Statistics Worksheet

1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

Answer : 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

Answer : c) 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

Answer 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

Answer : d) All of the mentioned

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

Answer : c Poisson

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

Answer :b) False

7. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Answer : b) 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

Answer 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

Answer : 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?

A normal distribution, sometimes called the bell curve, is a distribution that occurs naturally in many situations. For example, the bell curve is seen in tests like the SAT and GRE. The bulk of students will score the average (C), while smaller numbers of students will score a B or D. An even smaller percentage of students score an F or an A. This creates a distribution that resembles a bell (hence the nickname). The bell curve is symmetrical. Half of the data will fall to the left of the mean; half will fall to the right. Many groups follow this type of pattern. That's why it's widely used in business, statistics and in government bodies like the FDA:

Heights of people.

Measurement errors.

Blood pressure.

Points on a test.

IQ scores.

Salaries.

The empirical rule tells you what percentage of your data falls within a certain number of standard deviations from the mean:

- 68% of the data falls within one standard deviation of the mean.
- 95% of the data falls within two standard deviations of the mean.
- 99.7% of the data falls within three standard deviations of the mean.

The standard deviation controls the spread of the distribution. A smaller standard deviation indicates that the data is tightly clustered around the mean; the normal distribution will be taller. A larger standard deviation indicates that the data is spread out around the mean; the normal distribution will be flatter and wider.

**Properties of a normal distribution**

The mean, mode and median are all equal.

The curve is symmetric at the center (i.e. around the mean, μ).

Exactly half of the values are to the left of center and exactly half the values are to the right.

The total area under the curve is 1.

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

Techniques to handle missing data are

1. Listwise
   Listwise deletion (complete-case analysis) removes all data for an observation that has one or more missing values. Particularly if the missing data is limited to a small number of observations, you may just opt to eliminate those cases from the analysis. However in most cases, it is often disadvantageous to use listwise deletion. This is because the assumptions of MCAR (Missing Completely at Random) are typically rare to support. As a result, listwise deletion methods produce biased parameters and estimates.
2. Pairwise
   pairwise deletion analyses all cases in which the variables of interest are present and thus maximizes all data available by an analysis basis. A strength to this technique is that it increases power in your analysis but it has many disadvantages. It assumes that the missing data are MCAR. If you delete pairwise then you'll end up with different numbers of observations contributing to different parts of your model, which can make interpretation difficult.

3. Dropping Variables
   In my opinion, it is always better to keep data than to discard it. Sometimes you can drop variables if the data is missing for more than 60% observations but only if that variable is insignificant. Having said that, imputation is always a preferred choice over dropping variables

4. 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

12. What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

13. Is mean imputation of missing data acceptable practice?

There are some drawbacks of mean imputation:

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.

Hence it is not a good practice

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.

15. What are the various branches of statistics?

Descriptive Statistics

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 that are so easy to creep into the experiment.

Inferential Statistics

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.