STATISTICS WORKSHEET-1

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

   a) True

   b) False

   Answer: a

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

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

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: d

5. _____ random variables are used to model rates.

   a) Empirical

b) Binomial

c) Poisson d

d) All of the mentioned

Answer: c

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

   a) True

   b) False

   Answer: b

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

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

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: b

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

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

Answer:

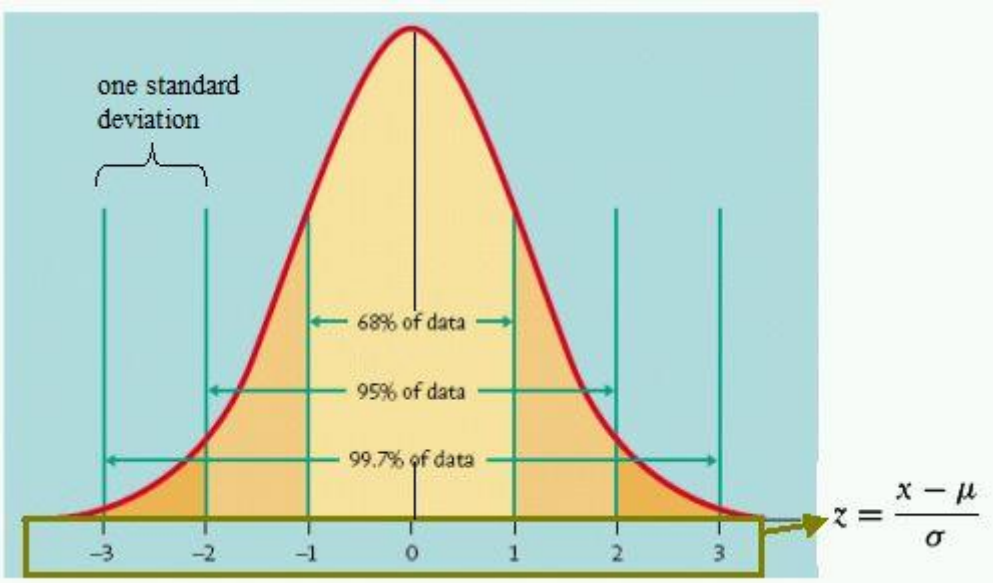
In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.Normal distributions are also called Gaussian distributions or bell curves because of their shape.

Normal distributions have key characteristics that are easy to spot in graphs:

- The mean, median and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.

The **empirical rule**, or the 68-95-99.7 rule, tells you where most of your values lie in a normal distribution:

- Around 68% of values are within 1 standard deviation from the mean.
- Around 95% of values are within 2 standard deviations from the mean.
- Around 99.7% of values are within 3 standard deviations from the mean.



Central Limit Theorem:

The central limit theorem is the basis for how normal distributions work in statistics.

In research, to get a good idea of a population mean, ideally you'd collect data from multiple random samples within the population. A sampling distribution of the mean is the distribution of the means of these different samples.

The central limit theorem shows the following:

- Law of Large Numbers: As you increase sample size (or the number of samples), then the sample mean will approach the population mean.
- With multiple large samples, the sampling distribution of the mean is normally distributed, even if your original variable is not normally distributed.

Formula of the normal curve:

Once you have the mean and standard deviation of a normal distribution, you can fit a normal curve to your data using a probability density function.

In a probability density function, the area under the curve tells you probability. The normal distribution is a probability distribution, so the total area under the curve is always 1 or 100%.

The formula for the normal probability density function looks fairly complicated. But to use it, you only need to know the population mean and standard deviation.

For any value of *x*, you can plug in the mean and standard deviation into the formula to find the probability density of the variable taking on that value of *x*.

Normal probability density formula	**Explanation**
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	• $f(x)$ = probability • x = value of the variable • μ = mean • σ = standard deviation • σ^2 = variance •

The **standard normal distribution**, also called the *z*-**distribution**, is a special normal distribution where the mean is 0 and the standard deviation is 1.

