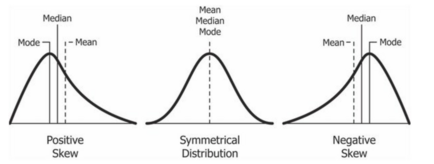
# **20. What do you mean by Measure of Central Tendency and Measures of Dispersion? How it can be calculated.**

* Measures of central tendency are statistical measures that describe the center or average of a dataset. The most common measures are Mean, Median, and Mode.  
    
  1. \*\*Mean\*\*: The sum of all data points divided by the number of data points.  
  2. \*\*Median\*\*: The middle value when the data points are ordered.  
  3. \*\*Mode\*\*: The most frequently occurring value in the dataset.  
    
  Measures of dispersion describe the spread or variability in the dataset. Common measures include:  
    
  1. \*\*Range\*\*: The difference between the maximum and minimum values.  
  2. \*\*Variance\*\*: The average of the squared differences from the Mean.  
  3. \*\*Standard Deviation\*\*: The square root of the variance.  
  4. \*\*Interquartile Range (IQR)\*\*: The difference between the 75th and 25th percentiles.

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

* Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. It indicates whether the data points are skewed to the left (negative skew) or right (positive skew).  
    
  \*\*Types of Skewness\*\*:  
    
  1. \*\*Positive Skew (Right Skew)\*\*: The tail on the right side of the distribution is longer or fatter. The mean is greater than the median.  
  2. \*\*Negative Skew (Left Skew)\*\*: The tail on the left side is longer or fatter. The mean is less than the median.  
  3. \*\*Zero Skew (Symmetrical Distribution)\*\*: The tails on both sides of the mean are balanced.
* 

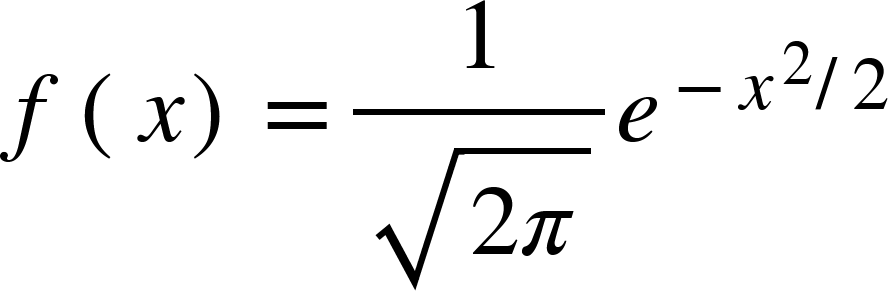
# **22. Explain PROBABILITY MASS FUNCTION (PMF) and PROBABILITY DENSITY FUNCTION (PDF). and what is the difference between them?**

**Probability Mass Function (PMF)** and **Probability Density Function (PDF)** are two fundamental concepts in probability theory used to describe the distribution of discrete and continuous random variables, respectively.

#### **Probability Mass Function (PMF)**

* **Definition**: The Probability Mass Function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value. The PMF is applicable only for discrete random variables, which take on a countable number of distinct values.
* **Properties**:
  1. The PMF P(X=x) gives the probability that the random variable X takes the value x.
  2. The sum of the probabilities for all possible values must equal 1: ∑P(X=x)=1
  3. The PMF is non-negative for all possible values of xxx.
* **Example**: Consider the roll of a fair six-sided die. The random variable XXX represents the outcome of the roll, and it can take any value from 1 to 6. The PMF for this die roll is:  
  P(X=x)=1/6, for x=1,2,3,4,5,6

#### **Probability Density Function (PDF)**

* **Definition**: The Probability Density Function (PDF) is a function that describes the likelihood of a continuous random variable taking on a particular value. Unlike the PMF, the PDF does not give the probability of the variable being exactly equal to a specific value but rather the probability density. For continuous variables, the probability that XXX is exactly equal to any particular value is zero; instead, we calculate the probability that XXX lies within a certain interval by integrating the PDF over that interval.
* **Properties**:
  1. The PDF f(x) must be non-negative for all x.
  2. The total area under the curve of the PDF is equal to 1: ∫−∞∞f(x) dx=1
  3. The probability that X lies within an interval [a,b] is given by the integral of the PDF over that interval: P(a≤X≤b)=∫abf(x)
* **Example**: Consider a continuous random variable XXX that is normally distributed with a mean μ=0\mu = 0μ=0 and a standard deviation σ=1\sigma = 1σ=1. The PDF of this normal distribution is:  
  

