Statistics for Data analytics Day 2

Content

1. Measure of central tendency
    a. mean
    b. median
    c. mode
2. choosing appropriate measure
3. Real world application in handling missing values
4. Measures of dispersion
    a. range
    b. variance
    c. Standard deviation
    d. Inter quartile range (IQR)
    e. coefficient of variation
5. Why sample variance is divided by (n-1)
6. Random variables
7. percentiles and quartiles
8. 5 number summary and box plot for removing outlier
9. skewness and kurtosis
10. covariance and correlation
11. Feature selection with covariance and correlation

Measure of central tendency

Measures of central tendency are statistical metrics that represent the center or typical value of a dataset. The three main measures are:

1. **Mean (Average): The sum of all values divided by the total number of values.**

   Mean= sum(data)/count(data)

   - **Sensitive to outliers.**
   - **Best for normally distributed data.**

2. **Median: The middle value when data is arranged in ascending order.**
   - **If there is an odd number of observations, the median is the middle value.**
   - **If even, it is the average of the two middle values.**
   - **Not affected by outliers.**

3. **Mode: The most frequently occurring value in the dataset.**
   - **A dataset may have no mode, one mode (unimodal), or multiple modes (bimodal/multimodal).**
   - **Useful for categorical data.**

Choosing appropriate measure

Choosing the appropriate measure of central tendency depends on the type of data and its distribution. Here's a guide to selecting the right measure:

1. Mean (Best for symmetric, continuous data)

Use when:

- **Data is normally distributed (no extreme outliers).**
- **You need to analyze overall trends (e.g., average income, test scores).**

❌ **Avoid when:**

- **The data has extreme outliers or a skewed distribution (it can be misleading).**

2. Median (Best for skewed data or outliers)

✅ **Use when:**

- **Data is skewed (e.g., house prices, salaries, rainfall).**
- **There are outliers that distort the mean.**
- **You need the middle value in a ranked dataset.**

❌ **Avoid when:**

- **You need to factor in all values for a true average.**

3. Mode (Best for categorical and discrete data)

✅ **Use when:**

- **Working with categorical data (e.g., most popular car color, preferred product).**
- **Finding the most frequent value in a dataset (e.g., most common shoe size).**

❌ **Avoid when:**

- **There’s no repeating value or too many modes.**

Real world application in handling missing values

- **We need to use mean to replace the null values if the data is normally Distributed.**
- **If the data is skewed and have outliers, we can replace the value with median**
- **If the data is categorical, we can replace the values with mode**

Measures of dispersion

Measures of dispersion describe how spread out the data is. They help understand variability, consistency, and distribution. The main measures include:

1. Range: The difference between the maximum and minimum values.

$$\text{Range} = \max(X) - \min(X)$$

✅ **Use When: You need a quick estimate of spread.**
❌ **Limitations: Affected by outliers and ignores data distribution.**

🔹 **Example: If test scores range from 40 to 90, then:**

Range=90-40=50

---

2. Variance: The average squared deviation from the mean.

$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$$

✅ **Use When: You need a precise measure of data spread.**
❌ **Limitations: The unit is squared, making interpretation difficult.**

🔹 **Example: If students' heights vary a lot, variance will be high.**

---

3. Standard Deviation : The square root of variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}$$

✅ **Use When: You need a meaningful spread measure in the same units as the data.**
❌ **Limitations: Affected by extreme values.**

🔹 **Example: In finance, low standard deviation in stock prices means stability.**

Interquartile Range (IQR): **The difference between the 75th percentile (Q3) and 25th percentile (Q1).**

$$\text{IQR} = Q3 - Q1$$

✅ **Use When: You need a robust measure that ignores outliers.**

❌ **Limitations: Not useful for small datasets.**

🔹 **Example: Used in box plots to detect outliers in salary distribution.**

Coefficient of Variation (CV): **The Coefficient of Variation (CV) is a standardized measure of dispersion that expresses the standard deviation as a percentage of the mean. It is useful for comparing variability across datasets with different units or scales.**

Formula

For a population:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100$$

Useful when comparing datasets with different units (e.g., comparing the risk of different financial assets).

Why sample variance is divided by (n-1)

When calculating sample variance, we estimate the population variance using a sample. However, using 'n' in the denominator underestimates the true variance because:

- **The sample mean $\bar{X}$ itself calculated from the sample data.**
- **Since $\bar{X}$ is not the true mean (μ), it tends to make the deviations smaller than those from the true mean.**
- **This underestimation leads to a bias in variance calculation.**

By dividing by (n−1) instead of n, we correct this bias, making the sample variance an unbiased estimator of the population variance. This is also called bessel's correction.

