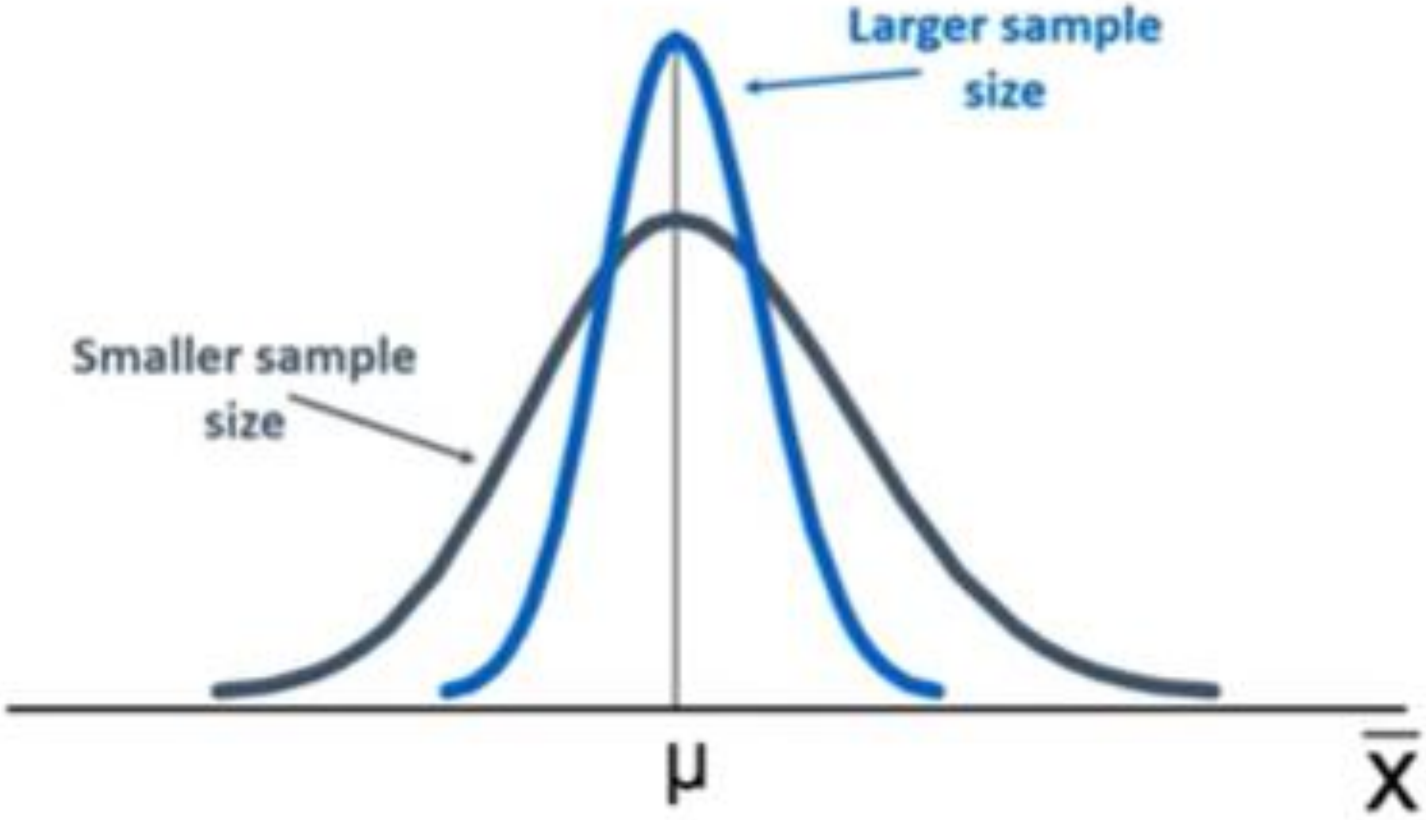
N

As the sample size get large, sampling distribution becomes almost Normal regardless of shape of population



Understanding Central Limit Theorem

As the sample size gets large enough…



the sampling distribution becomes almost normal regardless of shape of population

Understanding Central Limit Theorem

Question : There are 250 cats at show, where average weight of 12 Kg, with standard deviation of 8 Kg. If we choose 4 samples, then what is the probability they have an average weight of greater than 10 Kg and less than 25 Kg?

Understanding Central Limit Theorem

Question : There are 250 cats at show, where average weight of 12 Kg, with standard deviation of 8 Kg. If we choose 4 samples, then what is the probability they have an average weight of greater than 10 Kg and less than 25 Kg?

$$\boxed{Z = \frac{X - \mu}{\sigma_x}} \quad \boxed{\sigma_x = \frac{\sigma}{\sqrt{n}}}$$

Sol : Mean of Population $\mu = 12$ Kg $\sigma = 8$ Kg $n = 4$ $\sigma_x = 8/\sqrt{4}$

Where,
μ = Population mean
σ = Population standard deviation
$\mu_{\bar{x}}$ = Sample mean
$\sigma_{\bar{x}}$ = Sample standard deviation
n = Sample size

$$Z_{25} = \frac{25 - 12}{8/\sqrt{4}} \qquad Z_{25} = \frac{13}{4} \qquad Z_{25} = 3.25$$

$Z_{\text{value } 3.25} = 0.994$

$Z_{\text{value } 3.25} = 0.994 - 0.5$

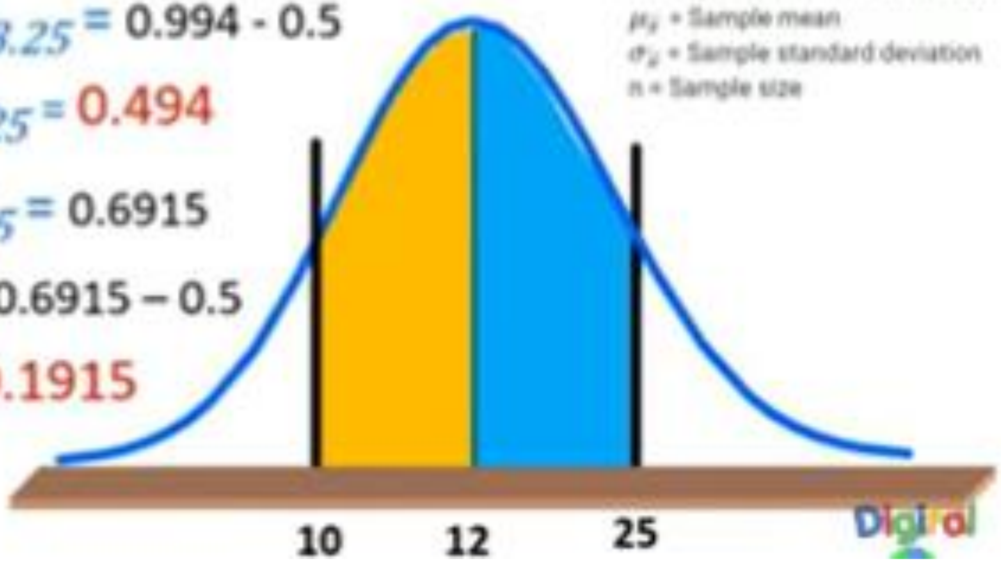
$Z_{\text{value } 3.25} = 0.494$

$$Z_{10} = \frac{10 - 12}{8/\sqrt{4}} \qquad Z_{10} = \frac{-2}{4} \qquad Z_{10} = -0.5$$

