Statistics worksheet 1

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

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

3. Which of the following is incorrect with respect to use of Poisson distribution?

b) Modeling bounded count data

4. Point out the correct statement.

d) All of the mentioned

5. _______ random variables are used to model rates

c) Poisson

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

b) False

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

b) Hypothesis

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.

a) 0

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

c) Outliers cannot conform to the regression relationship

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

Normal Distribution-

The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.

The normal distribution is a continuous probability distribution that is symmetric around its mean, with most observations clustered around the central peak, and probabilities of values far from the mean taper off equally in both directions. Extreme values in both ends of the distribution are not equally likely. While the normal distribution is symmetric, not all symmetric distributions are normal. For example, Student's t, Cauchy and logistic distributions are symmetric.

Parameters of the Normal Distribution-

As with any probability distribution, the parameters for the normal distribution define its shape and probabilities entirely. The normal distribution has two parameters, the mean and standard deviation.

- **Mean**

The mean is the central tendency of the normal distribution. It defines the location of the peak for the bell curve. Most values cluster around the mean. On a graph, changing the mean shifts the entire curve left or right on the X-axis.

- **Standard deviation**

The standard deviation is a measure of variability. It defines the width of the normal distribution. The standard deviation determines how far the values fall from the mean. It shows the specific distance between the observations and the average.

The Empirical Rule for the Normal Distribution-

When you have normally distributed data, the standard deviation becomes especially valuable. You can use it to determine the proportion of values that fall within a specified number of standard deviations from the mean. For example, in a normal distribution, 68% of observations fall within +/- 1 standard deviation from the mean. This property is part of the empirical rule, which describes the percentage of data that falls within a specific number of standard deviations from the mean of bell-shaped curves.

Standard Normal Distribution and Standard Scores-

As we saw above, the normal distribution has many different shapes depending on the parameter values. However, the standard normal distribution is a special case of the normal distribution where the mean is zero and the standard deviation is 1. This distribution is also known as the Z-distribution.

A value on a standard normal distribution is known as a standard score or Z-score. A standard score represents the number of standard deviations above or below the mean that a specific observation falls. For example, a standard score of 1.5 indicates that the observation is 1.5 standard deviations above the mean. On the other hand, a negative score represents a value below average. The Z-score of the mean is 0.

Standardization: How to Calculate Z-scores-

The standard score is a great way to understand where a specific observation falls relative to the overall normal distribution. They allow you to take observations taken from normally distributed populations that have different means and standard deviations and place them on a standard scale. This standard scale enables you to compare observations that would otherwise be difficult.

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

Ans.

Loss of data is an everyday problem that a data professional needs to deal with. Missing data is defined as a missing value, Missing data can be from anything missing sequence Incomplete feature, missing files, incomplete information, data entry error etc.

Handle missing data - In the case of multivariate analysis, if there are a large number of missing values, it may be better to discard those cases (rather than impute) and replace them. On the other hand, in univariate analysis, imputation can reduce the amount of bias in the data if the values are missing at random.

There are two forms of randomly missing values:

- MCAR: Missing completely at random – The first form is completely at random (MCAR) missing. This form exists when the missing values are randomly distributed across all observations.
- MAR: Missing at random - That second form is missing at random (MAR). In MAR, the missing values are not randomly distributed across observations but rather in one or more sub-samples. This form is more common than the previous one.

Imputation techniques –

I. Complete Case Analysis(CCA)- This is a very simple way to handle missing data, which directly deletes those rows in which data is missing i.e. we consider only those rows where we have complete data i.e. data is not missing. This method is also popularly known as "list wise deletion".

When to Use:-

- Data is MAR(Missing At Random).
- Good for Mixed, Numerical, and Categorical data.
- Missing data is not more than 5% – 6% of the dataset.
- Data doesn’t contain much information and will not bias the dataset.

II. Arbitrary Value Imputation- This is an important technique used in Imputation as it can handle both the Numerical and Categorical variables. This technique states that we group the missing values in a column and assign them to a new value that is far away from the range of that column. Mostly we use values like 99999999 or -9999999 or “Missing” or “Not defined” for numerical & categorical variables.

When to Use:-

- When data is not MAR(Missing At Random).
- Suitable for All.

III. Frequent Category Imputation- This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column. This technique is also referred to as Mode Imputation.

When to Use:-

- Data is Missing at Random(MAR)
- Missing data is not more than 5% – 6% of the dataset.

12. What is A/B testing?

A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.

The concept is similar to the scientific method. If you want to find out what happens when you change one thing, you have to create a situation where only that one thing changes.Think about the experiments you conducted in elementary school. If you put 2 seeds in 2 cups of dirt and put one in the closet and the other by the window, you'll see different results. This kind of experimental setup is A/B testing.

A/B testing important-

- A/B tests give you the data that you need to make the most of your marketing budget. Let's say that your boss 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, you're running the same number of ads for each option.
- A/B tests let you evaluate the impact of changes that are relatively inexpensive to implement. Running an AdWords campaign can be costly, so you want every aspect to be as effective as possible.
- A/B testing is not only cost effective, it's time efficient. You test 2 or 3 elements and get your answer. From there, it's easy to decide whether to implement a change or not. If real-life data doesn't hold up to your test results, it's always possible to revert back to an older version.

