Chapter1- Central Tendency (MEAN, MEDIAN, MODE)

1. ARITHMETIC MEAN:

Suppose the principal of your school asks your class teacher that how was the score of your class this time? What do you think is the teacher going to do? Do you think that the teacher is going to actually read out the individual score of all the students? NO!!! What the teacher does is, the teacher will tell the average score of the class instead of saying the individual score. So the principal gets an idea regarding the performance of the students.

It can be denoted as $\frac{\sum n(X)}{n}$ **where** $\sum$ means summation of the given data and n denotes the number of data given.

It can also be denoted as: $\mu = \frac{x1+x2+x3+\cdots+xn}{n}$ where μ stands for Arithmetic Mean

So, it stands for Summation of given data divided by number of data resulting in the average or mean value of the data collectively.

For Example:

In 2016, the tourist population of Destination A is 10,000. In 2017, it increased by 1000 due to addition of a new adventure activity and In 2018, the annual data showed the population grew for increment of 2000 than its previous year 2017. Now, the stakeholder wants to know their progress in attracting more tourists in these 3 years.

So, we have 3 data for 3 consecutive years 2016, 2017 and 2018. Now the average population the destination is attracting will be:

X1 =10,000, X2 = 11,000, X3 = 13,000 and n = 3. So, we will calculate arithmetic mean using formula:

$$\mu = \frac{x1+x2+x3+\cdots+xn}{n}$$

$$\text{So, } \mu = \frac{10000+11000+13000}{3}$$

$$\mu = \frac{24000}{3} = 8000$$

So, the average number of tourist population coming to the destination is 8000.

Partial Exercise:

These numbers are taken from the number of people that attended a particular church every Friday for 7 weeks: 62, 18, 39, 13, 16, 37, 25. Find the mean.

X1 = , X2 = , X3 = , X4= , X5= , X6= , X7 = and n = . So we will calculate arithmetic mean using formula:

$$\mu = \frac{x1+x2+x3+\cdots+xn}{n}$$

$$\text{So, } \mu = \frac{+ \quad + \quad + \quad + \quad + \quad +}{n}$$

$$\mu = — \quad =$$

Exercise 1: The accountant wants to find out a Travel Firm's performance to set new target for the firm for next year, by analyzing its each month sale. There are 12 months starting from financial Year April. Calculate the Average performance by the firm.

April – 1000
May – 1300
June- 2000
July- 2400
August- 1500
September - 800
October- 2000
November - 2100
December - 1600
January - 1800
February- 2000
March - 1900

Exercise 2: A group of travelers set for their Rafting activity but the raft can hold certain weight as per mentioned by the tour guide. The boat can carry weight of people near 70 KG. Before letting people climbing the raft, the guide measured each individual's and got the average 82 KG. The number of people in raft is 6. A solo traveler later joined them, having weight of 100 KG. Will the guide allow him to tagalong.

---

2. GEOMETRIC MEAN:

The geometric mean is an average that is useful for sets of positive numbers that are interpreted according to their product and not their sum (as is the case with the arithmetic mean); e.g., rates of growth.

It can be denoted as $\left(\sum_{i=1}^{n} \boldsymbol{xi}\right)^{\frac{1}{n}}$ n denotes the number of data given.

It can also be denoted as: $\boldsymbol{\mu} = \mathbf{(x1.x2.x3...xn)}^{\frac{1}{n}}$ where μ stands for Geometric Mean

So Geometric Mean is the nth root of the product of n numbers

For Example:

The population in 2016 is 100. In 2017 it is increased by 10% and in 2017 it is increased by 10%.

The Geometric Mean will be:

μ = $\mathbf{(100*110*121)^{\frac{1}{3}}}$ **= 110**

So

the population increased with growth of 110.

Partial Exercise:

Find the geometric mean of 4, 10 and 25.

There are three numbers. So, the geometric mean of the three numbers is the cube root of their product.

Geometric mean = $\sqrt[3]{\quad}$

= $\sqrt[3]{\quad}$

---

3. HARMONIC MEAN:

The harmonic mean is an average which is useful for sets of numbers which are defined in relation to some unit, for example speed (distance per unit of time).

$$\mu = n\left(\sum_i^n \frac{1}{xi}\right)^{-1}$$

For example, the harmonic mean of the five values: 4, 36, 45, 50, 75 is

$$\mu = \frac{5}{\frac{1}{4}+\frac{1}{36}+\frac{1}{45}+\frac{1}{50}+\frac{1}{75}} = \frac{5}{\frac{1}{3}} = 15.$$

Partial Exercise:

Given the following frequency distribution of first year students of a particular college, calculate the harmonic mean.

Age (Years): 13, 14, 15, 16, 17

$$\mu = \frac{n}{\frac{1}{x1}+\frac{1}{x2}+\frac{1}{x3}+\frac{1}{x4}+\frac{1}{x5}} = \frac{n}{\frac{1}{x}} =$$

4. MEDIAN:

Median is the middle number in a sorted list of numbers. To determine the median value in a sequence of numbers, the numbers must first be arranged in value order from lowest to highest.

For a given set of number of tourists staying in a 3 star hotel is given below:

We need to find the approximate number of tourists staying.

Day 1: 10

Day 2: 20

Day 3: 05

Day 4: 01

Day 6: 15

So, we need to arrange them from lowest to highest and find the central value.

It would be like: 01, 05, 10, 15, 20.

The data is in odd number and central value is 10 which would be our Median.

It can be calculated as Median (M) $= \frac{n+1}{2}$.

But when the value is in order of even numbers, we need to find the two mid values and get the average.

Partial Exercise:

Find the median of this data:

10, 40, 20, 50

Put the data in order first:

There is an even number of data points, so the median is the average of the middle two data points.

Median (M) $= \frac{n1+n2}{2} =$

Exercise 1: The Patil family drove through 7 states on their summer vacation. Petrol prices varied from state to state. What is the median petrol price?

The petrol price for each state is given below:

$1.79, $1.61, $1.96, $2.09, $1.84, $1,75, $2.11

Exercise 2: The annual snowfall in inches for Manali for the last 9 years is listed below. Find the median snowfall.

20.5 in, 26.1 in, 32.5 in, 18.9 in, 33.4 in, 29.7 in, 19.8 in, 25.6 in, 34.3 in

---

5. MODE:

The most frequent number—that is, the number that occurs the highest number of times in a given set of data is known as mode of the data.

Suppose, for development of a new activity, a travel researcher collected some data regarding number of people looking forward to volunteering. He got the result from two data (Yes or No). The majority of either answer he will get determine his Mode.

For Example- Mr. Lokesh has asked his five co-workers to rate Rishikesh from 1-5 according to the adventurous activities.

C1: 2

C2: 5

C3: 1

C4: 5

C5: 4

So, the Mode will be 5.