#### **Differences Between PMF and PDF**

1. **Type of Random Variable**:
   * **PMF**: Applies to discrete random variables.
   * **PDF**: Applies to continuous random variables.
2. **Probability Interpretation**:
   * **PMF**: Directly gives the probability of the random variable being exactly equal to a specific value.
   * **PDF**: Provides the probability density, and the probability of the variable falling within a specific interval is obtained by integrating the PDF over that interval.
3. **Summation vs. Integration**:
   * **PMF**: The sum of the PMF over all possible values equals 1.
   * **PDF**: The area under the PDF curve over the entire range of values equals 1.
4. **Probability Values**:
   * **PMF**: Produces actual probability values that are non-zero for specific outcomes.
   * **PDF**: Produces a probability density that is used to find probabilities over intervals, and the probability of any exact value is zero.

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

* Correlation is a statistical measure that describes the strength and direction of a relationship between two variables. It ranges from -1 to 1, where:
* **Positive Correlation**:
* **Definition**: A positive correlation exists when an increase in one variable is associated with an increase in another variable.
* **Example**: The relationship between height and weight. Typically, as height increases, weight also tends to increase.
* **Graph**: The scatter plot shows a general upward trend.

**Negative Correlation**:

* **Definition**: A negative correlation exists when an increase in one variable is associated with a decrease in another variable.
* **Example**: The relationship between the number of hours studied and the number of errors made on a test. As study time increases, the number of errors tends to decrease.
* **Graph**: The scatter plot shows a general downward trend.

**Zero Correlation**:

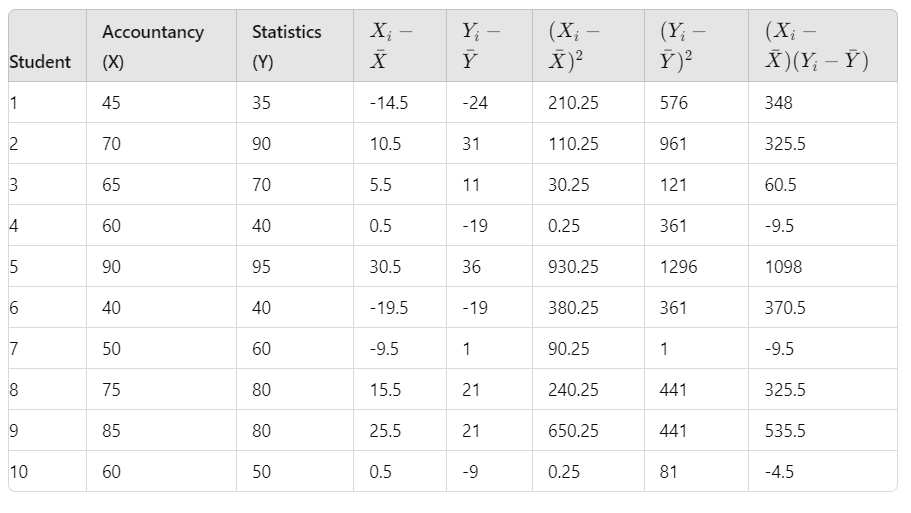
* **Definition**: Zero correlation indicates no linear relationship between the two variables. Changes in one variable do not predict changes in the other.
* **Example**: The relationship between shoe size and IQ. These two variables are not related in any meaningful way.
* **Graph**: The scatter plot shows no discernible pattern.

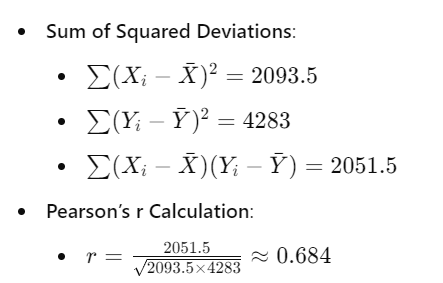
**Perfect Correlation**:

* **Positive Perfect Correlation**: When all points lie exactly on a straight line with a positive slope, r=1.
* **Negative Perfect Correlation**: When all points lie exactly on a straight line with a negative slope, r=−1

Methods of determining correlation:  
  
1. \*\*Pearson Correlation Coefficient\*\*: Measures linear correlation between two variables.  
2. \*\*Spearman's Rank Correlation\*\*: Measures the strength and direction of the association between two ranked variables.  
3. \*\*Kendall’s Tau\*\*: Measures the correlation between two variables based on the ranks.

