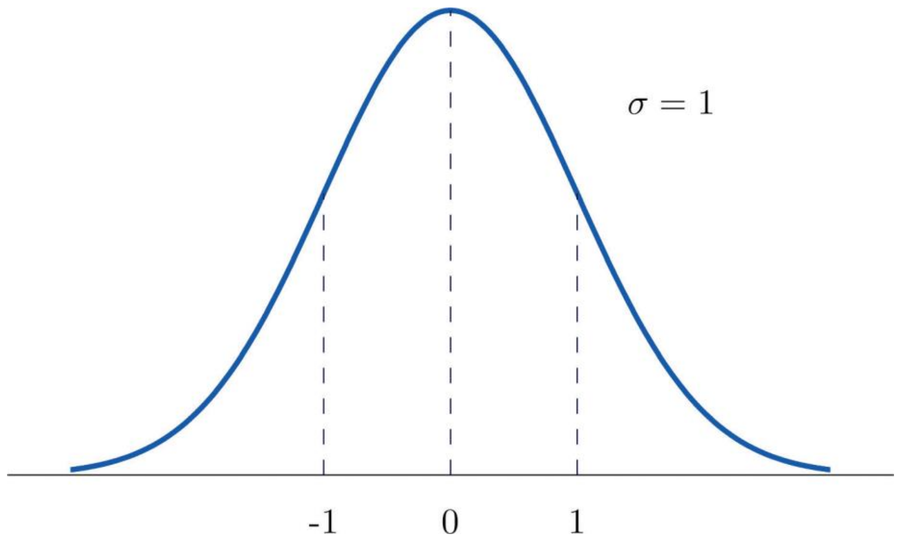
Statistical Tests

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 **$z = (x-\mu)/\sigma$**
- 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

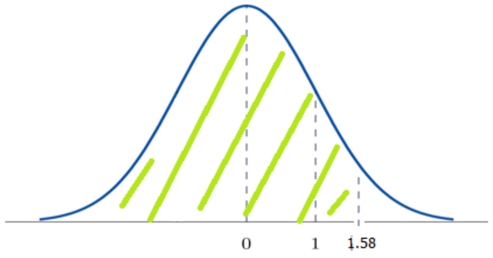
Calculating probabilities from z-scores

Example 1:

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

Answer

1) $z = (1.58-0) / 1 = 1.58$
2) From the z-table, find the probability value for 1.58
3) Answer = .9429 i.e. 94.29%
4) ∴ The probability of a thermometer reading less than 1.58 is 94.29%

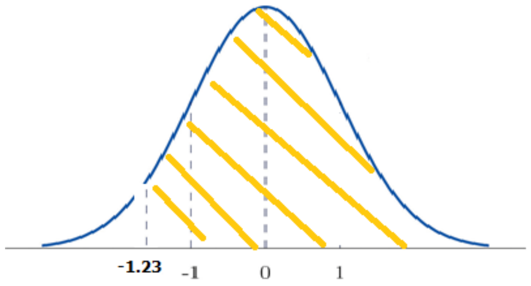


Example 2:

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

Answer

1) $z = (-1.23-0) / 1 = -1.23$
2) From the z-table, find the probability value for -1.23
3) Answer = .1093
4) This is area to the left. The area greater than -1.23 is to the right. i.e. 1-0.1093 = .8907
5) ∴ The probability of a thermometer reading greater than -1.23 is 89.07%

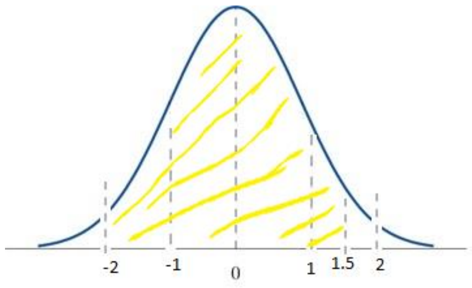


Example 3:

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

Answer

1) $z_1 = (-2-0) / 1 = -2$ **p(-2)** = 0.0228
2) $z_2 = (1.5-0) / 1 = 1.5$ **p(1.5)** = 0.9332
3) Area between z_1 and z_2 => 0.9332 – 0.0228 = 0.9104
4) ∴ The probability of a thermometer reading between -2 and 1.5 is 91.04%



Calculating z-scores from probabilities

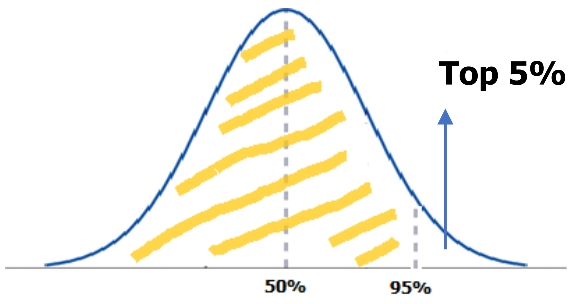
Example 4:

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

Answer

Bottom 95% == Top 5%

1) Probability = 0.9500 (represents the area)
2) From the Z-table, z-score (p = .9500) = 1.645
3) ∴ 95% of the thermometers will have a value of 1.645 or less



Example 5:

