WORKSHEET-1 STATISTICS

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.

Ans. a) 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?

Ans. a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans. b) Modeling bounded count data

4. Point out the correct statement.

Ans. d) All of the mentioned-

5. ________________random variables are used to model rates.

Ans. c) Poisson

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

Ans. b) False

7. Which of the following testing is concerned with making decisions using data?

Ans. b) Hypothesis

8. Normalized data are centered at______________and have units equal to standard deviations of the original data.

Ans. a) 0

9. Which of the following statement is incorrect with respect to outliers?

Ans. c) Outliers cannot conform to the regression relationship

Q10 to Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans. 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 graph form, normal distribution will appear as a bell curve.

A 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. For Example, pricing distributions are not perfectly normal throughout any financial year. The normal distribution model is motivated by the Central Limit Theorem. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

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

Ans. First we need to analyse and understand the nature of missing data, since it is critical in determining which method or treatment we have to apply/use to impute the missing data. Data can be missing in the following ways:

- **Missing Completely At Random (MCAR):** When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.
- **Missing At Random (MAR):** The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.
- **Not Missing At Random (NMAR):** When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

In the MAR & MCAR cases, it is safe to remove the data with missing values depending upon their occurrences, while in the MNAR case removing observations with missing values can produce a bias in the model. So we have to be really careful before removing observations. Note that imputation does not necessarily give better results.

The most common impute technique which we used are as follows:

I. **Mean or Median Imputation:** When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible. There may not be enough observations with non-missing data to produce a reliable analysis

   In predictive analytics, missing data can prevent the predictions for those observations which have missing data External factors may require specific observations to be part of the analysis. In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

II. **Random Forest:** Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

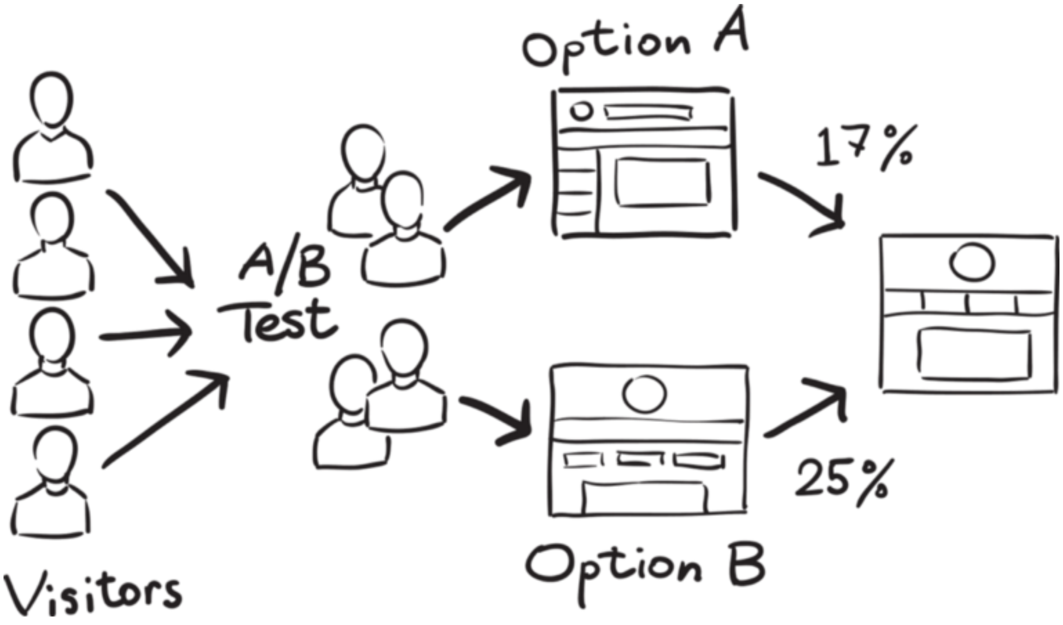
III. **Imputation by Chained Equations (MICE):** MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, Bayesian Polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data

12. What is A/B testing?

Ans. 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. An A/B test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

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

We 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, we try to decide which is performing better.



To put this in more practical terms, a prediction is made that Option B will perform better than Option A. Then, data sets from both pages are observed and compared to determine if Option B is a statistically significant improvement over Option A. 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.

The point of AB testing has absolutely nothing to do with how Option A or Option B will perform. We don't care about that. What we care about is how our page will ultimately perform with our entire audience.

For example, we have no way of knowing with 100% accuracy how the next 100,000 people who visit our website will behave. That is information we do not know today, and if we were to wait until those 100,000 people visited our site, it would be too late to optimize the user experience.

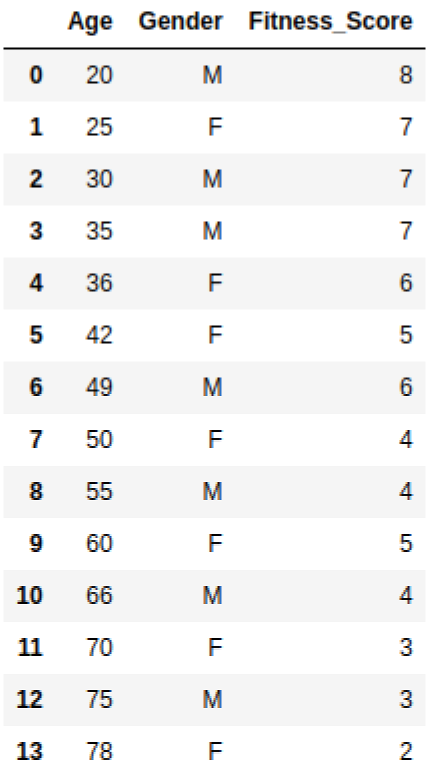
So, what we can do is observe the next 1,000 people who visit our site and then use statistical analysis to predict how the following 99,000 will behave/expect.

