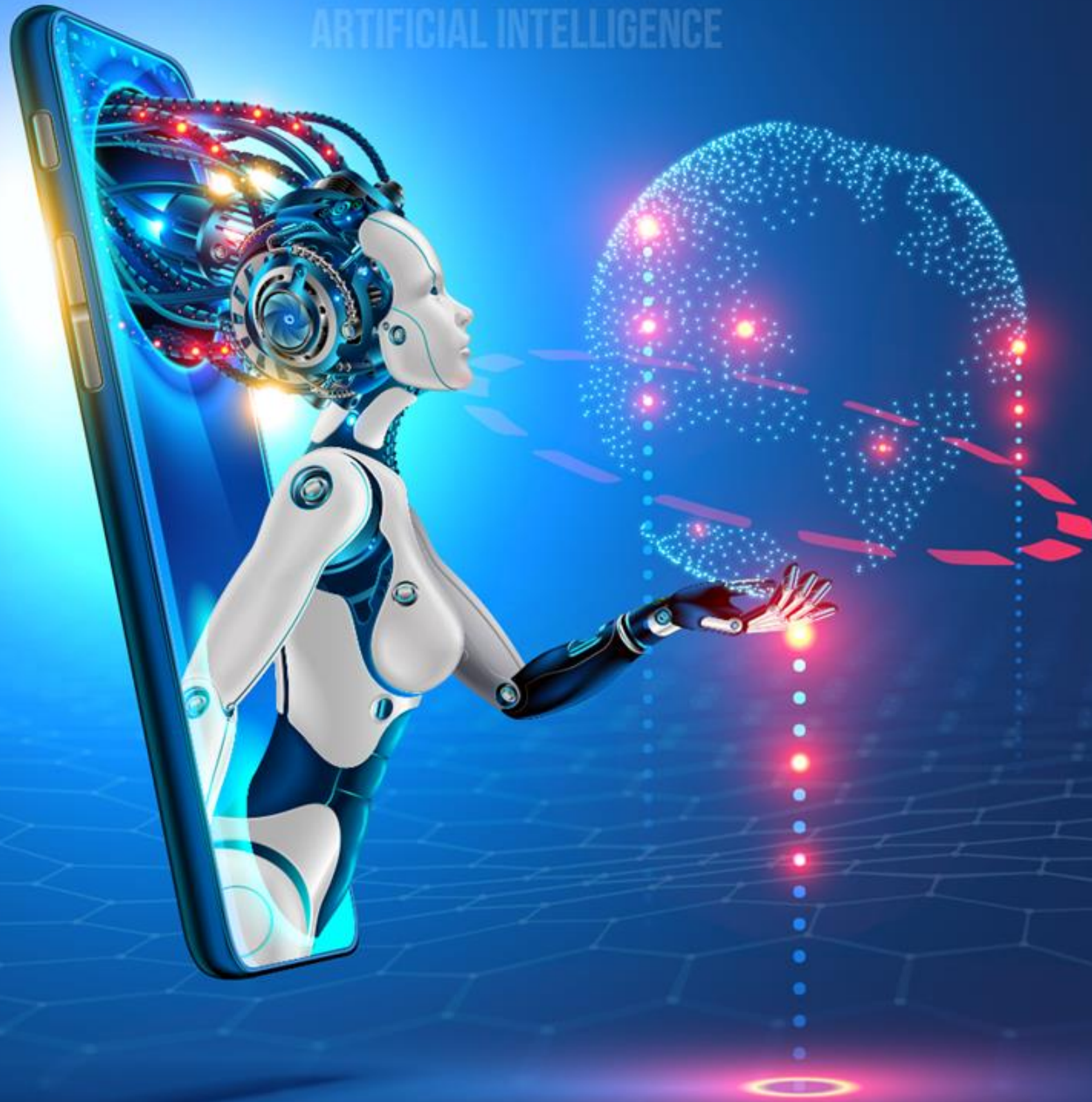


DATA AND
ARTIFICIAL INTELLIGENCE



Programming Basics and Data Analytics with Python

DATA AND ARTIFICIAL INTELLIGENCE



Statistical Computing

Learning Objectives

By the end of this lesson, you will be able to:

- 🕒 Interpret probability distributions
- 🕒 Perform hypothesis testing using z-scores
- 🕒 Infer distributions with respect to interval estimate
- 🕒 Perform A/B testing
- 🕒 Optimize your pages using results from A/B test



Statistics

Introduction to Statistics

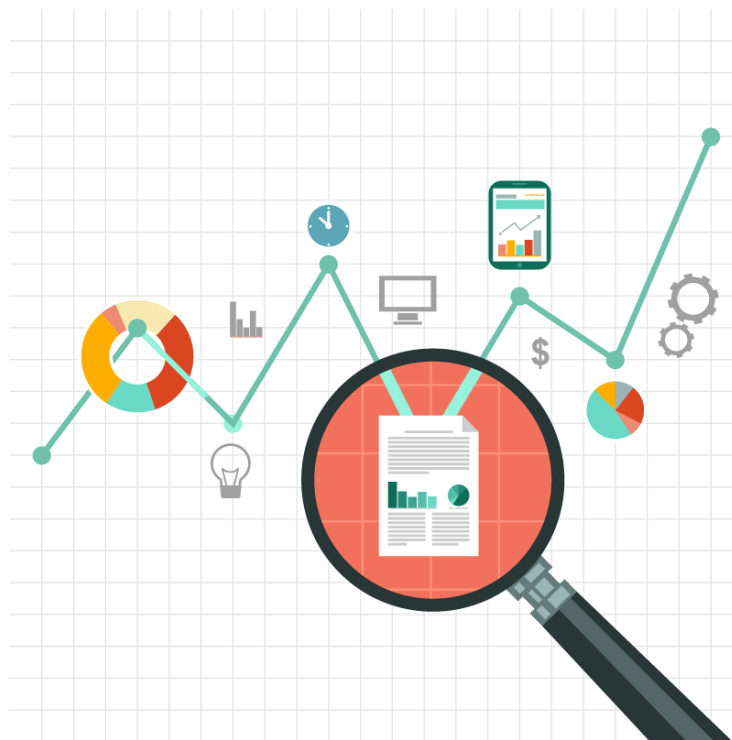
Statistics is the study of collection, analysis, interpretation, presentation, and organization of data.



Introduction to Statistics

Techniques available to analyze data:

- Statistical principles
- Functions
- Algorithms



What you can do using statistical tools:

- Analyze the primary data
- Build a statistical model
- Predict the future outcome

Statistical and Nonstatistical Analysis

Statistical Analysis



Statistical analysis is:

- scientific
- based on numbers or statistical values
- useful in providing complete insight of the data

Nonstatistical Analysis



Nonstatistical analysis is:

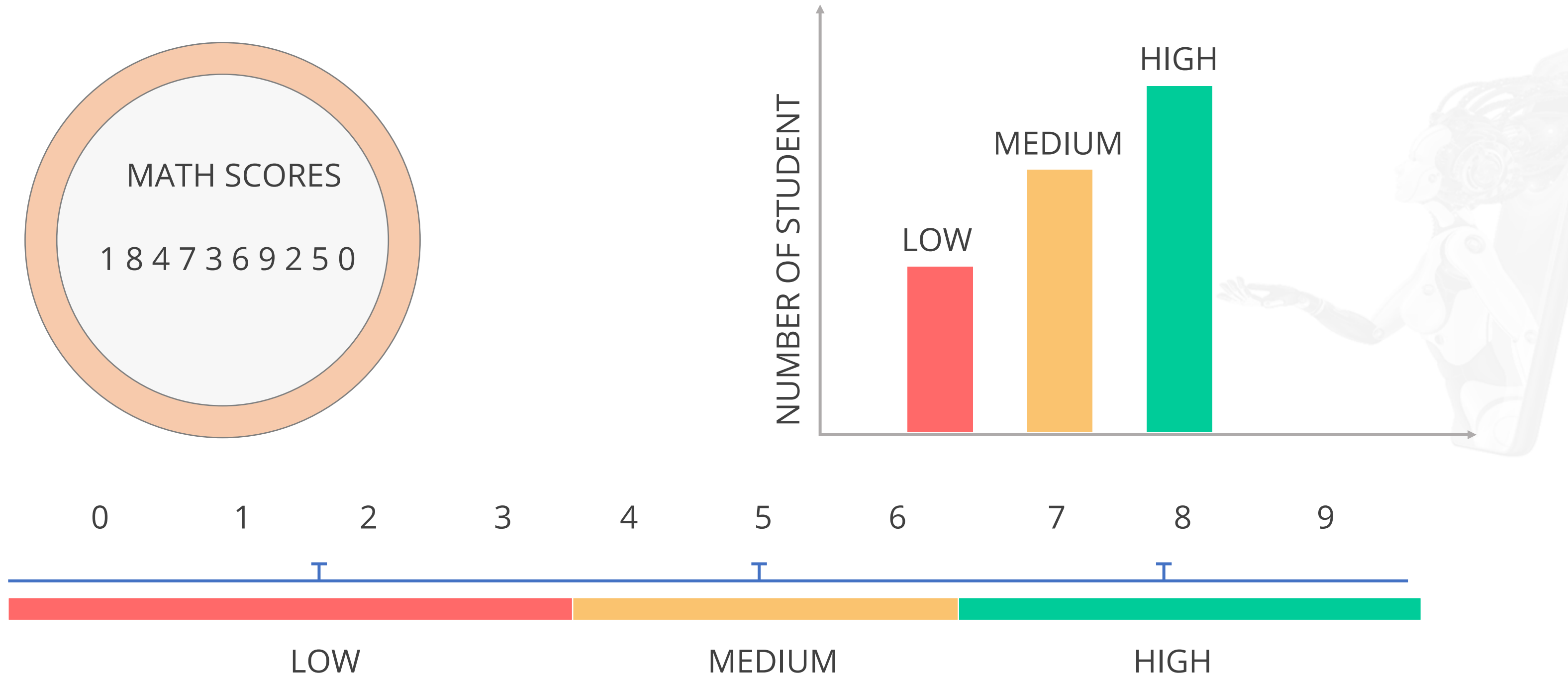
- based on generic information
- exclusive of statistical or quantitative analysis



Although both forms of analysis provide results, quantitative analysis provides more insight and a clearer picture. This is why statistical analysis is important for business.

Major Categories of Statistics

There are two major categories of statistics: descriptive analytics and inferential analytics. Descriptive analytics organizes the data and focuses on the main characteristics of the data.



Major Categories of Statistics: Inferential Analytics

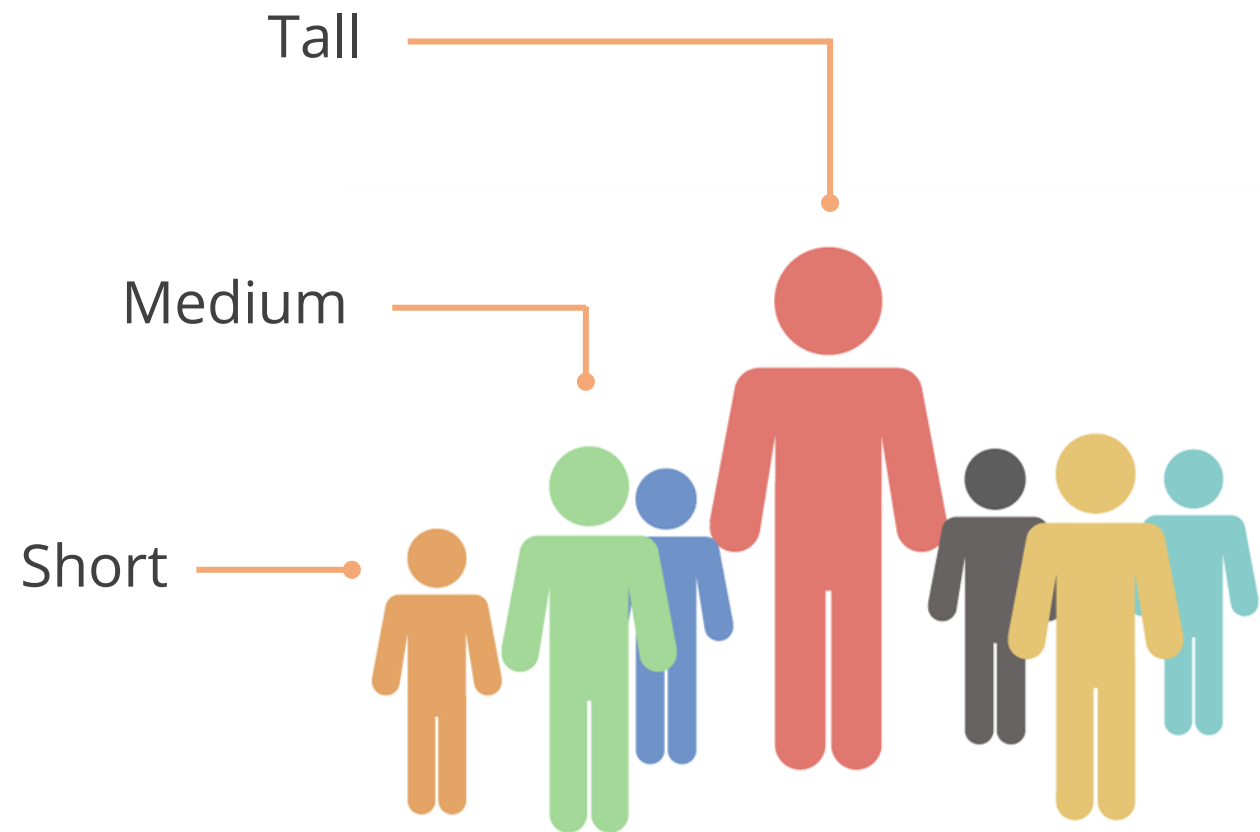
Inferential analytics is valuable when it is not possible to examine each member of the population.



- Random sample is drawn from the population
- Used to describe and make inferences about the population

Major Categories of Statistics: Example

Study of height in the population



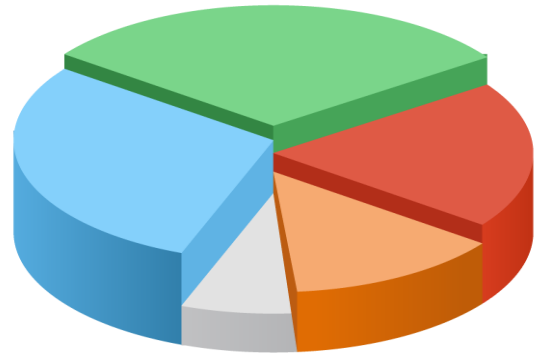
Inferential Method

Categorize height as tall, medium, and short. Take a sample from the population to study.

Descriptive Method

Record the height of each and every person. Provide the tallest, shortest, and average height of the population.

Statistical Analysis Considerations



Purpose

Make it clear and well-defined



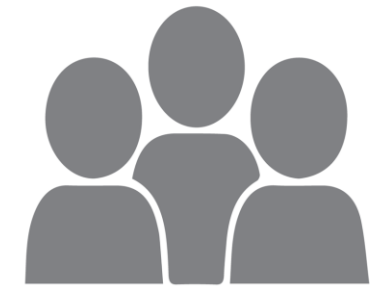
Document Questions

Prepare a questionnaire in advance



Define Population of Interest

Select population based on the purpose of analysis



Determine Sample

Select based on the purpose of study

Probability Density Function

Probability: Definition

Ratio of number of desired outcome to number of total possible outcomes.



Consider rolling a die:

$$P(\text{getting any of the outcome}) = 1/6$$



Probability: Sums to 1

The probability of all outcomes always sums to 1.




Consider rolling a die:
 $P(\text{all possible outcomes}) =$
 $1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1$

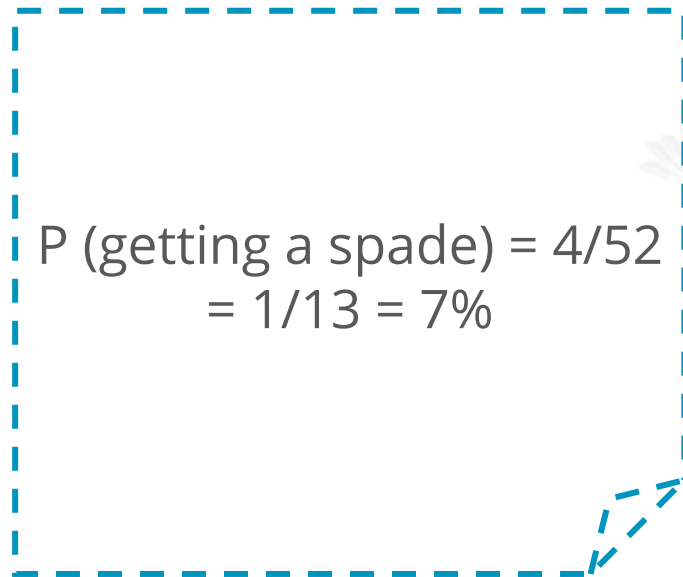


Probability: Use Case

A gambler wants to find out the occurrence of a spade in a deck of cards.