$Z_{\text{value } -0.5} = 0.6915$

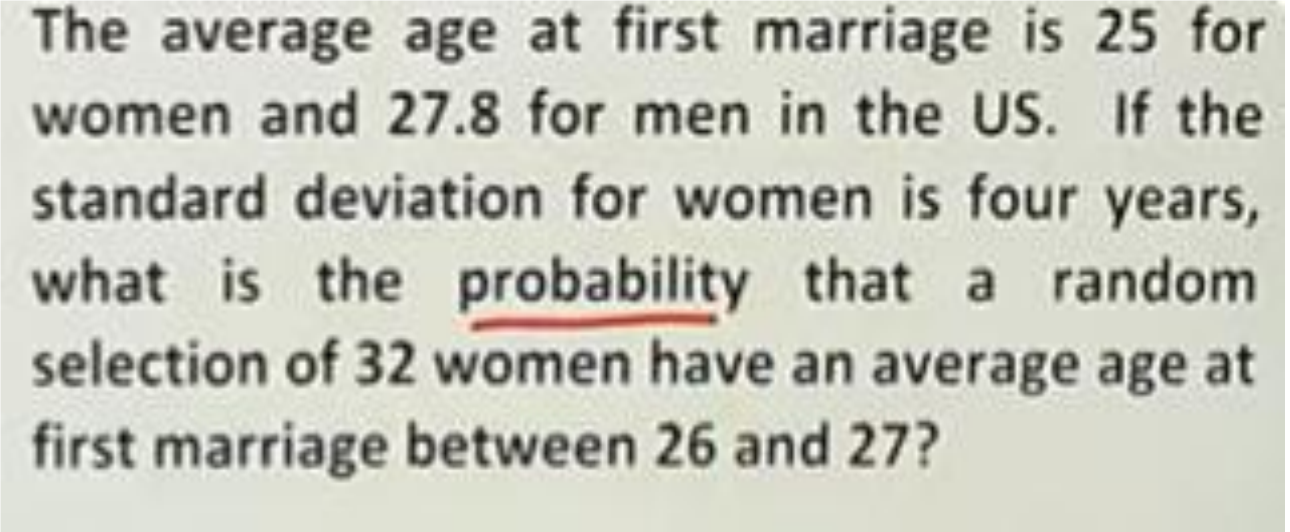
$Z_{\text{value } -0.5} = 0.6915 - 0.5$

$Z_{\text{value } -0.5} = 0.1915$



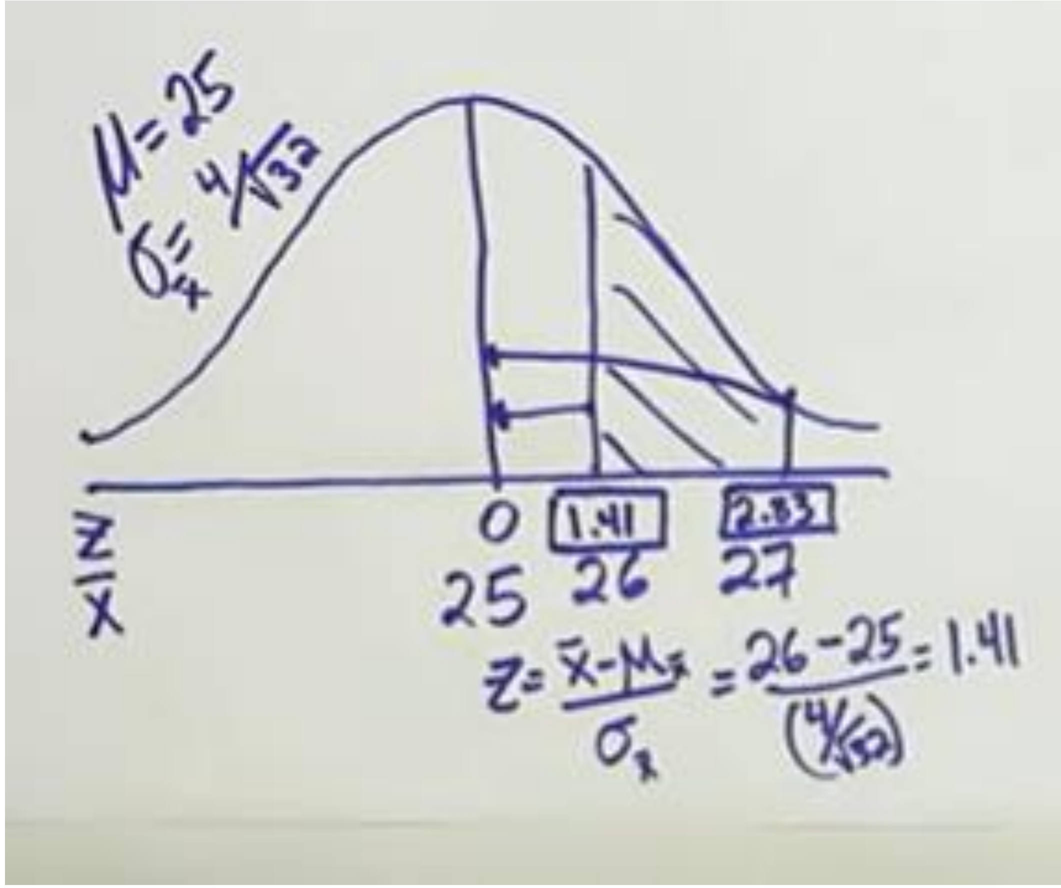
Adding = 0.494 +0.1915 = .6855 => 68.5%

Central Limit Theorem



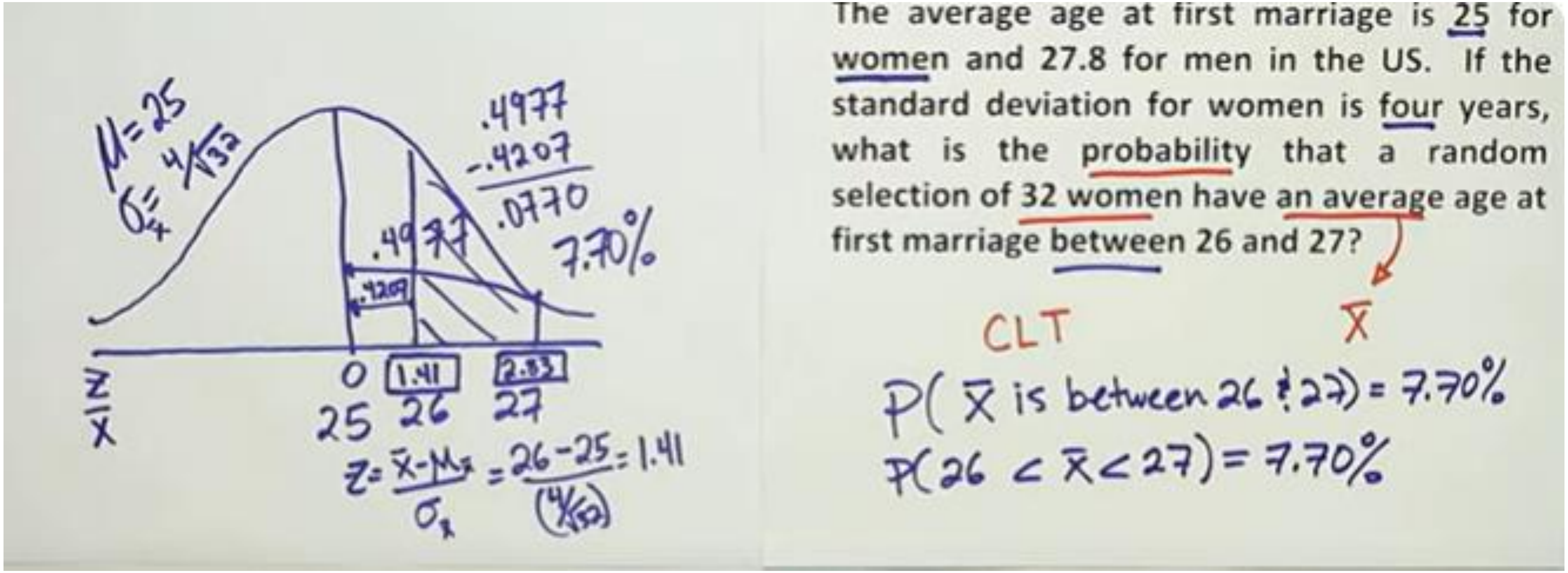
The average age at first marriage is 25 for women and 27.8 for men in the US. If the standard deviation for women is four years, what is the probability that a random selection of 32 women have an average age at first marriage between 26 and 27?

Central Limit Theorem



$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{26 - 25}{(4/\sqrt{32})} = 1.41$$

Central Limit Theorem



$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{26-25}{(4/\sqrt{32})} = 1.41$

The average age at first marriage is 25 for women and 27.8 for men in the US. If the standard deviation for women is four years, what is the probability that a random selection of 32 women have an average age at first marriage between 26 and 27?