Partial Exercise:

Ms. Norris asked students in her class how many siblings they each had.

Find the mode of the data:

0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 5

Let's look for the value which occur the most.

MODE =

Exercise 1: A marathon race was completed by 5 participants. What is the mode of these times given in hours?

2.7 hr, 8.3 hr, 3.5 hr, 5.1 hr, 4.9 hr

Exercise 2: In a crash test, 11 cars were tested to determine what impact speed was required to obtain minimal bumper damage. Find the mode of the speeds given in miles per hour below.

24, 15, 18, 20, 18, 22, 24, 26, 18, 26, 24

MEASURES OF DISPERSION

A measure of statistical dispersion is a non-negative real number that is zero if all the data is same and increases as the data become more diverse.

Suppose, there are two tourists visiting different destinations number of times:

	D1	D2	D3	D4	D5	D6
Tourist A:	4	4	5	6	4	7
Tourist B:	1	8	7	2	4	8

Here, we can see the two data variate to each other as first column gives more structured or less scattered than the column two data, even though they have approx. common mean i.e. 5. So, to measure this scatteredness, we use Measures of Deviation.

1. Range:

In the given above example, we can see

Tourist A has visited destinations with lowest 4 times and highest 6 times,

Tourist B has visited destinations with lowest 0 times and highest 8 times.

So, the range of Tourist A is 2 (6-4) and Range of Tourist B is 8 (8-0)

$$R = H - L$$

Partial Exercise:

Find the range of these distances run by 6 marathon runners:
10 km, 15 km, 12 km, 14 km, 8 km, 16 km

The range is calculating the difference between lowest and highest value set of data.

So, Range will be:

2. Quartile Deviation:

Suppose, there is a traveler experiencing many tourist activities and spending money on them unevenly, then to value its spending on the lowest and highest would do injustice to the rest of the experiences he had.

In such a situation, if the entire data is divided into four equal parts, each containing 25% of the values, we get the values of Quartiles.

Inter-Quartile Range is based upon middle 50% of the values in a distribution and not affected by extreme values. Half of the Inter-Quartile Range is called Quartile Deviation (Q.D.).

Thus, Q.D. = $(Q_3 - Q_1)/2$ where Q_3 is Upper Quartiles and Q_1 is Lower Quartiles.

Example:

Calculate the range and Quartile Deviation of the following observations. (Discreet series)

20, 25, 29, 30, 35, 39, 41, 48, 51, 60, 70.

So, Range = 70 – 20 = 50

Now, For Q.D., Let's first arrange the data in ascending order which it is already arranged.

$Q_1 = (n + 1)/4 = (11 + 1)/4 = 3$.

So, the 3rd value in the data will be $Q_1 = 29$.

For $Q_3 = 3(n+1)/4 = 3(11 + 1)/4 = 9^{th}$ value which will be 51.

So, $Q.D. = (51-29)/2 = 11$.

Partial Exercise:

Calculate the Range and Quartile Deviation for the given below continuous series.

A group of 40 travelers joined for hiking, there data is below:

Age level of Travelers (A1)	No. of Travelers (T1)
0-10	5
10-20	8
20-40	16
40-60	7
60-80	4
	40

So, to calculate we will simply do the formula, R = H – L

So, Range =

Now, to calculate Q.D in a continuous series, we need to calculate Q3 and Q1.

But first we need to calculate cumulative frequency.

Age level of Travelers (A1)	No. of Travelers (T1)	Cumulative Frequency
0-10	5	05
10-20	8	13
20-40	16	29
40-60	7	36
60-80	4	40
	N = 40	

For Continuous series, Q1 = (n)/4th value and Q3 = 3n/4th value.

Q1 = nth /4 = nth value. So, the class 10-20 contains the value.

To calculate exact value, the formula is:

$$Q1 = L + \frac{\left(\frac{n}{4}\right) - c.f}{f} * i$$

Where L is Lowest limit of the Quartile class

n is number of data.

c.f is the cumulative frequency of the Quartile class.

f is the frequency of the Quartile class.

i is the interval of the Quartile class.

SO, placing these values in the given formula, we get Quartile Deviation1.

$$Q1 = L + \frac{\left(\frac{n}{4}\right) - c.f}{f} * i$$

Similarly, calculating Q3 = 3(n)/4th value = = the value comes in class of 40-60.

Then, to get exact value, we follow the above formula.

$$Q3 = L + \frac{\left(\frac{3n}{4}\right) - c.f}{f} * i$$

$$Q3 = L + \frac{\left(\frac{3n}{4}\right) - c.f}{f} * i \quad \text{(Put the value)}$$

Exercise:

Calculate the range and Quartile deviation of number of tents available for women travelers group of 50. The data is given below:

Age level of Travelers (A1)	No. of Tents (T1)
20-30	7
30-40	18
40-50	16
50-60	7
60-80	2
	50

3. Mean Deviation from Mean:

The measures which are based upon deviation of the values from their average are Mean Deviation and Standard Deviation.

Here, the average of any data is its central value and rest of the data are deviations from its central/mean value which can either be positive or negative. Also, the summation of all these deviations result to 0. So, we try to take every deviation with positive value and calculate Mean Deviation from it.

Suppose a college is proposed for students of five towns A, B, C, D and E which lie in that order along a road. Distances of towns in kilometers from town A and number of students in these towns are given below:

Town	Distance from town A	No. of students
A	0	90
B	2	150
C	6	100
D	14	200
E	18	80
		620

So, we need to find a location so that the average distance travelled by students is minimum. In that case, we calculate Mean Deviation.

- **Calculation of Mean Deviation from Arithmetic Mean for ungrouped data.**

Example: **Calculate Mean Deviation of the given data:**

2, 4, 7, 8,9.

So, the Average Mean = (2+4+7+8+9)/5 = 6.

Now, we need to calculate deviations from the mean.

X	|d|
2	4
4	2

7	**1**
8	**2**
9	**3**
	12

So, the Mean Deviation will be = $(\sum |d|)\ /5$

$= 12/5 = 2.4$

Partial Exercise: Calculate Mean Deviation for ungrouped data of data given:

12. 50, 2, 6, 7

So, calculating Average Mean =

Now to calculate Mean Deviation first we calculate Deviations.

X	|d|
12	
50	
2	
6	
7	

So, the Mean Deviation will be = $(\sum |d|)\ /5$

$= \ / \ =$

- **Calculation of Mean Deviation from Median for ungrouped data**

Please find the data below of ungrouped data:

2, 4, 7, 8,9.

So, the Median = 7 (Odd number)

Now, we need to calculate deviations from the mean.