Random variables:-

A random variable is a function that assigns numerical values to outcomes of a random experiment. It represents uncertain quantities in probability theory and statistics.

1. Discrete Random Variable

A random variable that takes on countable values.

It uses probability mass function to find its probability distribution

Ex: probability of 1 after tossing die

2. Continous Random Variable

A random variable that can take any value within a given range (uncountable).

It uses probability definite function to find its probability distribution

Ex: time takes to complete a task

Percentiles

A percentile indicates the value below which a certain percentage of the data falls.

For a dataset of size n, the index i for the p-th percentile is:

$$i = \frac{p}{100} \times n$$

Example: Suppose we have the scores [50, 60, 65, 70, 75, 80, 85, 90, 95, 100], and we want to find the 30th percentile.

i = 30/100×10=3

Quartiles

Quartiles divide the data into four equal parts, each containing 25% of the data.

Quartile	Percentile Equivalent	Meaning
Q1 (First Quartile)	25th percentile	25% of the data is below this value
Q2 (Second Quartile, Median)	50th percentile	50% of the data is below this value
Q3 (Third Quartile)	75th percentile	75% of the data is below this value

Five-Number Summary

The five-number summary provides a quick overview of the data distribution and is essential for detecting outliers using a box plot.

---

1. Five-Number Summary

The five-number summary consists of:

1. **Minimum (min) → Smallest value in the dataset.**
2. **First Quartile (Q1) → 25th percentile (lower quartile).**
3. **Median (Q2) → 50th percentile (middle value).**
4. **Third Quartile (Q3) → 75th percentile (upper quartile).**
5. **Maximum (max) → Largest value in the dataset.**

✅ **Example: Suppose we have the dataset:**
(5,7,8,10,12,15,18,20,22,30)

Statistic	Value
Minimum	5
Q1 (25%)	8
Median (Q2, 50%)	12
Q3 (75%)	20
Maximum	30

---

2. Outlier Detection Using the Interquartile Range (IQR)

The Interquartile Range (IQR) helps detect outliers.

Formula for IQR:

IQR=Q3−Q1

Outlier Boundaries:

- **Lower Bound: Q1−1.5×IQR**
- **Upper Bound: Q3+1.5×IQR**

✅ **Example Calculation:**

- **Q1=8, Q3=20**
- **IQR=20−8=12**
- **Lower Bound: 8−(1.5×12)=-10**
- **Upper Bound: 20+(1.5×12) = 38**

Since all values in the dataset fall between −10 and 38, no outliers exist. However, if a value was outside this range, it would be considered an outlier.

Box Plot for Outlier Detection

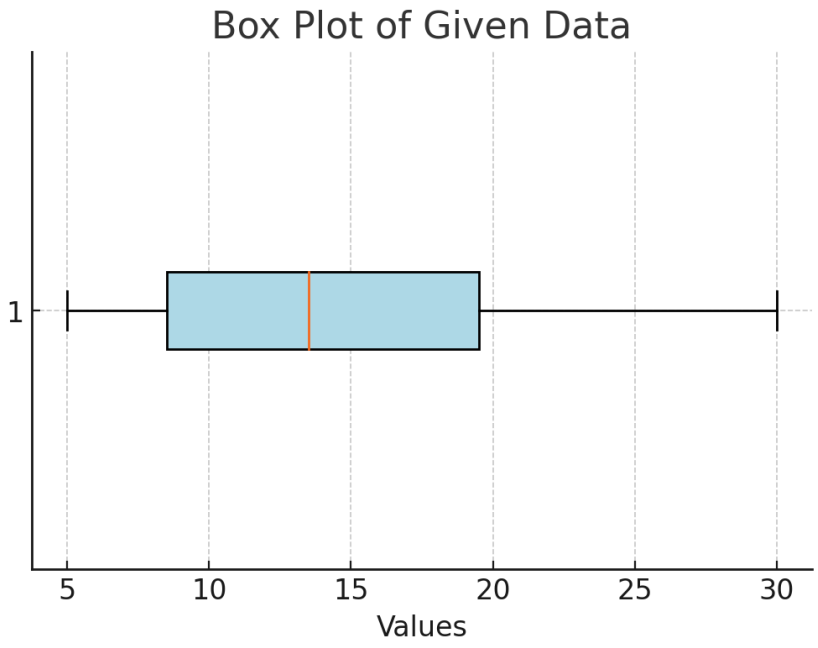
A box plot visually represents the five-number summary and helps identify outliers.

