STATISTICS WORKSHEET 1

Ans 1). (b) Bernoulli Random variables take only the values 1 and 0 is FALSE as it takes any value in the range (0,1) for success , I.e. p ∈(0,1) and so, failure or q which is p-1 can also take any value between 0&1.

Ans 2). (a) Central Limit Theorem states that if a distribution is normal , then the distribution of the samples or sample distribution is always normal irrespective of sample size.

In case of non-normal distribution, if the sample size is adequate (approx. >30 samples), then this sample distribution also becomes normal.

Ans 3). (b) As the basic assumption of Poisson Distribution is that it gives discrete probability distribution of the events which are independent of each other.

So, the statement 'Modeling bounded count data' is irrelevant in case of Poisson Distribution as the count data should be Unbounded to be analysed by Poisson Distribution (or somewhere, it can be weakly bounded but can't be properly Bounded).

Ans 4). (d) All of the above mentioned.

The exponent of a normally distributed random variables follow what is called Log-normal distribution. I.e. if x is log-normally distributed then y=ln(x) has a Normal distribution. Log-normal distribution is also called Galton's distribution.

The sum of normally distributed random variables are again normally distributed if the variables are dependent.

The squares of normally distributed variables follow a chi-square distribution.

Ans 5). (c) Poisson random variables are used to model rates.

Ans 6). (b) CLT is Computational Learning Theory. The main points on which CLT emphasis on (i) Sample Complexity, (ii) Computational Complexity and (iii) Mistake Bound.

$SE_x = \sigma/\sqrt{n}$ where SE is standard error.

Usually replacing the standard error by its estimated value does not change the CLT

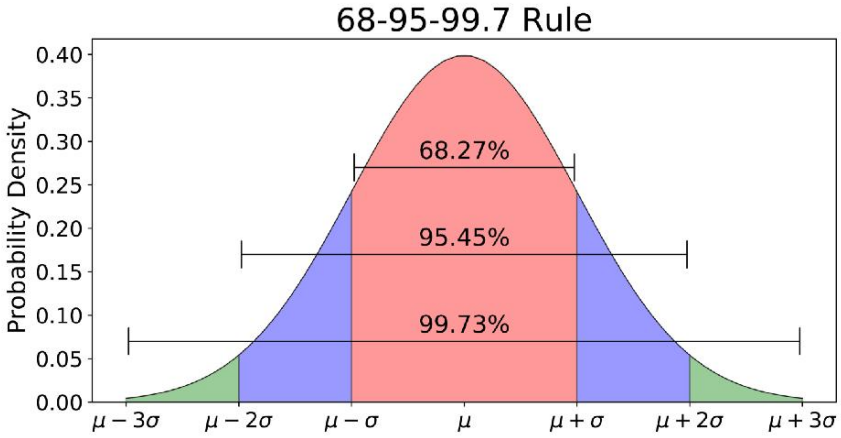
Ans 7). (b) Hypothesis testing is concerned with making decisions using data. In this testing, we make H0 as null hypothesis which is assumed to True and we check for Alternate Hypothesis (Ha) that whether we can reject it or not.

Ans 8). (a) Normalized data are centered at 0 and have units equal to standard deviations of the original data.

Ans 9). (c) The statement Outliers cannot conform to the regression relationship is incorrect as it is very well known that outliers can conform to any relationship, either it is regression or classification.

Ans 10). **<u>Normal Distribution</u>** :

Normal Distribution is also called Gaussian Distribution. It is of Bell shape. It is a probability distribution which is centered about mean and its data is mainly confined near the mean and very less spread far from mean. 68% of data resides in its 1 standard deviation of mean, 95%of data resides in 2 standard deviations of mean and 99.7% data resides in 3 standard deviations of mean. Mean, Median and Mode are equal in this distribution.



Normal Distribution is symmetrical but it is not necessary that every symmetrical distribution is normal. It can be converted into Standard Normal distribution by subtracting mean from its values and dividing them by standard deviation, I.e. Z = (x- μ)/ σ. The mean of Standard Normal Distribution is zero and its standard deviation is 1.

