STATISTICS WORKSHEET-1

Question No.1:- Bernoulli random variables take (only) the values 1 and 0.

a) True

b) False

Answer:- (A) True

Question No.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.

Question No. 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:- (B) Modeling bounded count data.

Question No. 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:- (C) The square of a standard normal random variable follows what is called chi-squared distribution.

Question No.5:- _____ random variables are used to model rates.

(A) Empirical

(B) Binomial

(C) Poisson

(D) All of the mentioned

Answer:- (C) Poisson.

Question No. 6:- Usually replacing the standard error by its estimated value does change the CLT.

(A) True

(B) False

Answer:- (B) False

Question No.7:- 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

Question No. 8:- 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

Question No.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

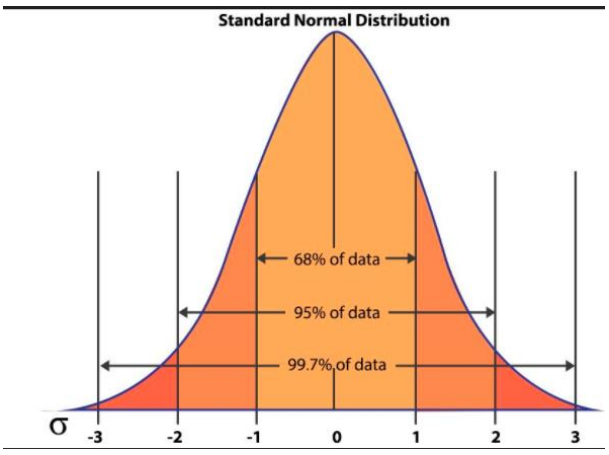
Answer:- (C) Outliers cannot conform to the regression relationship.

Question No. 10:- What do you understand by the term Normal Distribution?

Answer:- Normal Disribution:-" A probability function that specifies how the values of a variable are distributed is called the normal distribution."

The normal distribution is the most common type of distribution assumed in tecchnical common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: The mean and the standard deviation.

In a normal distribution, data is symmetrically distributed with no skew. Most values cluster around a centeral region, with values tapering off as they go further away from the center.



The simplest case of a normal distribution is known as the *standard normal* distribution or unit normal distribution. This is a special case when sigma=1 and mu=0.

In above graph is plot of normal distribution. In it we can easily interpret that a normal distribution is quite symmetrical about its center. That means the left side (0,-3) of the center of

the peak is a mirror image of the right side(0,+3). There is also only one peak i.e. one mode in a normal distribution.

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

Figure: Normal distribution in a bell curve

The random variables are distributed in the form of a symmetrical, bell-shaped curve.

Properties of Normal Distribution are as follows;

1. Unimodal -one mode

2. Symmetrical -left and right halves are mirror images

3. Bell-shaped -maximum height (mode) at the mean

4. Mean, Mode, and Median are all located in the center

5. Asymptotic

Formula:-

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

Where –

- μ = Mean
- σ = Standard Deviation
- $\pi \approx 3.14159$
- $e \approx 2.71828$

Question No. 11:- How do you handle missing data? What imputation techniques do you recommend?

Answer:- In Basic or I would say earlier we did use fillna method to treat nan or missing data. But now we have some advance technique to handle Nan and replacing the missing values.

(A)- KNN Imputer:- * KNN imputer will try to find relation with other columns and impute the data according the relation with other columns.

* In this case Age Nan is depending on the similarity with fare columns.

In [32]:
```
df=pd.DataFrame({'salary':[25000,48000,71000,85000,90000,55000],
                 'city':['Bengaluru','Delhi','Hyderabad','Bengaluru','Hyderabad','Bengaluru'],
                 'gender':['male', 'female','female','female','male','male'],
                 'exp':[1,3,5,6,9,None]})
df
```

Out[32]:

	salary	city	gender	exp
0	25000	Bengaluru	male	1.0
1	48000	Delhi	female	3.0
2	71000	Hyderabad	female	5.0
3	85000	Bengaluru	female	6.0
4	90000	Hyderabad	male	9.0
5	55000	Bengaluru	male	NaN

Now we can see that how knn imputation technique work in below example.

In [33]:
```
from sklearn.impute import KNNImputer
```

In [36]:
```
knn_imp=KNNImputer(n_neighbors=3)
knn_imp2=pd.DataFrame(knn_imp.fit_transform(df[['salary','exp']]), columns=['salary','exp'])
knn_imp2
```

Out[36]:

	salary	exp
0	25000.0	1.0
1	48000.0	3.0
2	71000.0	5.0
3	85000.0	6.0
4	90000.0	9.0
5	55000.0	3.0

(B) Iterative Imputer:- In above example we saw how knn imputation techque work, similarly iterative imputer technique work, but difference is that this method treat other columns (which does not have nulls as feature and train on them and treat null column as label. Finally it will predict the nan data and impute. Its just like regression problem. Here null column is lable.

We take above example's data set.

In [32]:
```
df=pd.DataFrame({'salary':[25000,48000,71000,85000,90000,55000],
                 'city':['Bengaluru','Delhi','Hyderabad','Bengaluru','Hyderabad','Bengaluru'],
                 'gender':['male', 'female','female','female','male','male'],
                 'exp':[1,3,5,6,9,None]})
df
```

Out[32]:

	salary	city	gender	exp
0	25000	Bengaluru	male	1.0
1	48000	Delhi	female	3.0
2	71000	Hyderabad	female	5.0
3	85000	Bengaluru	female	6.0
4	90000	Hyderabad	male	9.0
5	55000	Bengaluru	male	NaN

In [38]:
```
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
```

In [39]:
```
iter_imp=IterativeImputer()
ite_im=pd.DataFrame(iter_imp.fit_transform(df[['salary','exp']]),columns=['salary','exp'])
ite_im
```

Out[39]:

	salary	exp
0	25000.0	1.000000
1	48000.0	3.000000
2	71000.0	5.000000
3	85000.0	6.000000
4	90000.0	9.000000
5	55000.0	3.864759

(C) **Simple Imputer-** In this technique we can treat the null with mean, mode & median, but default it takes mean.