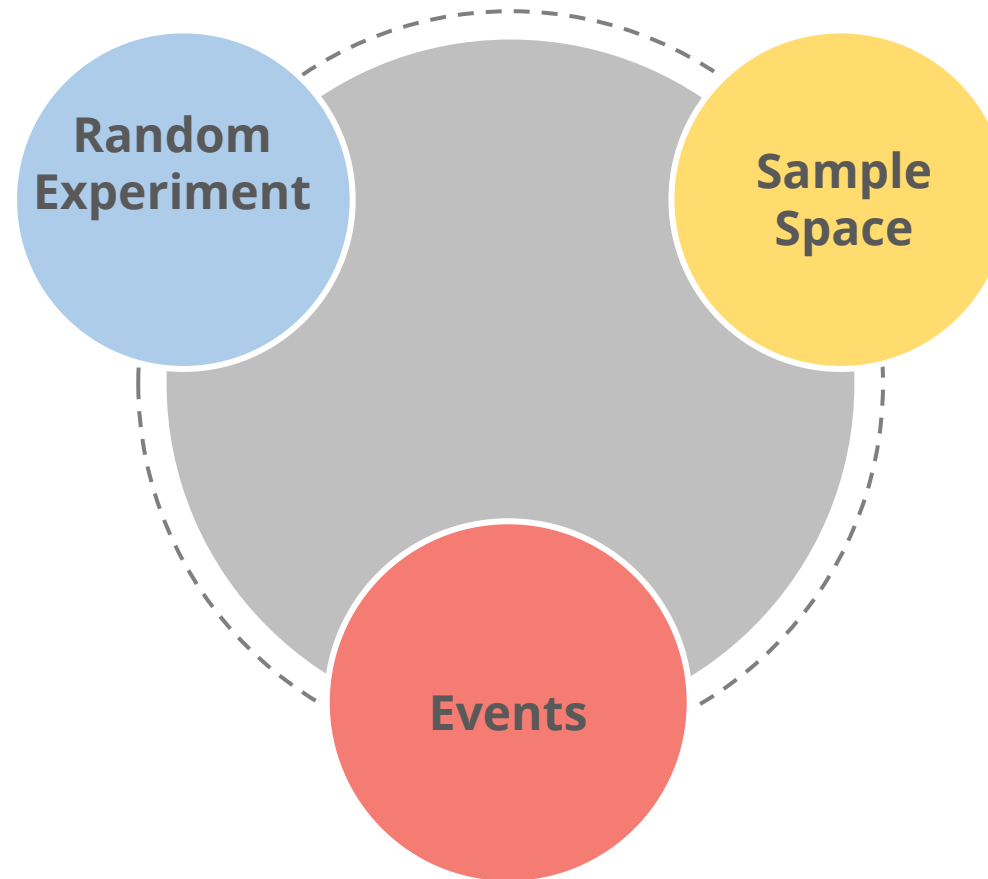


What is the chance that I will get a spade?


$$\begin{aligned} P(\text{getting a spade}) &= 4/52 \\ &= 1/13 = 7\% \end{aligned}$$

Probability: Terminologies

A process conducted to determine occurrence of outcomes, for example, rolling a die



A set of all possible outcomes, denoted by S , for example, 1 to 6 in case of die roll

A set of outcomes of an experiment, which is a subset of S



Probability: Rules

If an event E is a subset of a sample space S , then the following are true:

- $0 \leq P(E) \leq 1$ The probability of an event E is between 0 and 1 inclusive
- $P(\emptyset) = 0$ The probability of an empty set is zero
- $P(S)=1$ The probability of the sample space is 1
- The rule for unions is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If A and B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$
- The rule for complimentary events is $P(E') = 1 - P(E)$

Probability Distribution Function: Use Case

Consider that we have tabulated the salaries of a group of employees in a company and we want to create a frequency distribution table out of it.

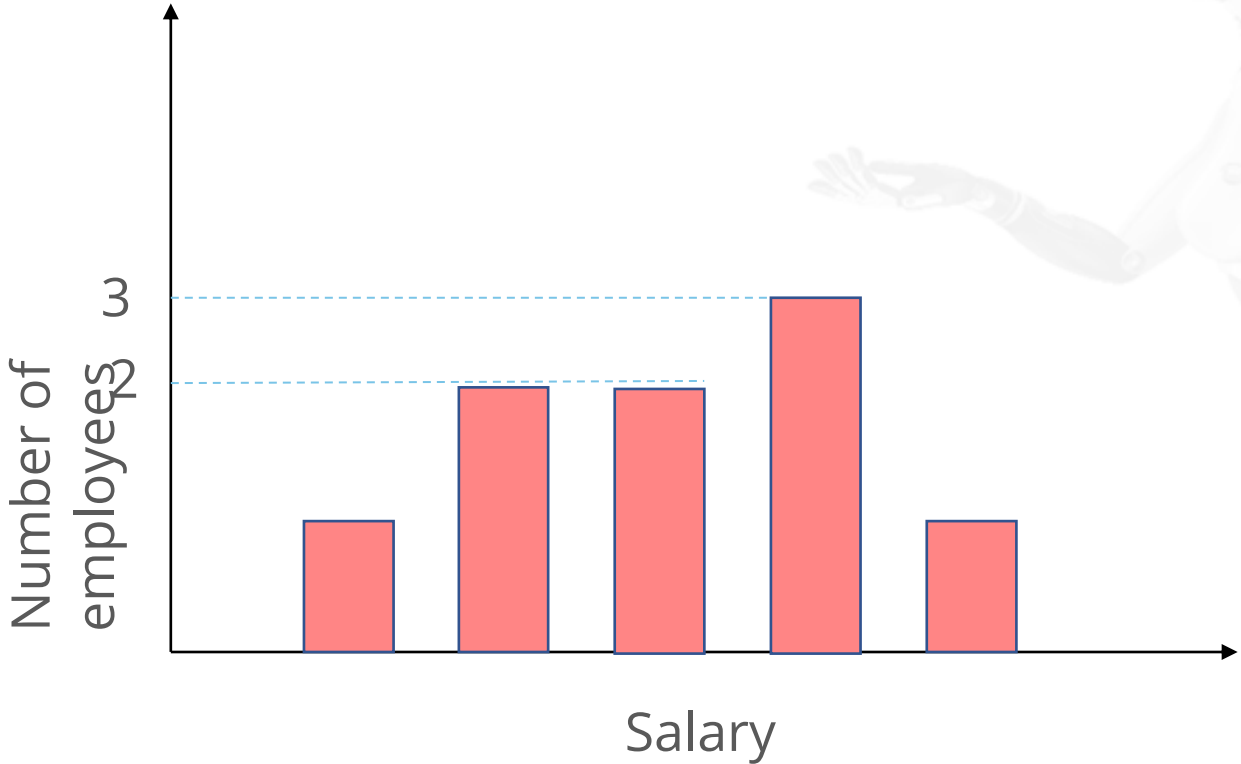


Salary (\$k)	ID
100	1
100	2
50	3
300	4
300	5
300	6
200	7
200	8
400	9

Probability Distribution Function: Use Case

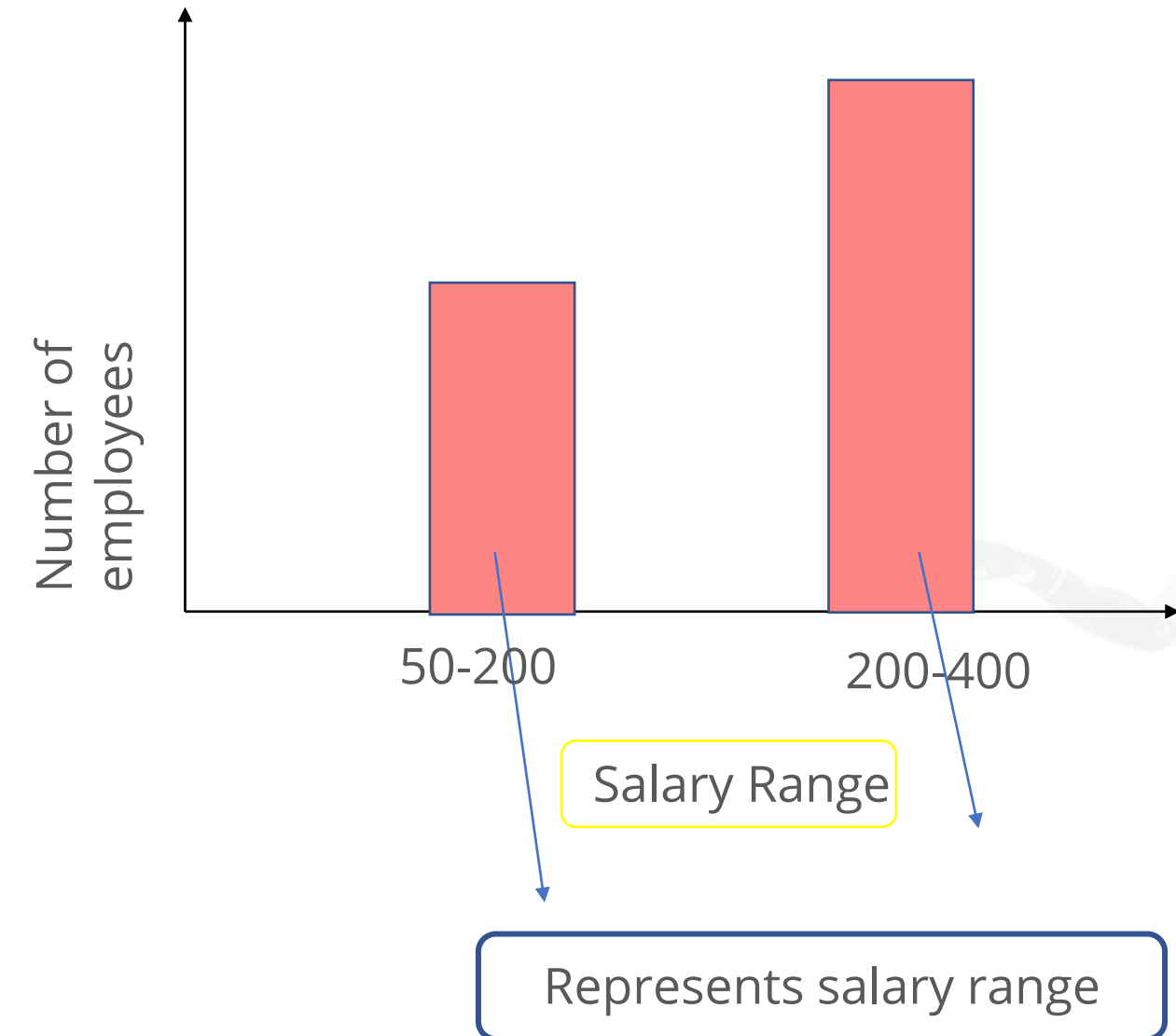
Plotting a histogram out of a frequency distribution table.

Salary (\$k)	No. of employees
50	1
100	2
200	2
300	3
400	1



Probability Distribution Function: Use Case

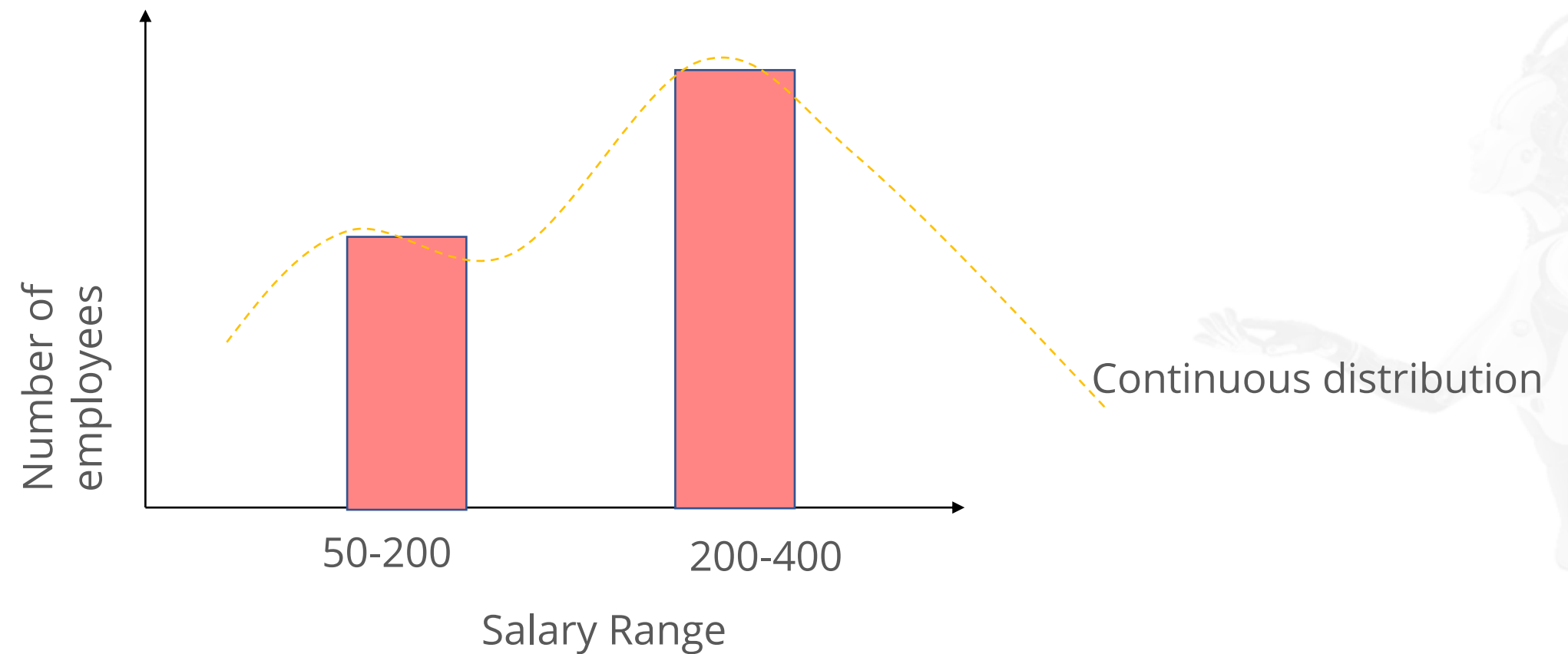
On a larger scale, this distribution can be substituted with a binned distribution.



Note: Binned distribution doesn't allow you to point out the salary of a particular employee.

Probability Distribution Function: Use Case

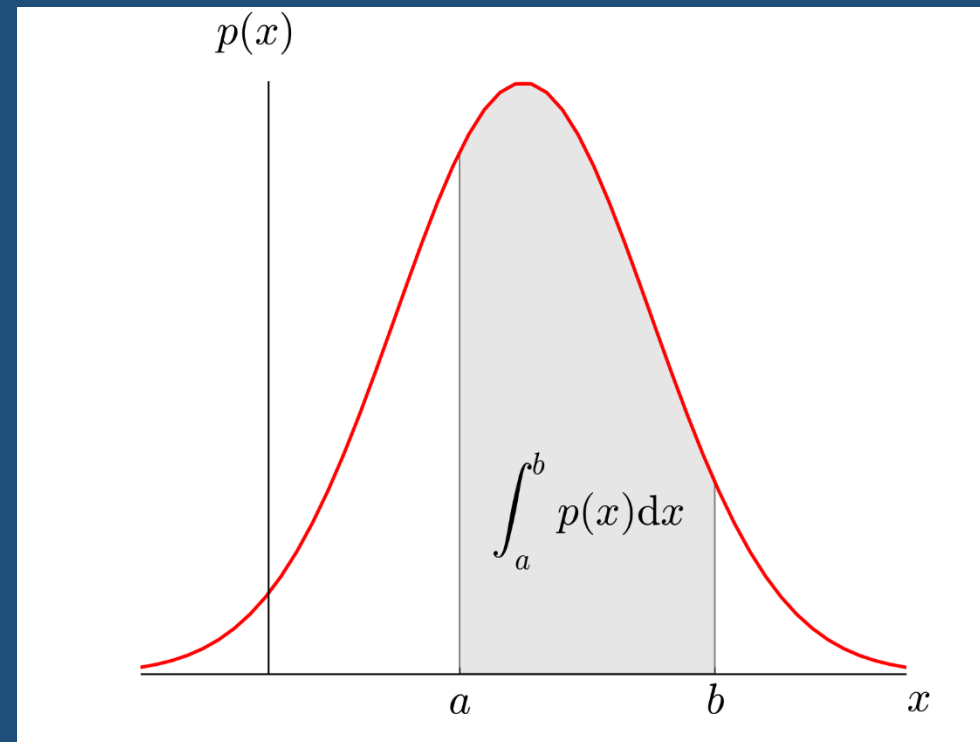
To sort out the problems with binned distribution and individual distribution, we use a distribution function.



