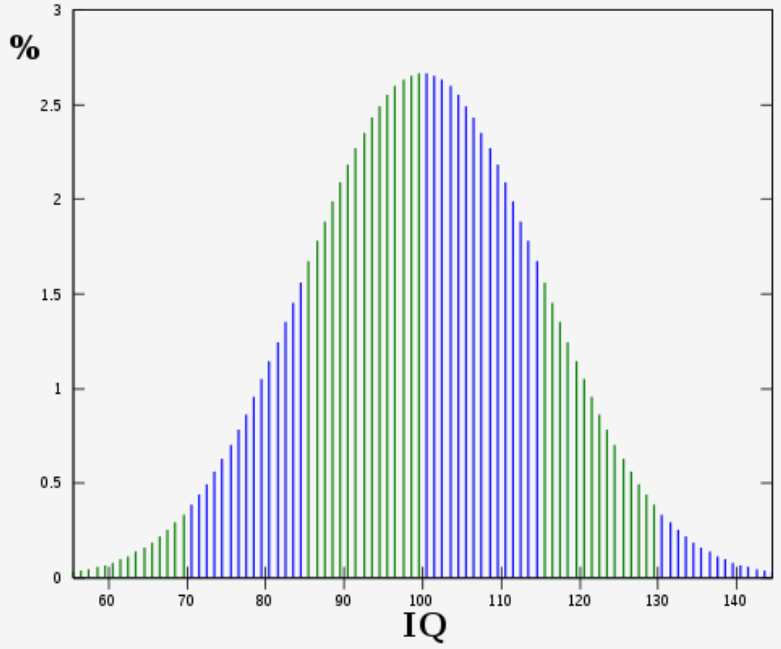
# *DATA SCIENCE*

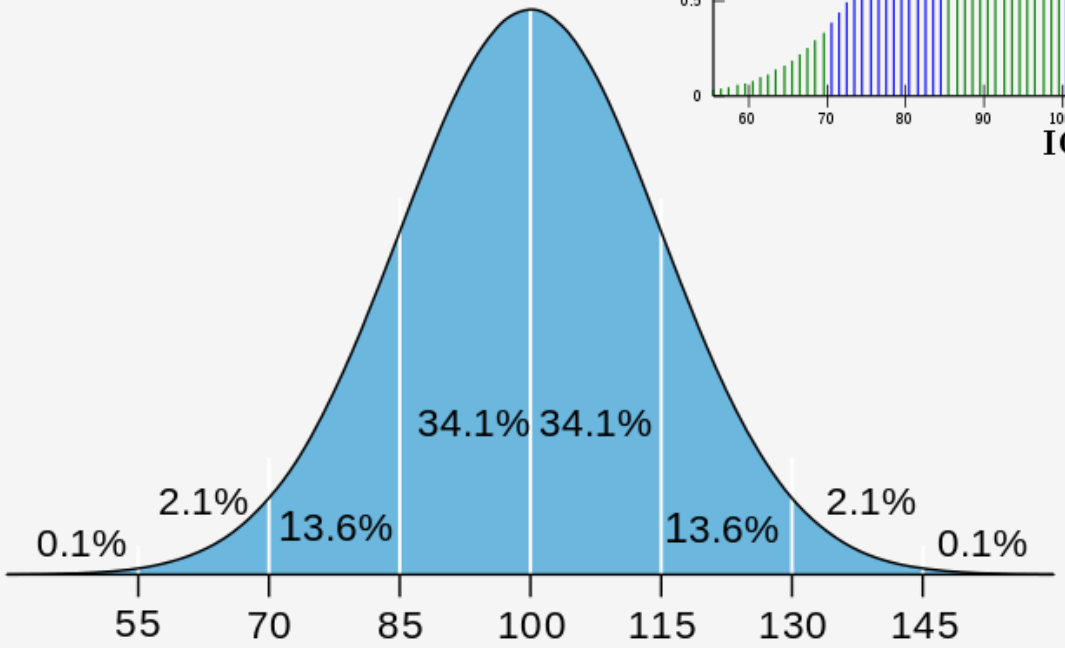
## *CHAPTER 1*

*BY AHMAD OBAIDAT*

# STATISTICS

# *Section 1*
# *Distribution*



# WHAT WE WILL LEARN

1. Continuous vs Discrete.
2. What is a Distribution?
3. Standard Deviation.
4. Normal Distribution.
5. Skewness
6. Mean, Median, Mode.

# CONTINUOUS VS DISCRETE

| Customer ID | Name | Surname | Gender | Age | Age Group | Height | Region | Job Classification | Tenure Months | Balance | Spend On Groceries |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 200000262 | Zoe | Clarkson | Female | 59 | 50 | 62 | Scotland | Other | 24 | 23550.89 | 70.77 |
| 200001214 | Carolyn | McDonald | Female | 58 | 50 | 61.2 | Scotland | Other | 24 | 69027.62 | 67.1 |
| 400000497 | Anna | Chapman | Female | 26 | 20 | 65.1 | Northern Ireland | White Collar | 46 | 5789.63 | 46.23 |
| 400001939 | Richard | Dowd | Male | 21 | 20 | 70.9 | Northern Ireland | White Collar | 23 | 10248.59 | 36.48 |
| 300002298 | Phil | Arnold | Male | 37 | 30 | 70.4 | Wales | Blue Collar | 15 | 80824.89 | 36.11 |
| 100003271 | Christopher | Hunter | Male | 32 | 30 | 67.9 | England | Blue Collar | 69 | 76492.8 | 81.24 |
| 200000684 | Leonard | Nash | Male | 23 | 20 | 68.6 | Scotland | Blue Collar | 27 | 57918.09 | 73.88 |
| 100000251 | Fiona | Jones | Female | 33 | 30 | 65.4 | England | White Collar | 5 | 57127.03 | 59.75 |
| 100002956 | Neil | Fisher | Male | 36 | 30 | 71.8 | England | White Collar | 10 | 33595.68 | 83.52 |
| 100002663 | Charles | Watson | Male | 46 | 40 | 67.4 | England | White Collar | 25 | 94033.87 | 68.42 |
| 100002307 | Nicola | Gray | Female | 43 | 40 | 71.8 | England | White Collar | 43 | 106071.02 | 70.67 |
| 100000901 | Rose | Butler | Female | 38 | 30 | 61.3 | England | White Collar | 33 | 95477.25 | 44.15 |
| 100002035 | Kevin | Rees | Male | 29 | 20 | 70.7 | England | White Collar | 33 | 12258.48 | 44.83 |
| 300003429 | Alison | Young | Female | 28 | 20 | 62.6 | Wales | White Collar | 23 | 3856.26 | 126.32 |
| 100001358 | Edward | Parsons | Male | 49 | 40 | 62.5 | England | Other | 42 | 1426.26 | 82.71 |
| 200003843 | Colin | Mackenzie | Male | 39 | 30 | 68.3 | Scotland | Blue Collar | 37 | 177.28 | 45.35 |
| 100003213 | Max | Quinn | Male | 45 | 40 | 68.1 | England | White Collar | 17 | 2479.52 | 40.61 |
| 100000400 | Jacob | Fisher | Male | 35 | 30 | 66.1 | England | White Collar | 10 | 2423.46 | 60.61 |
| 100002749 | Sean | Paterson | Male | 28 | 20 | 72.2 | England | White Collar | 8 | 69126.18 | 76.78 |
| 200001964 | Nathan | Stewart | Male | 42 | 40 | 68 | Scotland | Blue Collar | 12 | 43767.15 | 41.33 |
| 100001960 | Olivia | Harris | Female | 31 | 30 | 63.6 | England | White Collar | 8 | 130717.22 | 119.63 |
| 100001418 | Adam | MacDonald | Male | 41 | 40 | 66.4 | England | White Collar | 38 | 14899.94 | 69.23 |
| 100003584 | Deirdre | Underwood | Female | 36 | 30 | 65.5 | England | White Collar | 52 | 8942.95 | 40.93 |
| 300000495 | Stephen | Poole | Male | 40 | 40 | 66.7 | Wales | White Collar | 3 | 8257.16 | 57.81 |
| 100000849 | Lily | Gill | Female | 25 | 20 | 58.8 | England | White Collar | 86 | 2703.33 | 44.16 |
| 200001403 | Adam | Martin | Male | 39 | 30 | 69.9 | Scotland | Blue Collar | 109 | 1255.52 | 42.61 |

# CONTINUOUS VS DISCRETE

| Customer ID | Name | Surname | Gender | Age | Age Group | Height | Region | Job Classification | Tenure Months | Balance | Spend On Groceries |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 200000262 | Zoe | Clarkson | Female | 59 | 50 | 62 | Scotland | Other | 24 | 23550.89 | 70.77 |
| 200001214 | Carolyn | McDonald | Female | 58 | 50 | 61.2 | Scotland | Other | 24 | 69027.62 | 67.1 |
| 400000497 | Anna | Chapman | Female | 26 | 20 | 65.1 | Northern Ireland | White Collar | 46 | 5789.63 | 46.23 |
| 400001939 | Richard | Dowd | Male | 21 | 20 | 70.9 | Northern Ireland | White C | | | 36.48 |
| 300002298 | Phil | Arnold | Male | 37 | 30 | 70.4 | Wales | Blue Co | | | 36.11 |
| 100003271 | Christopher | Hunter | Male | 32 | 30 | 67.9 | England | Blue Co | | | 81.24 |
| 200000684 | Leonard | Nash | Male | 23 | 20 | 68.6 | Scotland | Blue Co | | | 73.88 |
| 100000251 | Fiona | Jones | Female | 33 | 30 | 65.4 | England | White C | | | 59.75 |
| 100002956 | Neil | Fisher | Male | 36 | 30 | 71.8 | England | White C | | | 83.52 |
| 100002663 | Charles | Watson | Male | 46 | 40 | 67.4 | England | White C | | | 68.42 |
| 100002307 | Nicola | Gray | Female | 43 | 40 | 71.8 | England | White C | | | 70.67 |
| 100000901 | Rose | Butler | Female | 38 | 30 | 61.3 | England | White C | | | 44.15 |
| 100002035 | Kevin | Rees | Male | 29 | 20 | 70.7 | England | White C | | | 44.83 |
| 300003429 | Alison | Young | Female | 28 | 20 | 62.6 | Wales | White C | | | 126.32 |
| 100001358 | Edward | Parsons | Male | 49 | 40 | 62.5 | England | Other | | | 82.71 |
| 200003843 | Colin | Mackenzie | Male | 39 | 30 | 68.3 | Scotland | Blue Co | | | 45.35 |
| 100003213 | Max | Quinn | Male | 45 | 40 | 68.1 | England | White C | | | 40.61 |
| 100000400 | Jacob | Fisher | Male | 35 | 30 | 66.1 | England | White C | | | 60.61 |
| 100002749 | Sean | Paterson | Male | 28 | 20 | 72.2 | England | White C | | | 76.78 |
| 200001964 | Nathan | Stewart | Male | 42 | 40 | 68 | Scotland | Blue Co | | | 41.33 |
| 100001960 | Olivia | Harris | Female | 31 | 30 | 63.6 | England | White C | | | 119.63 |
| 100001418 | Adam | MacDonald | Male | 41 | 40 | 66.4 | England | White C | | | 69.23 |
| 100003584 | Deirdre | Underwood | Female | 36 | 30 | 65.5 | England | White Collar | 52 | 8942.95 | 40.93 |
| 300000495 | Stephen | Poole | Male | 40 | 40 | 66.7 | Wales | White Collar | 3 | 8257.16 | 57.81 |
| 100000849 | Lily | Gill | Female | 25 | 20 | 58.8 | England | White Collar | 86 | 2703.33 | 44.16 |
| 200001403 | Adam | Martin | Male | 39 | 30 | 69.9 | Scotland | Blue Collar | 109 | 1255.52 | 42.61 |

**Discrete Variables**

Scotland = 0

Northern Ireland = 1

Wales = 2

England = 4

# CONTINUOUS VS DISCRETE

| Customer ID | Name | Surname | Gender | Age | Age Group | Height | Region | Job Classification | Tenure Months | Balance | Spend On Groceries |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 200000262 | Zoe | Clarkson | Female | 59 | 50 | 62 | Scotland | Other | 24 | 23550.89 | 70.77 |
| 200001214 | Carolyn | McDonald | Female | 58 | 50 | 61.2 | Scotland | Other | 24 | 69027.62 | 67.1 |
| 400000497 | Anna | Chapman | Female | 26 | 20 | 65.1 | Northern Ireland | White Collar | 46 | 5789.63 | 46.23 |
| 400001939 | Richard | Dowd | Male | 21 | 20 | 70.9 | Northern Ire | | 23 | 10248.59 | 36.48 |
| 300002298 | Phil | Arnold | Male | 37 | 30 | 70.4 | Wales | | 15 | 80824.89 | 36.11 |
| 100003271 | Christopher | Hunter | Male | 32 | 30 | 67.9 | England | | 69 | 76492.8 | 81.24 |
| 200000684 | Leonard | Nash | Male | 23 | 20 | 68.6 | Scotland | | 27 | 57918.09 | 73.88 |
| 100000251 | Fiona | Jones | Female | 33 | 30 | 65.4 | England | | 5 | 57127.03 | 59.75 |
| 100002956 | Neil | Fisher | Male | 36 | 30 | 71.8 | England | | 10 | 33595.68 | 83.52 |
| 100002663 | Charles | Watson | Male | 46 | 40 | 67.4 | England | | 25 | 94033.87 | 68.42 |
| 100002307 | Nicola | Gray | Female | 43 | 40 | 71.8 | England | | 43 | 106071.02 | 70.67 |
| 100000901 | Rose | Butler | Female | 38 | 30 | 61.3 | England | | 33 | 95477.25 | 44.15 |
| 100002035 | Kevin | Rees | Male | 29 | 20 | 70.7 | England | | 33 | 12258.48 | 44.83 |
| 300003429 | Alison | Young | Female | 28 | 20 | 62.6 | Wales | | 23 | 3856.26 | 126.32 |
| 100001358 | Edward | Parsons | Male | 49 | 40 | 62.5 | England | | 42 | 1426.26 | 82.71 |
| 200003843 | Colin | Mackenzie | Male | 39 | 30 | 68.3 | Scotland | | 37 | 177.28 | 45.35 |
| 100003213 | Max | Quinn | Male | 45 | 40 | 68.1 | England | | 17 | 2479.52 | 40.61 |
| 100000400 | Jacob | Fisher | Male | 35 | 30 | 66.1 | England | | 10 | 2423.46 | 60.61 |
| 100002749 | Sean | Paterson | Male | 28 | 20 | 72.2 | England | | 8 | 69126.18 | 76.78 |
| 200001964 | Nathan | Stewart | Male | 42 | 40 | 68 | Scotland | | 12 | 43767.15 | 41.33 |
| 100001960 | Olivia | Harris | Female | 31 | 30 | 63.6 | England | | 8 | 130717.22 | 119.63 |
| 100001418 | Adam | MacDonald | Male | 41 | 40 | 66.4 | England | | 38 | 14899.94 | 69.23 |
| 100003584 | Deirdre | Underwood | Female | 36 | 30 | 65.5 | England | White Collar | 52 | 8942.95 | 40.93 |
| 300000495 | Stephen | Poole | Male | 40 | 40 | 66.7 | Wales | White Collar | 3 | 8257.16 | 57.81 |
| 100000849 | Lily | Gill | Female | 25 | 20 | 58.8 | England | White Collar | 86 | 2703.33 | 44.16 |
| 200001403 | Adam | Martin | Male | 39 | 30 | 69.9 | Scotland | Blue Collar | 109 | 1255.52 | 42.61 |

**Continuous Variables**

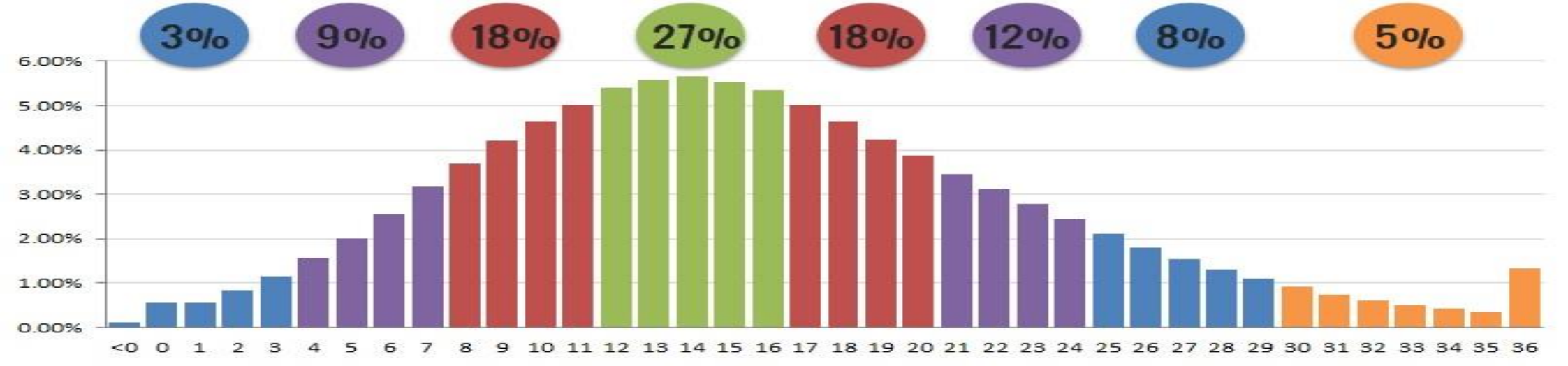
23550.89

235464.545

69027.62

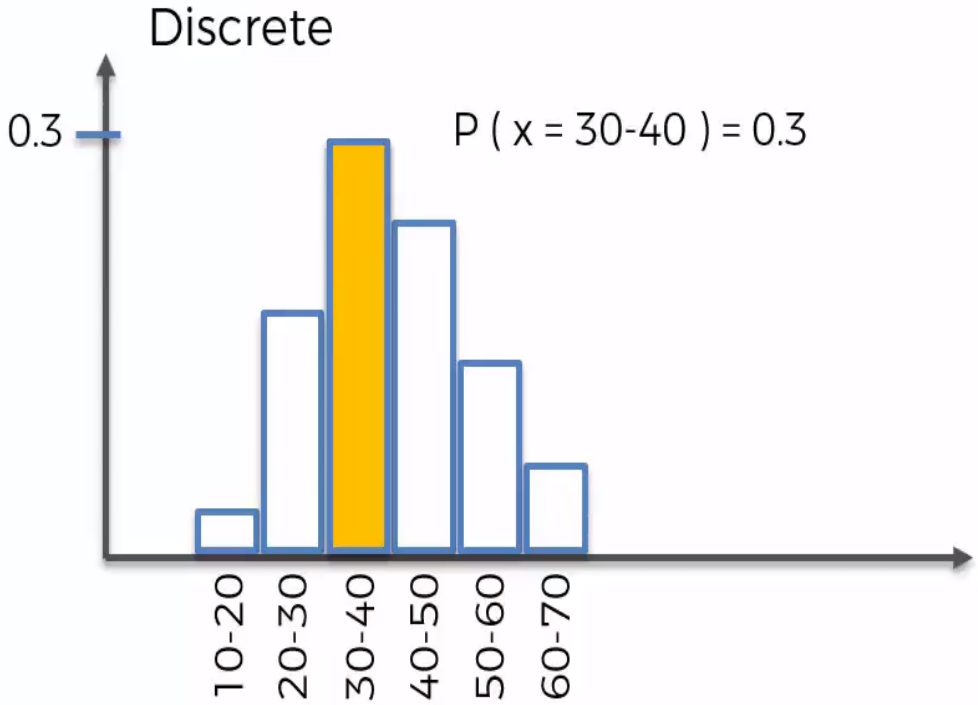
# WHAT IS A DISTRIBUTION?

