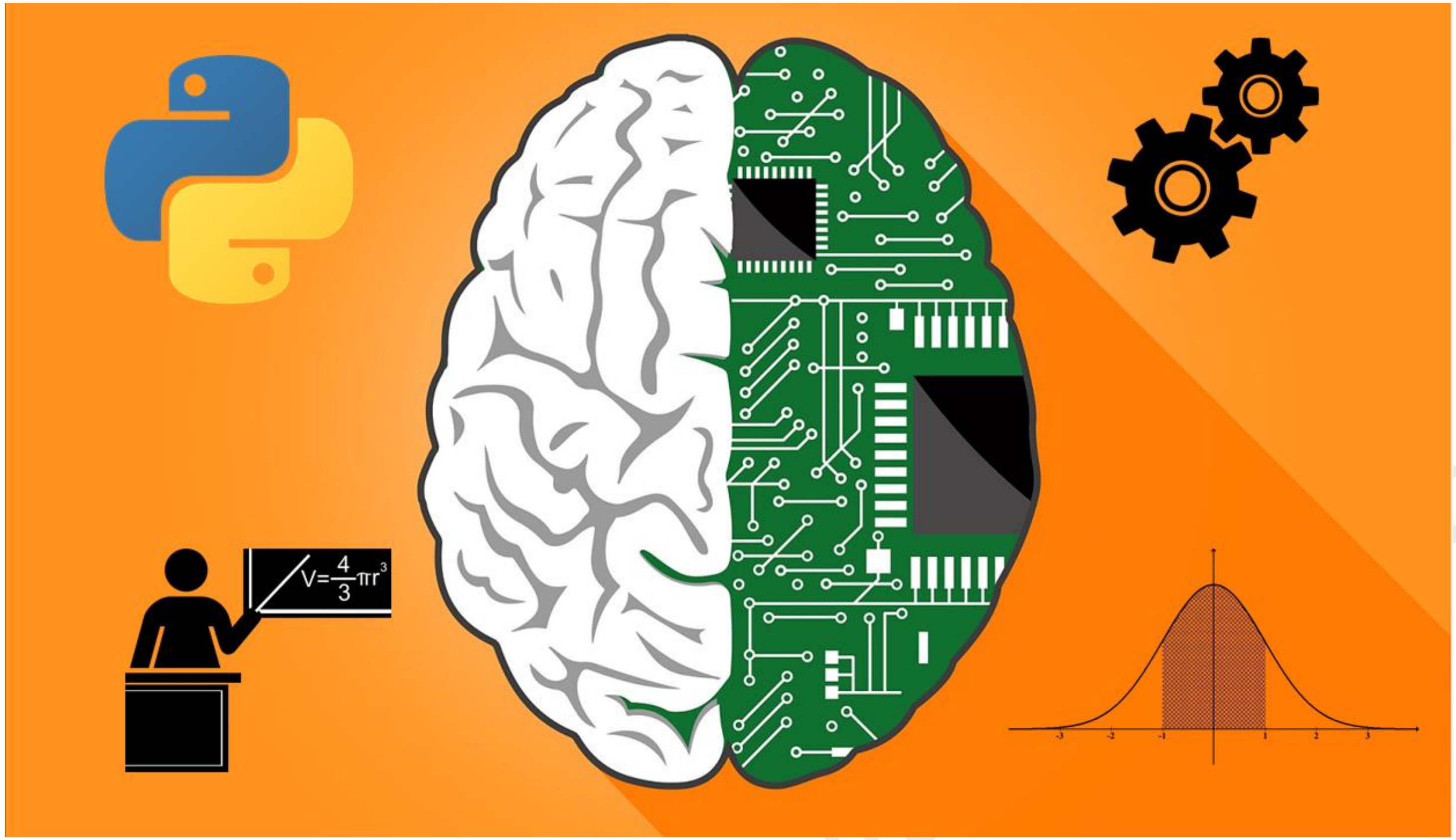


Complete Data Science and Machine Learning Using Python

By
Jitesh Khurkhuriya

Probability Distribution

What is a Distribution?

Distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.

