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

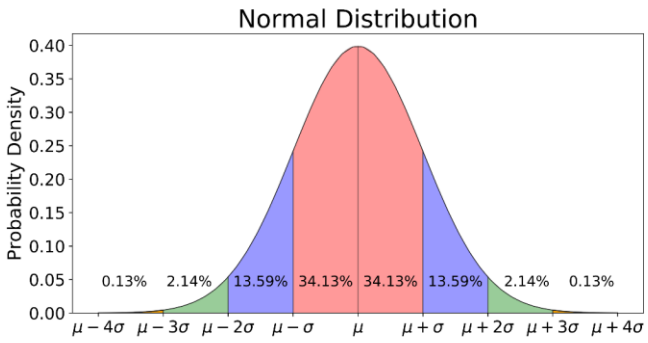
10. What do you understand by the term Normal Distribution?
Ans
Normal distribution is known as the Gaussian distribution, it is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
The normal distribution 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

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, 2.14% of the observations will appear within plus or minus one standard deviation of the mean; 13.59.% of the observations will fall within +/- two standard deviations; and 34.13% 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.



The general formula for the probability density function of the normal distribution is given

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

where,

μ is the location parameter, and

σ is the scale parameter

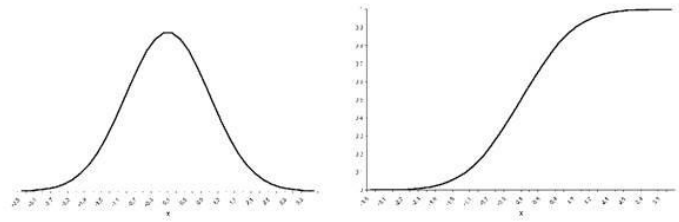
cumulative Density Function (CDF)

The formula for the cumulative distribution function of the normal distribution is given by,

$$F(x) = p(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

Remember that the above integral does not exist in a simple closed formula. It is computed numerically.



a) Normal Probability Density Function b) Normal Cumulative Distribution Function

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

Ans

Missing data can skew anything for data scientists, from economic analysisto clinical trials. After all, any analysis is only as good as the data. A data scientist doesn't want to produce biased estimates that lead to invalid results. The concept of missing data is implied in the name: it's data that is not captured for a variable for the observation in question. Missing data reduces the statistical power of the analysis

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.

12. What is A/B testing?

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.

One way to perform the test is to calculate **daily conversion rates** for both the treatment and the control groups. Since the conversion rate in a group on a certain day represents a single data point, the sample size is actually the number of days. Thus, we will be testing the difference between the mean of daily conversion rates in each group across the testing period.

When we run our experiment for one month, we noticed that the mean conversion rate for the Control group is 16% whereas that for the test Group is 19%..

13. Is mean imputation of missing data acceptable practice?

Ans

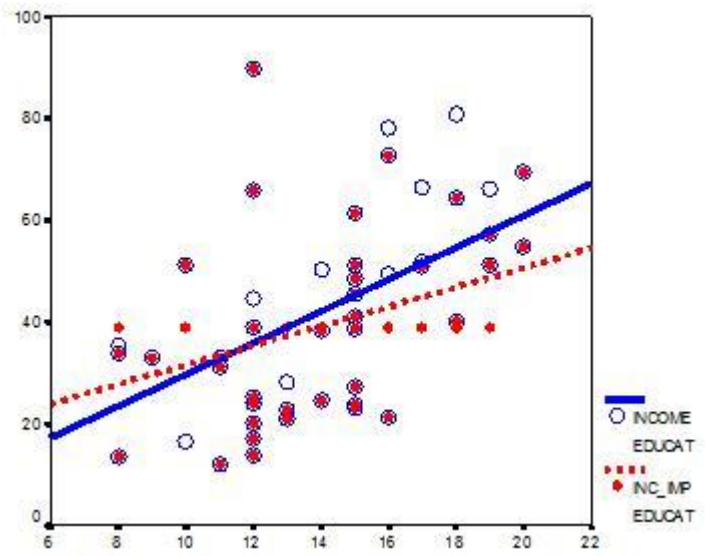
It’s a popular solution to missing data, despite its drawbacks. Mainly because it’s easy. It can be really painful to lose a large part of the sample you so carefully collected, only to have little power.

