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 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 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.

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?

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

## Time-Series Specific Methods

* **Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)**  
  This is a common statistical approach to the analysis of longitudinal repeated measures data where some follow-up observations may be missing. Longitudinal data track the same sample at different points in time. Both these methods can introduce bias in analysis and perform poorly when data has a visible trend
* **Linear Interpolation**  
  This method works well for a time series with some trend but is not suitable for seasonal data
* **Seasonal Adjustment + Linear Interpolation**  
  This method works well for data with both trend and seasonality

## Mean, Median and Mode

Computing the overall mean, median or mode is a very basic imputation method, it is the only tested function that takes no advantage of the time series characteristics or relationship between the variables. It is very fast, but has clear disadvantages. One disadvantage is that mean imputation reduces variance in the dataset.

**Linear Regression**

To begin, several predictors of the variable with missing values are identified using a correlation matrix. The best predictors are selected and used as independent variables in a regression equation. The variable with missing data is used as the dependent variable. Cases with complete data for the predictor variables are used to generate the regression equation; the equation is then used to predict missing values for incomplete cases. In an iterative process, values for the missing variable are inserted and then all cases are used to predict the dependent variable. These steps are repeated until there is little difference between the predicted values from one step to the next, that is they converge.  
It “theoretically” provides good estimates for missing values. However, there are several disadvantages of this model which tend to outweigh the advantages. First, because the replaced values were predicted from other variables they tend to fit together “too well” and so standard error is deflated. One must also assume that there is a linear relationship between the variables used in the regression equation when there may not be one.

Multiple Imputation

1. **Imputation**: Impute the missing entries of the incomplete data sets *m* times (*m*=3 in the figure). Note that imputed values are drawn from a distribution. Simulating random draws doesn’t include uncertainty in model parameters. Better approach is to use Markov Chain Monte Carlo (MCMC) simulation. This step results in m complete data sets.
2. **Analysis**: Analyze each of the *m* completed data sets.
3. **Pooling**: Integrate the *m* analysis results into a final result

This is by far the most preferred method for imputation for the following reasons:  
- Easy to use  
- No biases (if imputation model is correct)

Imputation of Categorical Variables

1. Mode imputation is one method but it will definitely introduce bias
2. Missing values can be treated as a separate category by itself. We can create another category for the missing values and use them as a different level. This is the simplest method.
3. Prediction models: Here, we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable (training) and another one with missing values (test). We can use methods like logistic regression and ANOVA for prediction
4. Multiple Imputation

KNN (K Nearest Neighbors)

There are other machine learning techniques like XGBoost and Random Forest for data imputation but we will be discussing KNN as it is widely used. In this method, k neighbors are chosen based on some distance measure and their average is used as an imputation estimate. The method requires the selection of the number of nearest neighbors, and a distance metric. KNN can predict both discrete attributes (the most frequent value among the k nearest neighbors) and continuous attributes (the mean among the k nearest neighbors)  
The distance metric varies according to the type of data:  
1. Continuous Data: The commonly used distance metrics for continuous data are Euclidean, Manhattan and Cosine  
2. Categorical Data: Hamming distance is generally used in this case. It takes all the categorical attributes and for each, count one if the value is not the same between two points. The Hamming distance is then equal to the number of attributes for which the value was different.  
One of the most attractive features of the KNN algorithm is that it is simple to understand and easy to implement. The non-parametric nature of KNN gives it an edge in certain settings where the data may be highly “unusual”.  
One of the obvious drawbacks of the KNN algorithm is that it becomes time-consuming when analyzing large datasets because it searches for similar instances through the entire dataset. Furthermore, the accuracy of KNN can be severely degraded with high-dimensional data because there is little difference between the nearest and farthest neighbor.

12. What is A/B testing

**A/B testing** (also known as **bucket testing** or **split-run testing**) is a user experience research methodology.  A/B tests consist of a [randomized experiment](https://en.wikipedia.org/wiki/Randomized_experiment) with two variants, A and B. It includes application of [statistical hypothesis testing](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing) or "[two-sample hypothesis testing](https://en.wikipedia.org/wiki/Two-sample_hypothesis_testing)" as used in the field of [statistics](https://en.wikipedia.org/wiki/Statistics). A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

13. Is mean imputation of missing data acceptable practice?

No, as they are prone to outliers.

14. What is linear regression in statistics?

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.

Naming the Variables.  There are many names for a regression’s dependent variable.  It may be called an outcome variable, criterion variable, endogenous variable, or regressand.  The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable.  Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes.  That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables.  A typical question is, “how much additional sales income do I get for each additional $1000 spent on marketing?”

Third, regression analysis predicts trends and future values.  The regression analysis can be used to get point estimates.  A typical question is, “what will the price of gold be in 6 months?”

15. What are the various branches of statistics?

We the students of statistics should know about the different branches of statistics to correctly understand statistics from a more holistic point of view. Often, the kind of job or work one is involved in hides the other aspects of statistics, but it is very important to know the overall idea behind statistical analysis to fully appreciate its importance and beauty.

[Descriptive statistics](https://explorable.com/descriptive-statistics) deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid [biases](https://explorable.com/research-bias) that are so easy to creep into the [experiment](https://explorable.com/conducting-an-experiment).

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

[Inferential statistics](https://explorable.com/inferential-statistics), as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Most [predictions](https://explorable.com/prediction-in-research) of the future and [generalizations](https://explorable.com/what-is-generalization) about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small [sample population](https://explorable.com/sample-group) that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to [draw conclusions](https://explorable.com/drawing-conclusions) relevant to his study.

While drawing conclusions, one needs to be very careful so as not to draw the [wrong](https://explorable.com/type-I-error) or [biased](https://explorable.com/research-bias) conclusions. Even though this appears like a science, there are ways in which one can [manipulate studies and results](https://explorable.com/science-fraud) through various means. For example, [data dredging](https://explorable.com/data-dredging) is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good [scientific methodology](https://explorable.com/research-methodology) needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.