While individual observations from normal distributions are referred to as *x*, they are referred to as *z* in the *z*-distribution. Every normal distribution can be converted to the standard normal distribution by turning the individual values into *z*-scores.

Z-scores tell you how many standard deviations away from the mean each value lies.

You only need to know the mean and standard deviation of your distribution to find the *z*-score of a value.

_Z_-score Formula Explanation

$$z = \frac{x - \mu}{\sigma}$$

- *x* = individual value
- μ = mean
- σ = standard deviation

We convert normal distributions into the standard normal distribution for several reasons:

- To find the probability of observations in a distribution falling above or below a given value.
- To find the probability that a sample mean significantly differs from a known population mean.
- To compare scores on different distributions with different means and standard deviations.

Finding probability using the *z*-distribution

Each *z*-score is associated with a probability, or *p*-value, that tells you the likelihood of values below that *z*-score occurring. If you convert an individual value into a *z*-score, you can then find the probability of all values up to that value occurring in a normal distribution.

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

    Answer: Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values. The following are some of the most prevalent methods:

    a) Mean imputation
    b) Substitution
    c) Hot deck imputation
    d) Cold deck imputation
    e) Regression imputation
    f) Stochastic regression imputation
    g) Single or Multiple Imputation

    In the statistical literature, the most advanced methodology for performing missing data imputation is multiple imputation. In multiple imputation we generate missing values from the dataset many times.

12. What is A/B testing?

    Answer:

It is a method to determine which set of varients is the best one. It is also known as split testing, is a marketing experiment where we split our audience to test a number of variations of compaign and determine which performs better. i.e.. we can show version A of marketing content to one half of our audience and version B to another.

A/B testing isn't difficult, but it requires marketers to follow a well-defined process. Here are these nine basic steps:

The fundamental steps to planning and executing an A/B test

- 1. Measure and review the performance baseline
- 2. Determine the testing goal using the performance baseline
- 3. Develop a hypothesis on how your test will boost performance
- 4. Identify test targets or locations
- 5. Create the A and B versions to test
- 6. Utilize a QA tool to validate the setup
- 7. Execute the test
- 8. Track and evaluate results using web and testing analytics
- 9. Apply learnings to improve the customer experience

BY Following the above steps—with clear goals and a solid hypothesis—will help you avoid common A/B testing mistakes.

A list of digital marketing elements that can be tested includes one or more of the items below:

- Navigation links
- Calls to action (CTAs)
- Design/layout
- Copy
- Content offer
- Headline
- Email subject line
- Friendly email "from" address
- Images
- Social media buttons (or other buttons)
- Logos and taglines/slogans

13. Is mean imputation of missing data acceptable practice?

Answer:

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

No, it is a terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Also, it decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Answer: Linear regresson is one of the most fundamental and widely known Machine Learning Algorithms which we start with. It depends on contionuous data. Building blocks of Linear Regression Model are:

a. Discrete/Continuous independent variables
b. A best_fit regression line
c. Continuous dependent variable. i.e.. the model predicts the dependent variable using a regression line based on independent variables. The equation is,

$$Y=a+b*x+e$$

Where a = intercept

b= slope of line

e= error

this equation is used to predict the values of the target variables based on the given predictor variables.

Here we have two types of regression model. i.e.. simple linear regression, multiple linear regression.

15. What are the various branches of statistics

Answer: There are four division on which statistics divided :

1) Mathematical or theoretical statistics: It helps in formatting statistical and experimental distribution.
2) Statistical methods or functions: It helps in collection, tabulation and interpretation of data. Also it helps in analyzing the data and insight from the data.
3) Descriptive statistics: Descriptive statistics include mean (average), variance, skewness, and kurtosis.
4) Inferential statistics: Inferential statistics include linear regression analysis, analysis of variance (ANOVA), logit/Probit models, and null hypothesis testing.

The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics:

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

i. Measure of frequency
ii. Measure of dispersion
iii. Measure of central tendency
iv. Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Inferential Statistics:

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to

give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.