Key Elements in a Box Plot:

- **Box: Represents the IQR (spanning from Q1 to Q3).**
- **Line inside the box: Represents the median (Q2).**
- **Whiskers: Extend to the minimum and maximum values within the outlier bounds.**
- **Outliers: Plotted as individual points beyond the whiskers.**

How to Interpret the Box Plot

- **Longer whiskers → More spread in the data.**
- **Outliers (dots) → Extreme values.**
- **Skewed Box:**
  - **If median is closer to Q1 → Right-skewed (positive skew).**
  - **If median is closer to Q3 → Left-skewed (negative skew).**



Skewness and Kurtosis in Statistics

Skewness:

Skewness measures the asymmetry of a probability distribution. It indicates whether the data is skewed to the left (negative) or right (positive).

- **Positive Skew (Right-Skewed, Skewness > 0): The tail on the right side is longer or fatter. Example: Income distribution.**
- **Negative Skew (Left-Skewed, Skewness < 0): The tail on the left side is longer or fatter. Example: Exam scores (if most students score high).**
- **Zero Skewness (Symmetric, Skewness = 0): The data is perfectly symmetrical, like a normal distribution.**

Kurtosis

Kurtosis measures the "tailedness" of a distribution, i.e., how much of the variance is due to extreme values (outliers).

- **Leptokurtic (Kurtosis > 3): Heavy tails, more extreme outliers (e.g., financial returns).**
- **Platykurtic (Kurtosis < 3): Light tails, fewer extreme values (e.g., uniform distribution).**
- **Mesokurtic (Kurtosis = 3): Similar to a normal distribution.**

Covariance

Covariance measures the direction of the relationship between two variables. It shows whether the variables move together (positive covariance) or in opposite directions (negative covariance).

$$[\text{Cov}(X, Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n}]$$

Advantages:

It will quantify the relation ship between x and y

Disadvantages:

It doesn't have boundaries can range from any value to any value

1. Pearson Correlation Coefficient (r)

- **Measures the linear relationship between two variables.**
- **Assumes both variables are normally distributed and have a linear relationship.**

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

Values range from -1 to 1:

- **1 → Perfect positive correlation**
- **0 → No correlation**
- **-1 → Perfect negative correlation**

Spearman Rank Correlation (ρ)

- **Measures the monotonic relationship between two variables.**
- **Does not assume normality or linearity.**
- **Converts data into ranks before computing correlation.**
- **Converts data into ranks before computing correlation.**

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Feature selection with covariance and correlation

We can directly delete the feature if covariance or correlation is almost zero.

Remove highly correlated features (multicollinearity): If two features have a high correlation (|r| > 0.8), one can be removed.

Section 1: Measures of Central Tendency

1. **Given the dataset: [4, 8, 6, 5, 10, 8, 7], calculate the mean, median, and mode.**
2. **If a dataset has extreme outliers, which measure of central tendency is most appropriate? Explain why.**

Section 2: Choosing the Appropriate Measure

3. **A company records employees' salaries, which include a few extremely high executive salaries. Which central tendency measure should they use to represent typical employee earnings?**

Section 3: Real-World Application in Handling Missing Values

4. **What are some ways to handle missing values in a dataset while ensuring minimal bias?**

Section 4: Measures of Dispersion

5. **Given the dataset [10, 15, 20, 25, 30], calculate the range, variance, standard deviation, and interquartile range (IQR).**
6. **Why is the coefficient of variation useful when comparing variability across datasets with different units?**

Section 5: Why Sample Variance is Divided by (n-1)

7. **Explain why sample variance is divided by (n-1) instead of n. What issue does this correct?**

Section 6: Random Variables

8. **Define a random variable and differentiate between discrete and continuous random variables with examples.**

Section 7: Percentiles and Quartiles

9. **Given the dataset [10, 20, 30, 40, 50, 60, 70, 80, 90, 100], determine the 25th percentile (Q1), 50th percentile (Q2), and 75th percentile (Q3).**

Section 8: 5-Number Summary and Box Plot for Outlier Detection

10. **Construct the 5-number summary for the dataset [3, 7, 7, 12, 15, 22, 29, 31, 45].**
11. **What is an outlier, and how does a box plot help in detecting outliers?**

Section 9: Skewness and Kurtosis

12. **What does a positive skew indicate about a dataset's distribution?**
13. **What does a high kurtosis value suggest about the presence of outliers in a dataset?**

Section 10: Covariance and Correlation

14. **How does correlation differ from covariance in measuring relationships between variables?**
15. **If two variables have a correlation coefficient of 0.95, what does that imply about their relationship?**

Section 11: Feature Selection with Covariance and Correlation

16. **Why is it important to remove highly correlated features when training a machine learning model?**
17. **How can Principal Component Analysis (PCA) use covariance for feature selection?**