statistics

**20. What do you mean by Measure of Central Tendency and Measures of Dispersion .How it can be calculated.**

**Measures of Central Tendency** are values that represent the center point or typical value of a dataset. They help to summarize a large set of data by identifying the central position within that dataset. The most common measures of central tendency are:

**Mean (Average):**

The mean is the sum of all values divided by the number of values. It gives an overall idea of the average value in a dataset.

Mean is calculated by the formulae

Mean= sum of all observations

No of all observations

Mean is represented by symbol μ(nu)

**Median:**

The median is the middle value in a list of numbers arranged in order. If there is an even number of observations, the median is the average of the two middle numbers.

Formulae if no of observations are odd = (n+1)/2 item will be the middle value

Formulae if no observations are even =( (n/2) + (n+1)/2)/2

Where n is the no of observations

Ex: 1,3,4,6,7

N is odd

So median =n+1/2 item that is 3rd item -🡪4

**Mode:**

The mode is the number that occurs most frequently in a dataset. A dataset can have more than one mode or no mode at all if all numbers are unique.

The most repeated item in the given data set

Ex:

Given dataset : 3,3,3,4,5,5,6,7,8,9

The mode will be 3 because it has repeated more times than the other items in given dataset

**Measures of Dispersion:**

Measures of Dispersion describe how the data is spread or variability of data . They indicate how much the data points differ from each other or from the central tendency.

There are different types:

**Variance:**

Variance describes how much the data points deviated from the mean. it gives an idea that how much the data points get dispersed from the mean . it doesn’t tells that how much a single datapoint on average dispersed from the mean . it will deals with whole data . Gives information of deviation of whole data from mean. it is one of the drawback.

Formulae:

Variance(σ2)= ∑(xi-μ)^2

N

->The average of the squared differences from the mean.

So it gives an idea of how much the data points deviate from the mean

**Standard Deviation:**

The square root of the variance. It provides a measure of the average distance of each data point from the mean.

Standard Deviation(σ) = sqrt(​)

This deals how a single datapoint on an average dispersed from the mean.

**Range:**

The difference between the highest and lowest values in the dataset

Formula = Range=Maximum Value−Minimum Value

It also indirectly gives dispersion or spread of data.

The range gives you a quick snapshot of how spread out the data is. If the range is large, it means there is a significant difference between the smallest and largest values. If the range is small, the values are clustered close together.

Another important one that tells the spread of data is iqr:

**Interquartile Range(IQR):**

The IQR tells you how spread out the middle half of your data is. By focusing on the central 50% of the data, the IQR gives a clearer picture of where most of the data points are, without being skewed by outliers. The IQR is commonly used to identify outliers. Values that fall below Q1−1.5×IQR or above Q3+1.5×IQR are often considered outliers. This helps in cleaning data and understanding unusual observations. Unlike the range, which is sensitive to extreme values, the IQR focuses on the middle of the data and is not influenced by outliers. This makes it a more reliable measure of spread when dealing with skewed data or datasets with outliers

**21.What do you mean by skewness.Explain its types.Use graph to show**

It is used to describe the direction and degree of asymmetry in a data distribution. It shows whether the data is evenly distributed around the mean or if it is uneven on both sides of mean , with more data points piling up on one side of the mean. it gives the shape of data based on that we can easily identify whether the given data is skewed or not.

Skewed types:

Positive skewness(right-skewed)

Negative skewness(left-skewness)

No-skewness(symmetric)

**Positive skewed:**

In a positively skewed distribution, the tail on the right side of the distribution is longer or fatter than the left side. The mean is typically greater than the median.

Reason behind is Most of the data values are concentrated on the left with a few larger values stretching out to the right.

Ex:

In india we have so called rich peoples and we have very poor peoples. If we drew a income distribution most of the people will lies in lower income , but a few people earn very hiogh incomes.

**Negative skewed:**

In a negatively skewed distribution, the tail on the left side of the distribution is longer or fatter than the right side. The mean is typically less than the median.

Reason behind is Most of the data values are concentrated on the right with a few smaller values stretching out to the left.

Ex:

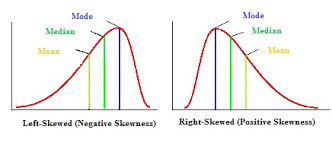
Age at retirement can be left-skewed, where most people retire around the same age, but a few retire much earlier

**Symmetric distribution:**

**We can also call this distribution as Zero Skewness.**

A distribution with zero skewness is perfectly symmetrical. The left and right sides of the distribution are mirror images of each other, and the mean, median, and mode are all equal.

Here in this distribution the data is evenly distributed around the mean, with no tail longer than the other.

****

**22. Explain PROBABILITY MASS FUNCTION (PMF) and PROBABILITY DENSITY FUNCTION (PDF). and what is the difference between them?**

**23. What is correlation. Explain its type in details.what are the  methods of determining correlation.**

It is a statistical measure that describes the strength and direction of the relationship between two variables. It tells us how changes in one variable are associated with changes in another variable.

Types of correlation:

Positive correlation

Negative correlation

No-correlation

**Positive correlation:**

A positive correlation means that as one variable increases, the other variable also increases. Conversely, if one variable decreases, the other also decreases. That means the both will move in same direction.

Correlation coefficient value ranges between 0 to 1 .

As the correlation coefficient approaching near to 1 means that indicates strength of relationship among variables are high in positive way.

**Negative correlation:**

A negative correlation occurs when one variable increases while the other variable decreases, or vice versa.

That means they both move in opposite direction.

