CHAPTER 5

# Data Analytics

www.ilmkidunya.com

**Q. What is data analytics and statistics?**

**Data Analytics**

**Data analytics** is the process of examining and analyzing data to find useful information, patterns or trends that can help in decision-making.

**Statistics**

**Statistics** is a branch of mathematics that is used to understand, analyze and interpret data. Statistics is important because it can summarize large sets of data in a simple way that makes it easier to draw useful conclusions and make decisions. It is widely used in various fields such as education, business, healthcare and economics etc.

**Q. What are measures of central tendency? Briefly describe the three main measures of central tendency with examples.**

**Measures of Central Tendency**

**Measures of central tendency** are statistical tools used to find the central or typical value in a dataset. The three main measures are **mean**, **median** and **mode**.

**1. Mean (Average)**

**Mean** is the average of all the values in a dataset. It is calculated by adding all the data values and then dividing by the number of values.

**Example**

Suppose the scores of five students in a test are 41, 12, 30, 25 and 9. The mean score is calculated by adding all the scores and then dividing by 5.

**Data set:** 41 12 30 25 9
**Mean:** (41 + 12 + 30 + 25 + 9) / 5 = 23.4

**2. Mode**

**Mode** is the value that appears most frequently in a dataset. It helps to identify the most common or repeated value in a dataset. There can be more than one mode if multiple values appear with the same highest frequency.

**Example 1: Single Mode**

Suppose the scores of five students are 50, 60, 70, 70 and 90. The number **70** appears twice whereas all other numbers appear only once. Therefore, the mode is **70** as it is the most frequent value.

**Example 2: Multiple Modes**

Suppose the scores of six students are 50, 60, 70, 70, 60 and 90. Both **60** and **70** appear twice which is the highest frequency in this dataset. Therefore, this dataset has two modes 60 and 70.

**3. Median**

**Median** is the middle value when the numbers are arranged in order. The median is the exact middle value if there is an odd number of values. The median is the average of the two middle values if there is an even number of values.

**Example 1**

Suppose the scores of five students are 50, 60, 70, 80 and 90. These scores are already in ascending order. The number of scores is odd with the middle value 70. Therefore, the median score is 70.

**Example 2**

Suppose the scores of four students are 50, 60, 70 and 80. These scores are already in ascending order. The number of scores is even. Therefore, the median score is the average of two middle values of 60 and 70 as (60 + 70) / 2 = 65.

**Q. What are measures of dispersion? List two commonly used measures of dispersion.**

**Measures of Dispersion**

Measures of dispersion describe how spread out or scattered the data values are in a dataset. They help to understand the degree of variation or inconsistency among the data points. These measures are important for identifying how much individual values differ from the mean (average).

Two commonly used measures of dispersion are as follows:

- Variance
- Standard deviation

**Q. What is variance. Explain how variance is calculated and interpreted**

**Variance**

**Variance** is a statistical measure that shows how much the values in a dataset differ from the mean (average). A higher variance means the values are more spread out from the mean. A lower variance means the values are closer to the mean. Variance is calculated using the following formula:

$$Variance(\sigma^2) = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Where

- $\sigma^2$ is the symbol for variance
- $X_i$ indicates each individual value in the dataset
- $\mu$ is the mean (average) of the dataset
- $\sum$ is the sum of all squared differences
- N indicates the total number of values in the dataset

www.ilmkidunya.com

**Example**

The following example calculates and compares the variance for two sets of test scores from Class A and Class B. Variance indicates how spread out the scores are in each class.

**Step 1: Variance for Class A**

**Class A Scores:** 50, 52, 55, 57, 60

**Step 1.1: Compute the Mean (μ)**

$$\mu = \frac{50 + 52 + 55 + 57 + 60}{5} = \frac{274}{5} = 54.8$$

**Step 1.2: Compute Each Squared Deviation $(xi-\mu)^2$**

**Xi−μ:** The difference between each data value (xi) and the mean (μ) is called the deviation. It shows how far a value is from the average.

**$(Xi-\mu)^2$:** This is the squared deviation. Squaring the deviation converts negative numbers into positive so that all deviations contribute equally. It also gives more weight to larger differences that helps to highlight the values that are far from the mean.

| Score (xi) | (xi−μ) | $(xi-\mu)^2$ |
|---|---|---|
| 50 | 50 − 54.8 = -4.8 | 23.04 |
| 52 | 52 − 54.8 = -2.8 | 7.84 |
| 55 | 55 − 54.8 = 0.2 | 0.04 |
| 57 | 57 − 54.8 = 2.2 | 4.84 |
| 60 | 60 − 54.8 = 5.2 | 27.04 |

**Step 1.3: Compute Variance**

$$\sigma^2 = \frac{23.04 + 7.84 + 0.04 + 4.84 + 27.04}{5} = \frac{62.8}{5} = 12.56$$

**Step 2: Variance for Class B**

**Class B Scores:** 30, 45, 55, 75, 90

**Step 2.1: Compute the Mean (μ)**

$$\mu = \frac{30 + 45 + 55 + 75 + 90}{5} = \frac{295}{5} = 59$$

**Step 2.2: Compute Each Squared Deviation $(xi-\mu)^2$**

| Score (xi) | (xi−μ) | $(xi-\mu)^2$ |
|---|---|---|
| 30 | 30 − 59 = -29 | 841 |
| 45 | 45 − 59 = -14 | 196 |
| 55 | 55 − 59 = -4 | 16 |
| 75 | 75 − 59 = 16 | 256 |
| 90 | 90 − 59 = 31 | 961 |

**Step 2.3: Compute Variance**

$$\sigma^2 = \frac{841 + 196 + 16 + 256 + 961}{5} = \frac{2270}{5} = 454$$

**Conclusion**

**Variance of Class A =** 12.56

Class A has a low variance of 12.56 that means the scores in Class A are close to the average (mean) of 54.8.

www.ilmkidunya.com

This suggests that:

- Most students in Class A performed similarly.
- There is less variation in performance.
- The class shows more consistency in test scores.

**Variance of Class B = 454**

Class B has a high variance of 454 that means the scores are widely spread out from the mean of 59. This indicates that:

- Students in Class B performed very differently from each other.
- There is a large gap between the lowest (30) and highest (90) scores.
- The data is less consistent and some students scored much higher or lower than others.

**Q. What is standard deviation? Explain how standard deviation is calculated and interpreted.**

**Standard Deviation**

**Standard deviation** is a statistical measure that indicates how much the values in a dataset deviate from the mean (average). It is calculated as the square root of the variance. A small standard deviation means that most values are close to the mean, indicating low variability. A large standard deviation means that values are spread far from the mean, indicating high variability.

Standard deviation provides a more practical and meaningful result because it is expressed in the same units as the original data. In contrast, variance is often considered less practical for real-life interpretation since its units are squared such as $ft^2$, $kg^2$ or $m^2$. These are different from the original units of measurement.

Standard deviation is calculated using the following formula:

$$Standard\ Deviation(\sigma) = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Where,

- σ is the symbol for standard deviation.
- Xi indicates each individual value in the dataset
- μ is the mean (average) of all the data values.
- N is the total number of values in the dataset
- $(xi-\mu)^2$ indicates the squared difference between each value and the mean
- ∑ is the sum of all squared differences

**Example**

The following example calculates and compares the standard deviation of test scores from two different classes such as Class A and Class B.