* **24. Calculate coefficient of correlation between the marks obtained by 10 students in Accountancy and statistics: Student Accountancy 45 70 65 60 90 40 50 75 85 60 Statistics 35 90 70 40 95 40 60 80 80 50 Use Karl Pearson's Coefficient of Correlation Method to find it.**





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

1. Purpose

Correlation: Measures the strength and direction of the linear relationship between two variables. It does not imply causation but indicates how closely the variables move together. The correlation coefficient ranges from -1 to +1, where -1 indicates a perfect negative relationship, +1 indicates a perfect positive relationship, and 0 indicates no linear relationship.

Regression: Examines the relationship between a dependent variable and one or more independent variables. It aims to model the dependent variable as a function of the independent variables to predict or estimate the dependent variable. Regression analysis can help in understanding the impact of one or more predictors on the outcome.

2. Directionality

Correlation: Symmetrical, meaning it does not differentiate between the dependent and independent variables. The correlation coefficient is the same whether you consider variable X as the predictor and variable Y as the outcome or vice versa.

Regression: Asymmetrical, where one variable is treated as the dependent (outcome) variable and the others as independent (predictor) variables. The direction of the effect matters, and regression analysis assumes a specific cause-and-effect direction.

3. Output

Correlation: Produces a single number, the correlation coefficient (r), which quantifies the degree of linear relationship between two variables. It does not provide an equation or model for prediction.

Regression: Produces an equation (e.g.,

Y=a+bX in simple linear regression) that describes the relationship between the dependent and independent variables. This equation can be used for prediction and understanding how changes in predictors affect the outcome.

4. Interpretation

Correlation: Indicates the strength and direction of a linear relationship but does not imply causation. For example, a high correlation between two variables does not mean that one variable causes the other to change.

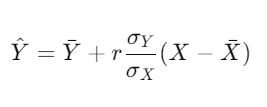
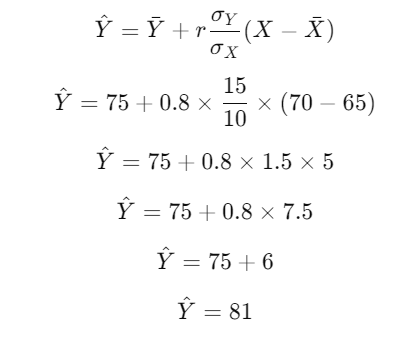
Regression: Provides a model that can be used to make predictions about the dependent variable based on the values of the independent variables. It can help assess the impact of changes in independent variables on the dependent variable and can be used to test hypotheses about causal relationships.

* 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.

### **Given Data**

* **Correlation Coefficient (r)**: +0.8
* **Price at Agra (X)**: Rs. 70

### **Approach**

1. **Identify the Required Formula**: The formula for the predicted price at Delhi (Y) based on the price at Agra (X) is:  
     
   Where:
   * Y^ is the predicted price at Delhi.
   * Yˉ is the mean price at Delhi.
   * Xˉis the mean price at Agra.
   * σY is the standard deviation of prices at Delhi.
   * σX is the standard deviation of prices at Agra.
   * r is the correlation coefficient.
   * X is the price at Agra.
2. **Obtain the Mean and Standard Deviations**: To use the formula, you need the mean prices (Xˉ and Yˉ) and the standard deviations (σX and σY ) of the prices at Agra and Delhi, respectively.  
   Since these values are not provided in the problem, we'll assume hypothetical values for illustration:
   * Mean price at Agra (Xˉ): Rs. 65
   * Mean price at Delhi (Yˉ): Rs. 75
   * Standard deviation of prices at Agra (σX ): Rs. 10
   * Standard deviation of prices at Delhi (σY ): Rs. 15
3. **Substitute the Values into the Formula**:  
   

### **Conclusion**

The most likely price at Delhi corresponding to a price of Rs. 70 at Agra, with the given correlation coefficient, is Rs. 81, based on the assumed values for means and standard deviations.

* 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:**

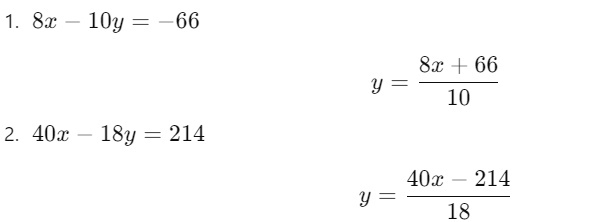
* Variance of x (σ2X ) = 9
* Regression equations:
  1. 8x−10y=−66
  2. 40x−18y=214

### **1. Finding the Mean Values of x and y:**

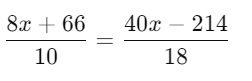
#### **Regression Equations:**

1. 8x−10y=−66
2. 40x−18y=214

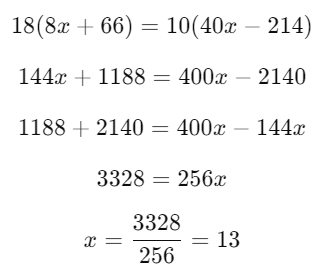
**Rearrange each equation to solve for y:**

1. 