Correlation coefficient value ranges between -1 to 0 .

As the correlation coefficient approaching near to -1 means that indicates strength of relationship among variables are high in negative way

**Zero correlation or no correlation:**

Zero correlation means there is no apparent relationship between the two variables; changes in one variable do not predict changes in the other.

The correlation coefficient is 0

Basically this arises when we want to get a relationship between two variables which are no-relevant to each other.

Ex:

Face color and intelligence typically have zero correlation; knowing someone’s face color doesn’t give information about their intelligence

**Perfect Positive Correlation:**

Perfect positive correlation occurs when two variables increase or decrease together in perfect proportion.

That means a unit change in one variable results a unit in change in other variable in positive direction .

Correlation coefficient value is 1

Ex:

The temperature in Celsius and Fahrenheit have a perfect positive correlation; if one increases, the other does in a linear fashion.

**Perfect negative Correlation:**

perfect negative correlation happens when one variable increases as the other decreases in perfect proportion..

That means a unit change in one variable results a unit in change in other variable in negative direction.

Correlation coefficient value is -1

Ex:

The amount of time left in a countdown timer and the countdown number itself have a perfect negative correlation; as time passes, the countdown decreases in a linear fashion.

**Methods of determining correlation:**

Pearson Correlation Coefficient (r)

Spearman's Rank Correlation Coefficient (ρ)

Kendall's Tau (τ) etc..,

**Pearson Correlation Coefficient (r):**

Iteasures the linear relationship between two continuous variables.

Correlation among the two variables are defined by the correlation coefficient.

Correlation coefficient value ranges between -1 to 1

Correlation coefficient(r) = ​Cov(X,Y)​/ σX​σY​

Cov(x,y) is the covariance between x and y

σ x is the standard deviation of ,x

σy is the standard deviation of y

​

Cov(x,y)= ∑(Xi​−X(mean))(Yi​−Y(mean))​/n-1

->Values close to 1 or -1 indicate a strong linear relationship, while values close to 0 indicate a weak or no linear relationship.

**Spearman's Rank Correlation Coefficient (ρ):**

Measures the strength and direction of the association between two ranked variables. It’s used for ordinal data or when the assumptions of Pearson’s correlation are not met.

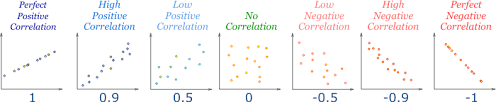
Formula = ρ=1-6∑di2​​/ n(n2−1)

-> di​ is the difference between the ranks of each pair of values.

The **Correlation Coefficient (ρ) ranges between -1 ro 1**

Spearman’s correlation operates on the ranks of the data rather than the raw data values. It is useful when the data does not meet the assumptions of normality or when dealing with ordinal variables

This graph shows show the correlation among two variables:

****

**24. Calculate coefficient of correlation between the marks obtained by 10 students in Accountancy and statistics:  Use Karl Pearson’s Coefficient of Correlation Method to find it.**

We know the Karl Pearson’s coefficient formulae

Correlation coefficient(r) = ​Cov(X,Y)​/ σX​σY​

Where cov(x,y) is the covariance between x and y variable it is given by formulae

Cov(x,y)= ∑(Xi​−X(mean))(Yi​−Y(mean))​/n-1

σX,​σY are the corresponding standard deviations of x and y variable

we know standard deviation formulae=

Variance(σ2)= ∑(xi-μ)^2

N

Std=sqrt(variance)

Now to solve the problem we need to find

∑X (Sum of Accountancy marks)

∑Y (Sum of Statistics marks)

∑X2 (Sum of squares of Accountancy marks)

∑Y2 (Sum of squares of Statistics marks)

∑XY (Sum of the product of Accountancy and Statistics marks)

Based on the given data

∑X(sum of accountancy marks) =45+70+65+30+90+40+50+75+85+60=630

∑Y (Sum of Statistics marks)=35+90+70+40+95+40+60+80+80+50=640

∑X2 (Sum of squares of Accountancy marks)= 45^2+70^2+65^2+30^2+90^2+40^2+50^2+75^2+85^2+60^2=

2025+4900+4225+900+8100+1600+2500+5625+7225+3600=47,650

∑Y2 (Sum of squares of Statistics marks)= 35^2+90^2+70^2+40^2+95^2+40^2+60^2+80^2+80^2+50^2

=1225+8100+4900+1600+9025+1600+3600+6400+6400+2500=57,350

∑XY (Sum of the product of Accountancy and Statistics marks)= (45\*35)+(70\*90)+(65\*70)+(30\*40)+(90\*95)+(40\*40)+(50\*60)+(75\*80)+(85\*80)+(60\*50)

=1575+6300+4550+1200+8550+1600+3000+6000+6800+3000=52,575

Cov(x,y)=

∑(Xi​−X(mean))(Yi​−Y(mean))​/n-1

3,591/10-1

=3,591/9

=399

**Std of x variable( accountancy marks) =**

To calculate standard deviation first we need to calculate variance of x

Variance of x=∑(xi-μ)^2

N-1

N=10

Mean(μ)=630/10=63

Sum of all items/no of items

On evaluating in the formula we get

Variance=3750/(10−1)=93570​=396.67

std=sqrt(variance)

std(x)=19.93

**Std of y variable( statistics marks) =**

**On following same procedure.**

Variance=4,395/(10-1)=488.33

Std(y)=sqrt(variance(y))

=22.09

Now correlation coefficient= Cov(X,Y)​/ σX​σY​

399/(19.93\*22.09)

=0.908

The Pearson’s correlation coefficient r is approximately 0.908. This indicates a strong positive linear relationship between the Accountancy and Statistics marks.