Note: Using probability distribution function allows you to find the probabilities of all the discrete points.

Probability Distribution Function: Properties

It maps the possible values of x against their respective probabilities of occurrence $p(x)$



$p(x)$ is a number from 0 to 1.0

The area under a probability function is always 1

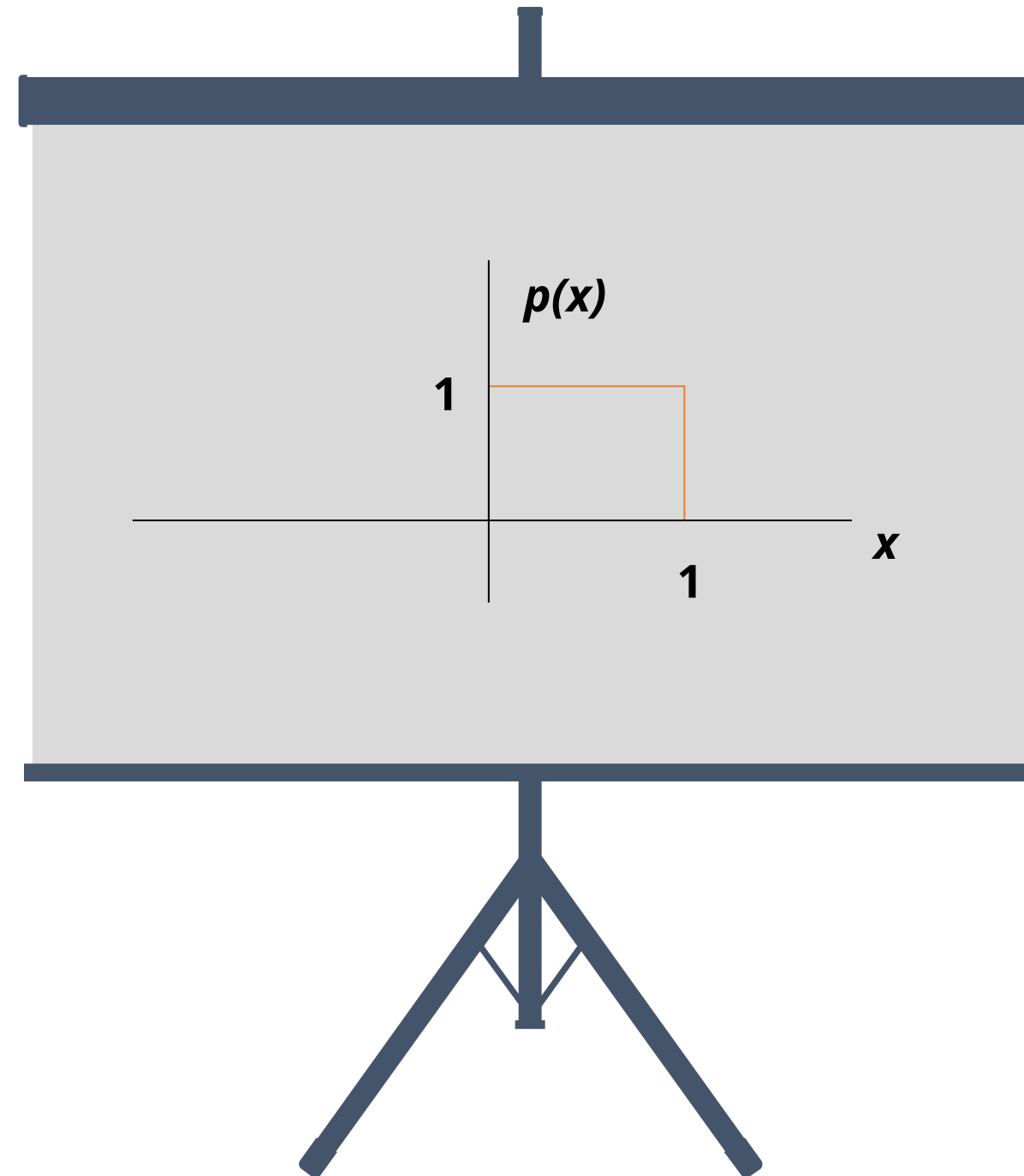
Example: The probability of X being between 0 and 1

$$P(0 \leq X \leq 1) = \int_0^1 f(x) dx$$

Types of Probability Density Function

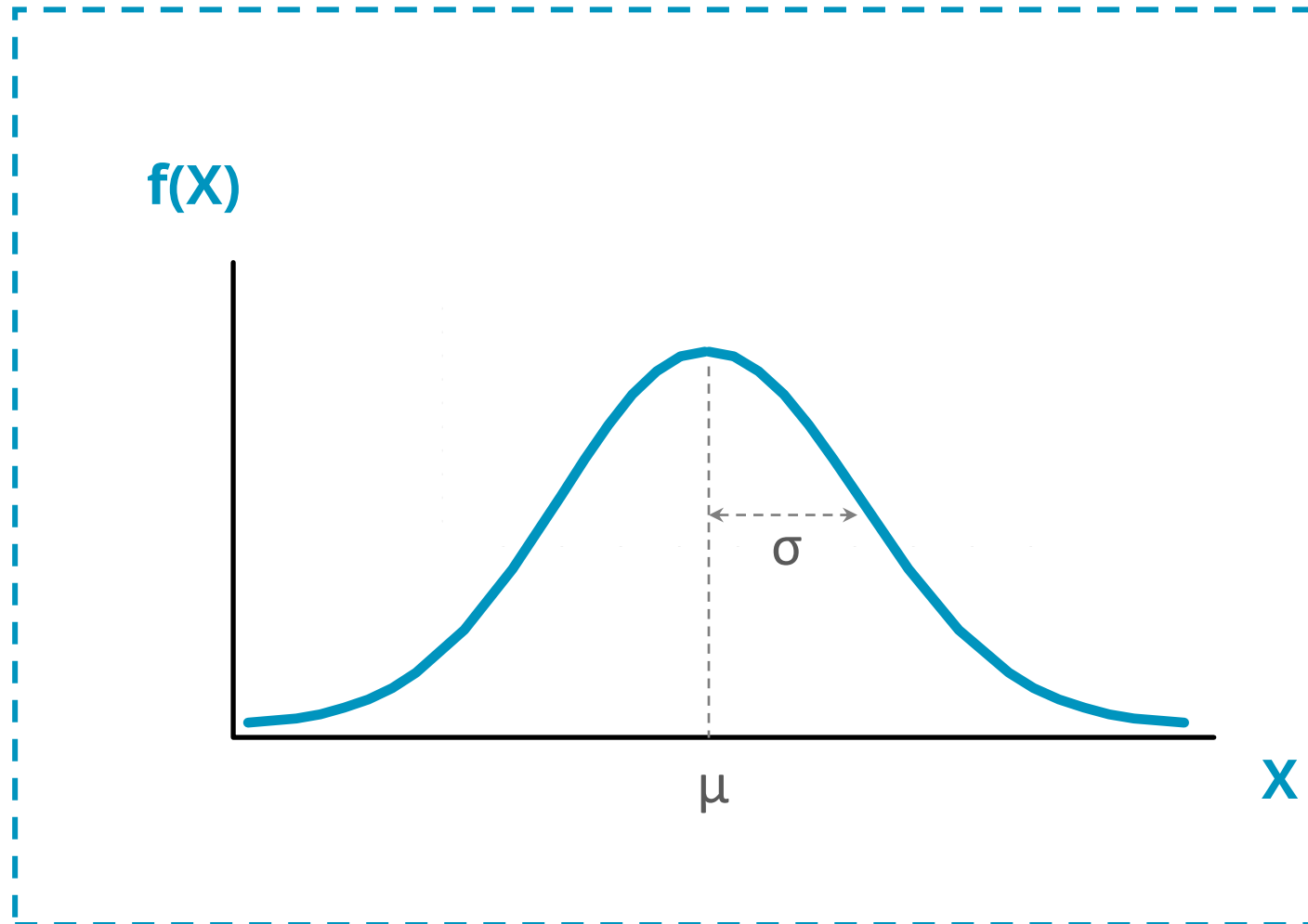
Uniform Distribution

Equal probabilities of occurrence of random variable



$$f(x) = 1, \text{ for } 1 \geq x \geq 0$$

Normal Distribution



- 01 Bell-shaped curve
- 02 Symmetric around the mean
- 03 Mean = Median
- 04 Total area under the curve is 1
- 05 Almost all the values fall within 3σ

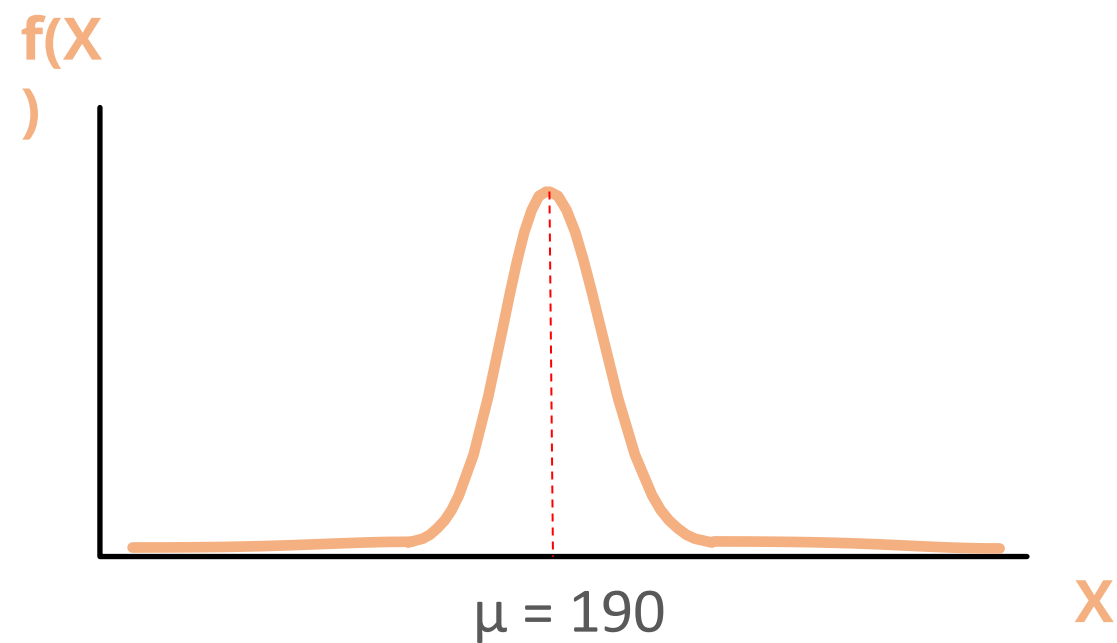
Normal Distribution: Use Case

Consider a scenario where John and Dany have 200 and 150 followers on Twitter and Facebook, respectively, and we are supposed to determine who is more popular.

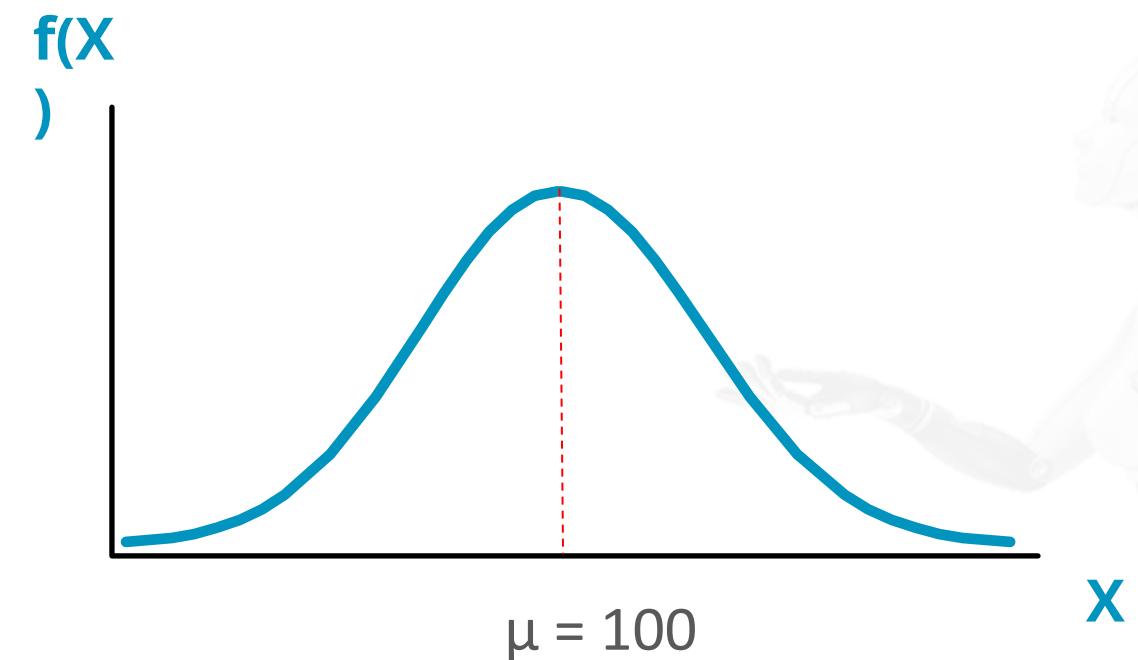


Normal Distribution

Probability distribution curves for Twitter and Facebook:



