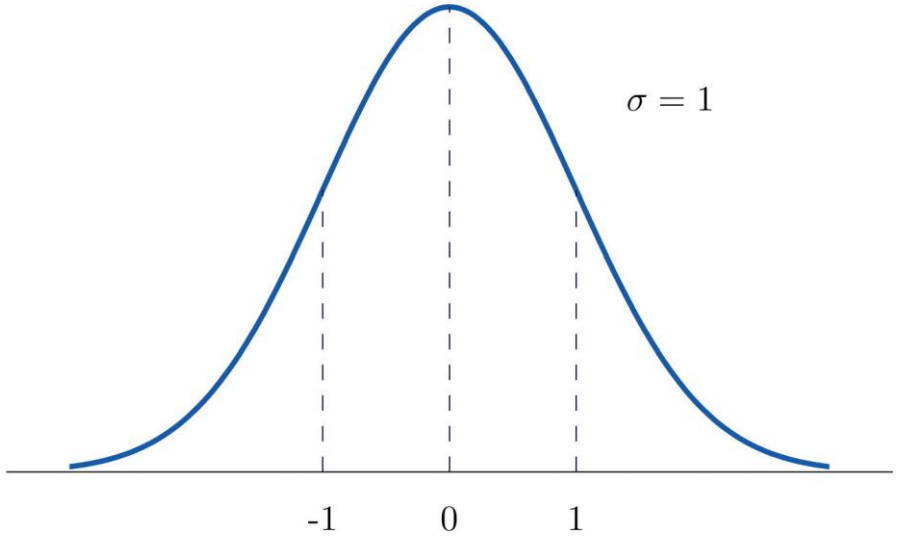
Statistical Tests

Z-score

- Z-score is a distance from the mean
- Z can also represent the area under the curve
  - Given a Z-score, area can be calculated and vice versa
- Converting a dataset into a standard data such that the **Mean=0** and **Standard Deviation=1**
- This will enable us to draw a bell-shape curve that represents the standard normal distribution
- Formula $\mathbf{z = (x-\mu)/\sigma}$
- There can be infinite number of random distributions, but only one standard normal distribution
- z-scores can be Positive(+) or Negative(–)
- Area represented by the z-scores indicate probabilities
- Area cannot be negative



Z-table gives the z-scores and areas

Calculating probabilities from z-scores

Example 1:

- Find the probability that a randomly selected thermometer will have a reading (x) of less than $1.58^{o,}$
  Given $\mu = 0$ and $\sigma = 1$

Example 2:

- Find the probability that a randomly selected thermometer will have a reading of greater than -1.23^{o}
  Given $\mu = 0$ and $\sigma = 1$

Example 3:

- Find the probability that a randomly selected thermometer will have a reading between -2^{o} and 1.5^{o},
  Given $\mu = 0$ and $\sigma = 1$

Calculating z-scores from probabilities

Example 4:

- Find the Z-score that represents the bottom 95% of the data

Example 5:

- Find the Z-score that represents the area between the top 2.5% and the bottom 2.5%

Confidence Interval

- A range used to fit a **population parameter** (***μ:*** population mean, ***σ:*** population std dev)
- Samples always vary with the actual value of a population and cannot guarantee the value
- Create a range of values that tells the value lies within that range
  - ✓ e.g: number of visitors in a resort during the rainy season is between 550 – 750
- Confidence Interval has a **confidence level (%)**
- The confidence level tells us with what confidence (%) we can tell that actual population parameter will fall in the given range.
  - ✓ e.g. The car company is (**95% / 97% / 99%)** sure that the new model will have a mileage of **21-25**
- Higher confidence levels produce higher confidence intervals
- Commonly used confidence intervals are: **.90**, **.95** and **.99**
- Complement of CI is represented by **ά (1-CI)**