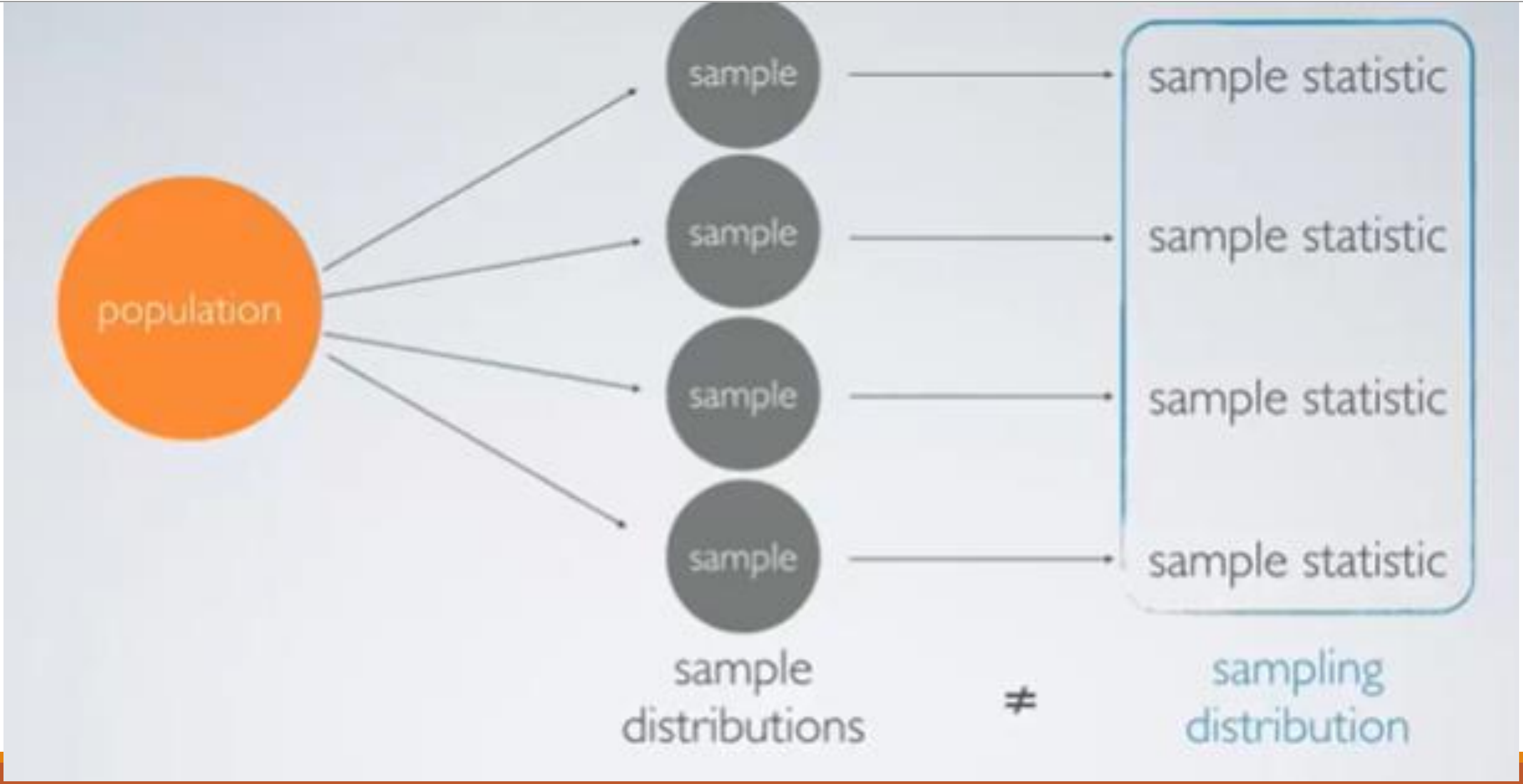
CLT

$\bar{x}$

$P(\bar{x}$ is between 26 & 27$) = 7.70\%$

$P(26 < \bar{x} < 27) = 7.70\%$

Sampling Variability and CLT



Sampling Variability

Sampling variability is the variation or fluctuation in sample statistics (such as the sample mean, sample variance, etc.) that occurs when different random samples are taken from the same population.

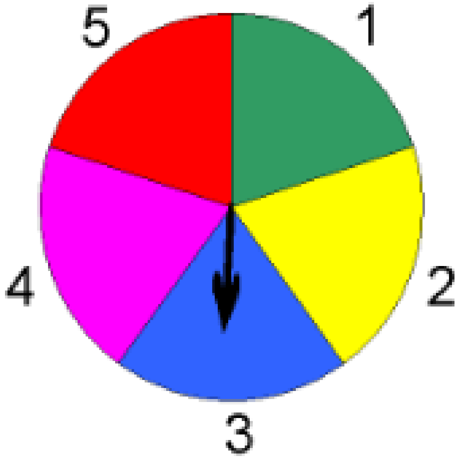
Sampling Distribution: Distribution of a statistics.

Sampling variability or Standard Error (SE) or Standard Error of mean:

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

Example

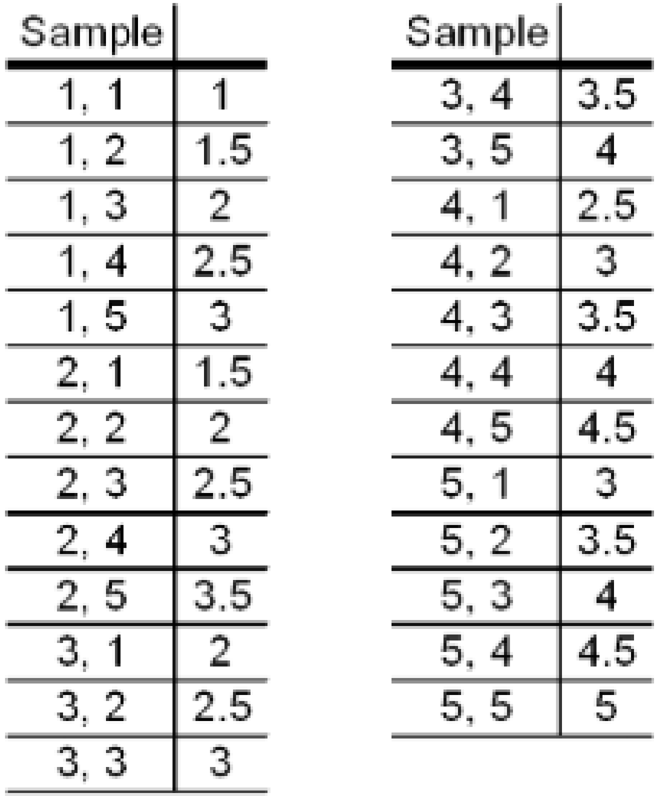
