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.

Ans- 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?

Ans- Central Limit Theorem

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

Ans Modeling bounded count data

4. Point out the correct statement.

Ans - Sums of normally distributed random variables are again normally distributed even if the variables are dependent

5. _____ random variables are used to model rates.

Ans- Poisson

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

Ans-False

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

Ans- Hypothesis

8 Normalized data are centered at_____and have units equal to standard deviations of the original data.
Ans – 0

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

Ans - Outliers cannot conform to the regression relationship

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 graph form, normal distribution will appear as a bell curve.

KEY TAKEAWAYS

- A 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.
- In reality, most pricing distributions are not perfectly normal.

1:13

Normal Distribution

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. 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 (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

Skewness and Kurtosis

Real life data rarely, if ever, follow a perfect normal distribution.
The skewness and kurtosis coefficients measure how different a given distribution is from a normal distribution. The skewness measures the symmetry of a distribution. The normal distribution is symmetric and has a skewness of zero. If the distribution of a data set has a skewness less than zero, or negative skewness, then the left tail of the distribution is longer than the right tail; positive skewness implies that the right tail of the distribution is longer than the left.

The kurtosis statistic measures the thickness of the tail ends of a distribution in relation to the tails of the normal distribution. Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution (e.g., five or more standard deviations from the mean). Distributions with low kurtosis exhibit tail data that is generally less extreme than the tails of the normal distribution. The normal distribution has a kurtosis of three, which indicates the distribution has neither fat nor thin tails. Therefore, if an observed distribution has a kurtosis greater than three, the distribution is said to have heavy tails when compared to the normal distribution. If the distribution has a kurtosis of less than three, it is said to have thin tails when compared to the normal distribution.

How Normal Distribution is Used in Finance

The assumption of a normal distribution is applied to asset prices as well as price action. Traders may plot price points over time to fit recent price action into a normal distribution. The further price action moves from the mean, in this case, the more likelihood that an asset is being over or undervalued. Traders can use the standard deviations to suggest potential trades. This type of trading is generally done on very short time frames as larger timescales make it much harder to pick entry and exit points.

Similarly, many statistical theories attempt to model asset prices under the assumption that they follow a normal distribution. In reality, price distributions tend to have fat tails and, therefore, have kurtosis greater than three. Such assets have had price movements greater than three standard deviations beyond the mean more often than would be expected under the assumption of a normal distribution. Even if an asset has went through a long period where it fits a normal distribution, there is no guarantee that the past performance truly informs the future prospects.

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

Ans-

Methods to Handle Missing Data

An analysis is only as good as its data, and every researcher has struggled with dubious results because of missing data. In this article, I will cover three ways to deal with missing data

Types of Missing Data

Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

- Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of

data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.

- Missing At Random (MAR): The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.
- Not Missing At Random (NMAR): When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.
-

Common Methods

1. Mean or Median Imputation

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

- There may not be enough observations with non-missing data to produce a reliable analysis
- In predictive analytics, missing data can prevent the predictions for those observations which have missing data
- External factors may require specific observations to be part of the analysis

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

2. Multivariate Imputation by Chained Equations (MICE)

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for

factor variables, and proportional odds model for ordered variables to impute missing data.
To set up the data for MICE, it is important to note that the algorithm uses all the variables in the data for predictions. In this case, variables that may not be useful for predictions, like the ID variable, should be removed before implementing this algorithm.
.
3. Random Forest
Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.
One caveat is that random forest works best with large datasets and using random forest on small datasets runs the risk of overfitting. The extent of overfitting leading to inaccurate imputations will depend upon how closely the distribution for predictor variables for non-missing data resembles the distribution of predictor variables for missing data. For example, if the distribution of race/ethnicity for non-missing data is similar to the distribution of race/ethnicity for missing data, overfitting is not likely to throw off results. However, if the two distributions differ, the accuracy of imputations will suffer.
The MICE library in R also allows imputations by random forest by setting the method to “rf”. The authors of the MICE library have provided an example on how to implement the random forest method here.
To sum up data imputations is tricky and should be done with care. It is important to understand the nature of the data that is missing when deciding which algorithm to use for imputations. While using the above algorithms, predictor variables should be set up carefully to avoid confusion in the methods implemented during imputation. Finally, you can test the quality of your imputations by normalized root mean square error (NRMSE) for continuous variables and proportion of falsely classified (PFC) for categorical variables.