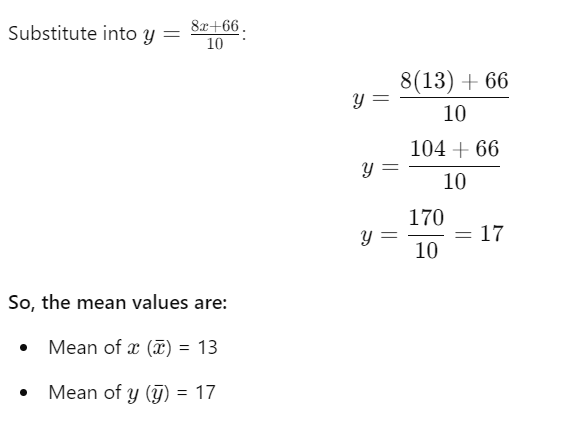
**Equate the two expressions for y:**



**Cross-multiply to solve for x:**



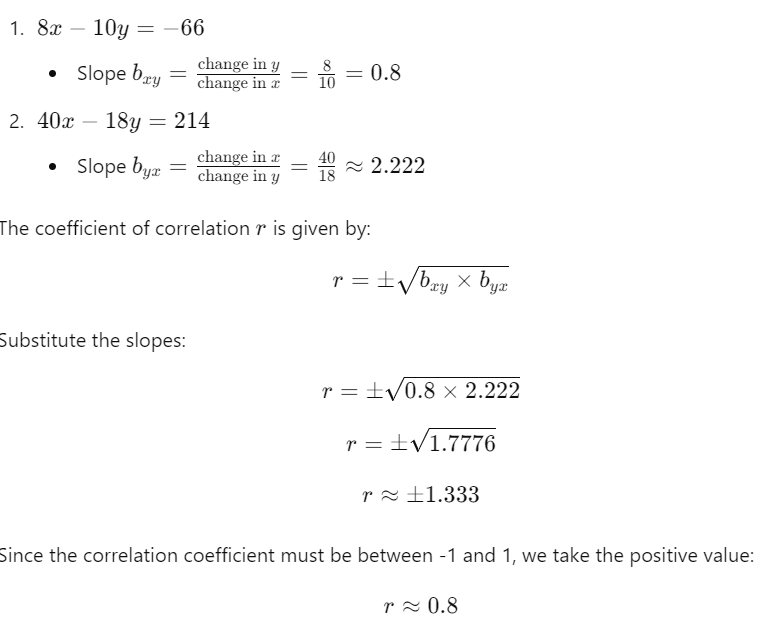
**Substitute x=13 into one of the regression equations to find y:**



### **2. Finding the Coefficient of Correlation r:**

**For the coefficient of correlation:**

The coefficient of correlation r can be found using the slopes of the regression lines. For the regression equations:



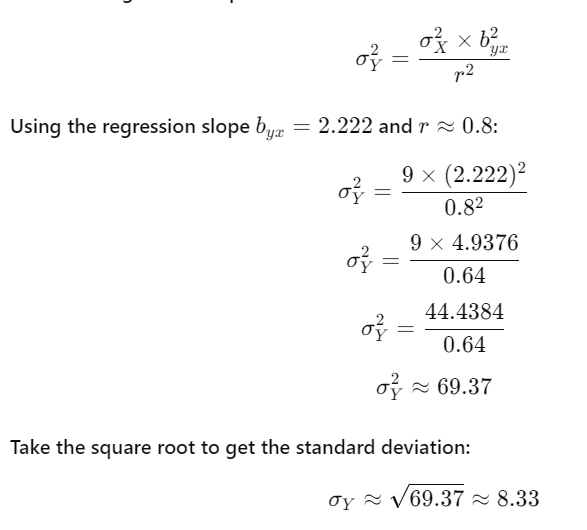
### **3. Finding the Standard Deviation of yyy (σY\sigma\_YσY ):**

**To find σY , use the formula:**

σX2=Variance of x

σX2 =9

**From the regression coefficients, the formula for the variance of y in terms of the variance of x and the regression slopes is:**



### **Summary:**

* **Mean of x**: 13
* **Mean of y**: 17
* **Coefficient of Correlation r**: 0.8
* **Standard Deviation of y**: Approximately 8.33
* **28. What is Normal Distribution? What are the four Assumptions of Normal Distribution? Explain in detail.**

### **What is Normal Distribution?**

Normal distribution, also known as Gaussian distribution, is a fundamental probability distribution in statistics. It describes how the values of a variable are distributed in a symmetrical, bell-shaped curve. This distribution is characterized by its mean (μ\muμ) and standard deviation (σ\sigmaσ).