Consider a population that consists of the numbers 1, 2, 3, 4 and 5 generated in a manner that the probability of each of those values is 0.2 no matter what the previous selections were. This population could be described as the outcome associated with a spinner such as given below with the distribution next to it.



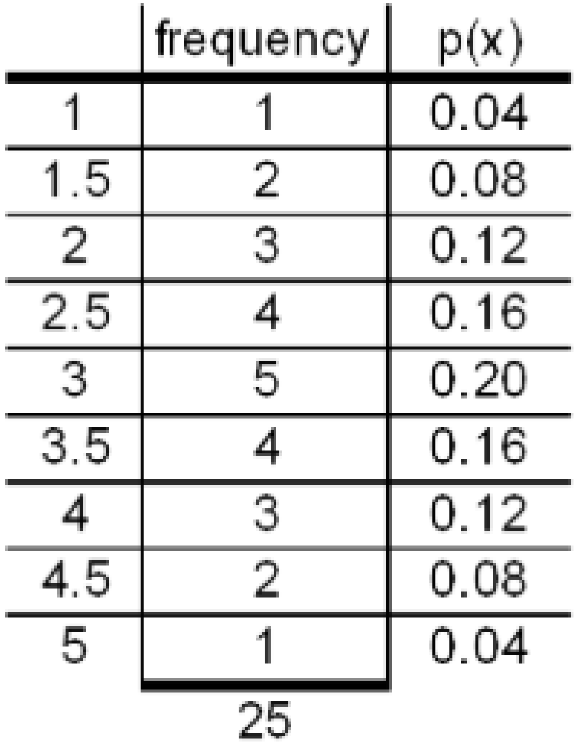
x	p(x)
1	0.2
2	0.2
3	0.2
4	0.2
5	0.2

Example

If the sampling distribution for the means of samples of size two is analyzed, it looks like



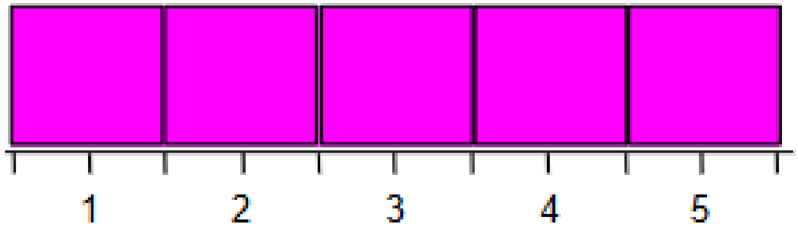
Sample	
1, 1	1
1, 2	1.5
1, 3	2
1, 4	2.5
1, 5	3
2, 1	1.5
2, 2	2
2, 3	2.5
2, 4	3
2, 5	3.5
3, 1	2
3, 2	2.5
3, 3	3
3, 4	3.5
3, 5	4
4, 1	2.5
4, 2	3
4, 3	3.5
4, 4	4
4, 5	4.5
5, 1	3
5, 2	3.5
5, 3	4
5, 4	4.5
5, 5	5