X	|d|
2	5
4	3
7	0
8	1
9	2
	11

So, the Mean Deviation will be = $(\sum |d|) /5$

$$= 11/5 = 2.2$$

- **Calculation of Mean Deviation from Mean for Continuous distribution**

Example:

Suppose we have a continuous series and to find Mean deviation in the data from average mean, we need to follow:

The profit of travel companies of a region is given:

Profit of companies (in lakhs)	No of companies
10-20	5
20-30	8
30-50	16
50-70	8
70-80	3
	40

So, to calculate Mean deviation, first we need to calculate:

1. Mean
2. the absolute deviations |d| of the class midpoints from the mean.

3. Multiply each |d| value with its corresponding frequency to get f|d| values. Sum them up to get Σ f|d|.
4. Apply the following formula,

   M.D = Σ f|d|/ Σ f

First, we will make another table,

Class Interval	f	m.p.	|d|	f|d|
10-20	5	15	25.5	127.5
20-30	8	25	15.5	124.0
30-50	16	40	0.5	8.0
50-70	8	60	19.5	156.0
70-80	3	75	34.5	103.5
	40			519.0

Now, to calculate Mean for a continuous distribution, we will follow the formula:

1. Mean = $\frac{\sum f*m.p}{\sum f} = \frac{\sum(5*15)+(8*25)+(16*40)+(8*60)+(3*75)}{40} = \frac{1620}{40}$

   μ =40.5

2. Now we need to calculate |d| of class midpoints from mean,

   |d|= m.p.- μ and put the value in the table.

3. Multiplying each |d| value with its frequency and fill the table.

4. **Applying the formula**

   M.D = Σ f|d|/ Σ f

   = 519/40 = 12.975

Partial Exercise:

The number of volunteers in a tribal location of a region is given:

Volunteers	No of countries
0-20	2
20-30	5
30-50	20
50-70	10
70-80	3
	40

Calculate Mean Deviation using Arithmetic Mean.

First let's draw a table and put the values on it by following the steps:

Class Interval	f	m.p.	|d|	f|d|
10-20				
20-30				
30-50				
50-70				
70-80				

Now, to calculate Mean for a continuous distribution, we will follow the formula:

1. Mean = $\frac{\sum f*m.p}{\sum f}$ = $\frac{\sum(*)+(*)+(*)+(*)+(*)}{40}$ = $\frac{}{40}$

   μ =

2. Now we need to calculate |d| of class midpoints from mean,

   |d|= m.p.- μ **and put the value in the table.**

3. Multiplying each |d| value with its frequency and fill the table.

4. **Applying the formula**

$M.D = \Sigma f|d| / \Sigma f$

$= \quad /40 \quad =$

- **Calculating Mean Deviation in a continuous series using Median.**

Using Median instead of Mean to calculate Mean Deviation is easy, all steps remain the same except instead of calculating mean in first step we calculate median of the mid points.

Example:

The profit of travel companies of a region is given:

Profit of companies (in lakhs)	No of companies
20-30	5
30-40	10
40-60	20
60-80	9
80-90	6
	40

First, we will draw a table:

Class Interval	f	m.p.	|d|	f|d|
20-30	5	25	25	125
30-40	10	35	15	150
40-60	20	50	0	0
60-80	9	70	20	180
80-90	6	85	35	210
	50			665

1. To get the Median of the mid points, Median will be 3rd value which is 50.
2. Calculate all the values, |d|, f|d| and put the value in the formula

$$M.D = \Sigma f|d|/\Sigma f$$

$$= 665/50 = 13.30$$

Exercise:

Calculate Mean deviation of the following ungrouped data:

3, 6, 6, 7, 8, 11, 15, 16

Exercise:

Calculate the mean deviation of number of travelers and activities they did:

No. of Travelers	0-5	5-10	10-20	20-30	30-40
Activities	2	5	1	3	12

Exercise:

Calculate Mean deviation through Median of following data:

Class	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	6	7	15	16	4	2

4. Standard Deviation

Suppose there are some data x1, x2, x3, x4, and x5. We first calculate their mean. Then, mean deviations and then, the deviations are squared. The Mean of these squared deviations is called Variance and positive square root of Variance is known as Standard Deviation.

Example:

Calculate Standard Deviation of following ungrouped data:

5, 10, 25, 30, 50

So, first we will make a table:

X	D(X-$\bar{X}$)	d^2
5	-19	361
10	-14	196
25	1	1
30	6	36
50	26	676
	0	1270

Let's first calculate Mean of the data:

Mean = 5+10+25+30+50/ 5 = 24

So, the Mean Deviation D(X-$\bar{X}$) will be put in the table.

Now the square of deviation is calculated which is the variance of each data.

So, The Variance is 254

For Standard Deviation,

$$\sigma = \frac{\sqrt{\sum d^2}}{\sqrt{n}}$$

$$= \sqrt[2]{\frac{1270}{5}} \quad = \sqrt[2]{254} \quad = 15.937$$

- **Calculation of Standard Deviation and variance of continuous data**

Exercise:

The profit of travel companies of a region is given:

Profit of companies (in lakhs)	No of companies
10-20	5
20-30	8
30-50	16
50-70	8
70-80	3
	40

Calculate Variance and Standard Deviation.

First, we will calculate Mean, then deviation and then Mean deviation.

Drawing the table,

Class Interval	f	m.p.	f.m	|d|	f|d|	$f\|d\|^2$
10-20	5	15	75	-25.5	-127.5	3251.5
20-30	8	25	200	-15.5	-124.0	1922
30-50	16	40	640	-0.5	-8.0	4
50-70	8	60	480	19.5	156.0	3042
70-80	3	75	225	34.5	103.5	3570.75
	40		1620		0	11790

Step 1: Mean = 1620/40 = 40.5

Step 2: Deviation for each class is written by the amount midpoint variate from its mean.

Step 3: Mean Deviation = $\sum f|d|/\sum f$

= 0/40 = 0

So, we will do square of |d| and then multiply with frequency (f). Then,

Applying the formula for standard deviation,

$$\sigma = \frac{\sqrt{\sum fd^2}}{\sqrt{n}}$$

$$= \sqrt[2]{\frac{11790}{40}} \quad = \sqrt[2]{294.75} \quad = \quad 17.168$$

And variance is 294.75.

Partial Exercise:

Calculate the Mean Deviation using mean and Standard Deviation for the following distribution.

Tourist age group	Frequencies
20-40	3
40-80	6
80-100	20
100-120	12
120-140	9
	50

Calculate first Mean, Deviation and mean deviation, then apply formula for Standard deviation.

Exercise 1:

To check the quality of two brands of lightbulbs, their life in burning hours was estimated as under for 100 bulbs of each brand.

