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

1. 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?
   1. Central Limit Theorem
   2. Central Mean Theorem
   3. Centroid Limit Theorem
   4. All of the mentioned
2. Which of the following is incorrect with respect to use of Poisson distribution?
   1. Modeling event/time data

b) Modeling bounded count data

1. Modeling contingency tables
2. All of the mentioned
3. Point out the correct statement.
   1. The exponent of a normally distributed random variables follows what is called the log- normal distribution
   2. Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   3. The square of a standard normal random variable follows what is called chi-squared distribution
   4. All of the mentioned
4. random variables are used to model rates.
   1. Empirical
   2. Binomial

c) Poisson

* 1. All of the mentioned

1. 10. Usually replacing the standard error by its estimated value does change the CLT.
   1. True
   2. False
2. 1. Which of the following testing is concerned with making decisions using data?
   1. Probability

b) Hypothesis

1. Causal
2. None of the mentioned
3. 4. Normalized data are centered at and have units equal to standard deviations of the original data.
   1. 0
   2. 5
   3. 1
   4. 10
4. Which of the following statement is incorrect with respect to outliers?
   1. Outliers can have varying degrees of influence
   2. 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?**

Normal distribution is a bell-shaped curve graph where majority of data is centered around the mean.

Hence, values further from the mean are less likely to occur (also known as outliers). It is a plot of the

probability function of a variable that has maximum data concentrated around one point and a few

points taper off symmetrically towards two opposite ends.

Following are the traits of a normal distribution

• The mean, mode and median are all equal.

• The curve is symmetric at the center

* Another peculiarity is 68, 95, 99.7 law, which states that for any normally distributed event, 68

percent of all outcomes fall within one standard deviation, 95 percent of all outcomes fall within

two standard deviations and 99.7 of all outcomes fall within three.

• The graph is symmetric - Half of the values are to the left of center and the other half of the

values are to the right.

• The total area under the curve is 1.

**The Standard Normal Model**

A standard normal model is a normal distribution with a mean as 0 and standard deviation as 1.

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

**Firstly, we need to determine if the data is** Missing completely at random (MCAR), Non-ignorable (NI) missing data or Missing at random (MAR).

By stating that data are MCAR, we assume that the missing values are not systematically different from the values we did observe. Non-ignorable (NI) missing data, also sometimes labeled not missing at random (NMAR) or informative missing data. Concerns about NI data may be raised when missing values are thought to systematically differ from observed values.

* If the volume of data is large, one can delete the blank data cells in columns with missing values.
* Delete the rows or columns that are not necessary and have lots of missing values which is not required while manipulating data
* One can also use plots such as box plots and distribution plots come very handily in deciding which central tendency measure techniques to use - mean, median or mode
* Replacing with next value – Backward fill. In backward fill, the missing value is imputed using the next value.
* Replacing with previous value – Forward fill. In forward fill, the missing value is imputed using the next value.

Interpolation

Missing values can also be imputed using interpolation. The missing values can be imputed with different interpolation methods like ‘polynomial’, ‘linear’, ‘quadratic’.

Using sci-kit learn Library

1. Univariate Approach

In a Univariate approach, only a single feature is taken into consideration. One can use the class SimpleImputer and replace the missing values with mean, mode, median or some constant value.

2. Multivariate Approach

In a multivariate approach, more than one feature is taken into consideration. There are two ways to impute missing values-. Using KNNImputer or IterativeImputer classes.

In case of a titanic dataset:

Suppose the feature ‘age’ is well correlated with the feature ‘Fare’ such that people with lower fares are also younger and people with higher fares are also older. In that case, it would make sense to impute low age for low fare values and high age for high fares values. So here we are taking multiple features into account by following a multivariate approach.

**12. What is A/B testing?**

A/B testing is, is one of the most popular ways for Businesses to test new UX features, new algorithm to decide whether your business should launch that new product/feature or not. new versions of a product  
  
The purpose of executing A/B testing is that you show the new version of the product to a sample of customers (the experimental group) and the existing version of the product to another sample of customers (the control group). The difference in product performance in between experimental/treatment versus the control group is tracked and measured, to identify the effect of this new version(s) of the product on the performance of the product. The idea is to find out whether there is a difference in the performance and if there is a difference what type of difference is it.

**13. Is mean imputation of missing data acceptable practice?**

Mean imputation is often frowned upon and rightly so as it can introduce bias into the data set.

However, it works well in following scenarios:

* Mean imputation works well when the observable data has symmetric distribution.
* One can also use plots such as box plots and distribution plots come very handily in deciding which central tendency measure techniques to use - mean, median or mode.
* Mean should especially not be sued in cases where there are many outliers in the data as outliers have significant impact on the mean and can distort values.
* In such cases using the mean imputation in case for a specific column in a particular data set can give better results rather than leaving the entire data missing.

For example: If the column has most values as the temperature between 39 to 45 Celsius. Mean imputation would work well for missing data.

* Using a specific group average for imputation is a more refined approach, as compared to using the overall dataset average.

For example, there are university student records which include both ‘height’ and ‘gender’ among their attributes. Suppose, that some of the height values are missing, whereas all the gender values are known, we can impute the missing height values by the per-gender category average (175 cm for men, and 163 cm for women), we are much likelier to approximate the ‘true’ values than would be the case with a ‘one-size-fits-all’ mean imputation of the overall dataset mean.

**14. What is linear regression in statistics?**

Simple linear regression is a statistical way if measuring the relationship between two continuous variables. The objective is to predict the value of an output variable (or response) based on the value of an input (or predictor) variable. It is the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables, denoted by X).

Simple linear regression models are models that have a certain fixed number of parameters that depend on the number of input features, and the output is a numeric prediction.

For example, if one is trying to improve process yield, the linear regression will be used to:

* Determine which variables contribute to high yields
* Predicting process yield for future production, given values of our predictors
* Identify factor settings that lead to optimal yields.

Formula:

Y=mx+b

Y =predictor

M= slope

X= input(value we don’t know)

B= bias

**15. What are the various branches of statistics?**

Statistcs concepts are used in data science apart from other fields. It is used to build various models to interpret an analyze data. Statistics has two wide branches - descriptive and inferential stats.

1. Descriptive Statistics: Descriptive Statistics summarizes a given data set, which can either be a representation of the entire or a sample of the population.

Techniques of Descriptive Statistics:

i) Measures of Central tendency (which include mean, median & mode)

ii) Measures of Variability (which include Standard deviation, variance, min & max variables, skewness, etc.)

Examples: A cricket player wants to find his score average for the last 20 games

Purpose: The main purpose of descriptive statistics is a brief summary of the samples.

1. Inferential Statistics: It deals with procedure for making inferences about the characteristics that describe the population or sample.

Techniques of Inferential Statistics: Estimation of Population Parameters (which include Confidence Interval CI), Testing of Statistical Hypothesis

Example: A cricket player wants to estimate his chance of scoring based on his current season average

Purpose: The purpose of statistical inference to estimate the uncertainty or sample to sample variation