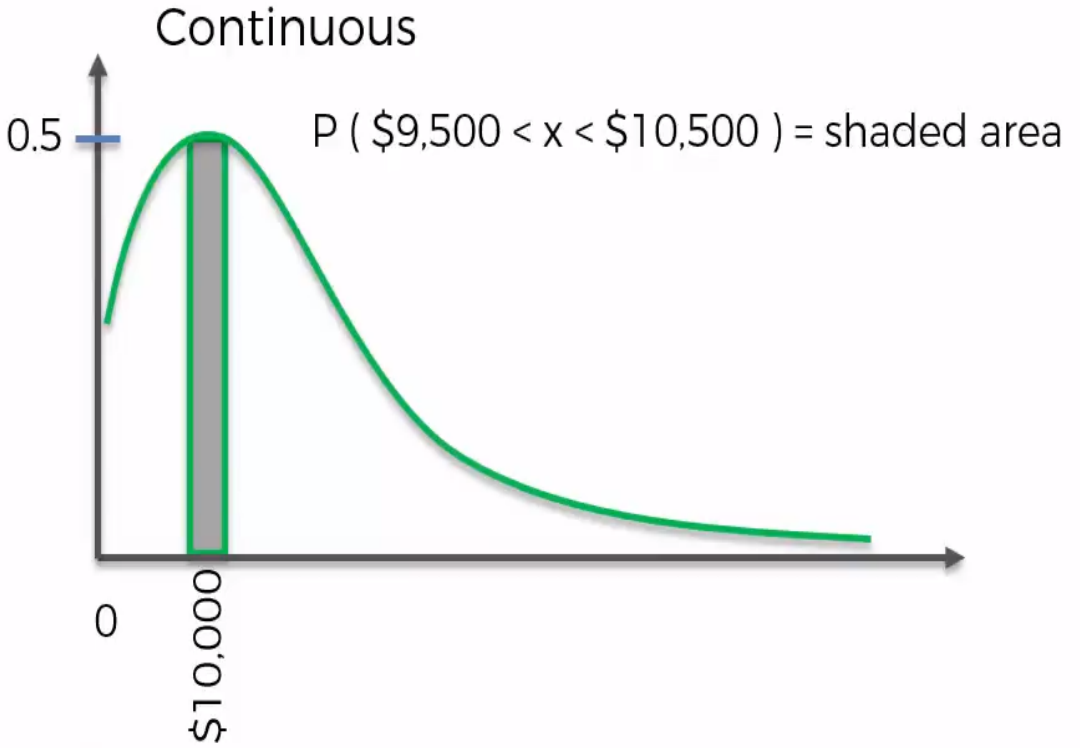
*A Probability distribution is a mathematical function that stated in a simple terms, can be thought of as providing the probability of occurrence of different possible outcomes in an experiment*



# WHAT IS A DISTRIBUTION?

| Customer ID | Name | Surname | Gender | Age | Age Group | Height | Region | Job Classification | Tenure Months | Balance | Spend On Groceries |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 200000262 | Zoe | Clarkson | Female | 59 | 50 | 62 | Scotland | Other | 24 | 23550.89 | 70.77 |
| 200001214 | Carolyn | McDonald | Female | 58 | 50 | 61.2 | Scotland | Other | 24 | 69027.62 | 67.1 |
| 400000497 | Anna | Chapman | Female | 26 | 20 | 65.1 | Northern Ireland | White Collar | 46 | 5789.63 | 46.23 |
| 400001939 | Richard | Dowd | Male | 21 | 20 | 70.9 | Northern Ireland | White Collar | 23 | 10248.59 | 36.48 |
| 300002298 | Phil | Arnold | Male | 37 | 30 | 70.4 | Wales | Blue Collar | 15 | 80824.89 | 36.11 |



# STANDARD DEVIATION

| Customer ID | Name | Surname | Gender | Age | Age Group | Height | Region | Job Classification | Tenure Months | Balance | Spend On Groceries |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 200000262 | Zoe | Clarkson | Female | 59 | 50 | 62 | Scotland | Other | 24 | 23550.89 | 70.77 |
| 200001214 | Carolyn | McDonald | Female | 58 | 50 | 61.2 | Scotland | Other | 24 | 69027.62 | 67.1 |
| 400000497 | Anna | Chapman | Female | 26 | 20 | 65.1 | Northern Ireland | White Collar | 46 | 5789.63 | 46.23 |
| 400001939 | Richard | Dowd | Male | 21 | 20 | 70.9 | Northern Ireland | White Collar | 23 | 10248.59 | 36.48 |
| 300002298 | Phil | Arnold | Male | 37 | 30 | 70.4 | Wales | Blue Collar | 15 | 80824.89 | 36.11 |

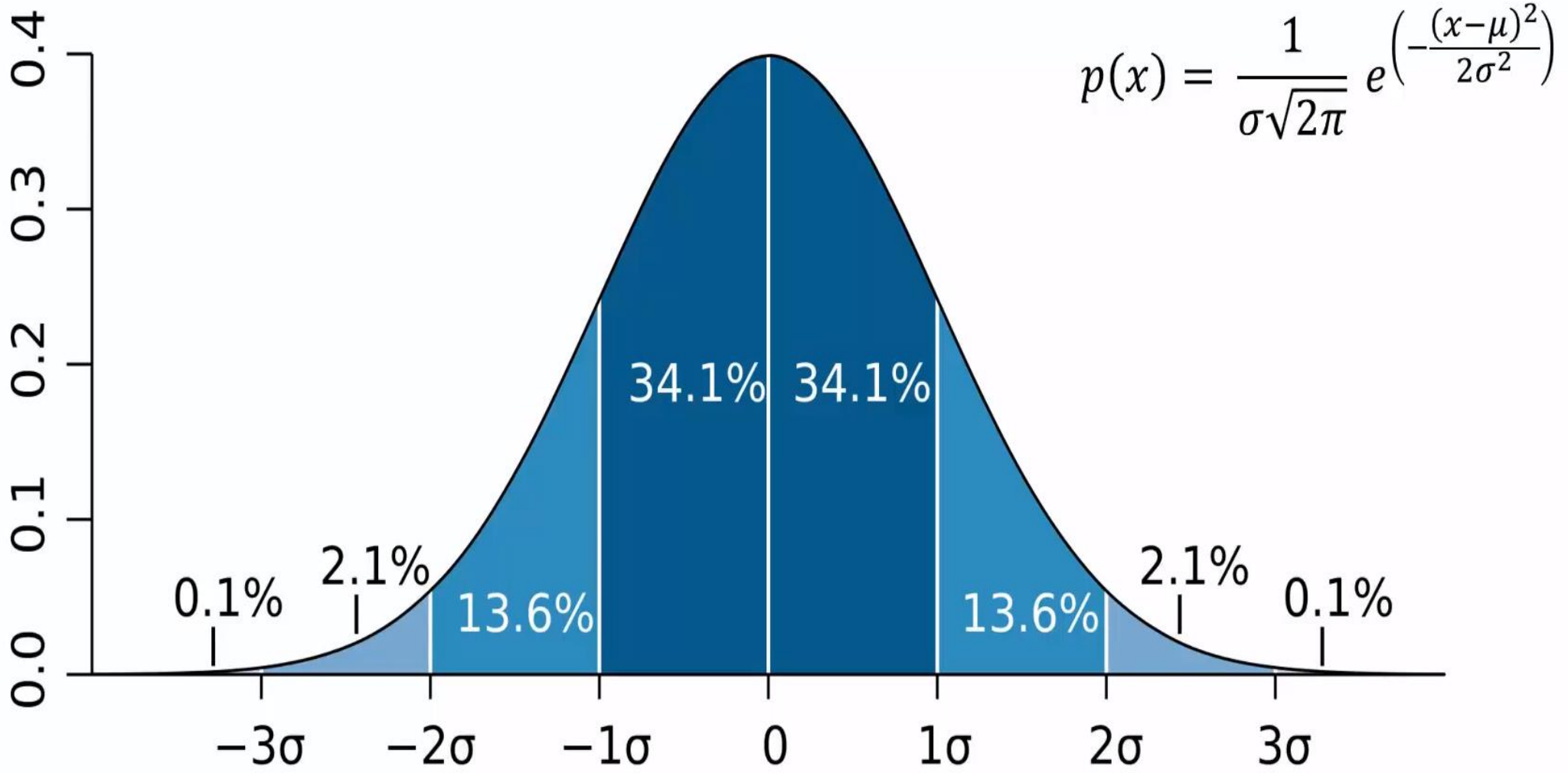
$$\{ 61.2, 62, 65.1, 70.4, 70.9 \}$$

$$\text{Mean} = \frac{61.2 + 62 + 65.1 + 70.4 + 70.9}{5} = 65.92$$

$$\text{Variance} = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = 16.64$$

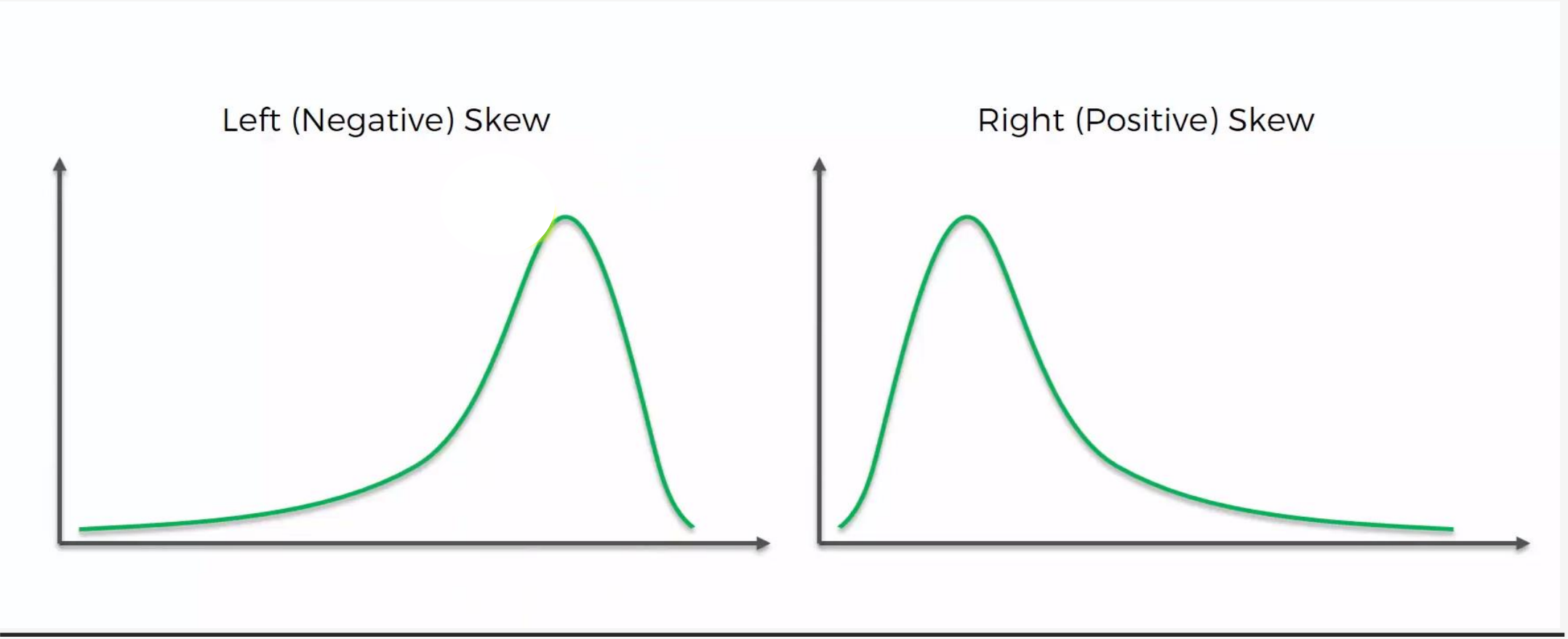
$$\text{STD.DEV} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} = 4.08$$

# NORMAL DISTRBUTION



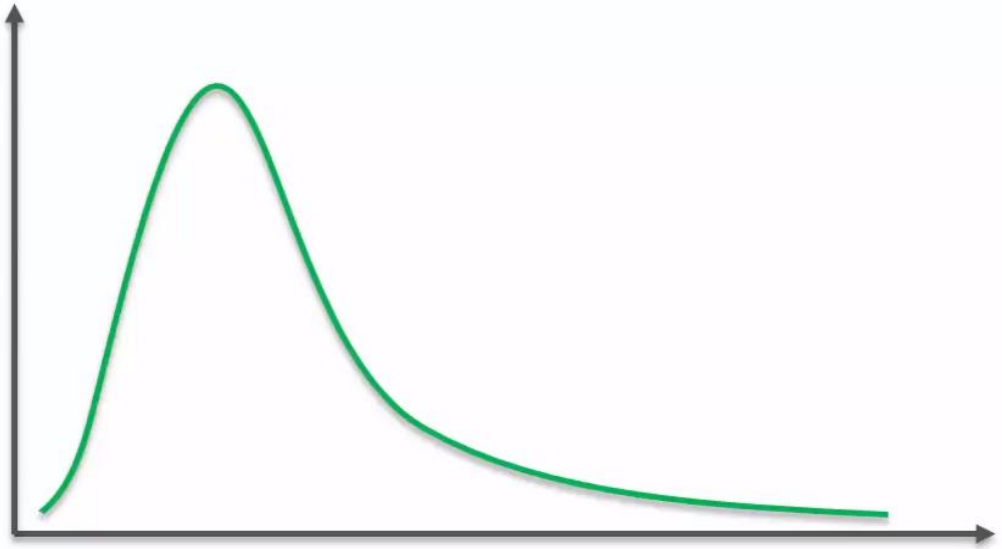
# SKEWNESS

**Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution**



# MEAN, MEDIAN, MODE

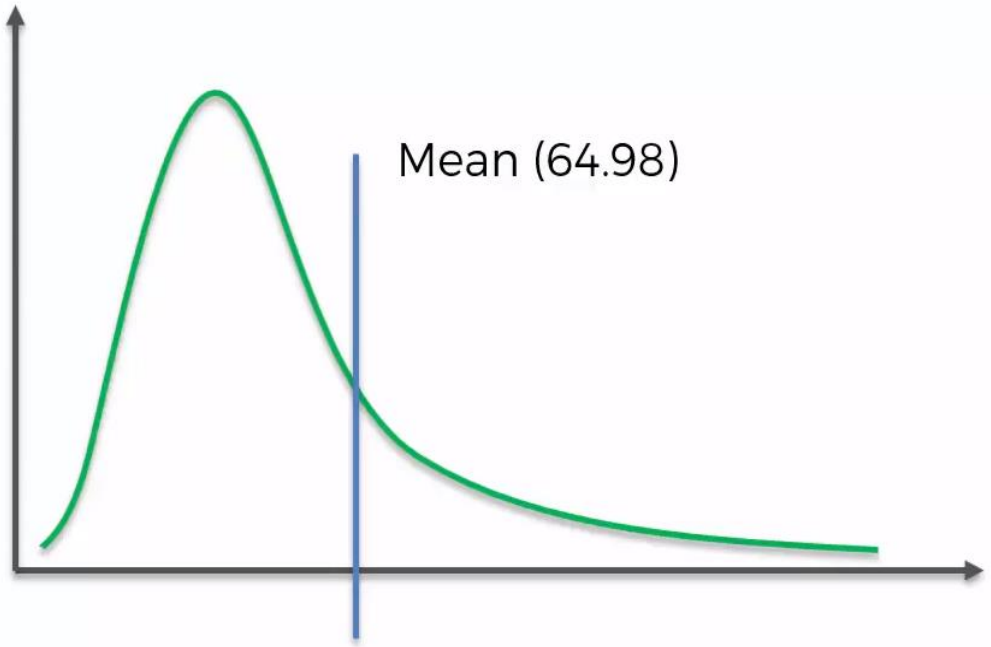
Height data set



{ 58.8 59.9 61.2 61.3 61.5 62 62.5 62.5 62.5 63 63 63.7 63.9 64 64 65.1 65.2 66.7 67.8 68.3 69.9 70.7 71.8 72.2 73 }

# MEAN, MEDIAN, MODE

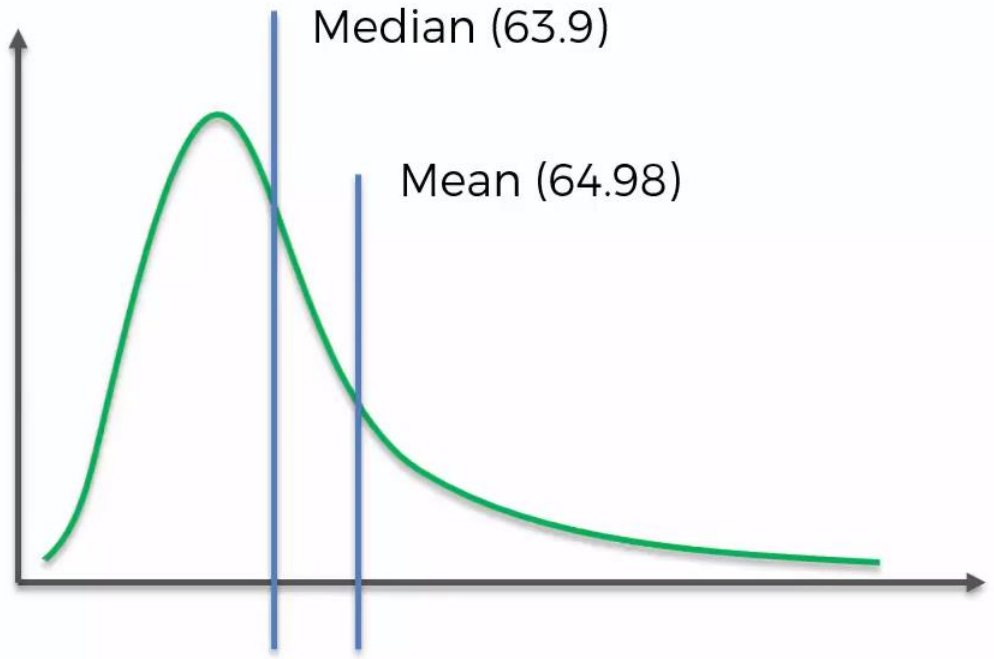
The **statistical mean** refers to the **mean** or average that is used to derive the central tendency of the data in question. It is determined by adding all the data points in a population and then dividing the total by the number of points. The resulting number is known as the **mean** or the average.