-- Wikipedia

Distribution of Discrete Variables

Dice1 →

← Dice2

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$P(2) = 1/36$ $P(3) = 2/36$

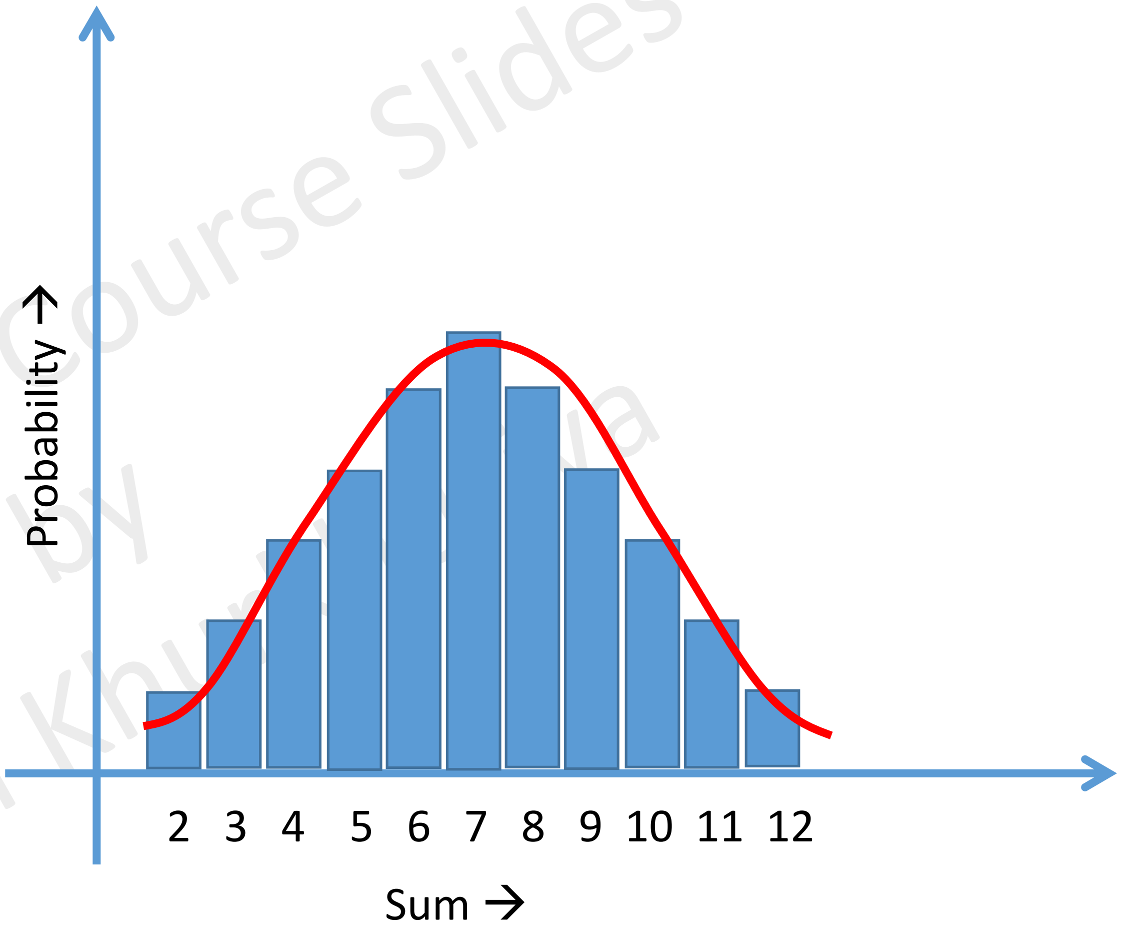
$P(4) = 3/36$ $P(5) = 4/36$

$P(6) = 5/36$ $P(7) = 6/36$

$P(8) = 5/36$ $P(9) = 4/36$

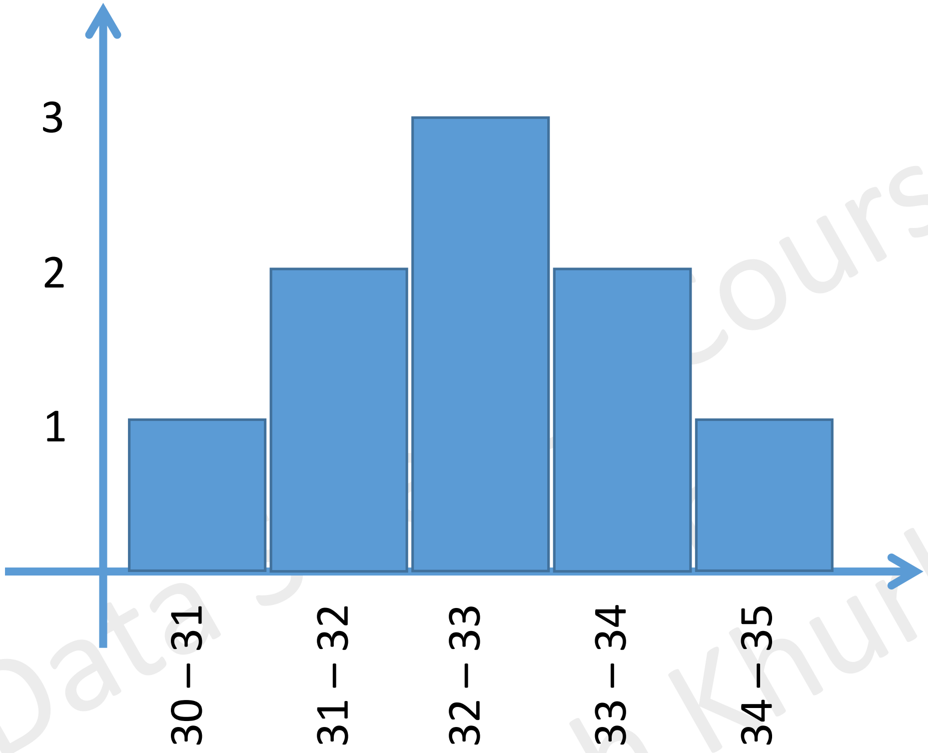
$P(10) = 3/36$ $P(11) = 2/36$

$P(12) = 1/36$



Distribution of Continuous Variable

Temperature
30.6
31.4
31.2
32.1
32.2
32.7
33.4
33.8
34.6



Frequency Distribution with Bins

What % of values are between

30-31 → 1/9 = 11.1%

31-32 → 2/9 = 22.2%

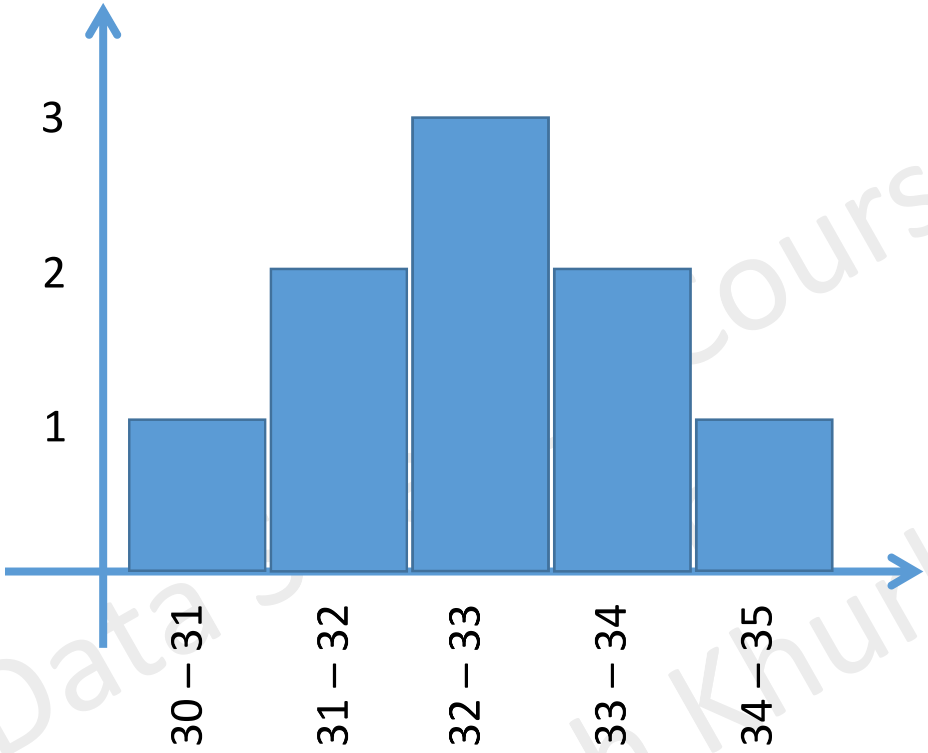
32-33 → 3/9 = 33.3%

33-34 → 2/9 = 22.2%

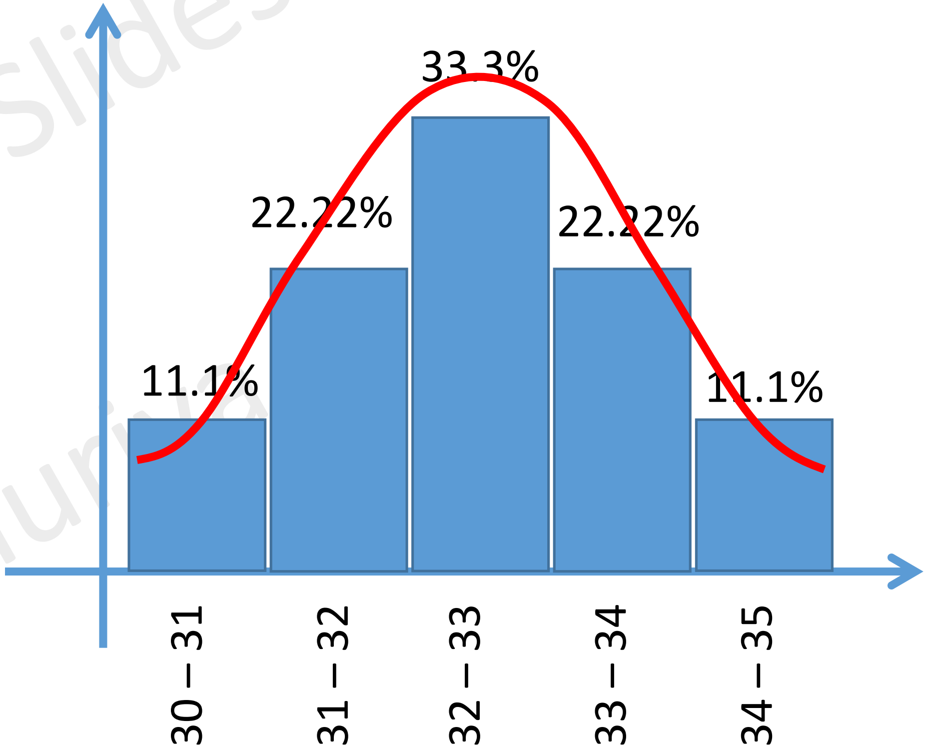
34-35 → 1/9 = 11.1%

Distribution of Continuous Variable

Temperature
30.6
31.4
31.2
32.1
32.2
32.7
33.4
33.8
34.6



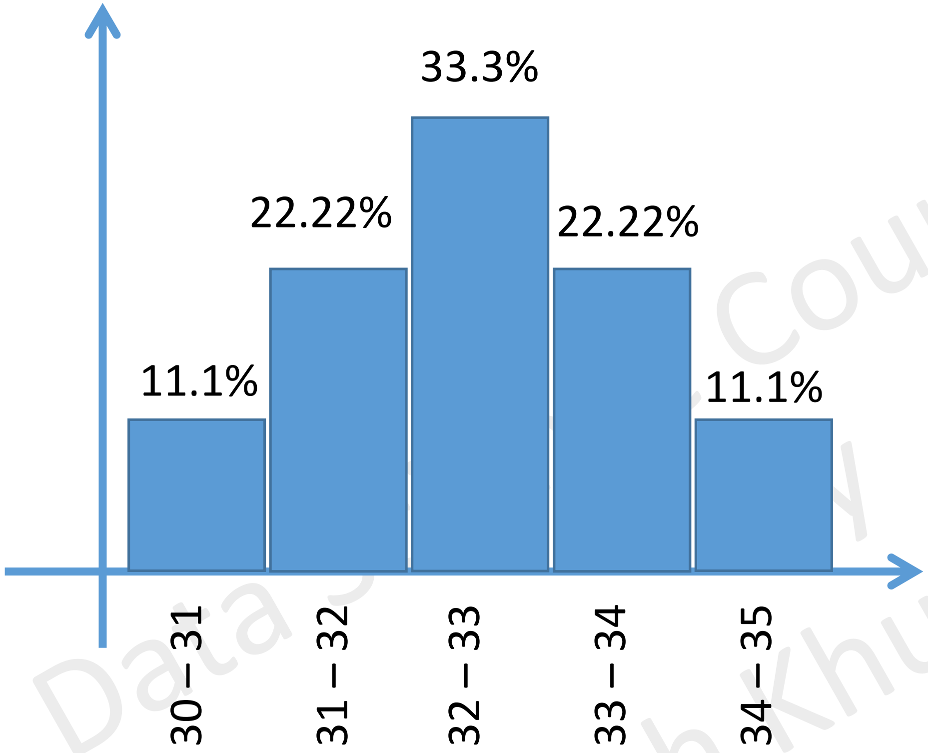
Frequency Distribution with Bins



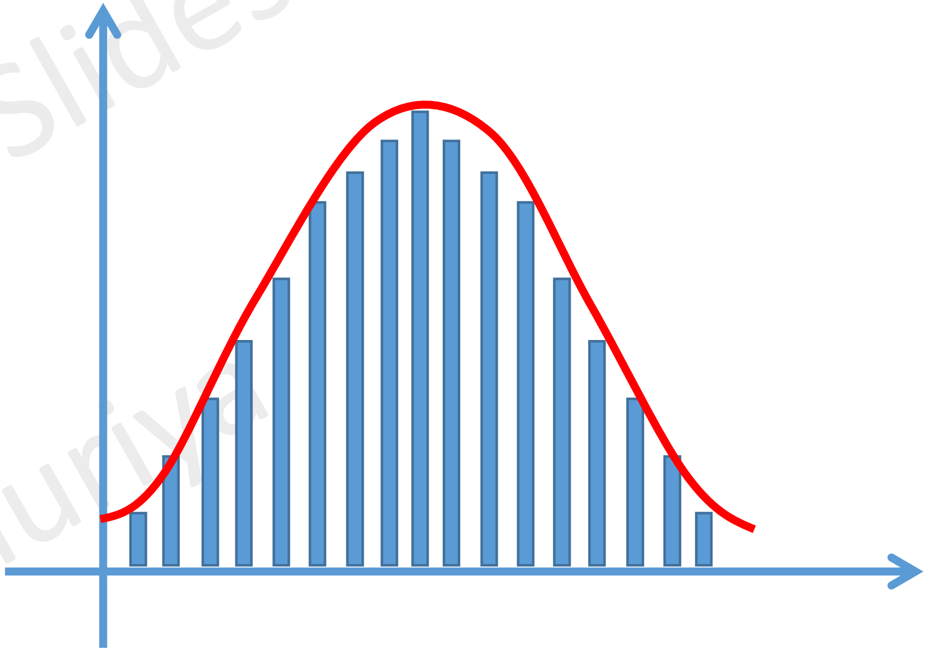
Probability of the Bins

Distribution of Continuous Variable

Temperature
30.6
31.4
31.2
32.1
32.2
32.7
33.4
33.8
34.6

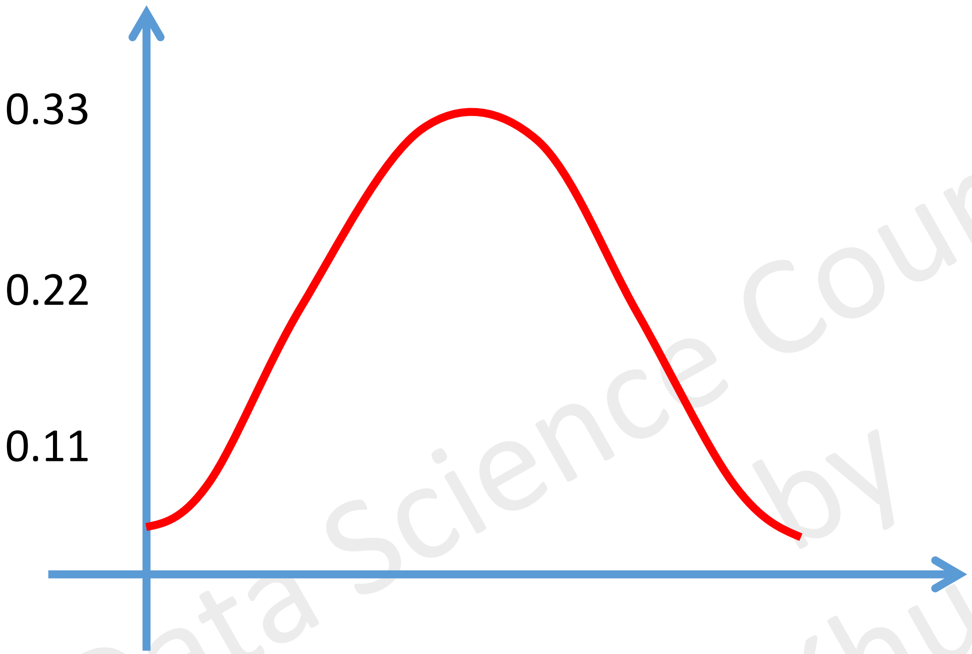


Probability of the Bins

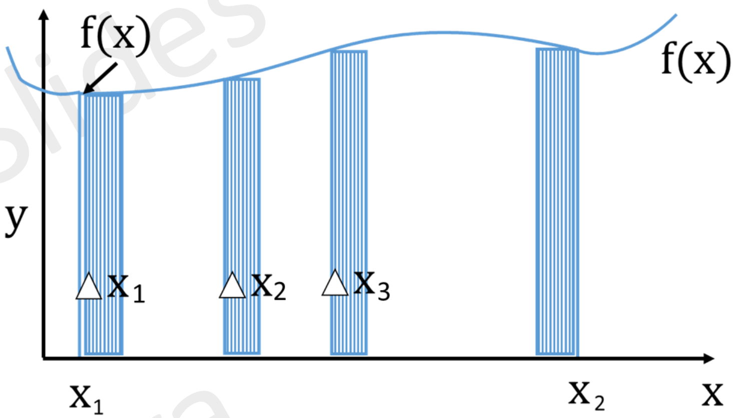


Probability Density

Distribution of Continuous Variable



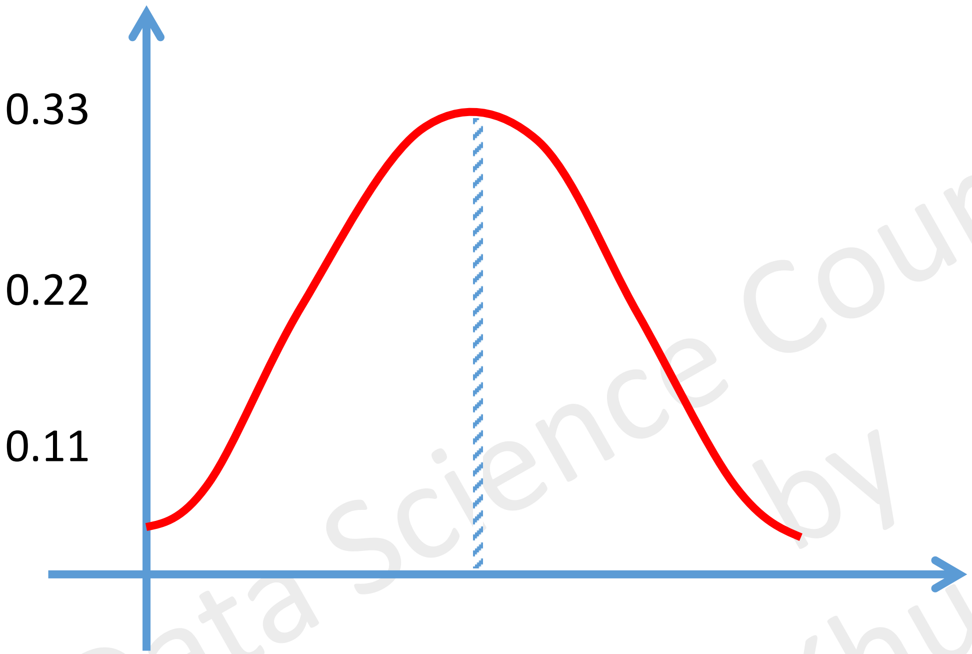
Probability Density Function



$$\text{Area} = \lim_{\triangle x \to 0} \sum_{i=1}^{n} f(x_i) * \triangle x_i$$

$$\boxed{\int_{x1}^{x2} f(x)dx}$$

Distribution of Continuous Variable



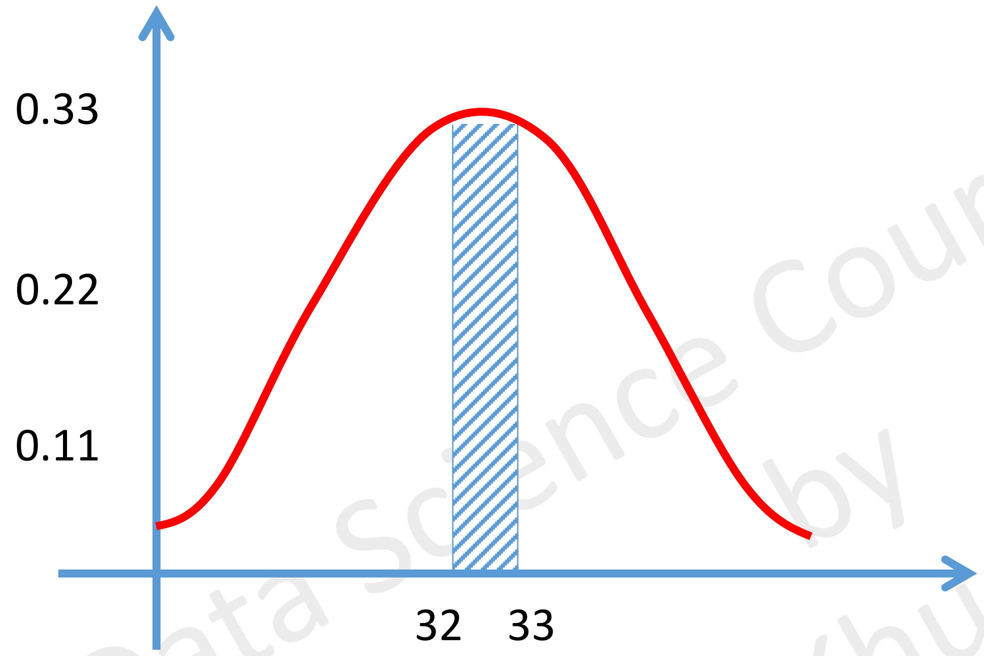
Probability Density Function

What is the probability that the temperature of the city will be exactly 32 degrees?

