Worksheet-Set 1: Statistics Assignment

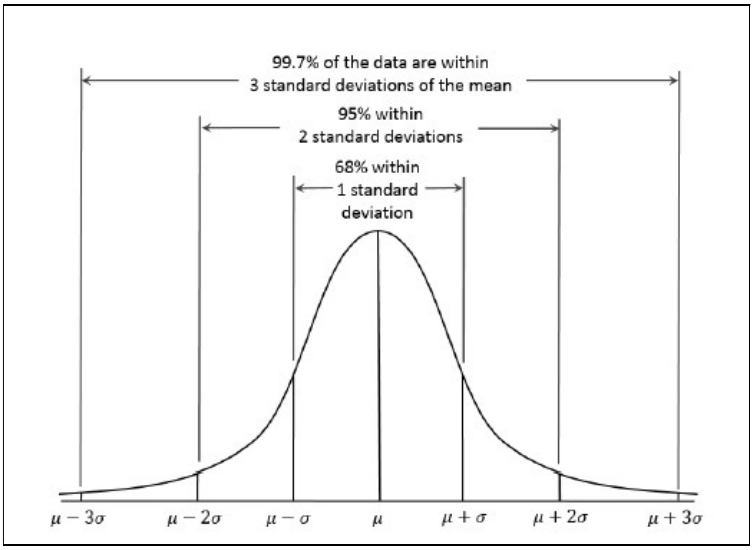
Answer keys for questions from Q1 to Q9:

Question Number	Answer keys
Q1 :	A) True
Q2 :	A) Central Limit Theorem
Q3 :	B) Modeling bounded count data
Q4 :	D) All of the mentioned
Q5 :	C) Poisson
Q6 :	B) False
Q7 :	B) Hypothesis
Q8 :	A) 0
Q9 :	C) Outliers cannot conform to the regression relationship

---

Q10. What do you understand by the term 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 graphical form, the normal distribution appears as a "bell curve".



Following are the key points about Normal Distribution:

- The normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1.
- It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- Many naturally-occurring phenomena tend to approximate the normal distribution.
- In a normal distribution 68% of the data lies within one standard deviation of the mean, and 95% lies within two standard deviations.

Q11. How do you handle missing data? What imputation techniques do you recommend?

There are two ways to deal with missing data, One should understand reasoning about missing data, whether the data is missing at random, missing not at random or missing completely at random.

1. Delete the rows/ columns with missing data. :

When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

- Dropping Variables: If data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.

2. Use of Imputation method to fill missing data.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. Imputation methods can deliver reasonably reliable results.

For Numerical variables, we use mean/median imputation, arbitrary value imputation, end of tail imputation & mode imputation.

For Categorical variables, we use frequent category imputation.

CCA – Complete case analysis is a method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing. This method is also popularly known as "Listwise deletion".

- Assumptions:-
  Data is Missing At Random (MAR); Missing data is completely removed from the table.
- Advantages:-
  Easy to implement; No Data manipulation required.
- Limitations:-
  Deleted data can be informative; Can lead to the deletion of a large part of the data; Can create a bias in the dataset; if a large amount of a particular type of variable is deleted from it; The production model will not know what to do with Missing data.
- When to Use:-
  Data is MAR (Missing At Random); Good for Mixed Numerical and Categorical data,;Missing data is not more than 5% – 6% of the dataset; Data doesn't contain much information and will not bias the dataset.

Arbitrary Value Imputation - This is an important technique used in Imputation as it can handle both the Numerical and Categorical variables. This technique states that we group the missing values in a column and assign them to a new value that is far away from the range of that column. Mostly we use values like 99999999 or -9999999 or "Missing" or "Not defined" for numerical & categorical variables.

- Assumptions:-
  Data is not Missing At Random; The missing data is imputed with an arbitrary value that is not part of the dataset or Mean/Median/Mode of data.
- Advantages:-
  Easy to implement; We can use it in production; It retains the importance of "missing values" if it exists.
- Disadvantages:-
  Can distort original variable distribution; Arbitrary values can create outliers; Extra caution required in selecting the Arbitrary value.
- When to Use:-
  When data is not MAR (Missing At Random); Suitable for All.

Frequent Category Imputation - This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column. This technique is

also referred to as Mode Imputation.

- Assumptions:-
  Data is missing at random; There is a high probability that the missing data looks like the majority of the data.
- Advantages:-
  Implementation is easy; We can obtain a complete dataset in very little time; We can use this technique in the production model.
- Disadvantages:-
  The higher the percentage of missing values; the higher will be the distortion; May lead to over-representation of a particular category; Can distort original variable distribution.
- When to Use:-
  Data is Missing at Random (MAR); Missing data is not more than 5% – 6% of the dataset.

Q12. What is A/B testing?

An A/B test is an experiment with two groups to establish which of two treatments, products, procedures, or the like is superior. Often one of the two treatments is the standard existing treatment, or no treatment. If a standard (or no) treatment is used, it is called the control. A typical hypothesis is that treatment is better than control.
Some examples of A/B testing include:

· Testing two soil treatments to determine which produces better seed germination
· Testing two therapies to determine which suppresses cancer more effectively
· Testing two prices to determine which yields more net profit
· Testing two web headlines to determine which produces more clicks
· Testing two web ads to determine which generates more conversions

---

Q13. Is mean imputation of missing data acceptable practice?

Mean imputation method does not consider feature correlation, which in turn affects the bias and variance of the data set column affecting the final result. For example; suppose we have a data with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.
Mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.
Using mean imputation all the cases/times is not a good practice.

Q14. What is linear regression in statistics?

Regression analysis involves identifying the relationship between a dependent variable and one or more independent variables. A model of the relationship is hypothesized, and estimates of the parameter values are used to develop an estimated regression equation. Various tests are then employed to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable given values for the independent variables.

Regression model:
In simple linear regression, the model used to describe the relationship between a single dependent variable y and a single independent variable x is $y = \beta0 + \beta1x + \varepsilon$. $\beta0$ and $\beta1$ are referred to as the model parameters, and ε is a probabilistic error term that accounts for the variability in y that cannot be explained by the linear relationship with x. If the error term were not present, the model would be deterministic; in that case, knowledge of the value of x would be sufficient to determine the value of y.

In multiple regression analysis, the model for simple linear regression is extended to account for the relationship between the dependent variable *y* and *p* independent variables *x*1, *x*2, . . ., *xp*. The general form of the multiple regression model is $y = \beta0 + \beta1x1 + \beta2x2 + \ldots + \beta pxp + \varepsilon$. The parameters of the model are the β0, β1, . . ., β*p*, and ε is the error term.

Q15. What are the various branches of statistics?

There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

1. **Data collection:** It is all about how the actual data is collected. For the most part, this needn't concern us too much in terms of the mathematics, but there are significant issues to consider when actually collecting data. For data such as marks in a class test, this is fairly straightforward. Each student has a defined mark associated with them, so the marks are simply collected together to make the data set. Sometimes, data is harder to collect. Counting the number of bees in a colony isn't easy, because they move and fly around; you may have to approximate in such cases.
2. **Descriptive statistics:** It is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on). Common visual techniques that we shall discuss in Chapter 2 include graphs, bar charts, pie charts and more, but we shall focus mainly on numerical techniques such as averages and spreads. The basic aim of descriptive statistics is to 'present the data' in an understandable way. If you simply write down every piece of data, it means little to someone who sees it; it needs to be summarized.
3. **Inferential statistics** is the aspect that deals with making conclusions about the data.