{ 58.8 59.9 61.2 61.3 61.5 62 62.5 62.5 62.5 63 63 63.7 63.9 64 64 65.1 65.2 66.7 67.8 68.3 69.9 70.7 71.8 72.2 73 }

# MEAN, MEDIAN, MODE

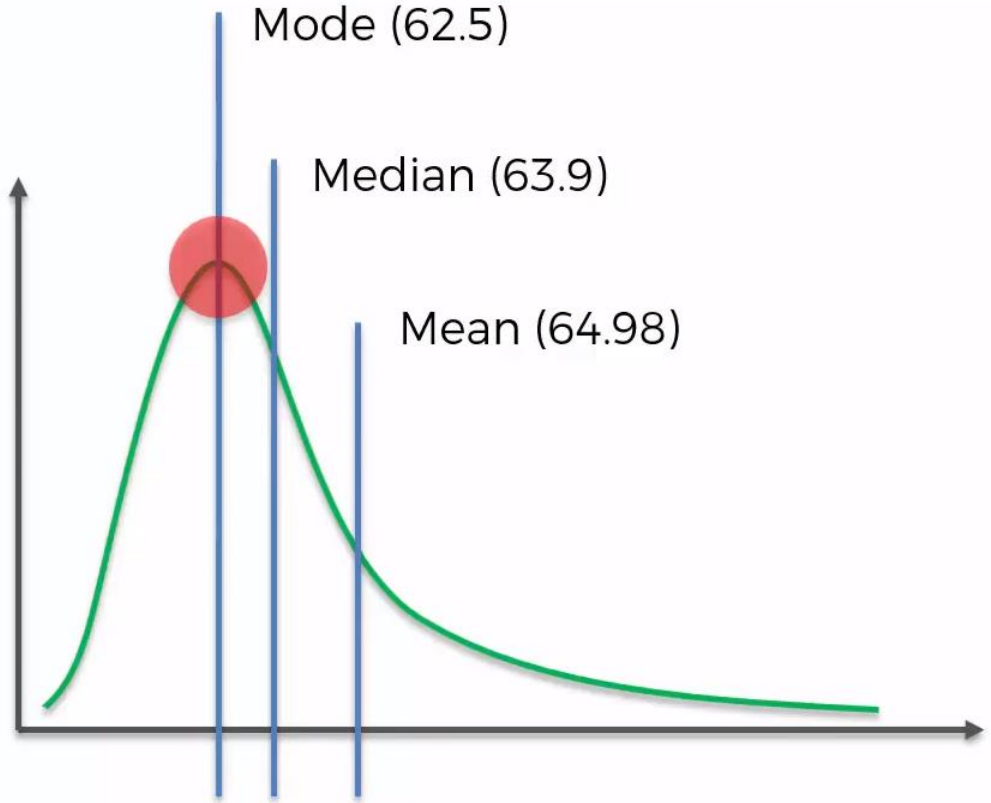
The **median** is a simple measure of central tendency. To find the **median**, we arrange the observations in order from smallest to largest value. If there is an odd number of observations, the **median** is the middle value. If there is an even number of observations, the **median** is the average of the two middle values.




{ 58.8 59.9 61.2 61.3 61.5 62 62.5 62.5 62.5 63 63 63.7 63.9 64 64 65.1 65.2 66.7 67.8 68.3 69.9 70.7 71.8 72.2 73 }

# MEAN, MEDIAN, MODE

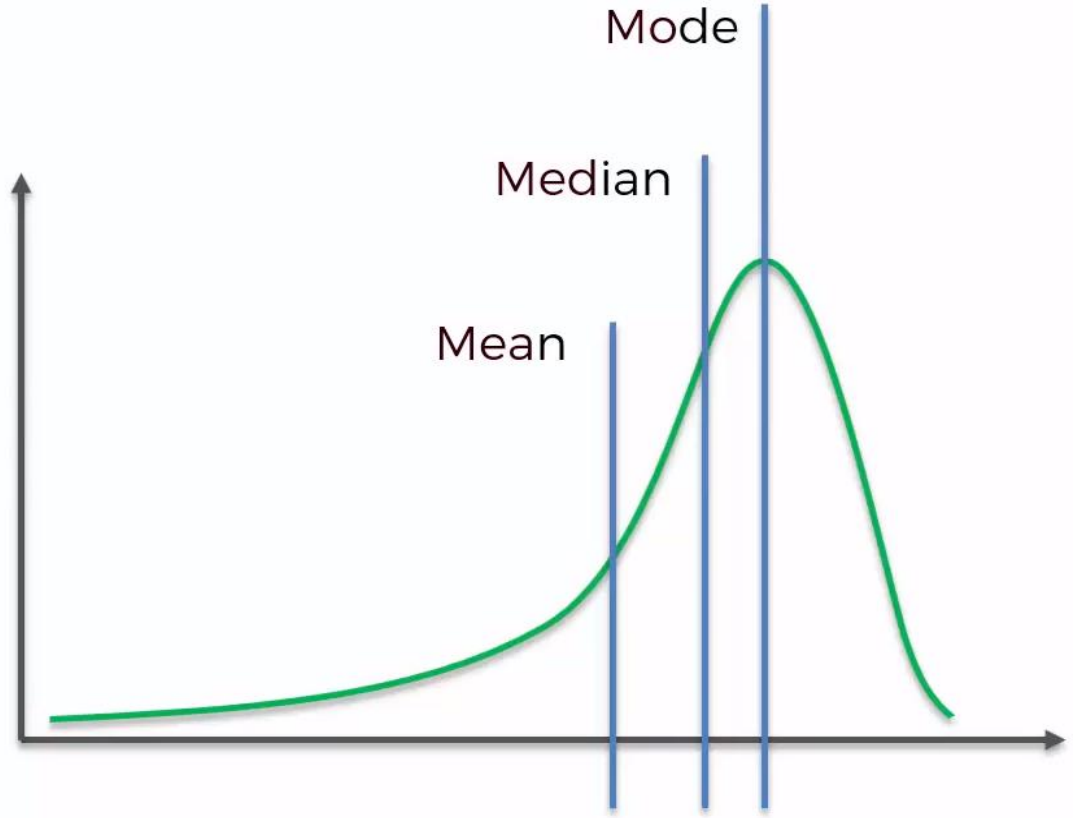
The **mode** is a **statistical** term that refers to the most frequently occurring number found **in** a set of numbers. The **mode** is found by collecting and organizing data **in** order to count the frequency of each result. The result with the highest number of occurrences is the **mode** of the set.




{ 58.8 59.9 61.2 61.3 61.5 62 62.5 62.5 62.5 63 63 63.7 63.9 64 64 65.1 65.2 66.7 67.8 68.3 69.9 70.7 71.8 72.2 73 }

# MEAN, MEDIAN, MODE

Left Skew



# HOMEWORK CHALLENGE

# *HOMEWORK CHALLENGE*

You are an Analyst working for a high-end clothes design boutique.

The company is developing a new line of clothes for very tall people. Your team is analyzing the viability of the project from a sales perspective and your manager has asked you to assist with some input variables to help test the financial forecast.

You need to create two distributions:

- A normal distribution of 1000 observations for heights of men in Jordan
- A normal distribution of 1000 observations for heights of women in Jordan

# HOMEWORK CHALLENGE

Also, for each of the two populations you have been asked to identify the minimum height of 2.2% of the tallest people in the population.

In Jordan, men's heights have a mean of 69.1 inches (175.5 cm) and standard deviation 2.9 inches (7.4 cm), while female's heights have a mean of 63.7 inches (161.8 cm) and standard deviation 2.7 inches (6.9 cm).

Hint 1 inch = 2.54 cm

# *HOMEWORK CHALLENGE*

Hint #1

Use the NORM.INV() function combined with RAND() in Excel

Example:

NORM.INV(RAND(), 69.1, 2.9)

# *HOMEWORK CHALLENGE*

## Hint #2

If you want to visualize your distributions, you will need to allocate your data to bins first. Check out the article on Microsoft for more information on how to do this in Excel

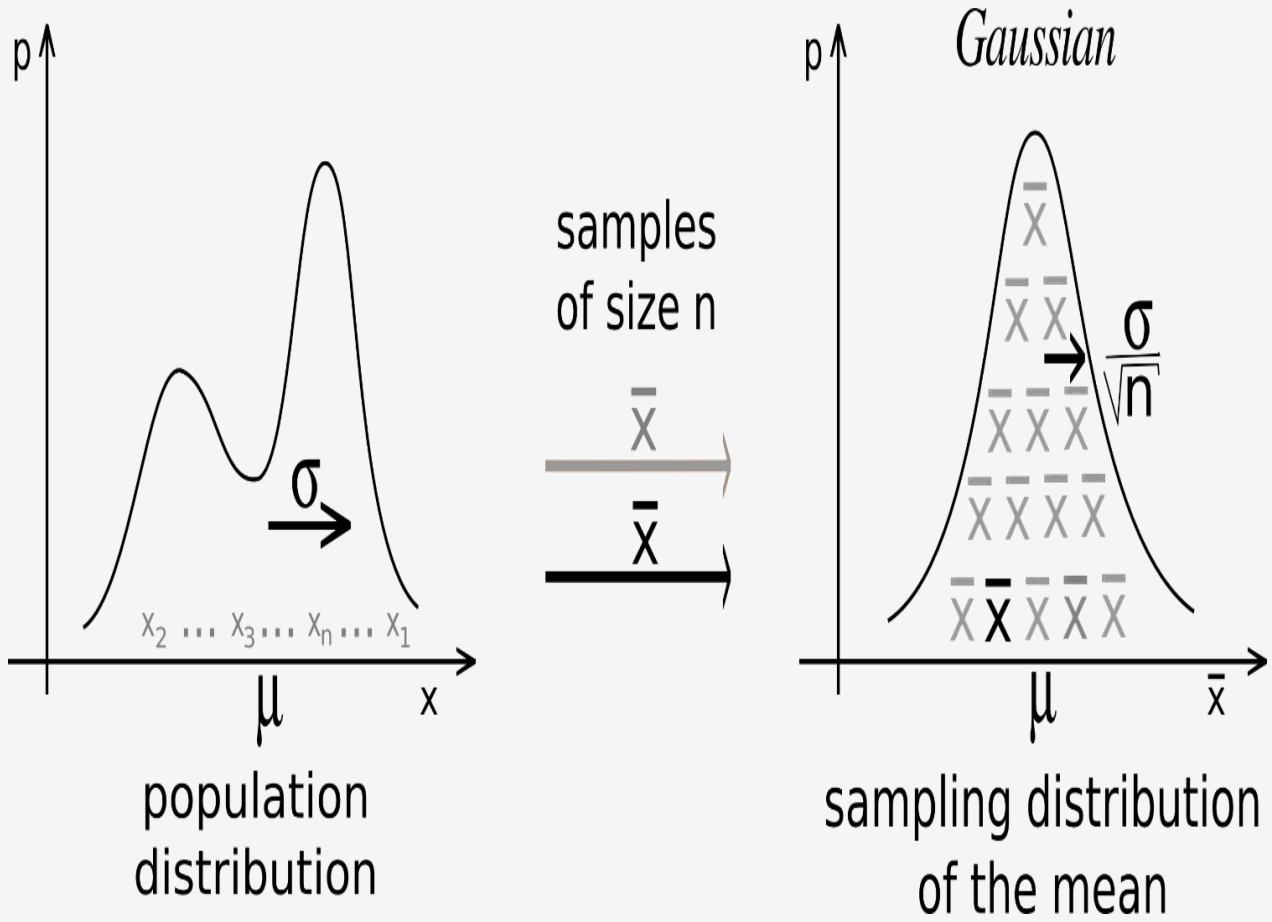
How to use the Histogram tool in Excel

https://support.microsoft.com/en-us/help/214269/how-to-use-the-histogram-tool-in-excel

Note: this histogram won't be dynamic. We will learn how to make the dynamic on in the homework solution

*Section 2*

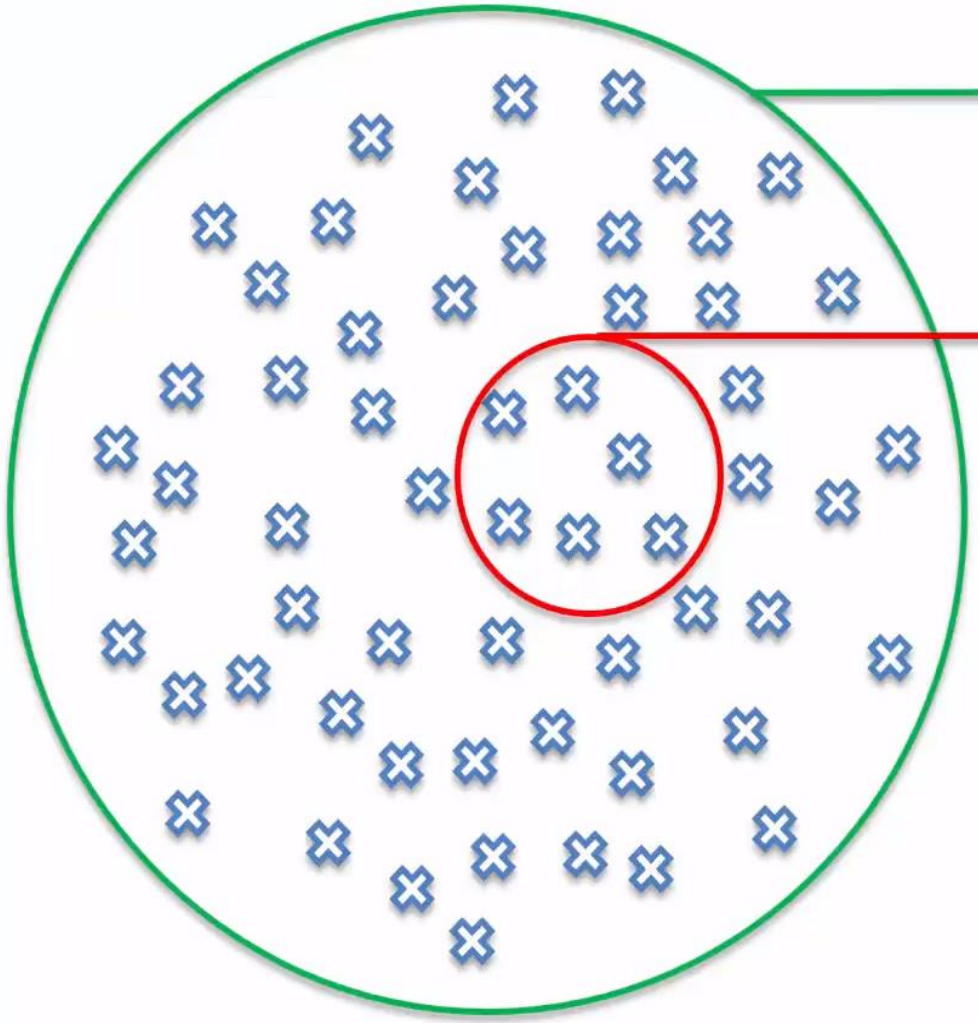
# Central Limit Theorem



# WHAT WE WILL LEARN

1. Population and Samples
2. Sampling Distribution
3. Central Limit Theorem
4. Central Limit Theorem – Intuition
5. Central Limit Theorem – Visualization
6. Z-Score
7. Hands-On CLT: An Analytics Challenge