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

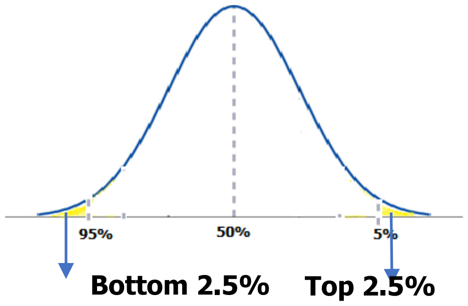
Answer

2.5% = 0.025

Bottom 2.5% = -1.96

Top 2.5% = 1-0.025 = 0.975 = +1.96

Area between top 2.5 and bottom 2.5 = 0.975 – 0.025 = **0.95 (z = 1.645)**



Example 6:

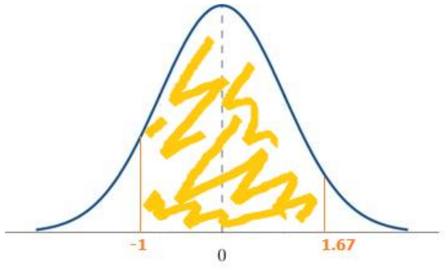
- What percentage of people have an IQ between 85 and 125, given that the IQ is normally distributed with mean=100 and SD=15

Answer

z1 = (85-100)/15 = -1. **p =** 0.1587

z2 = (125-100)/15 = 1.67 **p =** 0.9525

Area between z1 and z2 = 0.9525 – 0.1587 = **0.7938 (z=0.82)**



Confidence Interval

Example

A poll is taken of 1000 customers asking for their experience relating to their interaction with the customer care. 450 people reported as good experience.

- **$\hat{p}$ (sample proportion)** = 450/1000 = 0.45 or 45% people reported as good experience.
- This **$\hat{p}$** estimates the population proportion **p** (which is the parameter, and is unknown)

How close is $\hat{p}$ to p ?

- What are the likely values of p ?
- Cannot say with certainty how close is $\hat{p}$ to p
- But need some kind of estimate that would determine how close are $\hat{p}$ and p
- This estimate is the "confidence interval"

Confidence Interval

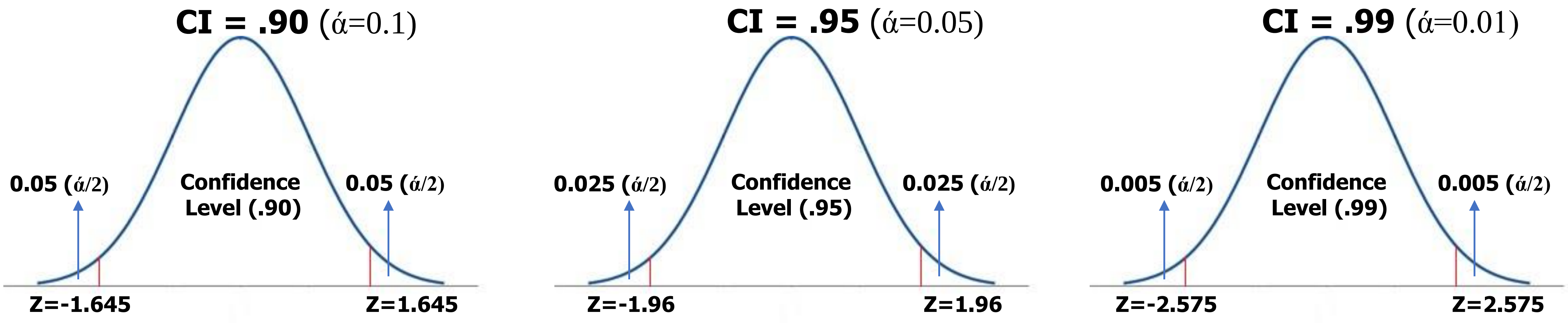
- A range used to fit a **population parameter** (***μ:*** population mean, ***σ:*** population std dev)
- Samples always vary with the actual value of a population and cannot guarantee the value
- Create a range of values that tells the value lies within that range
  - ✓ e.g: number of visitors in a 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**)