Consider the mean for Twitter distribution as 190 and standard deviation as 5

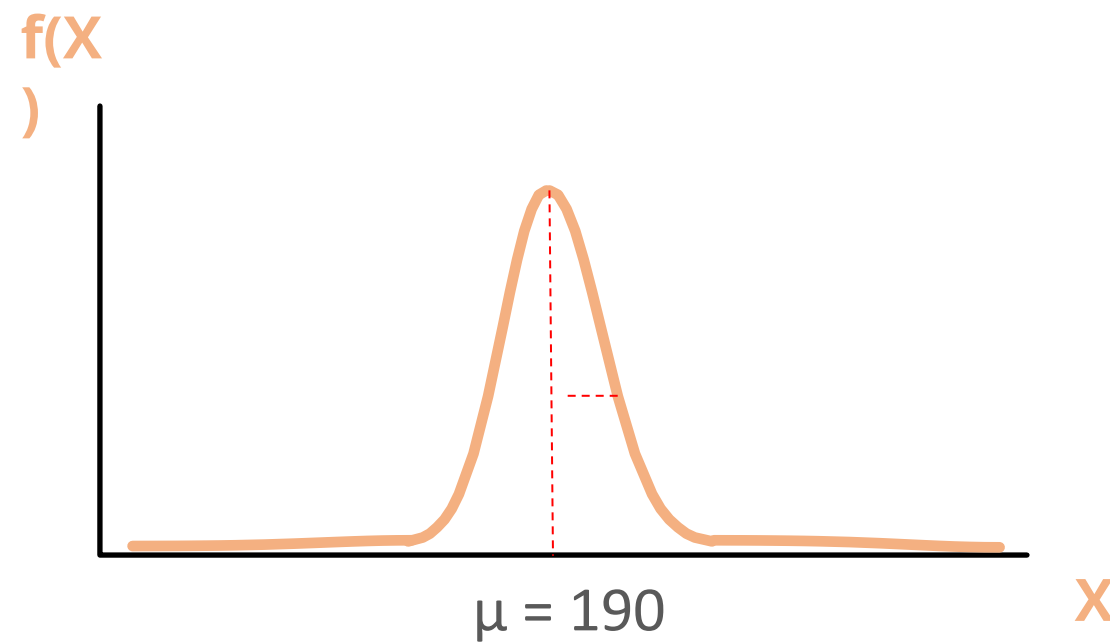


Consider the mean for Facebook distribution as 100 and standard deviation as 10

Z-Scores

Indicate how many standard deviations away from the mean does the point x lies

Consider the mean for twitter distribution as 190 and standard deviation as 5

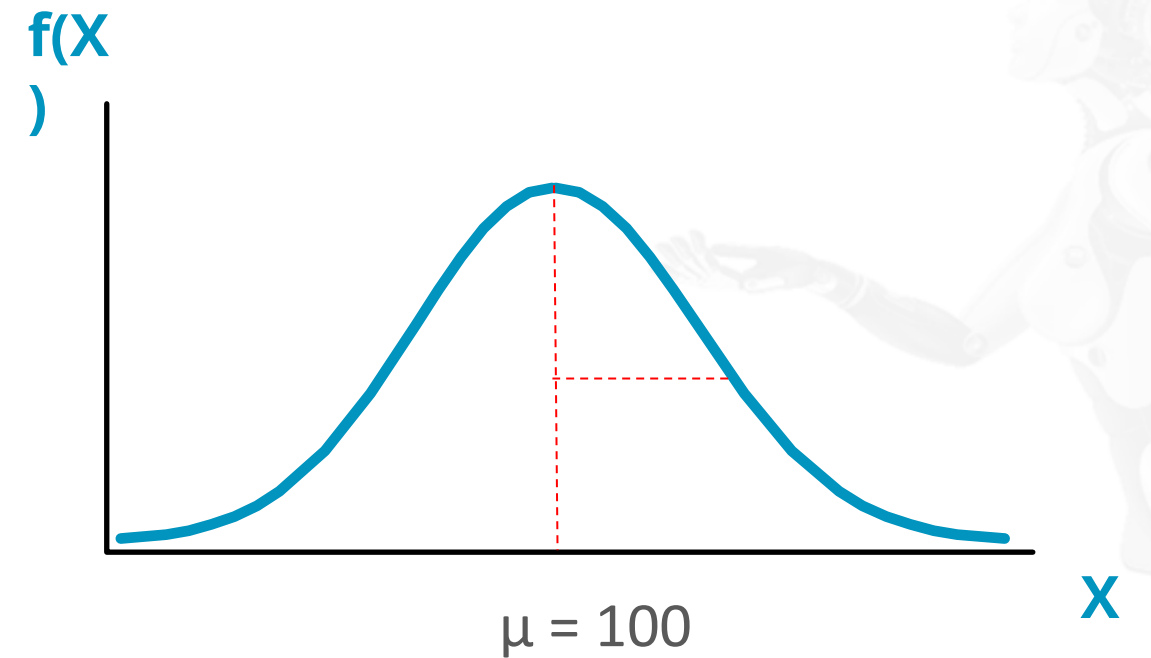


$$Z = \frac{X - \mu}{\sigma}$$



$$Z = 2$$

Consider the mean for Facebook distribution as 100 and standard deviation as 10



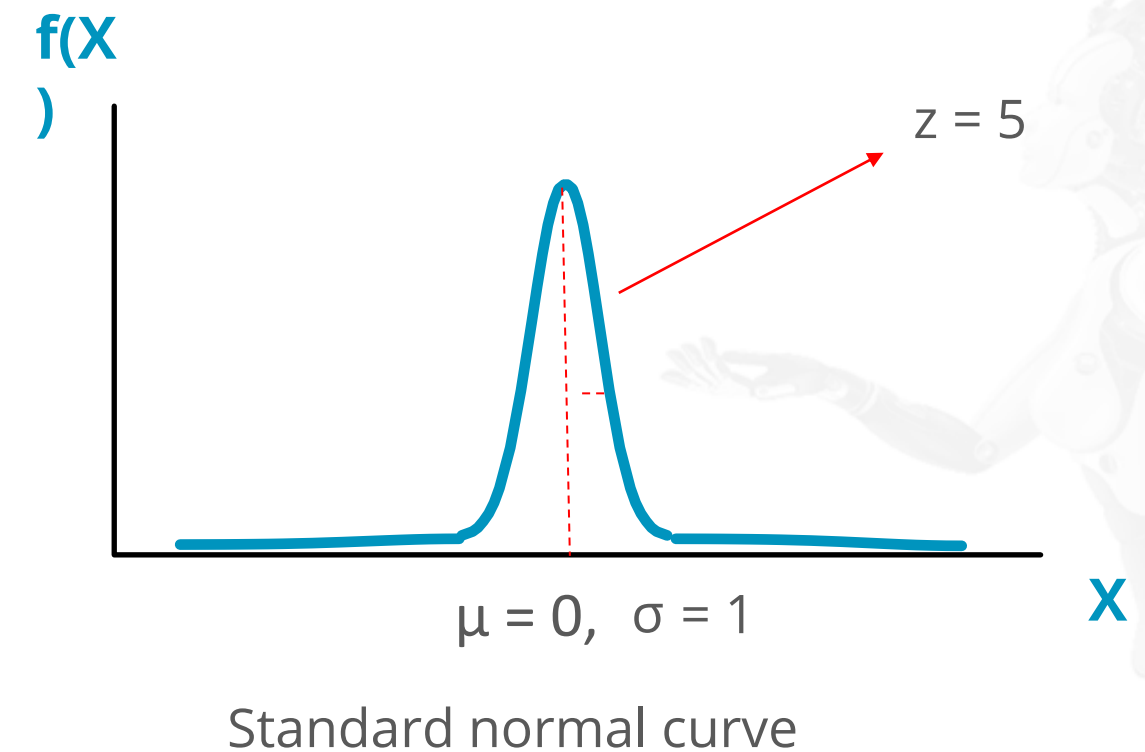
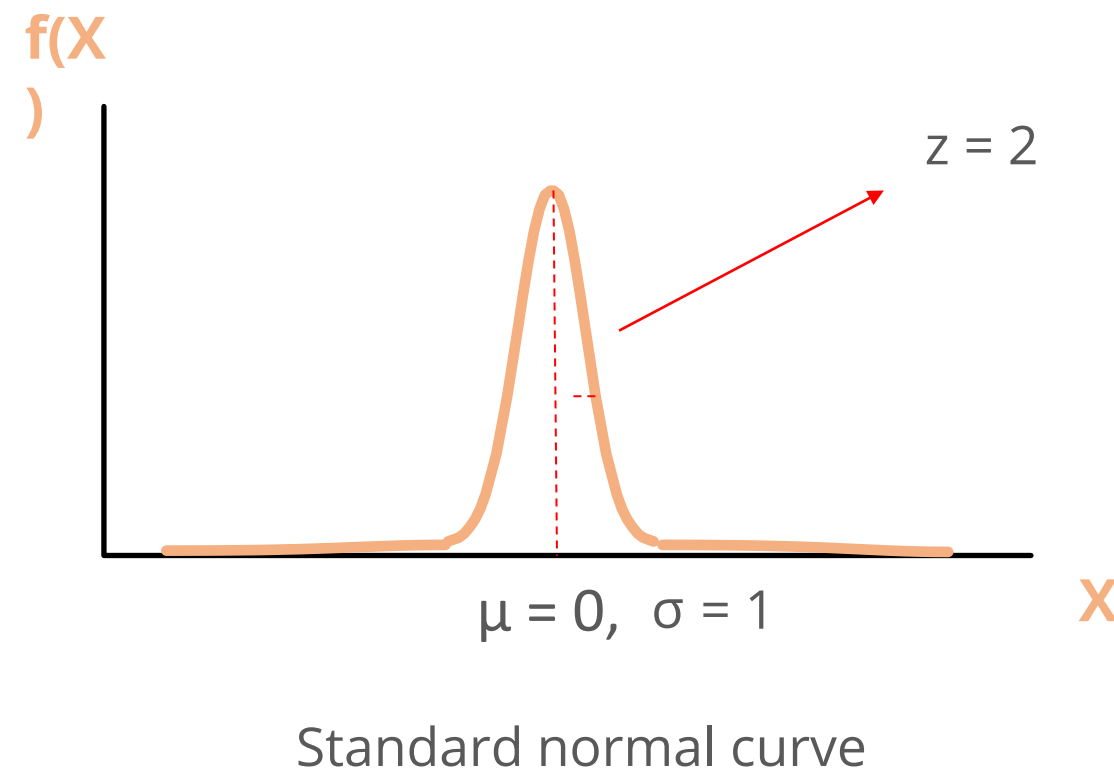
$$Z = \frac{X - \mu}{\sigma}$$



$$Z = 5$$

Standard Normal Distribution

It is a normal distribution curve with mean at 0 and standard deviation 1. They are most commonly used to compare two independent events.

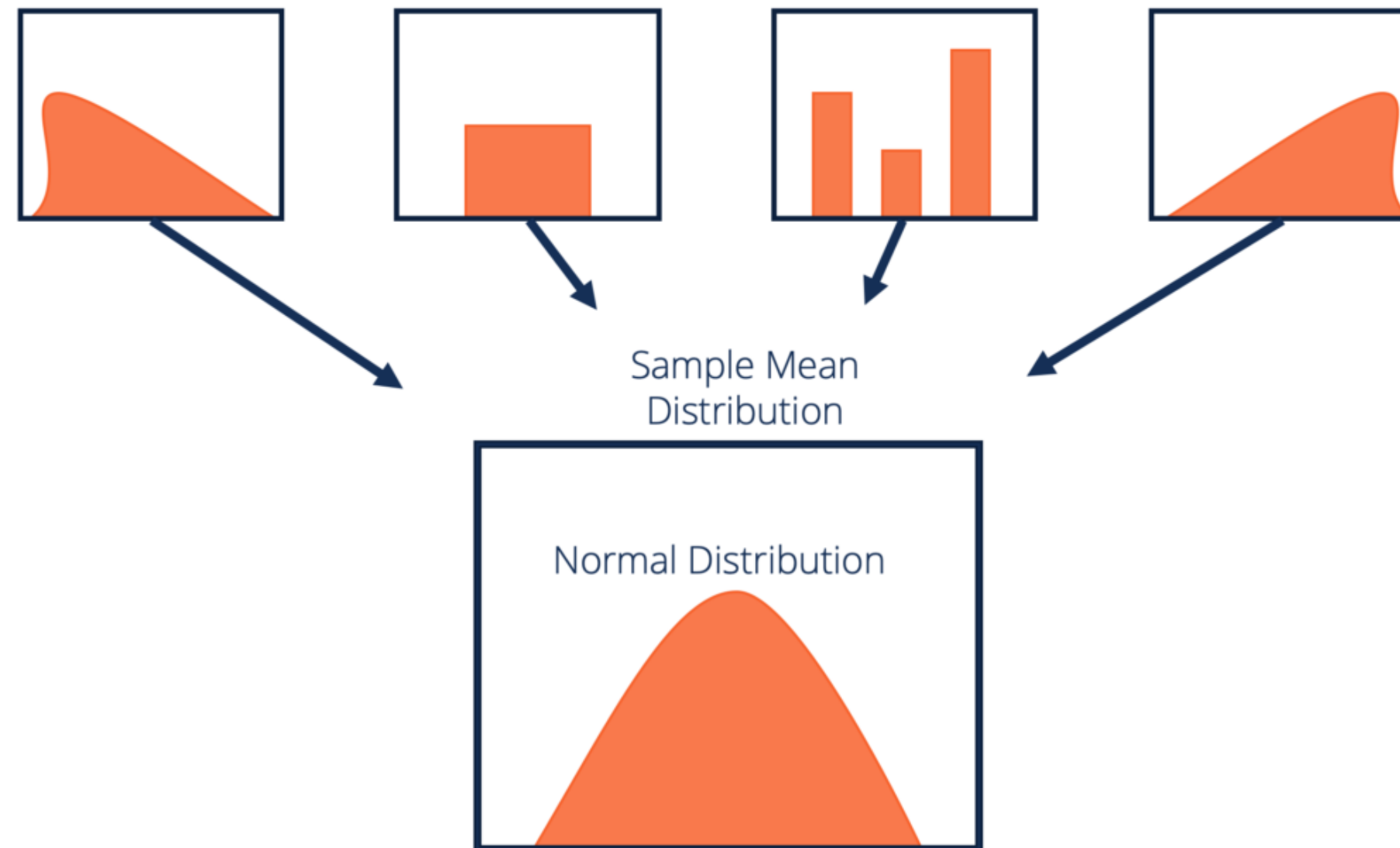


Note: From the above z-scores, we can infer that Dany is far more popular than John as she has a z-score of 5 which constitutes 99 percentile of the follower data.

Central Limit Theorem

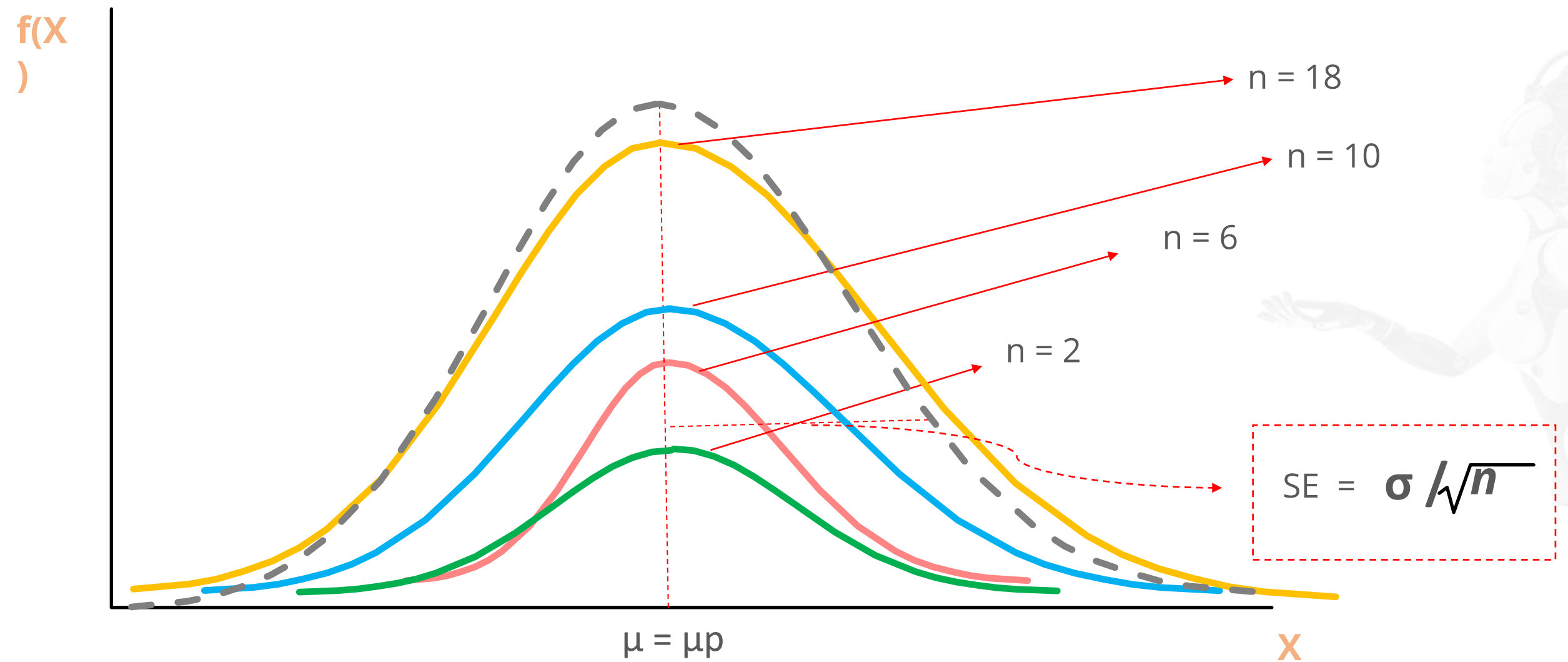
Central Limit Theorem

Sampling distribution of sample means is normal if the sample size is large enough.



Sample Mean vs. Population Mean

The mean of sampling distribution of sample means can be used to infer the population mean.