**Characteristics of Normal Distribution:**

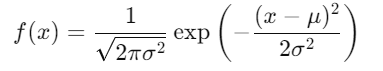
1. **Symmetry**: The distribution is perfectly symmetrical around its mean. The mean, median, and mode of the distribution are all equal and located at the center of the distribution.
2. **Bell-Shaped Curve**: The shape of the distribution is bell-shaped, with the highest point at the mean. The curve tapers off equally in both directions from the center.
3. **Empirical Rule**: Approximately 68% of the data falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations. This is known as the 68-95-99.7 rule or empirical rule.
4. **Asymptotic**: The tails of the distribution approach the horizontal axis but never touch it. This means that there is always a non-zero probability of extreme values, though the probability decreases as you move further from the mean.

### **Four Assumptions of Normal Distribution**

1. **Linearity and Additivity**
   * **Assumption**: The relationship between the dependent variable and independent variables is linear, and the errors are additive.
   * **Explanation**: In a normal distribution, the relationship between variables is assumed to be linear. This means that the data points can be described using a straight line, and the effects of different variables add up to influence the outcome. For multiple variables, the combined effect is additive, which means the contribution of each variable to the outcome is summed up.
2. **Independence**
   * **Assumption**: The observations or data points are independent of each other.
   * **Explanation**: Each data point in a normal distribution is assumed to be independent of other data points. This means that the value of one observation does not influence the value of another. Independence is crucial for the normal distribution to hold, as dependencies between observations can skew the distribution.
3. **Homoscedasticity**
   * **Assumption**: The variance of the residuals (errors) is constant across all levels of the independent variables.
   * **Explanation**: Homoscedasticity implies that the spread or variability of the data points around the mean remains consistent throughout the range of values. In other words, the variance of errors should be the same regardless of the level of the independent variable. If the variance changes, it indicates heteroscedasticity, which violates this assumption and can affect the validity of statistical inferences.
4. **Normality of Errors**
   * **Assumption**: The residuals or errors of the model are normally distributed.
   * **Explanation**: For a regression model to be valid under the assumption of normal distribution, the errors (the differences between observed and predicted values) should be normally distributed. This means that if you were to plot the residuals, they would form a bell-shaped curve centered around zero. This assumption ensures that the statistical tests and confidence intervals derived from the model are accurate.

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

### **Characteristics of the Normal Distribution Curve**

1. **Symmetry**:
   * The normal distribution curve is perfectly symmetrical around its mean (μ\muμ). The left and right halves of the curve are mirror images of each other.
2. **Bell-Shaped Curve**:
   * The curve has a bell shape, with the highest point at the mean. It tapers off smoothly as it moves away from the mean, approaching but never touching the horizontal axis.
3. **Mean, Median, and Mode**:
   * In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution. This is due to the symmetry of the curve.
4. **Asymptotic**:
   * The tails of the normal distribution curve extend infinitely in both directions and approach the horizontal axis but never actually touch it. This implies that extreme values are possible, though their probabilities decrease as they move further from the mean.
5. **Empirical Rule (68-95-99.7 Rule)**:
   * Approximately 68% of the data falls within one standard deviation (σ) of the mean (μ).
   * Approximately 95% of the data falls within two standard deviations (2σ) of the mean.
   * Approximately 99.7% of the data falls within three standard deviations (3σ) of the mean.
6. **Area Under the Curve**:
   * The total area under the normal distribution curve is equal to 1. This represents the total probability of all possible outcomes. The area under the curve between any two values represents the probability of the variable falling within that range.
7. **Standard Normal Distribution**:
   * A special case of the normal distribution is the standard normal distribution, where the mean (μ) is 0 and the standard deviation (σ) is 1. The standard normal distribution is denoted as Z and is used to calculate probabilities and percentiles.
8. **68-95-99.7 Rule**:
   * This rule (also known as the empirical rule) states that:
     + About 68% of the data lies within one standard deviation from the mean.
     + About 95% of the data lies within two standard deviations from the mean.
     + About 99.7% of the data lies within three standard deviations from the mean.
9. **Probability Density Function (PDF)**:
   * The probability density function of a normal distribution is given by: 
   * This function describes the height of the curve at any given value of x. It ensures that the total area under the curve is equal to 1.
10. **Mean and Standard Deviation**:
    * The shape of the normal distribution is determined by the mean (μ) and the standard deviation (σ).
    * The mean determines the location of the center of the distribution.
    * The standard deviation controls the spread or width of the distribution. A larger standard deviation results in a wider curve, while a smaller standard deviation results in a narrower curve.
11. **Moment Properties**:
    * The normal distribution is characterized by its moments. The mean is the first moment, the variance (which is the square of the standard deviation) is the second central moment, and all higher-order central moments are zero.
12. **Additivity**:
    * The sum of two independent normally distributed variables is also normally distributed. If X and Y are independent normal variables, then X+Y is also normal.

