STATISTICS WORKSHEET-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

Explanation: Bernoulli distribution is the simplest case of Binomial distribution which takes 1 and 0.

---

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

Explanation: As you increase the sample size. You are going to have a frequency plot that looks very close to normal distribution.
----

3. Which of the following is incorrect with respect to the use of Poisson distribution?
a) Modeling event/time data
b) Modeling bounded count data
c) Modeling contingency tables
d) All of the mentioned

Answer: (B) Modeling bounded count data

Explanation: Poison Distribution is a probability distribution that is used to show how many times an event can occur in a given period of time.

----
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

Answer: (D) All of the mentioned

Explanation:

----
5. _______ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

Answer: (C) Poisson

Explanation: Poisson distribution shows how many times an event has occurred in a specific time means a rate no. of times an event occur / time

----

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 Testing

Explanation: Null Hypothesis is assumed to be True and we have to either accept Null Hypo or Reject Null Hypo based on the p values.

----

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

Answer: (A) 0

Explanation: Normalized data centered at 0 with a standard deviation of 1.

----

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

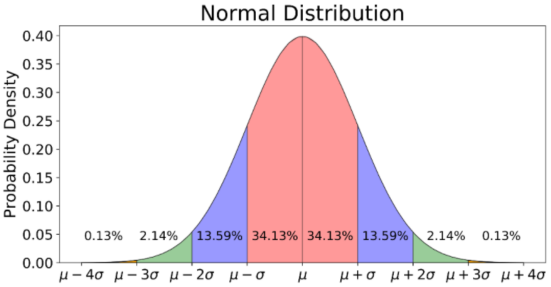
Answer: (C) Outliers can’t conform to the regression relationship

Explanation: Outliers are

WORKSHEET

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.
10. What do you understand by the term Normal Distribution?

Explanation: Normal Distribution, also called the Gaussian Distribution. Normal Distribution can be seen in very natural phenomena. It has been developed as a standard of reference for many probability problems.
Many things actually are normally distributed or very close to it. Example heights of people, Blood pressure, Salaries are normally distributed.
In Normal Distribution, the mean, median, mode all line up such that the centre of the distribution is the mean. Because of this .exactly half of the result falls to either side of the mean. The normal Distribution shape is likely to be Bell Curve. We usually work with the standardized normal distribution where μ =0 and σ=1.



This is the diagram of Normal Distribution which describes how the values of a variable are distributed and almost all values (99.7%) lie within 3 σ (Standard Deviation) of the mean as per Empirical rule.

Unfortunately, population parameters are unknown because it is generally impossible to measure an entire population. However, we can use random samples to calculate estimates of these parameters. For Sample, the sample mean is $\overline{x}$ and the Standard deviation is ‘s’.
Properties of Normal Distribution:

A. Symmetric bell curve
B. Mean, Median and Mode are all equal
C. Empirical rule allows determining the proportion of values that fall within certain distances from the mean.

While the Normal Distribution is essential in statics, it is just one of many probability distributions and doesn’t fit all populations.

---

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

Explanation: Missing data or missing values, occur when no data value is stored for the feature in an observation that is missed. It reduces the statistical power of the analysis which affect the model prediction.
There are mainly 3 types of missing data.

1. MCAR (Missing completely at random): a variable is missing completely at random. There is no relationship between the missing data and any other observed values.
   If values for observations are missing completely at random, then disregarding those cases would not bias the interferences.
2. MNAR (Missing not at random): these are systematic missing values. There is absolutely some relationship between the data missing and other values.
3. MAR (Missing at Random): Due to hesitation or embarrassment. Like someone didn’t mention his salary.

Handling Missing data is completely depends on the nature of a variable. Below are some techniques:

a. **Mean/Median/Mode replacement**: Mostly used for MCAR missing values.
   Mean is the average, Median is the central value, Model is the frequently used input for the variable.
   Mean is highly affected by outliers so if we have outliers in the feature, we can replace the NaN values by Median or Mode.
   We can compare the standard deviation of the feature with NaN and after NaN replacement. SD should not be changed much.

   Advantage:
   1. Easy to implement

2. Faster way to obtain the complete dataset

Disadvantage:

1. Change in the original variance, standard variation
2. It impacts the correlation of features with dependent and independent variables.

b. **Random Sample Imputation:** Mostly used for MCAR missing values.
It takes the random observation from the feature and uses this observation to fill NaN within the feature with missing values.
Advan: Easy to implement and variance distortion less happens
Dis-advan: Randomness won't work in every situation.

c. **Capturing NaN values with a new feature:** It works well when Data is MNAR. Because here NaN has some relationship with other variables. We have to capture all NaN importance here.
Let's suppose F1 has some missing values, we will create a new feature (f_nan) as a way to have NaN replaced by 1 and other records as 0. This is to capture the importance of records having Missing values. Now we can replace NaN values in feature F1 by mean or median because we have already captured the importance.