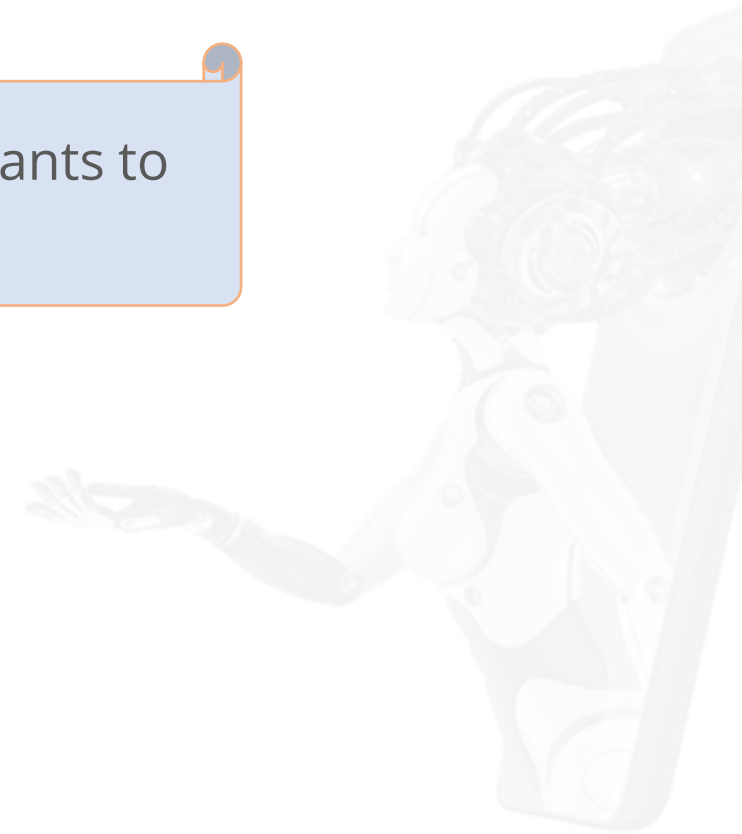
Note: μ_p : population mean, μ : sample mean of the sampling distribution, n : sample size, σ : population standard deviation.

Confidence Intervals

Hypothesis Test: Use Case

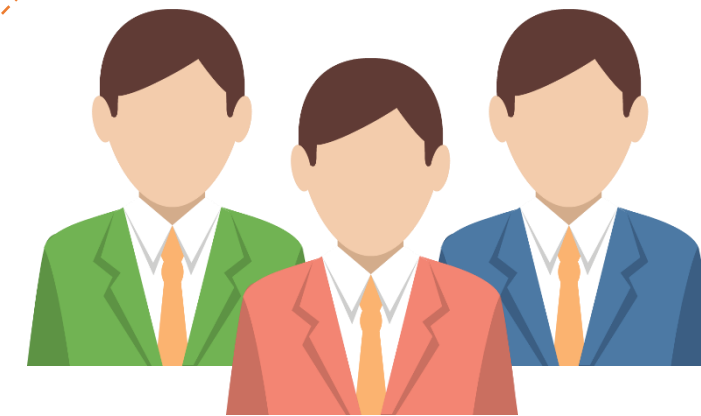
The mean of sampling distribution of sample means can be used to infer the population mean.

XYZ company has finished the development of its most awaited product and wants to test this product in the market on a certain set of customers.



Hypothesis Test: Use Case

The management decided to test NPS score of its new product out of a sample of learners.



Random sample of learners with
a sample size of 25

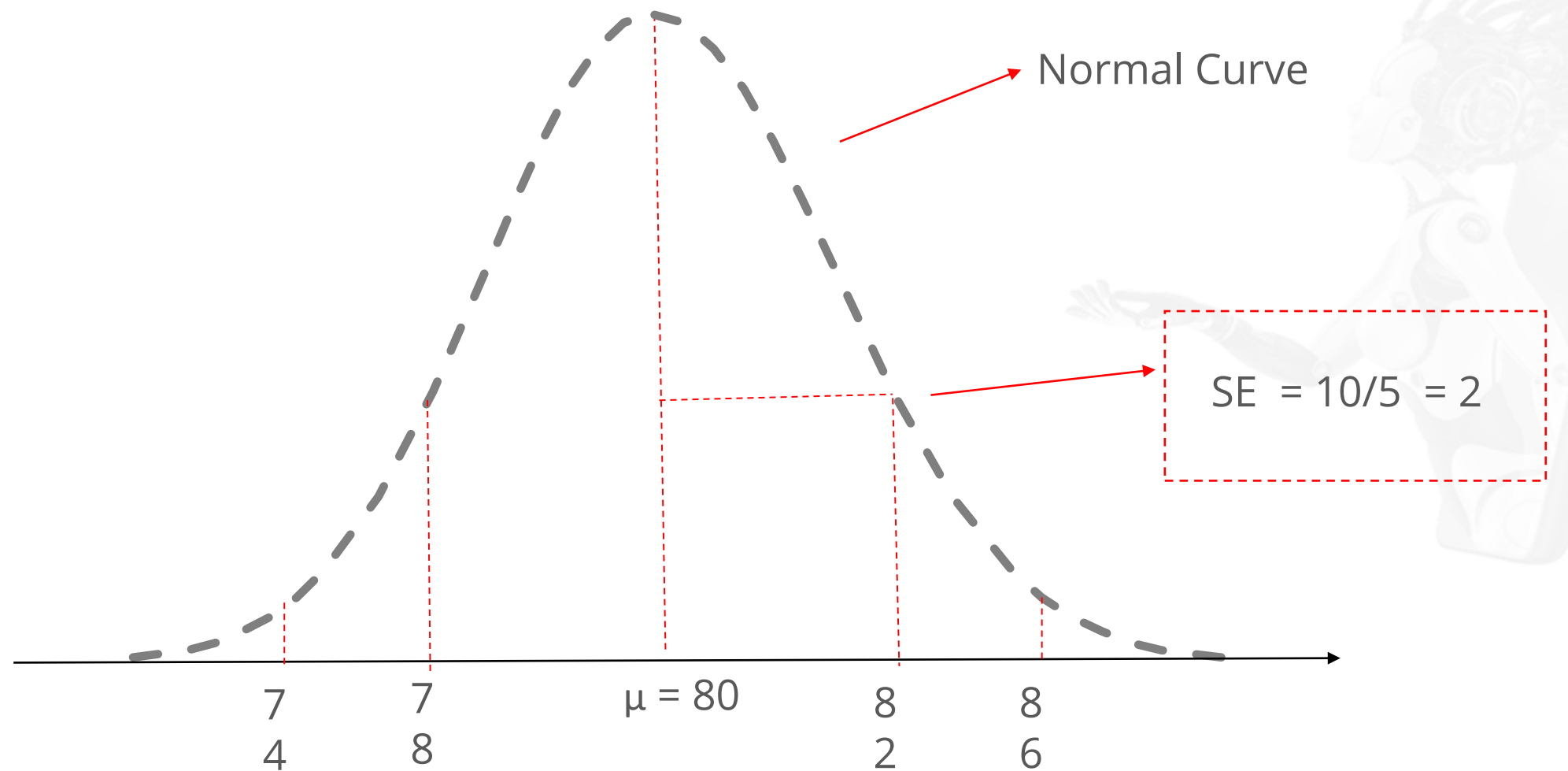
Mean of the NPS Score of the population
before product launch= 80 and standard
deviation = 10

Mean of the NPS Score of the
sample= 90

This leads the management to proceed with a point estimate that the mean NPS of the population will be 90 approximately.

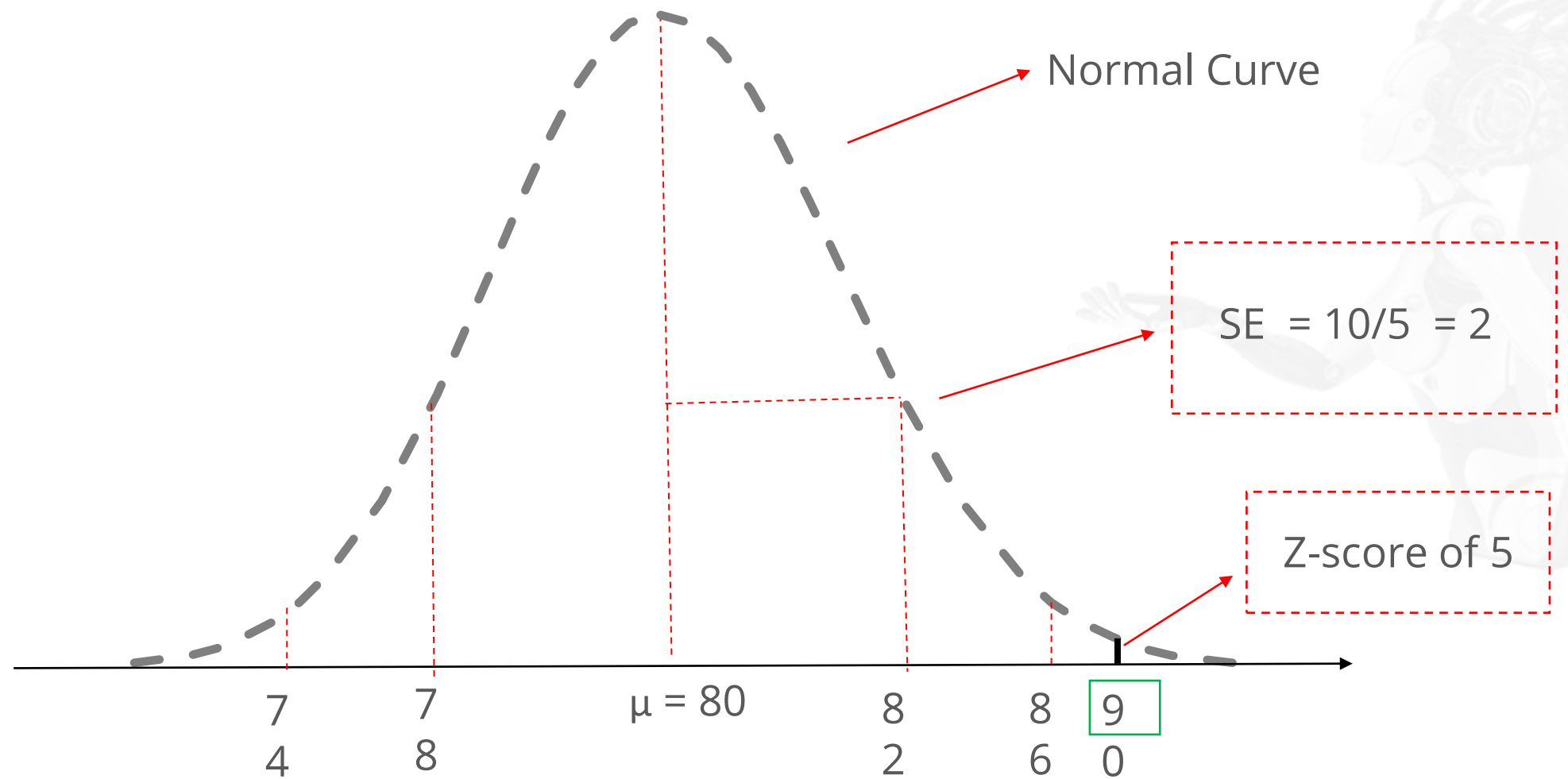
Hypothesis Test: Use Case

The management now took n samples of size 25 each such that plotting the sampling distribution of all these sample means resulted in a normal distribution with population mean of 80.



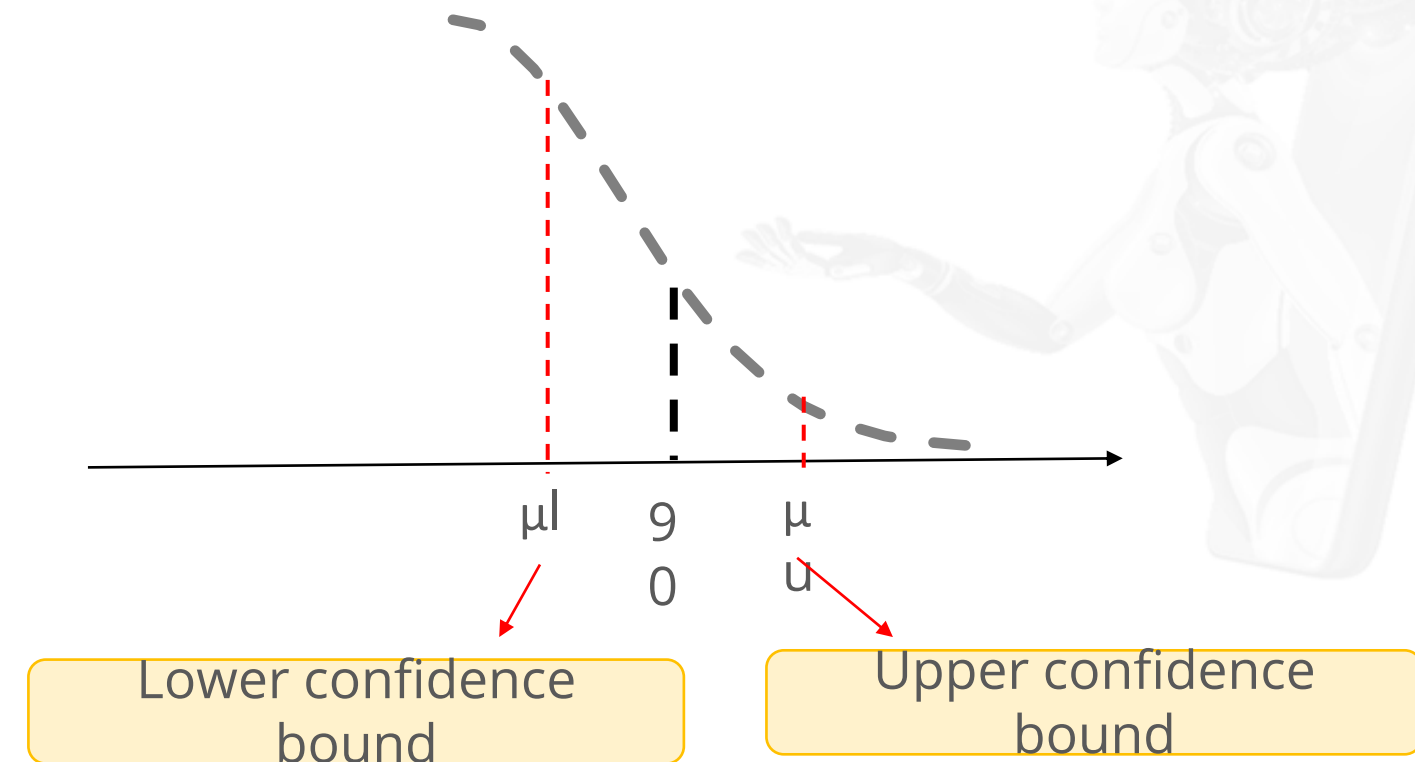
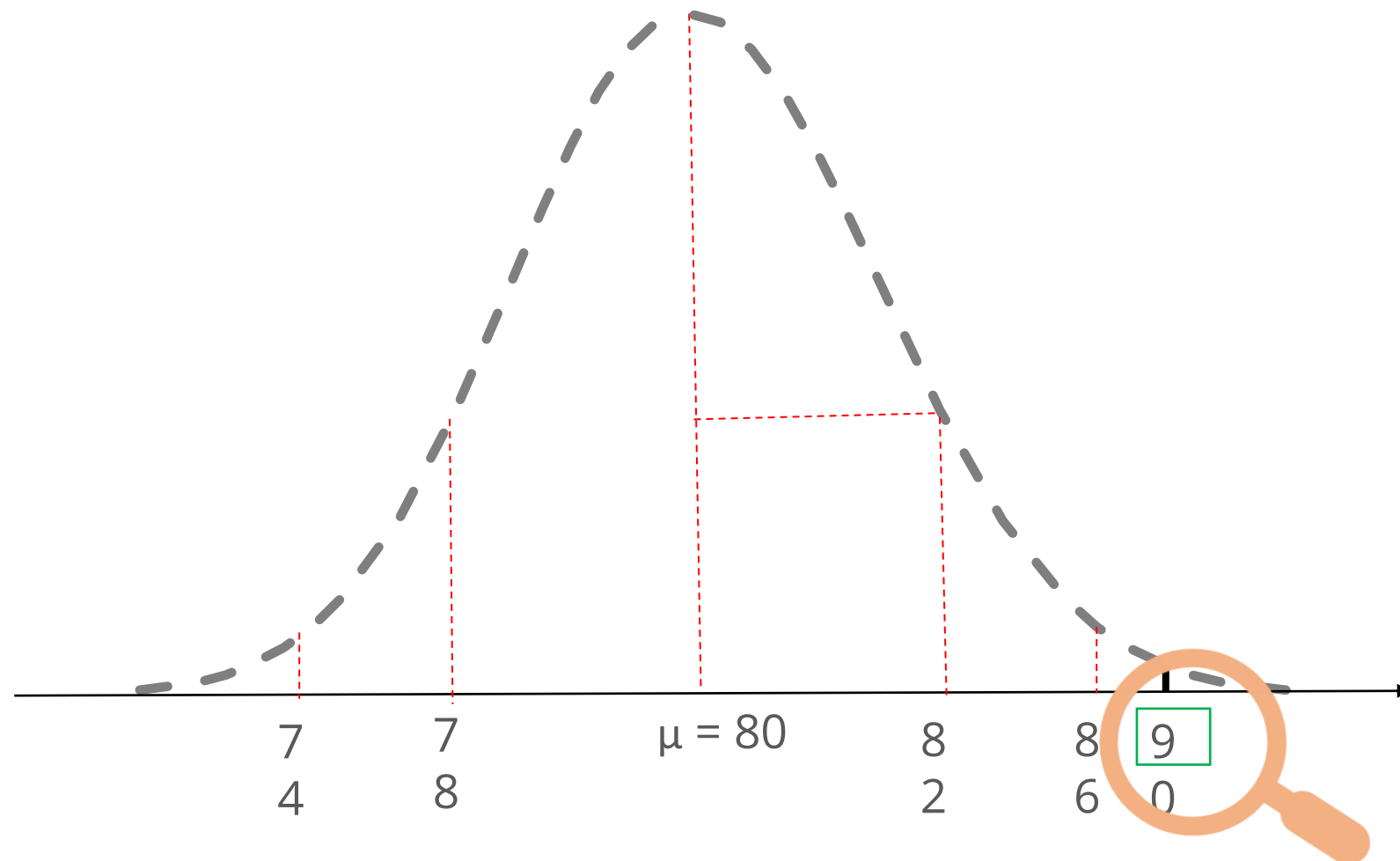
Hypothesis Test: Use Case

