What is Statistics?

Statistics is the branch of mathematics dealing with the Collection, Analysis, Interpreting & presenting from the numerical data we have in hand.

|  |
| --- |
| pic1.jpg |

Need of Statistics:

Knowledge in Statistics will help us to arrive at a conclusion easily from the versatile date we have.

Once concluded we can present the data in a constructive manner

Eg1:

News reporter makes a prediction of winner for elections based on political campaigns. Here statistics plays a strong part in who will be our government.

Eg2: We need to find which Country have occurred many number of times.

In comparing the 2 tables below, we could see using Statistics if we compute, we can arrive at a conclusion easily

|  |  |  |  |
| --- | --- | --- | --- |
| US | Canada | US | Canda |
| India | Japan | NewYork | China |
| Japan | India | India | New Jersey |
| China | New York | China | India |

|  |  |
| --- | --- |
| **Place** | **Frequency** |
| US | 2 |
| India | 4 |
| China | 3 |
| Japan | 2 |
| New York | 2 |
| New Jersey | 1 |
| Canada | 2 |

Classification of Statistics:

Inferential Statistics

Descriptive Statistics

Statistics

Measures of Central Tendency

Measures of Frequency

Measures of Position

Measures of Variability

**Descriptive Statistics:** The process of Organizing, Summarising and presenting data from the whole population (or) from the sample population taken.

1. **Measures Of Central Tendency**: Uses Mean, Median & Mode
2. **Measures Of Variability**: Uses Variance, Standard Deviation, Range – Minimum/Maximum Values, Kurtosis & Skewness
3. **Measures Of Position**: Percentile, Quartile
4. **Measures of Frequency**: Count, Frequency, Percent

**Measures of Central Tendency**:

1. Mean –> Arithmetic Mean – Sum of all Numbers/Total Numbers
2. Median -> Central Point of the Data
3. Mode -> Most frequently occurring Value

Unimodal -> The Dataset which has only 1 mode

{2, 9, 6, 4, 9, 6, 9} - > 9 is the mode

Bi Modal -> The Dataset which has 2 modes

**Measures Of Variability:**

1. Standard Deviation: The measure of Deviation from Mean is called Standard Deviation.

Its symbol is σ (the greek letter **sigma**), The formula is the square root of the Variance.

σ = Square Root Of Variance [OR]

SD = (x-µ)2/n

1. Range: Maximum Value – Minimum Value

**Measures Of Position:**

1. **Percentile:** If we say X is 85th Percentile, it implies there are 15% of people above X
2. **Quartile:** Dividing Data into 4 parts

**Measures of Frequency:**

1. Frequency: How frequent the data is occurring / Can be called as Count too

**Inferential Statistics:**

This is the process of inferring information about a population based on a sample from that population. As the sample size is typically smaller than the size of the population, such inferred information is subject to a measure of uncertainty.

|  |
| --- |
|  |

Measures used to describe **Population** are called **Parameters**

Measures used to describe **Sample** are called **Statistics**



**Variables in Statistics:**

Statistical Variables

Quantitative

Qualitative

Nominal – Data that are Unordered

Eg: Gender, Marital Status

Measurable – Data that are continuous

Eg: Height, Weight

Countable – Data that are discrete

Eg: No. of Cars,

No. of Children

Ordinal – Data that are Ordered

Eg: Rank, Size of Clothing

Independent

Variables

Dependent – also called as Target / Class Variable

|  |
| --- |
|  |

**Characteristics of Frequency Distribution:**

|  |
| --- |
| Characteristics Of  Frequency Distribution |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Modality | |  | Symmetry | |  | CentralTendency |  | Variability |
| Characteristics Of  Frequency Distribution | | |