Advan: Easy to implement ad captures the importance of missing values
Dis-advan: Creating an additional feature (Curse of Dimensionality)

d. **End of Distribution Imputation:** we will make a distribution graph of the variable and find out the values which are far away from the mean. Value from the end of the distribution.
We will take a value after 3 Standard Deviation.
Extreme= df.Age.mean() + 3* df.Age.std()
Fill NaN by Extreme value
Advan: Easy to implement and Capture the importance of missingness
Dis-advan: Distorts the original distribution of the variable.

e. **Arbitrary Value Imputation:** This is useful for both numerical and categorical features. Data should not (MAR)

For categorical Variables NaN can be replaced by an additional label called 'Missing' which is very common practice.
For Numerical Variable: Nan can be replaced within a variable by an arbitrary value. Typically used 0,999,-999, combinations of 9 instead of replacing those occurences with the mean or the median.

Advan: Easy to implement and capture the missingness
Dis-advan: distortion of original distribution and hard to decide arbitrary value.

f. **Frequent Category Imputation:** This imputation is for the categorical features. For MAR.
Select the most frequent values and replace NaN with them. It is like the Mode of the column.
Advan: Easy to implement and faster way to implement
Dis-advan:

1. Higher the % of missing data, the higher will be the distortion
2. May lead to over-representation of a particular category.
3. Can distort the original variable distribution

g. **Forward filling and Back Filling:**
Forward filling means filling missing values with previous data and
Backward filling means filling missing values with the next data points.

h. Most easy and un-useful way to drop the missing values. If total missing values are 1-2% of the complete dataset.

12. What is A/B testing?

Explanation: A/B testing is a popular way to test a product over comparison. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. For instance, It allows decision-makers to choose the best design for a website by looking at the analytics results obtained with the two best alternatives A and B.

Divide the product into 2 parts A and B, Here A will remain unchanged while you make significant changes in B's Packing. Now on the basis of the response from customer groups who used A and B respectively. You try to decide which is performing better.

It is a hypothetical testing method for making decisions that estimate population parameters based on sample statistics.

Population refers to all the customers buying your product while sample refers to the no. of customers that participated in the test.

How does A/B Testing work?

Let's say there is an e-commerce company XYZ. It wants to make some changes in its newsletter format to increase the traffic on its website. It takes the original newsletter and mark it with A and thereafter, make some changes in Language, style and marks it with B.

We will use A/B testing and collect data to analyze which newsletter performs better.

Let's take Null Hypothesis H0: There is no difference between newsletter A and B

While the Alternate Hypothesis says, yes there is a difference and newsletter B has a higher conversion rate.

Through A/B testing, we have to collect enough evidence to reject the Null Hypothesis.

Next, decide the group of customers that will participate in the test like Group 1 and Group 2.

The test is to calculate the daily conversion rate for both groups for 1 month. Now, take the mean conversion rate for both groups. It will decide which newsletter is performing well.

When to perform A/B testing:

A/B testing works best when testing incremental changes such as new features, ranking and page load time. Here we can compare pre and post-modification results to decide whether the changes are working as desired or not.

13. Is mean imputation of missing data acceptable practice?

Explanation: Mean imputation means filling NaN values of a feature with the mean of non-missing observations.

Mean Imputation is so simple and yet, so dangerous

It's a very popular technique to handle missing data but it should be a last resort.

Because:

Mean imputation reduces the variance of the imputed variables:

Let's have a feature f1 which have missing values and now instead of filling NaN with Mean, create a new feature (f_nan) with missing values replaced by the mean.

Compare both the features now.

F1 and f_nan both have the same mean values but the standard deviation (variance) of the f_nan is smaller because all missing values are centred to mean now.

Mean imputation shrinks standard errors, which invalidate most hypothesis tests and the calculation of confidence interval.

As we have seen already, an imputed variable always has a smaller variance than the original variable. The estimated variance is used to compute many other statistics, which are also shrunk like

1. Standard error of the mean
2. Confidence interval will be shorter based on mean-imputed data
3. The standard t-test for a mean uses the standard error to compute a p-value for the null hypothesis. If the standard error is shrunk by mean imputation, then the standard one-sample t-test is not valid and the p-value is too small. We will potentially reject a null hypothesis that might be true (a Type I error)

Mean imputation doesn’t prevent the relationship among variables such as correlation.
As mean imputation affects univariate statistics, it also distorts multivariate relationship and affect statistics such as correlation.

The conclusion is mean imputation should be avoided when possible.