$$\text{Area} = \lim_{\triangle x \to 0} \sum_{i=1}^{n} f(x_i) * \triangle x_i$$

$$\boxed{\int_{x1}^{x2} f(x)dx}$$

Distribution of Continuous Variable



Probability Density Function

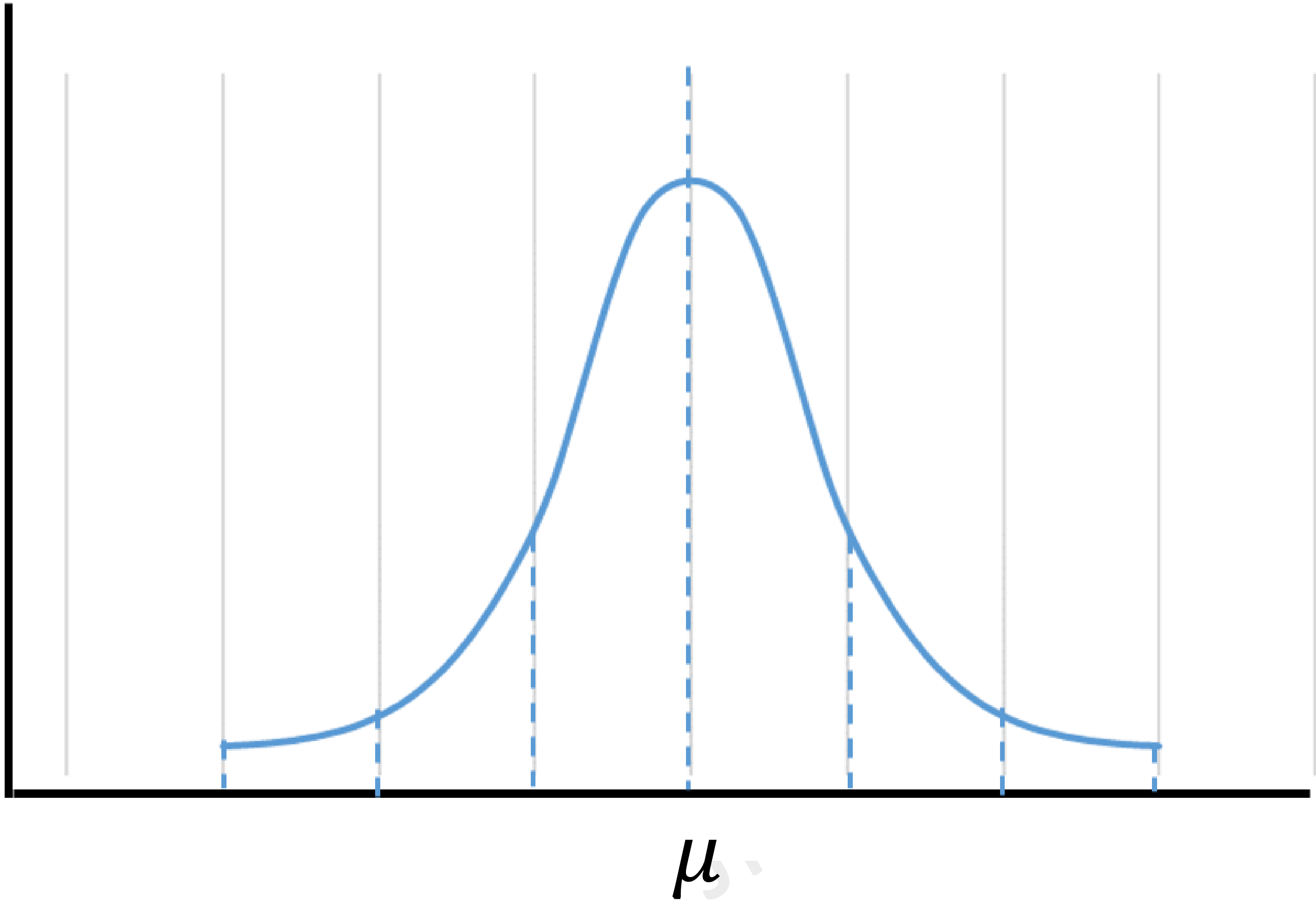
What is the probability that the temperature of the city will be between 32 and 33 degrees?

$$\text{Area} = \lim_{\triangle x \to 0} \sum_{i=1}^{n} f(x_i) * \triangle x_i$$

$$\boxed{\int_{x1}^{x2} f(x)dx}$$

Normal Distribution

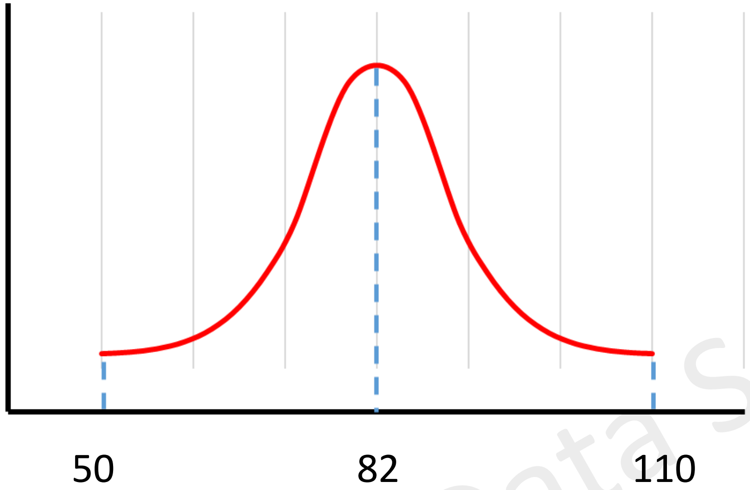
Normal Distribution – Bell Curve – Gaussian Distribution