|  |
| --- |
| Modality |

**Example:**

|  |  |  |
| --- | --- | --- |
| Unimodal |  | Bimodal |

|  |
| --- |
|  |

|  |
| --- |
| Characteristics Of  Frequency Distribution |

|  |
| --- |
| Symmetry |

|  |  |  |
| --- | --- | --- |
| Symmetric |  | Asymmetric |

|  |  |  |
| --- | --- | --- |
| Positive Skewness  If tail is more towards +ve side |  | Negative Skewness  If tail is more towards -ve side |

|  |
| --- |
|  |

|  |
| --- |
| Characteristics Of  Frequency Distribution |

|  |
| --- |
| Central Tendency |

|  |  |  |
| --- | --- | --- |
| Mean |  | Median |

|  |
| --- |
| Mode |

1. Mean –> Arithmetic Mean – {Sum of all Numbers /Total Numbers}, ∑X /n
2. Median -> Central Point of the Data – {Arrange the numbers in Increasing order and find the mid point}
3. Mode -> Most frequently occurring Value

|  |
| --- |
| Characteristics Of  Frequency Distribution |

|  |
| --- |
| Variability |

|  |  |  |
| --- | --- | --- |
| Range |  | Standard Deviation |

1. Range: Minimum Value – Maximum Value
2. Standard Deviation: Data points away from Mean

**Percentile & Quartile:**

1. **Percentile:** If we say X is 85th Percentile, it implies there are 15% of people above X

=percentile.exc (A1:Ax,K)

=percentile.exc(array, what percentile we require)

1. **Quartile:** Dividing Data into 4 parts

108,207,306,405,504,603,702,801,900

Dividing it into 4 parts: Q1, Q2, Q3 and IQR (Inter Quartile Range) -> Q3 – Q1

Q1 –> 1st Quartile – 25th Percentile

Q2 –> 2nd Quartile – 50th Percentile

Q3 –> 3rd Quartile - 75th percentile

IQR – > Q3 – Q1

**Box Plot:**

Five number summary of a data is called Box Plot

**Outliers** are extreme values

Minimum in below pic – Lower Whisker – That will be greater than the minimum value

Maximum in below pic – Upper Whisker – That will be lesser than the maximum value

|  |
| --- |
|  |

**Coefficient Of Variation:**

**Association between two Variables:**

1. Covariance 🡪 Due to Causation 🡪 Limit is 🡪 -∞ to +∞

Eg: When travel increases, Petrol decreases

1. Correlation Coefficient 🡪 Limit is 🡪 -1 to +1 🡪 So always this is preferred.

This can be explained as changes to the value of one variable predicts the changes to the value of another is called Correlation Coefficient.

Types Of Correlation Coefficient:

* **Positive Correlation**: As one variable increases, other increases
* **Negative Correlation**: As one variable increases, the other decreases
* **No Correlation**: There is no apparent relationship between the variables

Example: --?

**Central Limit Theorem:**

When we take >=30 samples from N numbers [0,1,2,3,….N]& take the mean of the samples and plot it, it will always be a Normal Distribution.

Greater the sample size, more accurate the mean will be present.

That is 🡪 the mean of Sample means 🡪 gives the Mean of Population

**Is Normal Distribution a Continuous Probability Distribution?**

**\*\* WriteUp**

**Z Score:**

The number of unit Standard deviation away from mean is called Z Score.

So based on Emperical Split as given below:

If Z score is 0 🡪 The data point is identical to the Mean

If Z score is +1 🡪 The data point is 1 unit of Standard deviation away from Mean

|  |
| --- |
|  |

Data Visualisation Plots: ???

**Standard Error: ??**

**Margin of Error:**

The maximum difference between the true population parameter and the sample parameter is called Margin of Error.

**Example:**

**Emperical Split:**

|  |
| --- |
|  |

**Confidence Interval:**

**Confidence Level:**

Will Accuracy Increases if Confidence Level Increases – Justify?

\* WriteUp

**Hypotheses Test:**

|  |
| --- |
| Tests |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Z Test |  | T Test |  | X2Chisquared Test |  | F Test |

|  |  |  |
| --- | --- | --- |
| Sampling distribution of  Means |  | Sampling distribution of  Variance |