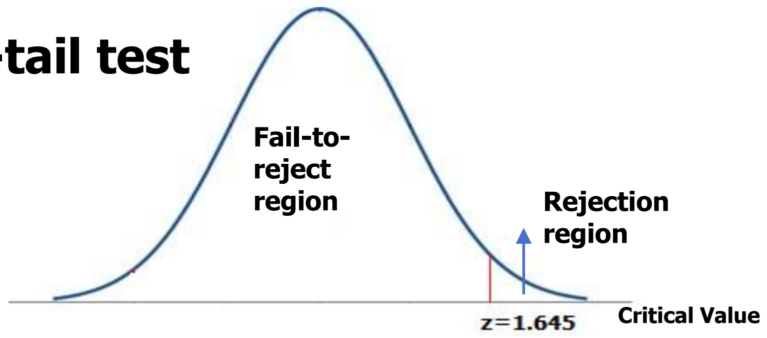
z-scores for Confidence Intervals



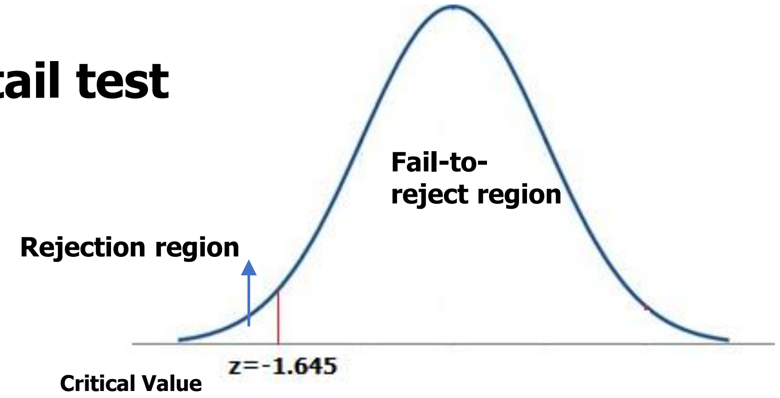
CI	ά (1-CI)	ά/2	Z-score ($Z_{\alpha/2}$)
.90	0.1	0.05	1.645
.95	0.05	0.025	1.96
.99	0.01	0.005	2.575

3 types of Hypothesis testing

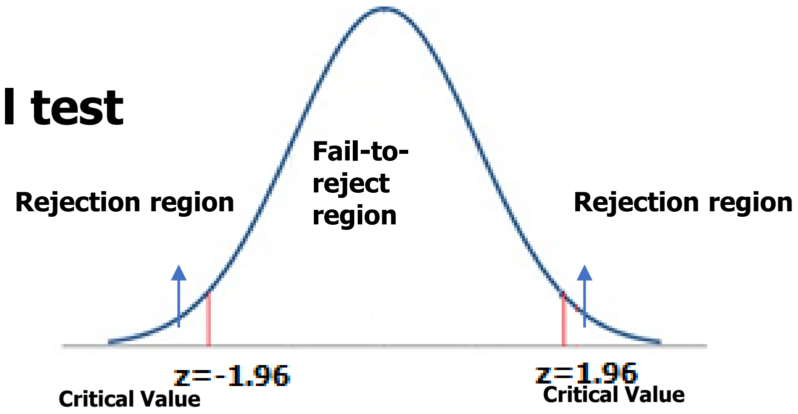
I) Right-tail test



II) Left-tail test



III) 2-tail test



Choice of tail depends upon the choice of :

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

Hypothesis testing

- Test whether a claim is valid or not ***<u>of a population</u>***
- **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)

- In Statistics, you cannot prove anything right. It is only wrong or not wrong. (just like a court verdict of 'guilty' or 'not guilty', but never 'innocent'

Parts of Hypothesis testing

NULL hypothesis

- NULL represented by $\mathbf{H_0}$

- H_0 states that the population parameter (mean, proportion) is **<u>EQUAL TO</u>** some value
- eg: $H_0 : \mu = 5.5$, $H_0 : p = 0.45$

ALTERNATE hypothesis

- ALTERNATE represented as $\mathbf{H_1}$

- H_1 states that the population parameter (mean, proportion) is **<u>DIFFERENT</u>** than H_0
  - ✓ $H_1: \mu > 5.5$ $H_1: p > 0.45$
  - ✓ $H_1: \mu < 5.5$ $H_1: p < 0.45$
  - ✓ $H_1: \mu \neq 5.5$ $H_1: p \neq 0.45$

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

Steps

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)

$\mu = 4$

4. Form the Alternate Hypothesis (H_1)

$\mu \neq 4$

5. Calculate the T-statistic

T-stat

6. Validate t-stat against Z-score for α

T-stat vs Z_α

Result

If t-stat > Z_α

1. Reject H_0 (NULL Hypothesis)
2. There **is enough evidence** to support the claim

If t-stat < Z_α

1. Fail To Reject H_0 (NULL Hypothesis)
2. 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

Formulas for 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)

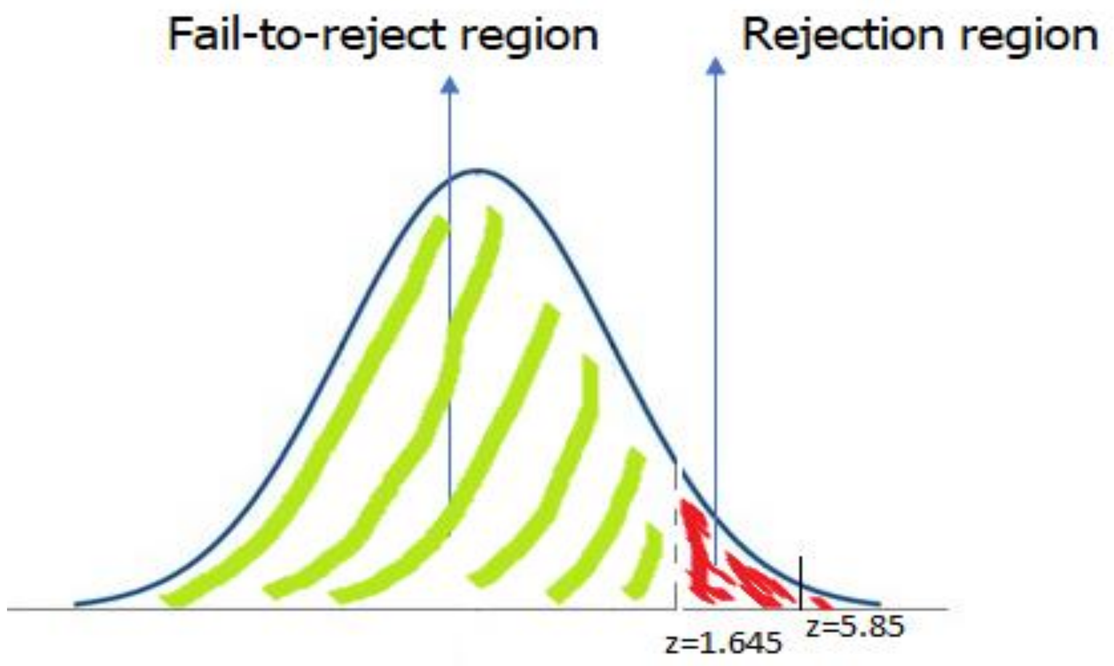
$$Z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$$