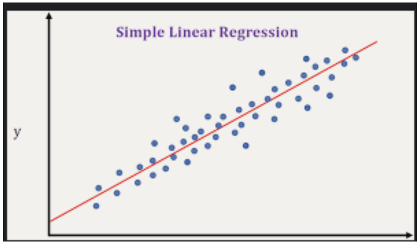
----

14. What is linear regression in statistics?

Explanation: The term ‘regression’ generally refers to predicting a real number. However, it can also be used for classification as well (predicting a category or class). Linear Regression is a basic and commonly used type of predictive analysis.
A Linear combination is an expression where one or more variables are scaled by a constant factor and added together. It reflects the relationship between x and y in the dataset. Then, with the help of the model, we can predict the value of Y at a given X.
The simplest form of the regression equation with one dependent and one independent variable is defined by a formula
y=c + m*x +e

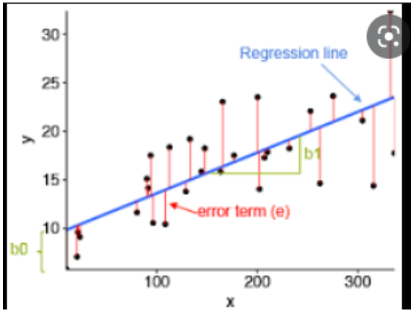
Where y=dependent variable, c=intercept, m=regression coefficient, x= independent variable

Let me try to describe it in simple words. We plot a graph between independent and dependent variables.



Linear Regression Model will try to make the best fit line through the data points. Either we would have 1 dependent variable and many independent variables y= c+m1x1+m2x2+m3x3+…….. mnxn + e.

The Model is designed to minimize the square error (e) to find the best fit line.



y=c + m*x +e
Error is the difference between the actual value and the predicted value. And the goal is to reduce this difference.
Error= actual value- predicted value
Sum of error= sum(actual – predicted)

Square of sum of error= (sum(actual-predicted values))2

SSE= $\sum \quad (y - y^{\wedge})2$

Bottom line is to minimize SSE to get a best-fit line.

Linear Regression Properties:

1. LR assumes that all independent variables are truly independent of each other. Only influence the dependent variable. But in practice, it is not possible. We call it a co-linearity issue.
   Let's suppose we have f1, f2, and f3 -> independent and y -> output. As per the Linear Regression assumption, there should not be any correlation between f1 and f2, f2 and f3 but in practice, there is a relationship between input variables. We call it multicollinearity issue or Curse of Dimensionality reduction. This can be handled by the Dimensionality reduction technique called PCA(Principal Component Analysis) which combines 2- correlated input features and create another new feature.

2. Mathematically, we need to know the relation between all input variables along with output variables. So Correlation Pearson helps to find the correlation between 2 variables how closely their relationship follows a straight line.
3. Correlation -1 and +1 indicates a perfect linear relationship between X and Y whereas a correlation of 0 indicates the absence of a linear relationship.
4. Coefficient of correlation ® is -1 to +1 where -1 indicates the Negative strong correlation between X and Y and +1 indicates the strong positive correlation between X and Y.
5. Gradient Descent method used to find the best model. It uses partial derivatives on the parameters (slope and intercept) to minimize the sum of squared errors.
6. Coefficient of determinant- R^2 determines the fitness of the linear model. Closer the points to the line, R^2 tends to 1. R^2 lies 0 to 1.

→Before we build the model, we have to evaluate each independent variables and see the ® values to identify good predictors.
Thereafter, how much your Model is reliable, R^2 helps to understand, how much of the total variance in our 'Y' has been explained by our model. R^2 is best at 1.

R^2 vs Adjusted R^2

Every time we add an independent variable to a model, R^2 increases to capture the variance even if the independent variable is insignificant. Whereas Adjusted R^2 increases only when the independent variable is significant and affects the dependent variable.
Adjusted R^2 will always be less than R^2.

Linear Regression Assumptions:

1. Assumption of Linearity: LR assumes that there is a linear relationship between the dependent variable and independent variables.
2. Assumption of normality of the error distribution: Errors should be normally distributed across the model
3. Very low or no Multicollinearity: no relationship between independent variables.
4. No auto-correlation of errors/residuals:
5. Homoscedasticity: The variation of errors across each of the independent variables should remain constant.
6. All observations are independent.

----

15. What are the various branches of statistics?

Explanation: Statistics is a study of Presentation, analysis, collection, interpretation and organization of data.
There are 2 main branches: a. Descriptive Statistic b. Inferential Statistic

A. **Descriptive Statistics**: It is the first part of statistics that deals with the collection of data. So simply we can say that it describes the data. We do analysis to understand the data with some statistical terms like, mean, median mode, Measure of central tendency. These measures help statisticians to analyze the distribution of data from a specific dataset.
   Some Tools and techniques:
   Bar Plot, Histogram, Pie-chart
   PDF, CDF
   Normal Distribution
   Measure of Central Tendency
   Mean, Median, Mode
   Measure of Variance
   Standard Deviation

B. **Inferential Statistics**: Inference statistics are a technique that enables statisticians to use the information collected from the sample to conclude, bring decisions or predict for the entire population.
Inferential statistics often speak in terms of probability by using Descriptive statistics. Besides, a statistician uses these techniques for data analysis, drafting, and making conclusions from limited information.
Most future predictions and generalizations on a smaller specimen population (sample) study are in the inference statistics scope.

Some Tools and techniques:
Regression analysis
Analysis of variance (ANOVA)
t-test
Correlation analysis
Confidence interval