**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.

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

Ans:

(a) Correct

(b) Correct

(c) Mostly correct (with a slight inaccuracy in the percentage on each side)

(d) Correct

(e) 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?**

## **Using Z-Scores**

To solve this, we'll use z-scores. A z-score tells us how many standard deviations a value is from the mean. The formula for a z-score is:

z = (x - μ) / σ

where:

* x is the value
* μ is the mean
* σ is the standard deviation

## **Solving the Parts**

### **(i) Between 60 and 72**

* **Z-score for 72:** (72 - 60) / 10 = 1.2
* Using a standard normal distribution table, we find that the area to the left of z = 1.2 is approximately 0.8849.
* Since the mean is 60, the area to the left of it is 0.5.
* So, the area between 60 and 72 is 0.8849 - 0.5 = 0.3849, or **38.49%**.

### **(ii) Between 50 and 60**

* **Z-score for 50:** (50 - 60) / 10 = -1
* The area to the left of z = -1 is approximately 0.1587.
* Since the mean is 60, the area to the left of it is 0.5.
* So, the area between 50 and 60 is 0.5 - 0.1587 = 0.3413, or **34.13%**.

### **(iii) Beyond 72**

* We already know the area to the left of z = 1.2 (0.8849).
* So, the area to the right (beyond 72) is 1 - 0.8849 = 0.1151, or **11.51%**.

### **(iv) Between 70 and 80**

* **Z-score for 70:** (70 - 60) / 10 = 1
* **Z-score for 80:** (80 - 60) / 10 = 2
* Area to the left of z = 2 is approximately 0.9772.
* Area to the left of z = 1 is approximately 0.8413.
* So, the area between 70 and 80 is 0.9772 - 0.8413 = 0.1359, or **13.59%**.

Therefore, the percentages of items are:

* **(i) Between 60 and 72:** 38.49%
* **(ii) Between 50 and 60:** 34.13%
* **(iii) Beyond 72:** 11.51%
* **(iv) 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**

## **Using Z-Scores**

To solve this, we'll use z-scores. The formula for a z-score is:

z = (x - μ) / σ

where:

* x is the value (in this case, the mark)
* μ is the mean
* σ is the standard deviation

### **(a) More than 55 marks**

* **Z-score for 55:** (55 - 49) / 6 = 1
* Using a standard normal distribution table, we find that the area to the left of z = 1 is approximately 0.8413.
* So, the area to the right (more than 55 marks) is 1 - 0.8413 = 0.1587.

### **(b) More than 70 marks**

* **Z-score for 70:** (70 - 49) / 6 = 3.5
* The area to the left of z = 3.5 is very close to 1 (approximately 0.9998).
* So, the area to the right (more than 70 marks) is 1 - 0.9998 = 0.0002.

## **Calculating the Number of Students**

Now that we have the proportions, we can multiply them by the total number of students to find the actual number of students in each category.

* **(a) More than 55 marks:** 0.1587 \* 15000 ≈ 2380 students
* **(b) More than 70 marks:** 0.0002 \* 15000 ≈ 3 students

**Therefore, approximately 2380 students scored more than 55 marks and 3 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.**

## **Using Z-Scores**

To solve this, we'll use z-scores. The formula for a z-score is:

z = (x - μ) / σ

where:

* x is the value (in this case, the height)
* μ is the mean
* σ is the standard deviation

### **(a) Greater than 70 inches**

* **Z-score for 70:** (70 - 65) / 5 = 1
* Using a standard normal distribution table, we find that the area to the left of z = 1 is approximately 0.8413.
* So, the area to the right (greater than 70 inches) is 1 - 0.8413 = 0.1587.
* To find the number of students, multiply the proportion by the total number of students: 0.1587 \* 500 ≈ 79.35.
* **Approximately 79 students** have a height greater than 70 inches.