# POPULATION AND SAMPLES



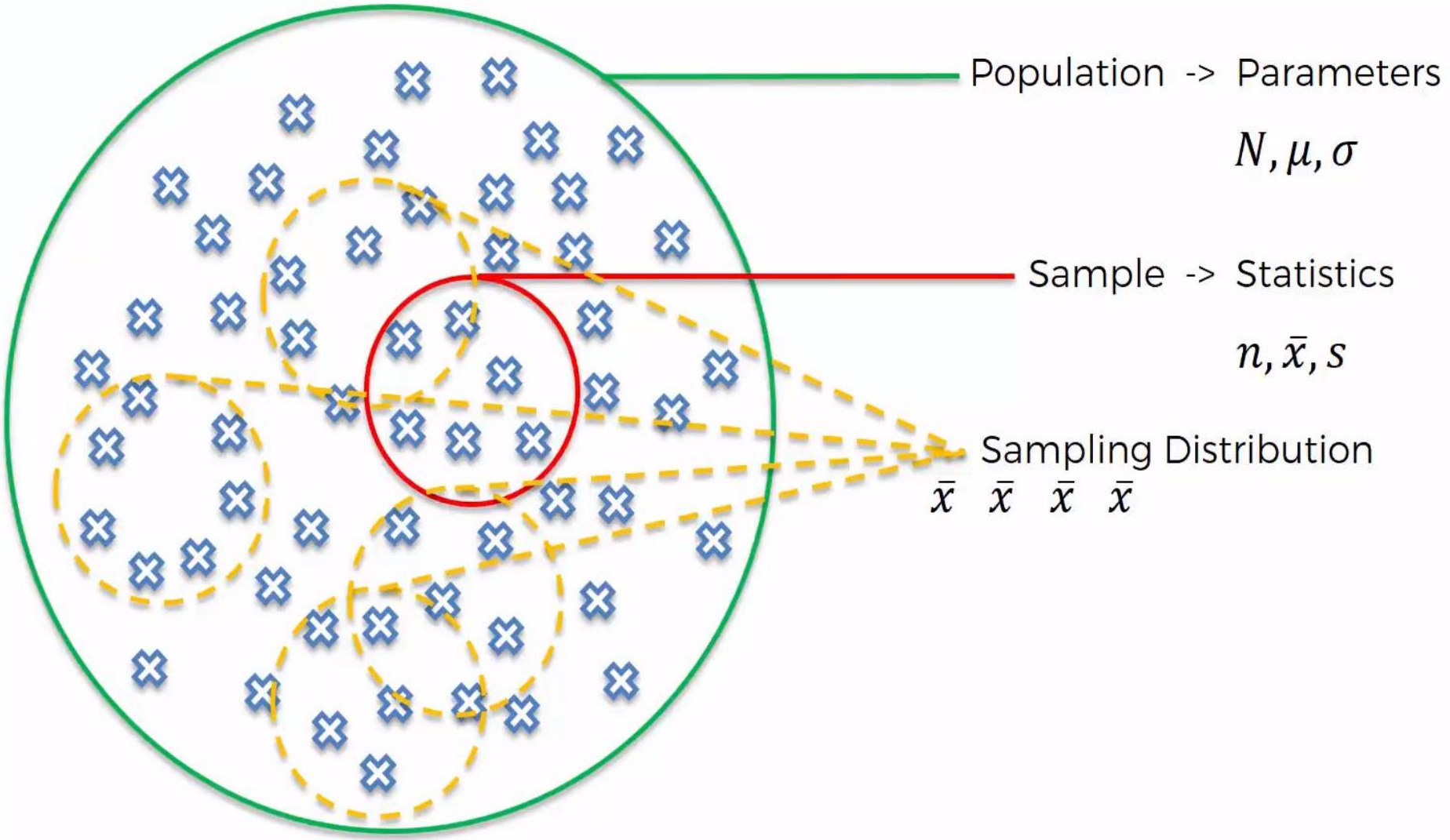
Population -> Parameters

$N, \mu, \sigma$

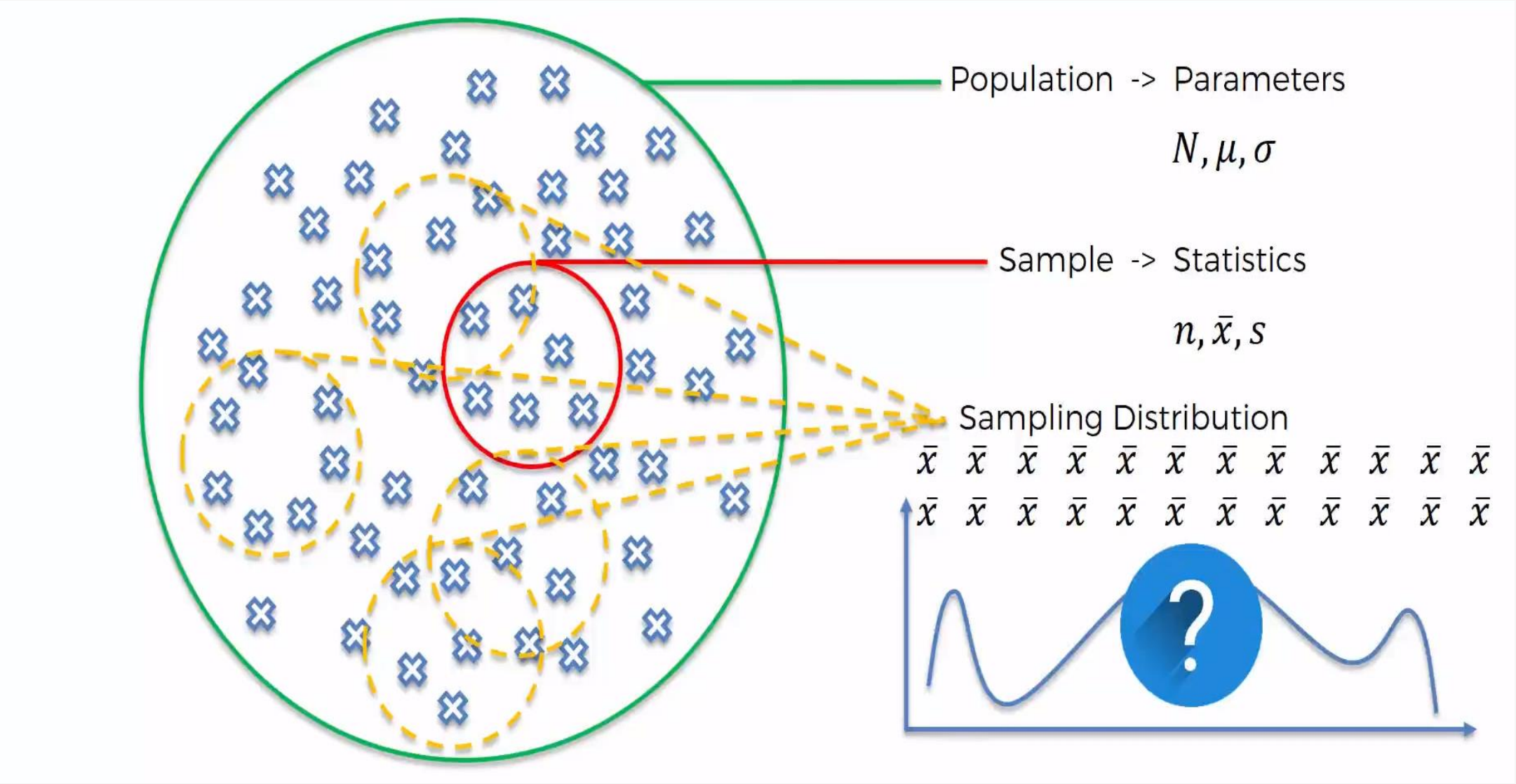
Sample -> Statistics

$n, \bar{x}, s$

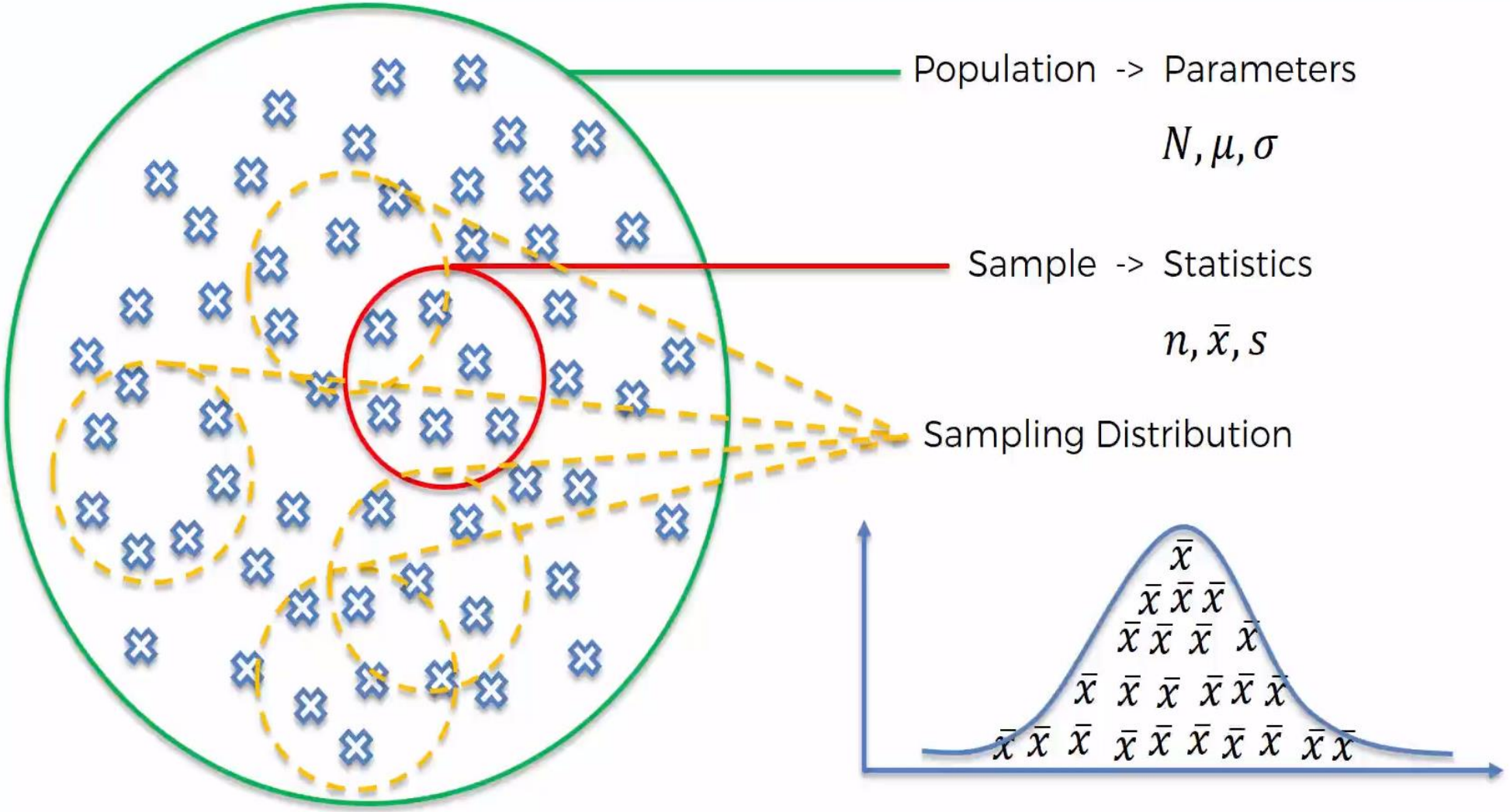
# SAMPLING DISTRIBUTION



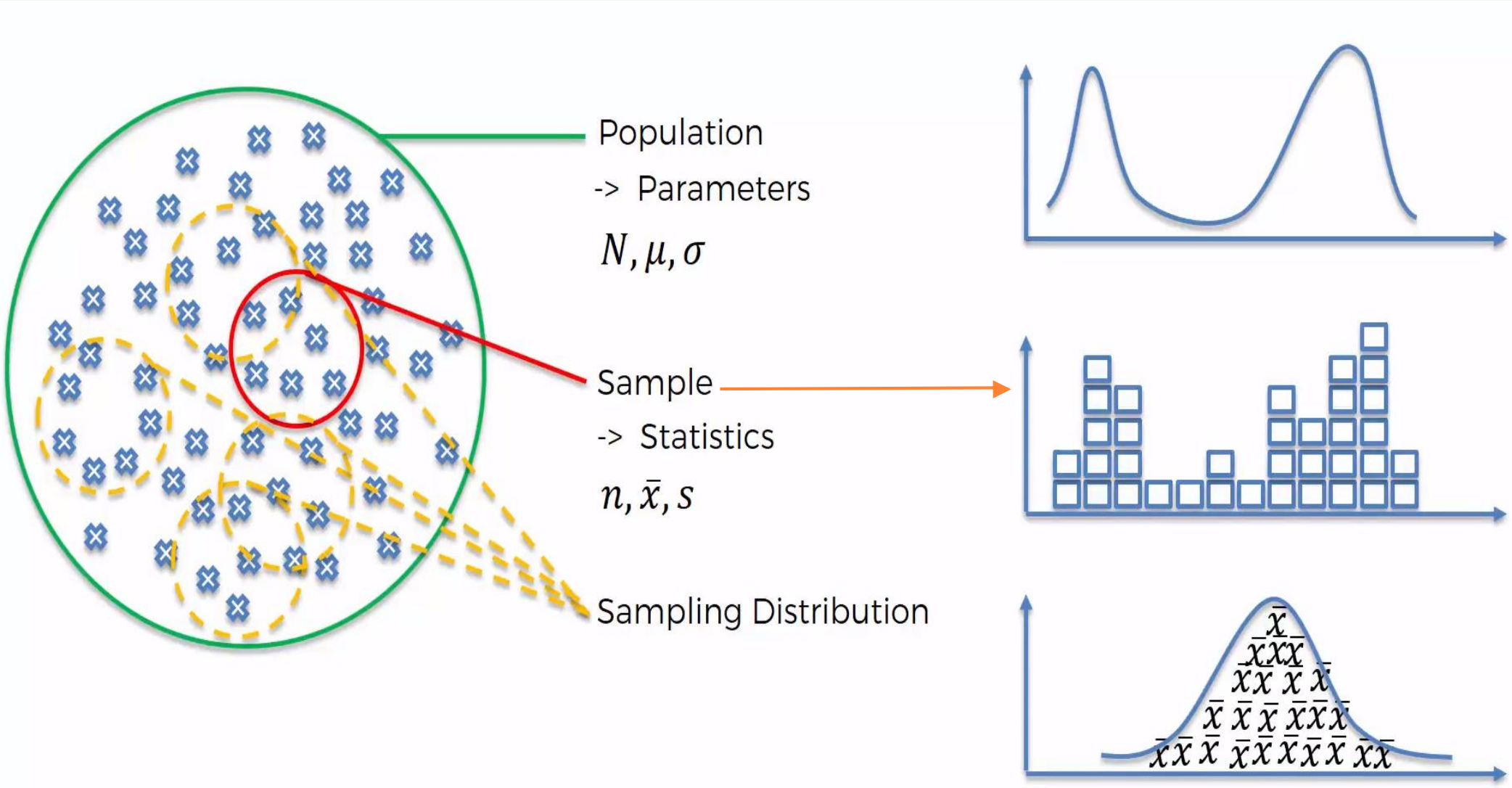
# SAMPLING DISTRIBUTION



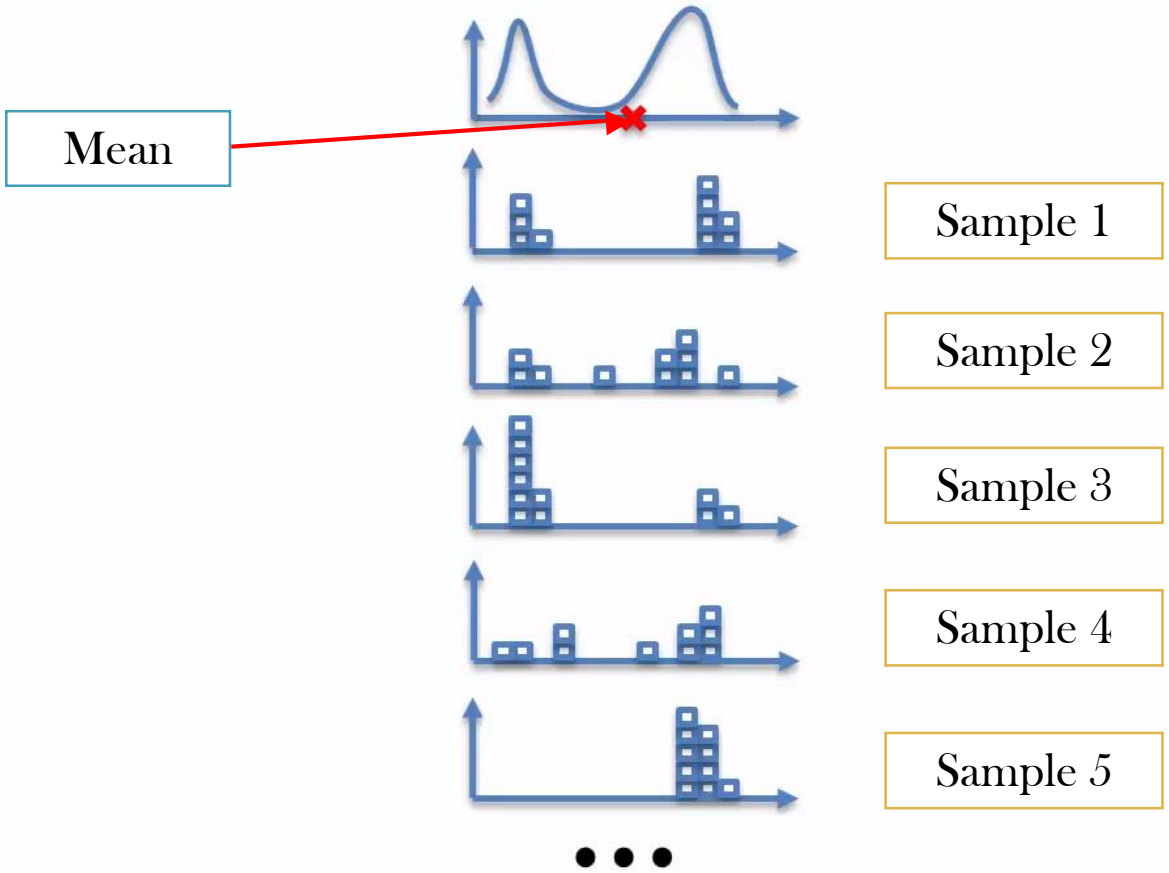
# CENTRAL LIMIT THEOREM



# CENTRAL LIMIT THEOREM

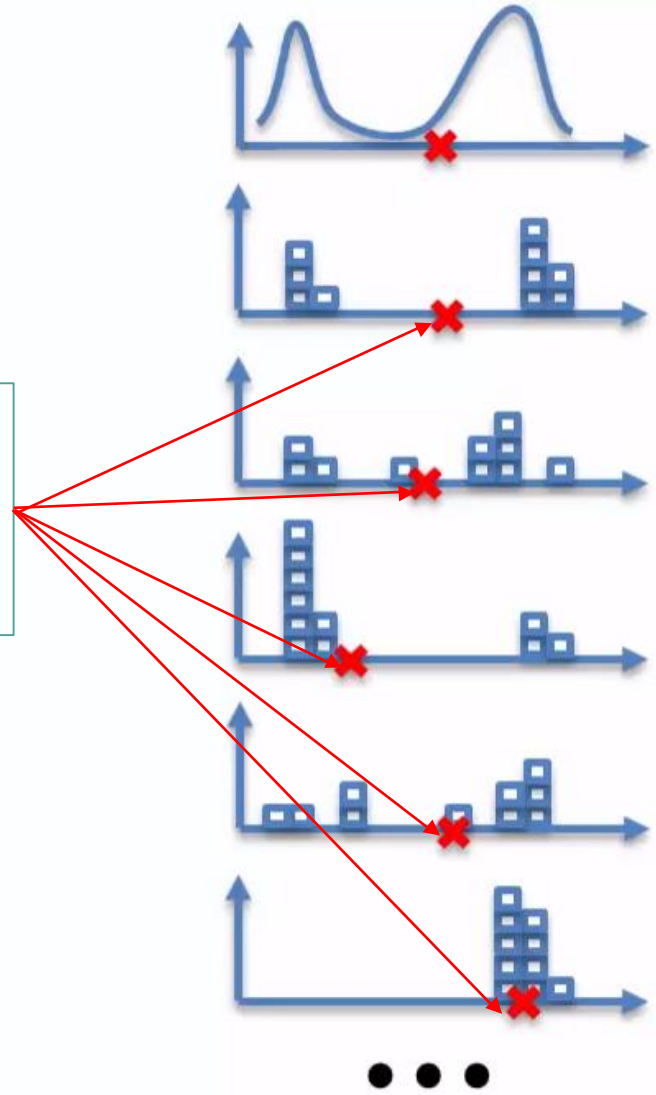


# CENTRAL LIMIT THEOREM – INTUITION

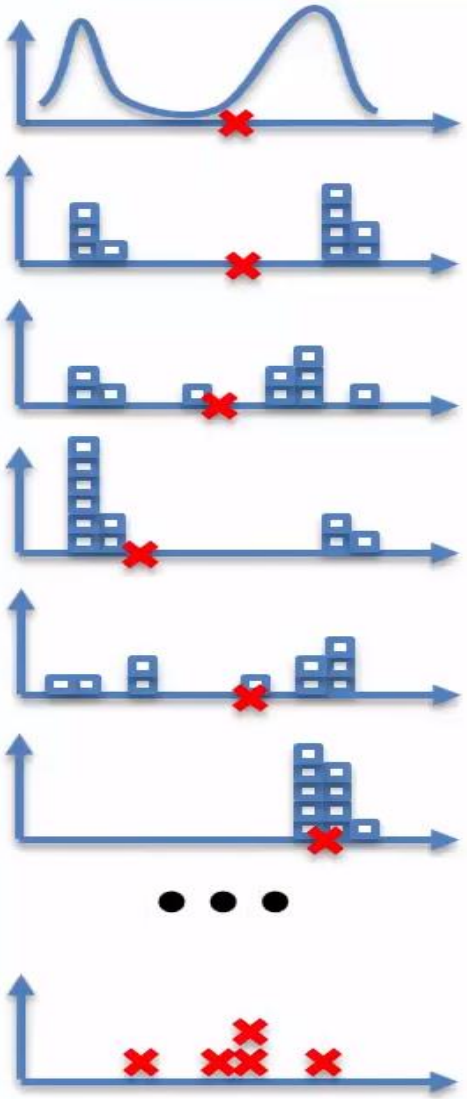


# CENTRAL LIMIT THEOREM – INTUITION

Lets take the x hat (mean) for each sample distribution



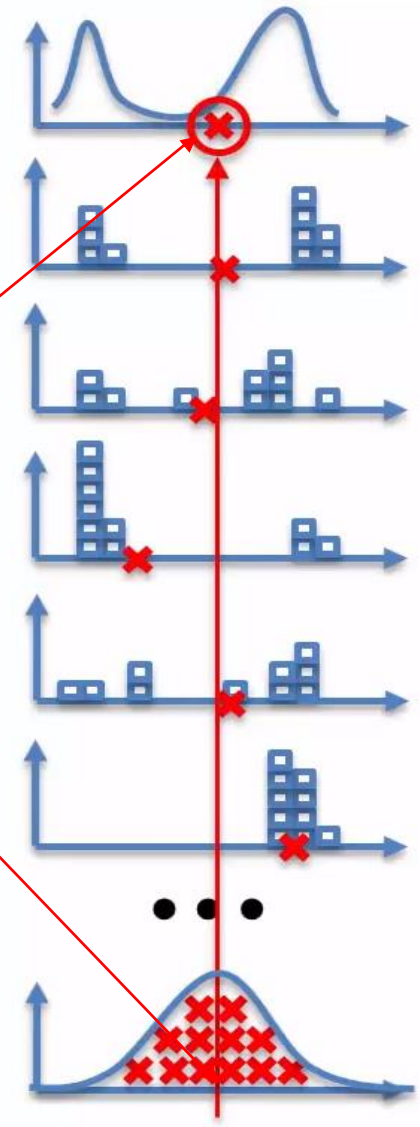
# CENTRAL LIMIT THEOREM – INTUITION



With more samples

# CENTRAL LIMIT THEOREM – INTUITION

The original distribution and the sampling distribution have the same mean



Sampling Distribution:

