LOKESH BAVISKAR
8/17/21

STATISTICS WORKSHEET 1 (WORKSHEET SET 1)

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
   Answer: 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?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
   Answer: 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
   Answer: 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 variablesare dependent
   c) The square of a standard normal random variable follows what is called chi-squareddistribution
   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. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

Answer: b) False

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
   Answer: b) Hypothesis

8. 4. Normalized data are centered at___and have units equal to standard deviations of theoriginal data.
   a) 0
   b) 5
   c) 1
   d) 10
   Answer: a) 0

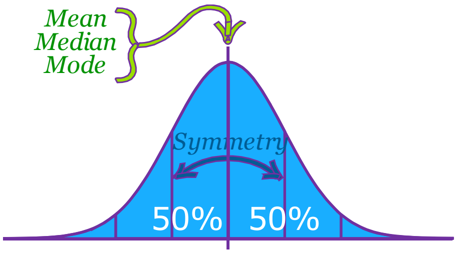
9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned
   Answer: c) Outliers 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?

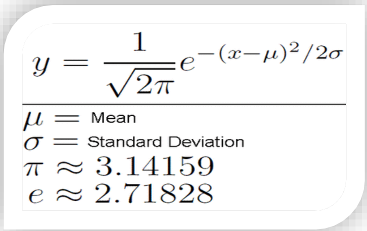
Answer**:**

- The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely.



- The distribution of data refers to the way the data is spread out. The distribution of a dataset shows us the frequency at which possible values occur within an interval.

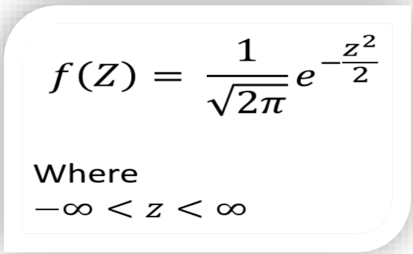
- *Mathematical Definition:*
  A continuous random variable “x” is said to follow a normal distribution with parameter μ(mean) and σ (standard deviation), if its probability density function is given by,



$$y = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation
$\pi \approx 3.14159$
$e \approx 2.71828$

- *Standard Normal Distribution:* The simplest case of the normal distribution, known as the Standard Normal Distribution, has an expected value of μ(mean) 0 and σ(std.) 1, and is described by this probability density function

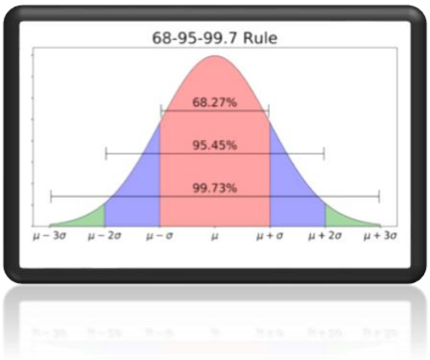


$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Where
$-\infty < z < \infty$

- *Distribution Curve Characteristics:*
  1) The total area under the normal curve is equal to 1.
  2) It is a continuous distribution.
  3) It is symmetrical about the mean. Each half of the distribution is a mirror image of the other half.
  4) It is asymptotic to the horizontal axis.
  5) It is unimodal.

- *Empirical Rule for Normal Distribution:*
  Almost all the data lies within 3 standard deviations. This rule enables us to check for Outliers and is very helpful when determining the normality of any distribution.



