**Statistics– WORKSHEET 4**

**Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.**

Q1. Bernoulli random variables take (only) the values 1 and 0.

a) True b) False

Ans. True

Q2. 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

Q3. 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

Q4. 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

Ans. D

Q5. \_\_\_\_\_\_\_\_\_\_ random variables are used to model rates.

a) Empirical b) Binomial c) Poisson d) All of the mentioned.

Ans. C

Q6. Usually replacing the standard error by its estimated value does change the CLT.

a) True b) False

Ans. B

Q7. 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

Q8. 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

Q9. 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

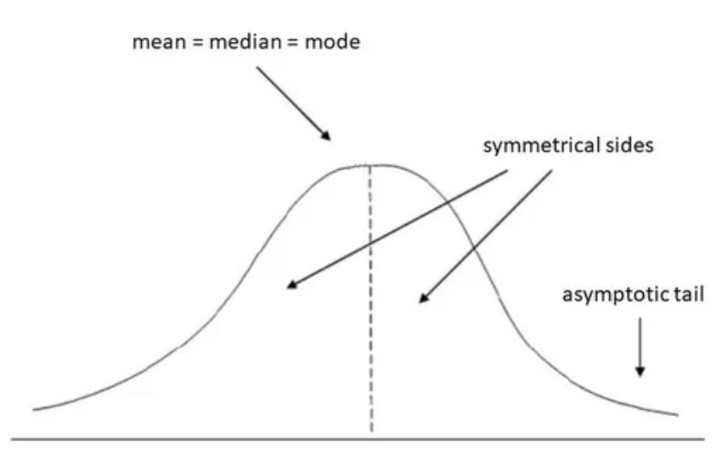
**Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.**

Q10. What do you understand by the term Normal Distribution?

Ans. The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the centre is a mirror image of the left side. The area under the normal distribution curve represents probability and the total area under the curve sums to one.

Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon (i.e. x-axis).

For a perfectly normal distribution the mean, median and mode will be the same value, visually represented by the peak of the curve.



The normal distribution is often called the bell curve because the graph of its probability density looks like a bell. It is also known as called Gaussian distribution, after the German mathematician Carl Gauss who first described it.

The normal distribution is the most important probability distribution in statistics because many continuous data in nature and psychology displays this bell-shaped curve when compiled and graphed. For example, if we randomly sampled 100 individuals we would expect to see a normal distribution frequency curve for many continuous variables, such as IQ, height, weight and blood pressure.

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

Ans. An analysis is only as good as its data, and every researcher has struggled with dubious results because of missing data.

Imputation is the process of replacing the missing data with estimated values. Instead of deleting any case that has any missing value, this approach preserves all cases by replacing the missing data with a probable value estimated by other available information. After all missing values have been replaced by this approach, the data set is analyzed using the standard techniques for a complete data.

In regression imputation, the existing variables are used to make a prediction, and then the predicted value is substituted as if an actual obtained value. This approach has a number of advantages, because the imputation retains a great deal of data over the listwise or pairwise deletion and avoids significantly altering the standard deviation or the shape of the distribution. However, as in a mean substitution, while a regression imputation substitutes a value that is predicted from other variables, no novel information is added, while the sample size has been increased and the standard error is reduced.

Q12. What is A/B testing?

A/B Testing is a tried-and-true method commonly performed using a traditional [statistical inference approach](https://en.wikipedia.org/wiki/A/B_testing) grounded in a hypothesis test (e.g. t-test, z-score, chi-squared test). In plain English, 2 tests are run in parallel:

1. Treatment Group (Group A) - This group is exposed to the new web page, popup form, etc.
2. Control Group (Group B) - This group experiences no change from the current setup.

The goal of the A/B is then to compare the conversion rates of the two groups using statistical inference.

The problem is that the world is not a vacuum involving only the experiment (treatment vs control group) and effect. The situation is vastly more complex and dynamic. Consider these situations:

* Users have different characteristics: Different ages, genders, new vs returning, etc
* Users spend different amounts of time on the website: Some hit the page right away, others spend more time on the site
* Users are find your website differently: Some come from email or newsletters, others from web searches, others from social media
* Users take different paths: Users take actions on the website going to different pages prior to being confronted with the event and goal

Often modeling an A/B test in this vacuum can lead to misunderstanding of the true story. The world is not a vacuum involving only the changes and effect. The situation is more complex.

Q13. Is mean imputation of missing data acceptable practice?

#### Ans. [Mean imputation](https://www.theanalysisfactor.com/em-imputation-and-missing-data-is-mean-imputation-really-so-terrible/) : Simply calculate the mean of the observed values for that variable for all individuals who are non-missing. It has the advantage of keeping the same mean and the same sample size, but many, many disadvantages. Pretty much every method listed below is better than mean imputation.

Substitution : Impute the value from a new individual who was not selected to be in the sample. In other words, go find a new subject and use their value instead.

Hot deck imputation : A randomly chosen value from an individual in the sample who has similar values on other variables. In other words, find all the sample subjects who are similar on other variables, then randomly choose one of their values on the missing variable. One advantage is you are constrained to only possible values. In other words, if Age in your study is restricted to being between 5 and 10, you will always get a value between 5 and 10 this way.

Another is the random component, which adds in some variability. This is important for accurate standard errors.

Cold deck imputation: A systematically chosen value from an individual who has similar values on other variables. This is similar to Hot Deck in most ways, but removes the random variation. So for example, you may always choose the third individual in the same experimental condition and block.

Regression imputation: The predicted value obtained by regressing the missing variable on other variables. So instead of just taking the mean, you’re taking the predicted value, based on other variables. This preserves relationships among variables involved in the imputation model, but not variability around predicted values.

Stochastic regression imputation: The predicted value from a regression plus a random residual value. This has all the advantages of regression imputation but adds in the advantages of the random component. Most multiple imputation is based off of some form of stochastic regression imputation.

Interpolation and extrapolation: An estimated value from other observations from the same individual. It usually only works in longitudinal data. Use caution, though. Interpolation, for example, might make more sense for a variable like height in children–one that can’t go back down over time. Extrapolation means you’re estimating beyond the actual range of the data and that requires making more assumptions that you should.

Q14. What is linear regression in statistics?

Ans. 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.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

### Q15. What are the various branches of statistics? Ans. **Descriptive Statistics and Inferential Statistics:** Every student 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:** [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.