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

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

Example 1: A sample of 706 companies found that 61% of CEO's were male. **Claim:** Most CEO's are males
Take significance level **($\acute{\alpha}$)** as 0.05

1	**Claim: Most CEO's are male ($p > 0.5$)**
2	**Opposite: Most CEO's are not males ($p <= 0.5$)**
3	**$H_0 : p = 0.5$ (NULL hypothesis)**
4	**$H_1 : p > 0.5$ (ALTERNATE hypothesis)**
5	**$\hat{p} = 0.61$**
6	**$q = 1-p = 0.5$**
7	**$n = 706$**
8	**$Z = (\hat{p} - p) / (\sqrt{p.q} / \sqrt{n})$ $= (0.61-0.5) / (\sqrt{(0.5*0.5)} / \sqrt{706})$ $= (0.11) / (0.0188)$ $= 5.85$ (Test statistic value)**
9	**Based on the value of $Z=5.85$, can we reject H_0 ? Need to make this decision based on evidence**



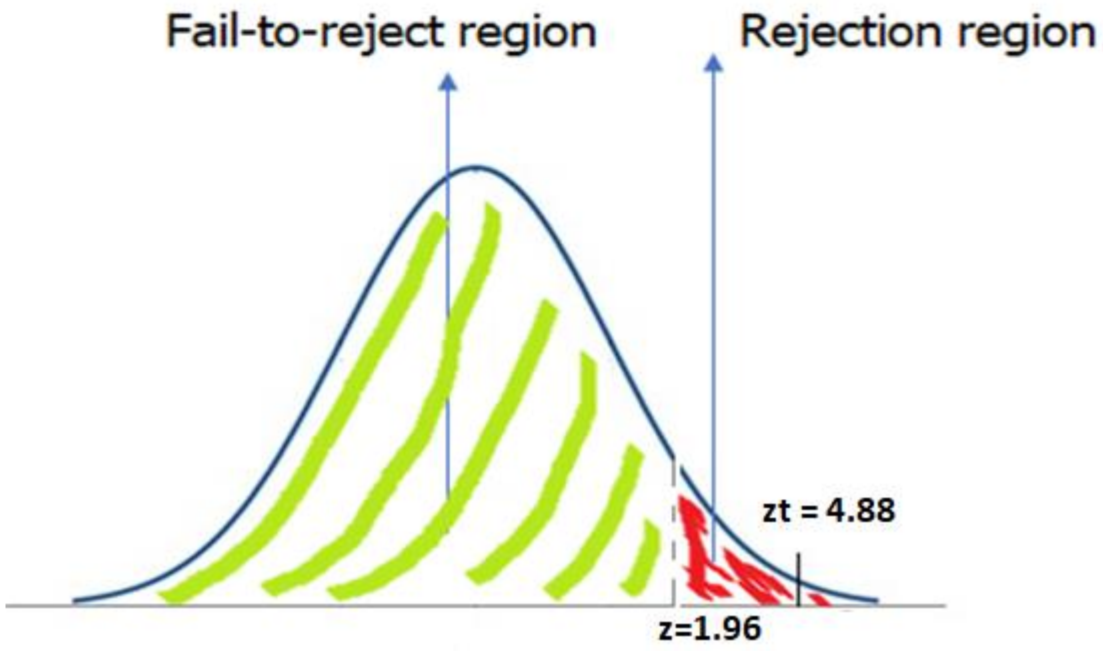
Since the test statistic **(5.85)** lies in the rejection region, we reject the NULL hypothesis.

i.e. there is enough evidence to support the claim that **most CEOs are males**

Exercise (identify the mistake)

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. Use 0.025 significance level

1	**Claim: Value of population mean is 70**
2	**Opposite: Value of mean is more than 70**
3	**$H_0 : \mu = 70$**
4	**$H_1 : \mu > 70$**
5	**$\bar{x} = 83$**
6	**$s = 12.5$**
7	**$n = 22$**
8	**$Z = (\bar{x} - \mu)/(s/\sqrt{n})$ $= (83-70) / (12.5 / \sqrt{22})$ $= 4.88$**
9	**Based on the value of Z=4.88, can we reject H_0 ? Need to make this decision based on evidence**