If we set things up properly, we can make that prediction with incredible accuracy, which allows us to optimize how we interact with those 99,000 visitors. This is why AB testing can be so valuable to businesses.

13. Is mean imputation of missing data acceptable practice?

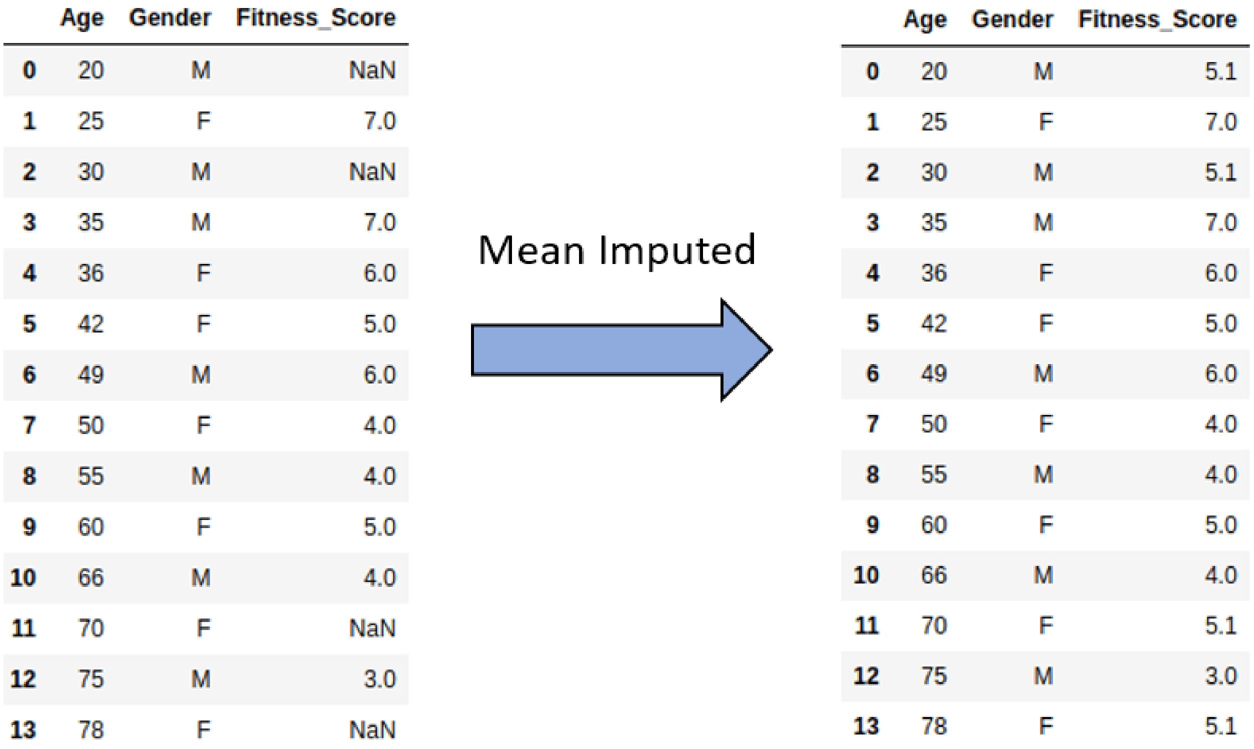
Ans. Firstly, we have to understand that there is no standard way to deal with the missing data. We have to use different methods depending upon which kind of problem we are solving, for example Time series, ML, regression etc. Mean imputation is a non-standard practice, but a fairly flexible imputation algorithm. It uses RandomForest at its core to predict the missing data. It can be applied to both continuous and categorical variables which makes it advantageous over other imputation algorithms. Mean imputation of missing values is not a recommended practice, If just estimating means then mean imputation preserves the mean of the observed data leads to an underestimate of the standard deviation Distorts relationships between variables by “pulling” estimates of the correlation toward zero.

For Example, Let’s have a look at a very simple example to visualize the problem. The following table have 3 variables: Age, Gender and Fitness Score. It shows a Fitness Score results (0–10) performed by people of different age and gender.



	Age	Gender	Fitness_Score
0	20	M	8
1	25	F	7
2	30	M	7
3	35	M	7
4	36	F	6
5	42	F	5
6	49	M	6
7	50	F	4
8	55	M	4
9	60	F	5
10	66	M	4
11	70	F	3
12	75	M	3
13	78	F	2

Now let’s assume that some of the data in Fitness Score is actually missing, so that after using a mean imputation we can compare results using both tables.



	Age	Gender	Fitness_Score
0	20	M	NaN
1	25	F	7.0
2	30	M	NaN
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	NaN
12	75	M	3.0
13	78	F	NaN

Mean Imputed

	Age	Gender	Fitness_Score
0	20	M	5.1
1	25	F	7.0
2	30	M	5.1
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	5.1
12	75	M	3.0
13	78	F	5.1

In the above table we can observe that the Imputed values don’t really make sense. In fact, they can have a negative effect on accuracy when training our ML model. For example, 78 year old women now has a Fitness Score of 5.1, which is typical for people aged between 42 and 60 years old. Mean imputation doesn’t take into account a fact that Fitness Score is correlated to Age and Gender features. It only inserts 5.1, a mean of the Fitness Score, while ignoring potential feature correlations.

So we can say that mean imputation is not a standard acceptable practice but depending upon the type of problem and where we are imputing the mean data we can use the mean imputation.

14. What is linear regression in statistics?

Ans. 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 modeller might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeller should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form **Y = a + bX**, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

15. What are the various branches of statistics?

Ans. The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for 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. 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, 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 of the future and generalizations 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 that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, 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.