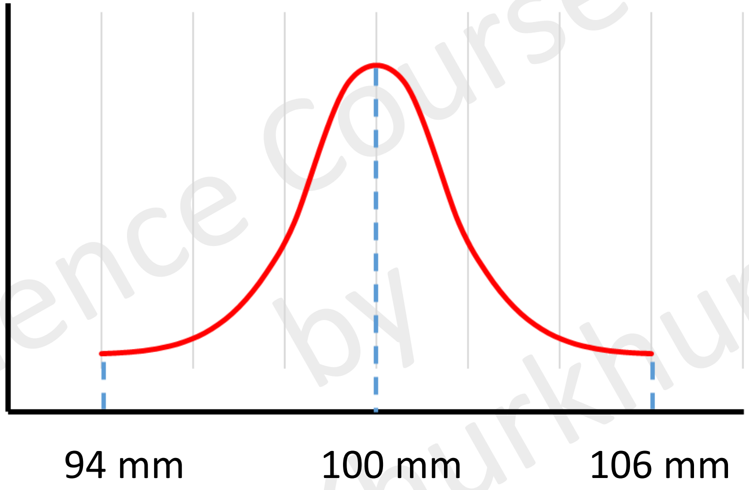
Carl Gauss

Examples of Normal Distribution

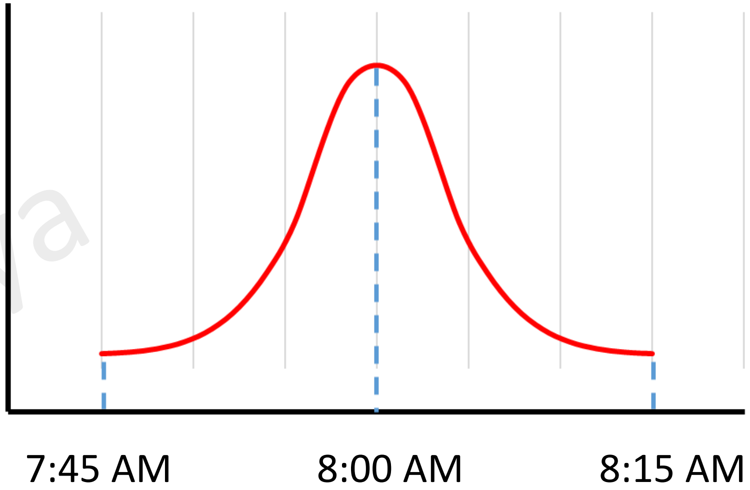
- Diastolic Blood Pressure



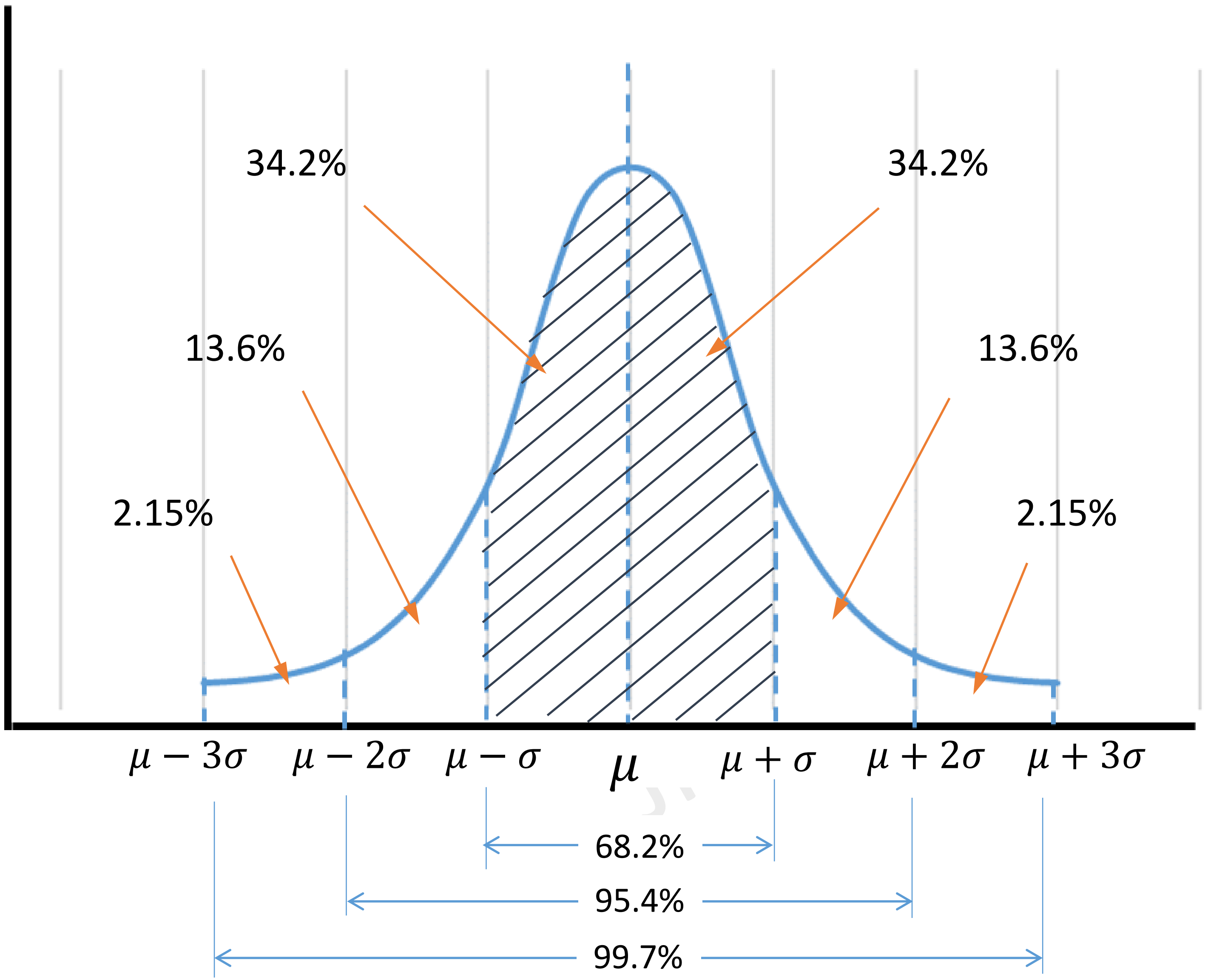
- Manufacturing



- Arrival Time at office

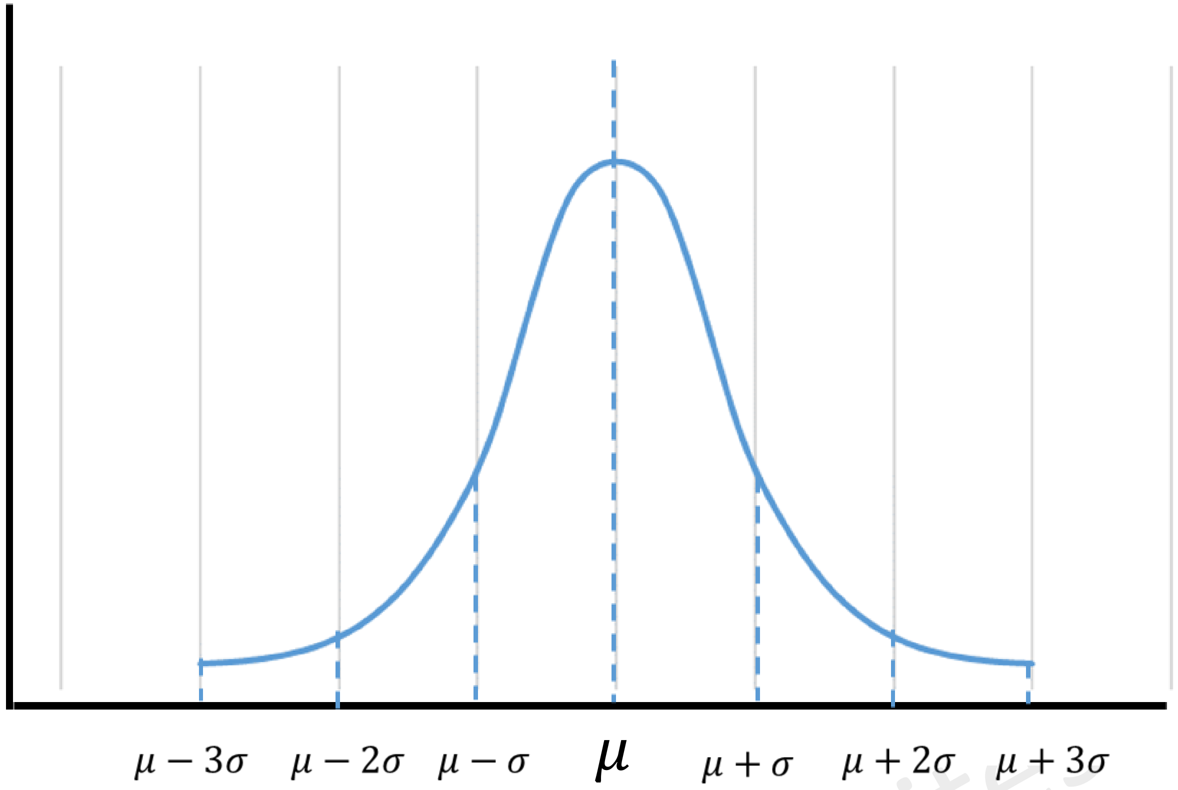


Normal Distribution – Bell Curve – Gaussian Distribution



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

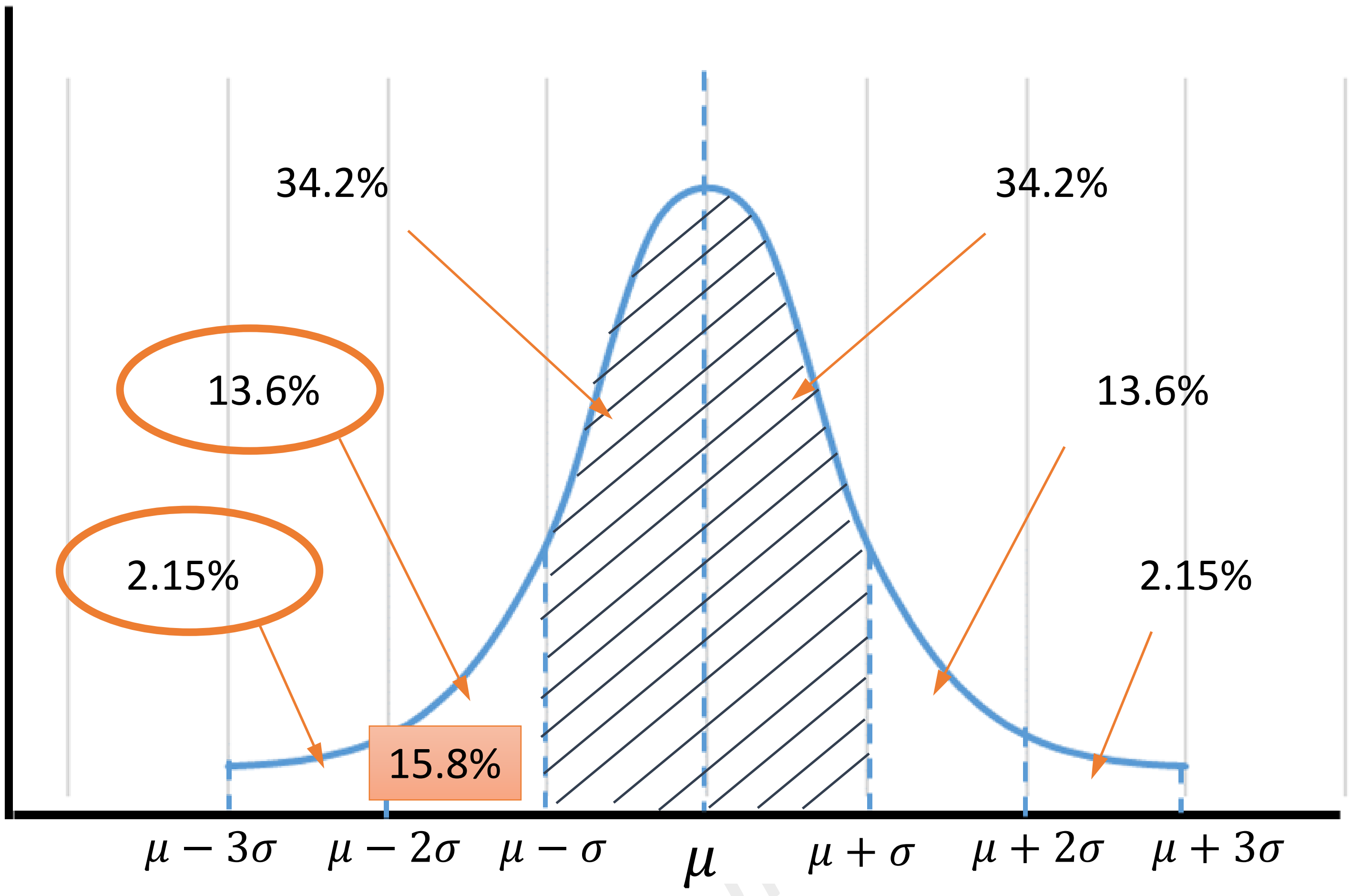
Characteristics of Normal Distribution



- Mean defines the centre of the graph
- Mean = Median = Mode
- Standard Deviation defines the width of the graph
- Entire distribution can be specified using mean and variance
- The total area under the curve is 1
- Probability at a given point is zero
- 68.2% of the area under the curve is within 1 σ of the mean
- 95.4% of the area under the curve is within 2 σ of the mean
- 99.7% of the area under the curve is within 3 σ of the mean

Standard Normal Distribution

Z-Score


34.2%
34.2%
13.6%
13.6%
2.15%
2.15%
15.8%
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Z-Score Table

- Standard Normal Table
- Provides Cumulative Distribution Function Values

	0.00	0.01	0.02	0.03
1.00	0.841345	0.843752	0.846136	0.848495
1.10	0.864334	0.866500	0.868643	0.870762
1.20	0.884930	0.886861	0.888768	0.890651
1.30	0.903200	0.904902	0.906582	0.908241
1.40	0.919243	0.920730	0.922196	0.923641
1.50	0.933193	0.934478	0.935745	0.936992

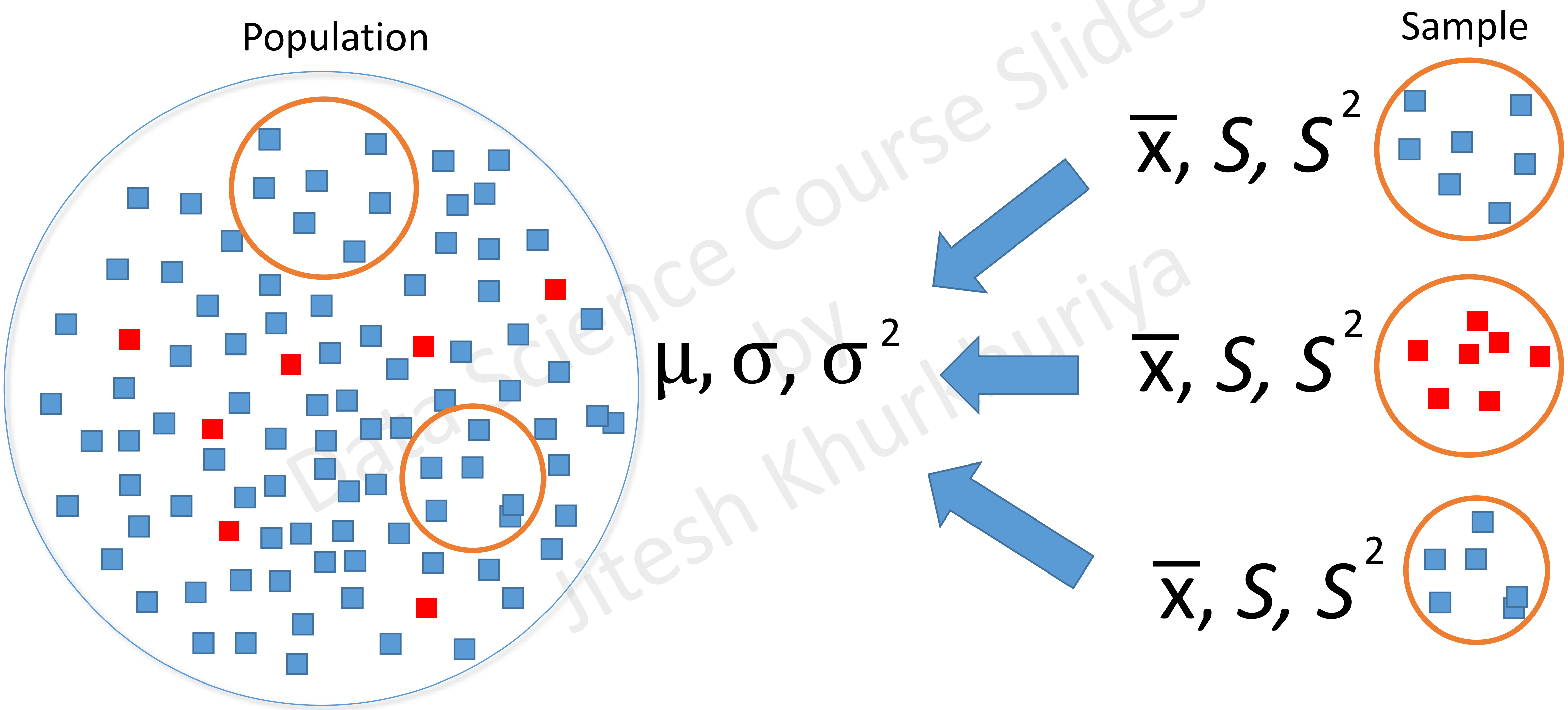
Importance of Standard Normal Distribution and Z-Score

- Standardises the readings or scores
- Calculate the probability within the normal distribution
- Comparison of two records from different normal distribution at two different scale

Experience in years	Salary
1	$ 4,500
4	$ 7,200
4	$ 6,500
6	$ 8,500
7	$ 8,900

Sampling Distribution

Population and Sample



Population and Sample

Yrs
3
5
6
7
7
8
9
9
10
10
7.4

Sample 1	
5	7.33
7	
10	

Sample 2	
3	7.33
9	
10	

Sample 3	
6	7
7	
8	

Sample 4	
8	8.67
9	
9	

Sample 5	
6	7.67
7	
10	

Sample 6	
5	8.33
10	
10	

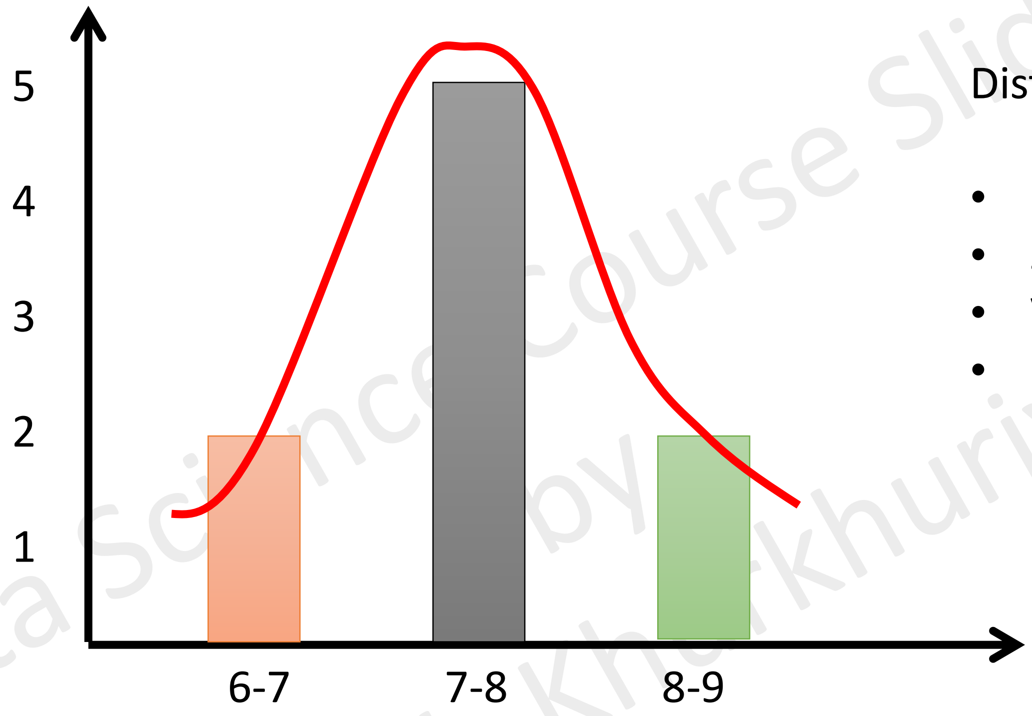
Sample 7	
3	7.66
10	
10	

Sample 8	
3	6
7	
8	

Sample 9	
5	6.33
6	
8	

Sampling Distribution

Sample Mean
7.33
7.33
7
8.67
7.67
8.33
7.66
6
6.33



Distribution of the Statistic of the Sample,

- Mean
- Standard Deviation
- Variance
- Range

Central Limit Theorem

Central Limit Theorem

When independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed.