Central Limit Theorem plays an important role in describing sampling distribution of any distribution. It states the same as described in Answer no. 2 above. Then the Mean of sample mean = Population mean. Also, Stdd deviation of sample = Stdd deviation of population/ square root of sample size.

If one of the tail becomes longer than the other, then the distribution is called Skew. It is of two types: Left-skewed distribution and Right-skewed distribution.

So, if some distribution is not normal, then filling NAN values, removing outliers, encoding etc. steps are done for transforming it into Normal Distribution.

Ans 11). <u>Handling missing data :</u>

Let us say that df is a data frame which holds the required data to be interpreted. First of all, we should check if there is any missing data. There are a lot of ways to check this. Some are mentioned below:

1). One is **<u>df.describe()</u>**: it tells that the count values of all the columns of df. So, u can compare these count values with the no. of rows (given by df.shape). If the count value of any column is less than that of the row value, it means that column is having some missing data.

2). Second is **<u>df.isna().sum() </u>:** this command gives the total number of missing/ NAN values w.r.t. each column of df. So, on getting this, some imputation techniques are done to fill this missing data.

Now, once missing data is found, there are lot of ways to handle this which are described below:

1). **Df.dropna():** firstly, check if some rows or columns having all the nan values/none values associated with them , it's better to delete that entire row/column instead of replacing or filling this.

2). **Df.fillna()**: secondly, if we want to fill these nan values, we can use fillna command. The parameters of this command are the values with which the missing data is to be replaced. If a particular column is having continuous values, then its missing data is replaced by mean of all the other values of that column.

as df['col1'].fillna(df['col1'].mean(), inplace = True)

And if a particular column is having discrerte data/ values, then its missing data is filled/ replaced by mode of all the values of that column.

as df['col2'].fillna(df['col2'].mode[0], inplace = True)

3). **Imputation techniques**: there are a lot of imputation techniques to fill the missing data. Some of these are listed below:

(a) **Simple Imputer:** it is defined in sklearn.impute. it's simply fill the missing data by mean, median, most_frequent or costant values, whichever we write in its strategy parameter. For mean, it simply fills data with the mean of that column. For median, it simply fills missing data with the median of that column. mean and median both can be used only for numeric data. Most_frequent can be used for numeric as well as string data. It fills the missing data with the most frequent data item of that column, means with its mode value. For constant, it fills the missing data with the constant value which we provided.

(b) **KNN Imputer:** it is also defined in sklearn.impute. Imputation for completing missing values using k-Nearest Neighbors.

Each sample's missing values are imputed using the mean value from n_neighbors nearest neighbors found in the training set. e.g. if there are four persons having salaries 100,120,180,200 and their corresponding experience is 4,5,nan,6 respectively. Then if we want to impute by knn imputer, with n_neighbors = 2 then it will check that the person with salary 180 (who has missing data in experience) has 2 neighbors person with salary 120 and 200. So, the experience will be calculated by knn imputer is (5 +6)/2 = 5.5, which on the other hand comes out to be 5 if we take fillna.mean or simple imputer. So, this is the basic working of knn Imputer.

(c) Iterative Imputer: it is defined in sklearn.impute. But for this, firstly, from sklearn.experimental enable_iterative_imputer should be done, means it should be enabled first.

It works in the principle of Round-robin-fashion. Here, the column having missing value will act as label and other column(s) act as feature(s). So, the predicted value for NAn value will be filled at the missing place after prediction from algorithm. Here, estimator is to be specified. At each step, a feature column is designated as output y and the other feature columns are treated as inputs X. A regressor is fit on (X, y) for known y. Then, the regressor is used to predict the missing values of y. This is done for each feature

