STATISTICS WORKSHEET-1

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

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

a) True b) False

ANS: 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: 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: 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: 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: c) Poisson

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

a) True b) False

ANS: 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

ANS: 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

ANS: 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: c) Outliers cannot conform to the regression relationship

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

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

ANS: Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

In graphical form, the normal distribution appears as a "bell curve".

- The normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- Many naturally-occurring phenomena tend to approximate the normal distribution.
- In finance, most pricing distributions are not, however, perfectly normal.

Understanding Normal Distribution

The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation.

The normal distribution model is important in statistics and is key to the Central Limit Theorem (CLT). This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance).

The normal distribution is one type of symmetrical distribution. Symmetrical distributions occur when where a dividing line produces two mirror images. Not all symmetrical distributions are normal, since some data could appear as two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

Properties of the Normal Distribution

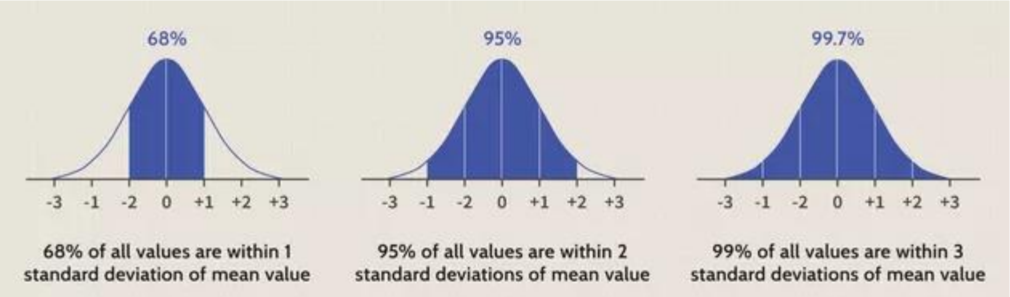
The normal distribution has several key features and properties that define it.

First, its mean (average), median (midpoint), and mode (most frequent observation) are all equal to one another. Moreover, these values all represent the peak, or highest point, of the distribution. The distribution then falls symmetrically around the mean, the width of which is defined by the standard deviation.

The Empirical Rule

For all normal distributions, 68.2% of the observations will appear within plus or minus one standard deviation of the mean; 95.4% of the observations will fall within +/- two standard deviations; and 99.7% within +/- three standard deviations. This fact is sometimes referred to as the "empirical rule," a heuristic that describes where most of the data in a normal distribution will appear.

This means that data falling outside of three standard deviations ("3-sigma") would signify rare occurrences.



The Formula for the Normal Distribution

The normal distribution follows the following formula. Note that only the values of the mean (μ) and standard deviation (σ) are necessary

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

where:

x = value of the variable or data being examined and f(x) the probability function

μ = the mean

σ = the standard deviation

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

ANS: Real-world data is messy and usually holds a lot of missing values. Missing data can skew anything for data scientists and, a data scientist doesn't want to design biased estimates that point to invalid results. Behind, any analysis is only as great as the data. Missing data appear when no value is available in one or more variables of an individual. Due to Missing data, the statistical power of the analysis can reduce, which can impact the validity of the results.

What are the reasons behind missing data?

Missing data can occur due to many reasons. The data is collected from various sources and, while mining the data, there is a chance to lose the data. However, most of the time cause for missing data is item nonresponse, which means people are not willing(Due to a lack of knowledge about the question) to answer the questions in a survey, and some people unwillingness to react to sensitive questions like age, salary, gender.

Types of Missing data

Before dealing with the missing values, it is necessary to understand the category of missing values. There are 3 major categories of missing values.

Missing Completely at Random (MCAR):

A variable is missing completely at random (MCAR) if the missing values on a given variable (Y) don't have a relationship with other variables in a given data set or with the variable (Y) itself. In other words, when data is MCAR, there is no relationship between the data missing and any values, and there is no particular reason for the missing values.

Missing at Random (MAR):

Let's understands the following examples:

Women are less likely to talk about age and weight than men.

Men are less likely to talk about salary and emotions than women.