According to the Empirical Rule for Normal Distribution:

1. 68.27% of data lies within 1 standard deviation of the mean
2. 95.45% of data lies within 2 standard deviations of the mean
3. 99.73% of data lies within 3 standard deviations of the mean

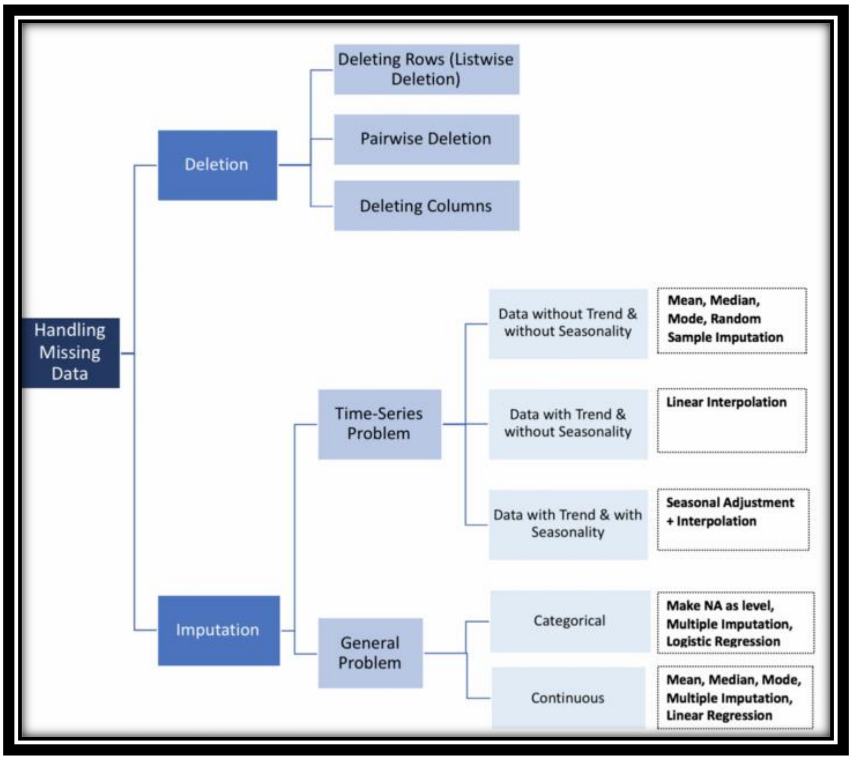
11. How do you handle missing data? What imputation techniques do you recommend?

Answer:

When we look at missing data, we come across two possibilities. Missing data can be totally random or there can be pattern behind missing data. Each of the two possibility need to handle in different way and need to address by different imputation techniques. Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Missing data is classified normally in three categories in following way:

1) *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.
2) *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.
3) *Not Missing at Random (NMAR):* When the missing data has a structure to it, we cannot treat it as missing at random.

Imputation methods are those where the missing data are filled in to create a complete data matrix that can be analyzed using standard methods. It is important to understand the nature of the data that is missing when deciding which algorithm to use for imputations. We can test the quality of your imputations by normalized root mean square error (NRMSE) for continuous variables and proportion of falsely classified (PFC) for categorical variables.


Handling Missing Data
Deletion
Deleting Rows (Listwise Deletion)
Pairwise Deletion
Deleting Columns
Imputation
Time-Series Problem
Data without Trend & without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend & without Seasonality
Linear Interpolation
Data with Trend & with Seasonality
Seasonal Adjustment + Interpolation
General Problem
Categorical
Make NA as level, Multiple Imputation, Logistic Regression
Continuous
Mean, Median, Mode, Multiple Imputation, Linear Regression

Deletion

It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings. Missing data can be deleted in three way i.e. listwise deletion, pairwise deletion, deleting column.

- **Listwise Deletion** - Listwise deletion (complete-case analysis) removes all data row wise for an observation that has one or more missing values. In this case assumptions of MCAR are typically rare to support. *As a result, listwise deletion methods produce biased parameters and estimates.*
- **Pairwise Deletion** - Pairwise deletion analyses all cases in which the variables of interest are present and thus maximizes all data available by an analysis basis. 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.
- **Deleting Columns**- A feature that has a high number of empty values is unlikely to be very useful for prediction. *If data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.* It can often be safely dropped. Dropping rare features simplifies your model, but obviously gives you fewer features to work with.