-- Wikipedia

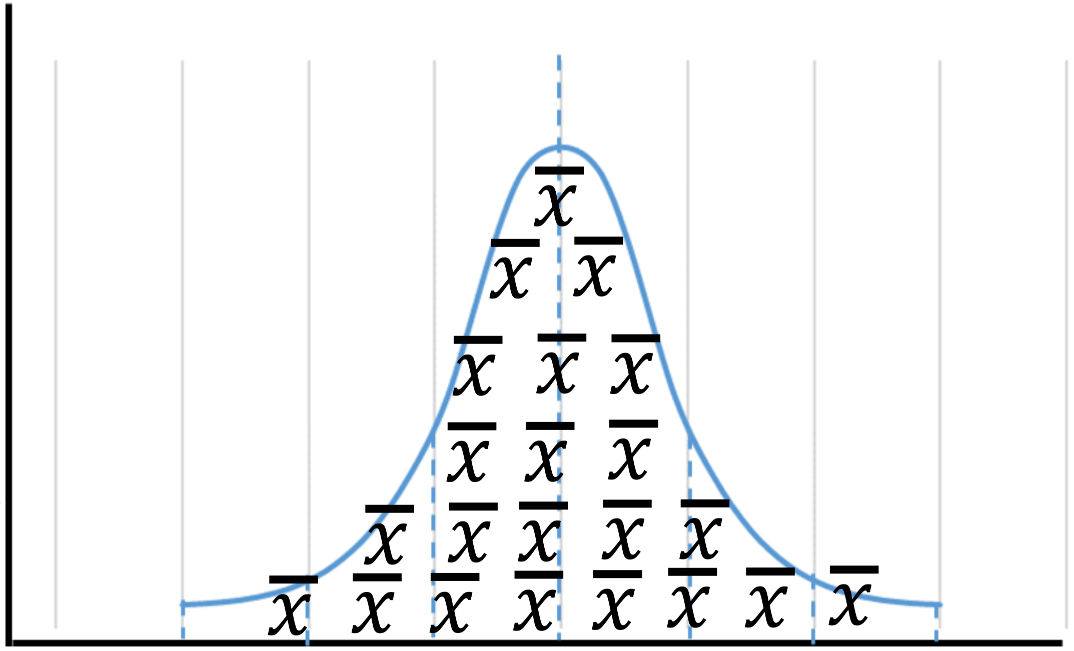
Central Limit Theorem

$\bar{x}$ $ 110,000
$\bar{x}$ $ 105,000
$\bar{x}$ $ 127,000
$\bar{x}$ $ 117,000
$\bar{x}$ $ 108,000
$\bar{x}$ $ 105,000
$\bar{x}$ $ 120.000
$\bar{x}$ $ 125,000
$\bar{x}$ $ 130,000
$\bar{x}$ $ 105,000

$ 140,000

$ 110,000

$ 121,000



$ 142,000

$ 128,000

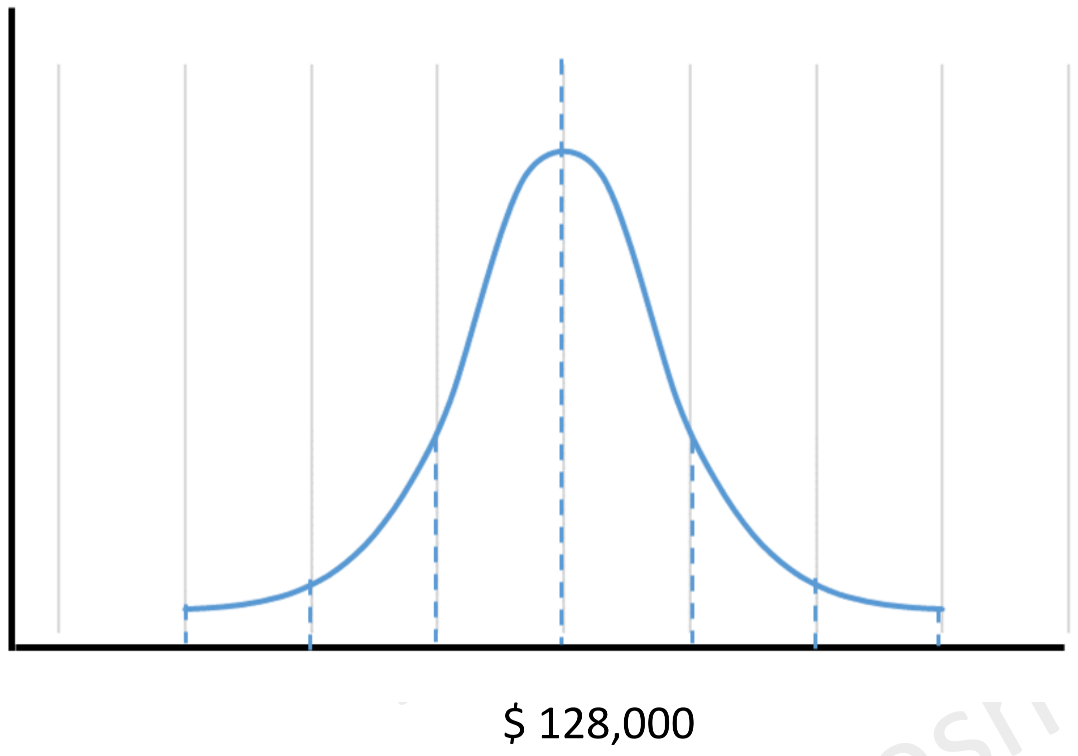
$ 111,000

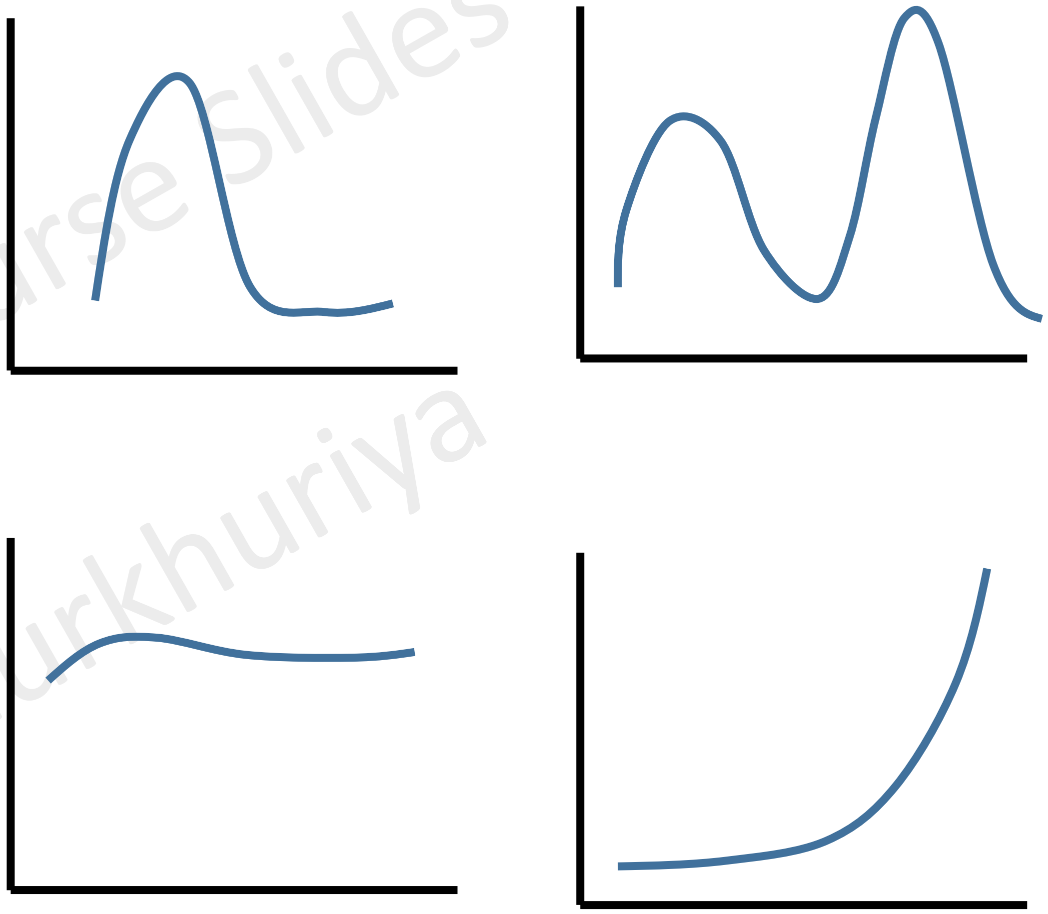
$\bar{x}$ $ 115,000
$\bar{x}$ $ 121,000
$\bar{x}$ $ 107,000
$\bar{x}$ $ 119,200
$\bar{x}$ $ 132,000
$\bar{x}$ $ 131,000
$\bar{x}$ $ 135.000
$\bar{x}$ $ 130,000
$\bar{x}$ $ 121,000
$\bar{x}$ $ 115,000

Importance of Sampling

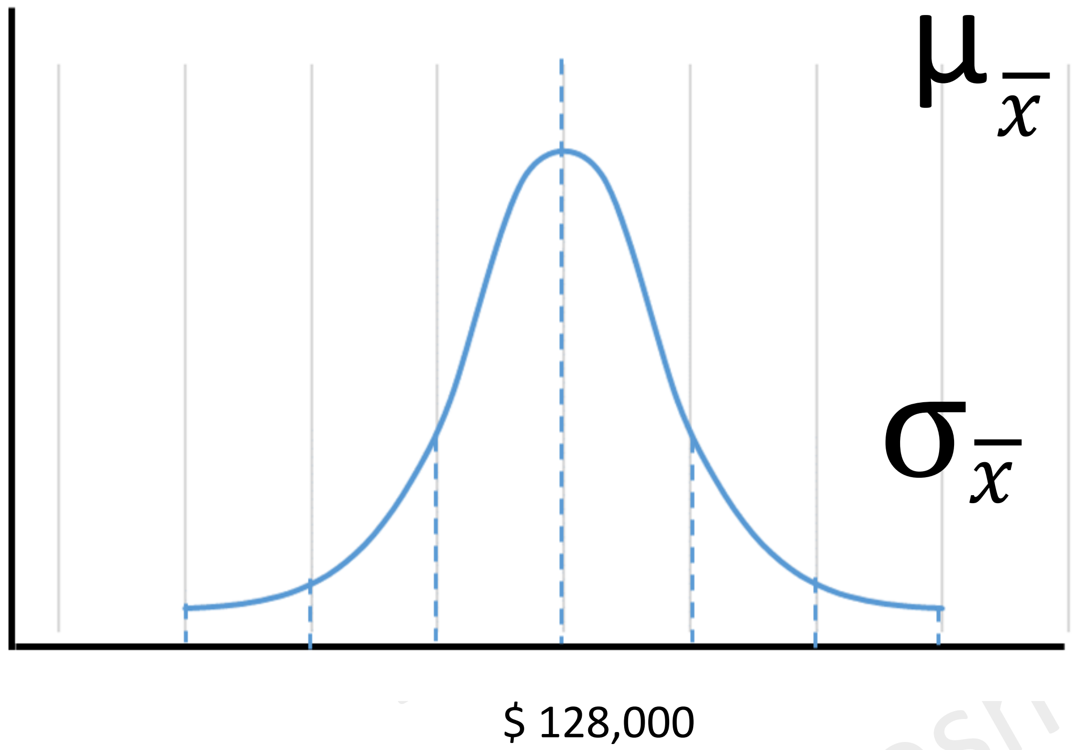
- Inferences about the population using a small subset
- Efficient in terms of time and money
- Flexible to approximate many sums and integrals in Machine Learning
- Sum or integral can be intractable/impossible or hard to define