Exp- Same Dataset.

Working method:-

```
In [41]: from sklearn.impute import SimpleImputer
         from sklearn.compose import make_column_transformer
```

```
In [44]: si=SimpleImputer()
         x=make_column_transformer((si,['exp']),
                                   remainder='passthrough') #passthrough to keep all other columns.
         impute=pd.DataFrame(x.fit_transform(df))
         df
```

Out[44]:

	salary	city	gender	exp
0	25000	Bengaluru	male	1.0
1	48000	Delhi	female	3.0
2	71000	Hyderabad	female	5.0
3	85000	Bengaluru	female	6.0
4	90000	Hyderabad	male	9.0
5	55000	Bengaluru	male	NaN

Question No. 12:- What is A/B testing?

Answer:- A/B tests give you the data that you need to make the most of your marketing budget. Let's say that your boss or manager has given you a budget to drive traffic to your site using google adwords. You set up an A/B test that tracks the number of clicks for 3 different article titles. You run the test for a week, making sure that on any particular day and at any particular time, youre running the same numberr of ads for each option.

A/B tests let you evaluate the impact of changes that are relatively inexpensive to implement.

A/B testing is not only cost effective,but also it's time effficient. You test 2 or 3 elements & get your answer. From here it is easy to deccide, whetther to implement a change or not. If red life data doesn't hold up to your test results, it's always possible to revert back to an older versions.

Question No. 13:- Is mean imputation of missing data acceptable practice?

Answer:- Mean imputations is a simple technique to treat null values with mean of it's particular feature. It is a good practice at that time untill we deal with the continous data, but for binary classification it is not a good practice because it doesn't take into account features correlation. That's overall we say that it is a bad practice to treat missing values by mean.

In other words, yes, we get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is

unbiased. Even so, the standard error of that mean will be too small. Because the imputations are themselves estimates, there is some error associated with them.

Question NO. 14:- What is linear regression in statistics?

Answer:- First of all we understand what is regression.

Regression:- Regression in statistics in the processs of predicting a label(or dependent variable) based of the featuress (Independent variables) at hand. Regression is used for time series model and finding the caused efffect relationship between the variables and forcasting.

For Example the relationship betweeen the stock prices of the company and various factors like customer reputaion and company annual perfomance etc. can be studies using regression.

***The uses of Regression:-** -It show the significant relationship between the lable(dependent variable) and the features (independent variable).

- It shows the extent of the impact of multiple independent variables on the dependent variable.

- It can also measure these effects even if the variables are on a different scale.

****Linear Regression:-** "Linear Regression is one of the supervised machine learning technic can be use for regression."

Linear regression is one of the most fundamental and widely known machine learning algorithms which people start with building blocks of a linear regression model are:-

- Independent variables could be Discreet and continous.
- A best-fit regression line.
- Continous dependent variable i.e. A linear regression model predicts the depndent variable using a regression line based on the independent variables.

The Equation of linear regression is

Y= a+b*x +E *Same like Y=Mx+c

Where, a=intercept

b =slope

x=independent variable, &

e = error.

Question No. 15:- What are the various branches of statistics?

Answer:- Statistics:- Statistics is a branch that deals with study of the collection, analysis , interp;retation, organisation, and presentation of data. Mathematically, statistics is defined as the set of equations, which are used to analysis the things. Statistics is the main branch of mathematics. Used to perform different operations like data collection, analysis and so on. In other words, statistics is a form of mathematical analysis that uses quantitative models to give a

set of experimental data or studies of real life. Statistics examine the methodology for collecting, reviewing, analyzing, and making data conclusions.

There are two types of statistics:-

Descriptive and Inferential

Descriptive:- With in term descriptive is describe. Lets understand with this simple real example. Let's assume that I am a teacher and my boss say to me tell me the strength of your students, here strength mean how good they are. In methamical term what is average marks in your class the students are getting? Suppose I have a 20 student, so that it is not a big task to me to describe my student because I can conduct a test and according to there average marks I can said that this is the strength of my students to boss. 20 students are not a big number, and I can intract to every student and every information I'll able to collect.

So that for me it is not a big task at all difficult to explain my boss that how good are my students. So what is happening here I am able to describe my student to my boss because the strength is very low.

Inferential:- In inferential Statistics, we take sample from the population and conduct the test on samples, and what ever the output we get it will be all about our whole dataset i.e. population.

In above example we take 20 stundent as the strength, but if I take 20000 students than I will not be able to keep track of all the students. It is very very difficult for me to identify each and every students. So that time for me inferential statistics comes into picture it helps us to describe us all the students.

In inferential statistics we are going to take samples from the population(here 20000 students) randomly. Now what we will do we take all of our sample students marks and get the average of it, than now I would say that this is the average marks of my all 20000 students.

IN inferential statistics we pick the samples from the population data, and conduct the test on sample data, and whatever the result we get we will apply it to all population data.

Analytics Methodology and how industry use statistics:

- In Weather forecasting
- Giving Insurance
- Stock Market
- Drug effectiveness before releasing for the public
- Diseased survival probablity.
- Election winning and Exit poll prediction
- Loan approval and fraud detection & so on.