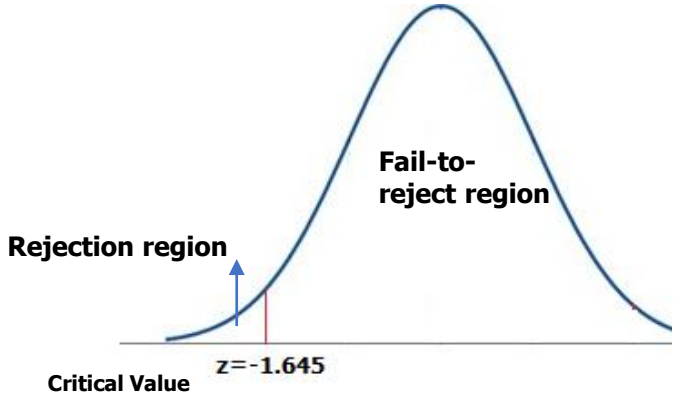
CI	ά (1-CI)
.90	0.1
.95	0.05
.99	0.01

Critical value

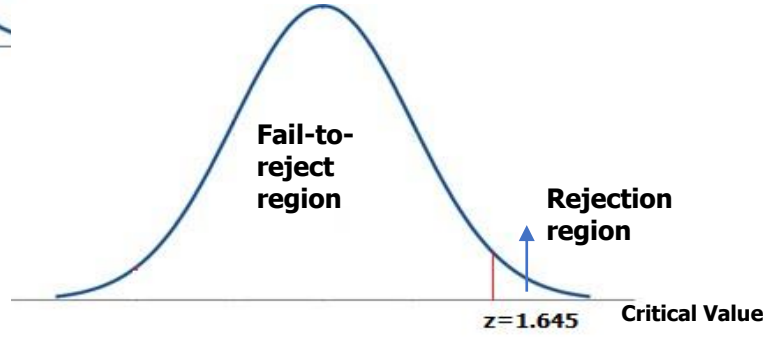
z-score that separates the likely region (**Reject region**) from the unlikely region (**Fail-to-reject region**)

CI	ά (1-CI)	Critical value
.90	0.1	1.289
.95	0.05	1.645
.99	0.01	2.575

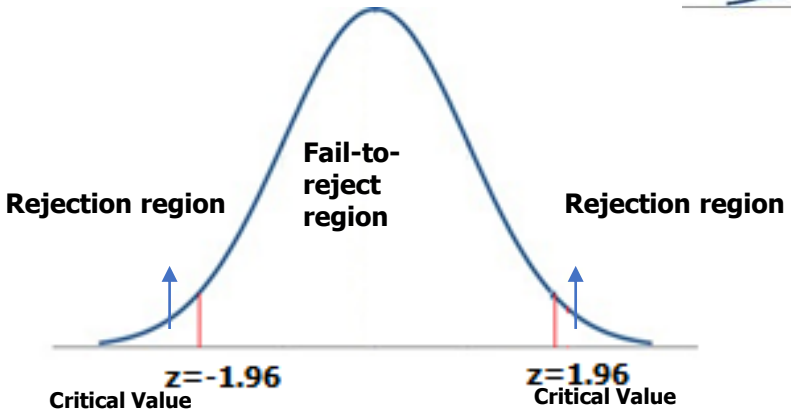
3 types of Hypothesis testing



I) Left-tail test



II) Right-tail test



III) 2-tail test

Choice of tail depends upon the choice of :

- **NULL Hypothesis**
- **ALTERNATE hypothesis**

Hypothesis testing

- Test whether a claim is valid or not **of a population**

- **Purpose of Hypothesis Testing**
  - Make a judgement about the difference between the sample statistic and hypothesized population parameter
  - (HT is not to question the computed value of the sample statistic)

- **Examples of claims**
  - Most people get jobs through networking (proportion)
  - The average number of trucks passing through this highway in a day is 355 (mean)

Claim	Hypothesis
It is a well-founded statement that is proven or is obvious	It is a statement where the truth is not known at the time of its formulation

Parts of Hypothesis testing

NULL hypothesis	ALTERNATE hypothesis
Represented by $\mathbf{H_0}$	Represented by $\mathbf{H_1}$
H_1 states that the population parameter (mean, proportion) is **EQUAL TO** some value $\mathbf{H_0}$: $\mu = 5.5$ $\mathbf{H_0}$: $p = 0.45$	H_1 states that the population parameter (mean, proportion) is **DIFFERENT** than H_0 $\mathbf{H_1}$: $\mu > 5.5$ $\quad$ $\mathbf{H_1}$: $p > 0.45$ $\mathbf{H_1}$: $\mu < 5.5$ $\quad$ $\mathbf{H_1}$: $p < 0.45$ $\mathbf{H_1}$: $\mu \neq 5.5$ $\quad$ $\mathbf{H_1}$: $p \neq 0.45$

