STATISTICS WORKSHEET-1

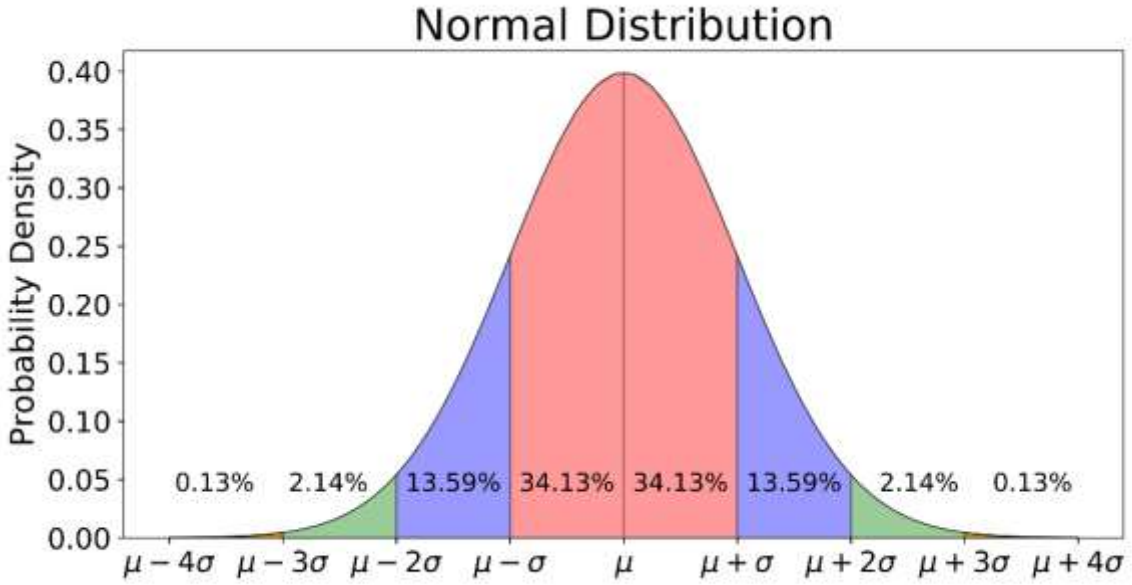
Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False
2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned
4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned
5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?

Ans: Normal distribution is also known as Gaussian distribution; it is a continuous probability distribution where in values lie in a symmetrical fashion mostly situated around the mean. It is the most common distribution function for independent, randomly generated variables. Its familiar bell-shaped curve is ubiquitous in statistical reports, from survey analysis and quality control to resource allocation.



The graph of normal distribution is characterized by two parameters. The mean, or average, which is the maximum of the graph and about which the graph is always symmetric; and the standard deviation, which determines the amount of dispersion away from the mean.
The normal distribution is produced by normal density function, f(x).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\mu - mean$$

$$\sigma - standard\ deviation$$

11. How do you handle missing data? What imputation techniques do you recommend?

Ans: Missing data can be dealt with in a variety of ways. They are often encoded as NaNs, blanks or any other placeholders. Training a model with a dataset that has a lot of missing values can drastically impact the machine learning model's quality. Some algorithms such as *scikit-learn estimators* assume that all values are numerical and have and hold meaningful value.

One way to handle this problem is to get rid of the observations that have missing data. However, one will risk losing data points with valuable information. A better strategy would be to impute the missing values. In other words, we need to infer those missing values from the existing part of the data. And this process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the

true observed values is called ***imputation.***
Let's discuss some of the imputation techniques:

1- Do Nothing: That's an easy one. You just let the algorithm handle the missing data. Some algorithms can factor in the missing values and learn the best imputation values for the missing data based on the training loss reduction. Some others have the option to just ignore them. However, other algorithms will panic and throw an error complaining about the missing values (i.e. Scikit learn — LinearRegression). In that case, you will need to handle the missing data and clean it before feeding it to the algorithm.

2- Imputation Using (Mean/Median) Values: This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.
Pros:
- Easy and fast.
- Works well with small numerical datasets.
- Maintain the same mean and sample size.