From the below z-score, the management can easily infer that the sample with sample mean of 90 came from a different distribution. (centered around 90)



Interval Estimate

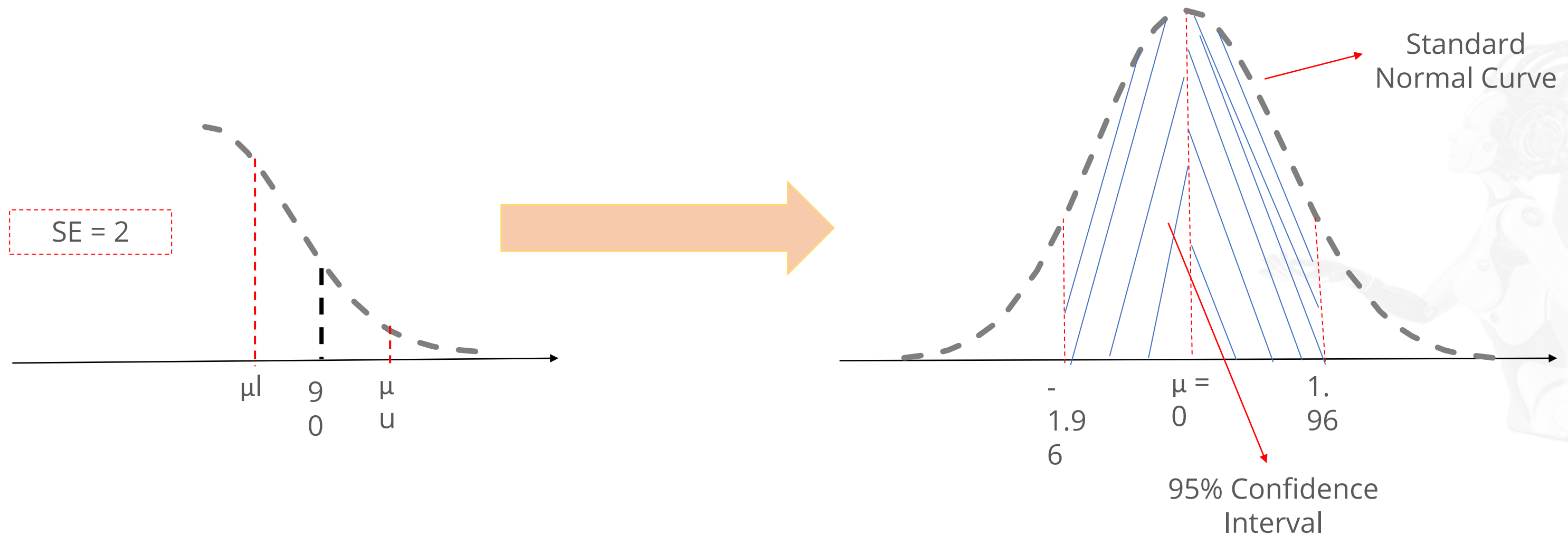
If the management opens the product for all its customers, the NPS score will definitely increase and will be somewhere near 90.



Note: One can never be sure with a point estimate. Therefore, to measure the certainty there is always a range associated with a point estimate.

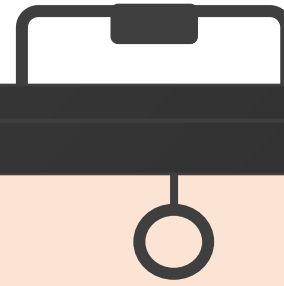
Confidence Intervals

The management wants to be 95% sure that if it launches its new product for the entire customer base, the new mean NPS will be 90.



$$\begin{aligned}\mu_l &= 90 - 1.96 SE = 86.08 \\ \mu_u &= 90 + 1.96 SE = 93.92\end{aligned}$$

Constructing Confidence Intervals

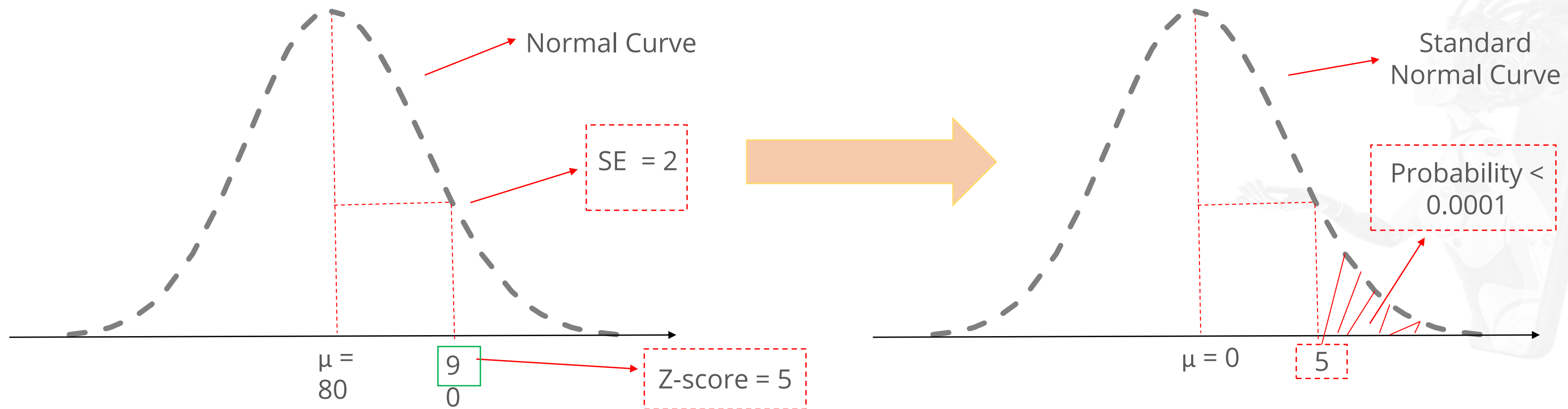


- ✓ Choose a statistic that you will use to estimate a population parameter.
- ✓ Select a confidence level.
- ✓ The confidence interval can be found out by:
confidence interval = sample statistic \pm Margin of error (z-score \times SE).

Hypothesis Testing: Parametric

Constructing Confidence Intervals

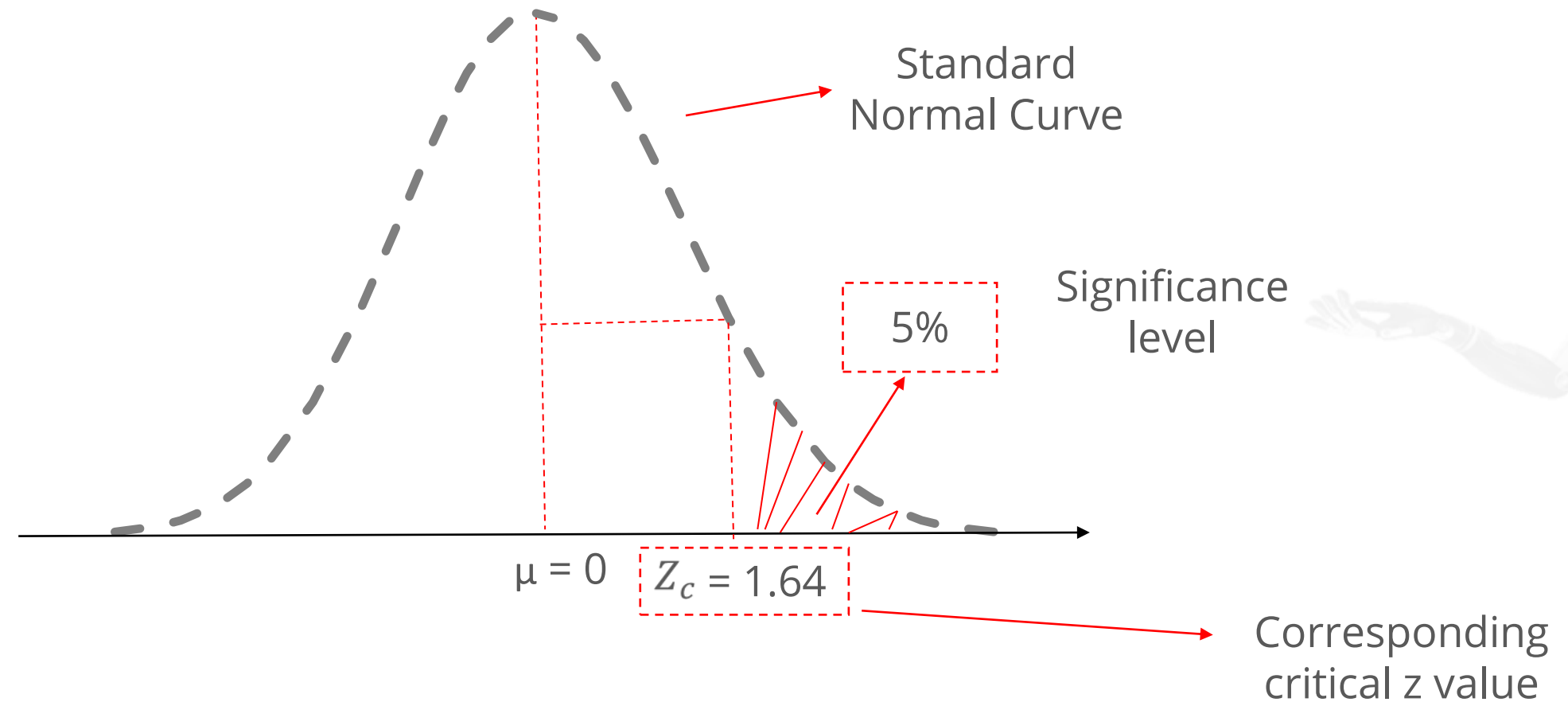
Consider the previous scenario where the company tested its new product against a sample, and the mean NPS came out to be 90.



From the above z-score and probability, the management can clearly state that this mean (sample mean of 90) is a part of different distribution.

Levels of Significance

The alpha levels represent the level of likelihood.



In the above case, if any sample mean has a z-score > 1.64 then it is considered unlikely, given the alpha levels as 5%.

Null Hypothesis vs. Alternate Hypothesis

Consider the previous scenario of product launch within a company.

Null Hypothesis

- New predicted mean NPS = 80, which is the same as old mean NPS
- $H_0: \mu_{New} = \mu_{old}$

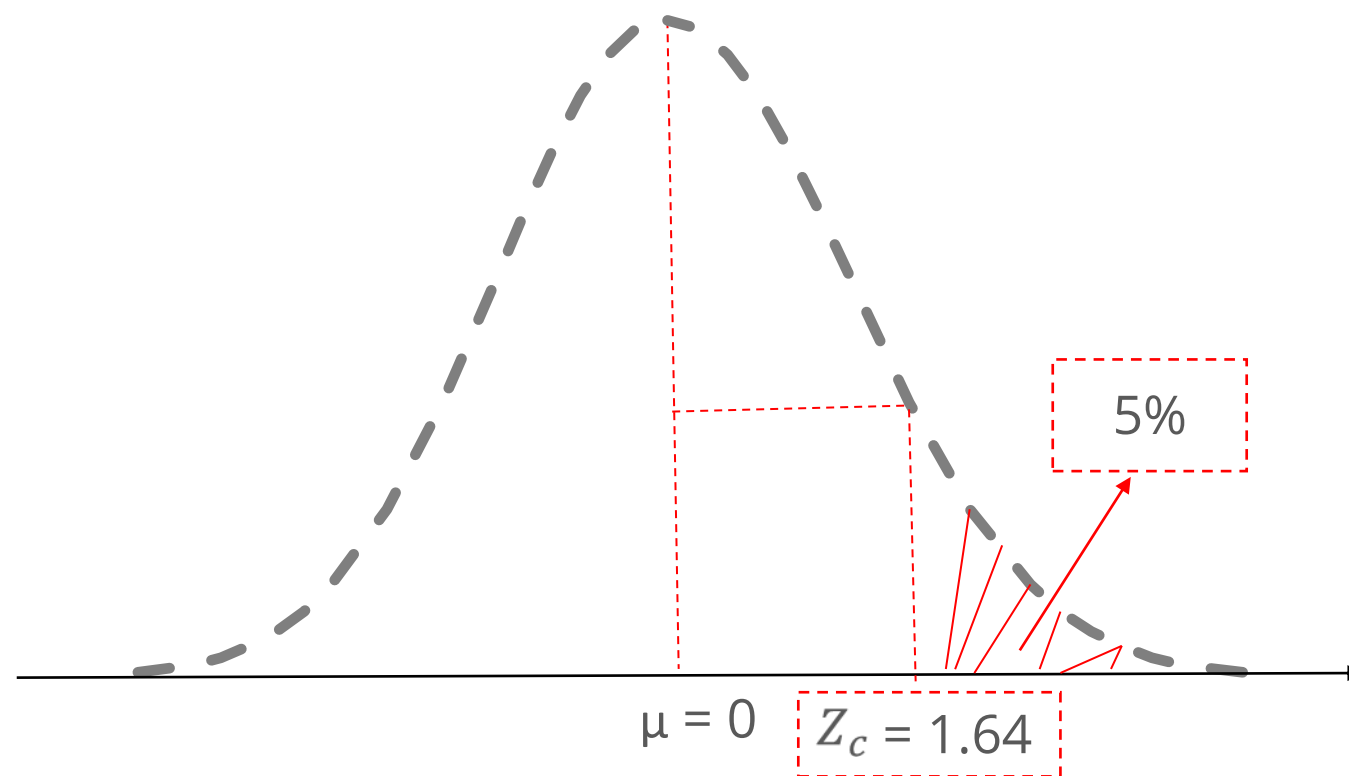
Alternate Hypothesis

- Either the new mean NPS is greater than the old mean NPS or the new mean NPS is lesser than the old mean or the new mean NPS is not equal to the old mean NPS
- $H_a: \mu_{New} > \mu_{old}, \mu_{New} < \mu_{old}, \mu_{New} \neq \mu_{old}$