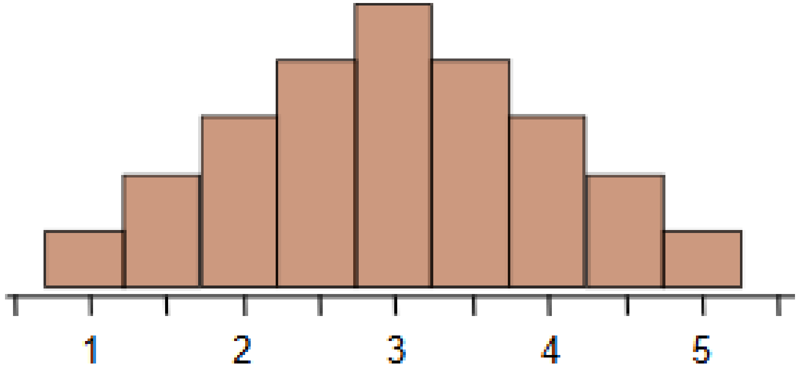
	frequency	p(x)
1	1	0.04
1.5	2	0.08
2	3	0.12
2.5	4	0.16
3	5	0.20
3.5	4	0.16
4	3	0.12
4.5	2	0.08
5	1	0.04
	25	

Example

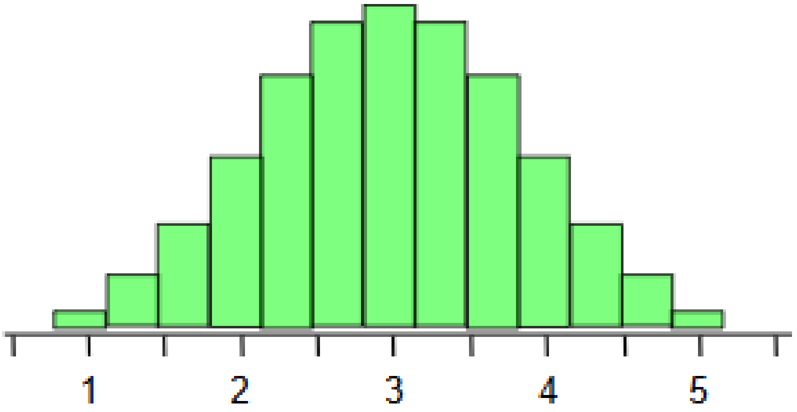
Sampling distributions for n=3 and n=4 were calculated and are illustrated below.



