Statistics

Reading Material

PW SKILLS

Topics:

1. Introduction to Basic Statistics Terms
2. Types of Statistics
3. Measures of Central Tendency
4. Measures of Dispersion
5. Measures of Skewness
6. Covariance and Correlation
7. Probability Distribution Function and Types

1. Introduction to Basic Statistics Terms

Statistics: The field that deals with gathering, organizing, examining, and understanding numerical information to make smart choices. It uses different methods to describe data and predict or draw conclusions about a group.

Population: The full set of all possible observations or measurements we could make. In studies, the population means the whole group we're interested in and want to know about.

Example: Every student in a school, every customer of a business.

Sample: A smaller group picked from the population for analysis. We choose a sample to stand for the whole population when it's not practical to collect information from everyone.

Example: A group of 100 students from the school, a survey of 500 customers.

Key Concepts

Mean (Average):
Definition: The mean has an impact on data analysis as the central point. It's the total of all values in a dataset divided by how many values there are.

Formula: Mean = (Total of all values) / (Number of values)
Example: Let's say we have [5 10 15]. To find the mean, we do (5 + 10 + 15) / 3 = 10.

Median:
Definition: The median refers to the middle number in a set of data when you line up the numbers from smallest to largest. If you have an even amount of numbers, you take the average of the two middle ones to get the median.

Example: Let's say you have these numbers: [3, 5, 7]. The median is 5. Now, if you have [3, 5 ,7, 9], the median is (5 + 7) / 2 = 6.

Mode:

Definition: The mode is the number that shows up most often in a set of numbers. A set of numbers might have one mode more than one mode, or no mode if all numbers appear often.

Example: In the set of numbers [2, 4, 4, 6, 8], 4 is the mode.

Variance:

Definition: Variance shows how much data points differ from the average. It's the mean of the squared differences from the average giving us a picture of how spread out the data is.

Formula: Variance (σ^2) = $\Sigma(X - \text{Mean})^2 / N$
Example: For the numbers [2, 4, 4, 6, 8], Variance = $[(2-4)^2 + (4-4)^2 + (4-4)^2 + (6-4)^2 + (8-4)^2] / 5 = 4$.

Standard Deviation:

Definition: The standard deviation shows the average difference of each number from the mean. It's the square root of the variance. Many people use it to measure how spread out data is.

Formula: Standard Deviation (σ) = √Variance

Example: Let's say we have these numbers: [2 4, 4, 6 ,8]. The Standard Deviation here is √4 = 2.
Example: A group of 100 students from the school, a survey of 500 customers.

2. Types of Statistics:

Descriptive Statistics:

Definition: Descriptive Statistics includes methods to sum up and organize data for easy understanding. It offers a way to describe the main features of a dataset using numbers.
Purpose: To sum up and describe the traits of a dataset. This helps us understand how the data is spread out, what's typical, and how much it varies.

Examples of Measures:

Mean: The average value.
Median: The middle value when you put the data in order.
Mode: The value that shows up most often.
Range: How far apart the highest and lowest value
Variance: The mean of the squared differences from the averages.
Standard Deviation: The variance's square root showing how values spread around the average.

Inferential Statistics:

Definition: Inferential Statistics makes predictions or draws conclusions about a larger group based on a data sample from that group. It goes beyond simply describing and tries to draw broader conclusions about the whole group.

Purpose: To predict or make general statements about a population using a sample, and to check how reliable these predictions are. This branch of statistics tests hypotheses and estimates population measures.

Examples of Techniques:

Hypothesis Testing: A way to check if sample data backs up a guess about a population measure.
Confidence Intervals: A set of numbers that includes the true population value showing how sure we can be.
Regression Analysis: A method to see how things relate to each other and guess one thing based on another.

3. Measures of Central Tendency

- **Mean (Average):**

Definition: The total of all numbers in a dataset divided by how many numbers there are.
Calculation: Add all numbers and divide by the count.
Use Case: Works best for data that spreads out without odd numbers that stand out. Gives a rough idea of the middle value.

- **Median:**

Definition: The number in the middle when you line up all the data. If you have an even amount of numbers, you take the average of the two middle ones.
Calculation: Put the data in order from lowest to highest and find the value in the middle
Use Case: Works best when the dataset has outliers or leans to one side. The median doesn't change with extreme values.

- **Mode:**

Definition: The value that shows up most often in a dataset.
Calculation: Find the value that appears more than any other.
Use Case: helpful for category-based data or to identify the most common value in a dataset.

Comparison and Significance

- **Mean:**

Importance: Shows the center point of a dataset, but outliers can sway it a lot.
Outlier Impact: Extreme values can pull the mean making it less accurate when outliers exist.

- **Median:**

Importance: Gives a better middle measure for lopsided data spreads.
Handles Outliers: The median stays steady even with outliers, so it's trustworthy in these cases.

- **Mode:**

Importance: Pinpoints the most frequent value, which helps to analyze category-based data.
Helps with Category-Based Info: When we're dealing with non-number data or want to find what shows up most often, the mode comes in handy.

4. Measures of Dispersion

Range:
Definition: The gap between the highest and lowest numbers in a set of data.
Calculation: Range = Highest Number - Lowest Number

Use Case: Gives a basic idea of how spread out the data is, showing the full span of the numbers.
Significance: It's easy to work out, but big outliers can throw it off. It might not tell the whole story about how the data varies.

Variance:
Definition: The average of the squared differences from the mean showing how much the data spreads from the average

Use Case: Variance plays a key role in statistical methods and models. It gives us a clear picture of how data points spread out.
Importance: Variance shows how much each data point differs from the mean. However, it uses squared units, which can make it harder to understand at first glance.