Importance of Central Limit Theorem





Importance of Central Limit Theorem



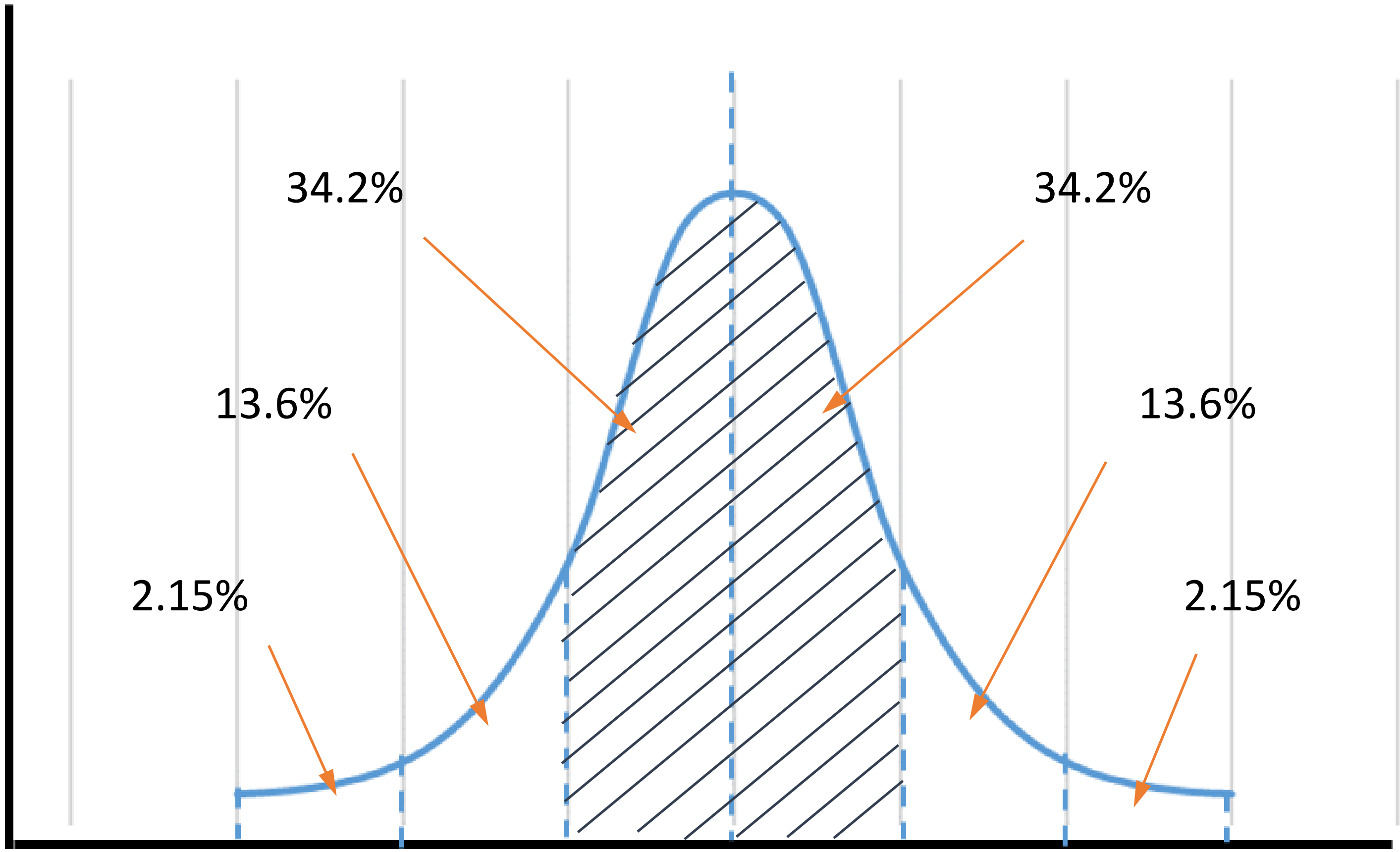
$$\mu_{\bar{x}} = \mu$$

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

- Valid Sample → Population Inferences
- Population Information → Valid Sample
- Population and Sample → Sample Verification
- Multiple Valid Samples → Infer the origin

Confidence Interval

Normal and Sampling Distribution



Sampling distribution of Mean

Point Estimate

A single value which is used to serve as a "best guess" or "best estimate" of an unknown population parameter.

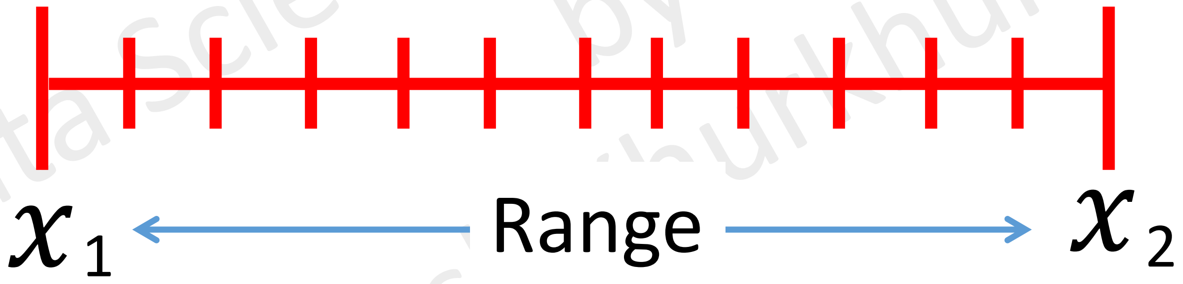
-- Wikipedia

$$\bar{x} \sim \mu$$

Interval Estimate

$$\bar{x} \quad \sim \quad \mu$$

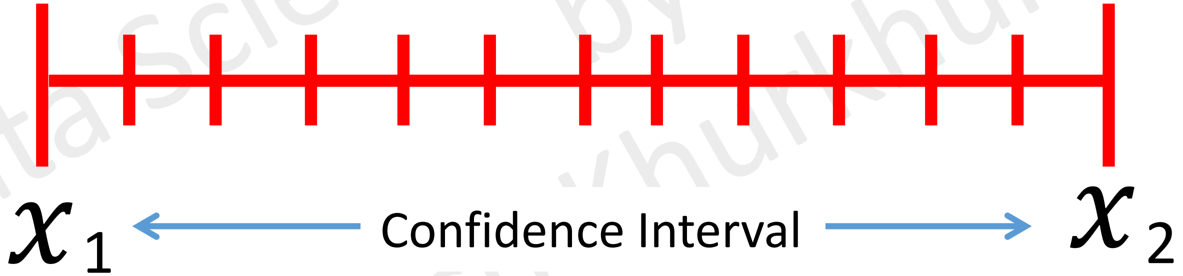
$$84 \qquad\qquad ?$$



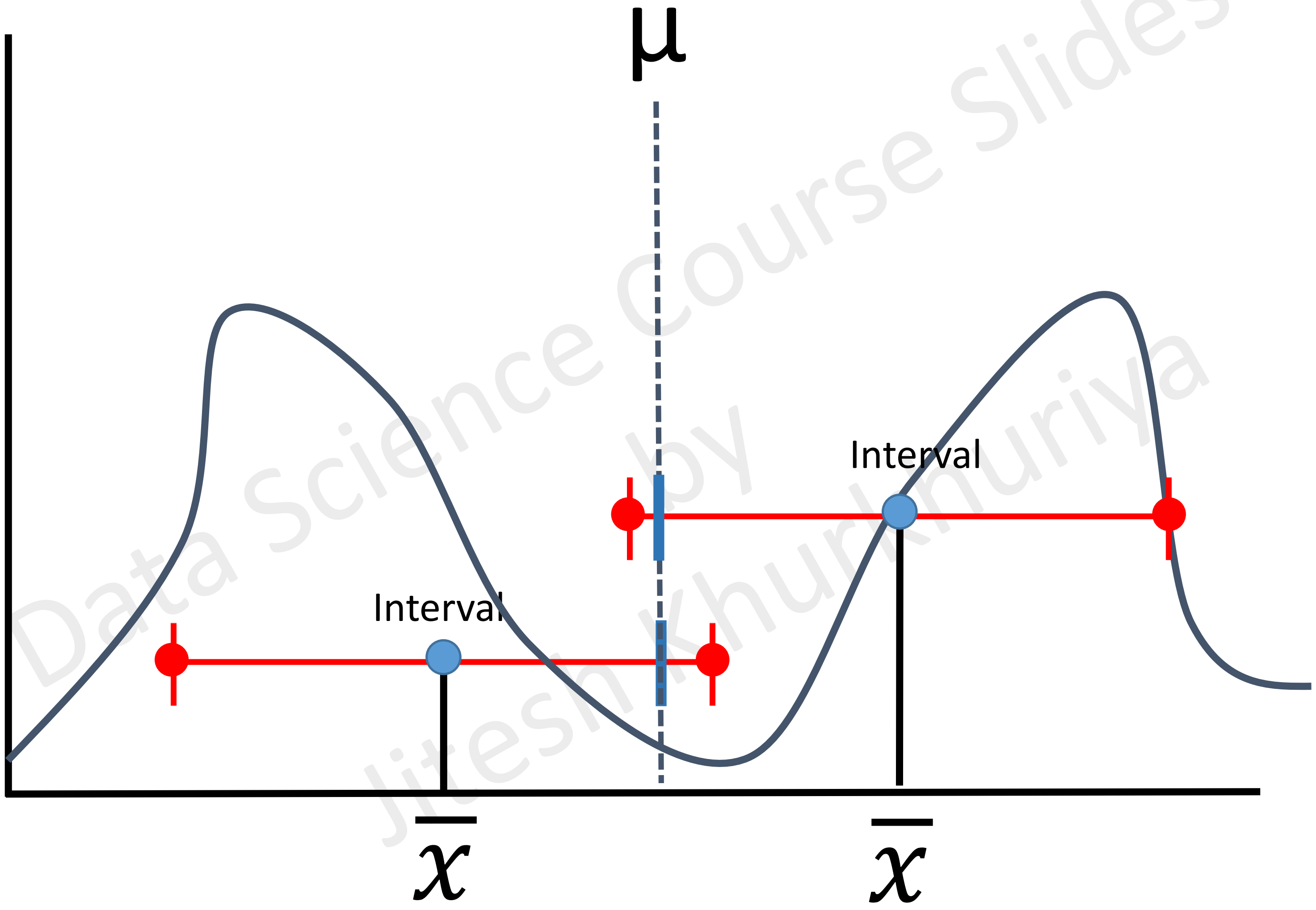
Interval Estimate

$\bar{x} \sim \mu$

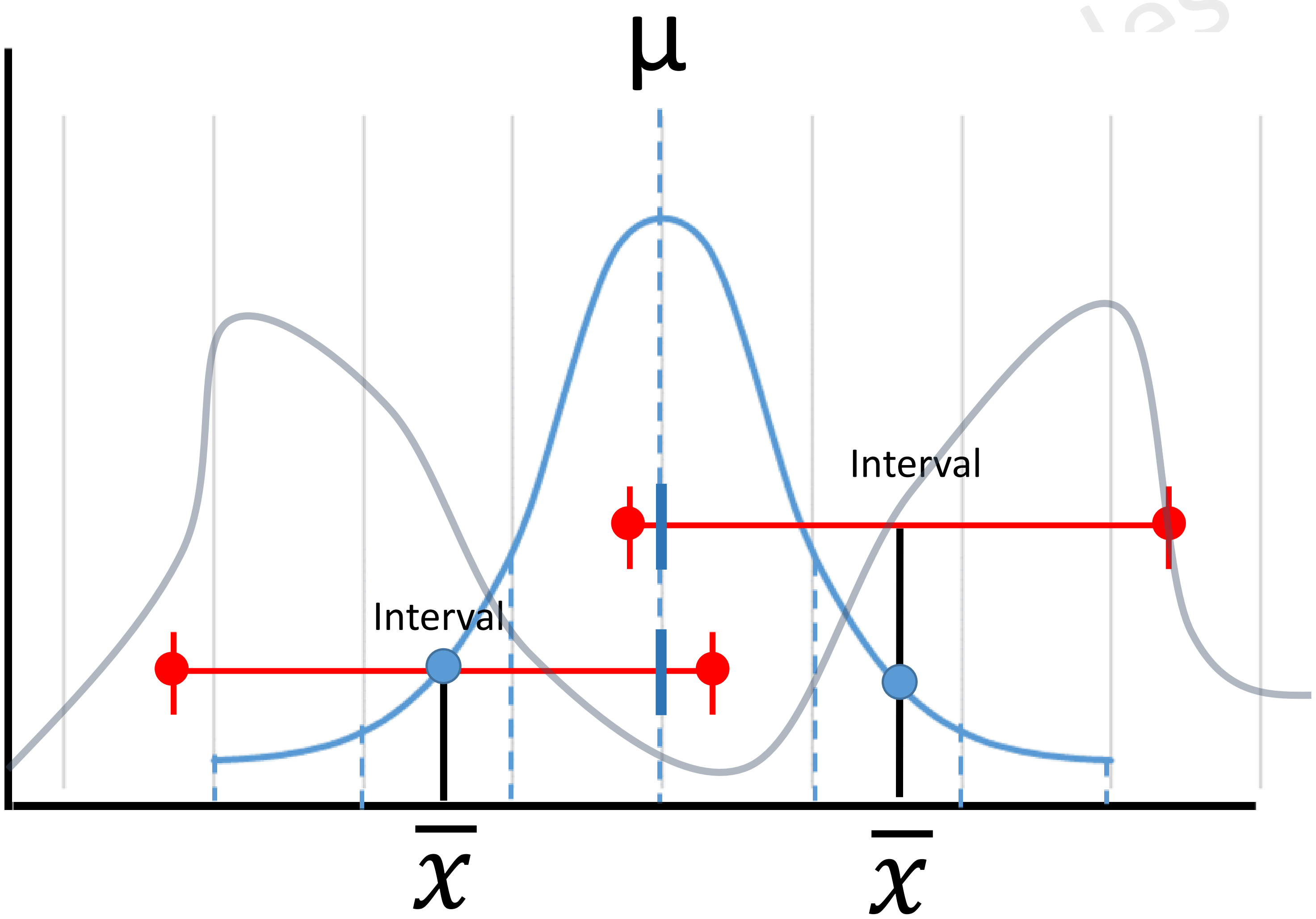
84 ?



