STATISTICS WORKSHEET-1

Answer of the questions are below.

1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

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

a) Central Limit Theorem

Picking random values to compare – values should be enough to compare

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

b) 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 variables are dependent
c) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned

All statement is correct and being follow

d) All of the mentioned

5. _ Poisson _____ random variables are used to model rates.

a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False
b) False
No because it's a theory and all values are random

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

b) Hypothesis

It's basically an assumption

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

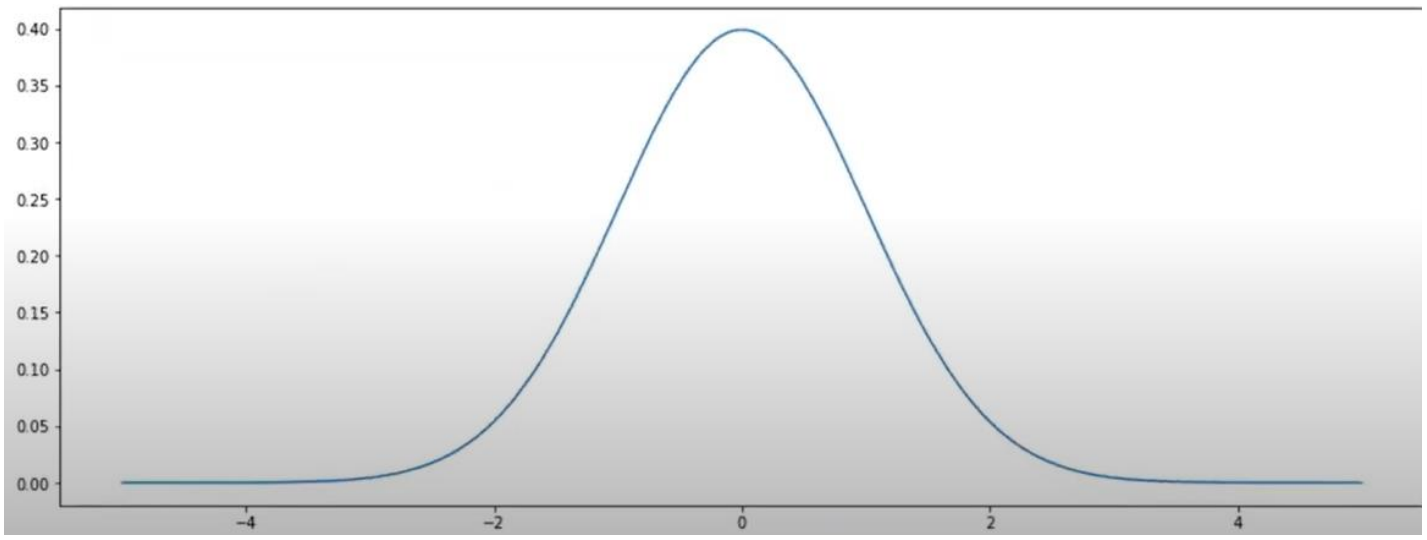
a) 0
Its normalized - has 0 unit

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

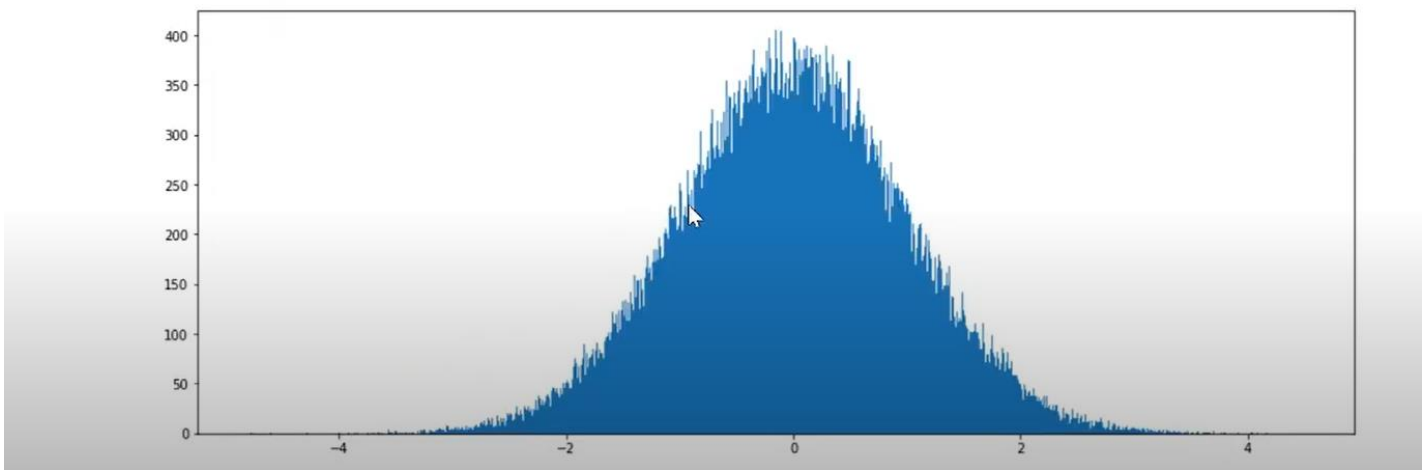
a) Outliers can have varying degrees of influence
b) Outliers can be the result of spurious or real processes
c) Outliers cannot conform to the regression relationship
d) None of the mentioned

c) Outliers cannot conform to the regression relationship

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




```
In [24]: fig, ax = plt.subplots(figsize=(16, 6))
         ax.hist(x_rvs, bins=1000)
         plt.show()
```



As we can see there is no outliers and data looks normally distributed – easy to identify.
Both side data –most of the data in center near to mean (mean of the distribution) - most data points cluster toward the middle of the range
Also called: Gaussian distribution

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

Imputation is getting used to replace data with related values calculating with other provided data, to retain almost every data in the dataset. Removing data from sheet is not good and feasible and it can reduce the size of the database as well. And the reduction can be very large it may change the database accuracy as well.
Doing that may give incorrect analysis or data may be messy as well.

So we can use Imputation to solve these problems, but now we need to know which the best option for Imputation is I recommend KNN imputers

We have 3 imputers
1 simple imputers
2 KNN imputers
3 iterative imputers

As I understand - KNN imputers is best for the imputation process as it working by finding related neighbors and giving data according to real-time data and by compering nearest or relatable value – as simple imputers is taking mean and iterative working as regression which is by comparing all data – either its nearest or retable to not

12. What is A/B testing?

It’s to comparison 2 version and check witch one is working better – as cx do not behave same they make different choices – to conduct a basic test we use A/B testing to check the most chosen version by cx

13. Is mean imputation of missing data acceptable practice?

I do not think mean imputation is acceptable practice because it has no correlation and considering all given data and getting mean which can be a terrible result.

14. What is linear regression in statistics?

Prediction algorithm to find relationship between dependent and independent value.

Like if we promoting out product on TV and want to see it will work or not, we can fetch previous year sales done by TV promotion we can get an idea it will do better in future or not.

15. What are the various branches of statistics?

Descriptive statistics and inferential statistics