MCQ

1 -> true

2-> Central Limit Theorem

3 -> Modeling bounded count data

4 -> All of the mentioned

5 -> Poisson

6 ->False

7 -> Hypothesis

8 -> 0

9 -> Outliers cannot conform to the regression relationship

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

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".

KEY TAKEAWAYS

- 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.

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.

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).1

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.

All normal distributions can be described by just two parameters: the mean and the standard deviation.

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

Normal Distribution Formula.

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

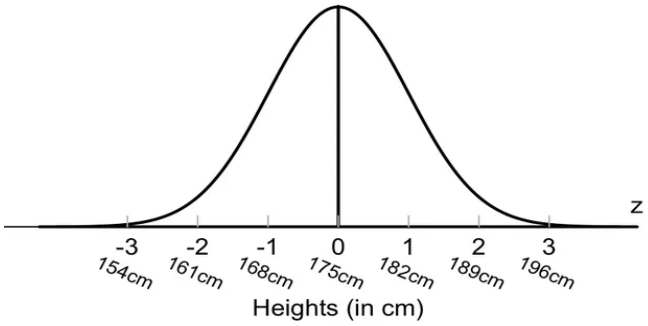
where:

- *x* = value of the variable or data being examined and f(x) the probability function
- μ = the mean
- σ = the standard deviation

Example of a Normal Distribution

Many naturally-occurring phenomena appear to be normally-distributed. Take, for example, the distribution of the heights of human beings. The average height is found to be roughly 175 cm (5' 9"), counting both males and females.

As the chart below shows, most people conform to that average. Meanwhile, taller and shorter people exist, but with decreasing frequency in the population. According to the empirical rule, 99.7% of all people will fall with +/- three standard deviations of the mean, or between 154 cm (5' 0") and 196 cm (6' 5"). Those taller and shorter than this would be quite rare (just 0.15% of the population each).



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

Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you.

Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.

Another common strategy among those who pay attention is imputation. 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:

1.Mean imputation

Calculate the mean of the observed values for that variable for all non-missing people.It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

2.Substitution

Assume the value from a new person who was not included in the sample.To put it another way, pick a new subject and employ their worth instead.

3.Hot deck imputation

A value picked at random from a sample member who has comparable values on other variables.To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

One benefit is that you are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in your research, you will always obtain a value between 5 and 10.Another factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

4.Cold deck imputation

A value picked deliberately from an individual with similar values on other variables.In most aspects, this is comparable to Hot Deck, but without the random variance. As an example, under the same experimental condition and block, you can always select the third individual.

5.Regression imputation

The result of regressing the missing variable on other factors to get a predicted value.As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

6.Stochastic regression imputation

The predicted value of a regression plus a random residual value.This has all of the benefits of regression imputation plus the random component's benefits.The majority of multiple imputation is based on stochastic regression imputation.

7.Interpolation and extrapolation

An estimate based on other observations made by the same person. It generally only works with data that is collected over time.Proceed with caution, though. For a variable like height in children–one that cannot be reduced through time–interpolation would make more sense. Extrapolation entails estimating beyond the data's true range, which necessitates making more assumptions than is necessary.

8.Single or Multiple Imputation

- Single and multiple imputation are the two forms of imputation. When people say imputation, they usually mean single.
- The term "single" refers to the fact that you only use one of the seven methods to estimate the missing number outlined above.
- It's popular since it's simple to understand and generates a sample with the same number of observations as the complete data set.
- When listwise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions.
- Unless the data is Missing Completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations. The bias is frequently worse than with listwise deletion, which is most software's default.
- The level of the bias is determined by a number of factors, including the imputation technique, the missing data mechanism, the fraction of missing data, and the information in the data set.

Furthermore, standard errors are underestimated by all single imputation approaches.Because the imputed observations are estimates, their values have a random error associated with them. However, your programme is unaware of this when you enter that estimate as a data point. As a result, it ignores the additional source of error, resulting in too-small standard errors and p-values.