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

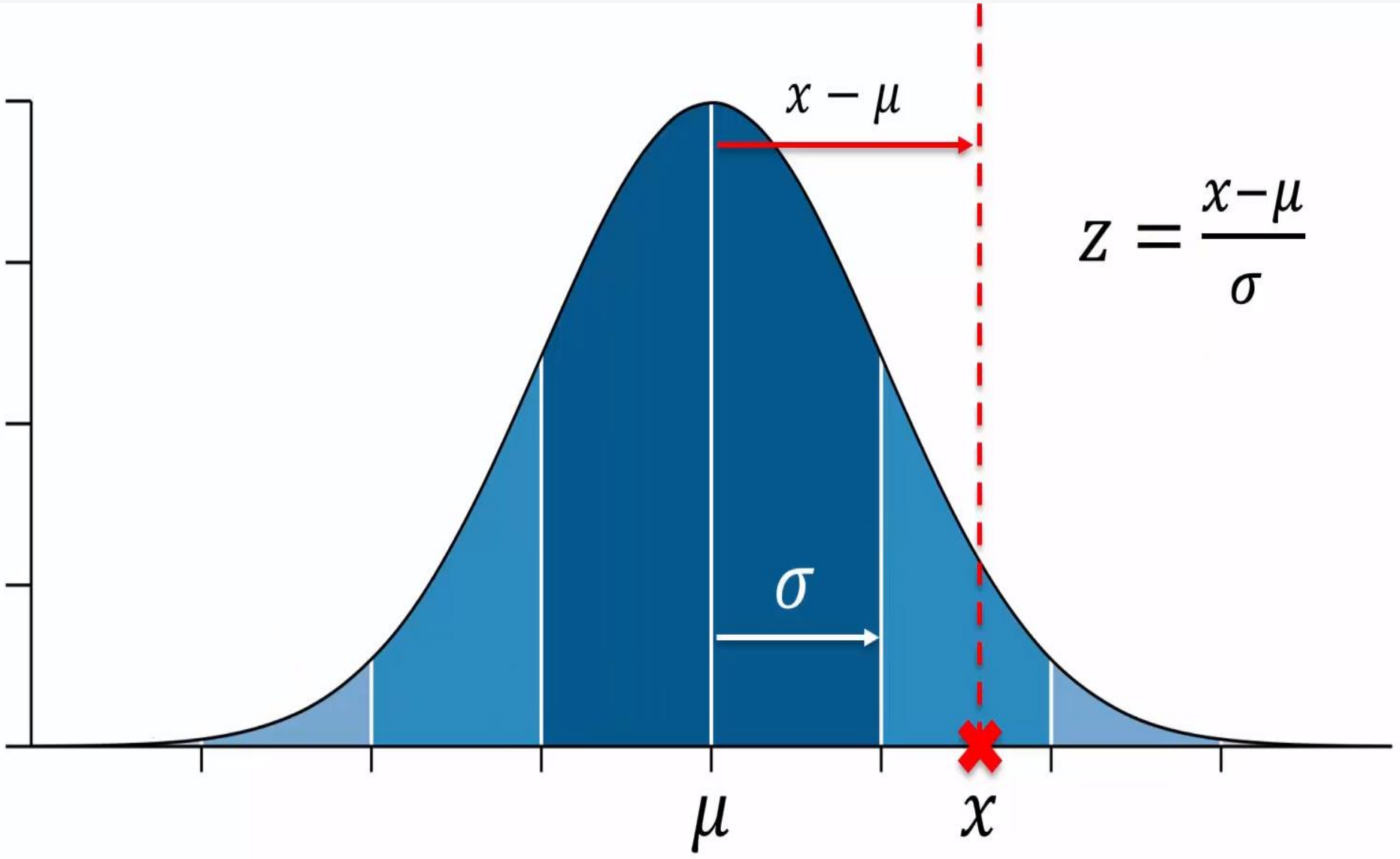
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

# CENTRAL LIMIT THEOREM – VISUALIZATION

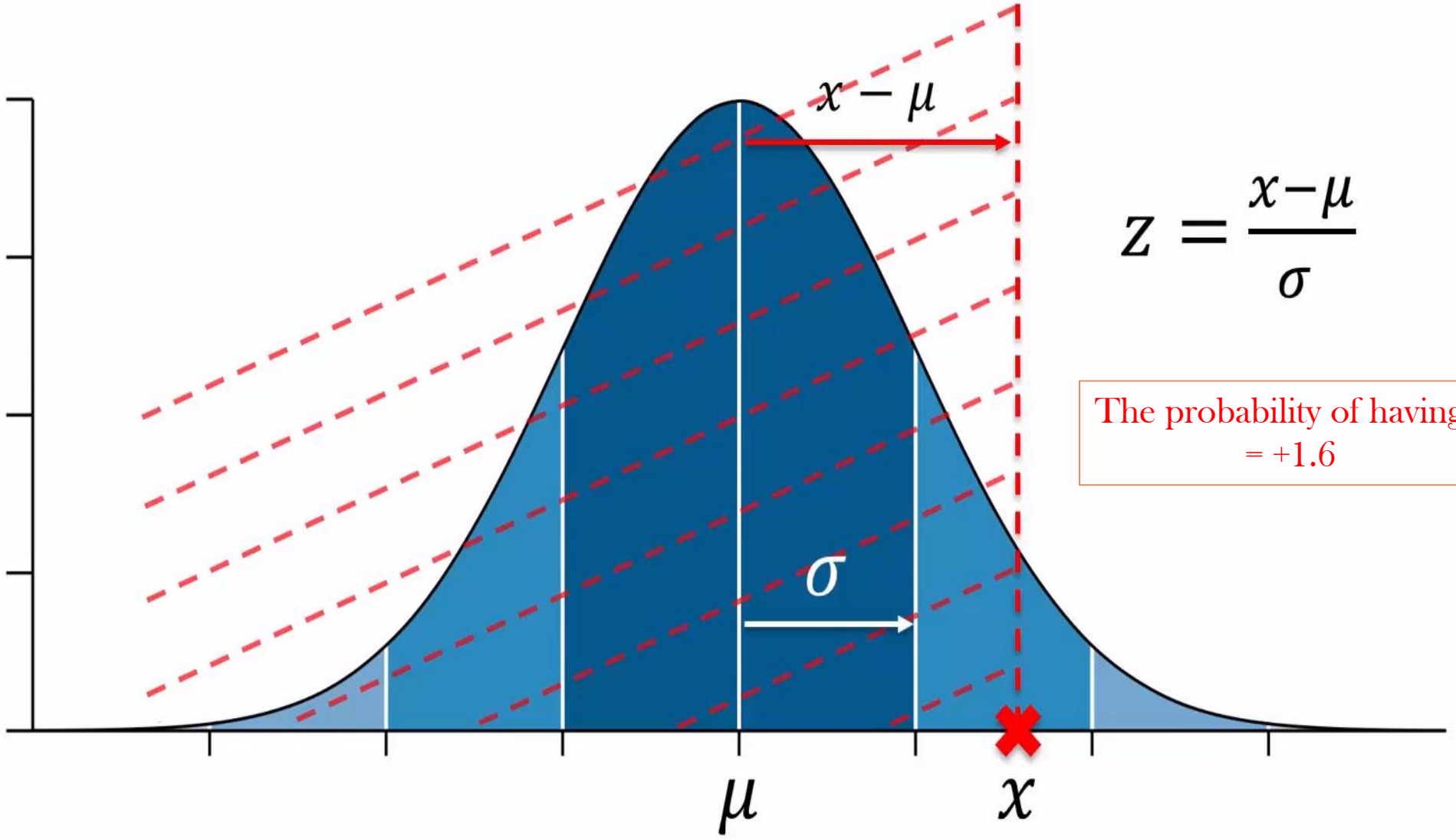
Go online and paste the website

http://www.onlinestatbook.com/stat_sim/sampling_dist/index.html

# Z-SCORE



# Z-SCORE




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

The probability of having z = +1.6

# *ANALYTICS CHALLENGE*

You are a Business Analyst working for FedEx*.

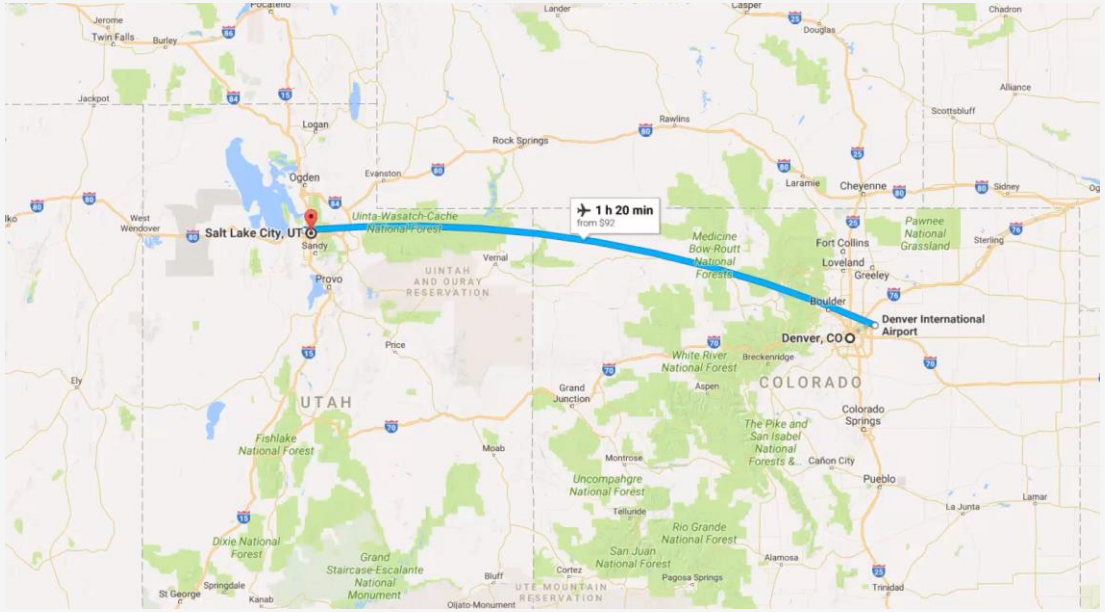
A business client has requested a large freight to be transported urgently from Denver  to Salt Lake City. When asked about the weight of the cargo they could not supply the exact weight. However they have specified that there are a total of 36 boxes.

From prior experience with this client you know that this type of cargo follows a distribution with a mean = 72 lb. (32.66 kg) and std.dev of 3 lb. (1.36 kg).

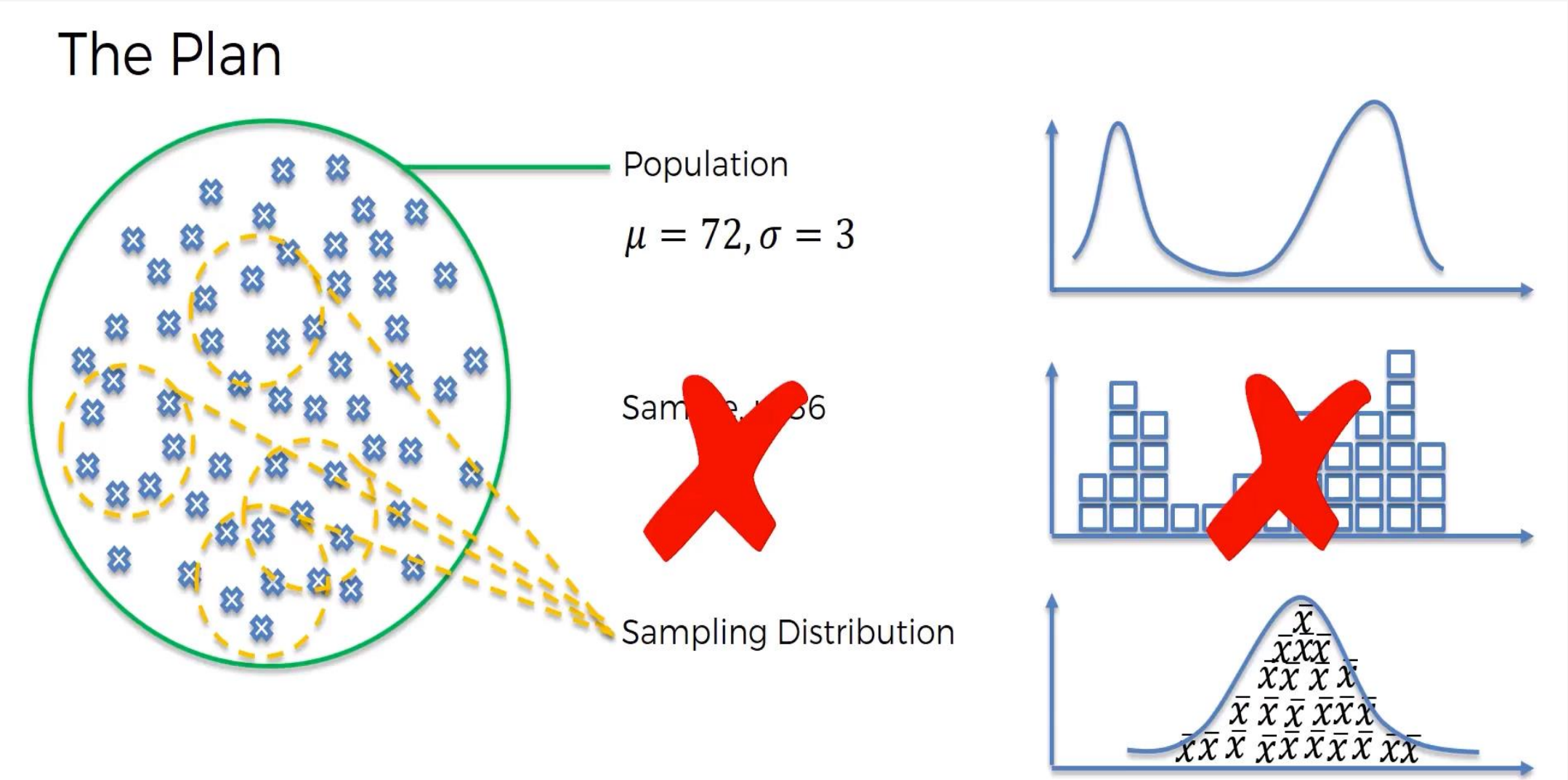
# *ANALYTICS CHALLENGE*

The only plane you currently have at Denver is a Cessna 208B Grand Caravan with a max cargo weight of 2,630 lb. (1,193 kg)

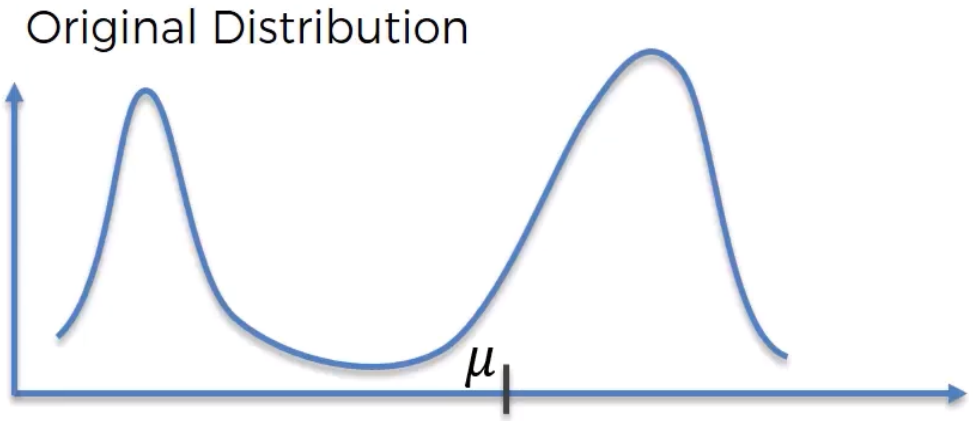
Based on this information what is the probability that all of the cargo can be safely loaded onto the plane and transported?

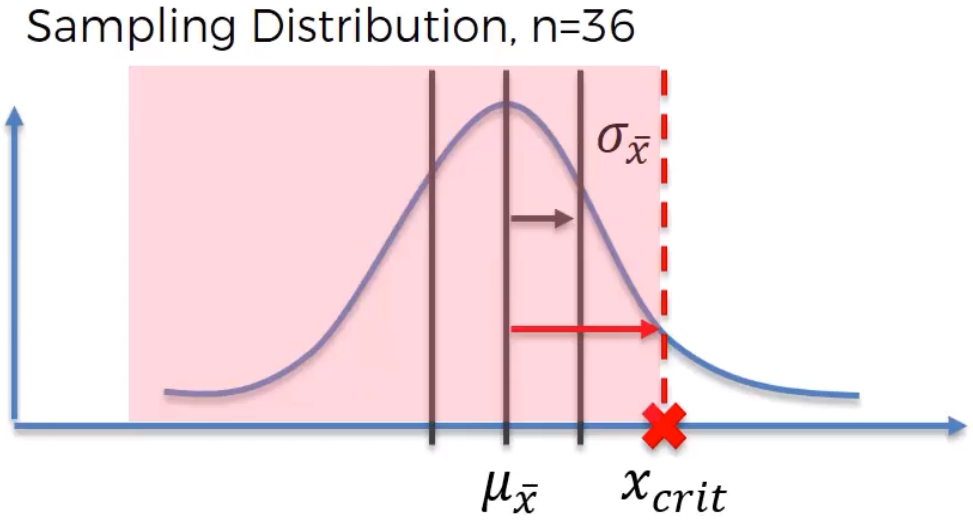


# ANALYTICS CHALLENGE



# ANALYTICS CHALLENGE




$$\mu_{\bar{x}} = \mu = 72$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{36}} = \frac{3}{6} = 0.5$$

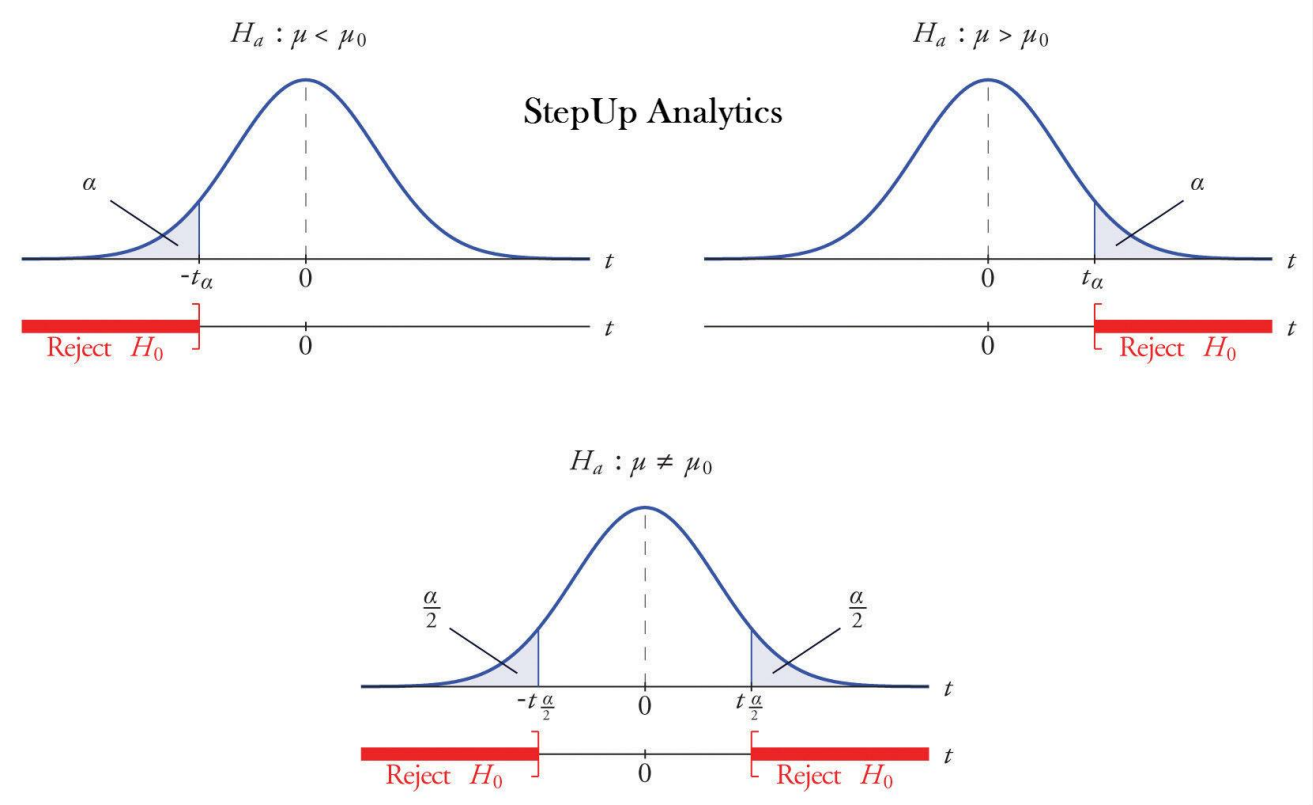
Plane Capacity = 2,630 lb.