H_0 and H_1
together cover all the possible values of the population parameter

Assume H_0 is TRUE
Unless there is enough evidence to prove H_1

Innocent unless proven guilty

H_0: Innocent
H_1: Guilty

Deducing the hypothesis
If **Reject** H_0, then Accept H_1
If **Fail to Reject** H_0, it is inconclusive
(fail to accept H_1)

How to test a hypothesis

1. State the Claim

The Average battery life is 4 years

2. State the Opposite Claim

The average battery life is not 4 years

3. Form the NULL hypothesis (H_0)

$H_0: \mu = 4$

4. Form the Alternate Hypothesis (H_1)

$H_1: \mu \neq 4$

5. Identify the tail (Left, Right, 2-tail)

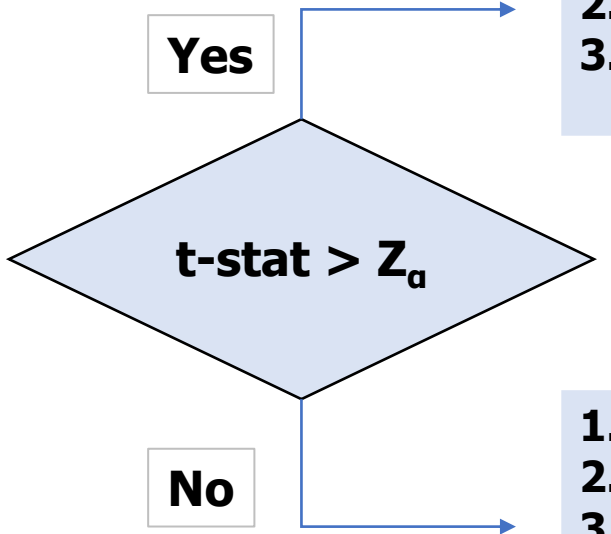
>: Right, <: Left, ≠: 2-tail

6. Calculate the T-statistic

T-stat

7. Validate t-stat against Z_α

T-stat vs Z_α



1. **Reject H_0**
2. **Accept H_1**
3. **There is enough evidence to support the (opposite) claim**

1. **Fail To Reject H_0**
2. **Cannot accept H_1**
3. **There is not enough evidence to support the claim**

Forming H_0 and H_1 for Hypothesis testing

Claim	Step 1 (State the claim)	Step 2 (State the opposite)	Step 3 (Identify H_0) (H_0 is where there is =)	Step 4 (Identify H_1)	Notes
The mean of a liquid is at least 12 oz in a can	$\mu >= 12$	$\mu < 12$	$H_0 : \mu = 12$	$H_1 : \mu < 12$	The claim is H_0
Most school principals are females	$p > 0.5$	$p <= 0.5$	$H_0 : p = 0.5$	$H_1 : p > 0.5$	The claim is H_1
The mean IQ score of a given class is 100	$\mu = 100$	$\mu \neq 100$	$H_0 : \mu = 100$	$H_1 : \mu \neq 100$	The claim is H_0

H_0: There is a fire

H_1: There is no fire

Fact	Decision	Result
H_0 is TRUE	Do not Reject H_0	Correct Decision
H_0 is TRUE	Reject H_0	Type I error
H_0 is FALSE	Do not Reject H_0	Type II error
H_0 is FALSE	Reject H_0	Correct Decision

Types of Tests

Parametric

- **Variable drawn from a normal distribution**
- **Variable is Interval / Ratio**

One – sample test
Z-test
t-test

2– sample test
Z-test
t-test
Paired t-test

>2-sample test
ANOVA

Non-Parametric

- **Variable does not follow any distribution**
- **Variable is Nominal**

Chi-Square

Assumption of Z-test and t-test

- Population has a normal distribution
- Random sampling from a defined population
- ***t-test is done when***
  - Sample size is less than 30
  - Population SD is unknown
  - t-critical value is determined from the ***t-table***

One – sample test

Formulas for 1-sample Test Statistic

Non-parametric test

Proportion (P)

$$Z = (\hat{p} - p) / (\sqrt{p.q} / \sqrt{n})$$

$\hat{p}$ = sample proportion
p = H_0
q = $1 - p$
n = sample size

Parametric test

Mean (μ)

$$Z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$$
$$T = (\bar{x} - \mu)/(s/\sqrt{n}) \text{ (when } \sigma \text{ not provided)}$$

$\bar{x}$ = sample mean
μ = H_0
σ = Population SD
s = Sample SD
n = sample size

Example 1

A sample of 40 sales receipts from a supermarket store has mean of Rs.1,410. The standard deviation of the sales is Rs. 302. Use these values to test whether or not the mean sales at the grocery store are different from Rs.1,500.

