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

Ans- True

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

Ans- Central Limit Theorem

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

Ans- Modeling event/time data

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

Ans- Sums of normally distributed random variables are again normally distributed even if the variables are dependent

5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson

d) All of the mentioned

Ans- Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

Ans- 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

Ans- Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10

Ans- 0

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

Ans- Outliers cannot conform to the regression relationship

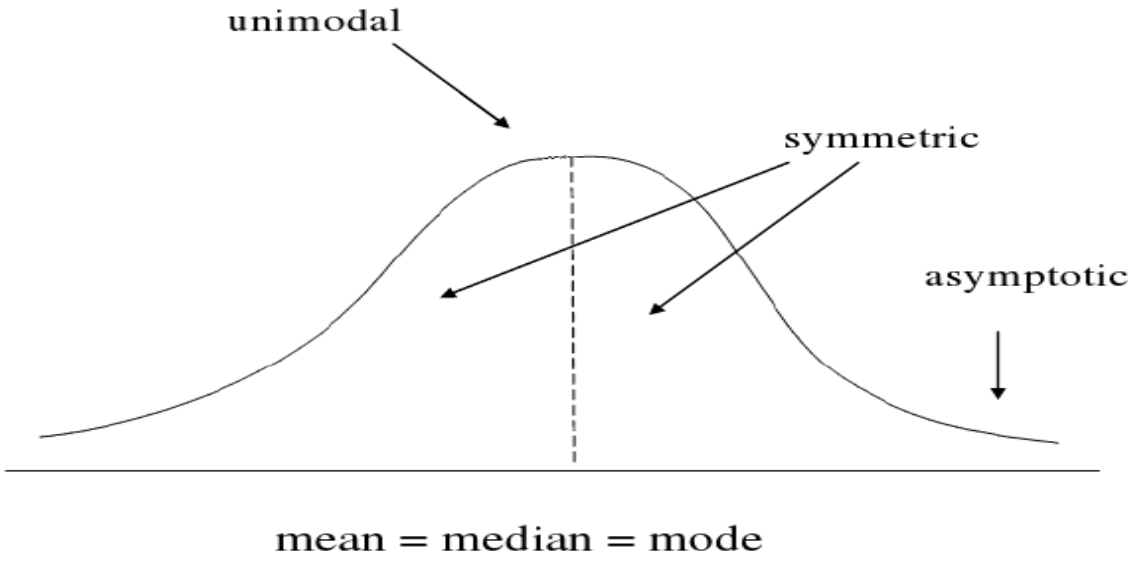
Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans- *The normal distribution is also known as the Gaussian Distribution,*
It is the most important probability distribution in statistics for independent, random variables.
The curve falls within data of standard deviation and mean.
In normal distribution:- mean = median = mode = 0.
In other words Normal distribution is also known as bell shaped curve.

Common Properties for All Forms of the Normal Distribution

- *They're all symmetric bell curves. The Gaussian distribution cannot model skewed distributions.*
- *The curve is symmetric at the center (i.e. around the mean, μ).*
- *The mean, median, and mode are all equal.*
- *Exactly half of the values are to the left of center and exactly half the values are to the right.*
- *The normal distribution doesn't even bother about the range. The range can also extend to –∞ to + ∞ and still we can find a smooth curve.*
- *Half of the population is less than the mean and half is greater than the mean.*
- *The Empirical Rule allows you to determine the proportion of values that fall within certain distances from the mean.*



---

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

Ans- *Missing data (or missing values) is defined as the data values that are not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data.*

Missing data present various problems :-

First,

The absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false.

Second,

The lost data can cause bias in the estimation of parameters.

Third,

It can reduce the representativeness of the samples.

Before we talk about the imputation method we must know the data is missing. So basically there are 3 reasons for that:-

- *Missing at Random (MAR)*
- *Missing Completely at Random (MCAR)*
- *Missing Not at Random (MNAR)*

The simple way to handle the missing data is: either with removal of the data i.e, deletion or dropping data but this method is not always possible so imputation method is

Some of the imputation techniques are :-

- *Multiple imputation*
- *K-Nearest neighbors*
- *few times mean,median and mode can be used.(by calculate the mean or median of the existing observations)*

---

12. What is A/B testing?

Ans- *A/B testing is also known as bucket testing or Split testing or also known as 2 Sample hypothesis testing includes 2 variants A and B used in the statistical hypothesis* ***A/B testing*** *is one of the most popular controlled experiments used to optimize web marketing strategies.*

It allows decision makers to choose the best design by looking at the analytics results obtained with two possible alternatives A and B. It allows to evaluate a product/feature with a subset of users to infer how the product may be received by all users.

