STATISTICS

(1) DESCRIPTIVE STATISTICS

(2) INFERENTIAL STATISTICS

* (1) DESCRIPTIVE * STATISTICS

(a) CENTERAL TENDENCY - MEAN, MEDIAN, MODE

(b) DISPERSION - VARIATION/VARIANCE, STANDARD DEVIATION (SD)

(c) DATA EXPLORATION - Histograms, pie charts, Bar charts.

* There are 3 Types of Descriptive statistics.

(a) MEASURES of central tendency

(b) MEASURES of variability (spread) -

(1) Range (2) variance (3) standard deviation

*(1) DESCRIPTIVE STATISTICS

(A) CENTERAL TENDENCY

MEAN | MEDIAN | MODE

(1) MEAN :-

MEAN is the average of a set of numbers. It is calculated by adding up all the numbers and dividing by the total count of numbers.

Formula :-

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where,

* $\bar{x}$ It represents the mean.
* $\sum_{i=1}^{n} x_i$ represents the sum of all the values in the set.
* 'n' represents the total number of values in the set.

Example :-

In the set [5, 10, 15, 20], the mean is calculated as :

$$\text{Mean} = \frac{5+10+15+20}{4} = \frac{50}{4}$$

$$= 12.5$$

(2) MEDIAN :-

The MEDIAN is the middle value in a sorted set of numbers. If there are an odd number of values, the median is the middle value.

If there are an even number of values, the median is the average of the two middle values.

Formula :-

(a) For an odd number of values.

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{value}$$

(b) For an even number of values:

$$\text{Median} = \left(\frac{n}{2}\right)^{th} \text{value} + \left(\frac{n}{2}+1\right)^{th} \text{value}$$

* 'n' if represent the total number of values in the set.

Example :-

* In the set [5, 10, 15, 20, 25].

The ~~me~~ middle value 15 is the median.

* In the set [5, 10, 15, 20]

The median is -

$$\text{Median} = \frac{10+15}{2}.$$

$$= 12.5 //$$

(3) MODE :

In Mode the value that appears most frequently in a set of numbers.

Example :

(a) In the set [5, 10, 15, 20, 10, 5]

The mode is 5 & 10, why because both appear twice, which is more than any other number.

(b) In the set [5, 10, 15, 20], there is no mode as each number appears only once.

Summary :-

(1) Mean :
The average of a set of numbers

(2) Median :
The middle value in a sorted set of numbers.

(3) mode :
The value that appears most frequently in a set of numbers.

Question :(B) DISPERSION

Q:- Find the standard deviation & variance of the following data:

[5, 9, 8, 12, 6, 10, 6, 8]

(1) variance :-

variance is the average of the squared differences from the mean.

Formula:

$$\text{Variance}(\sigma^2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2}$$

where,

* σ^2 it represents the variance
* x_i it represents each ~~each~~ value in the dataset
* $\bar{x}$ it represents the mean of the dataset

*n It represents the total number of values in the dataset.

(2) Standard Deviation:

Standard deviation (σ) = $\sqrt{\text{variance}}$

*Calculation steps:

(a) Calculate the mean ($\bar{x}$).
(b) Substract the mean from each value of find the deviation of each value.
(c) Square each deviation.
(d) Find the average of these squared deviations (this is the variance).
(e) Take the square root of the variance to get the standard deviation.

* Example dataset :

$[5, 9, 8, 12, 6, 10, 6, 8]$

(a) Calculate the Mean ($\bar{x}$) :

$$\bar{x} = \frac{5+9+8+12+6+10+6+8}{8} = \frac{64}{8}$$

$$= 8_{//}$$

(b) Calculate each deviation from the Mean and square it :

$(5-8)^2 = (-3)^2 = 9$

$(9-8)^2 = (1)^2 = 1$

$(8-8)^2 = (0)^2 = 0$

$(12-8)^2 = (4)^2 = 16$

$(6-8)^2 = (-2)^2 = 4$

$(10-8)^2 = (2)^2 = 4$

$(6-8)^2 = (-2)^2 = 4$

$(8-8)^2 = (0)^2 = 0$

(c) Sum of squared deviations:

$9+1+0+16+4+4+4+0 = 38$

(d) Calculate the variance:

$\text{variance}(\sigma^2) = \frac{38}{8}$

$= 4.75$ //

(e) Calculate the standard deviation:

Standard deviation $(\sigma) = \sqrt{4.75} \approx 2.18$ //

Conclusion

* variance is 4.75
* standard deviation ≈ 2.18

So, for the dataset [5, 9, 8, 12, 6, 10, 6, 8] the variance is 4.75 and the standard deviation is approximately 2.18 //

(B) MEASURES of VARIABILITY (Spread)

These measures indicate the ~~spread~~ or spread or dispersion of the dataset.

(1) Range (2) variance (3) standard deviation

(1) Range :
The difference between the highest and lowest values.