Life (in hrs)	No. of Bulbs	
	Brand A	Brand B
0-50	15	2
50-100	20	8
100-150	18	60
150-200	25	25
200-250	22	5
	100	100

I. **Find which brand gives higher life?**
II. **Which brand is more dependable?**

Exercise 2:

The yield of wheat and rice per acre for 5 districts of a state is as under:

District	1	2	3	4	5
Wheat	12	10	15	19	21
Rice	22	29	12	23	18

Calculate for each crop:

i. **Range**
ii. **Quartile Deviation**
iii. **Mean deviation about Mean**
iv. **Mean Deviation about Median**
v. **Variance and Standard Deviation.**

Exercise 3:

You grow 20 crystals from a solution and measure the length of each crystal in millimeters. Here is your data:

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

Calculate the sample standard deviation of the length of the crystals.

Ans: 3.061

Exercise 4:

Consider the following three data sets A, B and C.
A = {9,10,11,7,13}
B = {10,10,10,10,10}
C = {1,1,10,19,19}
a) Calculate the mean of each data set.
b) Calculate the standard deviation of each data set.
c) Which set has the largest standard deviation?
d) Is it possible to answer question c) without calculations of the standard deviation?

Solution:

a) 10, 10, 10
b) 2, 0, 8.05
c) Set C
d) Yes, it has value farther aways than Set A and B.

Exercise 5:

The frequency table of the monthly salaries of 20 people is shown below.

salary(in $)	**frequency**
3500	5
4000	8
4200	5
4300	2

a) Calculate the mean of the salaries of the 20 people.
b) Calculate the standard deviation of the salaries of the 20 people.

Solution:

a) 3955
b) 282

Exercise 6:

The following table shows the grouped data, in classes, for the heights of 50 people.

height (in cm) - classes	**frequency**
120 <- 130	2
130 <- 140	5
140 <- 150	25
150 <- 160	10
160 <- 170	8

a) Calculate the mean of the salaries of the 20 people.
b) Calculate the standard deviation of the salaries of the 20 people.

Solution:

a) 148.40
b) 9.9

Exercise 7:

Calculate variance of the following data:

Class interval	Frequency
4 - 8	3
8 - 12	6
12 - 16	4

16 - 20	7

Solution:

Variance: 19

Exercise 8:

Life of bulbs produced by two factories A and B are given below:

Length of life	**Factory A**	**Factory B**
(in hours)	(Number of bulbs)	(Number of bulbs)
550 - 650	10	8
650 - 750	22	60
750 - 850	52	24
850 - 950	20	16
950 - 1050	16	12
	120	120

The bulbs of which factory are more consistent from the point of view of length of life?

Solution:

C.V. of factory B > C.V. of factory A ⇒ Factory B has more variability which means bulbs of factory A are more consistent.

Chapter 2: Introduction to Probability- Probability Distributions:

1. **Discreet: It consists of Binomial and Poisson's Distribution.**

A discreet random variable takes on discrete values that can be counted and assume values from a distinct predetermined set.

Mean of the probability distribution is also known as expected value and Variance is sum of the products of the square deviations between the mean and all individual values of the random variable, one at a time.

- ## Binomial Distribution:

Where there is only two mutually outcomes of each trial of an event. For example. In tossing a coin, the resulting outcome will be either Head or Tail.

Also, on Checking quality of a product, the production either be good or defective so the outcomes have equal possibilities and occurs in with two possible outcomes.

For Example:

If we toss a coin 5 times, the total possibilities of getting two heads out of five outcomes will be 10. So, the probability of the sequence of 2 heads and three tails will be written as:

$10 * p^2 * q^3$

Now, we have p is 0.5

q is 0.5

So, the probability of two heads out of five tosses is $10*(0.5)^2(0.5)^3$

Is 0.3125

It can also be represented by $P(x) = {}^{n}_{x}C\ p^x\ q^{(n-x)}$

Partial Exercise:

If a drug is found to be effective 40% of the time, then what is the probability that in a random sample of 4 patients, it will be effective on 2 of them?

Solution: Let assume effective as success and non-effective as failure.

Then, p = 0.4, q = 0.6

x = 2, n = 4

So, putting the values on said formula,

$$\mathbf{P(x) = {}_{x}^{n}C\ p^{x}\ q^{(n-x)}}$$

$$P(x) = {}_{2}^{4}C\ (0.4)^{2}\ (0.6)^{(2)}$$

=

The mean and Standard deviation of Binomial Distribution will written as

Mean = np and Standard Deviation = $\sqrt{npq}$

For example:

In a manufacturing process, a packaging machine 5% defective packages. Find the mean and standard deviation of the number of defective packages in a random sample of 60 packages.

Solution:

Mean (μ) = 60*0.05 = 3

Standard Deviation (σ) = $\sqrt{60*0.05*0.95}$ = 1.69

POISSON DISTRIBUTION:

It is another discreet distribution. In this, we can find out the average number of successes in a given unit of time and space.

For Example, Incase of number of patients coming to the hospital for emergency treatment, we can calculate the number of patients arriving in any given hour and it can be called as successes though we cannot calculate number of patients not coming in any given hour so failures cannot be counted.

Continuous Distribution: NORMAL DISTRIBUTION

The Normal Distribution has many uses in practical world as many experimental results often follow normal distribution. It can be represented as Bell-Shaped curve.

The normal curve is symmetrical and defined by its mean and standard deviation.

The number of standard deviations Z for an observation, which is distance between value x and mean is defined by:

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

Where x = value of the observation

μ = the mean of the distribution

σ = standard deviation of the distribution

For Example:

The IQ of students is normally distributed with mean of 120 and standard deviation of 20. What proportion of students have:

a. An IQ between 100 and 130
b. An IQ above 140
c. An IQ below 150
d. An IQ between 140 and 150

Solution:

We have Mean(μ) = 120, σ = 20.

a. An IQ between 100 and 130 will be calculated by finding the area between 100 and 130.

For this we will add 100 to 120 and 120 to 130.

Then, Z1 = $\frac{100-120}{20}$ = -1. The area from table is 0.3413

Z2 = $\frac{130-120}{20}$ = 0.5. The area from table is 0.1915

So, the total area will be 0.3413 + 0.1915 = 0.5328

This means 53,28% of students have IQ between 100 and 130.

b. IQ above 140

Z = $\frac{140-120}{20}$ = 1. The area from table is 0.3413. So for above of 140 we will,

Substract 0.5- 0.3413 = 0.1587

Or 15.87% students have IQ above 140.

c. IQ below 150.

Z = $\frac{150-120}{20}$ = 1.5. The area from table is 0.4332. So, for below of 150 we will,

Add 0.5 + 0.4332 = 0.9332

Or 93.32% students have IQ below 150.

d. IQ between 140 and 150.

Z1 = $\frac{140-120}{20}$ = 1. The area from table is 0.3413.

Z2 = $\frac{150-120}{20}$ = 1.5. The area from table is 0.4332.

Difference = 0.0919

Or 9.19% of students have an IQ between 140 and 150.

Exercise 1:

The IQ of students is normally distributed with mean of 120 and standard deviation of 20. What proportion of students have:

a. Find the score so that 20% of the students have an IQ above this score.
b. What are the limits within which the central 50% of the scores lies.