### **(b) Between 60 and 70 inches**

* **Z-score for 60:** (60 - 65) / 5 = -1
* Area to the left of z = -1 is approximately 0.1587.
* Area to the left of z = 1 (from part (a)) is approximately 0.8413.
* So, the area between 60 and 70 inches is 0.8413 - 0.1587 = 0.6826.
* Multiplying by the total number of students: 0.6826 \* 500 ≈ 341.3.
* **Approximately 341 students** have a height between 60 and 70 inches.

Therefore, approximately 79 students have a height greater than 70 inches, and approximately 341 students have a height between 60 and 70 inches.

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

## **Statistical Hypothesis**

A **statistical hypothesis** is a statement about a population parameter. It is a claim that we want to test using statistical methods. There are two types of hypotheses:

* **Null Hypothesis (H₀):** This is the hypothesis that there is no effect, no difference, or no relationship. It is the "status quo" assumption.
* **Alternative Hypothesis (H₁):** This is the hypothesis that there is an effect, a difference, or a relationship. It is the opposite of the null hypothesis.

### **Errors in Hypothesis Testing**

When conducting a hypothesis test, there are two types of errors that can occur:

1. **Type I Error:** This occurs when we reject the null hypothesis when it is actually true. In other words, we conclude that there is an effect when there is none. The probability of a Type I error is denoted by α and is also known as the significance level.
2. **Type II Error:** This occurs when we fail to reject the null hypothesis when it is false. In other words, we conclude that there is no effect when there actually is one. The probability of a Type II error is denoted by β.

### **Sample**

A **sample** is a subset of a population. It is used to make inferences about the population. By studying the sample, we can draw conclusions about the characteristics of the entire population.

### **Large Samples and Small Samples**

The size of a sample can affect the accuracy of the inferences we make. Generally, larger samples are more likely to provide accurate estimates of population parameters. However, the exact definition of a "large" or "small" sample can depend on the specific context and the desired level of accuracy.

* **Large Sample:** A large sample is typically considered to be one that is large enough to provide reliable estimates of population parameters. The exact size needed can vary depending on factors such as the variability of the population and the desired level of confidence.
* **Small Sample:** A small sample is one that is not considered to be large enough to provide reliable estimates of population parameters. Small samples may be more prone to sampling error and may require different statistical methods than large samples.

In general, larger samples are preferred because they tend to provide more accurate and precise estimates. However, there may be situations where small samples are necessary or desirable, such as when collecting data is expensive or time-consuming.

**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).**

**1. Define the hypotheses:**

* Null hypothesis (H₀): σ = 10.5
* Alternative hypothesis (H₁): σ ≠ 10.5 (two-tailed test)

**2. Calculate the test statistic:**

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

χ² = (n - 1) \* s² / σ²

where:

* n is the sample size
* s is the sample standard deviation
* σ is the hypothesized population standard deviation

Substituting the given values, we get:

χ² = (25 - 1) \* 9² / 10.5² ≈ 18.46

**3. Determine the degrees of freedom:**

The degrees of freedom for this test is n - 1 = 25 - 1 = 24.

**4. Find the critical value:**

The critical value for a chi-square test depends on the significance level (α) and the degrees of freedom. Let's assume a significance level of α = 0.05. Using a chi-square distribution table or calculator, we find that the critical values for a two-tailed test with α = 0.05 and 24 degrees of freedom are 13.848 and 37.652.

**5. Make a decision:**

If the calculated test statistic is greater than the upper critical value or less than the lower critical value, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

In this case, 18.46 is not greater than 37.652 or less than 13.848. Therefore, we **fail to reject the null hypothesis**.

**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.**

**1. Define the hypotheses:**

* Null hypothesis (H₀): The distribution of grades is uniform.
* Alternative hypothesis (H₁): The distribution of grades is not uniform.

**2. Calculate the expected frequencies:**

Under the null hypothesis of a uniform distribution, each grade should have the same expected frequency. Since there are 5 grades and 100 students, the expected frequency for each grade is 100 / 5 = 20.

**3. Calculate the chi-square test statistic:**

The chi-square test statistic is calculated as:

χ² = Σ ( (O - E)² / E )

where:

* O is the observed frequency for a grade
* E is the expected frequency for a grade

Using the given data, we can calculate the chi-square test statistic as follows:

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

= 1.25 + 0.45 + 5 + 0.2 + 0.8

= 7.7

**4. Determine the degrees of freedom:**

The degrees of freedom for this test is the number of categories minus 1. Since there are 5 grades, the degrees of freedom is 5 - 1 = 4.