A/B testing is a shorthand for a simple randomized controlled experiment, in which two samples (A and B) of a single vector-variable are compared. These values are similar except for one variation which might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment.

However, by adding more variants to the test, its complexity grows.

A/B tests are useful for understanding user engagement and satisfaction of online features like a new Feature or product. Large social media sites like LinkedIn, Facebook, and Instagram use A/B testing to make user experiences more successful and as a way to streamline their services.

A/B tests are being used also for conducting complex experiments on subjects such as network effects when

users are offline, how online services affect user actions, and how users influence one another. Many professions use the data from A/B tests. This includes data engineers, marketers, designers, software engineers, and entrepreneurs. Many positions rely on the data from A/B tests, as they allow companies to understand growth, increase revenue, and optimize customer satisfaction.

---

13. Is mean imputation of missing data acceptable practice?

Ans- *Mean Imputation of missing data means filling the missing data with the mean data of that particular row or any desired column. But I can't say that mean imputation of missing data is acceptable practice because there are some Advantages as well as some Disadvantages Of this.*

So, First I would like to go on with the Advantages of this :-

- *Missing values in your data **do not reduce your sample size**, as it would be the case with listwise deletion (the default of many statistical software packages, e.g. R, Stata, SAS or SPSS). Since mean imputation replaces all missing values, you can keep your whole database.*
- *Mean imputation is very **simple to understand and to apply** (more on that later in the R and SPSS examples). You can explain the imputation method easily to your audience and everybody with basic knowledge in statistics will get what you've done.*
- *If the response mechanism is MCAR, the **sample mean of your variable is not biased**. Mean substitution might be a valid approach, in case that the univariate average of your variables is the only metric your are interested in.*

Now, I would like to Go for the Disadvantages of this :-

- *Mean substitution leads to **bias in multivariate estimates** such as correlation or regression coefficients. Values that are imputed by a variable's mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.*
- ***Standard errors and variance** of imputed variables are biased. For instance, let's assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, the confidence interval around the point estimation of our mean would be too narrow.*
- *If the response mechanism is MAR or MNAR, even the **sample mean of your variable is biased** (compare that with point 3 above). Assume that you want to estimate the mean of a population's income and people with high income are less likely to respond; Your estimate of the mean income would be biased downwards.*

---

14. What is linear regression in statistics?

Ans- *Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:*

- *Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?*
- *Which variables in particular are significant predictors of the outcome variable, and in what way do they– indicated by the magnitude and sign of the beta estimates–impact the outcome variable?*

*These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + m*x, where y = estimated dependent variable score, c = constant, m = regression coefficient, and x = score on the independent variable.*

Three major uses for regression analysis are as follow:-

- *Determining the strength of predictors,*
- *Forecasting an effect, and*
- *Trend Forecasting.*

Types of Linear Regression :-

- *Simple Linear Regression*
  1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

- *Multiple Linear Regression*
  1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

- *Logistic Regression*
  1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

- *Ordinal Regression*
  1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

- *Multinomial Regression*
  1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

- *Discriminant analysis*
  1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio

---

15. What are the various branches of statistics?

Ans- *Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data. The mathematical theories behind statistics rely heavily on differential and integral calculus, linear algebra, and probability theory. Statisticians, people who do statistics, are particularly concerned with determining how to draw reliable conclusions about large groups and general events from the behavior and other observable characteristics of small samples. These small samples represent a portion of the large group or a limited number of instances of a general phenomenon.*
The two major areas of statistics are known as descriptive statistics, which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions.

Some common statistical tools and procedures include the following:

- *Descriptive*
- *Mean (average)*
- *Variance*
- *Skewness*
- *Kurtosis*
- *Inferential*
- *Linear regression analysis*
- *Analysis of variance (ANOVA)*
- *Logit/Probit models*
- *Null hypothesis testing*

The two main branches of statistics are :-

- *Descriptive Statistics*
- *Inferential Statistics.*

Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics
Descriptive Statistics is summarizing the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. It does not involve any generalization or inference beyond what is available. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

Commonly Used Measures

1. *Measures of Central Tendency*
2. *Measures of Dispersion (or Variability)*

Inferential Statistics
Inferential statistics is one of the two main branches of statistics. It use a random sample of data taken from a population to describe and make inferences about the population. Inferential statistics are valuable when examination of each member of an entire population is not convenient or possible. For example, to measure the diameter of each nail that is manufactured in a mill is impractical. We can measure the diameters of a

representative random sample of nails.
We can use the information from the sample to make generalizations about the diameters of all of the nails.
Inferential statistics are used to make generalizations about large groups, such as estimating average demand for a product by surveying a sample of consumers' buying habits or to attempt to predict future events, such as projecting the future return of a security or asset class based on returns in a sample period.