Soln: a. 136.8

b. Limits are 106.6 and 133.40.

Exercise 2:

The heights of soldiers are normally distributed. If 11.51% of the soldiers are taller than 70.4 inches and 9.68% are shorter than 65.4 inches, find the mean and standard deviation for the data.

Solution: $\sigma = 2$ and $\mu = 68$ inches.

Exercise 3:

A fair coin is tossed 16 times. What is the probability of getting no more than 2 heads?

Exercise 4:

The GPA of Tourism students is 3 out of a total of 4 points, with a standard deviation of 0.5. If a policy has been adopted that all students with GPA of 2 or less during any semester will be put on probation, what percentage of students are expected to put on probation.

Exercise 5:

The heights of soldiers is normally distributed with a mean o 68 inches and a variance of 9 sq. inches. What is the probability that a soldier picked up at random is:

a. Less than 5 feet 1 inch tall

b. Between 63 inches and 66 inches

c. Taller than 6 feet

d. What must the height of a soldier be so that only 30% of the soldiers are taller than him?

Exercise 6:

The small business center of a city reports that 30% of all small businesses (20 or fewer employees) are owned by women. What is the probability that a random sample of 50 small businesses will show:

a. More than 42 owned by men

b. Less than 30 owned by women

c. Between 30 and 42 owned by men.

Exercise 7:

The life of Sears Die Hard battery is normally distributed with an average life of 5 years and a standard deviation of 60 days. How long should the guarantee period be if the company wishes to replace on more than 15% of the batteries.

Chapter 3: Sampling Distribution:

The concept of sampling distribution differs from sample distribution or population distribution. It is distribution of a specific descriptive measure, such as mean.

We can see the relationship between sample mean and population mean as follows:

For Example:

An accounting firm has 5 branch offices in the 5 areas of the city. These offices employ 2,3,4,5,6, employees respectively.

a) Find the mean and standard deviation of this finite population
b) List the 20 possible samples of size n equals 3 that can be drawn from this population without replacement.

Solution:

a. To calculate mean and standard deviation:

Branch	Employees	$(X-\mu)^2$
1	2	4
2	3	1
3	4	0
4	5	1
5	6	4
Total	20	10

So, Mean = $\frac{2+3+4+5+6}{5}$ = 4

And Standard Deviation = $\sqrt{\frac{10}{5}}$ = $\sqrt{2}$ = 1.414

b. Sample size of 3 without replacement will be

X1,X2,X3 – (2,3,4) μ1= 3

(2,4,5) μ2= 3.66

(2,5,6) μ3= 4.33

(2,3,5) μ4=3.33

(2,3,6) μ5= 3.66

(2,4,6) μ6= 4

(3,4,5) μ7= 4

(3,4,6) μ8= 4.33

(4,5,6) μ9 =5

So, taking any random sample may or may not be close to the total mean we got earlier which was 4.

Therefore, taking Grand Mean = $\frac{3+3.66+4.33+3.33+3.66+4+4+4.33+5}{9}$ = 3.92 approximate of mean 4.

So, the probability distribution of sample means also known as Sampling distribution of the mean is:

Sample Mean	Frequency	Re. Frequency	Probability
3	1	1/10	0.1
3.66	2	2/10	0.2
4.33	2	2/10	0.2
3.33	1	1/10	0.1
4	2	2/10	0.2
5	1	1/10	0.1

We have seen that the grand mean of sample means equals to the total mean of the given population. But it is not possible to take all the samples and we take few and watch how close the sample mean is to the population mean.

CENTRAL LIMIT THEOREM states that the sample size as increased will make it closer to the actual population mean.

STANDARD ERROR OF MEAN:

Similarly, we will calculate Standard deviation of Sample means and compare it to actual population standard deviation.

Actual standard deviation was 1.414

Now,

Sample Mean	Mean	(Sample mean-mean)^2
3	4	1
3.66	4	0.1156
4.33	4	0.1089
3.33	4	0.4489
3.66	4	0.1156
4	4	0
4	4	0
4.33	4	0.1089
5	4	1
		2.8979

Standard Sample Deviation = $\sqrt{\frac{2.8979}{9}}$ = 0.567

Hence, the sample deviation error will be always lesser than actual standard deviation.

The relationship between both deviations will be:

$$\sigma_x = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

For Example,

The IQ scores of students are normally distributed with mean 120 and standard deviation 10.

If a random sample of 25 students taken, what is the probability that the mean of the sample will be between 120 and 125.

Solution:

We have mean= 120

σ = 10

So, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{5} = 2$

Calculating Z = $\frac{(X-\mu)}{\sigma x} = \frac{125-120}{2} = 2.5$

The area of Z = 2.5 is 0.4938, or 49.38%. This shows that there is 49.38% chance that the sample mean will be in between 120 and 125.

Exercise 1:

The statistics class has a total of 60 students. Their average score in their mid-term exams was 70 with standard deviation of 8. A sample of 36 of these students taken at random. Calculate standard error of the mean for this sample.

Ans: 0.268

Exercise 2:

The average account receivable in a ledger is INR 125 with a standard deviation of INR 20. A random sample of 25 receivable accounts is selected from this ledger. What is the probability that the average of this sample will be less than INR 115?

Ans: 0.0062

Exercise 3:

The average IQ score of students in a school for gifted children is 165 with a standard deviation of 27. A random sample of 36 students taken, what is the probability that

a. The sample mean is greater than 170 (Ans: 0.1335)

b. The sample mean is less than 158 (Ans: 0.0594)

c. The sample mean is in between 155 and 160. (Ans: 0.1203)

SAMPLING DISTRIBUTION OF PROPORTIONS:

If a sample of students taking a statistics course suggests that 30% of these students want to become statisticians from all students taking the elementary course.

It can be defined as a distribution of proportions of all possible random samples of a fixed size n.

Mean of the Sampling distribution of proportions

p is sample proportion

π is population proportion

The sample proportion is defined as x/n where x is no. of successes and n is sample size.

Similarly, population proportion is defined as x/N where N is total population size.

For Example:

There is total 5 students who are asked if they want to be statisticians. Their answers are below:

Students	Answers
1	Y
2	N
3	N
4	Y
5	N

The number of students who want to become is 2 so N = 5 and x =2

So, the population proportion will be: **π= 2/5 =** 0.4 or 40%.

Now, let us take all possible samples of 4 from this population of size 5 so sample proportion who wants to be statisticianswill be:

1. 1,2,3,4 - 2 out of 4 will be 0.50

2. 1,2,3,5 - 1 out of 4 will be 0.25

3. 1,2,34,5 - 2 out of 4 will be 0.50

4. 1,3,4,5 - 2 out of 4 will be 0.50

5. 2,3,4,5 - 1 out of 4 will be 0.25

So, the Summation of the proportions will be 2.0.