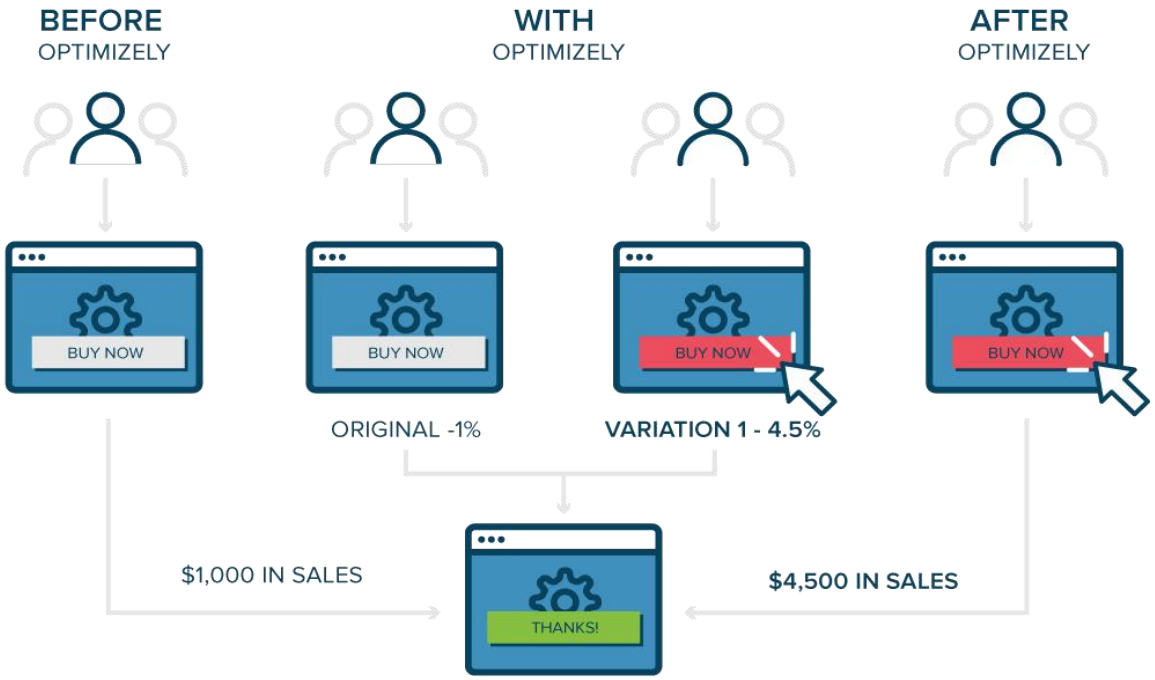
in an iterative fashion, and then is repeated for max_iter imputation rounds. The results of the final imputation round are returned.

Ans 12). A/B Testing:

It is also called Split Testing or Bucket Testing. It is the most popular way for Businesses to test new UX features, new versions of a product, or an algorithm to decide whether your business should launch that new product/feature or not. It is a method of comparing two versions of a webpage or app against each other to determine whichever performs better.

How does it work:

In business scenario, the variated version of the product is shown to some sample of customers, it's called Experimental group and the original version is shown to other group of customers. It's called Control group. Then the difference in the product performance is noted in both the cases and make the decision if the new version of the product has any positive impact on the customers and hence improving performance or not. So, in this way, the business firm gets the idea that how will this new version will make impact on the revenue and let them decide that if they need to launch it or not. Similarly, in the webpage or app case, two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.



so, as the above figure states that on applying some variation, there's an increase in the revenue of business. So, by A/B testing, we come to conclusion here that the new version of produvt/ app etc. Has positive impact on the market. Hence, can be produced further.

Benefits of A/B testing

- Allows to learn what works and what doesn't in a quick manner
- You get feedback directly from actual/real product customers
- Since the users are not aware that they are being tested, the results will be unbiased

Demerits of A/B testing

- Presenting different content/price/features to different customers especially in the same geolocation might potentially be dangerous resulting in **Change Aversion.**
- Requires a significant amount of Product, Engineering, and Data Science resources
- Might lead to wrong conclusions if not conducted properly

Common A/B test metrics

Popular performance metrics that are often used in A/B testing are the Click Through Rate, Click Through Probability and Conversion Rate.

1: Click-Through Rate (CTR) for usage

$$CTR = \frac{\#\ total\ clicks * 100\%}{\#\ total\ clicks + \#\ total\ views}$$

where the number of total views or sessions is taken into account. This number is the percentage of people who view the page (impressions) and then actually click on it (clicks).