Choose α= 5%.

Will your answer change if you choose α= 10%?

Example 2

A sample of 706 companies found that 61% of CEO's were male. Test the hypothesis that the claim (Most CEO's are males) is appropriate.

Take significance level **(ά)** as 0.05

Example 3

Given a sample mean of 83, sample standard deviation of 12.5 and a sample size of 22, test the Hypothesis that the value of the population mean is 70 against the alternative that it is more than 70. Use 0.025 critical value

Example 4

The average number of goals at a national level scored by a team in a game is 5.7. A coach selects five random games of his team and the scores were 5,8,11,4 and 9. Are these scores similar to the national average ? (Alpha level is 0.05)

Example 1

A sample of 40 sales receipts from a supermarket store has mean of Rs.1,410. The standard deviation of the sales is Rs. 302. Use these values to test whether or not the mean sales at the grocery store are different from Rs.1,500.

Choose α= 5%.

Will your answer change if you choose α= 10%?

Example 2

A sample of 706 companies found that 61% of CEO's were male. Test the hypothesis that the claim (Most CEO's are males) is appropriate.

Take significance level **(ά)** as 0.05

Practice questions

In the population, the average IQ is 100 with a standard deviation of 15. A team of scientists want to test a new medication to see if it has either a positive or negative effect on intelligence, or not effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect intelligence?

A professor wants to know if her introductory statistics class has a good grasp of basic math. Six students are chosen at random and given a math proficiency test. The professor wants the class to be able to score above 70 on the test. The six students get the following scores:62, 92, 75, 68, 83, 95.
Can the professor have 90% confidence that the mean score for the class on the test would be above 70.

Two – sample test

Testing equality of 2 population proportions

- To test whether P1 = P2
- When 2 samples are taken from 2 distinct populations
- To determine if the difference between the 2 sample proportions is negligible

- **Formula**

$$t_{stat} = \frac{p_1 - p_2}{\sqrt{\bar{P}\bar{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$p_i = \frac{x_i}{\eta_i} \qquad \bar{P} = \frac{x_1 + x_2}{n_1 + n_2} \qquad \bar{Q} = 1 - \bar{P}$$

where

p_1 = sample 1 proportion
p_2 = sample 2 proportion
n_1 = sample size 1
n_2 = sample size 2
$\overline{P}$, $\overline{Q}$ = pooled estimate

Example 1:

In a random sample of 800 people from rural area, 200 were found to be smokers. From 1000 people from urban, 350 were smokers.

Test whether the proportion of smokers is same for both the populations at 95% CI

Testing equality of 2 population Means - 1

- To test whether $\mu_1 = \mu_2$ (if means are equal)
- When 2 samples are taken from 2 distinct populations
- The Standard Deviation (σ) of the 2 populations are given
- To determine if the difference between the 2 sample means is negligible
- **The two variances are not equal**

- **Formula**

$$t_{stat} = \frac{\overline{x_1} - \bar{x}_2 - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where

$\bar{x}_1$ = sample 1 mean
$\bar{x}_2$ = sample 2 mean
d_0 = hypothesized difference (0 if $\mu_1 = \mu_2$)
σ_1 = Std Deviation of first population
σ_2 = Std Deviation of second population
n_1 = sample size 1
n_2 = sample size 2

Example

A trace element in blood sample varies by 14.1 and 9.5 ppm in males and females respectively. Random samples of 75 male and 50 females gives a mean of 28 and 33 ppm.

What is the likelihood that the population means of the concentrations of elements are same for both ? Test the significance at 95% Confidence level.

Testing equality of 2 population Means - 2

- To test whether $\mu_1 = \mu_2$ (if means are equal)
- When 2 samples are taken from 2 distinct populations
- The Standard Deviation (σ) of the 2 populations are given
- To determine if the difference between the 2 sample means is negligible
- **The two variances are equal (or assumed to be equal)**

- **Formula**

$$t_{stat} = \frac{\overline{x_1} - \bar{x}_2 - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s^2{}_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

$$sp = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}}$$

where

$\mathbf{\bar{x}_1}$ = sample 1 mean
$\mathbf{\bar{x}_2}$ = sample 2 mean
$\mathbf{d_0}$ = hypothesized difference (0 if $\mu_1 = \mu_2$)
$\mathbf{s^2_p}$ = Spooled Variance
$\mathbf{n_1}$ = sample size 1
$\mathbf{n_2}$ = sample size 2

Example

The data set contains miles per gallon for U.S. cars (sample 1) and for Japanese cars (sample 2). The summary statistics for each sample are shown below.

SAMPLE 1: Observations = 249 Mean = 20.14458 SD = 6.41470
SAMPLE 2: Observations = 79 Mean = 30.48101 SD = 6.10771

Test the hypothesis that the population means are equal for the two samples.
We assume that the variances for the two samples are equal.

Paired (Dependent) t-test

- To test whether $\mu_1 = \mu_2$ (if means are equal) before and after a certain procedure. Eg:
  - Student's scores before and after a test
  - Blood sugar level before and after a fast

- **Formula**

$$t_{stat} = \frac{\sum D}{\sqrt{\frac{n \sum D^2 - (\sum D)^2}{n-1}}}$$

where

D = differences between the two observations

n = number of samples

Example

The table shows student's marks before and after a certain course. Determine if there is any relation between the 2 samples.

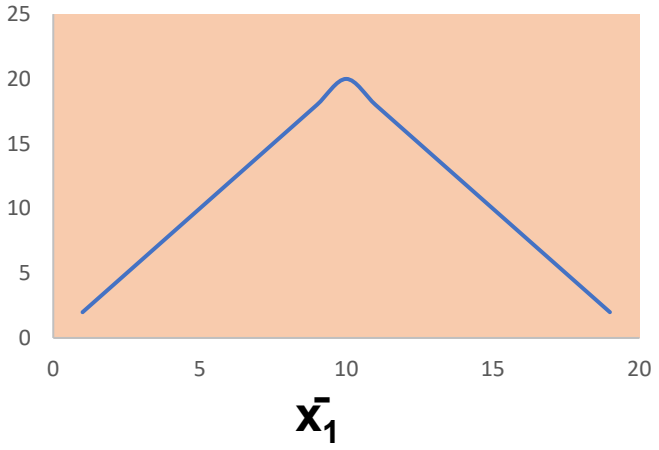
Test the significance at 95% Confidence level.

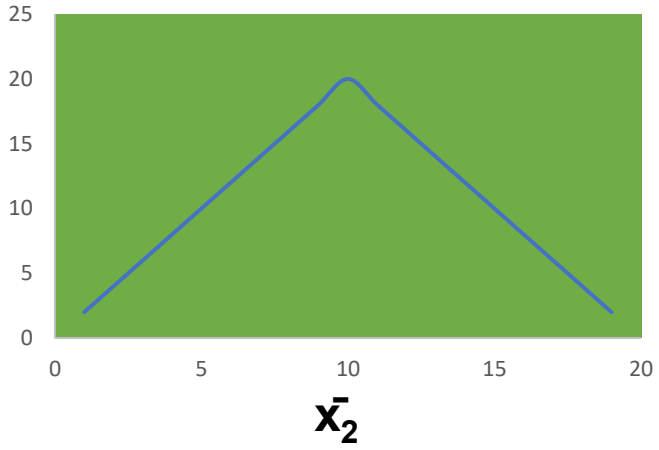
pre-test	post-test
23	35
25	40
28	30
30	35
25	40
25	45
26	30
25	30
22	35
30	40
35	40
40	35
35	38
30	41