Familiar right… This sort of missing content indicates missing at random.

MAR occurs when the missingness is not random, but there is a systematic relationship between missing values and other observed data but not the missing data.

Let me explain to you: you are working on a dataset of ABC survey. You will find out that many emotion observations are null. You decide to dig deeper and found most of the emotion observations are null that belongs to men's observation.

Missing Not at Random (MNAR):

The final and most difficult situation of missingness. MNAR occurs when the missingness is not random, and there is a systematic relationship between missing value, observed value, and missing itself. To make sure, if the missingness is in 2 or more variables holding the same pattern, you can sort the data with one variable and visualize it.

Detecting missing data

Detecting missing values numerically:

First, detect the percentage of missing values in every column of the dataset will give an idea about the distribution of missing values.

Imputation techniques:

The imputation technique replaces missing values with substituted values. The missing values can be imputed in many ways depending upon the nature of the data and its problem. Imputation techniques can be broadly they can be classified as follows:

Imputation with constant value:

As the title hints — it replaces the missing values with either zero or any constant value.

We will use the **Simple Imputer class from sklearn.**

Imputation using Statistics:

The syntax is the same as imputation with constant only the Simple Imputer strategy will change. It can be “Mean” or “Median” or “Most_Frequent”.

“Mean” will replace missing values using the mean in each column. It is preferred if data is numeric and not skewed.

“Median” will replace missing values using the median in each column. It is preferred if data is numeric and skewed.

“Most_frequent” will replace missing values using the most_frequent in each column. It is preferred if data is a string (object) or numeric.

Before using any strategy, the foremost step is to check the type of data and distribution of features (if numeric).

Advanced Imputation Technique:

Unlike the previous techniques, advanced imputation techniques adopt machine learning algorithms to impute the missing values in a dataset. Followings are the machine learning algorithms that help to impute missing values.

K_Nearest Neighbor Imputation:

The KNN algorithm helps to impute missing data by finding the closest neighbours using the Euclidean distance metric to the observation with missing data and imputing them based on the non-missing values in the neighbours.

The fundamental weakness of KNN doesn’t work on categorical features. We need to convert them into numeric using any encoding method. It requires normalizing data as KNN Imputer is a distance-based imputation method and different scales of data generate biased replacements for the missing values.

12. What is A/B testing?

ANS: A/B testing is a type of experiment in which you split your web traffic or user base into two groups, and show two different versions of a web page, app, email, and so on, with the goal of comparing the results to find the more successful version. With an A/B test, one element is changed between the original (a.k.a, “the control”) and the test version to see if this modification has any impact on user behaviour or conversion rates.

From a data scientist’s perspective, A/B testing is a form of statistical hypothesis testing or a significance test.

A/B Testing need-to-know terms

The data science behind A/B testing can get complex pretty quickly. But, we’ve highlighted a few need-to-know terms to start with the basics.

Null hypothesis

The null hypothesis, or H0, posits that there is no difference between two variables. In A/B testing, the null hypothesis would assume that changing one variable on a web page (or marketing asset) would have no impact on user behaviour.

Alternative hypothesis

On the flip side, an alternative hypothesis suggests the opposite of the null hypothesis: that changing an element will impact user behaviour. Take the example below:

Null hypothesis: The size of a call-to-action button does not impact click rates.

Alternative hypothesis: Larger call-to-actions buttons result in higher click rates.

Statistical significance

Statistical significance is meant to signify that the results of an A/B test are not due to chance (rejecting the null hypothesis).

This is calculated by measuring the p-value, or probability value. So, if the p-value is low, it is saying that it’s unlikely the results of the A/B test were random.

A rule of thumb tends to be that when the p-value is 5% or lower, the A/B test is statistically significant.

Confidence level

Think of the confidence level as the inverse of the p-value. The confidence level is the indication of how likely it is that the results of your experiment are due to the changed variable (that is, these results are not random or a fluke occurrence).

If a test is considered statistically significant when the p-value is at 5%, then the confidence level would be 95%

13. Is mean imputation of missing data acceptable practice?

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

Mean imputation is typically considered 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.