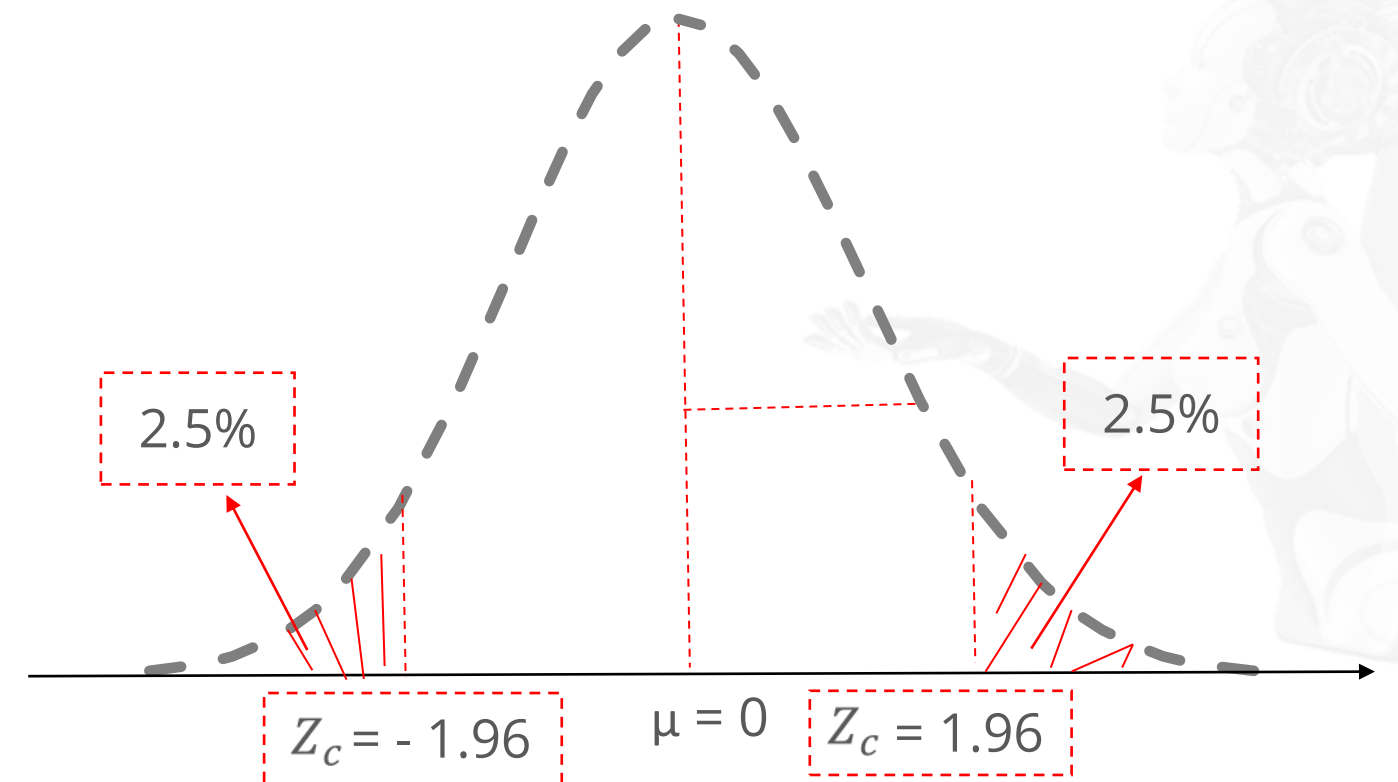
One Tail Test vs. Two Tail Test

Null hypothesis can be rejected or accepted based on one tailed or two tailed tests.

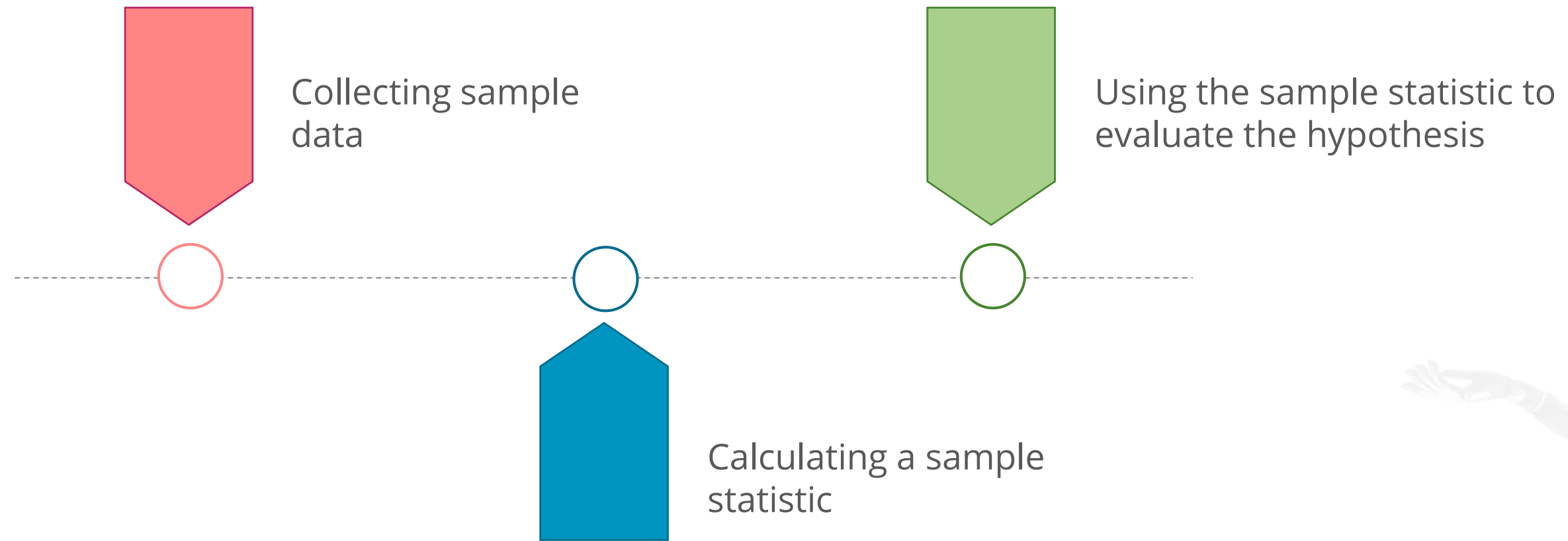
One tail



Two tail



Hypothesis Testing: Process Flow

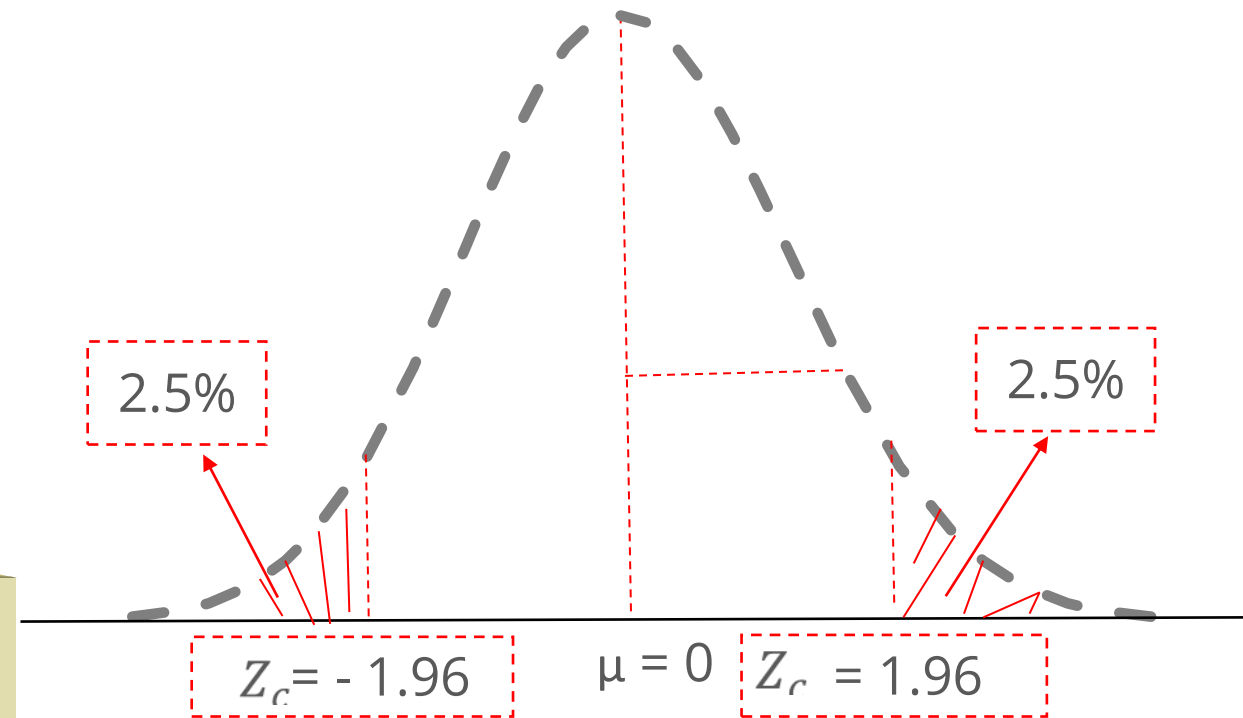


Type I and Type II Errors

Hypothesis testing is prone to errors.

Type I Error

The null hypothesis is rejected, when it is true



Type II Error

The null hypothesis is not rejected, when it is false

Hypothesis Testing: Nonparametric

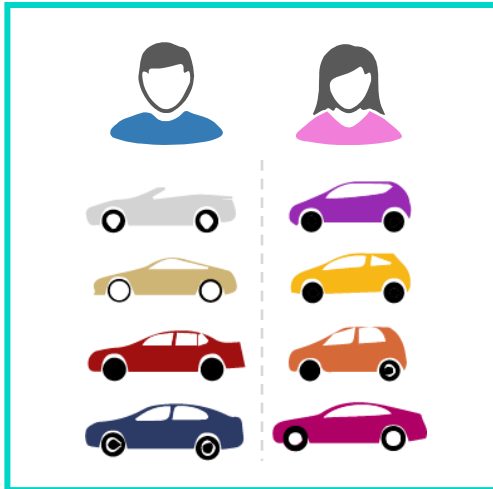
Chi-Square Test

It is a hypothesis test that compares the observed distribution of your data to an expected distribution of data.



Test of Association:

To determine whether one variable is associated with a different variable
Example: Determine whether the sales for different cell phones depend on the city or country where they are sold.



Test of Independence:

To determine whether the observed value of one variable depends on the observed value of a different variable
Example: Determine whether the color of the car that a person chooses is independent of the person's gender.



The test is usually applied when there are two categorical variables from a single population.

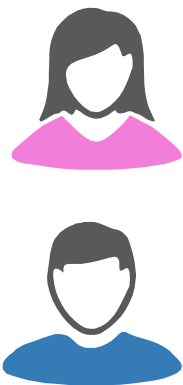
Chi-Square Test: Example

Null Hypothesis

- There is no association between gender and purchase.
- The probability of purchase does not change for 500 dollars or more, whether female or male.

Alternative Hypothesis

- There is association between gender and purchase.
- The probability of purchase over 500 dollars is different for female and male.



	Items <\$500	Items >\$500
fe	.55	.45
fo	.75	.25

Types of Frequencies

Expected and observed frequencies are the two types of frequencies.

Expected Frequencies (fe)

The cell frequencies that are expected in a bivariate table if the two tables are statistically independent

Observed Frequencies (fo)

The actual frequency that is obtained from the experiment



	Items <\$500	Items >\$500
fe	.55	.45
fo	.75	.25

No Association

Observed Frequency = Expected Frequency

Association

Observed Frequency \neq Expected Frequency

Features of Frequencies

The formula for calculating expected and observed frequencies using chi-square:

$$\sum \frac{(f_e - f_o)^2}{f_e}$$

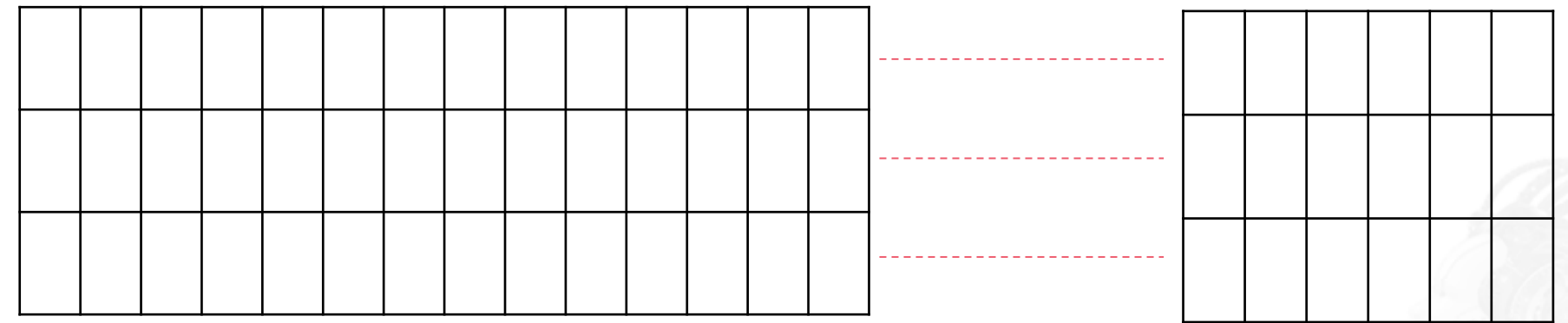
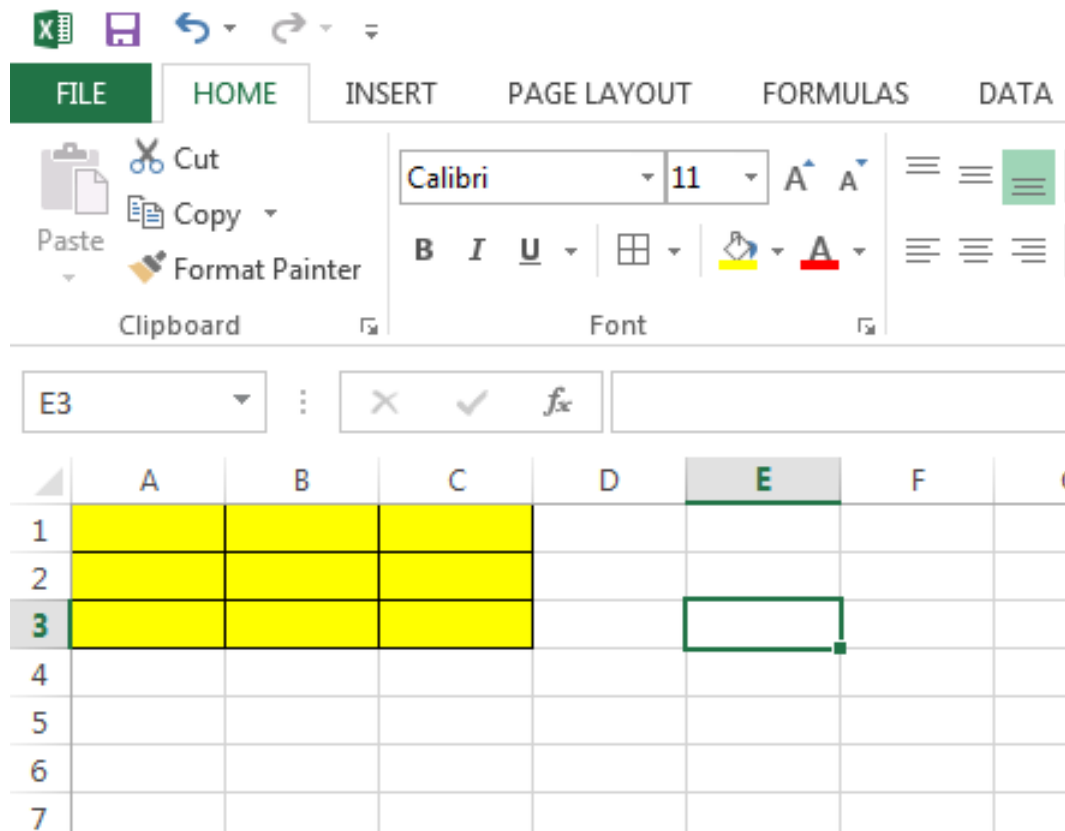
Features of expected and observed frequencies:

- Requires no assumption of the underlying population
- Requires random sampling



Correlation Matrix

A correlation matrix is a square matrix that compares large number of variables.



