Name: Arup Singha

Project: Batch DS2304

# STATISTICS WORKSHEET

**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: 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
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned
**Ans: 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
**Ans: b) Modeling bounded count 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: d) All of the mentioned**

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned
**Ans: c) Poisson**

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

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
**Ans: a) 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: 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?**

Ans: The term "Normal Distribution," also known as the Gaussian distribution or bell curve, refers to a continuous probability distribution that is symmetric and bell-shaped. It is a fundamental concept in statistics and probability theory.
A normal distribution is characterized by its mean (μ) and standard deviation (σ). The mean represents the center or average value of the distribution, while the standard deviation quantifies the dispersion or spread of the data around the mean.

The shape of the normal distribution is symmetric, with the highest point of the curve located at the mean. The curve tapers off on both sides, asymptotically approaching the x-axis without ever touching it. The standard deviation determines the width of the curve, with larger standard deviations leading to wider and flatter curves.

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

Ans: Handling missing data is an important aspect of data analysis, and there are several approaches to deal with missing data. The choice of imputation technique depends on various factors, such as the nature of the data, the amount of missingness, and the underlying assumptions of the analysis. Here are some common approaches to handling missing data and recommended imputation techniques:

1. Complete Case Analysis (CCA): This approach involves excluding any observations with missing values from the analysis. While CCA is straightforward, it may lead to loss of information if the missing data are not missing completely at random.
2. Mean/Mode/Median Imputation: In this technique, missing values are replaced with the mean (for continuous variables) or mode (for categorical variables) of the observed data. However, this method can lead to biased estimates and underestimation of variance.
3. Last Observation Carried Forward (LOCF): LOCF imputation fills in missing values with the most recently observed value. This approach is commonly used in longitudinal studies where missing values are assumed to remain constant between successive measurements. However, LOCF may not be appropriate if the assumption of data stability is violated.
4. Multiple Imputation (MI): MI is a more advanced technique that involves creating multiple imputed datasets, where missing values are filled in based on a predictive model. These imputed datasets are then analyzed separately, and the results are combined using specialized algorithms. MI handles uncertainty related to missing data and provides unbiased estimates.
5. Maximum Likelihood Estimation (MLE): MLE is a statistical technique that estimates the missing values based on the likelihood function. MLE estimates the parameters of a model while simultaneously imputing the missing values. This method is suitable when the missing data mechanism is assumed to be ignorable.
6. Model-Based Imputation: Model-based imputation involves fitting a model to the observed data and using the model to impute missing values. This can include regression models, decision trees, or other machine learning algorithms. Model-based imputation takes into account the relationships between variables and can provide more accurate imputations.

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

Ans: A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a 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 guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.

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

Ans: Mean imputation of missing data is a common and simple approach to handle missing values. However, its acceptability as a practice depends on the specific context and the nature of the data. While mean imputation is easy to implement, it has some limitations and potential drawbacks that should be taken into consideration:

1. Bias: Mean imputation can introduce bias in the data because it assumes that the missing values have the same mean as the observed values. This assumption may not hold true if the missingness is related to the value itself or to other variables in the dataset. This can lead to an underestimation or overestimation of the true values.
2. Loss of Variability: By replacing missing values with the mean, the imputed data will have less variability compared to the original data. This can lead to an underestimation of the standard deviation and can affect subsequent analyses or modeling.
3. Distortion of Relationships: Mean imputation does not account for the relationships between variables, potentially distorting the associations between variables. Imputing missing values with the mean can artificially strengthen or weaken correlations or other statistical relationships.
4. Ignoring Uncertainty: Mean imputation does not consider the uncertainty associated with the imputed values. It assumes that the imputed values are known with certainty, which can be problematic when conducting statistical analyses or drawing conclusions.

Despite these limitations, mean imputation may be acceptable in certain situations, such as:

- The missingness is random or missing completely at random (MCAR) and there is no relationship between the missing values and other variables.
- The missingness is minimal, and imputing with the mean does not substantially alter the results or conclusions.
- The imputed variable is not a crucial predictor or outcome in subsequent analyses.

In cases where the missingness is not random or missing at random (e.g., missing values are related to other variables or have a specific pattern), mean imputation is generally not recommended. In such situations, more advanced imputation techniques like multiple imputation or model-based imputation are preferable as they account for the underlying relationships and uncertainties in the data.

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

Ans: Linear regression is a type of statistical analysis used to predict the relationship between two variables. It assumes a linear relationship between the independent variable and the dependent variable, and aims to find the best-fitting line that describes the relationship

**15. What are the various branches of statistics?**
Ans: The various branches of statistics are:

1. Descriptive Statistics: Descriptive statistics involves summarizing and presenting data in a meaningful way. It includes measures such as mean, median, mode, standard deviation, and graphical representations like histograms and box plots.
2. Inferential Statistics: Inferential statistics involves drawing conclusions and making inferences about a population based on a sample of data. It includes techniques such as hypothesis testing, confidence intervals, and regression analysis.
3. Probability Theory: Probability theory is the foundation of statistics. It deals with the mathematical study of randomness and uncertainty. It includes concepts such as probability distributions, random variables, and probability calculations.
4. Biostatistics: Biostatistics applies statistical methods to biological and health-related research. It involves designing clinical trials, analyzing epidemiological data, and conducting studies in genetics and public health.
5. Econometrics: Econometrics applies statistical methods to economic data. It involves analyzing economic relationships, estimating economic models, and making forecasts.
6. Social Statistics: Social statistics focuses on statistical analysis and interpretation of social phenomena. It includes analyzing survey data, studying social trends, and conducting social research.
7. Multivariate Analysis: Multivariate analysis deals with the analysis of datasets that involve multiple variables. It includes techniques such as principal component analysis, factor analysis, and cluster analysis.
8. Time Series Analysis: Time series analysis involves analyzing and modeling data that are collected over time. It includes techniques for studying trends, seasonality, and forecasting future values.
9. Bayesian Statistics: Bayesian statistics is a branch of statistics that incorporates prior knowledge or beliefs into the analysis. It involves using Bayesian inference to update beliefs based on observed data.
10. Nonparametric Statistics: Nonparametric statistics is a branch of statistics that does not assume specific probability distributions for the data. It includes methods such as rank-based tests, bootstrapping, and permutation tests.