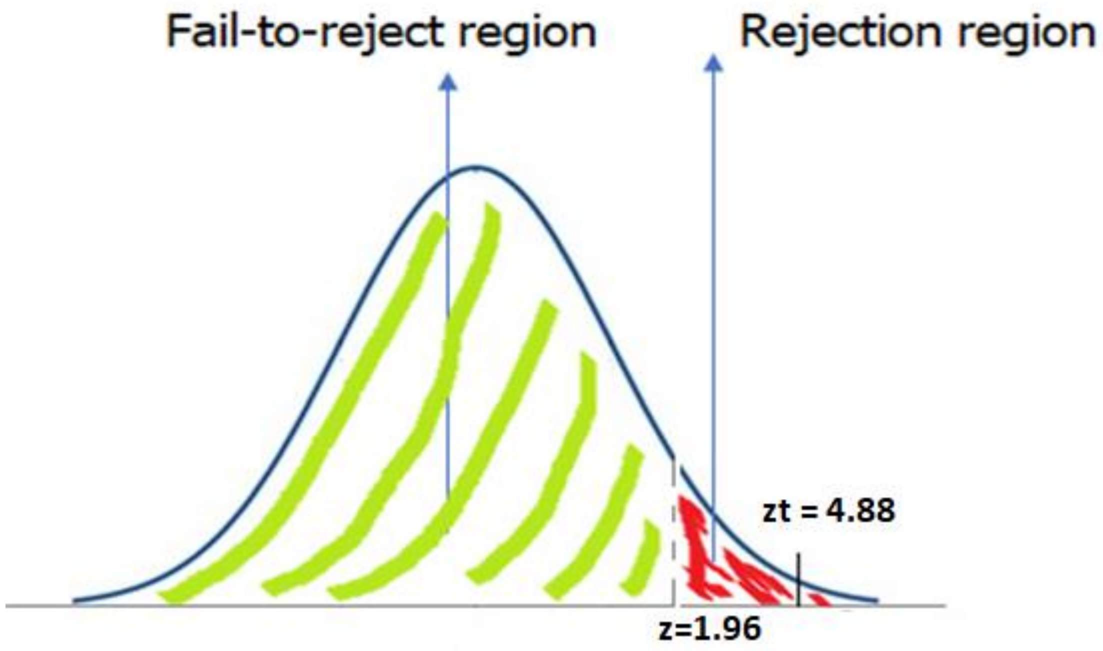
Since the test statistic **(4.88)** lies in the rejection region, we reject the NULL hypothesis.

i.e. there is enough evidence to prove the **population mean is more than 70**

Exercise

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 significance level

1	**Claim: Value of population mean is 70**
2	**Opposite: Value of mean is more than 70**
3	$H_0 : \mu = 70$
4	$H_1 : \mu > 70$
5	$\bar{x} = 83$
6	$s = 12.5$
7	$n = 22$
8	$Z = (\bar{x} - \mu)/(s/\sqrt{n})$ $= (83-70) / (12.5 / \sqrt{22})$ $= 4.88$
9	**Based on the value of Z=4.88, can we reject H_0 ?** **Need to make this decision based on evidence**



Since the test statistic **(4.88)** lies in the rejection region, we reject the NULL hypothesis.

i.e. there is enough evidence to prove the **population mean is more than 70**

Type I and II errors

- Does our conclusion of our hypothesis testing match the actual value ?
- More often that not, there will be some deviation from the actual.
- This deviation is known as the "Type I and Type II errors"
- These errors can be very critical

Type I error

- It is the ***rejection*** of the NULL hypothesis when ***it should not have been rejected***

Type II error

- ***Fail to reject*** of the NULL hypothesis when ***it should have been rejected***

Type I and II errors

NULL hypothesis
Everything is fine

Fire | **Reality** | No fire

Serious

Reject NULL hypothesis

Get help

Investigate

Result: No fire

Type I error

Rejecting "everything is fine" when "everything was fine"

False alarm

Not serious

Do not Reject NULL

Walk away

Investigate

Result: Fire

Type II error

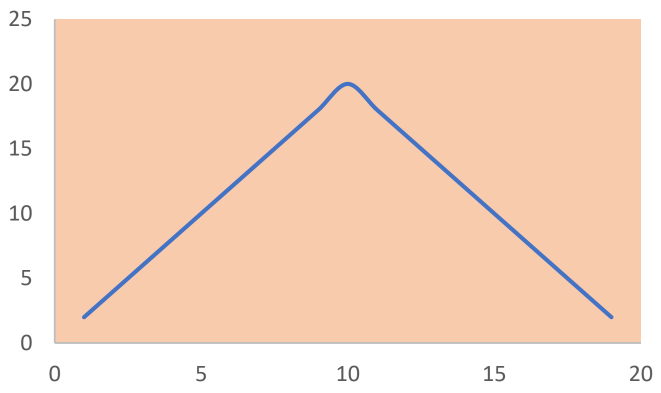
Not Rejecting "everything is fine" when "everything was not fine"