Taking average, $\mu_p = \sum p$ / N = 2/5 = 0.4

So, the sample mean proportion is same as population mean proportion.

To Calculate Standard Deviation of Sample distribution of Proportions

The formula for calculating standard deviation will be:

$$\sigma_p = \sqrt{\frac{\pi\,(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}}$$

$\sigma_{p=}$ standard error of the proportion

π = proportion of successes

N = proportion size

n = sample size

So, **$Z = (p- \pi)/ \sigma_p$**

For Example:

It is known that 65% of all Indian voters favor the Congress government.

a. What is the probability that a simple random sample of 100 Indian voters will reveal the proportion of voters favoring the Congress government to be 60% or less.

b. How would this probability change if the sample size is increased to 500.

Solution:

We have sample size n is 100.

E (p)= π = 0.65

To calculate standard error we have, $\sigma_p = \sqrt{\frac{\pi\,(1-\pi)}{n}} = \sqrt{\frac{0.65*0.35}{100}} = 0.0477$

Now, $\mathbf{Z = (p- \pi)/ \sigma_p}$ = (0.6-0.65)/ 0.0477 = -1.05

The area under the curve of the value for Z is 0.3531. So, less than 60% will be calculated as

0.5 – 0.3531 = 0.1469

b. When the sample size is increased to 500, we will follow the same process.

To calculate standard error we have, $\mathbf{\sigma_p = \sqrt{\frac{\pi\,(1-\pi)}{n}} = \sqrt{\frac{0.65*0.35}{500}} = 0.0213}$

Now, $\mathbf{Z = (p- \pi)/ \sigma_p}$ = (0.6-0.65)/ 0.0213 = -2.35

The area under the curve of the value for Z is 0.4906. So, less than 60% will be calculated as

0.5 – 0.4906 = 0.0094.

Exercise 1:

45% of all graduate students pursuing their doctorate degree at DU are married. If a sample of 200 graduate students is selected at random, what is the probability that the proportion of married students in this sample would be between 40% and 48%.

Solution: P = 0.7287

Exercise 2:

15% of the people in the small community of Sands Point have type B blood. A random sample of 500 persons is selected. What is the probability that the sample proportion of people with blood type B is?

a. More than 17.5%

b. Less than 14%

c. Between 16% and 18%.

Solution: a. 0.594, b. 0.2676, c. 0.2369

Statistical Inference: Estimation

- **Point Estimation:**

A point estimate uses a single sample value to estimate the desired population parameter. So, sample mean would be considered much closer to actual population mean. Similar, the case of variance and standard deviation.

The parameters of point estimate are:

$$\mu = \frac{\sum x}{n}$$

$$\sigma = \sqrt{\frac{(X-\mu)2}{n-1}}$$

$p = p_s = \left(\frac{x}{n}\right)$, where p_s is sample proportion.

- **Interval Estimation:**

The point estimator has major drawbacks regarding its accuracy to the total population and is less reliable. So, Interval estimate is used where.

First a point estimate is found, then, using this estimate to construct an interval on both side of the point estimate.

For Example:

We need to find the average salary of professors at the university who had served for 5 years. Let the average we got from a sample is INR 90,000, though it could be accurate or inaccurate though by using interval of between INR 80,000 to INR 1,00,000. Our assumption will be more accurate.

SO, we can write is as,

$$\mu = x \pm Z\sigma_x$$

Now, we need to find confidence interval around the sample mean which is expected to be 95% of the time.

Example:

The sponsor of a TV programmer wants to find out average amount of time children (4-10 yrs.) spend. A random sample of 100 indicated average time per week be 27.2 hrs. From previous experience, the population standard deviation is 8 hrs. A confidence level of 95% is considered accurate.

Solution:

We have,

(sample mean)x = 27.2

Z = 1.96 (47.5%)

σ = 8

n = 100

So, $\sigma_x = \frac{8}{\sqrt{100}} = 0.8$

Then,

$X1 = x - Z\,\sigma_x$

$= 27.2 - (1.96*0.8) = 25.632$

$X2 = x + Z\,\sigma_x$

$= 27.2 + (1.96*0.8) = 28.768$

So, the range for a child to be watching TV will be between 25.632 – 28.768.

Exercise 1:

It is desired to estimate the average age of students who graduate with MBA. A random sample of 64 students taken with average 27 years and standard deviation of 4 years.

a. Estimate a 95% confidence interval estimate of the true average.

b. How would the confidence interval limits change if the level was increased from 95% to 99%.

Ans: a. 26.02 – 27.98

b. 25.71 – 28.29

- **Sample Size Distribution for Estimating Population Mean**

The degree of accuracy and degree of confidence are two factors we need to determine sample size.

We know that,

$$\mathbf{Z = \frac{(X-\mu)}{(\sigma/\sqrt{n})}}$$ **where (X-μ) can be supposed as E.**

So, we will get,

$\sqrt{n}$ = Zσ/E or n= $(Z\sigma/E)^2$.

Exercise 2:

Mike wants to buy a second hand car. He decides to buy a 1989 model of Buick and selected 100 sale advertisements from local newspaper over 1 month and found average price will be USD 4500. He knows the standard deviation would be USD 520. Establish a 95% CONFIDENCE INTERVAL ESTIMATE OF THE TRUE AVERAGE PRICE FOR ALL USED CARS.

Ans: 4398.08-4601.92

Exercise 3:

An efficiency expert is interested to determine average time it takes a worker to assemble a laptop computer with available parts. How large a sample will he need to be 98% certain that

his sample mean will not differ from true mean by more than 10 min. The standard deviation is 40 min.

Ans: 86.88

Exercise 4:

The Plaza Pizza would like to determine the average delivery time. How large would the sample size be if it wants to be 95% confident that the sample would not differ than actual average delivery time by more than 1.5 min. the previous studies shows standard deviation of 7 min.

Exercise 5:

A sample of single mother students with more than two children was taken to find the age of youngest child. The average age of these 160 children was 6.7yrs with sample standard deviation is 2.3 yrs. Calculate standard error of the mean.

Interval Estimation of Population Proportion

We have, $\boldsymbol{\sigma_p} = \sqrt{\frac{\boldsymbol{\pi}\,(\mathbf{1}-\boldsymbol{\pi})}{\boldsymbol{n}}}$

In the cases, we expect 95% of all sample proportions to fall within the following range:

$\pi \pm 1.96\ \sigma_p$

For 99% of all sample proportion it will be: $\pi \pm 2.58\ \sigma_p$

For Example:

A survey of 500 people shopping at a mall, showed that 350 of them used credit cards for their purchases and 150 of them used cash.

a. Construct a 95% confidence interval estimate of the proportion of all persons at the mall who used credit card.

b. Determine confidence interval if shoppers using credit card is between 67% to 73%.

