STATISTICS WORKSHEET-1

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

Ans- Option-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

Ans- Option-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

Ans- Option-D (All of the mentioned)

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

Ans- Option-C (Poisson)

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

Ans- Option-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

Ans- Option-B (Hypothesis)

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

Ans- Option-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

Ans- Option-C (Outliers cannot conform to the regression relationship)

10. What do you understand by the term Normal Distribution?
Ans- The normal distribution, or Gaussian distribution, is a key idea in statistics and probability. Imagine a bell-shaped curve where data clusters around the middle, or mean, with fewer observations as you move away from it. This curve is symmetrical, meaning the chances of getting values on one side of the mean are the same as the other.

In simpler terms, if you were to graph data that follows a normal distribution, it would look like a gentle hill, peaking at the mean, and tapering off evenly on both sides.

One cool thing about the normal distribution is that its mean, median, and mode all sit at the center, making it a balanced distribution.

The spread of the data is determined by something called the standard deviation. This measures how much individual data points deviate from the mean.

There's this handy rule called the 68-95-99.7 rule that helps understand the spread better. Roughly speaking, it says that about 68% of the data falls within one standard deviation from the mean, about 95% falls within two standard deviations, and about 99.7% falls within three standard deviations.

Another useful tool is the Z-score, which tells us how many standard deviations a particular data point is from the mean. It helps compare different normal distributions.

Why does all this matter? Well, the normal distribution is used in many fields like finance, physics, social sciences, and engineering. It's like a common language for describing randomness in nature and human behavior. So, understanding it helps us make sense of the world around us and make better decisions based on data.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans- Handling missing data is super important in data analysis because missing values can mess up our conclusions. There are lots of ways to deal with them, depending on what kind of data we're working with and what we want to find out.

1. *Mean/Median/Mode Imputation*: If we have some missing numbers, we can just replace them with the average (mean), middle value (median), or most common value (mode) of the numbers we do have. It's easy and quick, but it might not be accurate if the missing numbers follow a pattern.

2. *Forward Fill/Backward Fill*: If we're dealing with data that's in order, like time series, we can fill in missing values with the last number we saw (forward fill) or the next one we expect (backward fill). This works if the missing numbers are probably close to the ones we know.

3. *Linear Interpolation*: Imagine we have a line connecting all our data points. We can guess the missing values based on where they'd fall on that line. This works well for data that changes smoothly over time.

4. *K-Nearest Neighbors (KNN) Imputation*: Let's say we have missing values for some people's ages. We can look at other people who are similar to them and use their ages as a guess. It's like asking neighbors for help.

5. *Multiple Imputation*: Sometimes, we're not sure what the missing values should be. So, we make a bunch of guesses based on what we know about the data. Then, we analyze all these guesses together to get a better idea.

6. *Regression Imputation*: If we know that some data points are related to each other, we can use that relationship to guess missing values. For example, if we know people's heights and weights are related, we can use one to guess the other.

7. *Expectation-Maximization (EM) Algorithm*: This one's a bit fancy. It's like a smart guess-and-check method. We make a guess about the missing values, then update our guess based on what we learn from the rest of the data. We keep doing this until our guesses are pretty close to the truth.

When we're picking a method, we have to think about what's best for our data and what we're trying to figure out. It's also important to be clear about how we filled in the missing values so others can understand and trust our results.

12. What is A/B testing?
Ans- A/B testing, also called split testing, is a core method in data science used to compare different versions of a product, webpage, marketing campaign, or any other business element to see which one works better. Instead of relying on hunches or guesses, it's all about using data to make decisions.

Here's a breakdown of how A/B testing usually goes:

1. **Coming Up with an Idea**: The first step in A/B testing is forming a hypothesis. This means figuring out a specific change (like a new design, layout, or wording) that you think might improve the thing you're testing. For example, you might guess that switching the color of a "Buy Now" button from green to red will get more clicks.

2. **Randomly Splitting the Audience**: Once you have your hypothesis, you randomly divide your audience or sample into two groups: Group A and Group B. Group A gets the original version (like the green button), while Group B gets the changed version (like the red button). This random split helps make sure that any differences in performance between the groups are because of the changes you're testing, not other factors.