**Step 1: Calculate mean**

**Class A Scores:** 50, 52, 55, 57, 60

$$Mean(\mu_A) = \frac{50+52+55+57+60}{5}$$

$$Mean(\mu_A) = \frac{274}{5} = 54.8$$

**Class B Score:** 30, 45,55,75,90

$$Mean(\mu_B) = \frac{30+45+55+75+90}{5}$$

$$Mean(\mu_B) = \frac{295}{5} = 59$$

www.ilmkidunya.com

**Step 2: Compute the squared deviations of the scores for each class**

**Class A**

| Score (xi) | (xi−μ) | $(xi-\mu)^2$ |
|---|---|---|
| 50 | 50 − 54.8 = -4.8 | 23.04 |
| 52 | 52 − 54.8 = -2.8 | 7.84 |
| 55 | 55 − 54.8 = 0.2 | 0.04 |
| 57 | 57 − 54.8 = 2.2 | 4.84 |
| 60 | 60 − 54.8 = 5.2 | 27.04 |
| Total | | 62.8 |

**Class B**

| Score (xi) | (xi−μ) | $(xi-\mu)^2$ |
|---|---|---|
| 30 | 30 − 59 = -29 | 841 |
| 45 | 45 − 59 = -14 | 196 |
| 55 | 55 − 59 = -4 | 16 |
| 75 | 75 − 59 = 16 | 256 |
| 90 | 90 − 59 = 31 | 961 |
| Total | | 2270 |

**Step 3: Use the information from Step 2 and apply the formula below to compute variance for each class**

**Class A**

$$Variance(\sigma_A^2) = \frac{\sum_{i=1}^{N}(X_i-\mu_A)^2}{N} = \frac{62.8}{5} = 12.56$$

**Class B**

$$Variance(\sigma_B^2) = \frac{\sum_{i=1}^{N}(X_i-\mu_B)^2}{N} = \frac{2270}{5} = 454$$

**Calculating Standard Deviation:**

**Class A:** Variance = 12.56

$$Standard\ Deviation(\sigma_A) = \sqrt{Variance(\sigma_A^2)} = \sqrt{12.56} = 3.55$$

**Class B:** Variance = 456

$$Standard\ Deviation(\sigma_B) = \sqrt{Variance(\sigma_B^2)} = \sqrt{456} = 21.26$$

**Interpretation of the Results**

**Class A** has a standard deviation of 3.55. It is low and indicates that most of the scores are close to the average. The students in this class performed at a similar level and there is less variation among their scores.

**Class B** has a standard deviation of 21.26. It is much higher and indicates that the scores are widely spread out. Some students scored much lower or higher than the average. It means that there is greater variation in performance within Class B.

**Q. Describe probability with an example.**

**Probability**

**Probability** is the study of how likely an event is to occur. It helps to predict the outcomes based on the available information and known possibilities. Probability is used to estimate the chances of various outcomes in everyday life as follows:

- **Weather forecasting:** It can be used to estimate the chance of rain or sunshine.
- **Business:** It can be used to assess the risks to make informed decisions.
- **Sports:** It can be used to predict match outcomes and player performance.

Mathematically, the probability of an event $A$ happening is given by:

$$P(A) = \frac{Number\ of\ favorable\ outcomes\ to\ Event\ A}{Total\ number\ of\ possible\ outcomes}$$

This formula is used when all possible outcomes are equally likely. Equally likely outcomes are those that have the same chance of occurring. For example, the two possible outcomes are head and tail when flipping a fair coin. Both outcomes have an equal probability of occurring that shown as follows:

$$P(\text{Head}) = P(\text{Tail}) = \frac{1}{2} = 0.5 = 50\%$$

www.ilmkidunya.com

**Q. What is data collection? Discuss different data collection methods.**

### Data Collection

**Data collection** is the process of gathering relevant information for a particular purpose. **Data collection methods** are the techniques used to gather reliable data for analysis or research.

### Data Collection Methods

Different methods can be used for data collection depending on the nature of research. These methods include surveys, observations and experiments. Each method has its own strengths and suitability. The selection of the right method depends on the research objective and the type of data required.

### 1. Surveys

**Surveys** are commonly used method for collecting large amounts of data in a structured way. They involve asking a predefined set of questions to a selected group of people known as a **sample**. Surveys can be conducted using various means such as online forms, telephone calls, or face-to-face interviews.

### Example

Suppose a small local grocery store in Islamabad needs to know which products the customers want to see more frequently. A short survey can be created to ask the required questions from the customers. It can be distributed to 50 customers over the weekend. The collected responses are then analyzed to stock products according to customer demand. This helps to improve customer satisfaction and enhance business operations.

### Customer Preference Survey

A sample Customer Preference Survey is as follows:

1. Which product categories do you buy most often?
   ☐ Fruits ☐ Vegetables ☐ Dairy ☐ Snacks ☐ Other: _____
2. Are there any products you would like to see more often?
3. How often do you shop at this grocery store?
   ☐ Daily ☐ Weekly ☐ Monthly ☐ Occasionally
4. What influences your purchasing decisions the most?
   ☐ Price ☐ Quality ☐ Availability ☐ Brand ☐ Promotions
5. Any additional comments or suggestions?

### 2. Observations

**Observation** involves collecting data by watching or monitoring a situation in the natural environment. This method is useful when researchers want to gather data on behaviors or events without interference. It is very effective because it does not directly depend on other participants.

### Example

Suppose a restaurant needs to know which tables are most frequently chosen by customers during lunchtime. A staff member observes the seating choices over a period of one week. The restaurant can rearrange seating to improve the comfort and flow of customers based on the observation. It helps in improving customer satisfaction and service efficiency.

www.ilmkidunya.com

### 3. Experiments

**Experiments** involve manipulating one or more variables to determine their effect on another variable. This method is particularly useful in scientific and engineering fields where controlled environments are necessary for accurate and reliable measurement.

**Example**

A school teacher wants to know whether the student performance improves in the exam if printed notes are provided. The teacher conducts an experiment with two groups of students. One group receives printed notes and other group only rely on lectures. Both groups take a test after one month. The teacher compares the results to see if printed notes had a positive impact on performance.

**Q. What is data preparation? Explain with an example.**

**Data Preparation**

It is important to prepare the data for analysis after it has been collected. Data preparation includes the following:

- Cleaning the data to remove errors or inconsistencies
- Organizing the data in a meaningful way
- Converting data into a format that is suitable for analysis

If data is missing or incorrect, the researchers use techniques such as interpolation or statistical adjustments to ensure accuracy. **Interpolation** involves estimating missing values based on the pattern of nearby known values. **Statistical adjustments** involve using mathematical or statistical methods to correct errors or fill in gaps.

**Example**

Suppose the survey responses contain incomplete information. The missing values can be estimated based on the available data. For example, a student did not report his favorite subject in a survey. It can be estimated by looking at his performance in different subjects and assuming that the subject with the highest marks is his favorite. Proper data preparation ensures that the analysis leads to reliable and valid results.

**Q. Define data cleaning and transformation. Why are they important steps in preparing data for analysis.**

**Data Cleaning and Transformation**

**Data cleaning and transformation** are important steps to prepare data for analysis. Raw data often has errors, missing values or incorrect formats which can affect the accuracy of the results. It is important to fix these issues to ensure that the data is reliable and ready for analysis.

**Data Cleaning**

**Data cleaning** is the process of identifying and correcting any problems in the data. These problems can include incorrect entries, missing values or duplicate data. The results of the analysis will be inaccurate or misleading if these errors are not fixed.

**Example**

Suppose a college is collecting data on student scores. Some students may have entered their names incorrectly or some scores may be missing from the records. In this case, data cleaning would involve correcting any wrong names and finding the missing scores to complete the dataset.

www.ilmkidunya.com

The following table shows data cleaning process for student scores in a college. It shows common issues such as incorrect names, missing scores and duplicate entries.

| Name | Score | Class | Section |
|---|---|---|---|
| Ali | 91 | 11 | B |
| Ahmed | 85 | 11 | B |
| Fahim | | 11 | B |

**Figure:** Original data with errors

| Name | Score | Class | Section |
|---|---|---|---|
| Ali | 91 | 11 | B |
| Ahmed | 85 | 11 | B |
| Fahim | 92 | 11 | B |

**Figure:** Data after data cleaning

### Data Transformation

**Data transformation** is the process of converting data into a format that is easy to work with. It is done after cleaning the data. This may include converting data into different formats, creating new columns or organizing data in a different way. These changes make the data more suitable for analysis.

### Example

Suppose a college has a dataset with students' marks in three subjects: Math, English and Science. A new column "Total Score" can be created by adding the marks from all three subjects. Additionally, date formats can be standardized to ensure consistency. These transformations make the data easier to analyze for performance trends or comparisons.