$$x_{crit} = \frac{2{,}630 \text{ lb.}}{36\ boxes} = 73.06\ lb/box$$

$$z = \frac{x_{crit} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{73.06 - 72}{0.5} = 2.12$$

$$P(x < x_{crit}) = 0.9830 = 98.3\%$$

# *Section 3* Hypothesis Testing / Statistical Significance




***Data Science- Statistics***
***Ahmad Obaidat***

# WHAT WE WILL LEARN

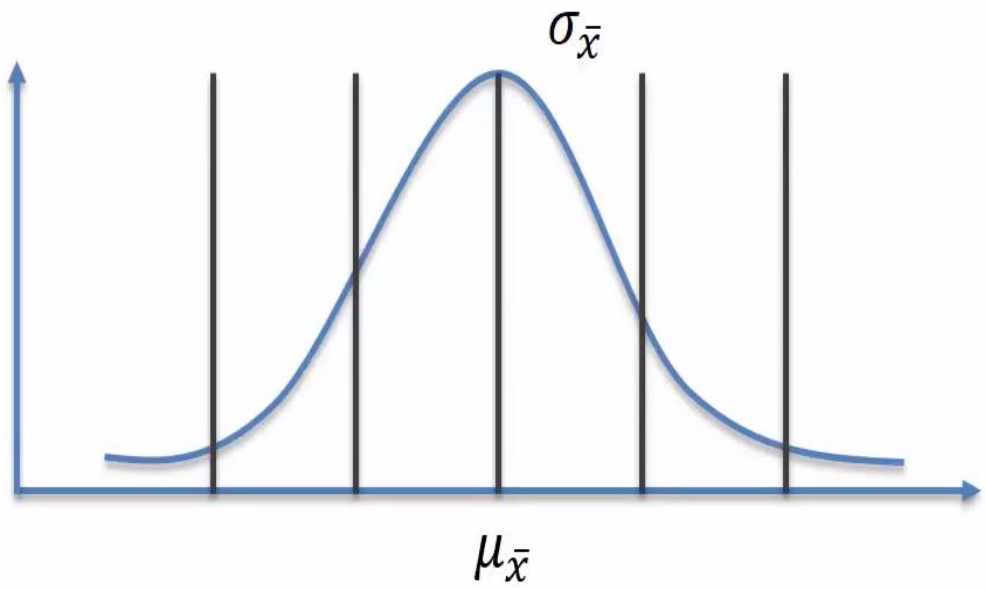
1. Hypothesis Testing – Steps
2. Statistical Significance
3. Hypothesis Testing – Rejection Region
4. Hypothesis Testing Assumptions
5. Proportion Testing

# HYPOTHESIS TESTING – STEPS

In 2015 millennials were watching 26.5 hours of TV a week with a std. dev. of 10 hrs. Today you surveyed 50 millennials and found that they watch 24 hours of TV per week. Has the parameter decreased?

$H_0$: It has not decreased, i.e. $\mu \geq 26.5$

$H_1$: It has decreased, i.e. $\mu < 26.5$

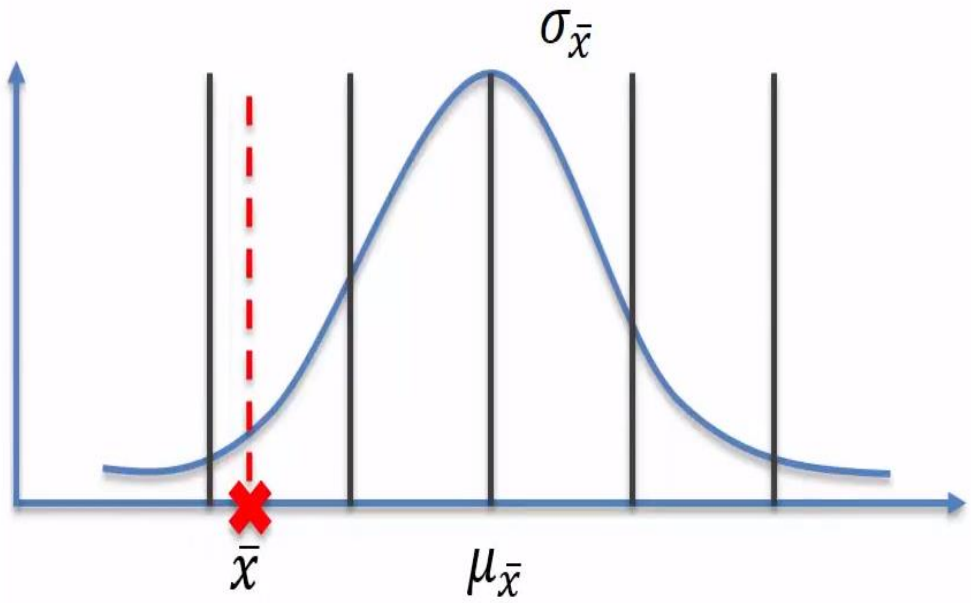


# HYPOTHESIS TESTING – STEPS

In 2015 millennials were watching 26.5 hours of TV a week with a std. dev. of 10 hrs. Today you surveyed 50 millennials and found that they watch 24 hours of TV per week. Has the parameter decreased?

$H_0$: It has not decreased, i.e. $\mu \geq 26.5$

$H_1$: It has decreased, i.e. $\mu < 26.5$




$$\mu_{\bar{x}} = \mu = 26.5$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{50}} = 1.41$$

$$\bar{x} = 24$$

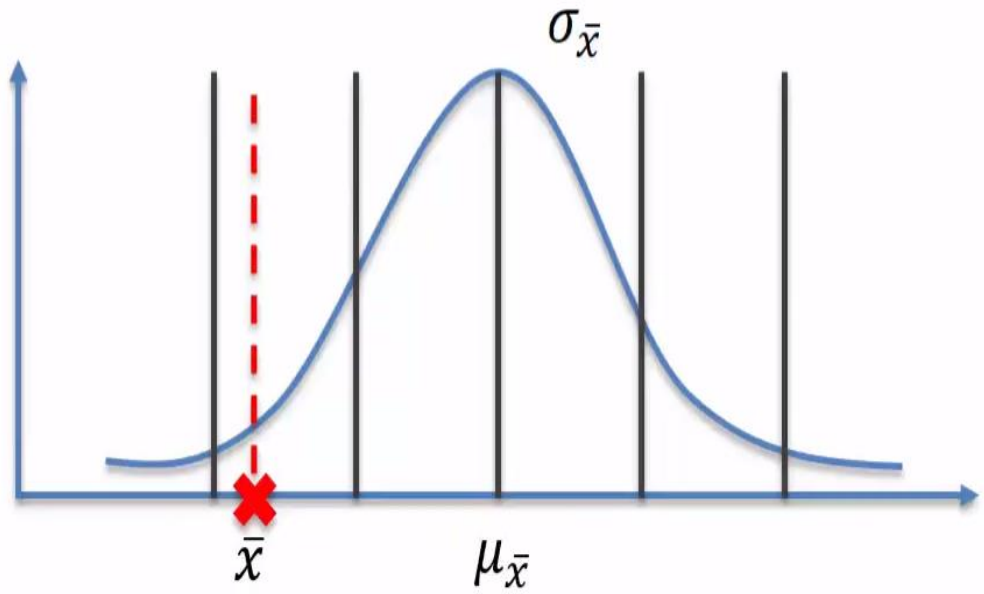
$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{24 - 26.5}{1.41} = -1.77$$

# HYPOTHESIS TESTING – STEPS

In 2015 millennials were watching 26.5 hours of TV a week with a std. dev. of 10 hrs. Today you surveyed 50 millennials and found that they watch 24 hours of TV per week. Has the parameter decreased?

$H_0$: It has not decreased, i.e. $\mu \geq 26.5$

$H_1$: It has decreased, i.e. $\mu < 26.5$




$$\mu_{\bar{x}} = \mu = 26.5$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{50}} = 1.41$$

$$\bar{x} = 24$$

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{24 - 26.5}{1.41} = -1.77$$

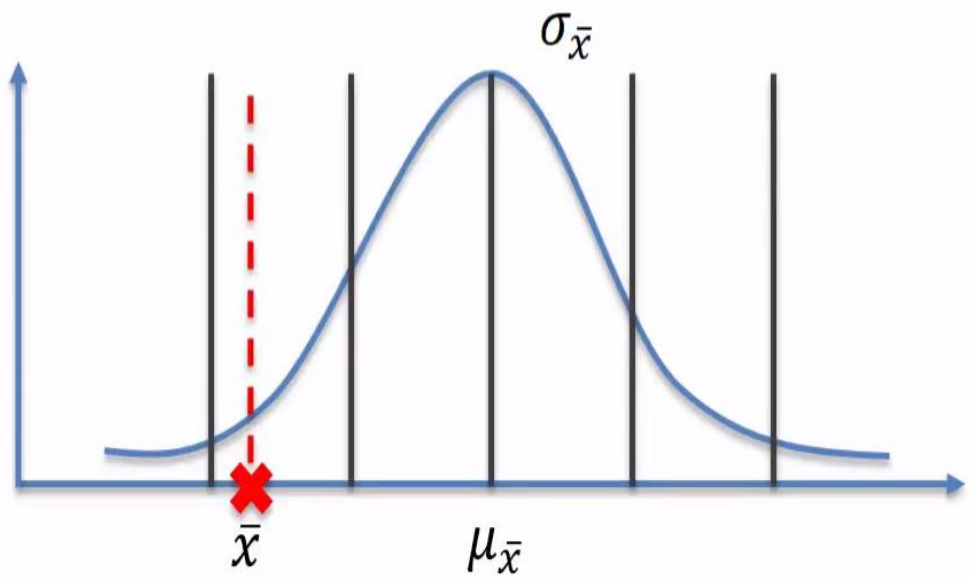
$$P = 0.0384 = 3.84\% < 5\%$$

# HYPOTHESIS TESTING – STEPS

In 2015 millennials were watching 26.5 hours of TV a week with a std. dev. of 10 hrs. Today you surveyed 50 millennials and found that they watch 24 hours of TV per week. Has the parameter decreased?

~~$H_0$: It has not decreased, i.e. $\mu \geq 26.5$~~

$H_1$: It has decreased, i.e. $\mu < 26.5$




$$\mu_{\bar{x}} = \mu = 26.5$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{50}} = 1.41$$

$$\bar{x} = 24$$

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{24 - 26.5}{1.41} = -1.77$$

$$P = 0.0384 = 3.84\% < 5\%$$

# *STATISTICAL SIGNIFICANCE*

## Coin Toss

# STATISTICAL SIGNIFICANCE

$H_0$: This is a fair coin
$H_1$: This is not a fair coin

# STATISTICAL SIGNIFICANCE

$H_0$: This is a fair coin
$H_1$: This is not a fair coin

 0.5

 0.25

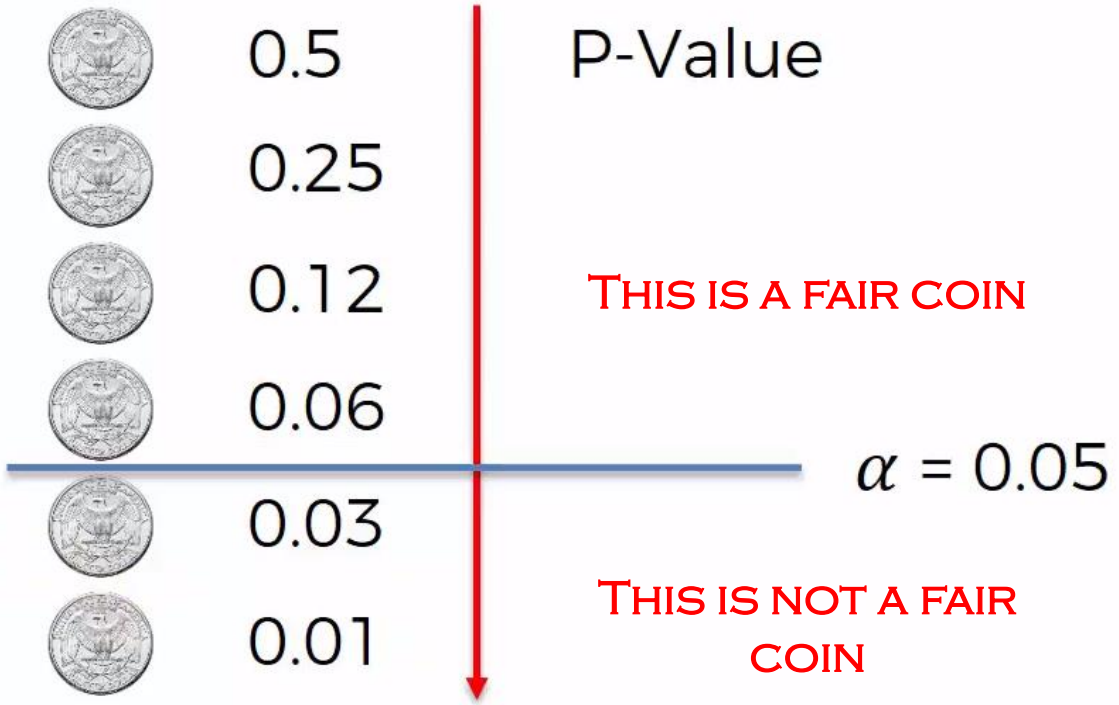
 0.12

 0.06

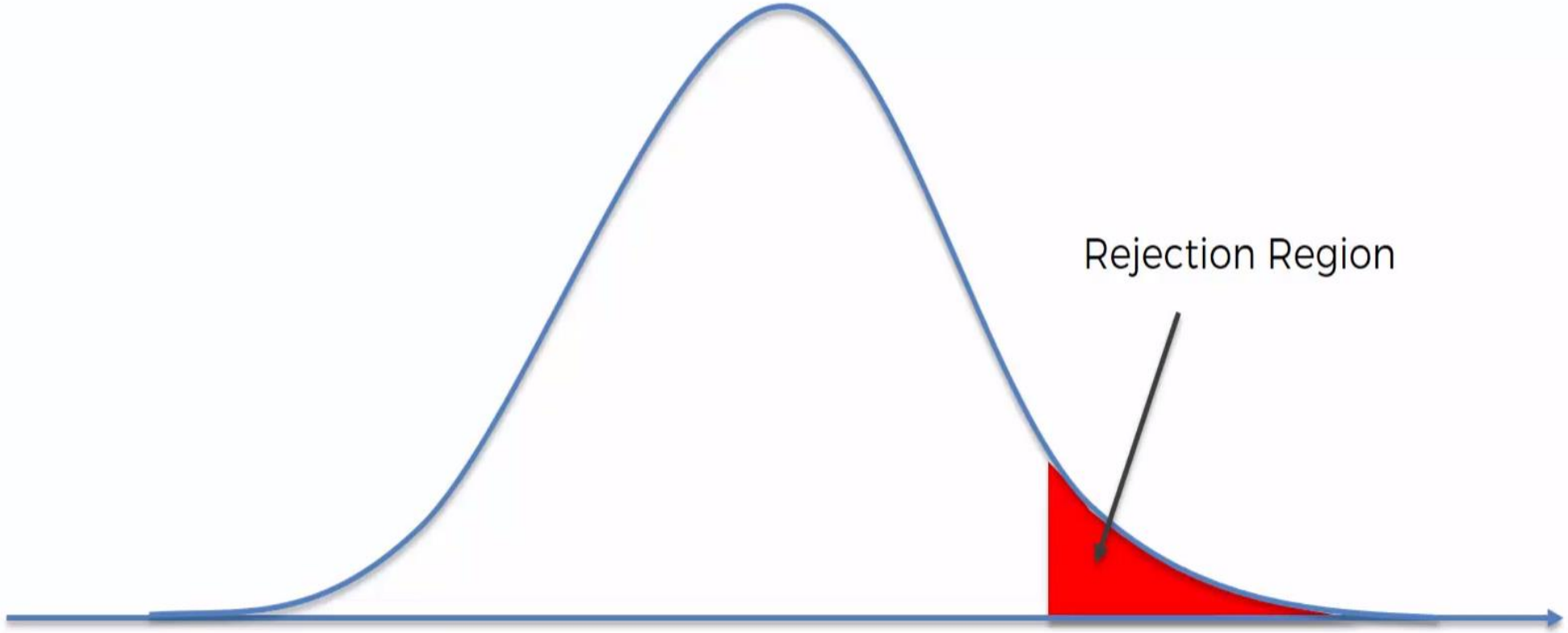
 0.03

# STATISTICAL SIGNIFICANCE

$H_0$: This is a fair coin
$H_1$: This is not a fair coin



# HYPOTHESIS TESTING – REJECTION REGION

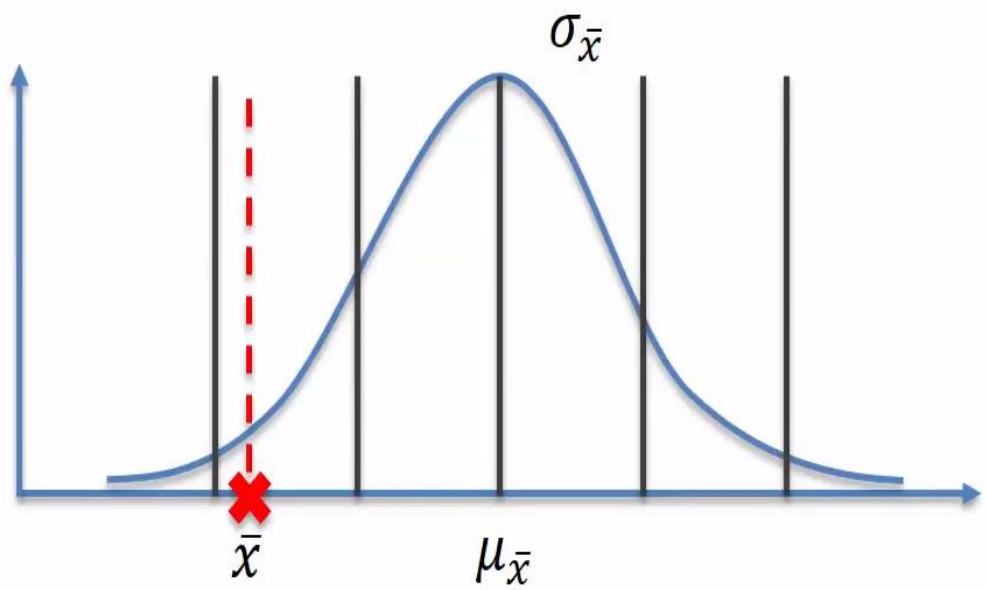


# HYPOTHESIS TESTING – REJECTION REGION

In 2015 millennials were watching 26.5 hours of TV a week with a std. dev. of 10 hrs. Today you surveyed 50 millennials and found that they watch 24 hours of TV per week. Has the parameter decreased?