(2) variance :
The average of the squared differences from the mean.

(3) standard deviation :
The square root of the variance, indicating how much the values in the dataset deviate from the mean.

Example :-

~~Contd~~

Consider the data set [10, 20, 20, 30, 40]

(1) Range :

Range = 40 - 10

= 30 //

(2) variance :

(i) Calculate the mean : 24

(ii) Then, find the squared deviations and their average :

$(10-24)^2 = 196$

$(20-24)^2 = 16$

$(20-24)^2 = 16$

$(30-24)^2 = 36$

$(40-24)^2 = 256$

(iii) Sum of square deviations :

$$= 196 + 16 + 16 + 36 + 256$$

$$= 520$$

(iv) variance $= \frac{520}{5}$

$$= 104_{//}$$

(3) standard deviation :

Standard deviation $= \sqrt{104} \approx 10.2_{//}$

* OTHER × DESCRIPTIVE × STATISTICS

(1) **Minimum and Maximum:**

The smallest and largest values in the dataset.

(2) **Quartiles:**

values that divide the dataset into four equal parts.

(3) **Interquartile Range (IQR):**

The range of the middle 50% of the data, calculated as Q3 - Q1.

* descriptive statistics provide a comprehensive summary of the dataset, allowing for a better understanding of the distribution, central value & spread of the ~~data~~ data.

Example:

Consider the dataset: [10, 20, 20, 30, 40]

* Minimum: 10
* Maximum: 40
* Quartiles:
  * Q1 (25th percentile): 20
  * Q2 (50th percentile or median): 20
  * Q3 (75th percentile): 30
* Interquartile Range (IQR):

$$IQR = Q3 - Q1 = 30 - 20 = 10$$

Conclusion

* Central Tendency: Mean (24), Median (20), Mode (20)
* variability: Range (30), variance (104), standard deviation ($\approx$ 10.2)
* other statistics: Minimum (10), Maximum (40), Quartiles (20, 20, 30), IQR 10.

* DESCRIPTIVE STATISTICS

(C) DATA EXPLORATION:

1. pie chart
2. Histogram
3. Bar chart

1. Pie chart:

A pie chart is a circular graph divided into slices to illustrate numerical proportions.

Each slice represents a category's proportion relative to the whole.

Purpose:

* Representation:
  Shows how individual parts contribute to a whole.

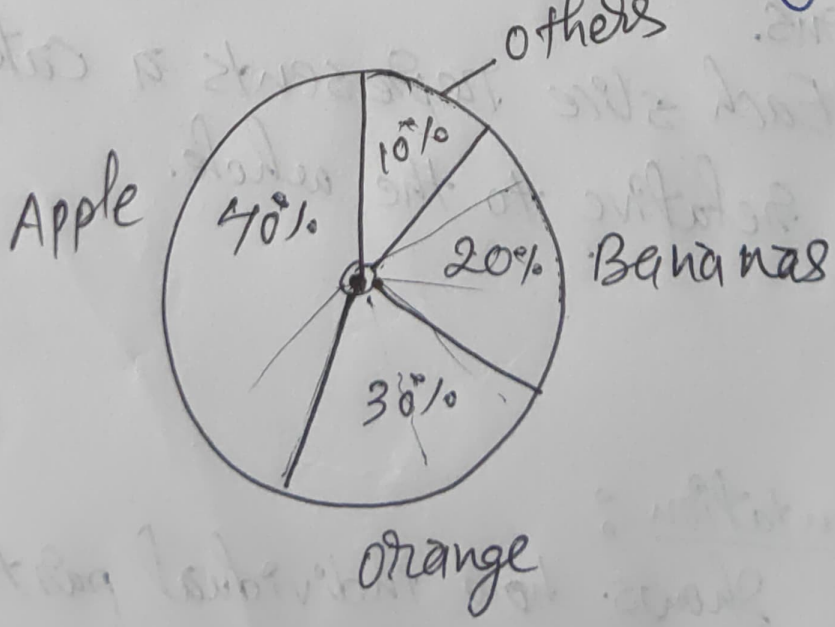
* Comparison:
  Facilitates visual comparison of -

different categories or parts.

Example:

Suppose we have survey data on favorite fruits.:
Apples (40%),
Oranges (30%),
Bananas (20%), &
others (10%).

A pie chart visually displays these proportions, with each slice representing the percentage of respondents choosing each fruit.



2. Histogram:

A Histogram is a graphical representation of the distribution of numerical data.

It consists of bars, where each bar represents a range (or bin) of values and the height represents the frequency of data points within that range.

purpose:

* Distribution:
  shows the shape, spread and center of data distribution.

* Frequency:
  visualize how often different values occur within specified intervals (bins).

~~Example~~:

* Example:

* Histograms are commonly used in statistics to display data distributions, such as exam scores grouped into bins like 0-10, 11-20 etc.

Each bar's height indicates how many students scored within that range.

