STATISTICS WORKSHEET-1

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: 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

Answer: b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

5. **_______ random variables are used to model rates.**

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer: d) All of the mentioned

6. **Usually replacing the standard error by its estimated value does change the CLT.**

a) True

b) False

Answer: b) False

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

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

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

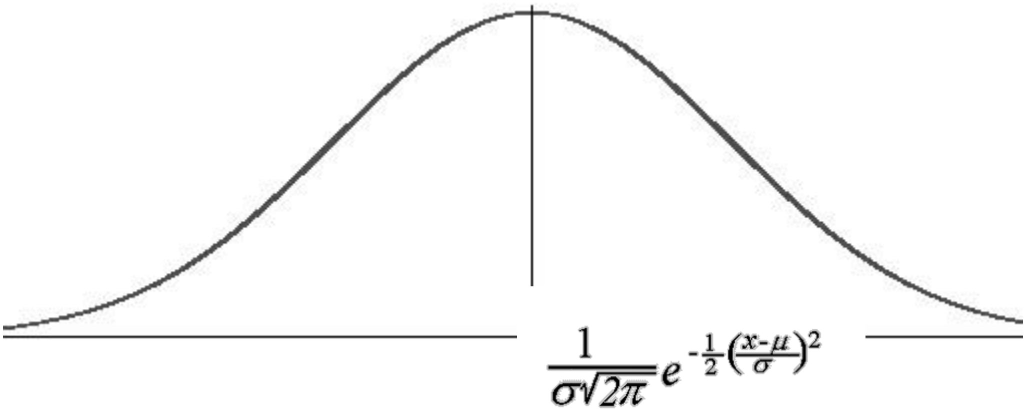
a) Outliers can have varying degrees of influence
b) Outliers can be the result of spurious or real processes
c) Outliers cannot conform to the regression relationship
d) None of the mentioned

Answer: c) Outliers cannot conform to the regression relationship

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

The normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where 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.

A normal curve can be represented as



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

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

Missing values can be handled using the following techniques:

- **Dropping null values:** If there are very less(less than 7-8%) null values in the dataset we can directly drop the null values
- **Imputing null values:** If the number of null values is more it is better to impute values rather than dropping values as dropping will lead to loss of data.

Various Imputing techniques are:

i. **Educated Guessing:** It sounds arbitrary and isn't your preferred course of action, but you can often infer a missing value. For related questions, for example, like those often presented in a matrix, if the participant responds with all "4s", assume that the missing value is a 4

ii. **Average Imputation:** Use the average value of the responses from the other participants to fill in the missing value. If the average of the 30 responses on the question is a 4.1, use a 4.1 as the imputed value. This choice is not always recommended because it can artificially reduce the variability of your data but in some cases makes sense.

iii. **Common-Point Imputation:** For a rating scale, using the middle point or most chosen value. For example, on a five-point scale, substitute a 3, the midpoint, or a 4, the most common value (in many cases). This is a bit more structured than guessing, but it's still among the riskier options. Use caution unless you have good reason and data to support using the substitute value.

iv. **Regression Substitution:** You can use multiple-regression analysis to estimate a missing value. We use this technique to deal with missing SUS scores. Regression substitution predicts the missing value from the other values. In the case of missing SUS data, we had enough data to create stable regression equations and predict the missing values automatically in the calculator.

v. **Multiple Imputation:** The most sophisticated and, currently, most popular approach is to take the regression idea further and take advantage of correlations between responses. In multiple imputation, software creates plausible values based on the correlations for the missing data and then averages the simulated datasets by incorporating random errors in your predictions. It is one of several examples where computers continue to change the statistical landscape. Most statistical packages like SPSS come with a multiple-imputation feature.

12. What is A/B testing?

- 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
- A/B testing includes 3 steps:

i. **Make a hypothesis**

- H_0(null hypothesis): this states that there will be no difference between results of control and variant groups
- H_a(alternative hypothesis): this states that there will be either positive or negative difference between results of control and test groups.

ii. **Create control group and test group**

- Control group receives a product or service without any changes
- Test group will receive a product or service with the proposed changes

iii. **Conduct A/B test and collect data**.

- The daily conversion rates are then measured for both control and test groups to study the impact of the changes made

13. Is mean imputation of missing data acceptable practice?

- Bad practice in general
- If just estimating means: mean imputation preserves the mean of the observed data.
- Leads to an underestimate of the standard deviation.
- Distorts relationships between variables by "pulling" estimates of the correlation toward zero.

14. What is linear regression in statistics?

- Linear regression is a basic and commonly used type of predictive analysis.
- The simplest form of linear regression with one dependent and one independent variable is y=mx+c where y is dependent variable, x is independent variable, m is slope of the regression line, c is the y intercept.
- This regression technique helps us establish direct relation between the independent and dependent variable.
- Following are the types of regressions:
- **Simple linear regression:** 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)
- **Multiple linear regression**: 1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)
- **Logistic regression**: 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)
- **Ordinal regression**: 1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)
- **Multinomial regression**: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)
- **Discriminant analysis**: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

15. What are the various branches of statistics?

2 main branches of statistics are

- Descriptive Statistics
  - ➢ Descriptive statistics deals with the presentation and collection of data
  - ➢ This is usually the first part of a statistical analysis

- This includes EDA, preprocessing and visualization normally

- Inferential statistics
  - This involves drawing right conclusion from the statistical analysis that has been performed using descriptive analysis.
  - Results for a sample data are studied to check how the data behaves in general.
  - This includes the prediction part of the process.