Correlation matrix: a square matrix

$n \times n$ Matrix

(very large number of rows and columns)

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

3 × 3 matrix (simple square matrix)

Correlation coefficient measures the extent to which two variables tend to change together.

The coefficient describes both the strength and direction of the relationship.

Correlation Matrix

Pearson product moment correlation

It evaluates the linear relationship between two continuous variables.

Linear relationship means that a change in one variable results in a proportional change in the other.

Spearman rank order correlation

It evaluates the monotonic relationship between two continuous or ordinal variables.

- Monotonic relationship means that the variables tend to change together though not necessarily at a constant rate.
- The correlation coefficient is based on the ranked values for each variable rather than the raw data.

Correlation Matrix: Example

An example of a correlation matrix calculated for a stock market:

U10		fx =CORREL(\$C\$9:\$C\$78,B\$9:B\$78)					
	T	U	V	W	X	Y	Z
8	Correlation	EQUITY 1	EQUITY 2	FX FORWARD 1	FX FORWARD 2	BOND 1	BOND 2
9	EQUITY 1	1.00	0.38	0.20	0.45	- 0.17	- 0.12
10	EQUITY 2	0.38	1.00	0.54	0.51	- 0.20	0.12
11	FX FORWARD 1	0.20	0.54	1.00	0.35	- 0.14	0.16
12	FX FORWARD 2	0.45	0.51	0.35	1.00	- 0.11	- 0.09
13	BOND 1	- 0.17	- 0.20	- 0.14	- 0.11	1.00	0.03
14	BOND 2	- 0.12	0.12	0.16	- 0.09	0.03	1.00



A correlation matrix that is calculated for the stock market will probably show the short-term, medium-term, and long-term relationships between data variables.

What Is A/B Testing?

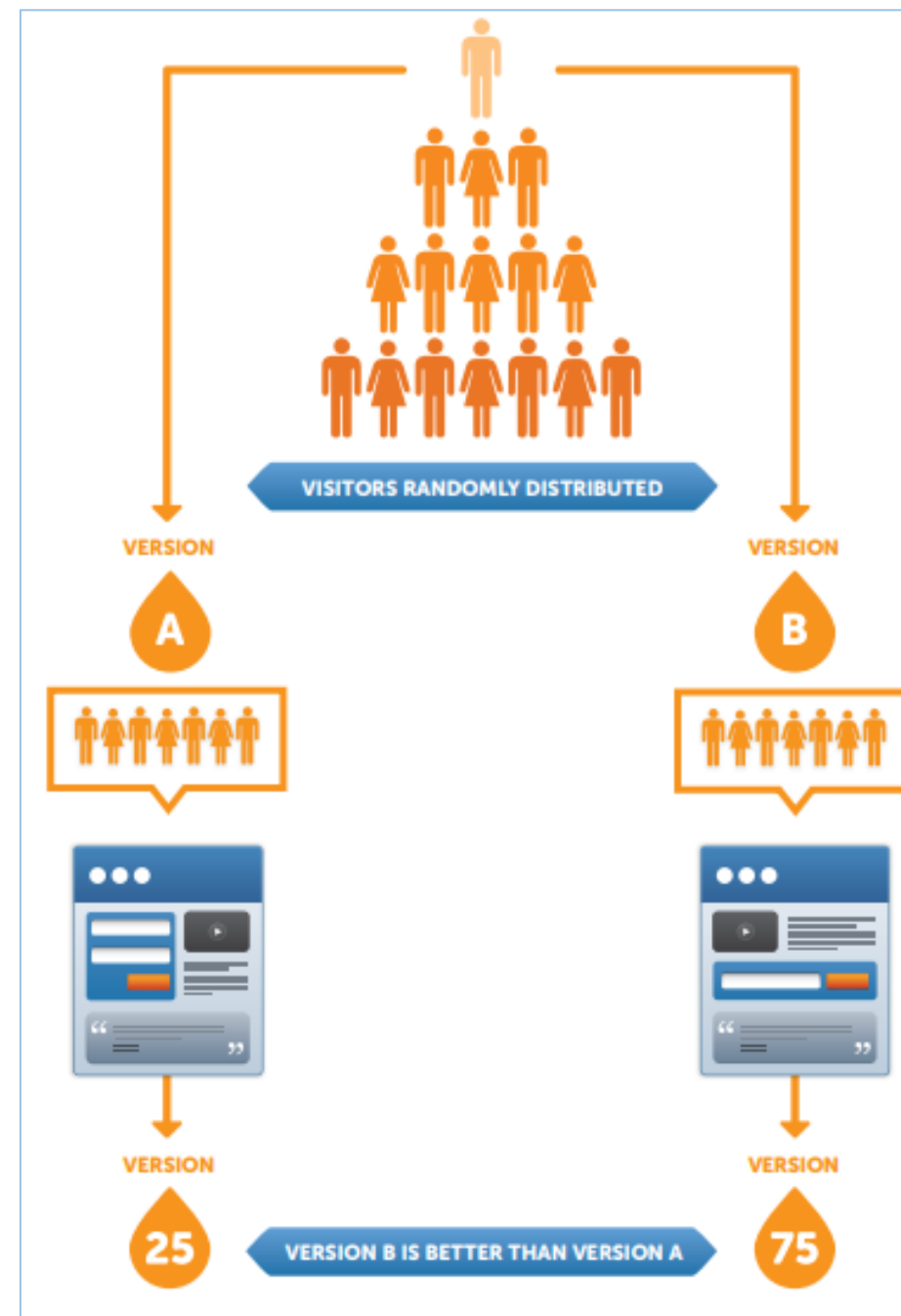
A/B Testing

A general methodology used online to test a new product or feature



A/B Testing

We compare one page against the other for a difference in an element of the control page.



A/B Testing: Steps

01

**Creating
control page**

A good control page is the foundation for testing. Popular structures cannot be right.

02

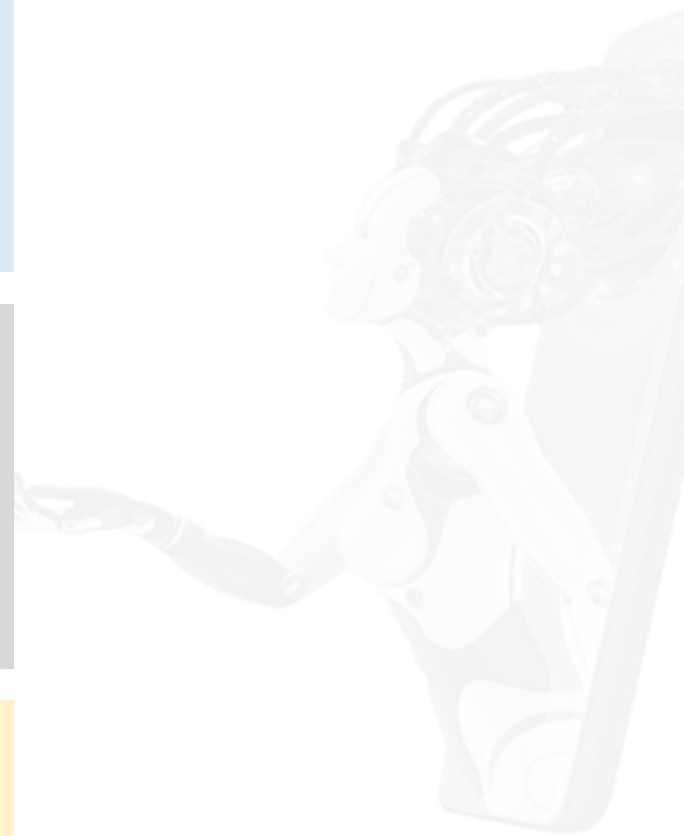
**Building page
hypothesis**

There should be only one hypothesis per variation, for example, color of background, different major image, different CTA in the headline, different CTA in the action button, etc.

03

**Determining test
winner**

The winner is the page variation that converts the best.



Stop A/B Testing



When you're nearing the end of a test, and your statistical confidence level is still low, but you're seeing a decent difference in conversion rate between the variants, make a gut-check call and remove the poorer performing pages.



A/B Testing: Advantages and Disadvantages

Advantages



- A/B testing is fast
- Each variation can be evaluated (Examples: click, tracking, heatmaps, etc.)
- Less traffic is required



Disadvantages

- More failures
- Effects of each element become less specific



Case Study: A/B Testing

Scenario

After the launch of the Noob Guide into online marketing, company XU created a landing page that used PayWithATweet.com to help expand the distribution of the guide.



Scenario

Data was gathered by inserting KISSinsights (now Qualaroo) widget in the page which prompted many people to provide their emails to get the pdf download.



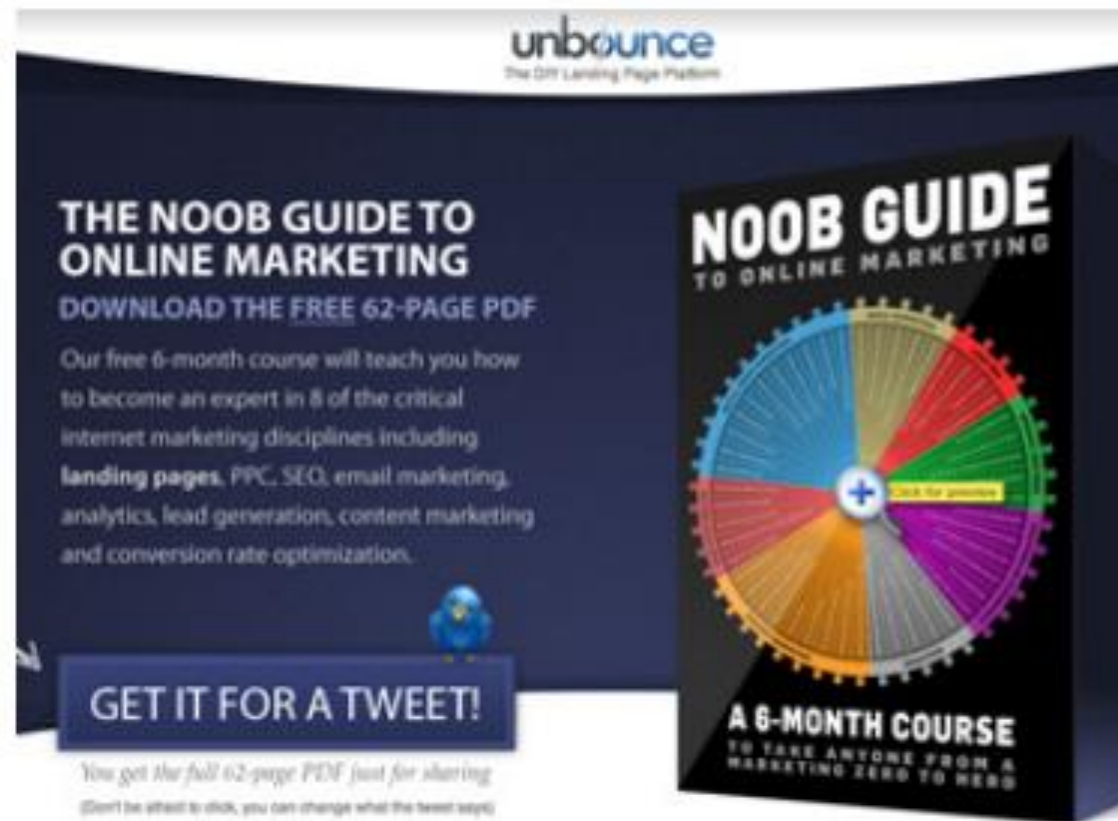
Formulating Test Hypothesis

Test Hypothesis:

Will allowing visitors to download the PDF by providing their email address work better than receiving it in exchange for a tweet, considering that not everyone has a Twitter account or is willing to share such information with their followers.

Perform A/B Testing

Tweet averaged at 18%
conversion rate



Email averaged at 22%
conversion rate



Interpreting A/B Testing

At the end of this test, XU decided to discuss a possible hybrid option that might produce the best of both worlds: a single version that lets the user decide whether they want to pay with a tweet or provide their email.

DATA AND ARTIFICIAL INTELLIGENCE



Knowledge Check

Knowledge Check

1

On an intelligence test with a mean of 100 and a standard deviation of 15, Cersei scored 85. What is Cersei's z-score?

- a. -2
- b. -1
- c. 1
- d. 2



Knowledge Check

1

On an intelligence test with a mean of 100 and a standard deviation of 15, Cersei scored 85. What is Cersei's z-score?

- a. -2
- b. -1
- c. 1
- d. 2



The correct answer is **b.**

The formula for z-score is: $\mu - x/\sigma$

Knowledge Check

2

A continuous random variable x can take any value between 0 and 1. Its probability density function is assumed to be uniform. What is the explicit form of its probability density function $f(x)$?

- a. $F(x) = 1$
- b. $F(x) = 1/2$
- c. $F(x) = x$
- d. $F(x) = \log(x)$



Knowledge Check

2

A continuous random variable x can take any value between 0 and 1. Its probability density function is assumed to be uniform. What is the explicit form of its probability density function $f(x)$?

- a. $F(x) = 1$
- b. $F(x) = 1/2$
- c. $F(x) = x$
- d. $F(x) = \log(x)$



The correct answer is **c.**

Since, domain of $f(x)$ is restricted to 0 and 1 and $f(x)$ is a uniform distribution, $f(0) = 0$ and $f(1) = 1$

Knowledge Check

3

In chi-square test, there is no association of variables if:

- a. Observed Frequency \neq Expected Frequency
- b. Observed Frequency = Expected Frequency
- c. Independent of observed frequencies
- d. Independent of expected frequencies



Knowledge Check

3

In Chi-Square test, there is no association of variables if:

- a. Observed Frequency \neq Expected Frequency
- b. Observed Frequency = Expected Frequency
- c. Independent of observed frequencies
- d. Independent of expected frequencies



The correct answer is **b**

Observed Frequency = Expected Frequency indicates no association.

Knowledge Check

4

What's the first step of running an A/B test?

- a. Creating variations for your landing page elements
- b. Establishing your success metric
- c. Making assumptions about what your customers will like
- d. Both a and b



Knowledge Check

4

What's the first step of running an A/B test?

- a. Creating variations for your landing page elements
- b. Establishing your success metric
- c. Making assumptions about what your customers will like
- d. Both a and b



The correct answer is **b.**

Before starting with A/B test, you must set your success metric.

Knowledge Check

5

Which of the following is a false statement about A/B testing?

- a. You only create two variations for your landing pages
- b. It improves your conversion rate
- c. You have to at least run the test for a couple of weeks to get correct results
- d. All of the above



Knowledge
Check

5

Which of the following is a false statement about A/B testing?

- a. You only create two variations for your landing pages
- b. It improves your conversion rate
- c. You have to at least run the test for a couple of weeks to get correct results
- d. All of the above



The correct answer is **a.**

When you run a multivariate test, you use one page and dynamically supply multiple versions of multiple elements.

A/B Testing



Problem Scenario: Proda, an electric utility company, manufactures devices that keep the electrical grid operating efficiently. It enables grid operators to respond to the demand for reliable energy with trends like digitalization, self-healing grids, and reinforced infrastructure.

Last month engineers in Proda developed two updated versions of EISR. It is an electrical insulator material that impedes the free flow of electrons. Due to their resistivity, these insulators are installed to prevent line damage from arcing. The CEO of Proda claimed that the updated versions of

EISR performed much better. To test the same, the delivery team of Proda released the two versions

among two customer groups (grouped based on random sampling) and found out that version 1 has

a mean feedback score of 50, whereas version 2 has a mean feedback score of 40.

However, to maintain standards, the company is trying to finalize on one of the two versions. You, as

a statistician, must do some level of statistical testing to help the management arrive on the final version.

A/B Testing

Consider the mean and standard deviation for version 1 distribution as 70 and 5 and for version 2 distribution as 80 and 8.

Objective: Arrive at the most relevant version.

Instructions to Perform:

- Initialize the null and alternate hypothesis
- Convert Version 1 distribution to standard normal distribution
- Convert Version 2 distribution to standard normal distribution
- Compare the Z-scores



Key Takeaways

Now, you are able to:

- Interpret probability distributions
- Perform hypothesis testing using z- tests
- Infer distributions with respect to interval estimate
- Perform A/B testing
- Optimize your pages using results from A/B test