3. **Gathering Data**: Next, you collect data on how each group interacts with the thing you're testing. This could mean tracking stuff like click-through rate, conversion rate, bounce rate, or any other important measure of performance. The goal is to get enough data so you can make solid conclusions about which version is better.

4. **Analyzing the Results**: Once you've got enough data, you crunch the numbers to see if there's a significant difference in performance between the two versions. This involves using statistical tools like hypothesis testing, confidence intervals, and p-values to figure out if any differences you see are real or just due to chance.

5. **Making a Call**: Based on what the data tells you, you decide whether to go with the changed version (if it performs better) or stick with the original (if there's no real difference or if the original is actually better). This decision should take

into account both statistical significance and practical considerations, like how big the difference is and any costs or risks involved in making the change.

A/B testing offers several advantages in data science and business decision-making:

1. **Using Data to Drive Decisions**: A/B testing lets organizations make choices based on actual data instead of gut feelings or opinions. This can lead to smarter strategies for improving performance.

2. **Fine-Tuning Everything**: By testing out different variations and seeing what works best, A/B testing helps optimize processes, products, and marketing efforts for better results.

3. **Testing the Waters**: A/B testing lets organizations try out changes on a small scale before committing to them fully. This helps minimize the risk of making big, expensive changes that might not pan out.

4. **Always Getting Better:** A/B testing is an ongoing process that can be used to keep improving over time. By regularly testing and tweaking things, organizations can stay competitive and adapt to changes in the market.

All in all, A/B testing is a powerful tool for data scientists, helping them optimize performance and drive business growth by making decisions based on evidence.

13. Is mean imputation of missing data acceptable practice?
Ans-

14. What is linear regression in statistics?
Ans- Linear regression is like drawing a straight line through a bunch of points on a graph. Imagine you have some data points scattered around on a graph. Linear regression helps you find the best-fitting line that goes through those points.

Here’s how it works:

1. **You have data**: You have some data points. For example, you might have the height and weight of a bunch of people.

2. **You want to find a relationship**: You suspect that there’s a relationship between these data points. For example, you might think that as people get taller, they tend to weigh more.

3. **Linear regression finds the line**: Linear regression helps you find the best-fitting straight line that shows the relationship between the variables. This line can help you predict one variable based on the other.

4. **Making predictions**: Once you have this line, you can use it to make predictions. For example, if you know someone’s height, you can predict their weight using the line you found.

So, linear regression is a tool that helps you understand and predict how one variable changes as another variable changes, using a straight line as a model.

15. What are the various branches of statistics?
Ans- There are some of the main branches of statistics explained in simple terms:

1. **Descriptive Statistics**: Descriptive statistics involve methods for summarizing and describing the features of a dataset. It includes measures like mean (average), median (middle value), mode (most common value), and measures of variability like standard deviation and range.

2. **Inferential Statistics**: Inferential statistics involves making predictions or inferences about a population based on a sample of data. It includes techniques like hypothesis testing, confidence intervals, and regression analysis.

3. **Probability**: Probability is the branch of statistics that deals with the likelihood of events occurring. It involves concepts like random variables, probability distributions, and calculating probabilities of various outcomes.

4. **Biostatistics**: Biostatistics applies statistical methods to biological and health-related data. It's used in fields like medicine, epidemiology, genetics, and public health to analyze data related to diseases, treatments, and health outcomes.

5. **Econometrics**: Econometrics applies statistical methods to economic data. It's used to analyze economic relationships, forecast economic trends, and evaluate the effects of policies or interventions.

6. **Actuarial Statistics**: Actuarial statistics involves using statistical methods to analyze risk and uncertainty in insurance, finance, and other fields. Actuaries use techniques like risk modeling, mortality tables, and financial forecasting to assess and manage risk.

7. **Social Statistics**: Social statistics applies statistical methods to data related to social phenomena, such as demographics, surveys, and social experiments. It's used in sociology, political science, and other social sciences to study patterns of behavior and social trends.

8. **Business Statistics**: Business statistics involves applying statistical methods to analyze data in business contexts. It's used for market research, sales forecasting, quality control, and other business-related tasks to make informed decisions and improve performance.

These are just some of the main branches of statistics, and there are many other specialized areas and applications within the field.