And, while imputation is straightforward in theory, it is difficult to master in reality. As a result, it isn't perfect, although it may suffice in some circumstances.

As a result of multiple imputation, numerous estimates are generated. In multiple imputation, two of the approaches indicated above–hot deck and stochastic regression–work as the imputation method.

The multiple estimates varied significantly because these two approaches contain a random component. This reintroduces some variance that your program can account for in order to provide reliable standard error estimates for your model.

About 20 years ago, multiple imputation was a big advance in statistics. It eliminates many (but not all) difficulties with missing data and, when done correctly, leads to unbiased parameter estimations and accurate standard errors.

12. **What is A/B testing?**

A/B testing is one of the most important concepts in data science and in the tech world in general because it is one of the most effective methods in making conclusions about any hypothesis one may have. It's important that you understand what A/B testing is and how it generally works.

A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric. Typically, two consumer groups are exposed to two different versions of the same thing to see if there is a significant difference in metrics like sessions, click-through rate, and/or conversions.

Using the visual above as an example, we could randomly split our customer base into two groups, a control group and a variant group. Then, we can expose our variant group with a red website banner and see if we get a significant increase in conversions. It's important to note that all other variables need to be held constant when performing an A/B test.

Getting more technical, A/B testing is a form of statistical and two-sample hypothesis testing. **Statistical hypothesis testing** is a method in which a sample dataset is compared against the population data. **Two-sample hypothesis testing** is a method in determining whether the differences between the two samples are statistically significant or not.

It's important to know what A/B testing is and how it works because it's the best method in quantifying changes in a product or changes in a marketing strategy. And this is becoming increasingly important in a data-driven world where business decisions need to be back by facts and numbers.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is typically considered terrible practice since it 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?

Simple linear regression is used to estimate the relationship between two quantitative variables. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

Simple linear regression formula

The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

- y is the predicted value of the dependent variable (y) for any given value of the independent variable (x).
- B_0 is the intercept, the predicted value of y when the x is 0.
- B_1 is the regression coefficient – how much we expect y to change as x increases.
- x is the independent variable (the variable we expect is influencing y).
- e is the error of the estimate, or how much variation there is in our estimate of the regression coefficient.

Linear regression finds the line of best fit line through your data by searching for the regression coefficient (B_1) that minimizes the total error (e) of the model.

15. What are the various branches of statistics?

There is two major areas of statistics are descriptive and inferential statistics.

1.Descriptive Statistics

Descriptive statistics is considered as the first part of statistical analysis which deals with collection and presentation of data. Scientifically, descriptive statistics can be defined as brief explanatory coefficients that are used by statisticians to summarize a given data set. Generally, a data set can either represent a sample of a population or the entire populations.

Descriptive statistics can be categorized into 2 parts.

- Measures of central tendency
- Measures of variability

To easily understand the analyzed data, both measures of tendency and measures of variability use tables, general discussions, and graphs.

Measures of Central Tendency

Measures of central tendency specifically help the statisticians to estimate the center of values distribution. These measures of tendency are:

- Mean

  This is the conventional method used in describing central tendency. Usually, to compute an average of values, you add up all the values and then divide them with the number of values available.

- Median

  This is the score found at the middle of a set of values. A simple way to calculate a median is to arrange the scores in numerical orders and then locate the score which is at the center of the arranged sample.

- Mode

  This is the frequently occurring value in a given set of scores.

Measures of Variability

The measure of variability help statisticians to analyze the distribution spread out of a given set of data. Some of the examples of measures of variability include quartiles, range, variance and standard deviation.

2.Inferential Statistics

Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population. Inferential statistics often talks in probability terms by using descriptive statistics. These techniques are majorly used by statisticians to analyze data, make estimates and draw conclusions from the limited information which is obtained by sampling and testing how reliable the estimates are.
The different types of calculation of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis