STATISTICS : *THINK THE ANALYST WAY*

Probability vs Statistics vs Computation ?
Statistics vs stats ?
Hint: Study vs Tools

Types of Statistics ?

1. ***DESCRIPTIVE STATS :COMPUTATION BASED***
2. ***INFERENTIAL STATS :ASSUMPTION BASED***

Population vs sample ?
Hint: Small portion of population

What sample size is best?
Sampling Techniques ?

1. ***Simple Random Sampling:***
2. ***Stratified Sampling : draw non-overlapped regions***
3. ***Systematic Sampling***
4. ***Convenience Sampling: Target-based sampling***

What do you mean by central tendency ?

Hint : there are three ways to analyze distribution of data

1. ***Examine the maximum area that is satisfied by maximum no of data points : MAJORITY WINS RULE***
2. ***Examine the Average distribution that is covered by data points: BALANCING FACTOR***
3. ***Examine the Probabilistic distribution of data-points :Finding the tendency of all to draw relevant conclusion (STATS WITH PROBABILITY ===>defines Central Tendency)***

What's the difference between mean,median & mode?

1. ***When median == mode==mean ; the dataset is normalized and evenly distributed .***
2. ***When median!=mean ; the plot is SKEWED LEFT OR RIGHT , it need handling***
    1. ***If mean <median , LEFT-SKEWED PLOT***
    2. ***If mean > median , RIGHT-SKEWED PLOT***

Measures of DISPERSION ?

1. ***Central point of data || maximum inclination of data : central tendency (MEAN+MEDIAN+MODE)***
2. ***Spread of data | limitations & boundary of data range :***
    (if only one parameter): VARIANCE
    (if two or more parameters): COVARIANCE & LINEAR REGRESSION

Probability Concepts 😁

Event
Trials
Frequencies

Dependent and independent events
Mutually exclusive events
Conditional Probability
Bayes Theorem
Prior Probability
Posterior Probability

Mathematical Concepts :

1. Population Mean
2. Deviation Population
3. Sample Mean
4. Sample variance
5. Percentile
6. Quartile
7. 1st quartile =25% =q1
8. 3rd quartile= 75%=q3
9. InterQuartile=q3-q1

Outliers =>

1. **Why is it dangerous , and needs handling ?**
   => The existence of more than one maxima in graph and existence of more than one minima in plot=>(error vs slope vs intercept)
   => creating a confused state .
   => disturbing the analysis by deviating the actual mean by a magnified gap.

2. **How to get rid of outliers?**
   => check sources of data
   => check the computed mean values and plug in each of them to real case study
   => check for extra zeros feeding
   => check type compatibility

=> if none solves , drop it out

=>Box Plot is the best plot to test n frame analysis for the outliers .

What are types of Distribution ?

Types of Distribution:

1. ***Normalized Distribution***
2. ***Log Normal Distribution***
3. ***Power Law Distribution***
4. ***Pareto Distribution***
5. ***Standard Normal Distribution***

DISTRIBUTION FUNCTIONS OF DATA 👍

1> Probability Distribution Function (PDF):
2> Random Variables ➕
3> Probability Density Function
4> Probability Mass Function

INFERENTIAL STATS :=>

=> HYPOTHESIS FORMULATION 👍

1. NULL HYPOTHESIS
2. ALTERNATE HYPOTHESIS

=> HYPOTHESIS TESTING 👏

1. Z test
2. T test
3. ANOVA TEST
4. CHI-SQUARE TEST

=> **P-value and Q-value** ???

=> **Types of ERRORS in Hypothesis Testing**:

1. ***Type 1 Error 😡 acceptance of hypothesis , in real it's false***
2. ***Type 2 Error 😡 rejection of hypothesis , in real it's true***
3. ***Make the sample size large with all variations under consideration***

BUILDING THE MODEL =>

1. **Collection of DATA for the given problem**
2. **Look for the target variable , the variable that need to be answered by this research**
3. **Perform Descriptive Statistics (PROBABILISTIC VIEW TO APPROACH DATASET)**
   Will give u the estimation of the distribution of data
   See the gap between Maximum & Minimum values of every parameters
   See the gap between mean and median …. Search for solution that can reduce the gap between mean and median
   See the gap between first quartile and third quartile results
   See the gap between 75% and maximum_value
   Plot the box plot
   Plot the correlation matrix

4. **Filter the dataset**
   Remove the parameters having correlation(r=0)
   Remove the fields with low variance rate
   Remove the duplicate sets
   Replace the missing values with (mean,median,mode,0)
   Replacement of missing values with median is the best solution
   Replacement of missing values with the considering the deviations of other parameters n formulating equation to extract predicted values for replacement .

5. **Explore data by drawing the graph**
   If the plot is normalized , its best fit model :)
   If the plot is not normalized ….
   Its skewed left or right
   It can be gaussian overlapping curve
   It can have outliers
   These all need a SOLUTION 🙂

6. **Now, u have to be INFERENTIAL STATS ZONE 🙂**

   MAKE ASSUMPTION OF NULL HYPOTHESIS AND ALTERNATE HYPOTHESIS

APPLY HYPOTHESIS TESTING TECHNIQUES 👍

1. ***Z-TEST : if the sample size is greater than 30***
2. ***T-TEST: if the sample size is less than 30***
3. ***Analyze p-value and critical values ratio***
4. ***After these computation u get to know whether the ASSUMED HYPOTHESIS IS TRUE OR FALSE***

7. Error Testing 😂

For a model that have linear plot , Y=m1x1+m2x2+m3x3…….(m nth).(x nth) + c
Would be the equation satisfying the model . now for this use the GRADIENT DESCENT technique to find the best fit model, Plot a graph with three parameters intercept(c), slope(m) n Error .
This would give you the Bowl shaped 3-D Visual plot having a global minima
The plot drawn will surely have MINIMA , because the equation format is a CONVEX FUNCTION .
The point of Global Minima would be a point (x_best,y_best) that point corresponds to the lowest Error estimation , and this is where you get the best suitable solution of the problem stated .

.....**Featured by ShadowCodGen** {*JyoTirMai Tiwari*)