**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: 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**

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**

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: d**

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**

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**

**SUBJECTIVE ANSWER TYPES**

**10) What do you understand by the term Normal Distribution?**

**Answer:** A Normal Distribution is also known as a Gaussian distribution or famously Bell Curve. It is a continuous probability distribution. The normal distribution density function simply accepts a data point along with a mean value and a standard deviation and throws a value which we call probability density. We can alter the shape of the bell curve by changing the mean and standard deviation.

We can quickly generate a normal distribution in Python by using the numpy.random.normal() function, which uses the following syntax: **numpy.random.normal(loc=0.0, scale=1.0, size=None)**

where: loc: Mean of the distribution. Default is 0.

scale: Standard deviation of the distribution. Default is 1.

size: Sample size.

Some excellent properties of a normal distribution:

* The mean, mode, and median are all equal.
* The total area under the curve is equal to 1.
* The curve is symmetric around the mean.

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

**Answer**: We often encounter missing values while we are trying to analyze and understand our data. It is very common in real-world data. There will be missing values because the data might be corrupted or some collection error. Missing values can cause bias and can affect the efficiency of how the model performs. There are many ways in which we can handle missing data.Imputation is the process of replacing missing values with substituted data. It is done as a pre-processing step.

If the data is numerical, we can use mean and median values to replace else if the data is categorical, we can use mode which is a frequently occurring value.

Other methods include Class label based imputation, Model based imputation. The methods to handle sometimes can be general/intuitive and can also depend on the domain.

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

**Answer:** Like any type of scientific testing, A/B testing is basically statistical hypothesis testing, or, in other words, statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics. A/B testing derives its power from random sampling.

When we conduct an A/B test (or multivariate), we distribute visitors randomly amongst different variations. We use the results for each variation to judge how that variation will behave.

**when we get some A/B testing results, we should check the following:**

* If sample size per variation is enough. The results that we get on a small sample size will have no relevance.
* If the number of conversions is enough. It should be at least 100, it is better to be around 200-300.
* How long the test runs.
* What is the margin of error (if testing engine provides this information). The smaller the margin of error the more accurate result we get.

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

**Answer:** It is definitely a terrible idea to miss a data. A very simple, and in many ways appealing, method devised to overcome these problems is mean imputation. While it’s true the mean doesn’t change by implementing this method of correction, the relationships with other variables do. And while the sample size remains at its full value, the standard error of that variable will be vastly underestimated–and this underestimation gets bigger the more missing data there are. Too-small standard errors lead to too-small p-values, leads to reporting results that should not be there.

There are other options to this. Multiple Imputation and Maximum Likelihood both solve these problems.

**14) What is linear regression in Statistics?**

**Answer:** Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. We then estimate the value of X (dependent variable) from Y (independent variable).

We can perform the linear regression method in a variety of programs and environments, including:

* R linear regression
* MATLAB linear regression
* Sklearn linear regression
* Linear regression Python
* Excel linear regression

**15) What are the various branches of statistics?**

**Answer:** The two main branches of statistics are [descriptive statistics](https://explorable.com/descriptive-statistics) and [inferential statistics](https://explorable.com/inferential-statistics). Both of these are employed in scientific analysis of data and both are equally important for the understanding of statistics.

* **Descriptive Statistics**

[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**

[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.