Cons:
- Doesn't factor the correlations between features. It only works on the column level.
- Will give poor results on encoded categorical features (do NOT use it on categorical features).
- Not very accurate.
- Doesn't account for the uncertainty in the imputations.

3- Imputation Using (Most Frequent) or (Zero/Constant) Values: Most Frequent is another statistical strategy to impute missing values and YES!! It works with categorical features (strings or numerical representations) by replacing missing data with the most frequent values within each column.
Pros:
- Works well with categorical features.

Cons:
- It also doesn't factor the correlations between features.
- It can introduce bias in the data.

Zero or Constant imputation — as the name suggests — it replaces the missing values with either zero or any constant value you specify.

4- Imputation Using k-NN: The *k* nearest neighbors is an algorithm that is used for simple classification. The algorithm uses '**feature similarity**' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. This can be very useful in making predictions about the missing values by finding the *k's* closest neighbours to the observation with missing data and then imputing them based on the non-missing values in the neighbourhood.
Pros:
- Can be much more accurate than the mean, median or most frequent imputation methods (It depends on the dataset).

Cons:
- Computationally expensive. KNN works by storing the whole training dataset in memory.
- K-NN is quite sensitive to outliers in the data (**unlike SVM**)

Other Imputation Methods:

- ***Regression Imputation:*** The result of regressing the missing variable on other factors to get a predicted value. As a result, instead of utilizing the mean, you are relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

- ***Stochastic regression imputation***: It is quite similar to regression imputation which tries to predict the missing values by regressing it from other related variables in the same dataset plus some random residual value.

- ***Extrapolation and Interpolation***: It tries to estimate values from other observations within the range of a discrete set of known data points

- ***Hot-Deck imputation***: Works by randomly choosing the missing value from a set of related and similar variables. One of the benefits is that you are limited to just feasible values. Another factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

- ***Cold deck imputation:*** A value picked deliberately from an individual with similar values on other variables. In most aspects, this is comparable to Hot Deck, but without the random variance.

In conclusion, there is no perfect way to compensate for the missing values in a dataset. Each strategy can perform better for certain datasets and missing data types but may perform much worse on other types of datasets. There are some set rules to decide which strategy to use for particular types of missing values, but beyond that, one should experiment and check which model works best for your dataset.

12. What is A/B testing?

Ans: A/B testing also known as split testing or bucket testing, refers to a randomized experimentation process where in two or more versions of variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics. Essentially, A/B testing eliminates all the guess work out of website optimization and enables experience optimizers to make data backed decision.
In A/B testing A refers to 'control' or original testing variable. Whereas B refers to 'variation' or new version of original testing variable. The version that moves your business metrics in positive direction is known as 'winner'. Implementing the change of the winning variation on elements can help in increase in business ROI.
It is one of the components of the overarching process of conversion rate optimization, using which you can gather both qualitative and quantitative user insights. And these insights can be further used in following fields:

- Solve visitors pain points
- Get better ROI from existing traffic
- Reduce bounce rate
- Make low risk modifications
- Achieve statistically significant improvements.
- Redesign website to increase future business gains.

Now the question is how to perform A/B test, like any other test following steps are involve in A/B testing:

1. Research
2. Observe and formulate hypothesis
3. Create variations
4. Run test
5. Analyze results and deploy changes
6. Documentation of the findings.

Types of A/B test:

- Split URL testing
- Multivariate testing (MVT)
- Multipage testing

13. Is mean imputation of missing data acceptable practice?

Ans: Mean imputation is acceptable but typically considered terrible practice since it ignores feature correlation. And also, this method can lead into severely biased estimates.

14. What is linear regression in statistics?

Ans: In statistics, linear regression is a linear approach for modelling the relationship between the scalar response and one or more explanatory variables. Or in other words, linear regression analysis is used to predict the value of a variable (dependent variable) based on the value of another variable(independent variable).
Linear Regression analysis estimates the coefficients of linear equation, involving one or more independent variables that best predict the value of the dependent variable and fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

15. What are the various branches of statistics?
Ans: There are basically two main branches of statistics –

1. Descriptive Statistics – It basically deals with collection of data, its presentation in various forms, such as tables, graphs, and diagrams and finding averages and other measures which would describe the data.

2. Inferential Statistics – It deals with the techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.