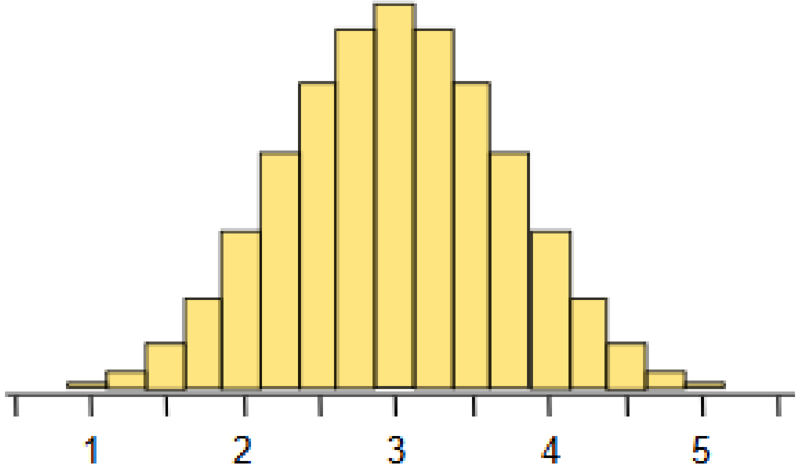
Original distribution



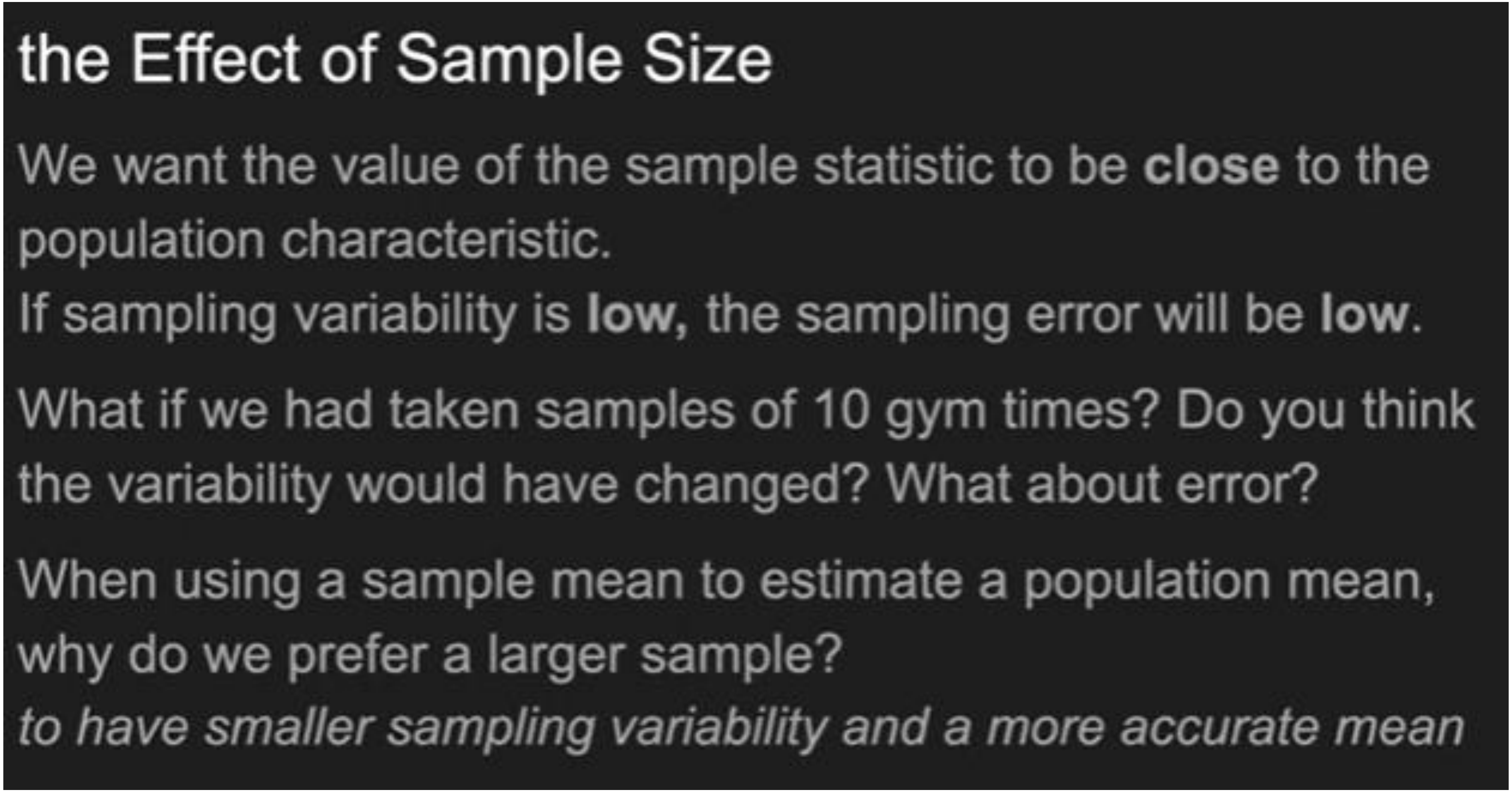
Sampling distribution n = 2



Sampling distribution n = 3



Sampling distribution n = 4



the Effect of Sample Size

We want the value of the sample statistic to be **close** to the population characteristic.
If sampling variability is **low,** the sampling error will be **low**.

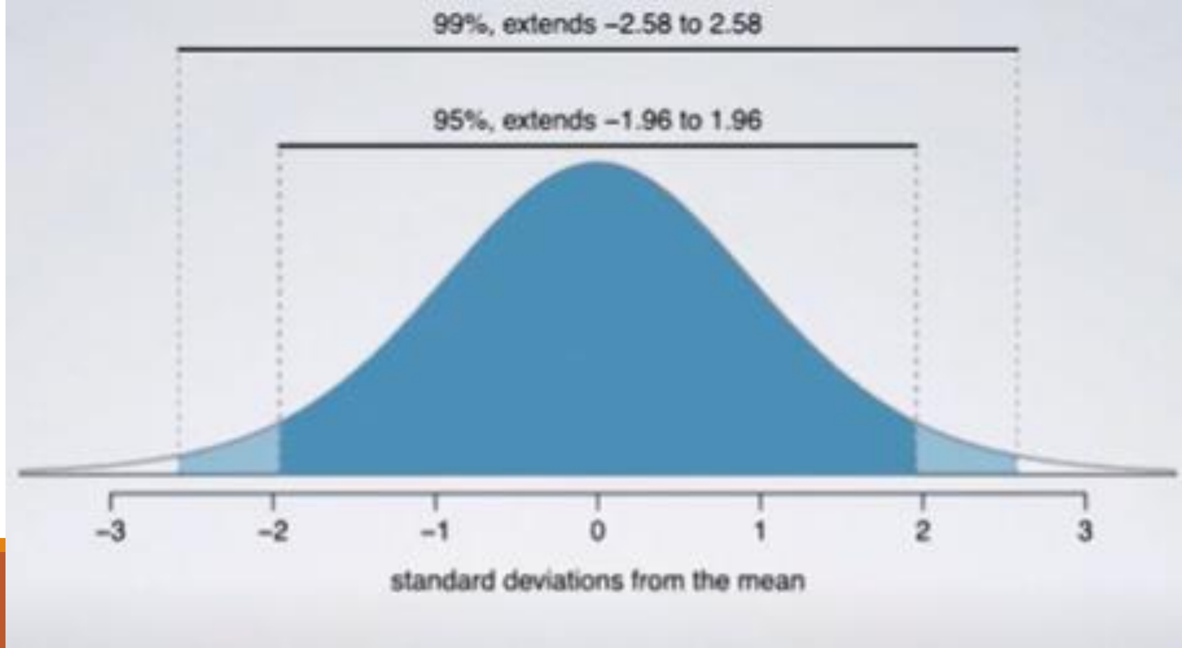
What if we had taken samples of 10 gym times? Do you think the variability would have changed? What about error?

When using a sample mean to estimate a population mean, why do we prefer a larger sample?
to have smaller sampling variability and a more accurate mean

Precision Vs Accuracy

Precision: How close together repeated measurements or estimates are regardless of whether they are close to the true value. It reflects the degree of variability in the sample data.

Narrow Confidence Interval gives Higher Precision, Wide Confidence Interval gives Lower Precision:



