**Population or Sample**

Descriptive Statistics is applied on both population or sample

from Population--- > No inferential statistics

from Sample--- > estimating the population : Inferential Statistics

Descriptive Stats : Measures on given data (multiple records)

for Discrete Variables within Descriptive Stats

-Mode (Measure)

-Probability

-Frequency Table & Frequency Distribution

-Probability Distribution

**1st Business Moment -- > Measures of Central Tendency**

Mean

Median

Mode

**2nd Business Moment -- > Measures of dispersion or spread**

Minimum

Maximum

Range

Deviation (How much data deviated from its mean)

Mean Deviation

absolute deviation

Mean Absolute Deviation

mean Square Deviation--- > Variance

Root Mean Square Deviation --- >std. deviation

Percentiles

Quartiles

Q1

Q2

Q3

Inter Quartile Range (IQR)

Lower Limit

Upper Limit

Outliers

**Probability**

Sum(P) = 1

Range : [0,1]

Probability for discrete Variable

Probability for continuous Variable (don't go with single value , you have to select only interval)

what is probability of a single value : 0 (only for continuous variable)

f**or Discrete variable**

Frequency Table (Categories (unique variable of discrete variable) ,Frequency)

Frequency Distribution

Probability Distribution

**for Continuous Variabl**e

Frequency Table (Interval, Frequency)

Frequency Distribution & Cumulative Frequency

Probability Distribution & Cumulative Probability Distribution

**3rd Business Moment - > Measures of Shape**

For continuous variable : Distribution (Histogram)

if distribution of data is symmetrical : Symmetrical Distribution or Normal Distribution

if distribution of data is not symmetrical : Unsymmetrical Distribution or Skewed Distribution

----left Skewed

----Right Skewed

**Skewness(Measure)**

If you have population data with you,then no need to estimate ---> directly calculate

No need of Inferential Statistics

then no need to estimate directly calculate

When you are not having population data,

you are having only sample data with you, based on sample data -- > estimate population

Kinley 1000ml water bottles (1000ml - 20/-)

Inspector : 10,00,000

sampling: selecting some data from population

Random Sampling : selecting some data from population randomly

Non Random Sampling (Biased Sampling) :

**Inferential Statistics**

Estimating or Inferencing the population parameter based on sample statistic value

**Hypothesis Testing**

Hypothesis: A statement about population parameter

Hypothesis Testing : A statement about population parameter should be accepted or rejected

Null Hypothesis (HO) : No difference / No action / Status Quo

Alternate Hypothesis (H1) : Difference / Action / Research Hypothesis

Formulate the Null Hypothesis & Alternate Hypothesis based on given question/ project

**Steps involved in Hypothesis Testing**

1. Formulate Null & Alternate Hypothesis

2. check for level of significance (error%)

--- value will be provide to you in question

--- if they have not given, default is 5%

Hypothesis Testing will be done using Statistical Tests

Statistical Tests are 2 types

1. Parametric Tests (Statistical Tests which are applied on Normal Distributed Data)

2. Non-Parametric Tests (Statistical Tests which are applied on Skewed Distributed Data or Categorical Data)

**for Continuous Variable**

we will apply Parametric Tests only

if given data is skewed, then first we will convert to normal distribution & later we apply Parametric Tests

1 Column (continuous)

-- > 1 Sample Test

1 Sample z test (if population Std.Deviation is known)

1 Sample t test (if population Std.Deviation is unknown)

-- > H0 : <= >= == (no action)

-- > H1 : > < != (action)

-- > right tail left tail Two tail

rejection rejection rejection

Area =1-p Area=p Rejection Area = (1-p)\*2

>1 Column (continuous)

--> ANOVA Test (F-test)

--> H0: no difference == (no difference,no action)

H1: difference != (difference,action)

1 Discrete

-- > 1 Proportion Test

-- > H0: statement no action

H1: statement action

>1 Discrete

-- > Chi-square Test

H0: statement no difference

H1: statement difference

In Job Description if they have used Statistical Tests in interview you will get questions on

Steps involved in Hypothesis Testing

Step-1: Formulate H0 & H1

step-2: level of significance (alpha) (error %) > default is 5% =0.05

Step-3: For Continuous, Check whether all variables are normally distributed

if any variable is skewed, convert to normal distribution

Step-4: (i) Select the Statistical Test based on given question

(ii) Apply Statistical Test

(iii) Calculate the P-Value

Step-5: Based on p-value & alpha, either accept H0 or Reject H0

p >= alpha Accept HO --- > P High Null Fly

p < alpha Reject HO --- > P Low Null go

Pizza Example

---------------

1. H0 : mean <=30 mins (no action)

H1 : mean >30 mins (action)

2. alpha =0.05

3. Data is Normal distributed (given in question)

4. Select test : 1 Sample z test

Apply test : z: sample mean - population mean/ std.dev/root(n) --> z=1.09

Calculate p value : Using z table, for z of 1.09, p-value is 0.86

: Since given question is right tail rejection, Rejected Area (1-p) = 1-0.8621 = 0.1379

5. 0.1379 >0.05 (p>alpha)

Accept H0 === Do not reject H0

sample size <30

1 sample z test (pop std.deviation is given)

1 sample t test (pop std.deviation is not given)

sample size >=30

1 sample t test or 1 sample z test

what is minimum sample size required : 30

when sample size = 30, Almost pop std.dev == sample std.dev

**Parametric Tests**

1 Sample Test

Anova Test

1 Sample z test with left side rejection

1 Sample z test with right side rejection

1 Sample z test with 2 side rejection

1 Sample t test with left side rejection

1 Sample t test with right side rejection

1 Sample t test with 2 side rejection

Law

-- -

H0 : person is innocent (no action)

H1 : person is criminal (action)

Judge -- > conclude based on evidences (sample data)

Error : Actual (real)

: your conclusion / prediction / estimation / accept (based on sample data)

Actual person is innocent, Predict person is innocent

Actual person is innocent, Predict person is criminal (wrong prediction)

Actual person is criminal, Predict person is innocent (wrong prediction)

Actual person is criminal, Predict person is criminal

|  |  |  |
| --- | --- | --- |
|  | HO (innocent) | H1 (criminal) |
| HO (innocent)  (Actual) | Actual HO  Predict HO  Correct Accept HO | Actual HO  Predict H1  Incorrect Reject HO  (Type-Il error) |
| H1 (criminal)  (Actual) | Actual H1  Predict HO  Incorrect Accept HO  (Type-l error) | Actual H1  Predict H1  Correct Reject HO |

Probability of a single value is zero

Probability of a continuous variable, will be calculate on a interval

**Central Limit theorem**

point estimate +- error

[mean +- error ] == > [mean-error, mean+error ]

Ex: 28+-2 [26,30]

**Confidence Interval**

refers to the probability that a [population](https://www.investopedia.com/terms/p/population.asp) parameter will fall between a set of values for a certain proportion of times

Univariate Plots : Plots applied on a single variable

Bivariate Plots : Plots applied on any 2 variables

Multivariate Plots : Plots applied on >2 variables

Types of Data

1. Structured Data

2. unstructured Data

Type of Variables (structured Data)

1. Continuous :Histogram, Box plot

2. Discrete Count

3. Discrete Categorical : Box Plot

4. Time Series

Continuous + Count : : Quantitative Variables

Categorical : Qualitative Variables

Bivariate Plots : Plots applied on any 2 variables

Multivariate Plots : Plots applied on >2 variables

Types of Data

1. Structured Data

2. unstructured Data

Type of Variables (structured Data)

1. Continuous

2. Discrete Count

3. Discrete Categorical