**25. Discuss the 4  differences between correlation and regression**

**Purpose of using them:**

**Correlation:** it measures the strength and direction of a linear relationship between two variables. It only indicates whether and how strongly pairs of variables are related.

**Regression:** it mainly focuses on modeling the relationship between a dependent variable and one or more independent variables. It not only assesses the strength of the relationship but also predicts the value of the dependent variable based on the independent variables.

**direction of relationship:**

**Correlation:** Does not imply causality; it simply tells us how two variables move together. The correlation coefficient (r) ranges from -1 to +1, where the sign indicates the direction of the relationship (positive or negative).

**Regression**: it implies a directional relationship where the independent variable(s) can predict the dependent variable. It assumes one variable influences or causes changes in another.

**Equation:**

**Correlation:** No equation is derived, only a single correlation coefficient is calculated to summarize the relationship.

**Regression:** Provides an equation, known as the regression equation, that quantifies the relationship. In simple linear regression, this equation is Y=a+bX, where Y is the dependent variable, X is the independent variable, a is the intercept, and b is the slope.

**Types of analysis :**

**Correlation:** Symmetrical, meaning it doesn't matter which variable you consider as independent or dependent; the correlation between X and Y is the same as between Y and X.

**Regression:** Asymmetrical, meaning it matters which variable is treated as the dependent variable and which is treated as the independent variable. The outcome and predictions change if the roles of the variables are swapped.

Top of Form

Bottom of Form

**26. Find the most likely price at Delhi corresponding to the price of Rs. 70 at Agra from the following data: Coefficient of correlation between the prices of the two places +0.8.**

Now we need to find the most likely price at Delhi corresponding to the price of Rs. 70 at Agra from the following data: Coefficient of correlation between the prices of the two places +0.8.

Correlation coefficient cant alone used to find the corresponding price at delhi .we need to use the concept of linear regression. The correlation coefficient provides information about the strength and direction of the relationship between the two sets of prices.

Given data:

Given correlation coefficient=+0.8

That means the both have highly positively correlated

Price at agra x=70

**Steps:**

**First we have to determine the linear relationship amog them:**

The correlation coefficient alone does not give the exact regression line parameters. To proceed with the calculation, you typically need additional information such as the means and standard deviations of prices at both locations, as well as the regression coefficients.

If we assume that the price at Delhi is a linear function of the price at Agra and we know the means and standard deviations, we can use the formula for the regression line:

Y=Y(mean)+b(X−X(mean))

where:

Y is the price at Delhi we want to find,

Y(mean) is the mean price at Delhi,

X(mean) is the mean price at Agra,

B is the slope of the regression line, which can be determined from the correlation coefficient.

to find the slope b:

The slope b of the regression line can be calculated using the correlation coefficient and the standard deviations of the two variables:

We can use the formula

**b=r \* (σX/​σY)​​**

where :

σX is the standard deviation of prices at Agra,

σY is the standard deviation of prices at Delhi.

Since we do not have the standard deviations or means, let's assume we have them for simplicity.

We can assume any value lets assume

Mean price at Agra (X(mean)) is Rs. 60 (assumed)

Mean price at Delhi (Y(mean)) is Rs. 80 (assumed)

Standard deviation at Agra (σX​) is Rs. 10 (assumed)

Standard deviation at Delhi (σY) is Rs. 15 (assumed)

Correlation coefficient r=+0.8 (given)

**Finding the value :**

Y=Y(mean)+b(X−X(mean))

b=r \* (σX/​σY)​​

b=0.8.(15/10)

=0.8\*1.5

=1.2

Y=80+1.2\*(70-60)

Y=92

92 is the most likely price at Delhi corresponding to the price of Rs. 70 at Agra from the following data: Coefficient of correlation between the prices of the two places +0.8

Note: I have assumed some params to derive the answer

**27. In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: Variance of x = 9, Regression equations are: (i) 8x−10y = −66; (ii) 40x − 18y = 214. What are (a) the mean values of x and y, (b) the coefficient of correlation between x and y, (c) the σ of y**

Given data

The variance of x is given as 9

Given two regression equations **8x−10y = −66** and **40x − 18y = 214.**

Now our main is by using the given data we have to find the unknown params

Steps:

1. First lets convert the given regression equations into standardize from.

We know the standard eaquation of regression is

Y=mx+c

Where y is dependent and x is independent

**First Regression Equation**:

8x−10y=−66

⇒ 10y=8x+66 ⇒ y=(8/10)x+(66/10)n ⇒ y=0.8x+6.6

**Second Regression Equation**: 40x−18y=214

⇒ 18y=40x−214 ⇒ y=(40/18)​x−(214​/18) ⇒ y=2.22x−11.89

but we need this in the form x=ny+dx . that means here we taking x as dependent.

so rearrange:

40x=18y+214 ⇒ x=(18/40)​y+(214/40) ​⇒ x=0.45y+5.35

1. **the mean values of x and y:**

to find the mean values x(mean) and y(mean) we use the fact that in a regression equation of the form y=mx+c at the point where x=x(mean) and y=y(mean) ,the mean values satisfy both regression equations.