**Q. How is missing data handled? Discuss different techniques to handle the missing data.**

### Handling Missing Data

Handling missing data is the process of dealing with incomplete or unavailable values in a dataset to ensure accurate and reliable analysis. Different techniques can be used to handle this problem. If only a small number of rows are affected, removing them may be a simple and effective solution. Alternatively, missing values can be filled in using methods such as calculating the average or using similar entries. The choice depends on the type of data and amount of missing information.

### Example

Suppose the dataset of student grades is missing the information about Ali. It creates a problem in assessing his performance. Different techniques to handle missing data are as follows:

### 1. Imputation

**Imputation** is the process of replacing missing data with estimated values to make a dataset complete and ready for analysis. For example, the college can calculate the average score of all students in Ali's class. The average score may be assigned to Ali temporarily. This approach allows the college to maintain a complete dataset while making a reasonable assumption about Ali's performance.

### 2. Flagging

**Flagging** is the process of marking specific data entries that meet certain conditions or need special attention. It is like adding a note to highlight important, unusual or problematic data. The college can keep track of Ali's missing score by adding a note in the dataset. This method indicates that Ali's score is not available. It ensures transparency while allowing the analysis to proceed without filling in the gap.

www.ilmkidunya.com

### 3. Removal

**Removal** is the process of deleting data that is incorrect or incomplete for the analysis. The college may choose to exclude Ali's record from specific analysis. This approach is useful if it does not significantly impact the overall understanding of student performance. However, it risks losing valuable information about Ali.

www.ilmkidunya.com

**Q. What is statistical modeling? Describe the steps involved in building a basic statistical model.**

### Statistical Modeling

**Statistical modeling** is a method of using data to understand patterns in the real world and make predictions about future events. It helps to make informed decisions based on past observations. Statistical models are widely used in areas like business, healthcare, economics and science to analyze data and support decision-making.

### Example

Statistical modeling can be used to estimate how much money will be spent on groceries next month. It can be calculated from the past grocery expenses. The past data can be analyzed to build a statistical model that can predict future spending more accurately. It helps individuals or families manage their budgets more effectively.

www.ilmkidunya.com

### Model Development Process

Building a statistical model involves several key steps designed to convert raw data into meaningful insights and reliable predictions. These steps are as follows:

### Step 1: Define the Problem

The first and most critical step is to clearly understand and define the problem to be solved. Suppose the objective is to predict monthly grocery expenses. It needs to identify the factors that influence those expenses such as family size, location, income and consumption habits etc.

### Step 2: Collect Relevant Data

The next step is to gather the necessary data after the problem is defined. In the grocery expense example, it involves collecting the data on past spending, family size, income levels and other relevant factors that may affect grocery costs.

### Step 3: Select an Appropriate Algorithm

A suitable algorithm needs to be selected to build the model. Algorithms are mathematical methods that detect patterns and relationships in data.

www.ilmkidunya.com

Some common examples include linear regression and logistic regression. The linear regression is used to predict numerical outcomes such as scores or prices whereas logistic regression is used to classify outcomes into categories such as yes/no or pass/fail. The best algorithm to use depends on the type of problem and the data.

### Step 4: Train the Model

This step is used to apply the selected algorithm to the training data. It allows the model to learn the underlying patterns and relationships. The model can understand how the input variables relate to the target outcome. For example, how the income or family size affect the grocery expenses.

**Step 5: Evaluate the Model's Performance**

Finally, the model is tested using new or unseen data to evaluate its accuracy and reliability. Different evaluation tools include prediction error and accuracy rate etc. They help to determine how well the model performs. This step ensures that the model makes valid predictions.

www.ilmkidunya.com

**Q. What is regression model? Give a detailed example where the regression model can be applied for making prediction.**

**Regression Model**

**Linear regression** is common statistical model that is used to understand the relationship between two variables. It predicts the value of one variable based on the known value of another variable.

- **Independent variable (x):** It is the variable used for prediction.
- **Dependent variable (y):** It is the variable to be predicted based on value of variable **x**.

**Example**

Suppose the user runs a small fruit stall in the town and needs to predict the daily earnings based on the number of customers who visit. In this scenario:

- The number of customers is the independent variable (X). It is also called **cause**.
- The daily earning is the dependent variable (Y). It is also called **effect**.

The linear regression can be applied to predict the earning as follows:

**Step 1: Collecting Data**

The first step is to collect relevant data. Suppose the user collected the number of customers and daily earnings for five consecutive days as follows:

| Day | Number of Customers (X) | Daily Earnings(Y) |
|---|---|---|
| 1 | 5 | 300 |
| 2 | 10 | 500 |
| 3 | 15 | 700 |
| 4 | 20 | 900 |
| 5 | 25 | 1100 |

**Step 2: Linear Regression Formula**

The formula for a simple linear regression model is as follows:

$Y = \beta_0 + \beta_1 X + \varepsilon$

Where:

- Y is the dependent variable that represents the daily earnings to be predicted.
- X is the independent variable that represents the number of customers.
- $\beta_0$ is the intercept that shows the earnings when there are zero customers.
- $\beta_1$ is the slope that shows increase in earnings with each additional customer.
- $\varepsilon$ is error term that shows difference between actual and predicted earnings.

**Step 3: Building the Linear Regression Model**

The slope and intercept must be calculated to build the model.

**Calculating the Slope $\beta_1$:**

The data shows that every increase of 5 customers results in an increase of Rs 200 in earnings:

$\beta_1 = \Delta Y / \Delta X = 200 / 5 = 40$

Where

- $\Delta Y$ = Change in the dependent variable (Y)
- $\Delta X$ = Change in the independent variable (X)
- $\Delta Y / \Delta X$ = Rate of change of Y with respect to X

The value of slope is 40 in the given scenario. It shows that every new customer will add Rs 40 to the earning.

**Calculating the Intercept $\beta_0$**

Using a data point where 10 customers lead to Rs 500 earnings. The customers can be represented by variable X and earning can be represented by variable Y.

$Y = \beta_0 + \beta_1 X$

$500 = \beta_0 + (40 \times 10)$

$\beta_0 = 500 - 400 = 100$

The intercept represents expected earnings even with zero customers. This amount may reflect fixed income sources such as standing orders.

**Regression Equation**

Earnings = $\beta_0 + \beta_1 \times$ Customers

Earnings = $100 + 40 \times$ Customers

This equation indicates a fixed amount of Rs 100 earned each day even if no customers visit. Additionally, Rs 40 is earned for every new customer.

**Step 4: Using the Model for Prediction**

The model can be used to make the predictions about future earnings based on the expected number of customers. Suppose the number of expected customers for tomorrow is 22. The predicted earnings is as follows:

Earnings = $100 + 40 \times 22 = 980$

It shows that the expected earning with 22 customers is Rs 980 approximately.

**Step 5: Testing the Model**

The new data is used to test the accuracy of the model. Suppose 28 customers visited on day 6 and the actual earning was Rs 1250.

The predicted earnings using the regression model from 28 customers will be:

Predicted Earnings = $100 + 40 \times 28 = 100 + 1120 =$ Rs 1220

The difference between the predicted earnings and the actual earning is known as the **prediction error**. It is calculated as follows:

Error = Predicted Earnings - Actual Earnings = 1220 - 1250 = Rs -30

The prediction of the model is close but it is not perfect. It shows that real-world often has some variation from predicted outcome.

**Q. What is logistic regression? Explain with example.**

**Logistic Regression**

**Logistic regression** is a powerful statistical tool predict outcomes that fall into two categories such as yes/no or pass/fail. It provides a probability value between 0 and 1. A value close to 1 means the event is more likely to happen. A value close to 0 means the event is less likely to happen.

www.ilmkidunya.com

Logistic regression is different from linear regression because it does not predict exact numerical values like scores or prices.

**Example**

Logistic regression can be used to predict whether a student will pass or fail an exam based on the number of study hours. The logistic regression helps to find the probability of passing instead of giving a specific score. For example, if the model gives a probability of 0.85, it means the student has 85% chance of passing the exam.

**Q. What are clustering techniques in data analysis? Explain the K-means clustering with an example.**

**Clustering**

**Clustering** is a technique of grouping similar items together based on their characteristics or features. For example, a teacher can use clustering to divide the students into groups based on Math and English scores. It helps the teacher to identify students with similar strengths or weaknesses. The teacher can design academic support accordingly to each group's needs.

**K-means Clustering**

**K-means clustering** is one of the most commonly used techniques to group data into clusters based on similarities. The process begins with deciding on the number of clusters (K) to form. It then then groups data points into these clusters by calculating the distance between them. A smaller distance indicates greater similarity. The letter **K** in K-means represents the number of clusters to be created.

**Example**

Suppose a teacher has the Math and English scores of following five students.

| Student | Math Score | English Score |
|---|---|---|
| Basim | 85 | 70 |
| Umer | 90 | 65 |
| Annie | 50 | 80 |
| Tallat | 40 | 85 |
| Maliha | 60 | 60 |

The teacher wants to use clustering technique to group students based on their performance in two subjects. The teacher divides the students into two clusters:

- Cluster 1 for students strong in math
- Cluster 2 for students strong in English

The K-means algorithm groups students based on their performance. It calculates the distance between their scores, identifies patterns and assigns them to the most suitable cluster. For example, the students Basim and Umer performed well in math and they are placed in one cluster. The students Annie and Tallat are good in English and they are placed in another cluster. This method helps the teachers to understand the pattern of student performance and offer targeted learning support to various groups.

www.ilmkidunya.com

**Q. What is meant by model evaluation? Describe different performance metrics used for model evaluation.**

## Model Evaluation

**Model evaluation** is the process of checking the performance of a model after it has been built. It helps to understand the accuracy and usefulness of the predictions made by the model. It is also important to measure its performance and make required improvements if needed. It ensures that the model is working correctly and reliably.

### 1. Performance Metrics

**Performance metrics** are used to measure the performance of a model. They help to understand whether the model's predictions are correct or close to the actual results. Two common types of metrics are as follows:

#### i. Error Metrics

**Error metrics** measure the difference between the model's prediction and the actual result. Smaller errors mean model is more accurate. Suppose a model predicts a grocery bill of Rs 8,000 but the actual bill is Rs 10,000. The difference is called the error.

#### ii. Accuracy Metrics

**Accuracy metrics** are used to know the number of correct predictions of the model. They are useful in classification problems like pass/fail or yes/no predictions. Suppose a model predicts whether a student will pass or fail. The accuracy shows how often the prediction was correct.

### 2. Interpreting Outputs

**Interpreting outputs** means understanding the actual meaning of model's results. It involves analyzing the output to draw conclusions, identify patterns or extract useful insights. This step helps in making informed decisions based on what the model reveals.

#### Example

Suppose a linear regression model shows that more hours studied lead to higher exam scores. The conclusion can be drawn that increasing study time helps students improve the scores of the students.

### 3. Ethical Considerations

**Ethical considerations** ensure the model is fair and respects privacy. The model should not harm people or treat them unfairly.

#### i. Fairness and Bias

A model should be fair and free from bias. It should ensure that all individuals or groups are treated equally. The outcomes become unfair when a model favors one group over another without valid reasons. For example, a model is biased if it approves loans for one group of people while unfairly rejecting others. Such models are considered unethical and must be corrected.

#### ii. Data Privacy

**Data privacy** means protecting personal information used in a model. It includes keeping data secure and not sharing it without permission. Respecting privacy builds user trust. For example, a company is using customer data to build models. It must ensure the data is secure and not shared without permission.

www.ilmkidunya.com

**Q. What is data visualization? Discuss the types of data visualization and their uses.**

## Data Visualization

**Data visualization** is the process of representing data in a visual format such as graphs or charts. It makes it easier to see and understand trends and patterns in the data. It is especially useful when analyzing large amount of data and make decisions in businesses and research etc. For example, a business can use data visualization to identify the products that are selling fast. It can help them to take decisions such as buying more stock of that product.



**Figure:** Data visualization

## Types of Visualizations

**Data visualization** is a powerful method for understanding complex information. Different types of visualizations serve different purposes. They make it easier to interpret, compare and analyze data more effectively.

### 1. Bar chart

A **bar chart** is used to compare different categories or groups. It displays rectangular bars that represent the values of different categories. The length or height of each bar indicates the value.

### Example

A bar chart can be used to compare the sales of different products in a store. This helps in understanding product performance.

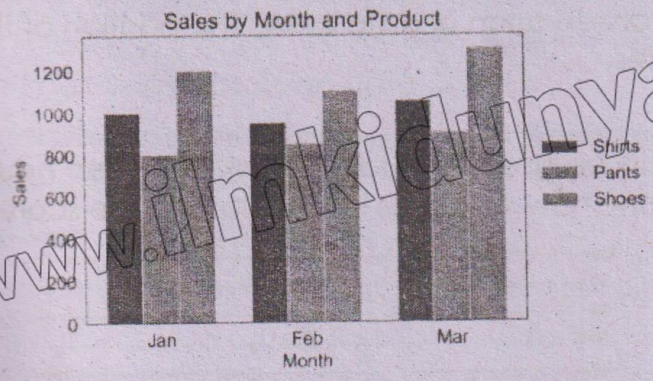



**Figure:** A bar chart showing sales of different products

### 2. Line Chart

A **line graph** is used to show trends or changes over time. It plots data points and connects them with lines to show how values increase or decrease. Each data point represents a specific value at a particular time or category.

www.ilmkidunya.com

**Example**

A line graph can show temperature changes over a week. It helps to visualize when the temperature went up or down each day. This makes it easier to understand daily trends.

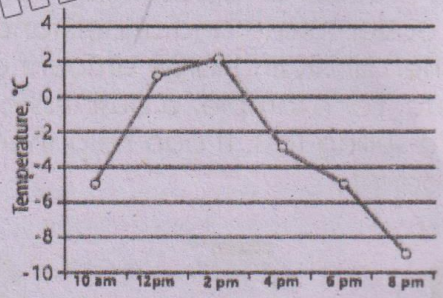



**Figure:** A line graph showing variation of temperature over time

### 3. Histograms

A **histogram** is used to show the distribution of data. It groups data into ranges or intervals called **bins**. It uses bars to display how frequently the values occur within each range.

**Example**

A histogram can be used to analyze the performance of students in a math exam. It can show the distribution of their scores that makes it easier to see which score ranges were most common.

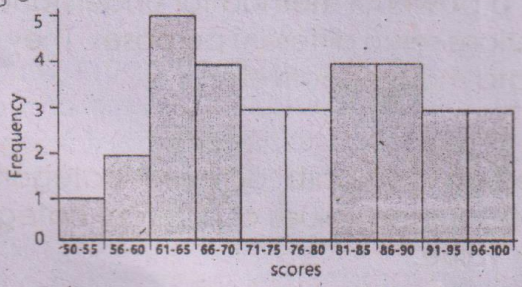



**Figure :** Histogram showing the distribution of exam scores

### 4. Scatterplots

**Scatterplots** show the relationship between two variables. Each dot or point on the graph shows one observation or data entry. The position of the dot is based on the values of both variables.

**Example**

A scatterplot can be used to study the link between hours of study and exam scores. Each point shows how student's hours relate to the score.

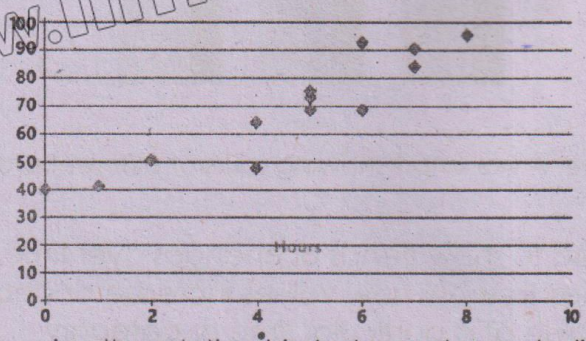



**Figure:** Scatterplot showing the relationship between hours studied and exam scores

www.ilmkidunya.com

### 5. Boxplot

A **boxplot** or **box-and-whisker plot** is a type of chart that shows the distribution and spread of data. It helps in identifying the median, quartiles and outliers. **Median** is the middle value of the data. **Quartiles** divide the data into four equal parts and the box represents the middle 50% of the data. **Outliers** are the unusual values that are much higher or lower than most of the data. It is shown as small circles or dots.