More serious

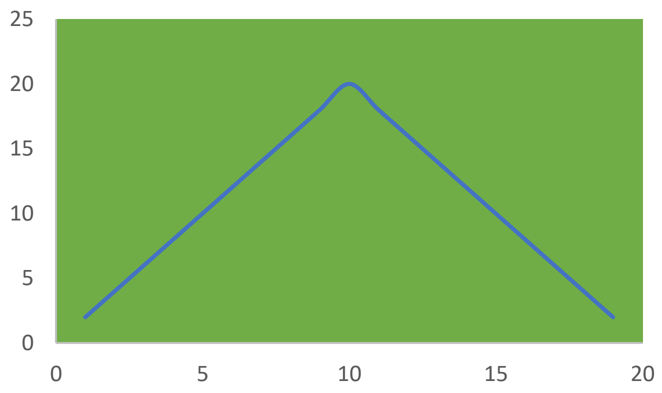
ANOVA

ANOVA – ANalysis Of VAriance

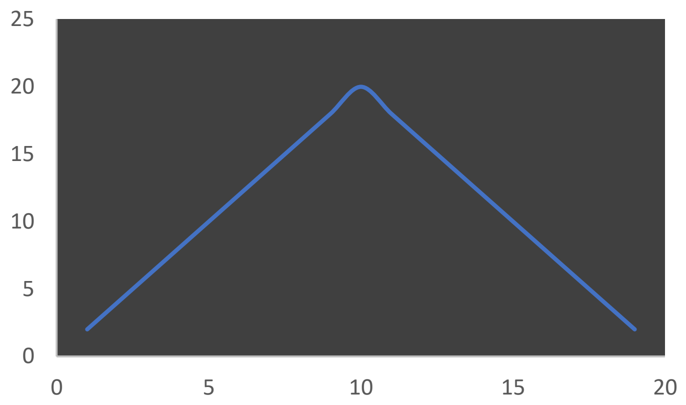
- To compare the means of more than two populations
- Test the significance of differences among more than 2 sample means



$\bar{x}_1$

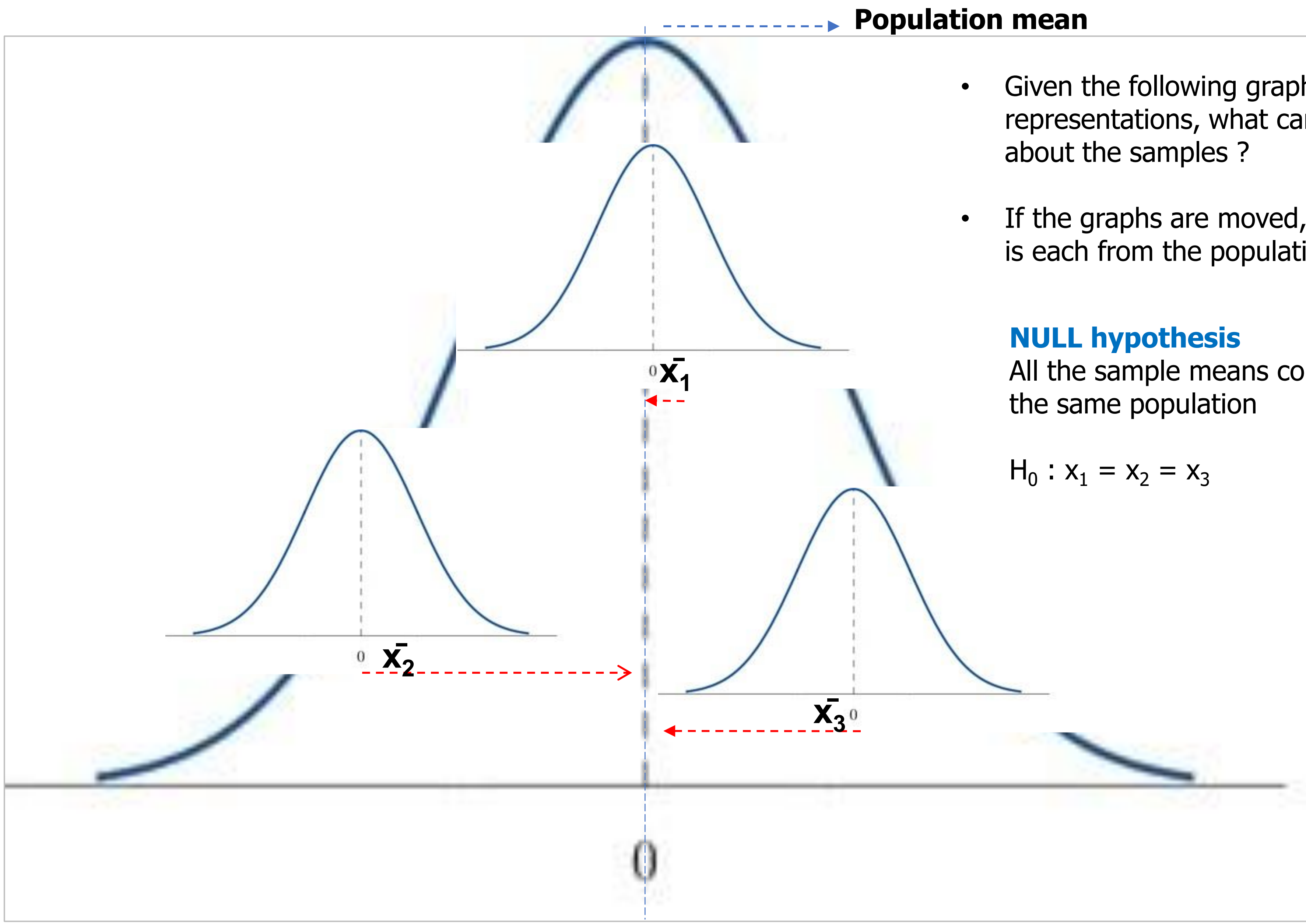


$\bar{x}_2$



$\bar{x}_3$

- 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 ?