So by using this

From 8x(mean)−10y(mean)=−66

From 40x(mean)−18y(mean)=214

So by solving this both equations we get

Eq1: 8x(mean)−10y(mean)=−66

Eq2: 40x(mean)−18y(mean)​=214

Multiply Equation 1 by 2 to eliminate y:

16x(mean)−20y(mean)​=−132

Subtracting this from Equation 2:

(40x(mean)−18y(mean)​)−(16x(mean)−20y(mean)​)=214+132

24x(mean)+2y(mean)​=346

**X(mean)=346/24​=14.42**

Now substitute x(mean) back into one of the original equations to find y(mean):

8(14.42)−10y(mean)​=−66

Then we get

**Y(mean)= 181.36​/10=18.14**

**b) the coefficient of correlation between x and y**

the correlation coefficient r between x and y can be find using geometric mean of the slopes of the two regression lines. The slope from the first regression equation is 0.80 and from the second is the reciprocal of the slope of y as a function of x, which is 1/0.45

r= sqrt(0.8\*0.45) = sqrt(0.36) = 0.6

**the correlation coefficient r between x and y is 0.6**

**(c) the σ of y**

to find the standard deviation of y ,

we can use a relationship

σy​ = σx​​/b

=3/0.45

**the σ of y**=6.67

so,

mean values=

x(mean) = 14.42

y(mean)=18.14

coefficient of correlation(r) =0.6

standard deviation of y ( σy ): 6.67

**28. What is  Normal Distribution? What are the four Assumptions of Normal Distribution? Explain in detail.**

**Normal distribution:**

the **Normal Distribution**, also known as the Gaussian distribution, is a fundamental probability distribution in statistics. It describes how the values of a random variable are distributed. Specifically, it is a continuous probability distribution that is symmetric around its mean, meaning that data near the mean are more frequent in occurrence than data far from the mean. The shape of the normal distribution is a bell curve, which is characterized by its mean (μ) and standard deviation (σ).

In machine learning we want the data distribution should be in normal distribution.

It is proven that that if the data is in normal distribution the model accuracy will be high .

There are some key features of normal distribution:

**Symmetry**: The normal distribution is symmetric about its mean. This implies that the left side of the distribution is a mirror image of the right side.

**Bell-shaped Curve**: The distribution forms a bell-shaped curve where most of the data points cluster around the mean.

**Mean, Median, and Mode**: In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.

**Asymptotic**: The tails of the distribution curve approach, but never touch, the horizontal axis. This means that the probability of extreme values is never zero but gets very small as you move further from the mean.

**Area under the Curve**: The total area under the curve of a normal distribution is 1. This area represents the total probability of all outcomes.

These are the main key features of normal distribution.

**Mathematical representation:**

The probability density function (PDF) of the normal distribution is given by:

f(x∣μ,σ) = (1/sqrt(2πσ^2​)) \* ​exp(−(x−μ)^2/2σ^2​)

where:

x is the variable,

μ is the mean,

σ is the standard deviation,

exp: denotes the exponential function

important thing is the normal distribution has some assumptions:

main four assumptions are:

**Linearity**:

**Assumption**: The relationship between the variables is linear.

The normal distribution assumes that any linear combination of normally distributed variables will itself be normally distributed. This means that the data should follow a straight-line relationship when plotted on a graph, and deviations from this could indicate that the data is not normally distributed.

**Independence**:

**Assumption**: Observations are independent of each other.

The values of the data points should not be influenced by each other. This assumption is critical because if data points are dependent, the statistical inferences made based on the normal distribution could be invalid. For example, the value of one observation should not predict or affect the value of another.

**Homoscedasticity**:

**Assumption**: Constant variance across the data.

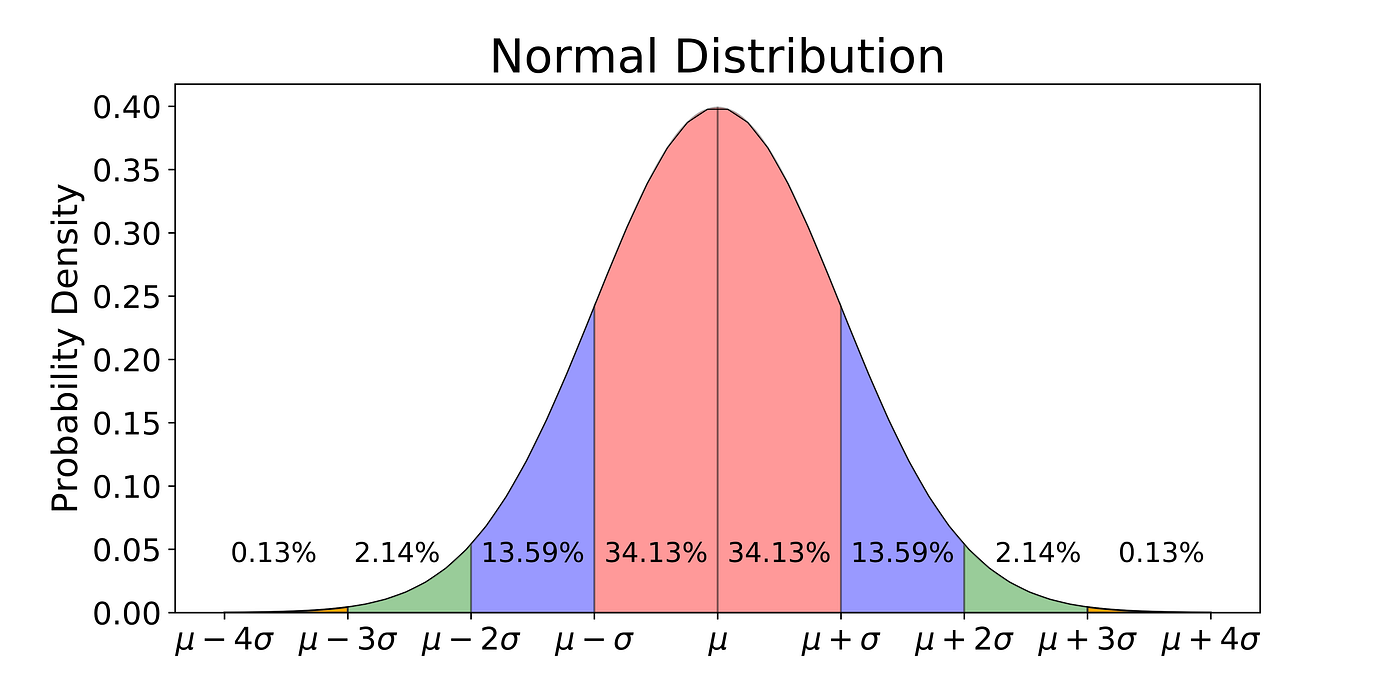
The variability in the data should be consistent across the entire range of data. This means that the spread of the residuals (the differences between observed and predicted values) should be the same regardless of the value of the independent variable. If the variance changes (heteroscedasticity), the data may not be normally distributed.