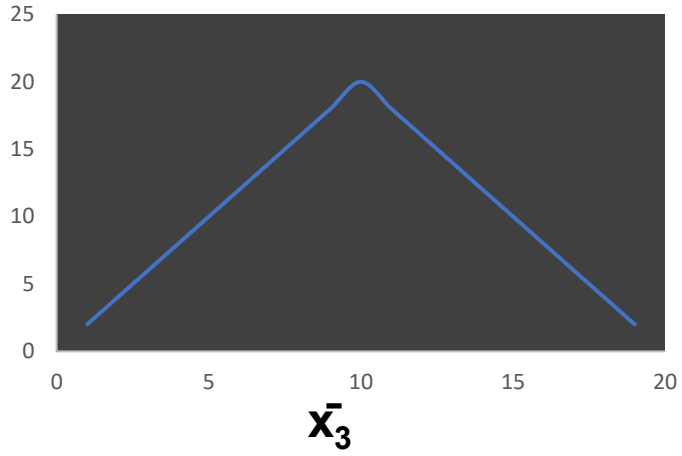
ANOVA

ANOVA – ANalysis Of VAriance

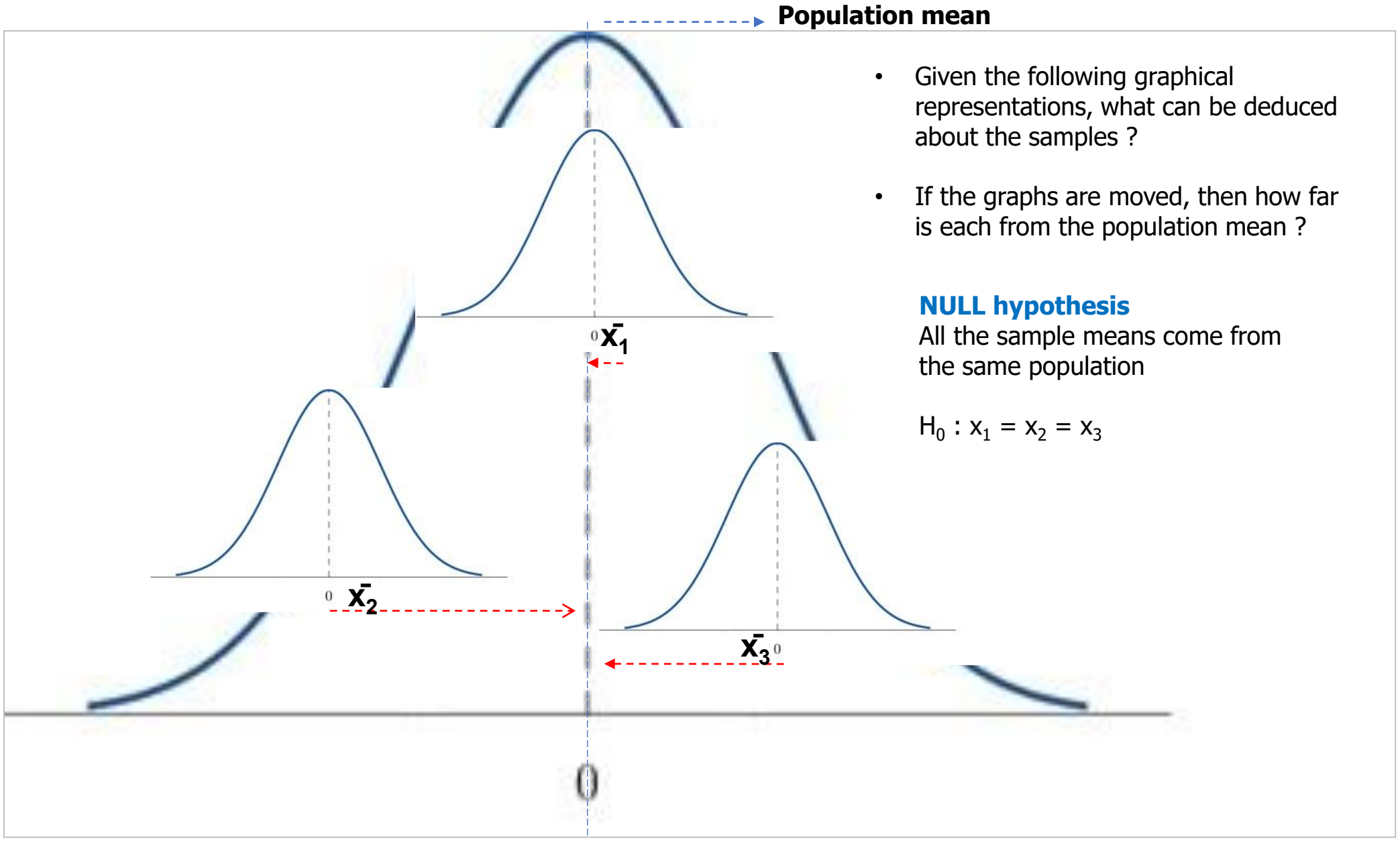
- To compare the means of more than two populations
- Test the significance of differences among more than 2 sample means







- Are these samples drawn from populations having the same mean ?
- Is there a difference in these means ?
- Calculating the relative difference between the means
- **Example:**
  - Comparing the mileage of five different vehicles
  - First-year earnings of graduates of a dozen different business schools
  - Comparing the average life expectancies of 10 different countries
  - etc



- Given the following graphical representations, what can be deduced about the samples ?
- If the graphs are moved, then how far is each from the population mean ?