3. Bar chart:

A bar chart represents categorical data with rectangular bars. The length or height of each bar is proportional to the frequency or value it represents.

purpose:

* Comparison:
  Compares different categories or groups.
* Frequency:
  shows absolute or relative ~~frequencies~~ frequencies of categorical data.

* Example:
  * Consider sales data for different months. January ($1000), February ($1500), March ($1200). A bar chart displays these values as bars of corresponding heights, making it easy to compare sales across months visually.

* CONCLUSION:

(1) pie chart shows proportions of a whole.

(2) Histogram displays distribution of numerical data.

(3) Bar chart compares categorical data with bars.

Usage in descriptive statistics:

* Descriptive statistics.
  These charts are essential tools for summarizing and visualizing data distributions, frequency and proportions.

* Insight Generation.
  They help analysts and researchers identify patterns, trends, outliers and relationships within datasets.

* Communication.
  Effective communication of data insights to stakeholders, clients or decision makers.

STATISTICS

(2) INFERENTIAL STATISTICS

Inferential statistics involves making inferences about populations using data drawn from the population. Instead of merely describing the data as in descriptive statistics, inferential statistics allow us to make predictions or generalizations about a larger group based on a sample of data.

* Key Concepts in Inferential statistics

(1) population vs. sample
(2) parameter vs. statistic
(3) Hypothesis Testing.
(4) Confidence Intervals
(5) Regression Analysis
(6) T-Tests and ANOVA

(1) population vs. sample.

* Population :

The entire group that we want to draw conclusions about.

* sample :

A subset of the population used to collect data and make inferences about the population.

(2) parameter vs. statistic :

* parameter :-

A numerical characteristic of a population (eg, population mean).

* statistic :

A numerical characteristic of a sample (eg, sample mean).

(3) Hypothesis Testing:

* Null Hypothesis (H_0):

The hypothesis that there is no effect or no difference.

* Alternative Hypothesis (H_a):

The hypothesis that there is an effect or a difference.

* p-value:

The probability of observing the data, or something more extreme, if the null ~~hypo~~ hypothesis is true.

* Significance level (α):

A threshold for determining wether the p-value is low enough to reject the null hypothesis (commonly 0.05).

(4) Confidence Intervals:

* A range of values that is likely to contain the population parameter with a certain level of confidence (e.g, 95%).

(5) Regression Analysis:

A statistical method for examining the relationship between two or more variables.

(6) T-tests and ANOVA:

* T-Test: used to compare the means of two groups.

* ANOVA (Analysis of Variance): used to compare the means of three or more groups.

* EXAMPLE OF INFERENTIAL STATISTICS

Q: Scenario:

A Company wants to know the average amount of time its employees spend on breaks during a work days.

Instead of asking all, 1,000 employees, the Company surveys a sample of 100 employees.

(1) Sample data collection:

* Suppose the average break time from the sample of 100 employees is ~~30 minutes~~ 30 minutes with a standard deviation of 5 minutes.

(2) Estimate population Mean:

* The Company uses the sample mean to estimate the population mean.

(3) Confidence Interval:

* Calculate the 95% confidence interval for the population mean break time.

* Formula :- $CI = \bar{x} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$

* where,

  * $\bar{x}$ is the sample mean (30 minutes)
  * Z is the ~~zero~~ Z-score corresponding level (1.96 for 95%).
  * σ is the sample standard deviation (5 minutes)
  * n is the sample size (100)

* Calculations:

$$CI = 30 \pm 1.96\left(\frac{5}{\sqrt{100}}\right)$$

$$CI = 30 \pm 1.96\ (0.5)$$

$$CI = 30 \pm 0.98$$

$$CI = [29.02, 30.98]$$

* Interpretation:

we are 95% confident that the true mean break time for all employees is between 29.02 and 30.38 minutes.

(4) Hypothesis Testing:

* Suppose the company has a policy that states the average break time should be 28 minutes.
* NULL Hypothesis (H_0): The average break time is 28 minutes.
* Alternative Hypothesis (H_a): The average break time is not 28 minutes.
* Calculate the test statistic (t-score):

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{30 - 28}{\frac{5}{\sqrt{100}}}$$

2

* Compare the t-score to the critical value from the t-distribution table with $\boxed{df = n - 1 = 99}$.

* If the t-score exceeds the critical value, reject the null hypothesis.

(5) Conclusion:

If the p-value corresponding to the t-score is less than 0.05, we reject the null hypothesis and conclude that the average break time is significantly different from ~~28 min~~ 28 minutes.

Summary

* Inferential statistics allows us to make predictions or generalizations about a populations based on a sample.

* Key techniques include hypothesis testing, Confidence intervals and regression analysis.

* Example:
  Estimating the average break time for all employee based on a sample and testing if it differs from a specified value.

~~Date~~.

Inferential statistics provide powerful tools to make decisions and predictions about populations based on sample data, allowing for informed decision-making in various fields.