**Normality of the Error Distribution**:

**Assumption**: The errors (residuals) of the model are normally distributed.

* The differences between observed and predicted values should be normally distributed with a mean of zero. This assumption is essential for making valid inferences about the population from the sample data. If the errors are not normally distributed, it may indicate that the data itself is not normally distributed or that the model used to fit the data is inappropriate.

**Graph of normal distribution:**



**29.Write all the characteristics or Properties of the Normal Distribution Curve.**

Normal distribution:

the **Normal Distribution**, also known as the Gaussian distribution, is a fundamental probability distribution in statistics. It describes how the values of a random variable are distributed. Specifically, it is a continuous probability distribution that is symmetric around its mean, meaning that data near the mean are more frequent in occurrence than data far from the mean. The shape of the normal distribution is a bell curve, which is characterized by its mean (μ) and standard deviation (σ).

**characteristics or Properties of the Normal Distribution Curve:**

the normal distribution curve we can also call it as Gaussian curve . there are some characteristics of normal distribution curve.

**Bell-Shaped Curve:**

The curve is symmetric and bell-shaped, with the highest point at the mean. The shape indicates that most of the data points cluster around the mean, with fewer points occurring as you move further away from the mean.

**Symmetry:**

The normal distribution is perfectly symmetric about the mean. This means that the left and right halves of the curve are mirror images of each other. Consequently, the mean, median, and mode are all equal and located at the center of the distribution.

**Mean, Median, and Mode are Equal:**

n a normal distribution, the mean, median, and mode all coincide at the center of the distribution. This single peak represents the most common value in the dataset

**Asymptotic Nature:**

The tails of the normal distribution curve approach but never actually touch the horizontal axis. This implies that the probability of extreme values (very far from the mean) is never zero but becomes increasingly smaller as you move further from the mean.

**Area Under the Curve:**

The total area under the curve of a normal distribution is equal to 1. This area represents the total probability of all possible outcomes. The area under the curve between any two points corresponds to the probability that a value will fall within that range.

**The most important characteristic of normal dist curve is:**

**Empirical Rule (68-95-99.7 Rule)**

 Approximately:

**68%** of the data falls within one standard deviation of the mean.

**95%** of the data falls within two standard deviations of the mean.

**99.7%** of the data falls within three standard deviations of the mean.

**->** This rule provides a quick way to understand the spread of data in a normal distribution

**No Skewness:**

Since the normal distribution is symmetric, it has a skewness of zero. Skewness measures the degree of asymmetry in a distribution, and in a normal distribution, there is no skewness because the distribution is perfectly symmetrical.