Time-Series Specific Imputation Methods

The time series methods of imputation assume the adjacent observations will be like the missing data. Time series data is characterized by trend and seasonality which need to consider while applying these techniques. However, these methods won't always produce reasonable results, particularly in the case of strong seasonality.

- **Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)**

These options are used to analyse longitudinal repeated measures data, in which follow-up observations may be missing. In this method, every missing value is replaced with the last observed value. Longitudinal data track the same instance at different points along a timeline. *However, this method may introduce bias when data has a visible trend.*

- **Linear Interpolation**

Linear interpolation is often used to approximate a value of some function by using two known values of that function at other points. The weights are inversely related to the distance from the end points to the unknown point. *When dealing with missing data, you should use this method in a time series that exhibits a trend line, but it's not appropriate for seasonal data.*

- **Seasonal Adjustment with Linear Interpolation**

When dealing with data that exhibits both trend and seasonality characteristics, use seasonal adjustment with linear interpolation. First you would perform the seasonal adjustment by computing a centered moving average or taking the average of multiple averages – say, two one-year averages – that are offset by one period relative to another. Then complete data smoothing with linear interpolation.

Continuous Variables Imputation Techniques:

There is different method to handle missing data for continuous variables.

- **Multiple imputation:**

Multiple imputation is considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point, the missing values are exchanged for values that encompass the natural variability and uncertainty of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures, and the multiple analysis results are combined to produce an overall result. *Multiple imputations can produce statistically valid results even when there is a small sample size or a large amount of missing data.*

- **Mean, Median, Mode Imputation-**

The simplest imputation method is replacing missing values with the mean or median values of the dataset at large, or some similar summary statistic. This has the advantage of being the simplest possible approach. Mean and Median are used to handle numerical missing data. Mode is used to replace categorical missing data by replacing null values with highest frequency occurring category variable i.e. mode of dataset. 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.

Imputation of Categorical Variables

1. Mode imputation is one method but it will definitely introduce bias
2. Missing values can be treated as a separate category by itself. We can create another category for the missing values and use them as a different level. This is the simplest method.
3. Prediction models: Here, we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable (training) and another one with missing values (test). We can use methods like logistic regression and ANOVA for prediction
4. Multiple Imputation

Other Imputation Techniques -

- **Regression imputation –**

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 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.

- **K Nearest Neighbors**

KNN is a machine learning algorithm which works on the principle of distance measure. This algorithm can be used when there are nulls present in the dataset. While the algorithm is applied, KNN considers the missing values by taking the majority of the K nearest values. In this method a distance measure for k neighbors, and the average is used to impute an estimate. KNN can identify the most frequent value among the neighbors and the mean among the nearest neighbors.

- **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.

12. What is A/B testing?

Answer:

- 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 Example, 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. 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.

- A/B testing is a form of statistical and two-sample hypothesis testing. **Statistical hypothesis testing** is a method in which a sample dataset is compared against the population data. **Two-sample hypothesis testing** is a method in determining whether the differences between the two samples are statistically significant or not.
- **A/B testing process -**

The following is an A/B testing framework you can use to start running tests:

1) *Make a Hypothesis -*

In hypothesis testing, we have to make two hypotheses i.e. Null hypothesis and the alternative hypothesis.

The **null hypothesis** is one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is no difference between the control and variant group.

The **alternative hypothesis** is one that states that sample observations are influenced by some non-random cause. From an A/B test perspective, the alternative hypothesis states that there is a difference between the control and variant group.

2) *Create your control group and test group -*

The next step is to create your control and test (variant) group. There are two important things to taken care in this step, random samplings and sample size.

Random sampling is a technique where each sample in a population has an equal chance of being chosen. Random sampling **eliminates sampling bias**, and lead A/B test to be representative of the entire population rather than the sample itself.

It is required that we determine the minimum sample size for our A/B test before conducting it so that we can **eliminate under coverage bias.** It is the bias from sampling too few observations.