Precision Vs Accuracy

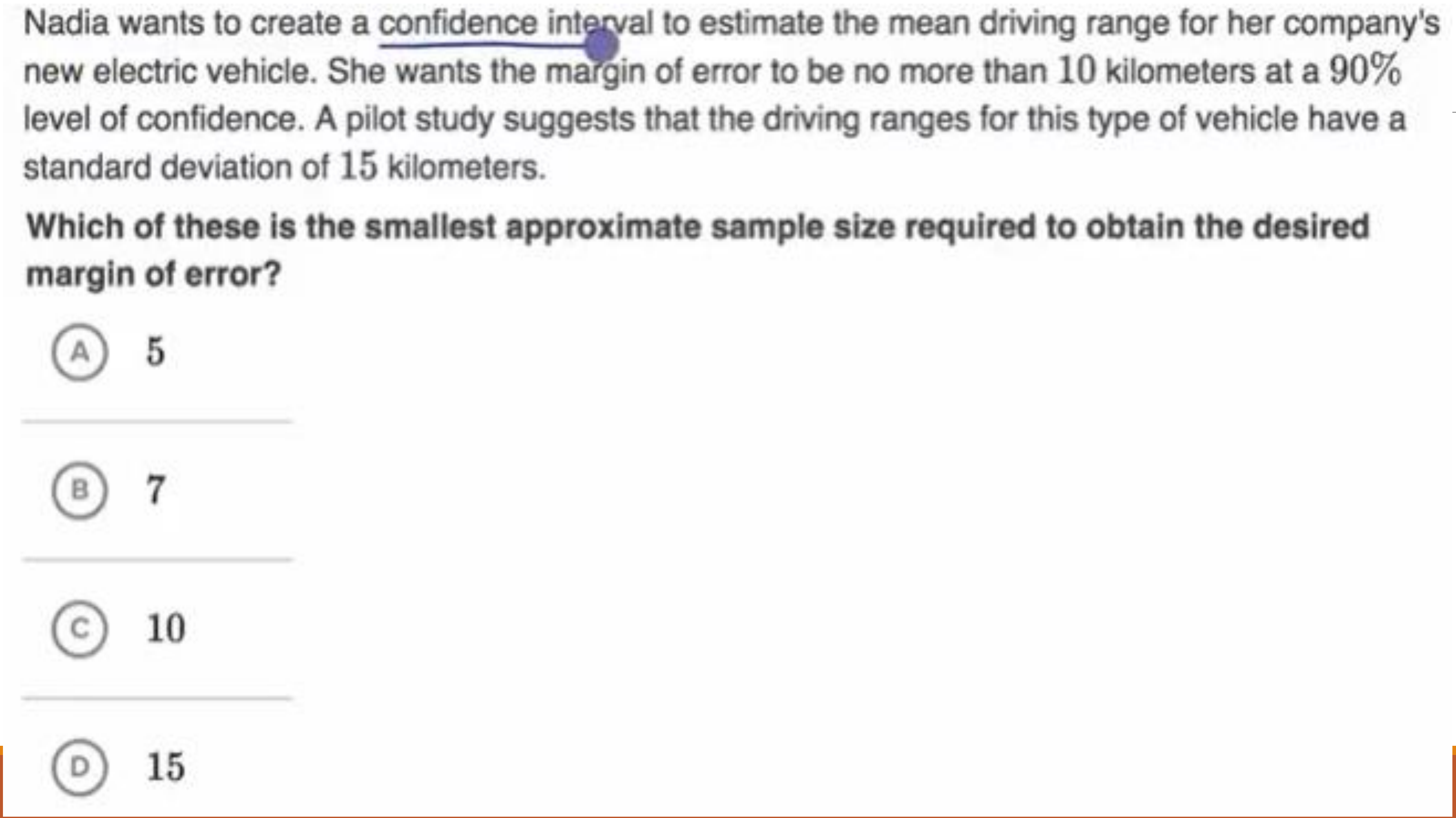
Accuracy: It refers to how close a measurement or estimate is to the true value or the true population parameter.

The confidence interval (CI) itself is a measure of uncertainty around an estimate.

A sample can have a **narrow confidence interval** but still be **inaccurate** if the sample is biased.

Similarly a sample can have a **wide confidence interval** and still be **accurate** if the sample mean is close to the true population value.

Required Sample Size for ME



Nadia wants to create a confidence interval to estimate the mean driving range for her company's new electric vehicle. She wants the margin of error to be no more than 10 kilometers at a 90% level of confidence. A pilot study suggests that the driving ranges for this type of vehicle have a standard deviation of 15 kilometers.

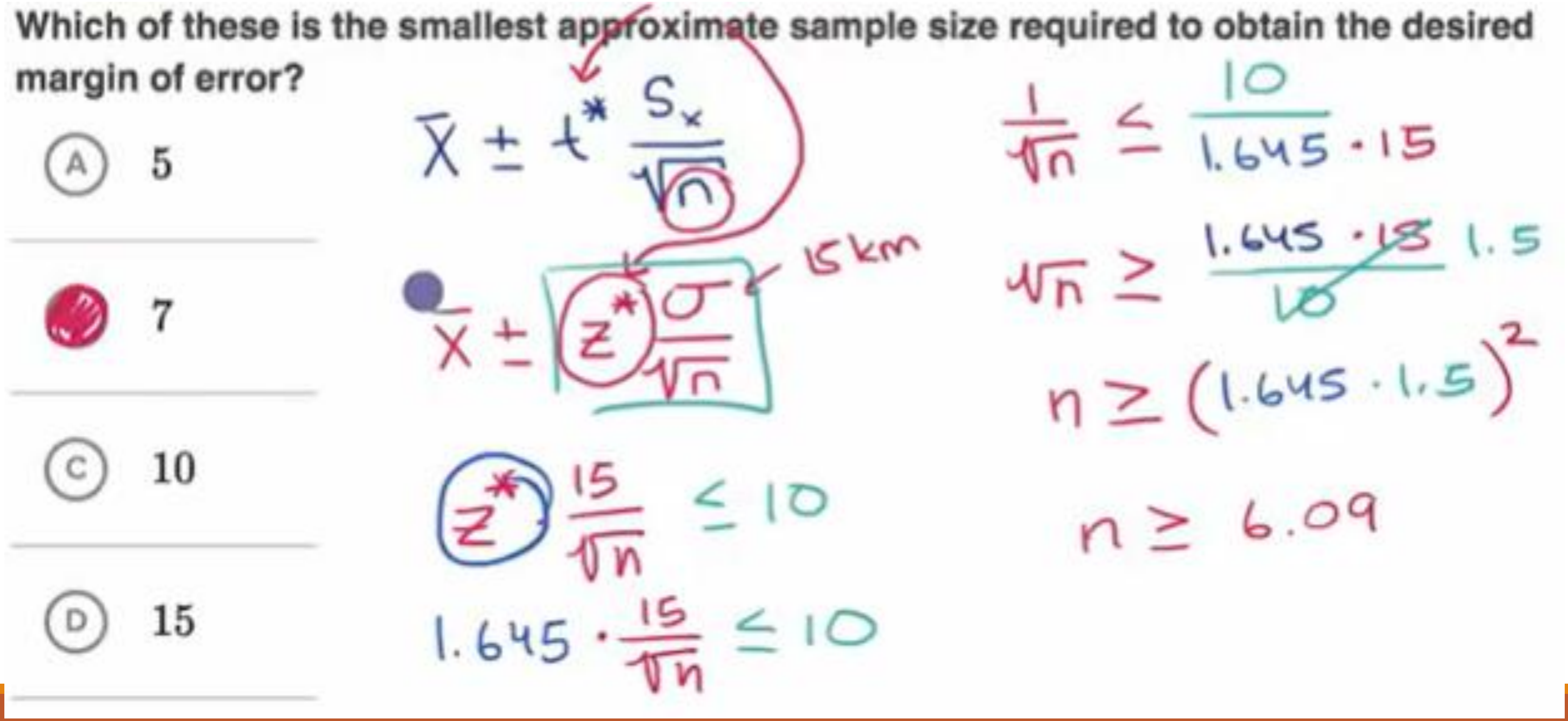
Which of these is the smallest approximate sample size required to obtain the desired margin of error?