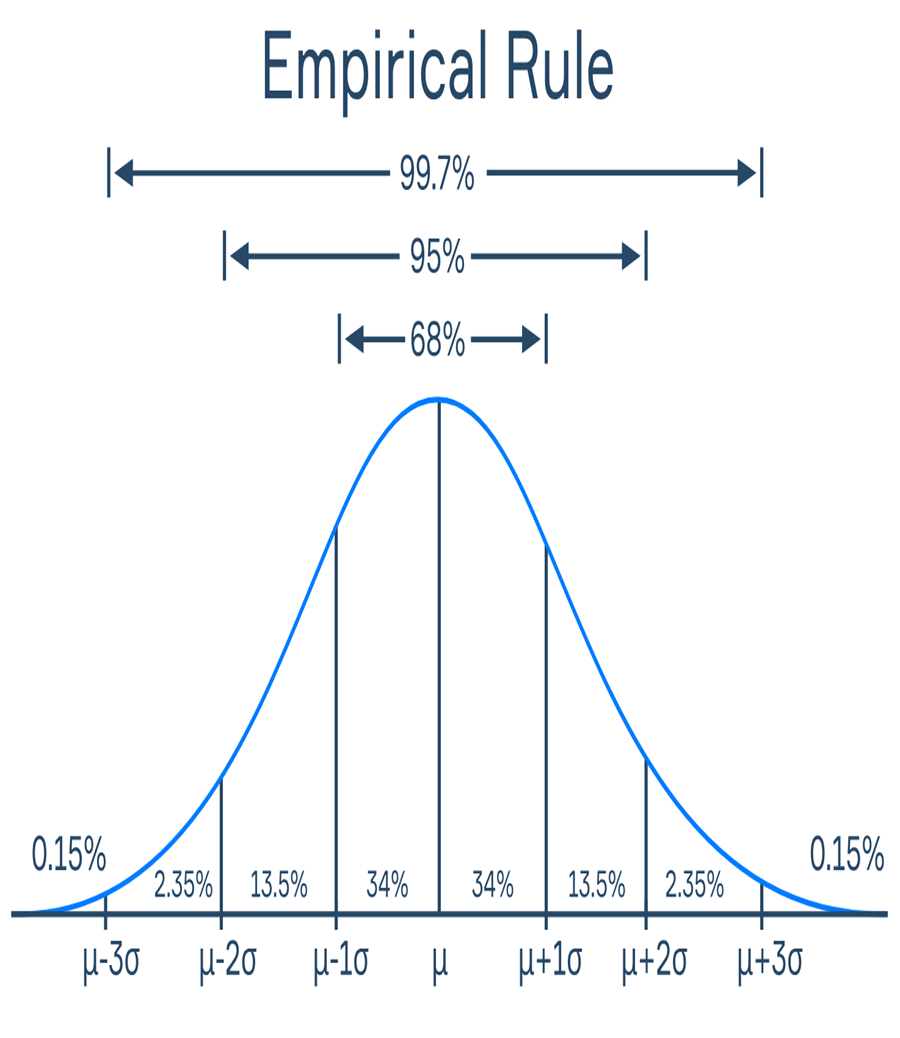
**Kurtosis:**

The normal distribution has a kurtosis of 3, which is often referred to as mesokurtic. Kurtosis measures the "tailedness" of a distribution. A normal distribution does not have heavy tails (which would make it leptokurtic) or light tails (which would make it platykurtic).

**Standard Normal Distribution:**

When a normal distribution has a mean of 0 and a standard deviation of 1, it is called the standard normal distribution. This standardized form allows for the easy calculation of probabilities and z-scores.

**GRAPH THAT REPRESENTS THE EMPIRICAL RULE:**



**30.Which of the following options are  correct about Normal Distribution Curve.**

**(a) Within a range 0.6745 of σ on both sides the middle 50% of the observations occur i.e. mean ±0.6745σ covers 50% area 25% on each side.**

**(b) Mean ±1S.D. (i.e. µ ± 1σ) covers 68.268% area, 34.134 % area lies on either side of the mean.**

**(c) Mean ±2S.D. (i.e. µ ± 2σ) covers 95.45% area, 47.725% area lies on either side of the mean.**

**(d) Mean ±3 S.D. (i.e. µ ±3σ) covers 99.73% area, 49.856% area lies on the either side of the mean Mean ±3 S.D. (i.e. µ ±3σ) covers 99.73% area, 49.856% area lies on the either side of the mean.**

**(e) Only 0.27% area is outside the range µ ±3σ.**

1. **Within a range 0.6745 of σ on both sides the middle 50% of the observations occur i.e. mean ±0.6745σ covers 50% area 25% on each side.**

**CORRECT:** The first option saying that if we look within 0.6745 times the standard deviation (σ) from the mean, we’ll find the middle 50% of all the data points. That’s correct. This range covers half of all the data points, with 25% on each side of the mean.

**b) Mean ±1S.D. (i.e. µ ± 1σ) covers 68.268% area, 34.134 % area lies on either side of the mean.**

**correct:** The second option is saying that if we go one standard deviation (σ) away from the mean, we’ll cover about 68% of all the data. That’s also correct! This means that about 34% of the data is on each side of the mean within this range.

c) **Mean ±2S.D. (i.e. µ ± 2σ) covers 95.45% area, 47.725% area lies on either side of the mean.**

**correct:** The third option is stating that going two standard deviations away from the mean covers about 95% of the data. This range includes almost all of the data, with about 47.7% on each side. So it is also correct.

d) **Mean ±3 S.D. (i.e. µ ±3σ) covers 99.73% area, 49.856% area lies on the either side of the mean**

**correct:** The fourth option is saying that going three standard deviations away from the mean covers 99.73% of the data. This is also correct. This means only a tiny amount of data is outside this range, with about 49.865% on each side.\

e) **Only 0.27% area is outside the range µ ±3σ.**

**Correct:** The last option is saying that only 0.27% of the data lies outside the range of three standard deviations from the mean. This is true . So, almost all data is within three standard deviations of the mean, with very little outside of it.

So we can say that given all options are correct.

**31. The mean of a distribution is 60 with a standard deviation of 10. Assuming that the distribution is normal, what percentage of items be**

**(i) between 60 and 72, (ii) between 50 and 60, (iii) beyond 72 and (iv) between 70 and 80?**

Given that the distribution is normal

And the mean of dist is 60 and standard deviation is 10

We can use z-score to find the percentage of items between the given values

z-score formula = (x- μ)/ σ​

where

X is the value in the distribution.

μ=60 (mean).

σ=10 (standard deviation)

1. **Percentage of items between 60 and 72:**

Finding the z-score for 60:

z-score=(x- μ)/ σ​

Z60 = (60-60)/10

= 0

Finding the z-score for 72:

Zscore for 72=( 72-60)/10

=1.2

Now looking in the standard distribution table:

Z-score of 0 corresponds to 50% (because it is at the mean).

Z-score of 1.2 corresponds to approximately 88.49%

Calculating the percentage of items between 60 and 72:

88.49%−50%=38.49%\

So, **38.49%** of the items are between 60 and 72.