3) *Conduct the test, compare the results, and reject or do not reject the null hypothesis* –

Determine if the difference between your control group and variant group is statistically significant. An experiment is considered to be statistically significant when we have enough evidence to prove that the result, we see in the sample also exists in the population. That means the difference between your control version and the test version is not due to some error or random chance.

The **two–sample t–test** is one of the most commonly **used** hypothesis **tests**. It is applied to compare whether the average difference between **the two** groups.

Set your alpha, the probability of making a type 1 error and determine the probability value (p-value). Lastly, compare the p-value to the alpha. **If the p-value is greater than the alpha, do not reject the null!**

13. Is mean imputation of missing data acceptable practice?

Answer:

The easiest way to impute is to replace each missing value with the mean of the observed values for that variable. imputing the mean preserves the mean of the observed data. So, if the data are missing completely at random, the estimate of the mean remains unbiased. Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. It is only reasonable idea if only few values are missing or there is small dataset size. But overall, it is not good idea to replace missing data with mean of dataset due to following reason:

- Mean of data is much more sensitive to outliers, so it can lead to underestimate or overestimate mean than actual mean of dataset if all data is present. Replacing missing values with median would be much more idea in that case.
- 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.
- Any statistic that uses mean from mean-imputed data that you would have gotten without the imputations will have a standard error that's too low. Because your standard errors are too low, so are your p-values. Now you're making Type I errors without realizing it.
- This strategy can severely distort the distribution for this variable, leading to underestimates of the standard deviation. Moreover, mean imputation distorts relationships between variables by "pulling" estimates of the correlation toward zero.

14. What is linear regression in statistics?

Answer:

- Regression analysis employs a model that describes the relationships between the dependent variables and the independent variables in a simplified mathematical form.
- In statistics, linear regression is a technique to model or find linear relationship between dependent and independent variable.
- When there is only one independent variable in the linear regression model, the model is generally termed as a simple linear regression model. When there are more than one independent variables in the model, then the linear model is termed as the multiple linear regression model.
- Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. There can be many line possible passing through data point but the best-fitting line is the line that minimizes the sum of the squared errors of prediction.
- In simple linear regression model equation of line is given by $y = ß_0 + ß_1x$

where y is termed as the dependent or study variable and x is termed as the independent or explanatory variable. The terms $ß_0$and $ß_1$ are the parameters of the model. The parameter $ß_0$is termed as an intercept term, and the parameter $ß_1$ is termed as the slope parameter. These parameters are usually called as regression coefficients.

- Although we minimize the sum of the squared distances of the actual y scores from the predicted y scores (y'), there is a distribution of these distances or errors in prediction which is important to discuss. We will define these directed (signed) distances (residuals) as e = (y-y'), where y' is our predicted value. Clearly both positive and negative values occur with a mean of zero.
- The square root of this value is the standard deviation and is known as the standard error of estimate.

15. What are the various branches of statistics?

Answer:

The field of statistics is divided into two major divisions: descriptive and inferential. Each of these segments is important, offering different techniques that accomplish different objectives.

(1) Descriptive statistics

Descriptive statistics is the part of statistics that deals with presenting the data we have. Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data. This can take two basic forms – presenting aspects of the data either visually or numerically.

Descriptive statistics is responsible for summarizing a statistical sample (set of data obtained from a population) Rather than learning about population Which represents the sample. Some of the measures commonly used in descriptive statistics to describe a set of data are the measures of central tendency and the Measures of variability or dispersion.

(2) Inferential Statistics

Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates. Most predictions of the future and generalization on a population study of a smaller specimen are in the scope of the inference statistics.

Different Techniques use to examine the relationships between variables and draw conclusion & predication in inferential statistics. Some of techniques are linear regression analyses, logistic regression analyses, ANOVA, correlation analyses, structural equation modeling, and survival analysis.