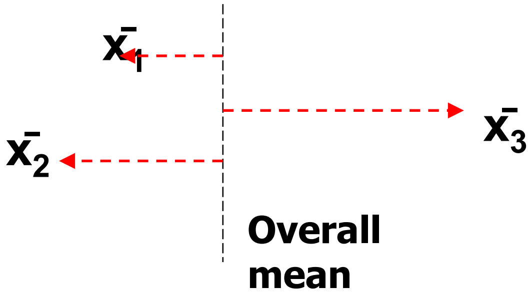
NULL hypothesis

All the sample means come from the same population

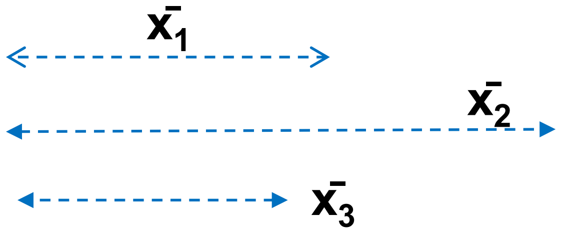
$H_0 : x_1 = x_2 = x_3$

- 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

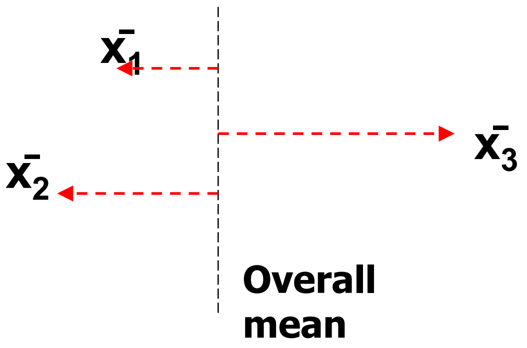


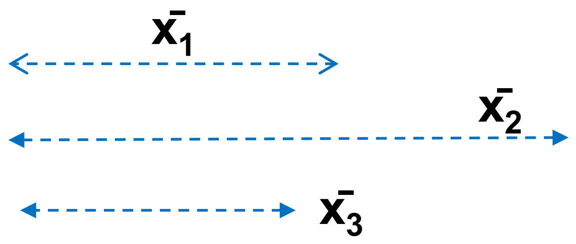
ANOVA (F) Formula

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

$$\sum(x_{1i} - \bar{x}_1)^2 = (n_1 - 1)\, s_1^2$$
$$\sum(x_{2i} - \bar{x}_2)^2 = (n_2 - 1)\, s_2^2$$
$$\sum(x_{3i} - \bar{x}_3)^2 = (n_3 - 1)\, s_3^2$$

where
n_n = number of observations of each sample
$\bar{x}_n$ = sample of each mean
X = mean of all sample means
S_n = standard deviation of each sample
N = total sample count
k = number of sample groups
k-1 = dof (numerator)
N-k = dof (denominator)





- **ANOVA = Variability between / Variability within**

- Hypothesis testing
  - ✓ VB/VW = Large / small → **Reject H_0**
    i.e. at least one mean is an outlier and each distribution is distinct from each other

  - ✓ VB/VW = similar / similar → **Fail to Reject H_0**
    i.e. sample means are probably closer to the overall mean, but cannot say for sure

  - ✓ VB/VW = small / Large → **Fail to Reject H_0**
    i.e. sample means are very closer to the overall mean

Example

- To evaluate 3 different training programs to determine if there are any differences in the effectiveness of the program
- 16 employees are chosen at random to attend these 3 training programs
- The daily production output is given in the table

Program1	Program2	Program3
-	-	18
15	22	24
18	27	19
19	18	16
22	21	22
11	17	15

Hypothesis

To decide whether these 3 samples are drawn from populations (total number of employees trained by the method) having the same mean ***(μ)***

- **Claim:** Training programs increase productivity
- **Opposite:** Training programs don't increase productivity

- **H_0:** $\mu_1 = \mu_2 = \mu_3$
- **H_1:** $\mu_1 \neq \mu_2 \neq \mu_3$

Calculations

ANOVA (F) Formula

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

Step 1 – calculate the mean of each group, count the total groups and mean of means

	TP1	TP2	TP3	
			18	
	15	22	24	
	18	27	19	
	19	18	16	
	22	21	22	
	11	17	15	
n	5	5	6	
Total	85	105	114	
$\bar{x}_n$	**17**	**21**	**19**	
K	total groups			**3**
$\bar{X}$	mean of means			**19**
N	Total sample count			**16**

Step 2 – calculate the "between columns" variance (numerator of F)

$n_1(\bar{x}_1 - \bar{X})^2$	$n_2(\bar{x}_2 - \bar{X})^2$	$n_3(\bar{x}_3 - \bar{X})^2$	between column variance
20	20	0	**20**

Step 3 – calculate the "within sample" variance (denominator of F)

$(x_1 - \bar{x}_1)^2$	$(x_2 - \bar{x}_2)^2$	$(x_3 - \bar{x}_3)^2$	Σ	within sample variance
0	0	1		
4	1	25		
1	36	0		
4	9	9		
25	0	9		
36	16	16		
70	**62**	**60**	**192**	**14.77**

Step 4 – calculate the F-score (numerator / denominator)