**Example**

A boxplot can be used to compare exam scores of different classes. It helps to see which class has higher or more consistent scores. It also shows if there were any unusual high or low scores.

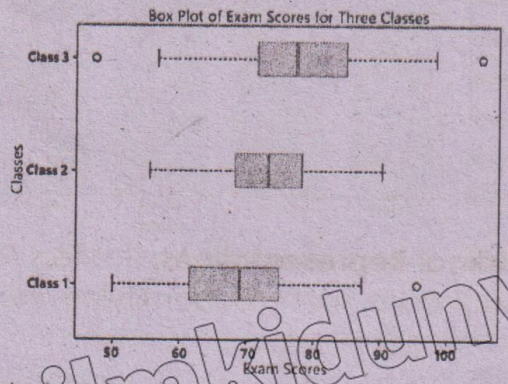



**Figure :** A Boxplot, showing class scores performance of three classes

**Q. What are different tools for data visualization? Give an example.**

**Tools for Data Visualization**

**Data visualization tools** are used to convert raw data into visual formats such as charts and graphs. It helps the users to understand and analyze data easily. There are many tools available for data visualization such as **Microsoft Excel**, **Google Sheets**, **Python (Matplotlib)** and **Tableau**. These tools support the creation of different types of charts and graphs such as bar charts, pie charts and line graphs.

**Example**

Suppose a small business wants to track the number of products sold each month. The monthly sales data can be entered in Excel or Google Sheets. A bar chart can be created with a few steps easily. It can show which month had the most sales. It helps to make better decisions based on data.

**Q. How can you create and interpret data visualizations using tools like Excel or Google Sheets?**

**Creating and Interpreting Visualizations**

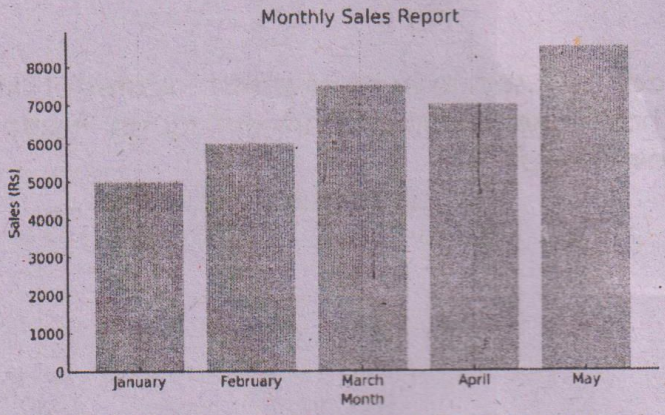
A simple guide to create a visualization in Excel or Google Sheets is as follows:

1. **Enter Data:** The first step is to enter data into the spreadsheet. For example, one column can have months and another column can have sales numbers.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | **Month** | **Sales (Rs)** | | |
| 2 | January | 5000 | | |
| 3 | February | 6000 | | |
| 4 | March | 7500 | | |
| 5 | April | 7000 | | |
| 6 | May | 8500 | | |

www.ilmkidunya.com

2. **Select Data:** The data can be selected by clicking and dragging mouse over the desired cells in the spreadsheet. This highlights the values to be visualized.
3. **Choose a Chart Type:** After selecting the data, go to **Insert** tab and choose the type of chart to be created such as a bar chart or a pie chart.
4. **Customize the Chart:** Customizing the chart helps add clarity. For example, the user can add axis labels, titles and colors to make the chart easier to read and interpret.




5. **Understanding Statistical Representations:** It refers to understanding the actual meaning of the visual data such as identifying the highest values or noticing changes over time.

www.ilmkidunya.com

## Exercise Solution

### Multiple-Choice Questions (MCQs)

1. **An example of a basic statistical model:**
   a) Linear Regression
   b) Neural Networks
   c) Decision Trees
   d) Support Vector Machines
2. **The activity involved in experimental design in data science:**
   a) Creating visualizations
   b) Collecting and analyzing data systematically
   c) Writing code for machine learning
   d) Building databases
3. **A commonly used tool for creating data visualizations:**
   a) MS Excel
   b) Python (Matplotlib)
   c) Tableau
   d) All of the above
4. **The meaning of the slope in a linear regression model:**
   a) The intercept of the model
   b) The change in the dependent variable for a unit change in the independent variable
   c) The error term
   d) The mean of the data
5. **An example of a real-world application of statistical models:**
   a) Predicting house prices
   b) Creating social media posts
   c) Designing websites
   d) Writing essays
6. **Option not considered a benefit of data visualization:**
   a) Identifying trends and patterns
   b) Communicating insights effectively
   c) Making data more complex
   d) Summarizing large datasets

7. **A primary goal of K-Means clustering:**
   a) To classify data into predefined categories
   b) To group data into clusters based on similarity
   c) To predict continuous outcomes
   d) To reduce the dimensionality of data
8. **The meaning of "K" in K-Means clustering:**
   a) Number of features in the dataset
   b) Number of clusters to be formed
   c) Number of iterations required for convergence
   d) Number of data points in the dataset

**Answers**

| | | | | | |
|---|---|---|---|---|---|
| 1. a | 2. b | 3. d | 4. b | 5. a | 6. b |
| 7. b | 8. b | | | | |

## Short Questions

**Q.1. What is the importance of building statistical models in real-world applications?**

Building statistical models is important in real-world applications because they help to analyze data, identify patterns and make accurate predictions. These models support informed decision-making in fields such as business, healthcare and education. They also assist to measure risk, improve efficiency and allocate resources effectively.

**Q.2. Name one basic statistical model used for predicting outcomes and explain its purpose.**

One basic statistical model used for predicting outcomes is linear regression. It helps to understand the relationship between two variables called the independent variable and the dependent variable. The model is used to predict the value of the dependent variable based on the value of the independent variable. For example, it can be used to predict sales based on the amount spent on advertising.

**Q.3. List two types of data visualizations and describe when you would use each.**

Two types of data visualizations include bar chart and line chart. A bar chart is used to compare different categories or groups. It can be used to compare the sales of different products in a store. This helps in understanding product performance. A line graph is used to show trends or changes over time. It can be used to show temperature changes over a week.

**Q.4. How does visualizing data help in understanding descriptive statistics?**

Visualizing data helps in understanding descriptive statistics by making patterns, trends and distributions easier to understand. The charts such as histograms, boxplots and bar graphs help users to quickly identify measures such as mean, median and range. Visuals simplify complex data and highlight outliers or unusual patterns.

## Long Questions

**Q.1. Explain the role and importance of statistical models in solving real-world problems.**

Statistical models are powerful tools for solving real-world problems. They play important role in understanding and solving real-world problems by analyzing data, identifying patterns and making accurate predictions. These models help individuals, organizations and governments to make better decisions in various fields.

**1. Understanding Relationships in Data**

Statistical models help to identify and understand the relationships between different variables. For example, they can show how study time affects student performance or how prices affect customer demand. These relationships are essential for drawing meaningful conclusions from data.

www.ilmkidunya.com

**2. Making Accurate Predictions**

A major advantage of statistical modeling is the ability to predict future outcomes. They use historical data to forecast sales, weather conditions, population growth or disease spread. These predictions are critical for planning, forecasting and resource management.

**3. Supporting Evidence-Based Decisions**

Statistical models provide a foundation for making decisions based on data rather than assumptions. Businesses, healthcare systems and governments use these models to create strategies, allocate resources and solve problems efficiently.