1. **between 50 and 60**

Finding the z-score for 50:

z-score=(x- μ)/ σ​

Z60 = (50-60)/10

= -1

Finding the z-score for 60:

Z-score for 60=( 60-60)/10

=0

Now looking in the standard distribution table:

Z-score of -1 corresponds to 15.87%

Z-score of corresponds to approximately 50%

Calculating the percentage of items between 50 and 60:

50%-15.87%=34.13%

So, **34.13%** of the items are between 50 and 60.

**iii)beyond 72 and**

for z-score 72 we have already calculated and it came as 88.49%

to calculate the percentage of items beyond 72:

100%-88.49%=11.51%

So,

11.51% of the items are beyond 72

**iv)percentage of items between 70 and 80**

z-score for 70:

=(70-60)/10

=1

z-score for 80:

=(80-60)/10

=2

Now looking in the standard normal distribution table

z-score of 1 corresponds to 84.13%

z-score of 2 corresponds to 97.72%

calculating the percentage of items between 70 and 80:

97.72%-84.13%=13.59%

S0,

13.59% of the items are between 70 and 80

**So the final answers:**

(i) Percentage between 60 and 72: **38.49%**

ii) Percentage between 50 and 60: **34.13%**

iii)Percentage beyond 72: **11.51%**

(iv) Percentage between 70 and 80: **13.59%**

**32. 15000 students sat for an examination. The mean marks was 49 and the distribution of marks had a standard deviation of 6. Assuming that the marks were normally distributed what proportion of students scored (a) more than 55 marks, (b) more than 70 marks**

**Given information:**

**Mean Marks** (μ) = 49

**Standard Deviation** (σ) = 6

**Total Students** = 15,000

a)finding proportion of students scored marks more than 55

given it is normally distributed

now finding the z-score for 55 marks:

=(55-49)/6

**=6/6=1**

A z-score of 1 means 55 is 1 standard deviation above the mean.

Percentage of students scoring more than 55

By seeing the z-score for 55 in z-table

We get a percentage of **84.13% of the students below 55.**

So we can simply say 15.87% scored more than 55.

On the basis of this percentage lets calculate no of students below 55 score

To find the number of students:

**15.87%×15000=0.1587×15000≈2381  students**

**b) Find the Proportion of Students Scoring More Than 70**

**z-score for 70:**

(70-49)/6 = 21/6 = 3.5

A z-score of 3.5 is far above the mean, indicating that 70 is much higher than the average score.

**Percentage of students scoring more than 70:**

z-score of 3.5 corresponds to **99.95%** of the students scoring below 70.

This means only **0.05%** scored above 70.

**No of students scoring more than 70:**

Finding how many students:

0.05%×15000=0.0005×15000=7.5 students

So, about **7 or 8 students** scored more than 70 marks

**So final answers:**

**2,381 students** scored more than 55 marks.

**7 or 8 students** scored more than 70 marks.

**33. If the height of 500 students are normally distributed with mean 65 inch and standard deviation 5 inch. How many students have height : a) greater than 70 inch. b) between 60 and 70 inch.**

Given the data is in normal distribution.

To solve this problem we use z-score.

Mean=65 inches

Standard deviation=5 inches

**a)The number of students with height greater than 70 inches**

the z-score is calculated by using the formulae

z=x-μ​/σ

where X is the value for which we're finding the probability.

For x=70 inches,

Z=70-65/5

=5/5

=1

Finding the probability of Z being greater than 1:

Using the standard normal distribution table or a calculator, find the probability for Z≤1, which is approximately 0.8413. This is the cumulative probability up to 70 inches.

Therefore, the probability of a height greater than 70 inches is:

P(Z>1)=1−P(Z≤1)=1−0.8413=0.1587

Next step is to calculate the number of students:

On multiplying the probability by the total number of students:

Number of students=0.1587\*500=79.35

Rounding to the nearest whole number, approximately 79 students have a height greater than 70 inches.

**B)The number of students with height between 60 and 70 inches**

Converting the heights to z-scores:

the z-score is calculated by using the formulae

z=x-μ​/σ

where X is the value for which we're finding the probability.

For x=60 inches,

Z=60-65/5

=-5/5

=-1

Next step is finding the probabilities for these z-scores:

From the Z-table or calculator:

The cumulative probability for Z≤−1 is approximately 0.1587.

The cumulative probability for Z≤1 is approximately 0.8413.

Therefore, the probability of being between 60 and 70 inches is:

P(−1<Z<1)=P(Z≤1)−P(Z≤−1)=0.8413−0.1587=0.6826

Next step is finding the no of students:

Multiply the probability by the total number of students:

Number of students=0.6826×500≈341.3

Rounding to the nearest whole number, approximately 341 students have a height between 60 and 70 inches.

🡪

**Number of students with height greater than 70 inches:** Approximately 79

**Number of students with height between 60 and 70 inches:** Approximately 341Top of Form

Bottom of Form

**34. What is the statistical hypothesis? Explain the errors in hypothesis testing.b)Explain the  Sample. What are Large Samples & Small Samples.**

**Statistical hypoithesis:**

A statistical hypothesis is a statement or assumption about a population parameter that is tested using statistical methods. It provides a basis for conducting statistical tests and drawing conclusions about the data

Types of hypothesis:

There are two types of hypothesis are there:

->Null hypothesis

->alternate hypothesis

Null hypothesis:

The null hypothesis is a statement that there is no effect, no difference, or no relationship between variables. It serves as the default or starting assumption that any observed effects in the data are due to chance or random variation. The goal of hypothesis testing is often to determine whether there is enough evidence to reject this hypothesis.