Q 12-What is A/B testing?

Ans --At its most basic, A/B testing, also known as split testing, is a way to compare different versions of something to see which performs better. In these experiments, you define a conversion goal to measure, like clicks or completed

transactions. Two variations of the same marketing asset (like a web page or email) are then shown to different users at random while measuring the difference in performance.

For example, let's say you wanted to increase the number of clicks on the "Buy now" button on your product pages. You could run an A/B test to find out how button color affects click-through rates, experimenting with a green button and a blue button. You would:

1. Define your conversion goal. In this example, you want to measure click-through rate.
2. Design the A/B test. How big of a sample size do you want? Who will participate, new customers or existing customers?
3. Gather data. Will you run your own test or use A/B testing software? For how long will the test run?
4. Analyze the results. Which variation had the biggest positive impact on the conversion metric that matters most?

At the end of the A/B test, you'll be able to confidently implement the winning variation without worrying about jeopardizing conversions.

Why you should A/B test your online store

A/B testing is often used across traditional marketing campaigns to increase email open rates or boost social media engagement. However, when you apply these same principles to your online store, the benefits are even more impactful. Simply generating a 1% increase in clicks on your buy button could result in thousands of dollars in extra sales.

A/B testing your online store can help you:

- Eliminate guesswork: Designing your online store can be an extremely subjective process, with everyone having their own visual preferences or

making suggestions to mimic industry leaders. By continuously running different A/B tests, you can offer a data-driven approach to design and remove subjectivity from the process.

- Customize design decisions for your business: Homepage design, product pages, and the checkout experience look different for every business depending on the items sold and target audience. And, what works for one store won't necessarily work for yours. A/B testing allows you to generate results that are hyper-targeted and specific to your business and customers.
- Experiment in a low-risk environment: Changing even a few design elements on your homepage can create a jarring experience for returning customers. A/B testing allows you to make these types of changes in a controlled environment without sacrificing long-term customer loyalty, retention, or brand awareness.

A/B testing examples

The most challenging part of A/B testing is often coming up with the experiments themselves. Which pages should you focus on? Which elements should you change? How big of a change should you make?

Sometimes, you'll be able to react to customer feedback. For example, if customers keep asking the same questions about your return policy, you could test different ways to highlight relevant return policy details (like including a link to your return policy vs. including the full policy details directly on the product page). Other times, you have to get creative and come up with experiments yourself.

Q13-Is mean imputation of missing data acceptable practice?

Ans-Missing values, common in epidemiologic studies, are a major issue in obtaining valid estimates. Simulation studies have suggested that multiple imputation is an attractive method for imputing missing values, but it is relatively complex and requires specialized software. For each of 28 studies in the Asia Pacific Cohort Studies Collaboration, a comparison of eight imputation procedures

(unconditional and conditional mean, multiple hot deck, expectation maximization, and four different approaches to multiple imputation) and the naive, complete participant analysis are presented in this paper. Criteria used for comparison were the mean and standard deviation of total cholesterol and the estimated coronary mortality hazard ratio for a one-unit increase in cholesterol. Further sensitivity analyses allowed for systematic over- or underestimation of cholesterol. For 22 studies for which less than 10% of the values for cholesterol were missing, and for the pooled Asia Pacific Cohort Studies Collaboration, all methods gave similar results. For studies with roughly 10–60% missing values, clear differences existed between the methods, in which case past research suggests that multiple imputation is the method of choice. For two studies with over 60% missing values, no imputation method seemed to be satisfactory.

Q14-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.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, "how much additional sales income do I get for each additional $1000 spent on marketing?"

Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, "what will the price of gold be in 6 months?"

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)

Discriminant analysis

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

When selecting the model for the analysis, an important consideration is model fitting. Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R^2). However, overfitting can occur by adding too many variables to the model, which reduces model generalizability. Occam's razor describes the problem extremely well – a simple model is usually preferable to a more complex model. Statistically, if a model includes a large number of variables, some of the variables will be statistically significant due to chance alone.

Q15-What are the various branches of statistics?

Ans- Branches of Statistics

Descriptive Statistics and Inferential Statistics

Every student of statistics should know about the different branches of statistics to correctly understand statistics from a more holistic point of view. Often, the kind of job or work one is involved in hides the other aspects of statistics, but it is very important to know the overall idea behind statistical analysis to fully appreciate its importance and beauty.

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

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment. Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In

the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.