Interpreting Measures of Dispersion

- **Range:**

Interpretation: A large range indicates significant variability in the dataset, while a small range suggests the data points are close to each other.
Limitation: The range is influenced by outliers and does not consider the distribution of values within the range.

- **Variance:**

Interpretation: A higher variance indicates that data points are more spread out from the mean, suggesting greater diversity in the dataset.
Limitation: Since variance is in squared units, it can be harder to interpret directly compared to standard deviation.

- **Standard Deviation:**

Interpretation: A lower standard deviation means data points are close to the mean, while a higher standard deviation indicates greater spread. It is often used in conjunction with the mean to summarize the dataset.
Application: Widely used in finance, quality control, and many other fields to assess risk, consistency, and variability.

5. Measures of Skewness

Introduction to Skewness

Definition: Skewness is a measure of the asymmetry or deviation from symmetry in the distribution of data.

- **Symmetric Distribution:** Data is evenly distributed around the mean, with the left and right sides of the histogram mirroring each other.
- **Skewed Distribution:** Data is not symmetrically distributed, causing the distribution to lean towards one side.

Types of Skewness:

Positive Skewness (Right-Skewed):
Description: The tail on the right side of the distribution is longer or thicker than the left side.
Interpretation: The mean exceeds the median showing that a few very high values pull the mean up.
Implication: Data clusters on the left, with outliers on the right.

Example: Income distribution where most people earn average wages, but a small group has high incomes.

Negative Skewness (Left-Skewed):

Description: The distribution's left side has a longer or thicker tail compared to its right side.
Interpretation: The mean falls below the median showing that a few very low values pull the mean down.
Implication: Data clusters on the right, with outliers on the left.

Example: Retirement age where most folks stop working around a specific age, but some quit much earlier.

Interpreting Skewness Values:

Skewness = 0:
Description: The distribution is perfectly symmetrical, resembling a normal distribution.
Implication: Mean and median are equal, and the distribution has no bias towards left or right.

Example: Heights of adult men in a given population.

Skewness > 0 (Positive Skewness):

Description: Distribution has a right tail longer or fatter than the left tail.
Implication: Mean is greater than the median, indicating a concentration of lower values with a few high outliers.
Application: Important in financial risk management where understanding the skewness of asset returns is crucial.

Skewness < 0 (Negative Skewness):

Description: Distribution has a left tail longer or fatter than the right tail.
Implication: Mean is less than the median, indicating a concentration of higher values with a few low outliers.
Application: Useful in fields like biology where certain characteristics may be negatively skewed, such as the lifespan of a species.

6. Covariance and Correlation

Covariance:

Definition: Covariance measures the degree to which two variables change together.
Types
Positive Covariance: Indicates that as one variable increases, the other also tends to increase.
Negative Covariance: Indicates that as one variable increases, the other tends to decrease.
Magnitude and Direction: Covariance indicates the direction of the linear relationship between two variables but does not standardize the magnitude.

Correlation:
Definition: Correlation is a standardized measure of the strength and direction of the linear relationship between two variables.

Ranges from -1 to 1:

1: Perfect positive correlation.
-1: Perfect negative correlation.
0: No correlation.

Magnitude and Direction: Correlation not only indicates the direction of the relationship but also provides a standardized magnitude, making it easier to compare across different datasets.

Types:

Positive Correlation (r > 0): A strong positive correlation indicates that as one variable increases, the other variable tends to increase in a proportional manner.
Example: Study hours and exam scores—a higher number of study hours is typically associated with higher exam scores.

Negative Correlation (r < 0): A strong negative correlation indicates that as one variable increases, the other variable tends to decrease proportionally.
Example: Price of a product and demand—if the price increases, demand might decrease.

Zero Correlation (r ≈ 0): No linear relationship between the variables.
Example: Shoe size and intelligence—there's no logical connection between these two variables.

7. Probability Distribution Function

- Probability Distribution Function (PDF):A Probability Distribution Function (PDF) shows how likely a random variable will have a specific value. It helps us grasp the probability distribution of a random variable giving us a clear picture of the chances for different results.

Discrete vs. Continuous Distributions:
Discrete distributions have outcomes you can count, like how many times you succeed in a set of tries. For instance, you can count the number of heads when you flip a coin several times. On the flip side continuous distributions have endless outcomes. Take a person's height as an example - it can be any of countless values within a certain range.

Discrete Distributions:

1. **Binomial Distribution:** The binomial distribution shows how many times you succeed in a set number of separate attempts. You use it when each try can end in one of two ways (win or lose).
- **Probability Mass Function (PMF):** The PMF tells you how likely each possible outcome is giving you a clear picture of your chances of getting different numbers of wins.

2. **Poisson Distribution:** The Poisson distribution has an impact on modeling the number of events that take place within a set period or area. People often use it to count happenings that occur randomly and on their own over time or space.
- **Key Concept:** The Probability Mass Function (PMF) plays a crucial role in discrete distributions because it tells you how likely each separate outcome is.

Continuous Distributions:

1. **Normal Distribution:** The normal distribution, also called the Gaussian distribution, has a bell-shaped curve.
- **Two factors define it:** the mean and standard deviation. These show how data points spread out around the mean.

2. **Exponential Distribution:** The exponential distribution models the time between events in a Poisson process. It helps to describe the time gaps between events that happen non-stop and on their own.
- **Key Concept:** The Probability Density Function (PDF) shows the chances of a variable falling within a specific range for continuous distributions. It offers a method to figure out the likelihood of outcomes over an interval.