Ⓐ 5

Ⓑ 7

Ⓒ 10

Ⓓ 15



Which of these is the smallest approximate sample size required to obtain the desired margin of error?

○ A 5

● B 7

○ C 10

○ D 15

$$\bar{X} \pm t^* \frac{S_x}{\sqrt{n}}$$

$$\bar{X} \pm \boxed{z^* \frac{\sigma}{\sqrt{n}}}$$

15km

$$z^* \frac{15}{\sqrt{n}} \le 10$$

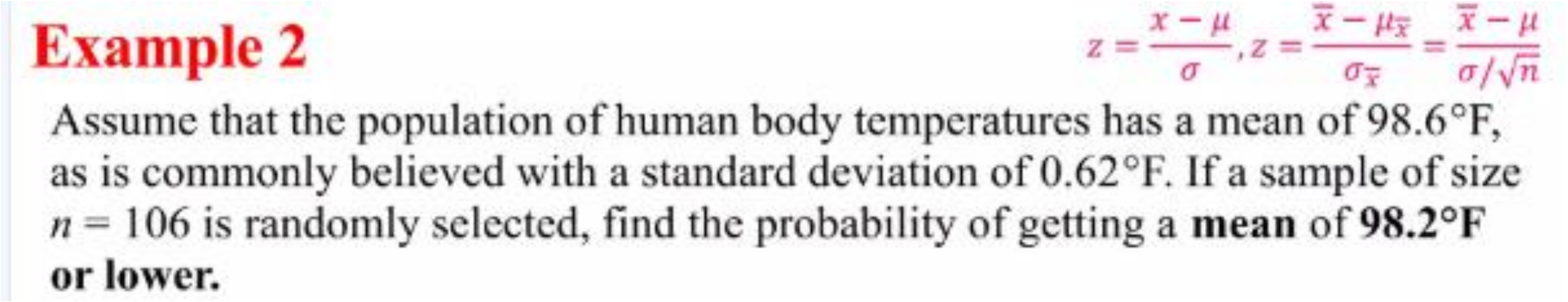
$$1.645 \cdot \frac{15}{\sqrt{n}} \le 10$$

$$\frac{1}{\sqrt{n}} \le \frac{10}{1.645 \cdot 15}$$

$$\sqrt{n} \ge \frac{1.645 \cdot \cancel{15}\ 1.5}{\cancel{10}}$$

$$n \ge (1.645 \cdot 1.5)^2$$

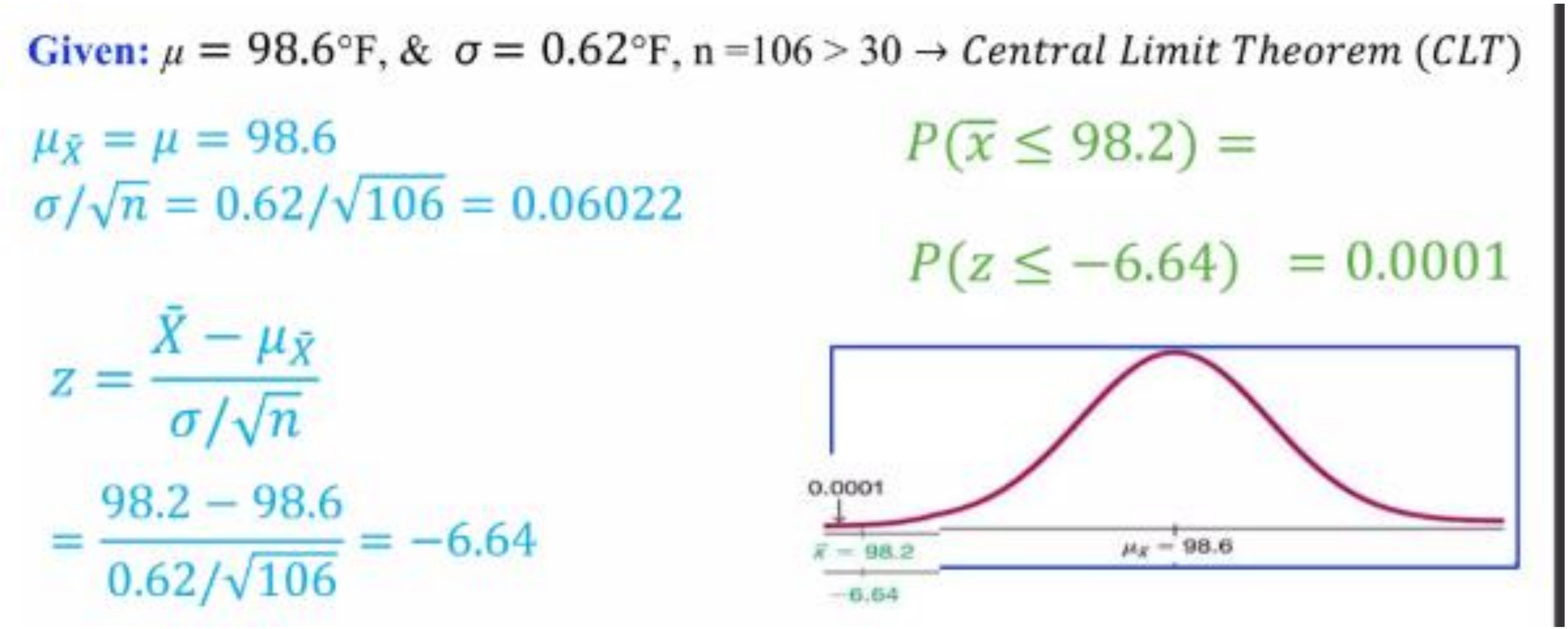
$$n \ge 6.09$$



Example 2

$$z = \frac{x-\mu}{\sigma}, z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

Assume that the population of human body temperatures has a mean of 98.6°F, as is commonly believed with a standard deviation of 0.62°F. If a sample of size $n = 106$ is randomly selected, find the probability of getting a **mean** of **98.2°F or lower.**



Given: $\mu = 98.6$°F, & $\sigma = 0.62$°F, n =106 > 30 → $Central\ Limit\ Theorem\ (CLT)$

$$\mu_{\bar{X}} = \mu = 98.6$$
$$\sigma/\sqrt{n} = 0.62/\sqrt{106} = 0.06022$$

$$P(\overline{x} \le 98.2) =$$

$$P(z \le -6.64) \quad = 0.0001$$

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma/\sqrt{n}}$$

$$= \frac{98.2 - 98.6}{0.62/\sqrt{106}} = -6.64$$