Statistics– WORKSHEET 4

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

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

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 lognormal 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.
   b) False

7. Which of the following testing is concerned with making decisions using data?
   b) Hypothesis

8.
   Normalized data are centered at ___ and have units equal to standard deviations of the original data.
   a) 0

FLIP ROBO WORKSHEET

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

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

10. What do you understand by the term Normal DIstribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

---

10. NORMAL DISTRIBUTION

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

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

- **Ignore the records with missing values.**

Many tools ignore records with missing values. When the percentage of records with missing values is small, we could ignore those records.

- **Substitute a value such as mean**.

When the percentage is large and also when it makes sense to do something to avoid bias modeling results, substituting a value (e.g. mean, median) is a commonly used way. But this method could cause bias distribution and variance. That's where the following imputation methods come in.

- **Predict missing values.**

Depending on the type of the imputed variable (i.e. continuous, ordinal, nominal) and missing data pattern (i.e. monotone, non-monotone), below are a few commonly used models. If you plan to do it in SAS, there are SAS codes that you can write to identify the missing data pattern.

- Logistic Regression
- Discriminant Regression
- Markov Chain Monte Carlo (MCMC)
- ...
- **Predict missing values - Multiple Imputation**. Although there are pros & cons, MI is considered to be superior to single imputation, and it better measures the uncertainty of the missing values.

In addition, there are a few required **statistical assumptions** for multiple imputation:

1. Whether the data is missing at random (MAR).
2. Multivariate normal distribution, for some of the modeling methods mentioned above (e.g. regression, MCMC).

12. **A/B testing** is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics.

13. **Imputation of missing data** is:

- Bad practice in general.
- If just estimating means: mean imputation preserves the mean of the observed data
- Leads to an underestimate of the standard deviation
- Distorts relationships between variables by “pulling” estimates of the correlation toward zero

14.

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) 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 + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

15.What are the various branches of statistics?

Statistics may be divided into two main branches:
(1) Descriptive Statistics **(2)** Inferential Statistics
(1) Descriptive Statistics
Descriptive statistics deals with the 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.

For example: Industrial statistics, population statistics, trade statistics, etc. Businessmen make use of descriptive statistics in presenting their annual reports, final accounts, and bank statements.

(2) Inferential Statistics

Inferential statistics deals with 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.

For example: Suppose we want to have an idea about the percentage of the illiterate population of our country. We take a sample from the population and find the proportion of illiterate individuals in the sample. With the help of probability, this sample proportion enables us to make some inferences about the population proportion. This study belongs to inferential statistics.