But that doesn't make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population.

On the other hand, there are many alternatives to mean imputation that provide much more accurate estimates and standard errors, so there really is no excuse to use it.

This post is the first explaining the many reasons not to use mean imputation (and to be fair, its advantages).

First, a definition: mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.



This graph illustrates hypothetical data between X=years of education and Y=annual income in thousands with n=50. The blue circles are the original data, and the solid blue line indicates the best fit regression line for the full data set. The correlation between X and Y is r = .53.

I then randomly deleted 12 observations of income (Y) and substituted the mean. The red dots are the mean-imputed data.

Blue circles with red dots inside them represent non-missing data. Empty Blue circles represent the missing data. If you look across the graph at Y =

39, you will see a row of red dots without blue circles. These represent the imputed values.

The dotted red line is the new best fit regression line with the imputed data. As you can see, it is less steep than the original line. Adding in those red dots pulled it down.

The new correlation is r = .39. That's a lot smaller that .53.

The real relationship is quite underestimated.

Of course, in a real data set, you wouldn't notice so easily the bias you're introducing. This is one of those situations where in trying to solve the lowered sample size, you create a bigger problem.

One note: if X were missing instead of Y, mean substitution would artificially *inflate* the correlation.

In other words, you'll think there is a stronger relationship than there really is. That's not good either. It's not reproducible and you don't want to be overstating real results.

This solution that is so good at preserving unbiased estimates for the mean isn't so good for unbiased estimates of relationships.

14. What is linear regression in statistics?

Ans

Linear regression strives to show the relationship between two variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable. For example, the weight of the person is linearly related to his height. Hence this shows a linear relationship between the height and weight of the person. As the height is increased, the weight of the person also gets increased.

It is not necessary that here one variable is dependent on others, or one causes the other, but there is some critical relationship between the two variables. In such cases, we use a scatter plot to imply the strength of the relationship between the variables. If there is no relation or linking between the variables, the scatter plot does not indicate any increasing or decreasing pattern. For such cases, the linear regression design is not beneficial to the given data.

Y = a + bX

where X is the independent variable and plotted along the x-axis

Y is the dependent variable and plotted along the y-axis

The slope of the line is b, and a is the intercept (the value of y when x = 0).

$$a = \frac{[(\sum y)(\sum x^2) - (\sum x)(\sum xy)]}{[n(\sum x^2) - (\sum x)^2]}$$

$$b = \frac{[n(\sum xy) - (\sum x)(\sum y)]}{[n(\sum x^2) - (\sum x^2)]}$$

Linear regression determines the straight line, called the least-squares regression line or LSRL, that best expresses observations in a bivariate analysis of data set. Suppose Y is a dependent variable, and X is an independent variable, then the population regression line is given by;

$Y = B_0 + B_1X$

Where

B_0 is a constant

B_1 is the regression coefficient

If a random sample of observations is given, then the regression line is expressed by;

$\hat{y} = b_0 + b_1x$

where b_0 is a constant, b_1 is the regression coefficient, x is the independent variable, and $\hat{y}$ is the predicted value of the dependent variable.

15. What are the various branches of statistics?

There are three branches of statistics: **data collection, descriptive statistics and inferential statistics**.

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. Examples of data sets are: marks in a class test: 9, 2, 5, 8, 10, 3, 5, 8, 8, 9

inflation rate: 2.1, 3.2, 4.1, 2.3, 5.1, 2.2, 0.5 voting intention in a referendum: Yes, No, No, Yes, Yes, No

Key term:….

Data collection

Descriptive statistics

Inferential statistics

Discrete and continuous data

Frequency distributions

Statistics is essentially the study of data. It is used in a huge variety of areas; virtually any subject will need some element of data analysis and study. Remember that there are various aspects to statistics: the actual data collection, the presentation of the data (descriptive statistics), and the conclusions that can be drawn (inferential statistics). We shall of course explore this in much more detail as we go through the book.