$H_0$: It has not decreased, i.e. $\mu \geq 26.5$

$H_1$: It has decreased, i.e. $\mu < 26.5$




$$\mu_{\bar{x}} = \mu = 26.5$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{50}} = 1.41$$

$$\bar{x} = 24$$

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{24 - 26.5}{1.41} = -1.77$$

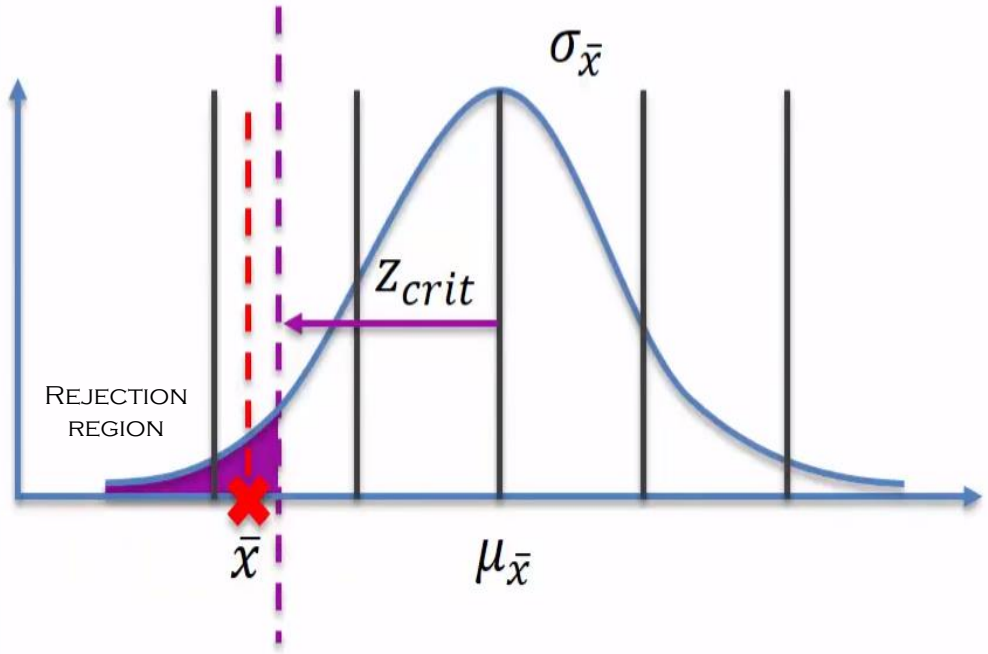
$$P_{crit} = 0.05 \Rightarrow z_{crit} = -1.65$$

# HYPOTHESIS TESTING – REJECTION REGION

In 2015 millennials were watching 26.5 hours of TV a week with a std. dev. of 10 hrs. Today you surveyed 50 millennials and found that they watch 24 hours of TV per week. Has the parameter decreased?

$H_0$: It has not decreased, i.e. $\mu \geq 26.5$

$H_1$: It has decreased, i.e. $\mu < 26.5$




$$\mu_{\bar{x}} = \mu = 26.5$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{50}} = 1.41$$

$$\bar{x} = 24$$

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{24 - 26.5}{1.41} = -1.77$$

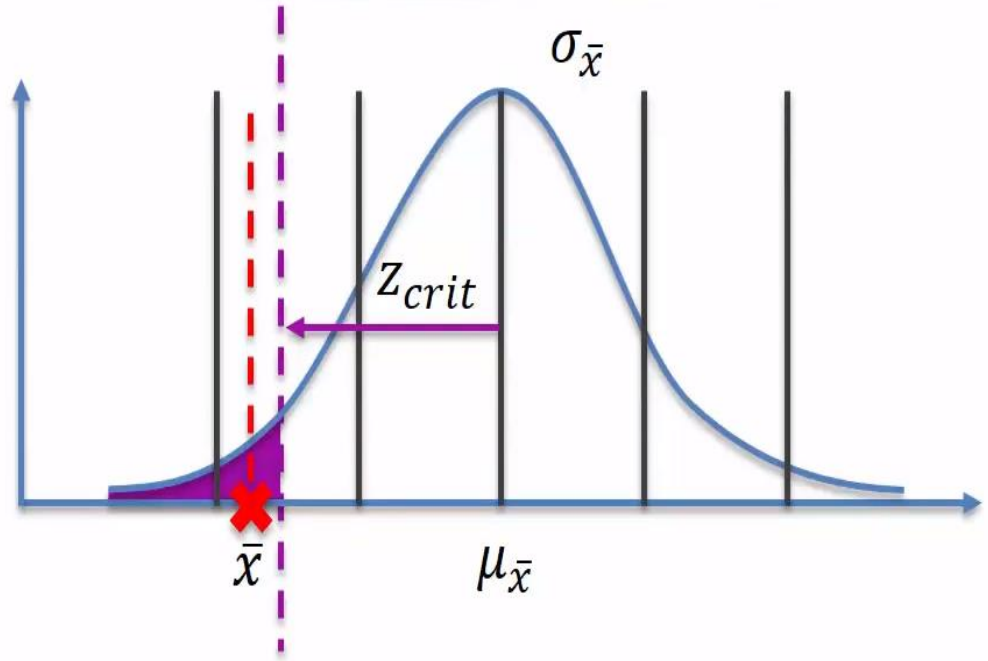
$$P_{crit} = 0.05 \Rightarrow z_{crit} = -1.65$$

# HYPOTHESIS TESTING – REJECTION REGION

In 2015 millennials were watching 26.5 hours of TV a week with a std. dev. of 10 hrs. Today you surveyed 50 millennials and found that they watch 24 hours of TV per week. Has the parameter decreased?

~~$H_0$: It has not decreased, i.e. $\mu > 26.5$~~

$H_1$: It has decreased, i.e. $\mu < 26.5$




$$\mu_{\bar{x}} = \mu = 26.5$$

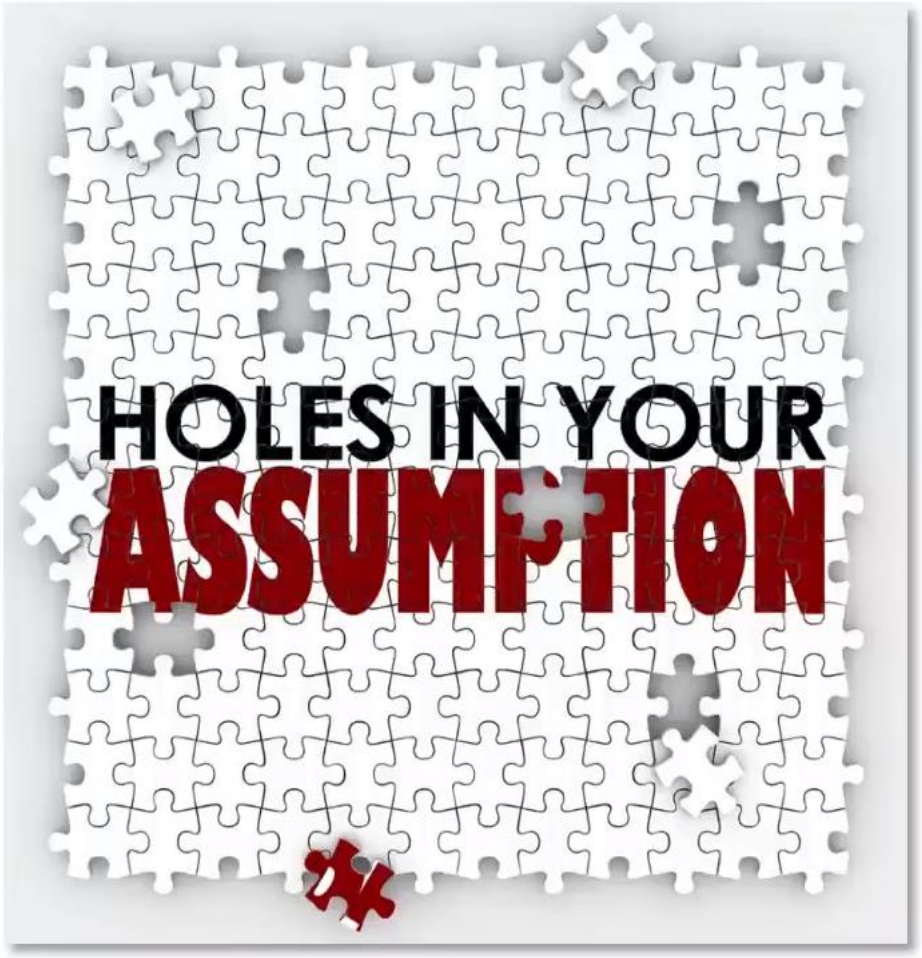
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{50}} = 1.41$$

$$\bar{x} = 24$$

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{24 - 26.5}{1.41} = -1.77$$

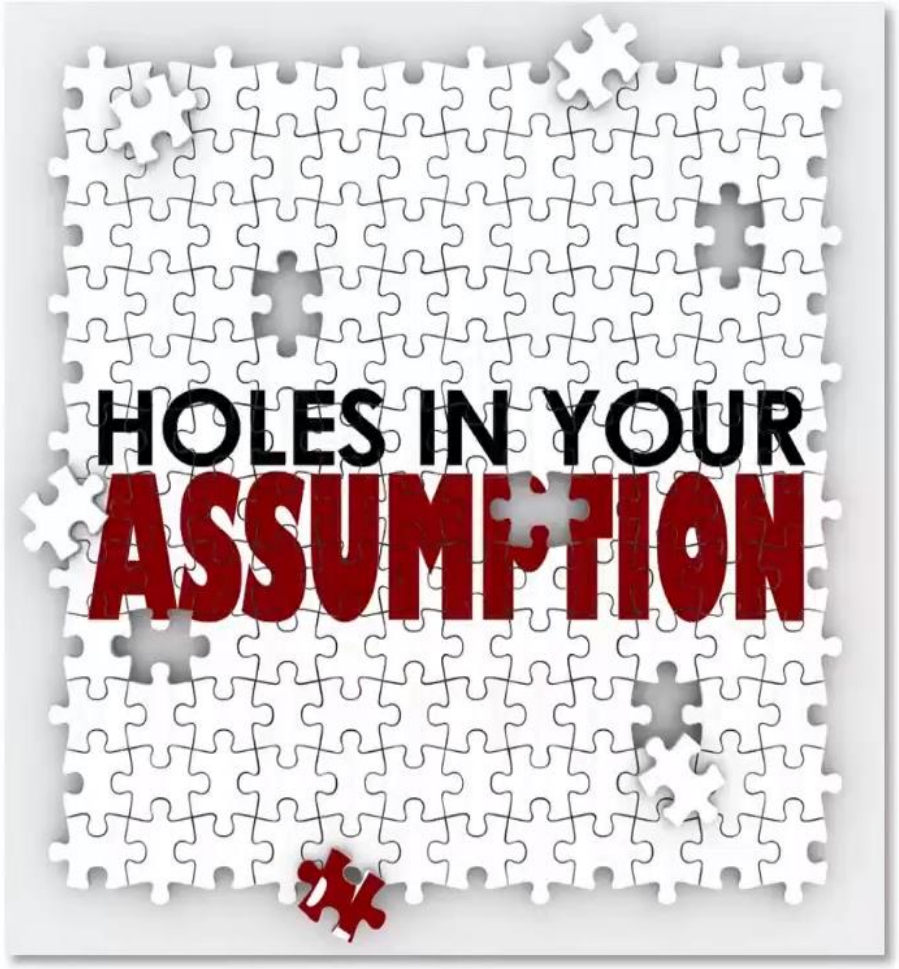
$$P_{crit} = 0.05 \Rightarrow z_{crit} = -1.65$$

# HYPOTHESIS TESTING ASSUMPTIONS




Z-Test Assumptions:

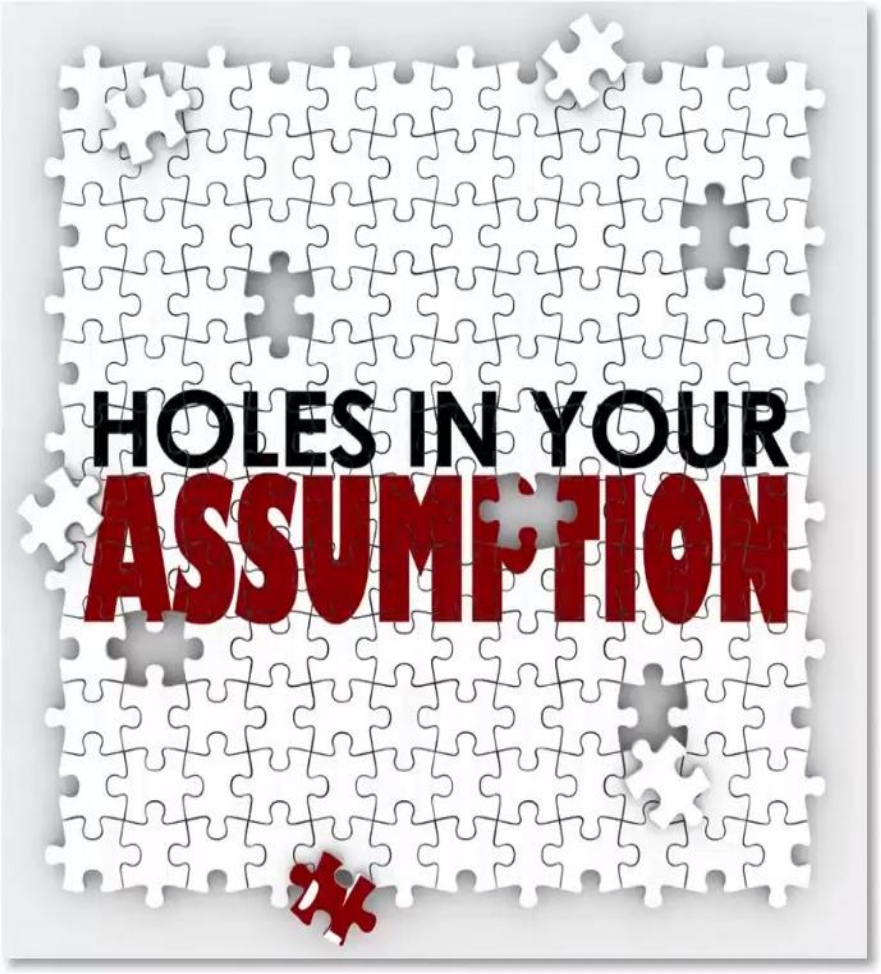
# HYPOTHESIS TESTING ASSUMPTIONS




Z-Test Assumptions:

1. Sample is selected at random

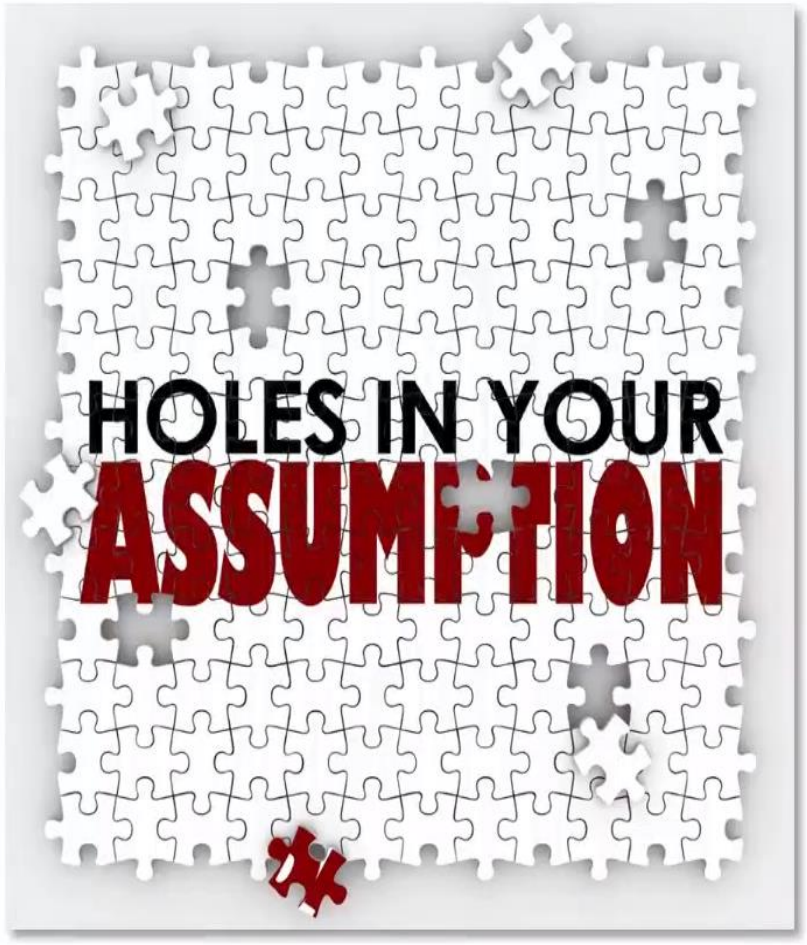
# HYPOTHESIS TESTING ASSUMPTIONS




Z-Test Assumptions:

1. Sample is selected at random
2. Observations are independent

# HYPOTHESIS TESTING ASSUMPTIONS



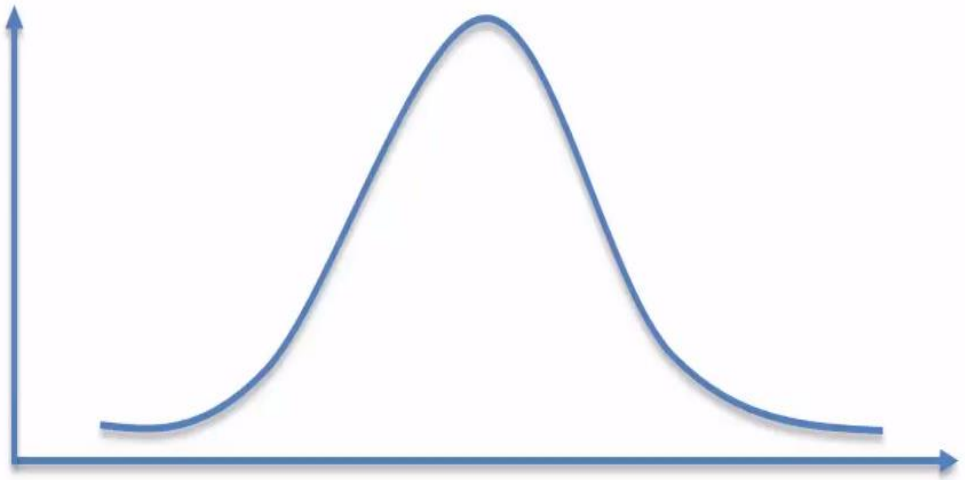

Z-Test Assumptions:

1. Sample is selected at random
2. Observations are independent
3. The population's standard deviation is known OR the sample contains at least 30 observations

OTHERWISE USE THE T-TEST

# *PROPORTION TESTING*

According to a 2016 survey, *"a staggering 58% of American households have tablets"*. Your manager has asked you to test the Hypothesis (with a 97% significance level) that only 58% of American homes have tablet devices. You surveyed 100 random households and found that 73 of them had tablets.