Solution:

Sample size n is 350 and N is 500. So, p = 350/500 =0.7

For 95% confidence interval, we have range $\pi \pm 1.96\ \sigma_p$

So, putting the value

Range is $0.7 \pm 1.96 \sqrt{\frac{\pi\,(1-\pi)}{n}} = 0.7 \pm 1.96 \sqrt{\frac{0.7*0.3}{500}} = 0.7 \pm 1.96*0.02$

The confidence limits are:

P_1= 0.7- 1.96*0.02 = 0.6608 = 66.08%

P_2= 0.7 + 1.96*0.02 = 0.7392 = 73.92%

This means the population using credit card is between 66.08% to 73.92%.

b. If the population is between 0.67 and 0.73 where sample proportion p is 0.7 given.

$P_1 = p - Z*\sigma_p$

0.67= 0.7- Z(0.02)

Z= 0.03/0.02 = 1.5

Similarly, $P_1 = p + Z*\sigma_p$

0.67= 0.7+ Z(0.02)

Z= 0.3/0.02 = 1.5

So, the area under the curve Z is 0.4332. So, the total area will be 0.4332 + 0.4332 which will be 0.8664 or 86.64%.

Sample Size determination for Estimating the population proportion

Here, we have $Z = (p-\pi)/\sigma p$ where $\sigma_p = \sqrt{\frac{\pi\,(1-\pi)}{n}}$

So, (p-π) is taken as E and we can write,

$$Z = E/\sqrt{\frac{\pi\,(1-\pi)}{n}}$$

For Example:

It is desired to estimate the proportion of children watching TV on Saturdays. We want confidence interval 95% confident that our estimate value within ± 2% of the true population proportion.

What sample size should we take if a previous survey showed 40% of children watching TV on Saturdays.

Solution:

In this case:

Z= 1.96 (95% confidence)

P= 0.4

E= 0.02

Substituting the values, $Z = E/\sqrt{\frac{\pi\,(1-\pi)}{n}}$

$$n = Z^2p(1\text{-}p)/E^2 = ((1.96)^2 * 0.4 * 0.6)/(0.02)^2 = 2304.96$$

So the sample size will be 2305.

Exercise 1:

In a survey, 1200 persons selected and asked their opinions whether a Congressman's term should be limited to 12 years in the Congress. Out of this sample, 780 persons answered affirmative. Construct a 95% confidence interval of the true proportion regarding such opinions.

Solution: Range: 0.614 to 0.686

Exercise 2:

The police department is concerned about drivers and want to estimate who exceed the speed limit by more than 5miles per hour. How large a sample will be needed so that the police can be at least 99% confident that the error in their estimate not exceed 0.04.

Solution: Sample size (n)= 1040.6

Chapter 4: Hypothesis Testing

In here, we will be testing our assumptions or hypothesis being correct or not. A claim or hypothesis about the values or population parameters is written as H_o (Null Hypothesis).

- The Null Hypothesis is then tested with available evidence to either accept it or reject it.
- H_1 is the Alternate hypothesis which become trues once Null hypothesis is rejected.
- Type 1 error(α): An error made in rejecting the null hypothesis, when in fact it is true.
- Type 2 error(β): An error made in accepting the null hypothesis, when in fact it is false.

One Tailed test: When area of rejection is on one extreme of the curve.

For example: A low calorie ice-cream is sampled and if mean is found higher than average then it's rejected though if found less then it will be accepted.

Two-Tailed Test: When the area of rejection is at both ends.

For example: If the mean of sample student's height is greater or less than average mean then the null hypothesis will be rejected.

I. Test Involving A Population Mean

Example 1:

Assume that the average annual income for government employees in the nation is reported by the Census Bureau to be 18750. A random sample of 100 employees taken and it was found that their average salary is 19240 with a standard deviation of 2610. At a level of significance α = 0.05, can we conclude that the average salary of the sample size is representative of national average?

Solution:

First, it is a two-tailed test as greater or lower than the given average will reject our hypothesis.

H_0: The average represents the national average.

H_1: The average does not represent national average.

According to Central Limit Theorem, the sample distribution od sample mean will be:

Calculating Z = $\frac{(X-\mu)}{\sigma x}$

$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{2610}{10} = 261$

So, $Z = \frac{(X-\mu)}{\sigma x} = \frac{(19240-18750)}{261} = \frac{490}{261} = 1.877$

Now, the level of significance is 0.05 so Z should be in within $\pm$1.96. Here, 1.877 clearly lies in the given range so, Null Hypothesis will be accepted.

Example 2:

The light bulbs of company X lasts on an average of 1600 hrs. Hypothesis is accepted if the average of the sample would be equal or more than 1600 hrs. but gets rejected if it is less than 1600hrs.

A sample of 100 bulbs taken with average 1570 hrs and standard deviation of 120 hrs. At α = 0.01, let us test the validity of the claim of this company.

Solution:

H_o: $\mu = 1600$

H_1: $\mu < 1600$

For 99% confidence level, Z= -2.33, on left tail.

Now,

$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{120}{10} = 12$

So, $Z = \frac{(X-\mu)}{\sigma x} = \frac{(1570-1600)}{12} = -\frac{30}{12} = -2.5$

Since our computed value Z is larger than required Z= 2.33. The null hypothesis is rejected.

Exercise 1:

An insurance company claims that it takes 14 days on average to process an auto accident claim. The standard deviation is 6 days. An investigator selected 36 people who have recently filed claims. The sample average resulted in 16 days. At 99% level of confidence, check if it takes the company more than 14 days on an average.

Solution: Null Hypothesis is accepted.

Exercise 2:

A law professor has told his students that in New York city, convicted embezzlers spend on an average 15 months in jail. A student takes a random sample of 35 such cases and got average sample was 13.8 months with standard deviation of 4.2 months. At 95% level of significance, check whether professor's information is correct or not.

Solution: Null Hypothesis is accepted.

Exercise 3:

An educator claims that the average IQ of City College students is no more than 110. A sample of 150 students was taken with average sample 111.2 IQ ad standard deviation 7.2. At level of significance 0.01, test if the claim of the educator is justified?

Solution: Null Hypothesis is accepted

b. Test involving a Single Proportion

For qualitative data, the parameter of interest will be population proportions or percentages.

For Example:

The sponsor of a television show believes that his studio audience is divided equally btw women and men. Out of 400 persons attending the show one day there are 230 men. At level of significance 0.05, test if the claim of the sponsor is justified?

Solution:

Now, it is two tailed test with,

H_o: $\pi = 0.5$

H_1: $\pi \neq 0.5$

For 95% confidence level, Z= 1.96,

Now,

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.5*0.5}{400}} = 0.025$$

Now, p is (230/400) which will be 0.575

$$\text{So, } Z = \frac{(p-\pi)}{\sigma p} = \frac{(0.575-0.5)}{0.025} = \frac{0.075}{0.025} = 3$$

Since our computed value Z is larger than required Z= 1.96The null hypothesis is rejected.