Second, mean imputation 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.

Mean imputation reduces the variance of the imputed variables. Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval. Mean imputation does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

ANS: 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 in predicting an outcome (dependent) variable?

(2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the 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 on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Generate predictions more easily

You can perform linear regression in Microsoft Excel or use statistical software packages such as IBM SPSS® Statistics that greatly simplify the process of using linear-regression equations, linear-regression models and linear-regression formula. SPSS Statistics can be leveraged in techniques such as simple linear regression and multiple linear regression.

You can perform the linear regression method in a variety of programs and environments, including:

R linear regression

MATLAB linear regression

Sklearn linear regression

Linear regression Python

Excel linear regression

Why linear regression is important

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

You'll find that linear regression is used in everything from biological, behavioral, environmental and social sciences to business. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a longestablished statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.

A proven way to scientifically and reliably predict the future

Business and organizational leaders can make better decisions by using linear regression techniques. Organizations collect masses of data, and linear regression helps them use that data to better manage reality — instead of relying on experience and intuition. You can take large amounts of raw data and transform it into actionable information.

You can also use linear regression to provide better insights by uncovering patterns and relationships that your business colleagues might have previously seen and thought they already understood. For example, performing an analysis of sales and purchase data can help you uncover specific purchasing patterns on particular days or at certain times. Insights gathered from regression analysis can help business leaders anticipate times when their company's products will be in high demand.

Key assumptions of effective linear regression

Assumptions to be considered for success with linear-regression analysis:

For each variable: Consider the number of valid cases, mean and standard deviation.

For each model: Consider regression coefficients, correlation matrix, part and partial correlations, multiple R, R2, adjusted R2, change in R2, standard error of the estimate, analysis-of-variance table, predicted values and residuals. Also, consider 95-percent-confidence intervals for each regression coefficient, variance-covariance matrix, variance inflation factor, tolerance, Durbin-Watson test, distance measures (Mahalanobis, Cook and leverage values), DfBeta, DfFit, prediction intervals and case-wise diagnostic information.

Plots: Consider scatterplots, partial plots, histograms and normal probability plots.

Data: Dependent and independent variables should be quantitative. Categorical variables, such as religion, major field of study or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables.

Other assumptions: For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and each independent variable should be linear and all observations should be independent.

Make sure your data meets linear-regression assumptions

Before you attempt to perform linear regression, you need to make sure that your data can be analyzed using this procedure. Your data must pass through certain required assumptions.

Here's how you can check for these assumptions:

- The variables should be measured at a continuous level. Examples of continuous variables are time, sales, weight and test scores.
- Use a scatterplot to find out quickly if there is a linear relationship between those two variables.
- The observations should be independent of each other (that is, there should be no dependency).
- Your data should have no significant outliers.
- Check for homoscedasticity — a statistical concept in which the variances along the best-fit linear-regression line remain similar all through that line.
- The residuals (errors) of the best-fit regression line follow normal distribution.

15. What are the various branches of statistics?

ANS: Statistics is the branch of mathematics that deals with data. Data (technically a plural word; the singular is 'datum') is a collection of values. For most of what we do, it will be numerical data (such as the inflation rate, the number of bees in a colony, or the marks in a class test), but it can also take other forms (such as the political party a voter intends to vote for, the football team they support, and so on). A collection of data is often referred to as a data set or set of data¸ but other words such as a list or simply collection are also often used. Don't worry too much about the words, just understand that we are referring to a collection of values. There are three real branches of statistics:

1. Data collection
2. Descriptive statistics
3. Inferential statistics

Data collection

Data collection is all about how the actual data is collected. For the most part, this needn't concern us too much in terms of the mathematics (we just work with what we are given), but there are significant issues to consider when actually collecting data.

Descriptive statistics

Descriptive statistics is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on). Descriptive statistics is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on).

Inferential statistics

Inferential statistics is the aspect that deals with making conclusions about the data. This is quite a wide area; essentially you are asking ‘What is this data telling us, and what should we do?’