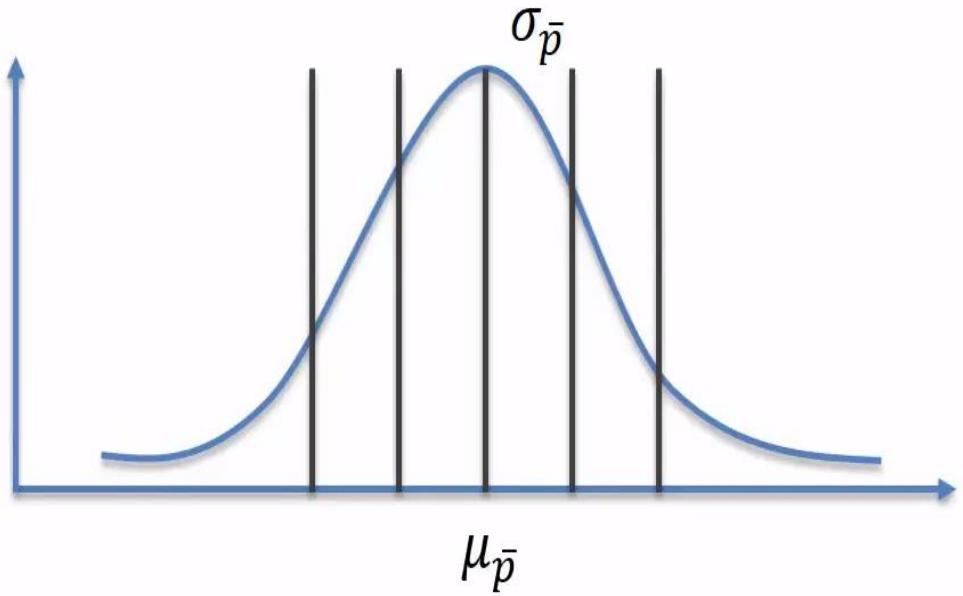


# PROPORTION TESTING

According to a 2016 survey, *"a staggering 58% of American households have tablets"*. Your manager has asked you to test the Hypothesis (with a 97% significance level) that only 58% of American homes have tablet devices. You surveyed 100 random households and found that 73 of them had tablets.

$H_0$: 58% of fewer hh have tablets, i.e. $p \leq 58\%$

$H_1$: More than 58% have tablets, i.e. $p > 58\%$

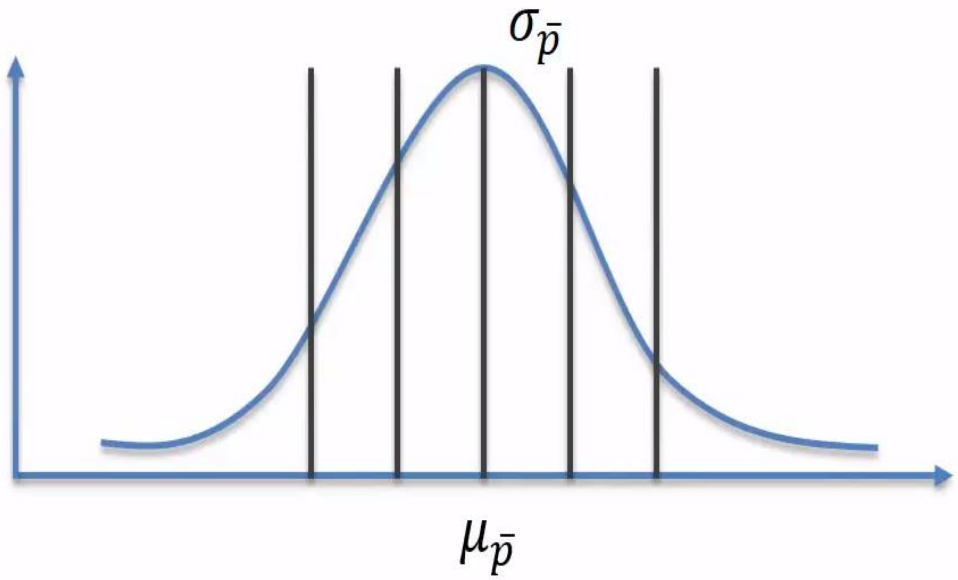


# PROPORTION TESTING

According to a 2016 survey, *"a staggering 58% of American households have tablets"*. Your manager has asked you to test the Hypothesis (with a 97% significance level) that only 58% of American homes have tablet devices. You surveyed 100 random households and found that 73 of them had tablets.

$H_0$: 58% of fewer hh have tablets, i.e. $p \leq 58\%$

$H_1$: More than 58% have tablets, i.e. $p > 58\%$




Check that:

- $n\bar{p} > 10$
- $n\bar{q} > 10$

q =1- p hat

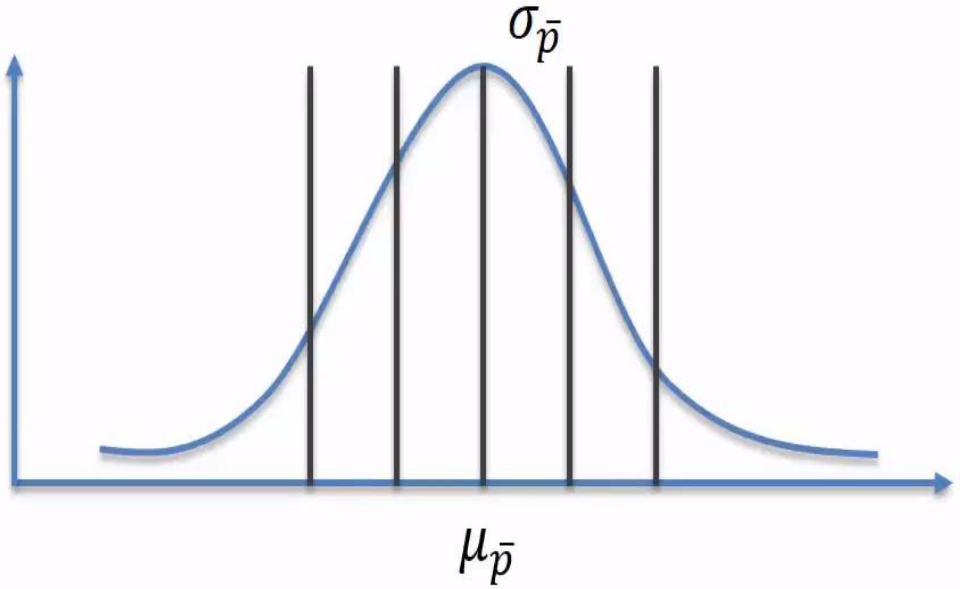
In case the check failed increase the random sample

# PROPORTION TESTING

According to a 2016 survey, *"a staggering 58% of American households have tablets"*. Your manager has asked you to test the Hypothesis (with a 97% significance level) that only 58% of American homes have tablet devices. You surveyed 100 random households and found that 73 of them had tablets.

$H_0$: 58% of fewer hh have tablets, i.e. $p \leq 58\%$

$H_1$: More than 58% have tablets, i.e. $p > 58\%$




Check that:

- $n\bar{p} > 10$
- $n\bar{q} > 10$

New formulas:

$$\mu = p$$

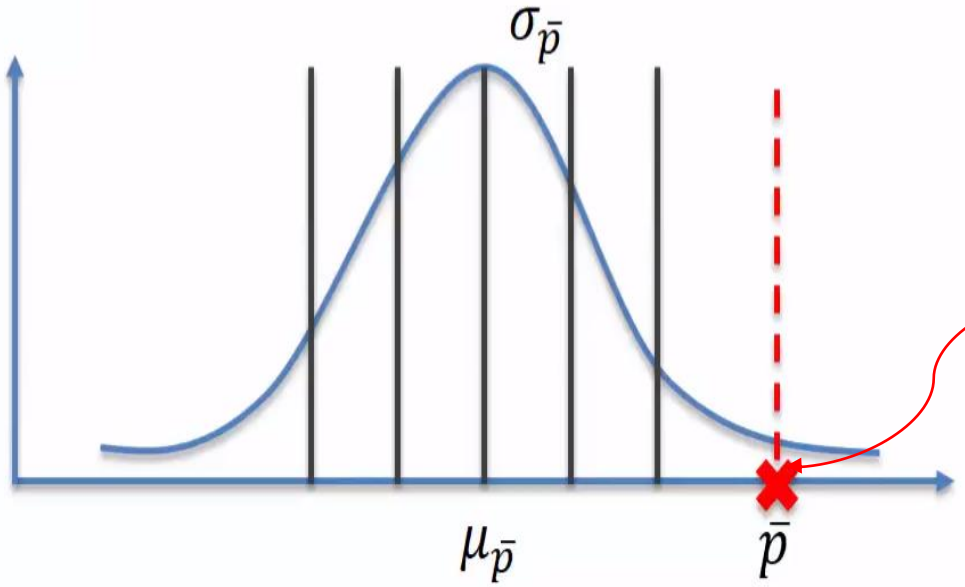
$$\sigma = \sqrt{pq}$$

# PROPORTION TESTING

Your manager has asked you to test the Hypothesis (with a 97% significance level) that only 58% of American homes have tablet devices. You surveyed 100 random households and found that 73 of them had tablets.

$H_0$: 58% of fewer hh have tablets, i.e. $p \leq 58\%$

$H_1$: More than 58% have tablets, i.e. $p > 58\%$




$$\mu_{\bar{p}} = \mu = p = 0.58$$

$$\sigma = \sqrt{pq} = \sqrt{0.58 * 0.42} = 0.49$$

$$\sigma_{\bar{p}} = \frac{\sigma}{\sqrt{n}} = \frac{0.49}{\sqrt{100}} = 0.049$$

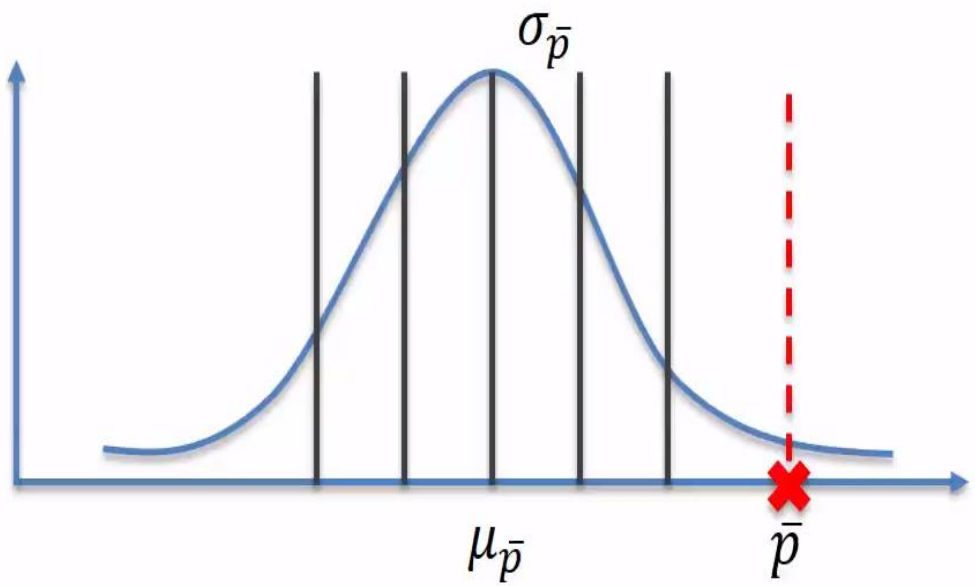
$$\bar{p} = \frac{73}{100} = 0.73$$

# PROPORTION TESTING

Your manager has asked you to test the Hypothesis (with a 97% significance level) that only 58% of American homes have tablet devices. You surveyed 100 random households and found that 73 of them had tablets.

~~$H_0$: 58% of fewer HH have tablets, i.e. $p < 58\%$~~

$H_1$: More than 58% have tablets, i.e. $p > 58\%$




$$\mu_{\bar{p}} = \mu = p = 0.58$$

$$\sigma = \sqrt{pq} = \sqrt{0.58 * 0.42} = 0.49$$

$$\sigma_{\bar{p}} = \frac{\sigma}{\sqrt{n}} = \frac{0.49}{\sqrt{100}} = 0.049$$

$$\bar{p} = \frac{73}{100} = 0.73$$

$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{0.73 - 0.58}{0.049} = 3.06$$

$$P = 1 - 0.9989 = 0.11\% < 3\%$$

# *CHALLENGE*

You are consultant engaged by a factory which manufactures spoons.

The factory executives recently spent $10,000,000 upgrading equipment and Processes in order to combat excessively high defects in manufacturing (23%) which were leading to high return rates from clients.

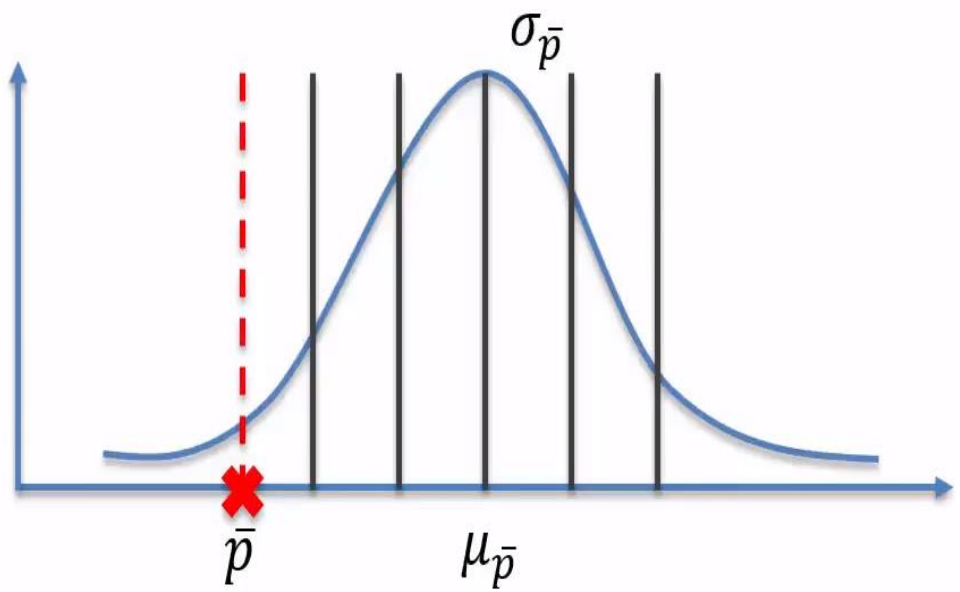
You have been asked to prove (with a confidence level of 95%) that new equipment has improved the situation and that the number of defective spoons has decreased to under 18% you have been supplied with a random sample of 150 spoons and found that 23 spoons have defects

# CHALLENGE

You have been asked to prove (with a confidence level of 95%) that … the number of defective spoons has decreased to under 18%. You have been supplied with a random sample of 150 spoons and found that 23 spoons have defects.

$H_0$: 18% or more have defects, i.e. $p \geq 18\%$

$H_1$: Less than 18% have defects, i.e. $p < 18\%$




$$\mu_{\bar{p}} = \mu = p = 0.18$$

$$\sigma = \sqrt{pq} = \sqrt{0.18 * 0.82} = 0.384$$

$$\sigma_{\bar{p}} = \frac{\sigma}{\sqrt{n}} = \frac{0.384}{\sqrt{150}} = 0.031$$

$$\bar{p} = \frac{23}{150} = 0.153$$

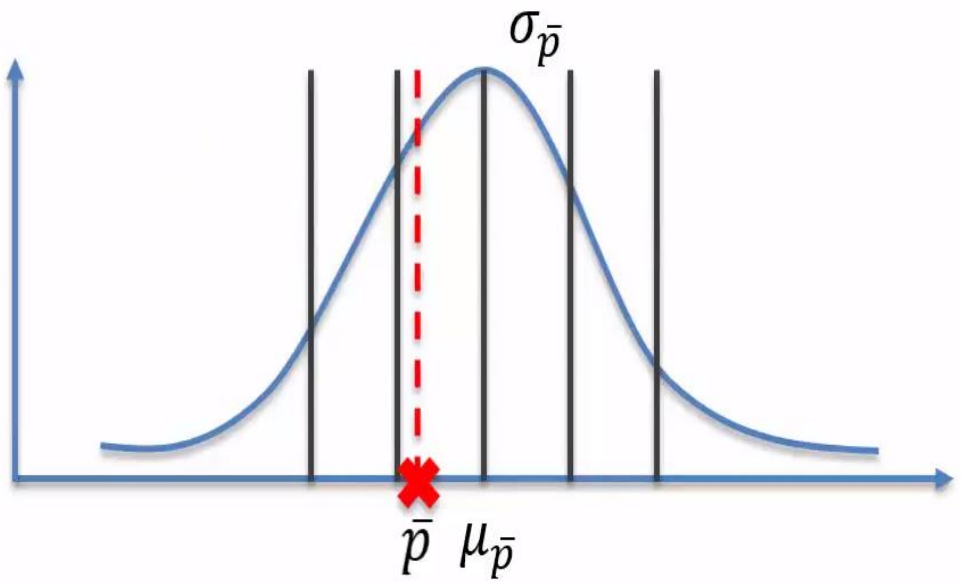
$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{0.153 - 0.18}{0.031} = -0.87$$

# CHALLENGE

You have been asked to prove (with a confidence level of 95%) that … the number of defective spoons has decreased to under 18%. You have been supplied with a random sample of 150 spoons and found that 23 spoons have defects.

$H_0$: 18% or more have defects, i.e. $p \geq 18\%$

$H_1$: Less than 18% have defects, i.e. $p < 18\%$




$$\mu_{\bar{p}} = \mu = p = 0.18$$

$$\sigma = \sqrt{pq} = \sqrt{0.18 * 0.82} = 0.384$$

$$\sigma_{\bar{p}} = \frac{\sigma}{\sqrt{n}} = \frac{0.384}{\sqrt{150}} = 0.031$$

$$\bar{p} = \frac{23}{150} = 0.153$$

$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{0.153 - 0.18}{0.031} = -0.87$$

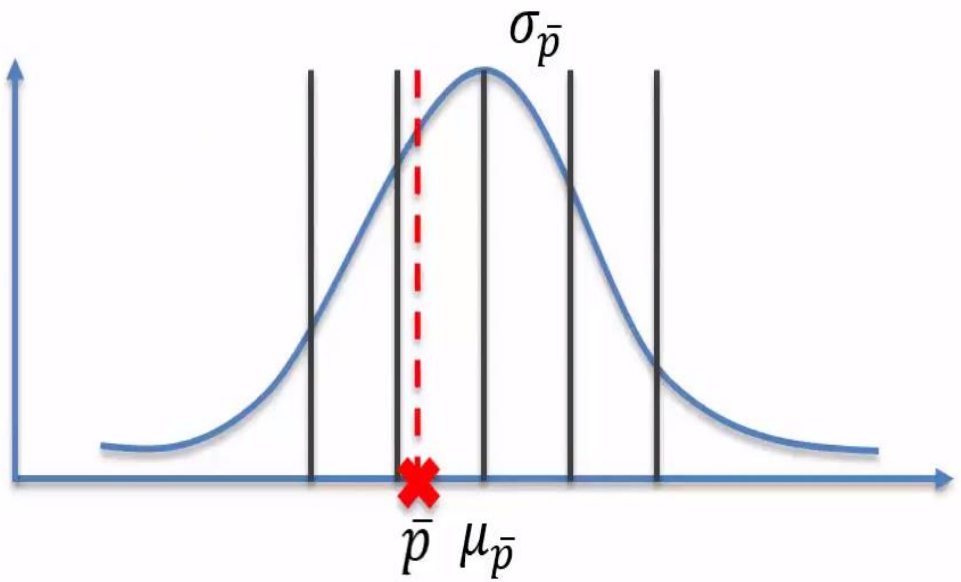
$$P = 0.1922 = 19.22\% > 5\%$$

# CHALLENGE

You have been asked to prove (with a confidence level of 95%) that … the number of defective spoons has decreased to under 18%. You have been supplied with a random sample of 150 spoons and found that 23 spoons have defects.

$H_0$: 18% or more have defects, i.e. $p \geq 18\%$

$H_1$: Less than 18% have defects, i.e. $p < 18\%$




$$\mu_{\bar{p}} = \mu = p = 0.18$$

$$\sigma = \sqrt{pq} = \sqrt{0.18 * 0.82} = 0.384$$

$$\sigma_{\bar{p}} = \frac{\sigma}{\sqrt{n}} = \frac{0.384}{\sqrt{150}} = 0.031$$

$$\bar{p} = \frac{23}{150} = 0.153$$

$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{0.153 - 0.18}{0.031} = -0.87$$

$$P = 0.1922 = 19.22\% > 5\%$$

# DATA SCIENCE

## CHAPTER 1

BY AHMAD OBAIDAT

# STATISTICS