2: Click-Through Probability (CTP) for impact

$$CTP = \frac{\#\ people\ with\ at\ least\ 1\ click * 100\%}{\#\ number\ of\ unique\ visitors\ per\ page}$$

Unlike the CTR, the CTP does take into account the duplicate clicks which means that if a user for some reason has performed multiple clicks, in a single session, on the same item for some reason (e.g. because of impatience), this multiple clicks is counted as a single click in CTP.

3: Conversion Rate

Conversion rate, defined as the proportion of sessions ending up with a transaction.

$$CR = \frac{\#converted}{\#converted + \#notconverted}$$

So, CTR can be used if it needs to measure the usability of the site and use CTP is used if it needs to measure the actual impact of the feature. CTR doesn't take care of the duplicate clicks, so if the user has impatiently pushed the same button multiple times, this will not be corrected to be equal to 1.

Its implementation in statistics:

Hypothesis Making:

H0 or null hypothesis: it assumes that there's no change in conversion rate before and after adding some variations in the product.

Ha or Alternate Hypothesis: It challenges the null hypothesis.

Now, on getting the result, let us firstly check if the results are affected by any error like Type1 error or Type 2 error. To check this, power analysis is done in which 'beta' specifies the probability of the type 2 error, (1-beta) is the power of A/B testing, alpha is the probability of type 1 error and delta is minimum detectable effect. All of this is done in Power analysis part. Type 1 error refers to the situation in which Ho is rejected but it is true otherwise and type 2 error refers to the situation of not rejecting H0 when it is false. Alpha is also called significance level. It's common practice to take power of A/B testing as 80% means 20% chances of type 2 error are fine. Generally, normal permissible value of alpha is 5%.

The most popular parametric tests that are used in A/B testing are:

A). 2 Sample T-test (when N < 30, metric follows student-t distribution, and you want to identify whether there exist a relationship and the type of relationship between control and experimental groups)

B). 2 Sample Z-test (when N > 30, metric follows asymptotic Normal distribution and you want to identify whether there exist a relationship and the type of relationship between control and experimental groups)

The most popular non-parametric tests that are used in A/B testing are:

A). Fishers Exact test (small N and if it needs to identify whether there exists a relationship between control and experimental groups)

B). Chi-Squared test (large N and if needs to identify whether there exists a relationship between control and experimental groups)