It is denoted by h0

Alternate hypothesis;

The alternative hypothesis is a statement that contradicts the null hypothesis. It suggests that there is an effect, a difference, or a relationship between variables. This hypothesis represents what researchers are trying to prove or find evidence for.

It is denoted by h1 or ha

**Errors in hypothesis testing:**

In hypothesis testing there are mainly two types of errors are there

->type 1 error

->type II error

Type 1 error: This occurs when the null hypothesis is incorrectly rejected when it is actually true. In other words, we conclude that there is an effect or difference when there isn’t one. The probability of a Type I error is denoted by the significance level (α\alpha), often set at 0.05.

Example: Concluding that a drug is effective when it actually is not.

Type II error:

This occurs when the null hypothesis is not rejected when it is actually false. In other words, you fail to detect an effect or difference when one exists. The probability of a Type II error is denoted by β, and the power of the test (1 - β) indicates the test’s ability to correctly detect an effect.

**Example:** Concluding that a drug is not effective when it actually is.

**Sample:**

A sample is a subset of individuals or observations drawn from a larger population. It is used to make inferences about the population based on the data collected from the sample.

**Types of samples:**

**->large sample**

**->small sample**

**Large sample:** A large sample is typically defined by a sample size n that is sufficiently large to approximate the population characteristics reliably. The exact size that constitutes a "large" sample can vary depending on the context and statistical methods used. In many cases, a sample size of 30 or more is considered large enough to apply the Central Limit Theorem, which allows the use of normal distribution approximations for statistical inference.

**Small samples**:

**:** A small sample is defined by a sample size n that is relatively small compared to the population. When working with small samples, the sampling distribution may not approximate normality, and specific statistical methods, such as t-tests, may be required to account for the small sample size.

The size of sample is les than 30 then it is often consider as a small sample

**35.A random sample of size 25 from a population gives the sample standard derivation to be 9.0. Test the hypothesis that the population standard derivation is 10.5. Hint(Use chi-square distribution)**

Given the size of sample is 25 and its standard deviation is given a s9.0

Population standard deviation=10.5

Hypothesis:

Null hypothesis:

The population standard deviation is 10.5

H0:σ=10.5

Alternate hypothesis:  
the population standard deviation is not 10.5

H1: σ ≠ 10.5

Test statistic:

The test statistic for testing the population standard deviation using the chi-square distribution is given by:

χ2=(n−1)s2​/ σ02

where:

n = sample size

ss= sample standard deviation

σ0 = hypothesized population standard deviation

calculating:

Substituting the given values into the formula:

χ2=(n−1)s2​/ σ02

24\*81/10.25=17.63

Degrees of freedom:

The degrees of freedom (df) for this test is n−1:

Df=25-1=24

Chi-square critical values:

to determine whether to reject the null hypothesis,we comparing the calculated χ2 value to the critical values from the chi-square distribution table at a chosen significance level (usually α=0.05).

For df=24 at α=0.05 , the critical values are:

Lower critical value: χ0.025,242≈13.848

Upper critical value: χ0.975,242≈36.415

Decision:

if χ2 falls between the lower and upper critical values (13.848 and 36.415), we **fail to reject** the null hypothesis.

If χ2 falls outside this range, we **reject** the null hypothesis.

Sinc χ2≈17.63 is between 13.848 and 36.415, we **fail to reject** the null hypothesis.

Conclusion:

There is not enough evidence to conclude that the population standard deviation is different from 10.5, based on the sample provided.

**37.100 students of a PW IOI obtained the following grades in Data Science paper : Grade :[A, B, C, D, E]  Total Frequency :[15, 17, 30, 22, 16, 100] Using the  χ 2 test , examine the hypothesis that the distribution of grades is uniform**

Now we have to find that the distribution of grades is uniform or not.

Given:

Grades:[A,B,C,D,E]

Observed frequencies:[15,17,30,22,16]

Total no of students(N) : 100

Formulating the hypothesis:

Null hypothesis(h0): The grades are uniformly distributed, meaning each grade is equally likely.

Alternate hypothesis(h1): The grades are not uniformly distributed.

Calculating the expected frequencies(E):

If the distribution is uniform, each grade should have the same frequency. Since there are 5 grades and 100 students, the expected frequency for each grade is:

E – 100/5 = 20

So, the expected frequencies for each grade are: E=[20,20,20,20,20]

Calculate the chi-square test statistic:

The chi-square statistic is calculated using the formula:

χ2=∑(oi-Ei)2/Ei)

where Oi and Ei​ are the observed and expected frequencies for each grade.

Substituting these values:

χ2= (15−20)2/20​+(17−20)2/20​+(30−20)2/20​+(22−20)2/20​+(16−20)2​/20

χ2=(−5)2/20​+(−3)2/20​+(10)2/20​+(2)2/20​+(−4)2/20

χ2=1.25+0.45+5+0.2+0.8=7.7​

determining the degrees of freedom:

df=Number of categories−1=5−1=4

finding the critical value:

At a significance level of α=0.05 and 4 degrees of freedom, the critical value from the chi-square distribution table is approximately 9.488.

Decision :

**If χ2 calculated <chi-square critical value,** fail to reject the null hypothesis.

**If χ2 calculated > chi-square critical value,** reject the null hypothesis.

Here, χ2=7.7 is less than the critical value of 9.488.

Conclusion:

We **fail to reject** the null hypothesis, which means there is not enough evidence to conclude that the grades are not uniformly distributed. Therefore, the distribution of grades can be considered uniform based on this sample.

**38: anova test**