**4. Measuring Risk and Uncertainty**

Many real-world situations involve uncertainty. Statistical models help to measure this uncertainty and assess risk. For example, the models in finance can estimate the possibility of market fluctuations or loan defaults that help in better risk management.

**5. Improving Operational Efficiency**

Statistical models also improve efficiency in daily operations. In agriculture, they can recommend the best time to plant crops. In manufacturing, they can help to detect defects early and improve production quality.

**Q.2. Describe the steps involved in building a basic statistical model (e.g., linear regression). Include details on data collection, model training, and evaluation.** (See chapter)

**Q.3. Discuss the types of data visualizations and their uses.** (See chapter)

**Q.4. Explain data collection methods.** (See chapter)

**Q.5. Discuss the concept of measure of tendency with example.** (See chapter)

## SHORT QUESTIONS

**Q.1. What is statistics and why is it important?**

Statistics is a branch of mathematics that is used to understand, analyze and interpret data. Statistics is important because it can summarize large sets of data in a simple way that makes it easier to draw useful conclusions and make decisions. It is widely used in various fields such as education, business, healthcare and economics etc.

**Q.2. What are measures of central tendency?**

Measures of central tendency are statistical tools used to find the central or typical value in a dataset. The three main measures are mean, median and mode.

**Q.3. Why are measures of central tendency important?**

The measures of central tendency are important as they help to summarize a large dataset into a single representative value. This makes it easier to understand the overall pattern or trend in the data. These measures are widely used in reports and analysis.

**Q.4. How do the three measures of central tendency differ?**

Mean is the average of all the values in a data set. Median is the middle value when the numbers are arranged in order. The mode is the value that occurs most frequently in the dataset.

**Q.5. What is mean and how is it calculated?**

Mean is the average of all the values in a data set. It is calculated by adding all the data values and then dividing by the number of values.

**Q.6. What is the median and how is it calculated?**

Median is the middle value when the numbers are arranged in order. The median is the exact middle value if there is an odd number of values. The median is the average of the two middle values if there is an even number of values.

**Q.7. Can a data set have more than one median?**

No, a data set has only one median. Even if the number of values is even, the median is the average of the two middle numbers.

www.ilmkidunya.com

**Q.8. What is the mode in statistics?**

Mode is the value that appears most frequently in a dataset. It helps to identify the most common or repeated value in a dataset. There can be more than one mode if multiple values appear with the same highest frequency.

**Q.9. Give an example of finding the mode in a dataset.**

Suppose the scores of five students are 50, 60, 70, 70 and 90. The number 70 appears twice but all other numbers appear only once. The mode is 70 as it is the most frequent value.

**Q.10. Can a dataset have multiple modes?**

Yes, there can be more than one mode if multiple values appear with the same highest frequency.

**Q.11. Why are measures of dispersion important in statistics?**

Measures of dispersion are important in statistics because they describe how spread out or scattered the data values are in a dataset. They help in understanding the degree of variation or inconsistency among the data points. They indicate how much individual values differ from the average. These measures are essential for comparing datasets and assessing their reliability.

**Q.12. What is variance?**

Variance is a statistical measure that shows how much the values in a dataset differ from the mean. It is calculated by finding the average of the squared differences between each value and the mean. A higher variance indicates that the data points are more spread out while a lower variance shows that they are closer to the mean.

**Q.13. What does a high variance indicate?**

A high variance indicates that the data points are widely spread out from the mean. This means the values are less consistent and show more variability.

**Q.14. What does a low variance indicate?**

A low variance indicates that most data values are close to the mean. It suggests that the data is consistent and shows little variability.

**Q.15. How is variance calculated?**

The variance is calculated by subtracting the mean from each value, squaring the result and taking the average of these squared differences. The formula is:

$$Variance(\sigma^2) = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

**Q.16. What is standard deviation in statistics?**

Standard deviation is a statistical measure that indicates how much the values in a dataset deviate from the mean (average). It is calculated as the square root of the variance. A small standard deviation means that most values are close to the mean, indicating low variability. A large standard deviation means that values are spread far from the mean, indicating high variability.

**Q.17. How do you calculate standard deviation?**

Standard deviation is calculated as square root of the variance. The formula is as follows:

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

**Q.18. Why is standard deviation preferred over variance in interpretation?**

Standard deviation is preferred over variance because it is expressed in the same units as the original data, making it easier to interpret. In contrast, variance is in squared units, which can be harder to understand.

**Q.19. What is probability?**

Probability is the study of how likely an event is to occur. It helps to predict the outcomes based on the available information and known possibilities. Probability is used to estimate the chances of various outcomes in everyday life such as weather forecasting and business etc.

**Q.20. How is probability calculated?**

Probability is calculated using the following formula:

$$\text{Probability} = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

www.ilmkidunya.com

**Q.21. What is the probability of getting heads when flipping a fair coin?**

The two possible outcomes are head and tail when flipping a fair coin. Both outcomes have an equal probability of occurring: P(head) = P(tail) = ½ = 0.5 = 50%.

**Q.22. What is the purpose of data collection in research?**

Data collection helps gather relevant information to answer research questions or solve problems. It ensures the information used in the study is accurate, complete and useful. Good data collection leads to reliable results.

**Q.23. Why is data preparation important before analysis?**

Data preparation ensures the collected data is clean, organized and in the correct format. It removes errors and inconsistencies that can affect accuracy of analysis. This step is crucial for obtaining meaningful insights.

**Q.24. Write common methods of data collection.**

Common methods include surveys, observations and experiments. Each method is chosen based on the research goals and the type of data needed. These methods help gather data systematically.

**Q.25. What are surveys?**

Surveys are commonly used method for collecting large amounts of data. They involve asking a predefined set of questions to a selected group of people known as a sample. Surveys can be conducted using various means such as online forms, telephone calls, or face-to-face interviews.

**Q.26. What is meant by data cleaning?**

Data cleaning is the process of identifying and correcting any problems in the data. These problems can include incorrect entries, missing values or duplicate data. The results of the analysis will be inaccurate or misleading if these errors are not fixed.

**Q.27. What is meant by data transformation?**

Data transformation is the process of converting data into a format that is easy to work with. It is done after cleaning the data. This may include converting data into different formats, creating new columns or organizing data in a different way. These changes help make the data more suitable for analysis or modeling.

**Q.28. Why data cleaning and transformation important?**

Data cleaning and transformation are important steps to prepare data for analysis. Raw data often has errors, missing values or incorrect formats which can affect the accuracy of the results. It is important to fix these issues to ensure that the data is reliable and ready for analysis.

**Q.29. What is imputation in handling missing data?**

Imputation is the process of replacing missing data with estimated values to make a dataset complete and ready for analysis.

**Q.30. What is flagging and how does it work in handling missing data?**

Flagging is the process of marking specific data entries that meet certain conditions or need special attention. It is like adding a note to highlight important, unusual or problematic data. It ensures transparency while allowing the analysis to proceed without filling in the gap.

**Q.31. Write the steps involved in building a statistical model.**

The steps involved in building a statistical model include defining the problem, collecting data, choosing an algorithm, training the model, and evaluating the model.

**Q.32. What is meant by training a model?**

Training a model meants to apply the selected algorithm to the training data. It allows the model to learn the underlying patterns and relationships. The model can understand how the input variables relate to the target outcome.

**Q.33. Why is evaluating a statistical model important?**

Evaluating a statistical model is important to esnure its accuracy and reliability. Different evaluation tools include prediction error and accuracy rate etc. They help to determine how well the model performs.

www.ilmkidunya.com

**Q.34. How does linear regression work?**

Linear regression works by finding the best line that explains the relationship between two variables. The equation for simple linear regression is $Y = \beta_0 + \beta_1 X + \varepsilon$, where Y is the dependent variable, X is the independent variable, $\beta_0$ is the intercept, $\beta_1$ is the slope, and $\varepsilon$ is the error term.

**Q.35. What is the purpose of the intercept in linear regression?**

The intercept ($\beta_0$) represents the starting value of the dependent variable when the independent variable is zero. It provides a baseline value for the dependent variable.