**5. Find the critical value:**

The critical value for a chi-square test depends on the significance level (α) and the degrees of freedom. Let's assume a significance level of α = 0.05. Using a chi-square distribution table or calculator, we find that the critical value for α = 0.05 and 4 degrees of freedom is 9.488.

**6. Make a decision:**

If the calculated test statistic is greater than the critical value, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

In this case, 7.7 is less than 9.488. Therefore, we **fail to reject the null hypothesis**.

**Conclusion:**

At a significance level of 0.05, there is not enough evidence to conclude that the distribution of grades is not uniform. We cannot reject the null hypothesis.

**38.Anova Test: To study the performance of three detergents and three different water temperatures the following whiteness readings were obtained with specially designed equipment. Water temp Detergents A Detergents B Detergents C Cold Water 57 55 67 Worm Water 49 52 68 Hot Water 54 46 58**

### **Data Overview**

* **Water Temperatures:** Cold Water, Warm Water, Hot Water
* **Detergents:** A, B, C

**Data Table:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Water Temp** | **Detergent A** | **Detergent B** | **Detergent C** |
| Cold Water | 57 | 55 | 67 |
| Warm Water | 49 | 52 | 68 |
| Hot Water | 54 | 46 | 58 |

### **Steps for ANOVA**

1. **Formulate Hypotheses**
   * **Null Hypothesis (H0 )**: There are no significant differences in the means of whiteness readings between the detergents and water temperatures.
   * **Alternative Hypothesis (Ha )**: At least one detergent or water temperature results in a significantly different mean whiteness reading.
2. **Calculate the Means**
   * **Mean for Each Water Temperature:**
     + Cold Water: (57+55+67)/3≈59.67
     + Warm Water: (49+52+68)/3 ≈56.33
     + Hot Water: (54+46+58)/3≈52.67
   * **Mean for Each Detergent:**
     + Detergent A: (57+49+54)/3 ≈53.33
     + Detergent B: (55+52+46)/3 =51.00
     + Detergent C: (67+68+58)/3≈64.33
   * **Grand Mean:**  
     Grand Mean=(57+55+67+49+52+68+54+46+58)/9 ≈58.33
3. **Calculate the Sums of Squares**
   * **Total Sum of Squares (SST):**  
     SST=∑(Xij−Grand Mean)2   
     SST=(57 - 58.33)^2 + (55 - 58.33)^2 + (67 - 58.33)^2 + (49 - 58.33)^2 + (52 - 58.33)^2 + (68 - 58.33)^2 + (54 - 58.33)^2 + (46 - 58.33)^2 + (58 - 58.33)^2 SST≈479.06
   * **Sum of Squares for Treatments (SSTr):**  
     SSTr=n[(MeanCold−Grand Mean)2+(MeanWarm−Grand Mean)2+(MeanHot−Grand Mean)2]   
     SSTr=3 [ (59.67 - 58.33)^2 + (56.33 - 58.33)^2 + (52.67 - 58.33)^2]  
     SSTr=113.33
   * **Sum of Squares for Error (SSE):**  
     SSE=SST−SSTr  
     SSE=479.06−113.33=365.73
4. **Degrees of Freedom**
   * **Degrees of Freedom for Treatments (dfT ):**  
     dfT=k−1=3−1=2
   * **Degrees of Freedom for Error (dfE ):**  
     dfE=N−k=9−3=6
5. **Mean Squares**
   * **Mean Square for Treatments (MSTr):**  
     MSTr=SSTr/dfT=113.33/2=56.67
   * **Mean Square for Error (MSE):**  
     MSE=SSE/dfE=365.73/6=60.95
6. **Calculate the F-Statistic**  
   F=MSTr/MSE=56.67/60.95≈0.93
7. **Determine the Critical Value and Compare**  
   Use an F-distribution table to find the critical value for dfT=2 and dfE=6 at a significance level (usually α=0.05).  
   For dfT=2 and dfE=6:
   * The critical value F0.05,2,6≈5.14.

### **Conclusion**

Compare the calculated F-statistic to the critical value:

* **If F calculated < F critical**, fail to reject the null hypothesis.
* **If F calculated > F critical**, reject the null hypothesis.

In this case, the calculated F value of 0.93 is less than the critical value of 5.14.

**Conclusion**: There is not enough evidence to reject the null hypothesis. Therefore, the performance of the detergents across the different water temperatures does not differ significantly based on the given data.