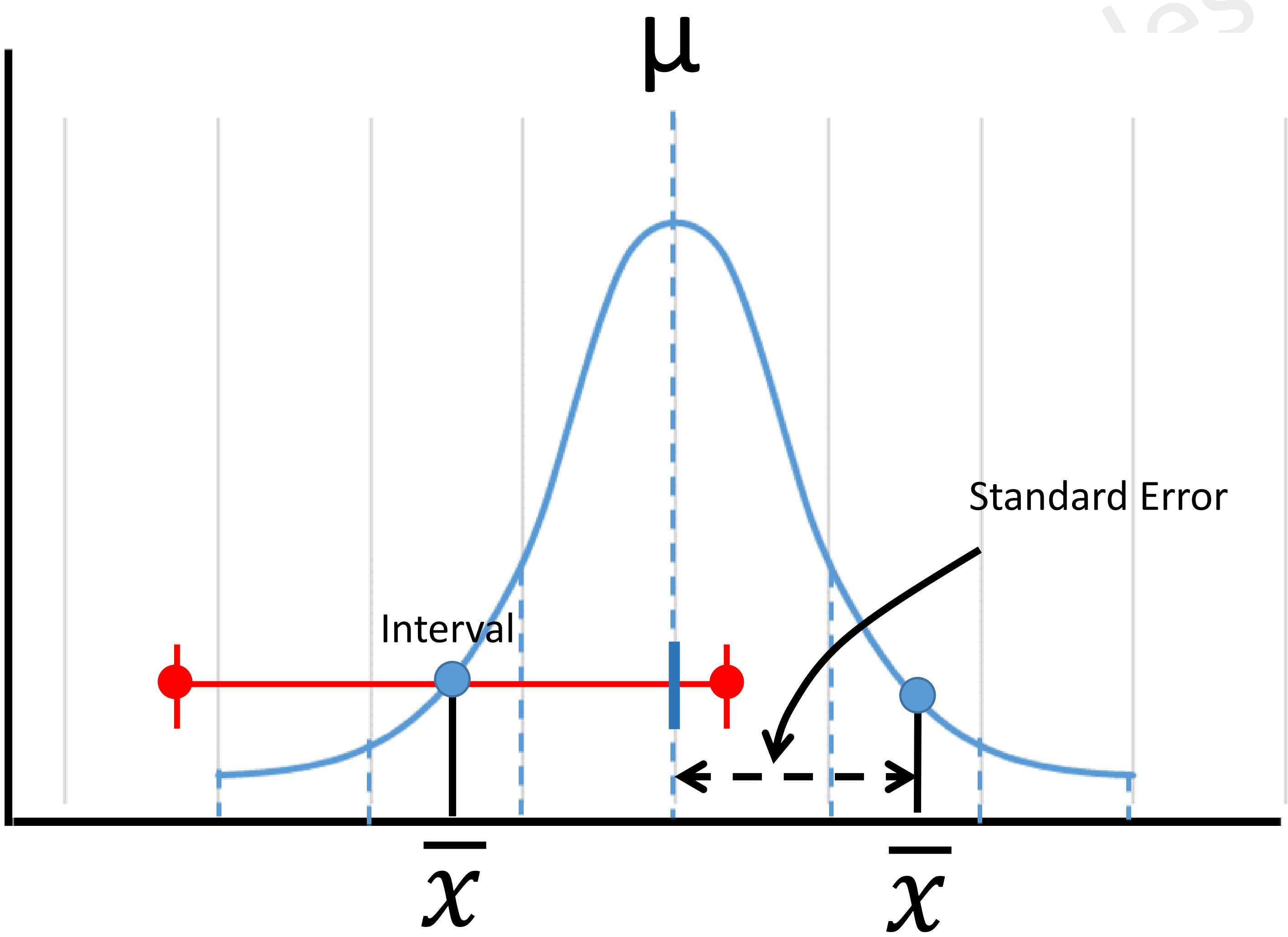
Interval Estimate



Interval Estimate

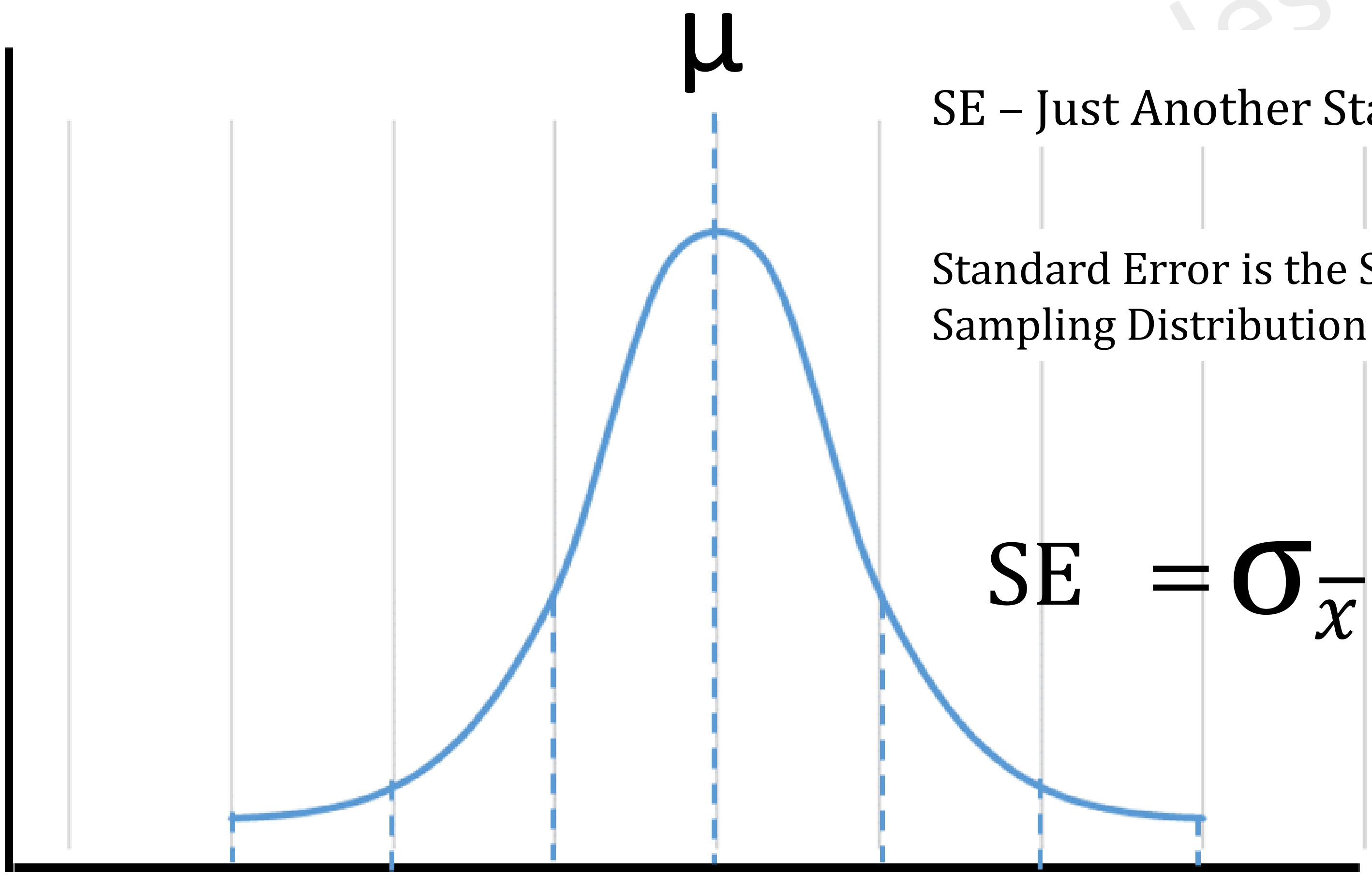


Interval Estimate



How Far Sample Mean is from the population Mean?

Standard Error



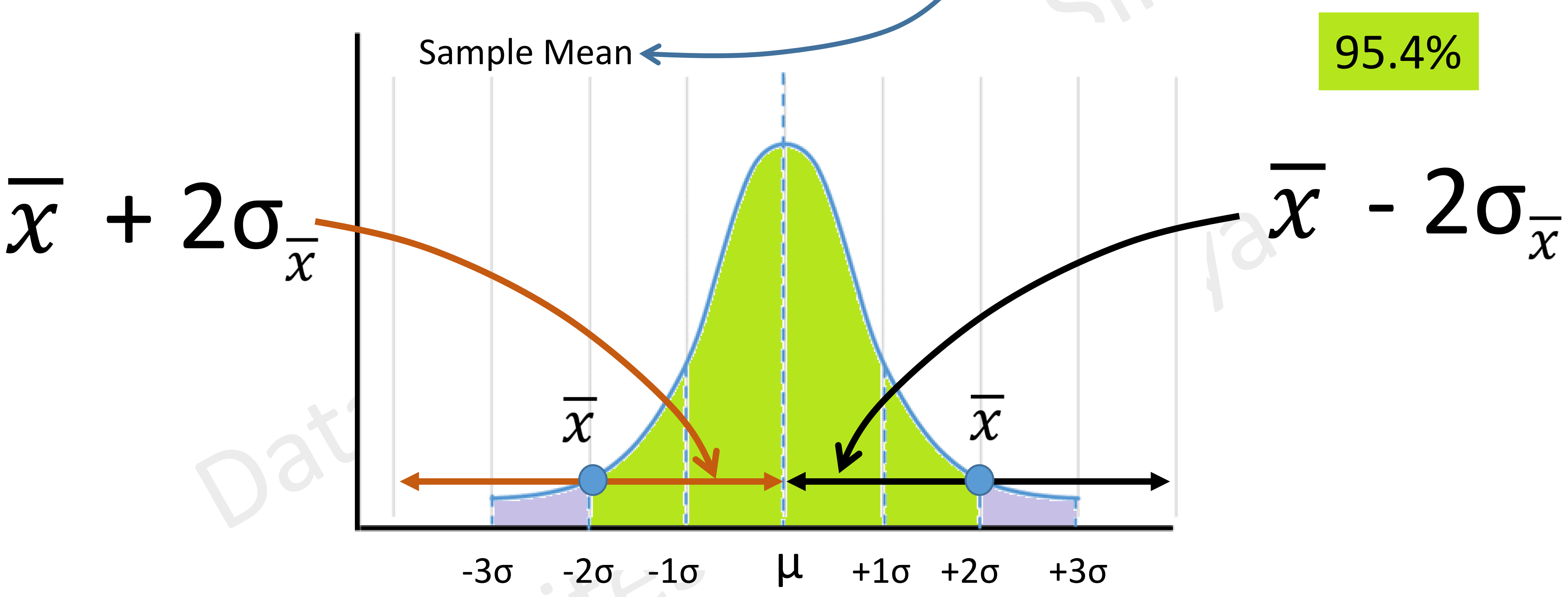
SE – Just Another Statistical Jargon

Standard Error is the Standard Deviation of the Sampling Distribution of the Mean of the Samples