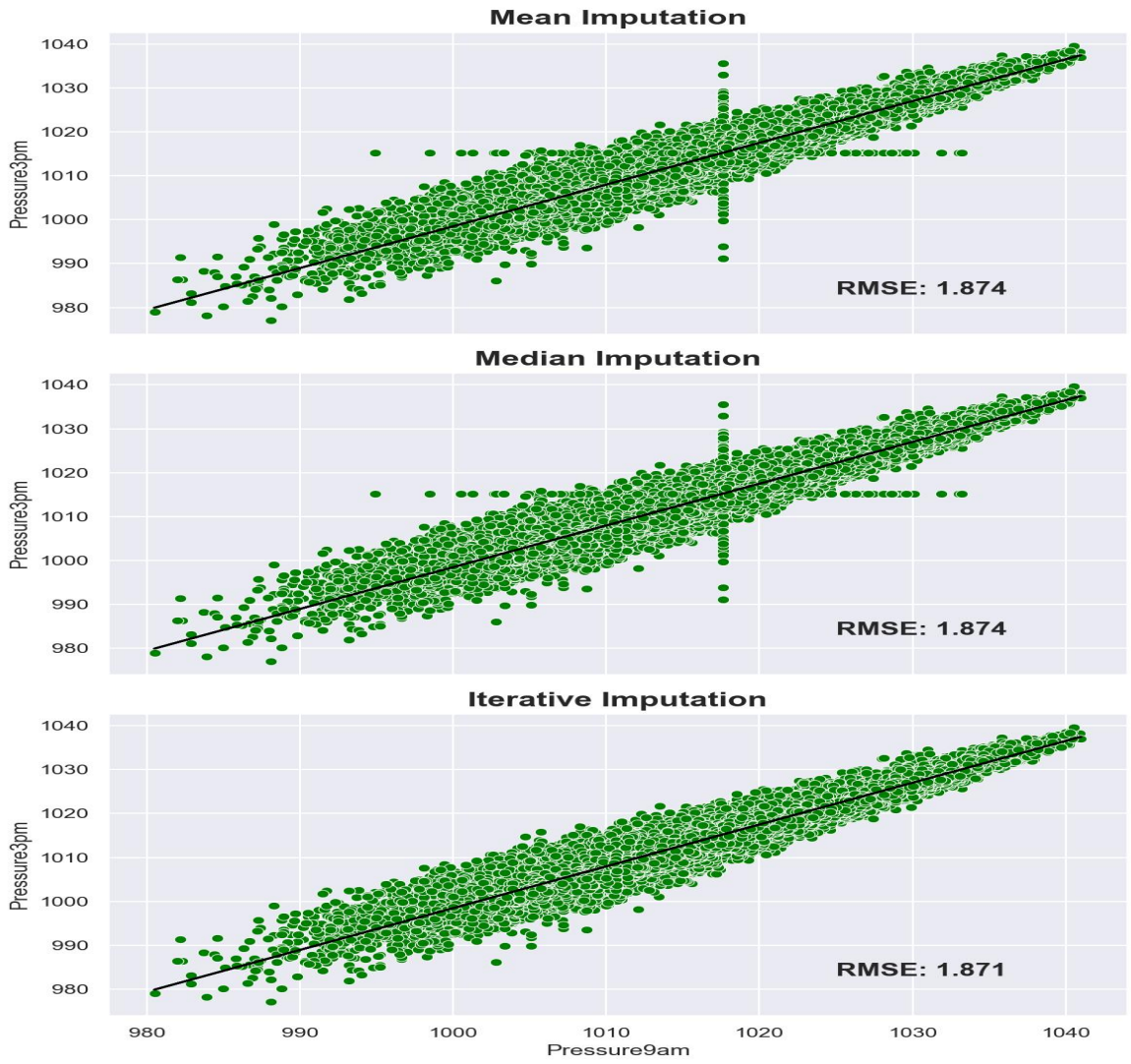
C). Wilcoxon Rank Sum/Mann Whitney test (small N or large N, skewed sampling distributions, testing the difference in medians between control and experimental groups).

Ans 13). **Mean Imputation:**

Yes, it is an acceptable practice but there are some restrictions under which it looses its importance. Generally, mean imputation is better in case of Univariate data but in multivariate data, advanced techniques like iterative imputer plas a better role. This can be described better by following article:

While creating a next-day prediction model for rain in Australia, Using the weatherAUS.csv file from the source dataset, the continuous features are imputed below using three different strategies:

1. Mean imputation
2. Median imputation
3. Iterative imputation



Pressure9am and Pressure3pm features are compared as they are directly related to one another and exhibit a linear relationship which will be useful for evaluation purposes.

On applying mean imputation , median imputation and iterative imputation, we got the following plots of the data along with a regression line, and then displays the root mean squared error (RMSE, lower is better).

There are a couple of things to keep in mind:

The amount of missing data in the Pressure9am and Pressure3pm features was only about 10%. As a result, the RMSE values can only improve so much compared to the mean and median strategies.

This comparison is only looking at two features while the full dataset contains many more. Small improvements across each of these features can lead to a large improvement overall when using all of the data in the modeling process.

Conclusion

Simple imputation strategies such as using the mean or median can be effective when working with univariate data. When working with multivariate data, more advanced imputation methods such as iterative imputation can lead to even better results.