**Q.36. How do you interpret the slope in linear regression?**

The slope ($\beta_1$) shows how much the dependent variable changes with each unit increase in the independent variable. It represents the rate of change between the two variables.

**Q.37. What is logistic regression?**

Logistic regression is a powerful statistical tool predict outcomes that fall into two categories such as yes/no. It provides a probability value between 0 and 1. A value close to 1 means event is more likely to happen. A value close to 0 means event is less likely to happen.

**Q.38. How is logistic regression different from linear regression?**

Linear regression is used to predict continuous numerical values such as temperature or sales figures. Logistic regression predicts the probability of a categorical outcome such as whether an event will occur or not.

**Q.39. What is clustering? Give an example.**

Clustering is a technique of grouping similar items together based on their characteristics or features. For example, a teacher can use clustering to divide the students into groups based on Math and English scores. It helps to identify students with similar strengths or weaknesses.

**Q.40. What is K-means clustering?**

K-means clustering is one of the most commonly used techniques to group data into clusters based on similarities. The process begins with deciding on the number of clusters (K) to form. It then then groups data points into these clusters by calculating the distance between them. A smaller distance indicates greater similarity.

**Q.41. What are performance metrics in model evaluation?**

Performance metrics measure how well a model works. Common metrics include error metrics and accuracy metrics. They help determine if the model gives reliable predictions.

**Q.42. What are accuracy metrics in modeling?**

Accuracy metrics are used to know the number of correct predictions of the model. They are useful in classification problems like pass/fail or yes/no predictions. Suppose a model predicts whether a student will pass or fail. The accuracy shows how often the prediction was correct.

**Q.43. How the outcomes of a model can be unfair? Give an example.**

The outcomes of a model can be unfair if it favors one group over another. For example, a model is biased if it approves loans for one group of people while rejecting others unfairly. Such models are considered unethical and must be corrected.

**Q.44. What is data visualization?**

Data visualization is the process of representing data in a visual format such as graphs or charts. It makes it easier to see and understand trends and patterns in the data. It is especially useful when analyzing large amount of data and make decisions in businesses and research etc.

**Q.45. What is bar chart?**

A bar chart is used to compare different categories or groups. It displays rectangular bars that represent values of different categories. The length or height of each bar indicates a value.

**Q.46. Give an example of when to use a bar chart.**

A bar chart can be used to compare the sales of different products in a store. This helps in understanding product performance.

**Q.47. What is a line graph?**

A line graph is used to show trends or changes over time. It plots data points and connects them with lines to show how values increase or decrease. Each data point represents a specific value at a particular time or category.

www.ilmkidunya.com

**Q.48. Give an example of when to use a line graph.**

A line graph can show temperature changes over a week. It helps to visualize when the temperature went up or down each day. This makes it easier to understand daily trends.

**Q.49. What is a histogram?**

A histogram is used to show the distribution of data. It groups data into ranges or intervals called bins. It uses bars to display how frequently the values occur within each range.

**Q.50. What is a scatterplot?**

Scatterplots show the relationship between two variables. Each dot or point on the graph shows one observation or data entry. The position of the dot is based on the values of both variables.

**Q.51. Give an example of using a scatterplot.**

A scatterplot can be used to study the link between hours of study and exam scores. Each point shows how student's hours relate to the score.

**Q.52. What is a boxplot?**

A boxplot is a type of chart that shows the distribution and spread of data. It helps in identifying the median, quartiles and outliers.

**Q.53. What is median, quartiles and outliers in boxplot?**

Median is the middle value of the data. Quartiles divide the data into four equal parts and the box represents the middle 50% of the data. Outliers are the unusual values that are much higher or lower than most of the data. It is shown as small circles or dots.

**Q.54. When is a boxplot useful?**

A boxplot is useful for comparing performance between different groups. For example, it can be used to compare exam scores of different classes to see which performed better overall.

**Q.55. What is the purpose of using data visualization tools?**

Data visualization tools are used to turn raw data into visual formats such as charts and graphs to understand and analyze data easily.

**Q.56. Name some tools used for data visualization.**

Some tools available for data visualization are Microsoft Excel, Google Sheets, Python (Matplotlib) and Tableau. These tools support the creation of different types of charts and graphs such as bar charts, pie charts and line graphs.

**Q.57. What types of charts can you create in Excel or Google Sheets?**

You can create different types of charts in Excel or Google Sheets such as bar charts, line graphs, pie charts etc. These charts help visualize data clearly and are useful for comparing, analyzing and identifying trends.

**Q.58. What is Matplotlib in Python used for?**

Matplotlib is a powerful data visualization library in Python. It is used to create a wide range of visualizations including line graphs, histograms, and scatter plots.

**Q.59. What is the use of Tableau?**

Tableau is a professional data visualization tool used to create interactive and detailed visual dashboards. It is widely used in business analytics and reporting.

www.ilmkidunya.com

## Multiple Choice Questions

1. **The primary purpose of _____ is to examine data to find useful information, patterns or trends.**
   a. Data analytics b. Data encryption
   c. Data warehousing d. Data modeling
2. **Which branch of mathematics helps in understanding and analyzing data?**
   a. Statistics b. Algebra c. Geometry d. Calculus

3. **Which statistical measure represents the "center" or typical value in a dataset?**
   a. Measure of central tendency b. Measure of skewness
   c. Measure of kurtosis d. Measure of variability
4. **Which of the following is NOT a measure of central tendency?**
   a. Mean b. Median c. Mode d. Range
5. **Which measure of central tendency is calculated by adding all numbers and dividing by the total count?**
   a. Mean b. Median c. Mode d. Range
6. **What is the mean of the numbers 50, 60, 70, 80 and 90?**
   a. 65 b. 70 c. 75 d. 85
7. **If the total of five numbers is 250, what is their mean?**
   a. 45 b. 50 c. 55 d. 60
8. **The value that appears most often in a dataset is called:**
   a. Mean b. Median c. Mode d. Range
9. **The mode of 2, 3, 4, 3, 5, 3, 6 is:**
   a. 2 b. 3 c. 4 d. 5
10. **What is the mode of the data set: 50, 60, 70, 70, 60 and 90?**
    a. 50 b. 60 and 70 c. 70 and 90 d. 50 and 60
11. **The middle value when data is arranged in order is called:**
    a. Mean b. Mode c. Median d. Range
12. **The median of 50, 60, 70, 80 and 90 is:**
    a. 30 b. 50 c. 70 d. 90
13. **When a data set is arranged in order and has an even number of values, how is the median calculated?**
    a. Smallest value b. Largest value
    c. Most frequent value d. Average of two middle values
14. **What is the median of the numbers: 50, 60, 70, 80?**
    a. 60 b. 70 c. 65 d. 75
15. **Which of the following measures requires the data to be arranged in order?**
    a. Mean b. Median c. Mode d. Range
16. **Which of the following is a measure of dispersion?**
    a. Mean b. Median c. Mode d. Variance
17. **Which measure of dispersion shows how much individual data points differ from mean?**
    a. Variance b. Median c. Mode d. Range
18. **What is the formula for calculating variance?**
    a. $\sigma^2 = (\Sigma(xi - \mu)^2) / N$ b. $\sigma^2 = (\Sigma(xi + \mu)^2) / N$
    c. $\sigma^2 = (\Sigma(xi - \mu)) / N$ d. $\sigma^2 = (\Sigma(xi + \mu)) / N$
19. **What does it indicate when the data points are spread out from the mean?**
    a. Low variance b. High variance
    c. Normal distribution d. Skewed distribution
20. **A lower variance in a dataset indicates that the data points are:**
    a. Spread far apart b. Far from the mean
    c. Closer to the mean d. Increasing rapidly
21. **What is the first mathematical operation needed when calculating variance?**
    a. Square each value b. Find the median
    c. Compute the mean d. Sort the values
22. **Which of the following is NOT required for calculating variance?**
    a. Mean b. Total number of values
    c. Mode d. Squared deviations
23. **Which of the following is calculated as the square root of variance?**
    a. Mean b. Median
    c. Standard deviation d. Range
24. **What is the mathematical study of how likely an event is to happen called?**
    a. Statistics b. Probability c. Calculus d. Algebra