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

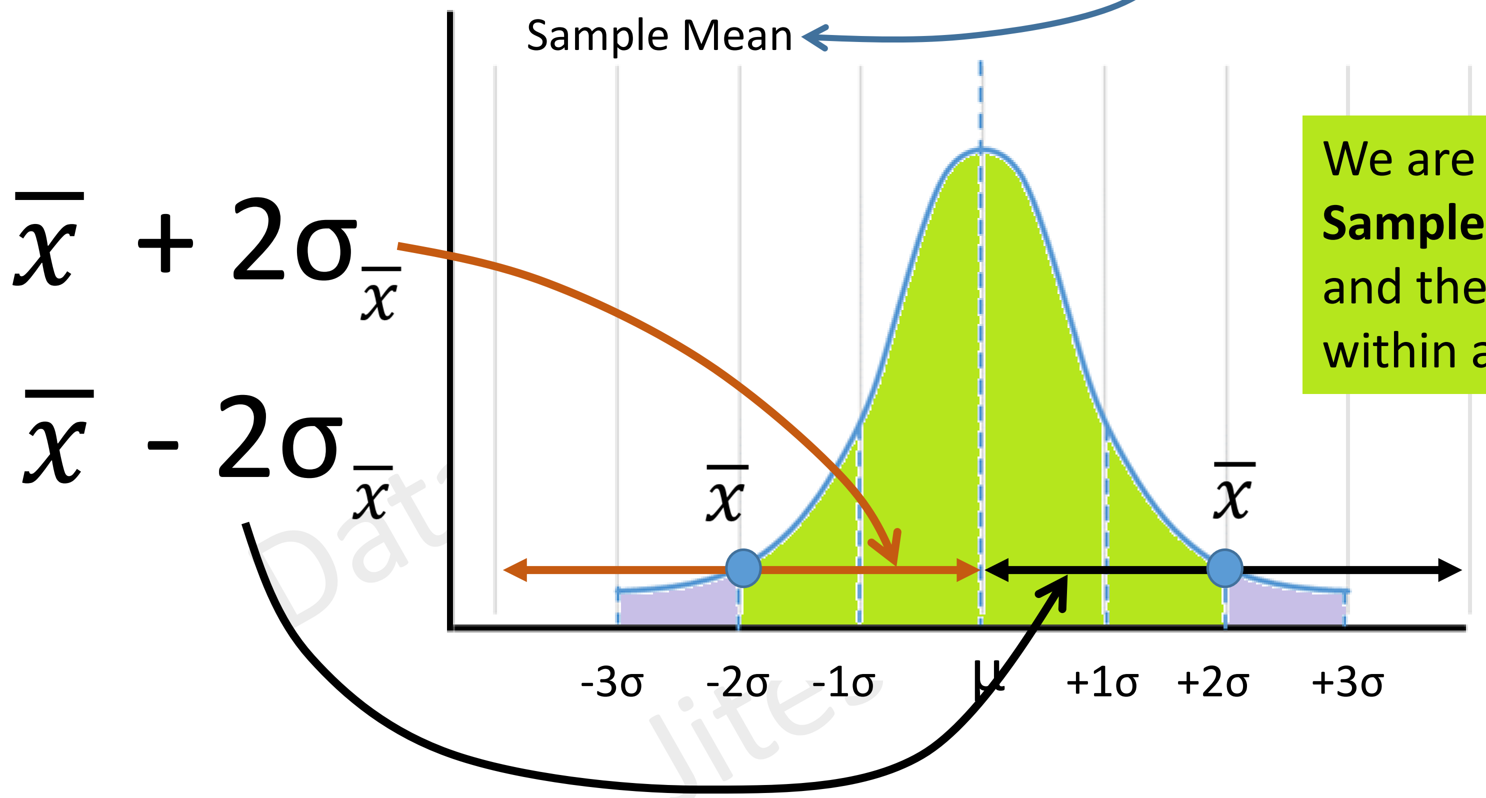
Reliability Factor

A Number based on the Sampling Distribution of the point estimate and the degree of confidence.



Reliability Factor

A Number based on the Sampling Distribution of the point estimate and the degree of confidence.



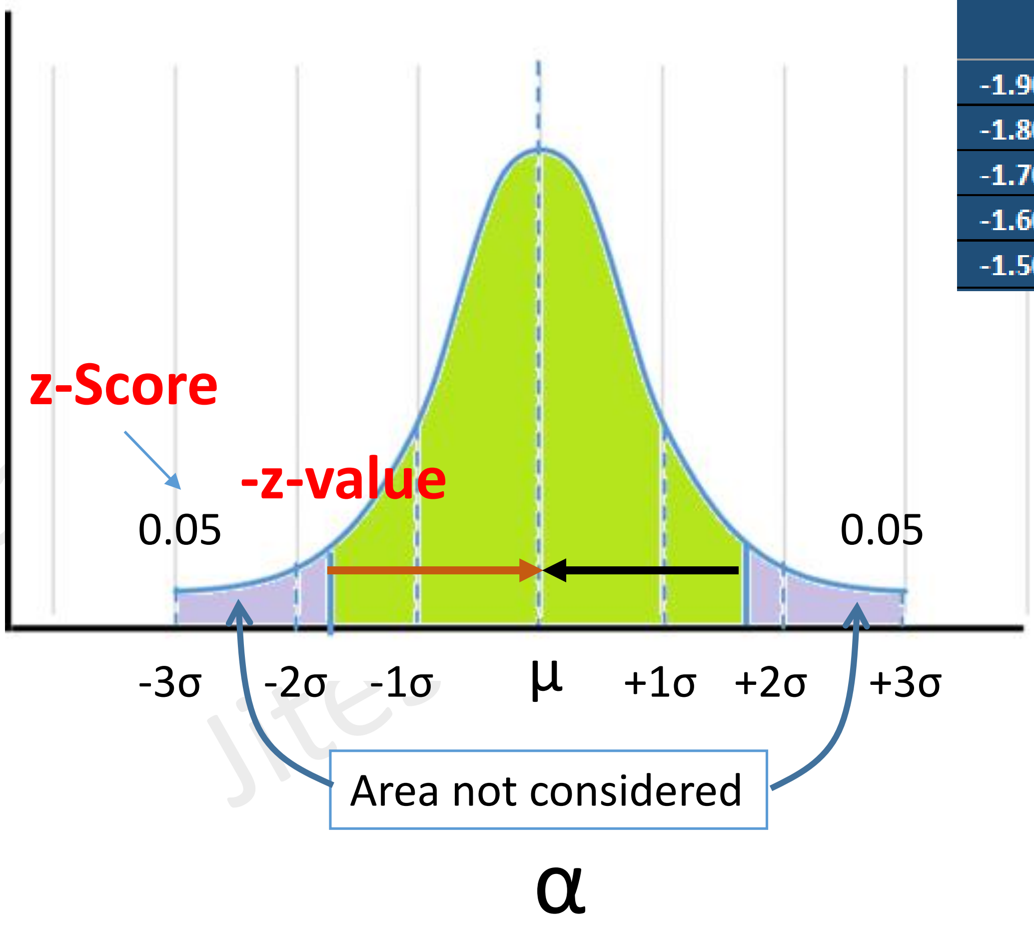
We are 95.4% confident that the **Sample Mean** will be within $\pm$ 2σ and the **Population Mean** will be within an **interval** of $\pm$ 2σ.

Reliability Factor

A Number based on the Sampling Distribution of the point estimate and the degree of confidence.

α = 1 – Confidence Level

Confidence Level = 1 – α



	0.04	0.05	0.06	0.07
-1.90	0.031443	0.032157	0.032884	0.033625
-1.80	0.039204	0.040059	0.040930	0.041815
-1.70	0.048457	0.049471	0.050503	0.051551
-1.60	0.059380	0.060571	0.061780	0.063008
-1.50	0.072145	0.073529	0.074934	0.076359

$$Z_{\alpha/2} = -1.7 + 0.05 = -1.65$$

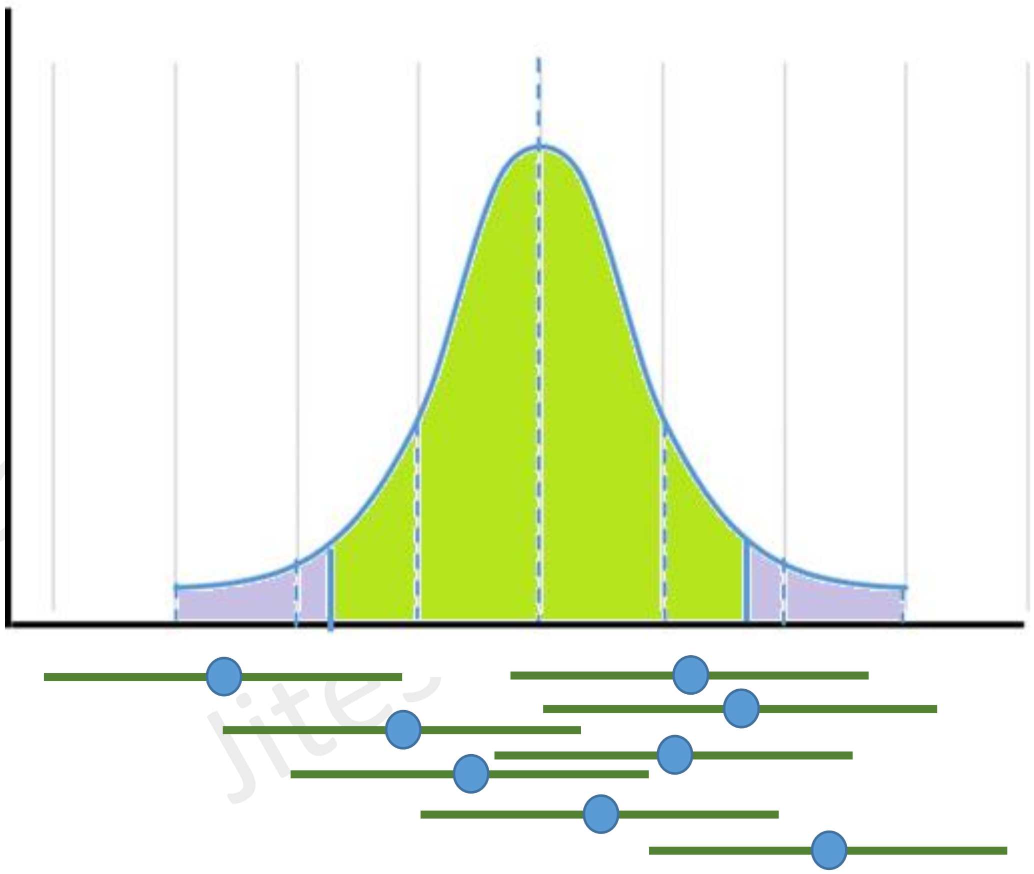
Reliability Factor

A Number based on the Sampling Distribution of the point estimate and the degree of confidence.

90% CI does not mean there is 90% probability that population mean will be in the given interval.

90% intervals will have population mean within the interval limits.

9 out of 10 random intervals will have population mean within the range.



If we draw a sample and calculate its mean, we are 90% confident that the population mean will be within an interval of,

$$\bar{x} \pm 1.65 * \sigma_{\bar{x}}$$

Confidence Interval

Confidence Interval $=$ Point Estimate $\pm$ Reliability Factor $*$ Standard Error

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

Lower Endpoint

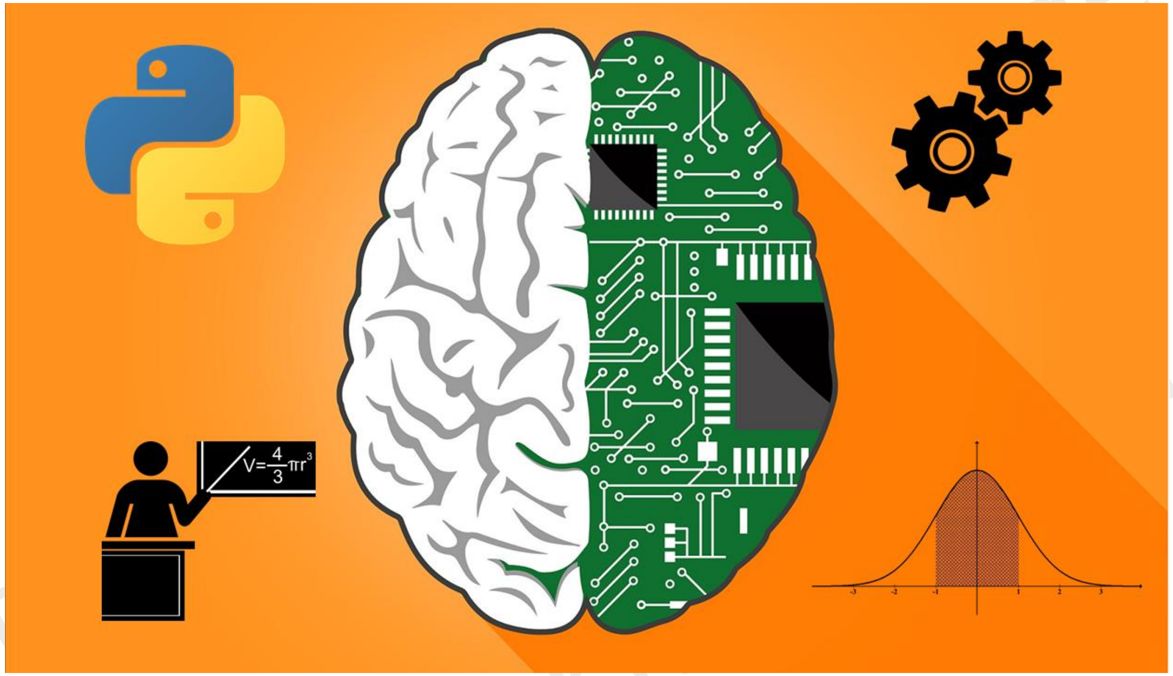
$$\bar{x} - Z_{\alpha/2} * \sigma_{\bar{x}}$$

Upper Endpoint

$$\bar{x} + Z_{\alpha/2} * \sigma_{\bar{x}}$$

α = 1 – Confidence Level

Complete Data Science and Machine Learning Using Python



Thank You!