F = (bet col / bet samp)
1.354

Excel calculation

Calculations

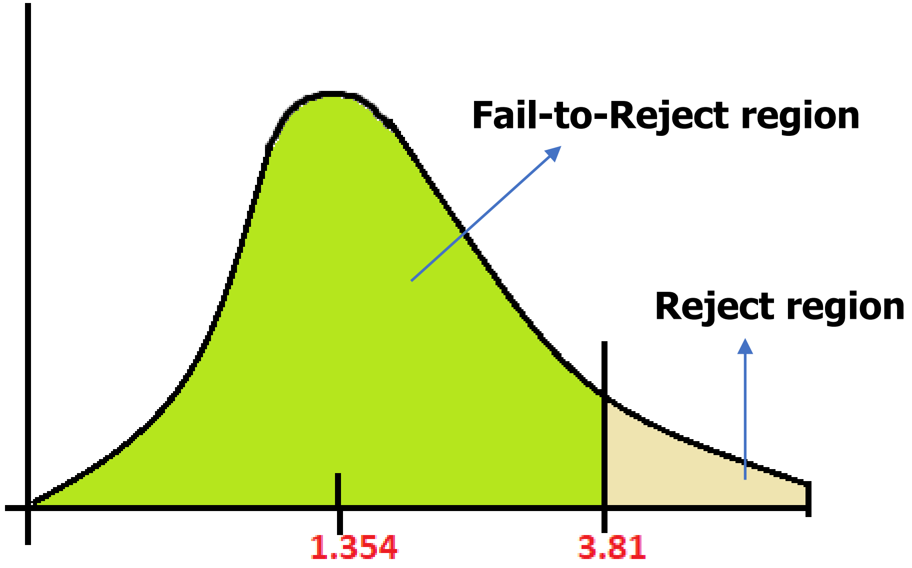
Step 5 – Calculate the degrees of freedom for *Numerator* in F-ratio

(Numerator)Number of groups – 1
3 – 1 = 2

Step 6 – Calculate the degrees of freedom for *Denominator* in F-ratio

(Denominator Σ(sample $_n$ – 1)
(5-1)+(5-1)+(6-1) = 13

Significance level = 0.05, using the Anova table, F = 3.81
$F_{statistic}$ = 1.354



Interpretation

As the F_{stat} (1.354) falls in the FTR region, there is no significant differences in the effects of the 3 training programs on employee productivity

Exercise

The following airline dataset is given with delay time of arrival.

Assumption: All airlines land at the same expected time.

airline1	airline2	airline3
40	56	92
28	48	56
36	64	64
32	56	72
60	28	48
12	32	52
32	42	64
36	40	68
44	61	76
36	58	56

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 not independent / 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

The following sample student dataset shows the credit cards owned by student of each year

Claim: Cards owned by students of each year is the same

	2007	**2008**	**2009**	**2010**	**2011**
Freshman	560	495	553	547	512
Sophomore	369	385	358	361	393
Junior	209	226	248	268	285
Senior	267	277	304	328	340
Unclassified	64	70	93	77	126

Step 1

observed	2007	2008	2009	2010	2011	Total
Freshman	560	495	553	547	512	**2667**
Sophomore	369	385	358	361	393	**1866**
Junior	209	226	248	268	285	**1236**
Senior	267	277	304	328	340	**1516**
Unclassified	64	70	93	77	126	**430**
Total	**1469**	**1453**	**1556**	**1581**	**1656**	**7715**

Step 2

expected	2007	2008	2009	2010	2011	Total
Freshman	507.82	502.29	537.89	546.54	572.46	**2667**
Sophomore	355.30	351.43	376.34	382.39	400.53	**1866**
Junior	235.34	232.78	249.28	253.29	265.30	**1236**
Senior	288.66	285.51	305.75	310.67	325.40	**1516**
Unclassified	81.88	80.98	86.72	88.12	92.30	**430**
Total	**1469**	**1453**	**1556**	**1581**	**1656**	**7715**

2007	2008	2009	2010	2011
(2667*1469)/7715	(2667*1453)/7715	(2667*1556)/7715	(2667*1581)/7715	(2667*1656)/7715

Step 3

(O-E)	2007	2008	2009	2010	2011
	52.18	-7.29	15.11	0.46	-60.46
	13.70	33.57	-18.34	-21.39	-7.53
	-26.34	-6.78	-1.28	14.71	19.70
	-21.66	-8.51	-1.75	17.33	14.60
	-17.88	-10.98	6.28	-11.12	33.70

Step 5

$(O-E)^2/E$	2007	2008	2009	2010	2011
	5.36	0.11	0.42	0.00	6.39
	0.53	3.21	0.89	1.20	0.14
	2.95	0.20	0.01	0.85	1.46
	1.63	0.25	0.01	0.97	0.65
	3.90	1.49	0.45	1.40	12.31

Step 4

$(O-E)^2$	2007	2008	2009	2010	2011
	2722.86	53.11	228.19	0.22	3655.77
	187.64	1126.81	336.51	457.57	56.71
	694.04	45.99	1.65	216.45	387.95
	469.11	72.50	3.08	300.43	213.03
	319.54	120.64	39.38	123.61	1135.82

Step 6

χ^2	46.78
p	0.05
r	5
c	5
df	16
Chi critical	26.30
Conclusion	Reject H_0