www.ilmkidunya.com

25. **What is the probability of getting heads when flipping a fair coin?**
    a. 25% b. 100% c. 50% d. 75%
26. **What is the probability of getting tails when flipping a fair coin?**
    a. 25% b. 100% c. 50% d. 75%
27. **In a single flip of a fair coin, how many possible outcomes count as "heads"?**
    a. 0 b. 1 c. 2 d. 3
28. **What is the total number of outcomes when flipping a fair coin?**
    a. 1 b. 2 c. 3 d. 4
29. **What is the formula for calculating probability?**
    a. Total outcomes / favorable outcomes
    b. Favorable outcomes / total outcomes
    c. Total outcomes × favorable outcomes
    d. Favorable outcomes + total outcomes
30. **Which of the following is NOT a method of data collection?**
    a. Survey b. Observation
    c. Experiment d. Prediction
31. **Which data collection method involves watching and recording behavior in a natural setting without interference?**
    a. Survey b. Observation
    c. Experiment d. Interview
32. **Which method is most efficient for collecting data from a large number of people?**
    a. Survey b. Observation
    c. Experiment d. Case study
33. **Surveys typically use:**
    a. Predefined questions b. Unscripted conversations
    c. Live participant monitoring d. Case study
34. **Which method would a teacher use to test if printed notes improve exam scores?**
    a. Survey b. Observation
    c. Experiment d. Interview
35. **Which of the following is NOT typically fixed during data cleaning?**
    a. Duplicate records b. Incorrect entries
    c. Software bugs d. Missing values
36. **The process of estimating missing values using existing data is called:**
    a. Forecasting b. Imputation
    c. Summarization d. Classification
37. **When is data transformation usually performed?**
    a. Before data collection b After data analysis
    c. Before data cleaning d. After data cleaning
38. **Which task is NOT part of data transformation?**
    a. Rearranging data b. Creating new columns
    c. Estimating missing values d. Changing data formats
39. **_____involves analyzing data to understand patterns and make predictions about future events or trends?**
    a. Statistical modeling b Data cleaning
    c. Data warehousing d. Data validation
40. **What is the first step in building a statistical model?**
    a. Train the model b. Collect data
    c. Define the problem d. Choose an algorithm
41. **What is the second step in statistical model development?**
    a. Collect data b. Train the model
    c. Interpret results d. Evaluate the Model
42. **Which statistical technique is commonly used to predict one variable based on another?**
    a. Linear Regression b. Logistic Regression
    c. Decision Trees d. Clustering

www.ilmkidunya.com

43. **In linear regression, the independent variable is also known as?**
    a. Dependent variable  b. Predictor variable
    c. Outcome variable  d. Response variable
44. **In linear regression, the variable which is used for prediction is called the:**
    a. Dependent variable  b. Independent variable
    c. Intervening variable  d. Response variable
45. **In linear regression, the variable that is being predicted is called the:**
    a. Dependent variable  b. Independent variable
    c. Intervening variable  d. Predictor variable
46. **Which equation correctly represents a simple linear regression model?**
    a. $Y = \beta0 + \beta1X + \varepsilon$  b. $Y = \beta0 - \beta1X + \varepsilon$
    c. $Y = \beta0 + \beta1X - \varepsilon$  d. $Y = \beta0 - \beta1X - \varepsilon$
47. **In the simple linear regression equation $Y = \beta_0 + \beta_1X + \varepsilon$, what does Y represent?**
    a. Independent variable  b. Dependent variable
    c. Error term  d. Slope
48. **In the simple linear regression equation $Y = \beta_0 + \beta_1X + \varepsilon$, what does X represent?**
    a. Independent variable  b. Dependent variable
    c. Error term  d. Slope
49. **What is the term for the difference between actual value and the predicted value in a regression model?**
    a. Slope ($\beta_1$)  b. Intercept ($\beta_0$)
    c. Error term ($\varepsilon$)  d. Independent variable (X)
50. **Which component of the linear regression equation $Y = \beta_0 + \beta_1X + \varepsilon$ determines the direction and rate of change of Y with respect to X?**
    a. Slope ($\beta_1$)  b. Intercept ($\beta_0$)
    c. Error term ($\varepsilon$)  d. Independent variable (X)
51. **Which statistical method is specifically designed for predicting binary categorical outcomes such as yes/no?**
    a. Linear regression  b. Decision trees
    c. Logistic regression  d. Polynomial regression
52. **What is the range of possible output values in logistic regression?**
    a. 0 to 100  b. 0 to 1  c. -1 to 1  d. $-\infty$ to $+\infty$
53. **_____ is a technique that groups similar data points based on their characteristics.**
    a. Clustering  b. Regression
    c. Classification  d. Outlier detection
54. **Which of the following is a clustering technique?**
    a. Linear Regression  b. Logistic Regression
    c. K-means  d. Classification
55. **_____ is the process of checking how well a model performs by comparing its predictions to the actual outcomes.**
    a. Model Training  b. Model Evaluation
    c. Model Deployment  d. Model tuning
56. **_____ is the process of representing data in a visual format such as graphs or charts.**
    a. Data visualization  b. Graphic design
    c. Infographics  d. Data mining
57. **Which of the following is NOT a type of data visualization?**
    a. Line graph  b. Bar chart
    c. Scatter plot  d. Spreadsheet
58. **What type of visualization is ideal for comparing different categories?**
    a. Line graph  b. Bar chart  c. Scatter plot  d. Pie chart
59. **What type of visualization is used to show trends over time?**
    a. Line graph  b. Histogram  c. Scatter plot  d. Bar chart
60. **What type of visualization is used to show the distribution of a dataset?**
    a. Line graph  b. Bar chart
    c. Scatter plot  d. Histogram

www.ilmkidunya.com

61. **A histogram groups data into:**
a. Time periods b. Bins or intervals c. Categories d. Rows

62. **What type of visualization summarizes data distribution by displaying the median, quartiles, and potential outliers?**
a. Line graph b. Histogram c. Boxplot d. Bar chart

63. **What type of visualization displays relationships between two variables?**
a. Line graph b. Histogram c. Scatter plot d. Bar chart

64. **Which of the following is a widely used tool for data visualization?**
a. Microsoft Word b. Microsoft Excel c. Google Sheets d. b and c

65. **Which of the following is NOT typically used for creating data visualizations?**
a. Microsoft Word b. Microsoft Excel
c. Python d. Tableau

66. **_____ is a data visualization library in Python, widely used to create charts and graphs like bar charts, line graphs, and histograms.**
a. Matplotlib b. NumPy c. SciPy d. Pandas

67. **What kind of charts can be created using Excel or Google Sheets?**
a. Bar charts b. line graphs c. pie charts d. All

68. **What is the first step in creating a visualization in Excel or Google Sheets?**
a. Select the data b. Choose a chart type
c. Enter your data d. Customize the chart

69. **Which tab in Excel or Google Sheets allows you to add a chart?**
a. Home b. Format c. Insert d. View

## Answers

| | | | | | |
|---|---|---|---|---|---|
| 1. a | 2. a | 3. a | 4. d | 5. a | 6. b |
| 7. b | 8. c | 9. b | 10. b | 11. | 12. c |
| 13. d | 14. c | 15. b | 16. d | 17. a | 18. a |
| 19. b | 20. c | 21. c | 22. c | 23. c | 24. b |
| 25. c | 26. c | 27. b | 28. b | 29. b | 30. |
| 31. b | 32. a | 33. a | 34. c | 35. c | 36. b |
| 37. d | 38. c | 39. a | 40. c | 41. a | 42. a |
| 43. b | 44. b | 45. a | 46. a | 47. b | 48. a |
| 49. c | 50. a | 51. c | 52. b | 53. a | 54. c |
| 55. b | 56. a | 57. d | 58. b | 59. a | 60. d |
| 61. b | 62. c | 63. c | 64. d | 65. a | 66. a |
| 67. d | 68. c | 69. c | | | |

www.ilmkidunya.com