When it comes to customer-facing content, there is so much you can evaluate with A/B testing. Common targets include-

- Email campaigns
- Individual emails
- Multimedia marketing strategies
- Paid internet advertising
- Newsletters
- Website design

13. Is mean imputation of missing data acceptable practice?

Mean imputation-

Maybe it's a bit dramatic, but median imputation (also known as mean substitution) really should be a last resort. It is a popular solution to missing data, despite its shortcomings. Mainly because it's easy. It can be really painful to lose a large part of the sample you collected so carefully, only to have a little power. But that doesn't make it a good solution, and it may not help you find relationships with strong parameter estimates. even if they are present in the population.

Mean imputation does not preserve the relationships among variables-

True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That's a good thing.

Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too.this is the original logic involved in mean imputation.

If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate.

Advantages and Drawbacks of Mean imputation –

- Missing values in your data do not reduce your sample size, as it would be in the case of listwise deletion (the default of many statistical software packages, such as R, Stata, SAS or SPSS). Since the mean imputation replaces all missing values, you can keep your entire database.
- Mean imputation is very easy to understand and implement (more on that later in R and SPSS examples). You can easily explain the imputation method to your audience and everyone with a basic knowledge of statistics will get what you did.
- If the response mechanism is MCAR, then the sample mean of your variable is not biased. Mean substitution may be a valid approach, if the univariate average of your variables is the only metric you are interested in.

Mean Imputation in R (Example)-

```
##### Create some synthetic data with missings #####

set.seed(87654)   # Reproducibility
N <- 1000        # Sample size

# Some random variables
x1 <- round(rnorm(N), 2)
x2 <- round(x1 + rnorm(N, 10, 5))
x3 <- round(runif(N, -100, 20))

# Insert missing values
x1[rbinom(N, 1, 0.2) == 1] <- NA  # 20% missingness
x2[rbinom(N, 1, 0.05) == 1] <- NA # 5% missingness
x3[rbinom(N, 1, 0.7) == 1] <- NA  # 70% missingness

# Indicator for missings (needed later)
x1_miss_ind <- is.na(x1)
x2_miss_ind <- is.na(x2)
x3_miss_ind <- is.na(x3)
```

```
# Store variables in a data frame
data <- data.frame(x1, x2, x3)
head(data)       # First 6 rows of our data
```

Mean Imputation of One Column-

The mean imputation. If we want to impute only one column of our data frame, we can use the following R code:

```
##### Imputation of one column (i.e. a vector) #####

data$x1[is.na(data$x1)] <- mean(data$x1, na.rm = TRUE)
```

14. What is linear regression in statistics?

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) Does a set of predictor variables do a good job of predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of outcome variables, and how do they—indicated by the magnitude and sign of beta estimates—affect outcome variables? These regression estimates are used to explain the relationship between a dependent variable and one or more independent variables. The simplest form of a regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score at independent variable.

Variable naming. The dependent variable of regression has many names. It may be called an outcome variable, criterion variable, endogenous variable, or regressor. The independent variable may be called the exogenous variable, the predictor variable, or the regressor.

The three major uses for regression analysis are (1) determination of the strength of predictors, (2) prediction of an effect, and (3) trend forecasting.

Types of Linear Regression-

- **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)

15. What are the various branches of statistics?

Various Branches Of Statistics-

Descriptive statistics and inferential statistics are the two main branches of statistics. Both of these are used in scientific data analysis.

- **Descriptive Statistics-**

  The first aspect of statistics is descriptive statistics, which deals with the presentation and collection of data. It is not as simple as it seems, and the statistician must know how to design and conduct experiments, select the appropriate focus group, and prevent biases that are very easy to introduce into the experiment.

  Generally, descriptive statistics can be categorized into-

  Measures of central tendency

  Measures of variability

- **Measures of Central Tendency-**

  Measures of central tendency are used by statisticians to examine the value distribution center. These are the measures of trend

  Mean

  The mean is a common approach for describing the central tendency. To calculate the average of several values, first count them all and then divide them by the number of possible values.

  Median

  It is an outcome that is found in the middle of a set of values. In numerical journals, edit the results and the result that is in the centre of the distributed sample finds that one is an easy technique to get the median.

  Mode

  In the given data set the value which occurs most frequently is the mode.

- **Measures of Variability-**

  The measure of variability helps the statisticians in analysing the distribution that comes from a particular data set. Quartiles, ranges, variances, and standard deviation are the variability variables.

- **Inferential Statistics-**

  Inference statistics are statistical techniques that allow statisticians to utilise data from a sample to conclude, predict the behaviour of a given population, and make judgments or decisions.

  Using descriptive statistics, inference statistics frequently talk in terms of probability. Furthermore, a statistician uses these techniques mainly for data analysis, writing, and drawing conclusions from the limited data. This is accomplished by taking samples and determining their reliability.