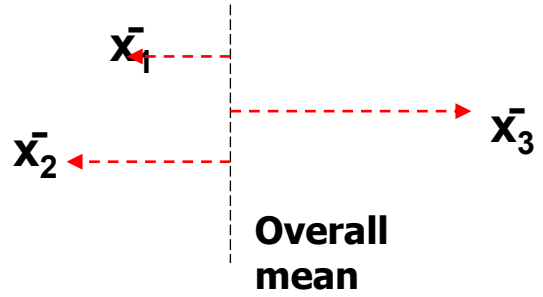
NULL hypothesis

All the sample means come from the same population

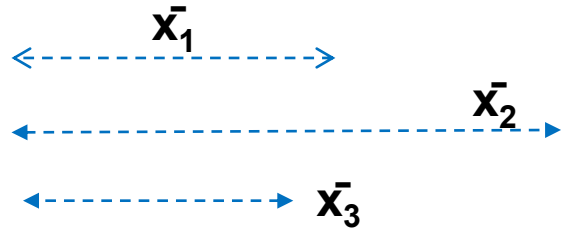
$H_0 : x_1 = x_2 = x_3$

- There will be variability between the sample means
- Every sample mean will be away by some distance from the overall population mean
- **ANOVA (F-ratio)** is a variability ratio

- Variability among/between the means
- Distance from overall mean
- AMONG variance



- Variability around/within distributions
- Internal distance
- AROUND variance



ANOVA (F) Formula

$$F = \frac{\frac{n_1(\bar{x}_1-\bar{X})^2+n_2(\bar{x}_2-\bar{X})^2+\cdots}{(k-1)}}{\frac{\sum(x_{1i}-\bar{x}_1)^2+\sum(x_{2i}-\bar{x}_2)^2+\cdots}{(N-k)}}$$

Among Variance

Around Variance

where
$\mathbf{n_n}$ = number of observations of each sample
$\bar{\mathbf{x}}_\mathbf{n}$ = sample of each mean
$\bar{\mathbf{X}}$ = mean of all sample means
k = number of sample groups
k-1 = dof (numerator)

$\mathbf{X_{nk}}$ = individual values
N = total sample count
N-k = dof (denominator)

Example

A factory manager in an engineering firm wants to evaluate the conveyor belt's productivity at different speeds for an 8-hour shift. The belts ran at different speeds and the number of defective pieces were recorded as follows:

At a significance level of 0.05, determine whether the four belts produce the same mean rate of defective products in a given 8-hour shift?

Defective unit			
Speed 1	**Speed 2**	**Speed 3**	**Speed 4**
37	27	32	35
35	32	36	27
38	32	33	33
36	34	34	31
34	30	40	29

Chi-Square test of Independence

- Chi-square test helps to understand the **relationship** between two **categorical variables**
  e.g. Does the **field of Education (X)** play any role in **Employee Attrition (Y)**
  Are these variables **"statistically"** independent ?

- **Hypothesis**
  - ✓ H_0 : The two categorical variables are independent / No relation exists
  - ✓ H_1: The two categorical variables are dependent / Relation exists

- Involves counting of categories (frequencies of events)
- Compares **Observed vs Expected** using population data
- Chi-Square helps in determining the role of random chance variation between the categorical variables
- Uses Chi-Square distribution and Critical value to reject / Fail-to-reject H_0

Formula

χ^2 = (**O**bserved – **E**xpected)^2 / **E**xpected
Degrees of freedom (DF) = (# columns -1) x (# rows – 1)

Exercise

A researcher is examining whether the incidence of Malaria is same in cities Mumbai and Delhi. Sample from both the cities are taken and are as follows.

Test the hypothesis that the incidence is same in both the cities at $\alpha = 5\%$

	Malaria	No Malaria
Mumbai	12	488
Delhi	7	393