Statistics

1. (a)
2. (a)
3. (c)
4. (d)
5. (c)
6. (b)
7. (b)
8. (a)
9. (c)
10. 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 graph form, normal distribution will appear as a bell curve. 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. For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations. The normal distribution model is motivated by the Central Limit Theorem. 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. Normal distribution is sometimes confused with symmetrical distribution. The random variables following the normal distribution are those whose values can find any unknown value in a given range. For example, finding the height of the students in the school. Here, the distribution can consider any value, but it will be bounded in the range say, 0 to 6ft. This limitation is forced physically in our query. Whereas, the normal distribution doesn't even bother about the range. The range can also extend to $-\infty$ to $+\infty$ and still we can find a smooth curve. These random variables are called Continuous Variables, and the Normal Distribution then provides here probability of the value lying in a particular range for a given experiment. Also, use the normal distribution calculator to

find the probability density function by just providing the mean and standard deviation value.

11. When dealing with missing data, we can use two primary methods to solve the error: imputation or the removal of data.The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model. The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required. Imputation is the process of replacing missing data with substituted values. In this post, different techniques have been discussed for imputing data with an appropriate value at the time of making a prediction. When imputed data is substituted for a data point, it is known as **unit imputation**; when it is substituted for a component of a data point, it is known as **item imputation**. At the time of model training/testing phase, missing data if not imputed with proper technique could lead to **model bias** which tends to degrade model performance.

    Some imputation techniques are as follows:

    - Predicted value imputation
    - Distribution-based imputation
    - Unique value imputation
    - Input data validation

    I will recommend distribution based imputation for data missing

12. A/B testing is a shorthand for a simple randomized controlled experiment, in which two samples (A and B) of a single vector-variable are compared. These values are similar except for one variation which might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, its complexity grows. A/B tests are useful for understanding user engagement and satisfaction of online features like a new feature or product. Large social media sites like LinkedIn, Facebook, and Instagram use A/B testing to make user experiences more successful and as a way to streamline their services. Today, A/B tests are being used also for conducting complex experiments on subjects such as network effects when users are offline, how online services affect user actions, and how users influence one another. Many professions use the data from A/B tests. This includes data engineers, marketers, designers, software

engineers, and entrepreneurs. Many positions rely on the data from A/B tests, as they allow companies to understand growth, increase revenue, and optimize customer satisfaction. Version A might be a version used at present (thus forming the control group), while version B is modified in some respect vs. A (thus forming the treatment group). For instance, on an e-commerce website the purchase funnel is typically a good candidate for A/B testing, since even marginal-decreases in drop-off rates can represent a significant gain in sales. Multivariate testing or multinomial testing is similar to A/B testing, but may test more than two versions at the same time or use more controls. Simple A/B tests are not valid for observational, quasi-experimental or other non-experimental situations - commonplace with survey data, offline data, and other, more complex phenomena

13. Bad idea according to me the pain when the dataset we want to use for Machine Learning contains missing data. The quick and easy workaround is to substitute a mean for numerical features and use a mode for categorical ones. Even better, someone might just insert 0's or discard the data and proceed to the training of the model. In the following article, I will explain why using a mean or mode can significantly reduce the model's accuracy and bias the results. I will also point you to few alternative imputation algorithms which have their respective Python libraries that you can use out-of-the-box. Key fact to note is that the drawbacks of using a mean apply when the missing data is MAR (Missing At Random).

14. Once the degree of relationship between variables has been established using co-relation analysis, it is natural to delve into the nature of relationship. Regression analysis helps in determining the cause and effect relationship between variables. It is possible to predict the value of other variables (called dependent variable) if the values of independent variables can be predicted using a graphical method or the algebraic method.It involves drawing a scatter diagram with independent variable on X-axis and dependent variable on Y-axis. After that a line is drawn in such a manner that it passes through most of the distribution, with remaining points distributed almost evenly on either side of the line. A regression line is known as the line of best fit that summarizes the general movement of data. It shows the best mean values of one variable corresponding to mean values of the other. The regression line is based on the criteria that it is a straight line that minimizes the sum of squared deviations between the predicted and observed values of the dependent variable.

15. There are so many branches of statistics. I once thought that every science, every decision making, and every analysis should integrate with the right statistics.

1. Econometric

Econometric is one of the branches of statistics where it takes parts to resolve economic models and problem. Along the way, there are tons of equation and statistics formula that we need to use in order to calculate and support economic theories.

2. Actuarial

Actuarial is another applied statistical branch that focuses on studying and analyzing risk in finance and insurance. Someone who masters this knowledge is expected to be the one who can take big parts in financial problem likes insurance, future risk, predicting a financial trend, etc. An actuary (call for actuarial experts) have a big contribution in deciding the fate of the company.

3. Psychometrics

Psychometrics is another interesting branch of statistics. This one focus one studying measurement technique and analyzing in the education world and psychology. This thing included attitude, personality, emotion, and many others. And the one that interests me so much, psychometrics is related to intelligence measurement. If you have take pars in IQ test or EQ test or personality test, that's when psychometrics hit the most.

4. Physics Statistics

Physics statistics is one of the statistical branches that focuses on solving physic science. Usually, statistics take part in measurement and calculation with particle. The combination of statistics and physics, there will be a settlement especially with the problem of modeling level of confidence atom or particle.

5. Population Statistics

The population of statistics is one of the most useful branches that study about many things related to society. It has many connections with another aspect of our life, such as health, education, migration, and so on. With this knowledge, the statistician can make a trend and predict what will happen to the population in the future.