Ans 14). Linear Regression in Statistics:

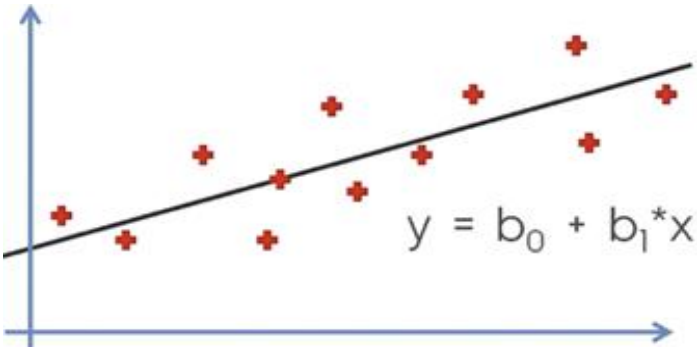
Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable to be predicted is called the dependent variable. The variable which is used to predict the other variable's value is called the independent variable.

It is a regression technique. It means the dependent variable should be continuous variable. It can't be a discrete type.

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 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. In brief, there are a lot of lines. The distance between the actual data points and the line is called Residual. So, the line which has the least residuals value is called Best Fit Line. The Gradiant descent is used to predict the 'best fit line' for this linear regression. It works on the principle The value of X (dependent variable) is then estimated from Y (independent variable). equation of linear regression is

$y = ax + b + e$

b is intercept, a is slope of line and e is error term.



How fit method works in this case:

Given datapoints (x,y), a and b should be chosen in such that the given line becomes a best fit line. the criteria for the best fit is that the sum of the squared differences between the data points and the line itself should be minimum.

I.e $\sum_{i=1}^{n} \text{square}(\text{ith } y - (a + b * \text{ith } x))$

Before performing linear regression, make sure that the data can be analyzed using this procedure. The data must pass through certain required assumptions.

Here's how these assumptions can be checked:

1. The variables should be measured at a continuous level. Examples of continuous variables are time, sales, weight and test scores.
2. Use a scatterplot to find out quickly if there is a linear relationship between those two variables.
3. The observations should be independent of each other (that is, there should be no dependency).
4. The data should have no significant outliers. Use boxplots to check if there are outliers. If they exist, try to delete them (if possible and required, otherwise leave them).
5. Check for homoscedasticity — a statistical concept in which the variances along the best-fit linear-regression line remain similar all through that line.
6. The residuals (errors) of the best-fit regression line follow normal distribution.

The accuarcy of Linear Regression is calculated by Rsquare value. Rsquare = (TSS - RSS)/TSS

Where TSS =Total sum of squares and RSS= Residual sum of squares.

Defined in Python:

Linear Regression is defined in sklearn.linear_model in the form LinearRegression

And rsquare is defined in sklearn.mterics in the form r2_score

And the errors are given by MAE, MSE and RMSE, explained in Answer no. 15 of Machine Learning Assignment - 39

Ans 15). Various Branches of Statistics:

Statistics is the branch of mathematics that deals with data.

Statistics is broadly classified into two categories:

A). Descriptive Statistics

B). Inferential Statistics

Before moving further, let's have a look on data and data collection. Data collection is all about how the actual data is collected. Sometimes, it is hard to collect the data as in the case of exit-poll after elections, it's not possible to ask from each n every voter about where they cast their vote. Another example is like counting the number of parrots in a city. As it's impossible to count them properly as some of them may

fly while counting, some of them may move to some other place etc. So, in these type of cases, the concept of sample and population comes into picture.

Population is the entire dataset like number of voters who cast their vote etc. And sample is a sub-part of population. it gives some fair idea about the population as it may be time-consuming and expensive to take data from entire population. Sample represents the small part of data where the statistics part can be done for the data analysis.

The branches of statistics are described below:

A). **Descriptive Statistics:** Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is the branch of statistics that focuses on collecting, summarizing, and presenting a set of data. In this part, the statistics is done on the whole population. Here, to find the mean, median n mode is called Measure of Central Tendency. And to find variance and standard deviation is called Measure of dispersion.

B). 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.

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.

During working on the samples rather than entire population, it is the Central Limit Theorem. It is explained in Answer no. 2 itself.

Pdf (Probability Density Function), Cdf (Cumulative Density Function), Hypothesis testing like student's ttest, ANOVA ets. all are statistical parts.