Assignment – 6

STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question. → Discrete Random

it takes ✓ 0, 1, 2, 3 - - - n
✓ 2.3, 3.4, 9.7 - - - n both
↳ Contineous Random

?① Bernoulli random variables take (only) the values 1 and 0.

a) True
[✓ b) False]

?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] – CLT
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

?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]

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]

?⑤ ______ random variables are used to model rates.

a) Empirical – 3 sigma
b) Binomial – 0, 1
[✓ c) Poisson] — independent
d) All of the mentioned

⑥ 10. Usually replacing the standard error by its estimated value does change the CLT. – Central Limit theorem

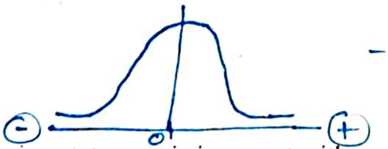
a) True CLT → Mean of all samples is approximately equals to the mean of the population.
[✓ b) False]

⑦ 1. Which of the following testing is concerned with making decisions using data?

~~a) Probability~~
[✓ b) Hypothesis] – P.value > 0.05 – Ho Accepted, HA Rejected; P.value < 0.05 – Ho – Rejected HA Accepted
c) Causal
d) None of the mentioned

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



- 0 = Mean = Median = Mode
- ideal Condition
- Bell shaped Curve

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]

✓ All 3 statements are correct with respect to outliers.

Ans. – ⑥ Mean of sample $\frac{\text{identical}}{}$ Mean of population

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

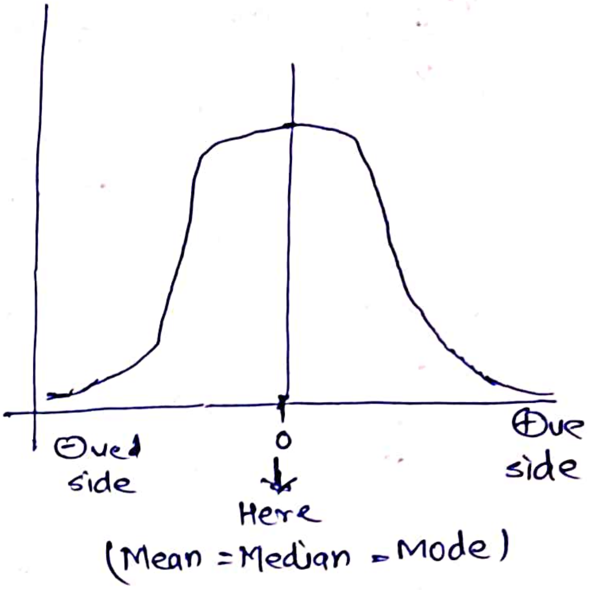
10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?

? 12. What is A/B testing?

13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

Answer - 10 Normal Distribution —

- Bell-shaped curve
- No Deviation in graph.
- Ideal Condition
- Normal Curve
- Data is symmetrically distributed around mean, median and mode.
- Here standard deviation is ±1.



Answer 11 - There are various techniques to handle missing values - Actually it depend's on the nature & how much missingness is present in datset.

(i) Mean / Median / Mode imputation.

(ii) drop data

(iii) Machine learning based imputation — KNN / Decesion tree

Answer 12 - Not TAUGT YET, SORRY

Answer-13 Actually the Acceptability of imputation techniques is depend's on the context and nature of the data. with respect to specific Context, Missingness, & the goal of Analysis we Can use diffrent techniques — Mean / median / mode where Mean Imputation is simple and Commonly used technique.

Answer-14 it Is a statistical Modeling technique which is used to Understand the relationship between Dependent Variable and independent Variable

$$\text{dependent Variable} \leftarrow \boxed{Y = \underset{\text{intercept}}{a} + \underset{\text{slope/ Cofficient}}{b}\,\underset{\text{Independent Variable}}{x} + e} \rightarrow \text{error}$$

Auswel-15

VARIOUS BRANCHES OF STATISTICS

STATISTICS

Discriptive statistics

Central Tendency

1. Mean (Numpy)

np.mean()

2. Median (Numpy)

np.median()

3. Mode (Pandas)

df [ ] . mode ()

Dispersion of Data

1. Range

max − Min

2. percentile

np . percentile (a, 80)

or df [ ] . quantile (a . 80)

3. Variance

$\sigma^2 = \frac{\Sigma (x - mean)^2}{n}$ np.var()

4 Standard deviation

$\sigma^2 = \sqrt{\frac{\Sigma (x - mean)^2}{n}}$ np.std()

5. skew

Inferential statistics

- We need to ~~bes~~ make the Decesion based on interpretation

1. Hypothesis testing
2. T − Testing.
3. correlation test
4. Chi − square test
5. Annove
6. AnnCovet

etc