Exercise 1:

The mayor of the city claims that 60% of the people of the city follow him and support his policies. We want to test whether his claim is valid or not. A random sample of 400 persons taken and found 220 of these supports him. At level of significance 0.01, test if the claim of the mayor is justified?

Solution: The answer is -2.04 which is less than requited -2.33 (99%). So, The claim of mayor is accepted.

Exercise 2:

In a psychology class, a professor read a report which noted that 30% of all women are afraid of flying. A student took a random sample of 150 women and found 50 of them afraid. At level of significance 0.05, test if the claim of the professor is justified?

Solution: The value of Z is calculated is 0.81 which is less than required range of 1.96 (95%). So the Null hypothesis is accepted.

Exercise 3:

An airline claims that at most 8% of its lost luggage is never found. A consumer advocacy agency took 200 cases and found 22 cases of lost baggage never found. At level of significance 0.01, test if the claim of the airline is justified?

Solution: The value of Z calculated is 1.58 which is less than required 2.33(99%). So the claim of airline is accepted.

II. Test Involving Difference Between Two Sample Means:

When there are two samples taken then a hypothesis is tested for any significant difference between the sample mean and the population mean.

Now,

Distribution of the differences in sample mean: (X1-X2)

a. The mean of sampling distribution is: $\mu_1 - \mu_2$

b. The standard error of differences is given by:

$$\sqrt{\frac{(\sigma 1)^2}{n1} + \frac{(\sigma 2)^2}{n2}}$$

c. $$Z = \frac{x1 - X2}{\sqrt{\frac{(\sigma 1)^2}{n1} + \frac{(\sigma 2)^2}{n2}}}$$

For Example:

A potential buyer of electric bulbs bought 100 bulbs each of two famous brands A and B. Upon testing, he found brand A had a mean life of 1500 hrs with a standard deviation of 50 hrs and Brand B had an average life of 1530 hrs with standard deviation of 60 hrs. Can it be concluded at 95% level of significance that the two brands differ in quality?

Solution:

Null Hypothesis: $H_0: \mu_1 = \mu_2$

Alternate Hypothesis: $H_1: \mu_1 \neq \mu_2$

So, $$Z = \frac{x1 - X2}{\sqrt{\frac{(\sigma 1)^2}{n1} + \frac{(\sigma 2)^2}{n2}}}$$

Where, $$\sqrt{\frac{(\sigma 1)^2}{n1} + \frac{(\sigma 2)^2}{n2}} = \sqrt{\frac{(50)^2}{100} + \frac{(60)^2}{100}} = \sqrt{61} = 7.81$$

Now, $$Z = \frac{x1 - X2}{\sqrt{\frac{(\sigma 1)^2}{n1} + \frac{(\sigma 2)^2}{n2}}} = \frac{1500 - 1530}{7.81} = -3.841$$

Since, the computed value of Z is larger than α which is 1.96. So Null hypothesis is rejected.

Exercise 1:

A civil group in the city claims that a female college graduate earns less than a male college graduate. Sample of salary of 60 male and 50 female graduates was taken and it was found that the average starting salary of female was 25000 with standard deviation 500 and average salary of men is 27000 with standard deviation 600. At 1% level of significance test this claim.

Solution: Left one tailed test and Null hypothesis is rejected.

Exercise 2:

The Dean of students wants to find out if there is any difference in mathematical ability of boys and girls. A random sample of 50 female and 100 and is taken. Average sample of Female is 70 with standard deviation 12 and average sample of male is 75 with standard deviation is 10. At 1% level of significance testify this.

Solution: Two tailed test, the null hypothesis is accepted.

Exercise 3:

A researcher claims that American 18 years old females are on average taller than 18 years old British females. A random sample of 50 American and 50 British females was taken with average sample American is 65.2 with standard deviation 2.5 and sample average of British is 64.5 with standard deviation is 2.8. Test this with $\alpha = 0,05$.

Solution: One tailed test, Null hypothesis is accepted.

b. Test Involving Difference Between Two Population Proportions:

In here we check whether the two population proportions are equal or not, so the difference in sample proportions is measured for their significance.

Now,

Distribution of the differences in sample proportions will be: $(p_1 - p_2)$

a. The mean of distribution of differences of proportions is given by: $\pi_1 - \pi_2$

b. The standard deviation of differences in proportions is given by:

$$\sigma_p = \sqrt{\pi 1(1-\pi 1)\left(\frac{1}{n1}+\frac{1}{n2}\right)}$$ where π1(pi hat) $= \frac{n1p1+n2p2}{n1+n2}$

c. $$Z = \frac{p1-p2}{\sqrt{\pi 1(1-\pi 1)\left(\frac{1}{n1}+\frac{1}{n2}\right)}}$$

For Example:

A sample of 200 students revealed that 18% of them were seniors. A sample of 400 students of other college revealed 15% seniors. We want to test the difference of these two sample proportions to conclude that these populations are indeed different at 5% significance level.

Solution:

We have, n1= 200, n2= 400, p1= 0.18 and p2= 0.15.

So, first we will calculate,

π1(pi hat) $= \frac{n1p1+n2p2}{n1+n2} = \frac{200*0.18+400*0.15}{200+400} = \frac{36+60}{600} = 0.16$

To put the value in calculating standard deviation,

$$\sigma_p = \sqrt{\pi 1(1-\pi 1)\left(\frac{1}{n1}+\frac{1}{n2}\right)} = \sqrt{0.16 * 0.84\left(\frac{1}{200}+\frac{1}{400}\right)}$$

$$= \sqrt{0.1344 * 0.0075} = 0.0317$$

So, calculating Z,

$$Z = \frac{p1-p2}{\sqrt{\pi 1(1-\pi 1)\left(\frac{1}{n1}+\frac{1}{n2}\right)}} = \frac{0.18-0.15}{0.0317} = 0.95$$

The value of Z is less than required value of 1.96 (95%). So, there is no significant difference between two college samples.

Exercise 1:

An insurance company believes there is higher risk of heart attack in men smokers over 50 yrs of age than non-smokers. For this, the company took sample of 200 men of which 80 men are smokers. The survey indicated that 18 smokers suffered from heart disease and 15 non-smokers suffered from heart disease. At 5% level of significance, justify the claim of the company.

Solution: The value of Z is 1.86 which is lesser than required 1.96 so, the claim is accepted.

Exercise 2:

An advertising agency wants to find out if there is any difference in the degree of loyalty between men and women for a cereal. A sample of 200 men and 200 men was taken and determined 58% of women and 65% of men showed brand loyalty. At 95% significance level, test the null hypothesis that there is no significant difference.

Solution: Z is 1.47 which is less than required 1.96 so, there is no significant difference.

END