WORKSHEET-1_STATISTICS

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 **ANSWER>>(A)** b) False

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

ANSWER>>(A)

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

a) Modelling event/time data b) Modelling bounded count data

c) Modelling contingency tables d) All of the mentioned

ANSWER>>(B)

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 **ANSWERS>>(D)**

5. _____ random variables are used to model rates.

a) Empirical b) Binomial

c) Poisson **ANSWER>>(C)** d) All of the mentioned

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

WORKSHEET-1_STATISTICS

a) True b) False

ANSWER>>>(B)

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

a) Probability b) Hypothesis

c) Causal **ANSWERS>>(B)** d) None of the mentioned

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 **ANSWER>>>(A)**

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 **ANSWER>>>(D)**

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

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

ANSWER>>>

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.

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

ANSWER>>

The problem of missing value is quite common in many real-life datasets. Missing value can bias the results of the machine learning models and/or reduce the accuracy of the model

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset

In the dataset, blank shows the missing values.

In Pandas, usually, missing values are represented by **NaN**.

Reasons for the missing data from the dataset affect the approach of handling missing data. So, it's necessary to understand why the data could be missing.

Some of the reasons are listed below:

- Past data might get corrupted due to improper maintenance.
- Observations are not recorded for certain fields due to some reasons. There might be a failure in recording the values due to human error.
- The user has not provided the values intentionally.
- Many machine learning algorithms fail if the dataset contains missing values. However, algorithms like K-nearest and Naive Bayes support data with missing values.
- You may end up building a biased machine learning model which will lead to incorrect results if the missing values are not handled properly.
- Missing data can lead to a lack of precision in the statistical analysis.

There are 2 primary ways of handling missing values:

1. **Deleting the Missing values**

    - Deleting the entire row

- Deleting the entire column

2.Imputing the Missing Values

- Replacing With Arbitrary Value
- Replacing With Mean
- Replacing With Mode
- Replacing With Median
- Replacing with previous value – Forward fill
- Replacing with next value – Backward fill
- Interpolation

➢ **Imputing Missing Values for Categorical Features**
Imputation of Missing Value Using sci-kit learn Library

- Univariate Approach
- Multivariate Approach
- Nearest Neighbors Imputations (KNNImputer)

End Notes

- It is critical to reduce the potential bias in the machine learning models and get the precise statistical analysis of the data.
- Handling missing values is one of the challenges of data analysis.
- Understanding different categories of missing data help in making decisions on how to handle it.
- We explored different categories of missing data and the different ways of handling it in this article.
- Missing values handling is a gigantic topic. In any case, it's very important to understand your data well and why it's missing, talk to the experts if possible to figure out what's going on with the data before blindly following any of the above methods.

12. What is A/B testing?

ANSWER>>>

A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

WORKSHEET-1_STATISTICS

Post learning about four different types of A/B testing experimentation methods, it's equally important to understand which statistical approach to adopt to successfully run an A/B test and draw the right business conclusion.

Ideally, there are two types of statistical approaches used by A/B/n experimenters across the globe: Frequentist and Bayesian. Each of these approaches has its own pros and cons. However, we, at VWO, use, support, and promote the Bayesian approach

Some goals of a media and publishing business may be to increase readership and audience, to increase subscriptions, to increase time spent on their website by visitors, or to boost video views and other content pieces with social sharing and so on. You may try testing variations of email sign-up modals, recommended content, social sharing buttons, highlighting subscription offers, and other promotional options.

13. Is mean imputation of missing data acceptable practice?

ANSWERS>>>

Mean imputation: So simple. And yet, so dangerous.

Perhaps that's a bit dramatic, but mean imputation (also called mean substitution) really ought to be a last resort.

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 is explaining for the many reasons not to use mean imputation (and to be fair, its advantages.

WORKSHEET-1_STATISTICS

Mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

Problem 1: Mean imputation does not preserve the relationships among variables.

True, imputing the mean preserves the mean of the observed data. So, if the data are missing completely at random, the estimate of the mean remains unbiased. That's a good thing.

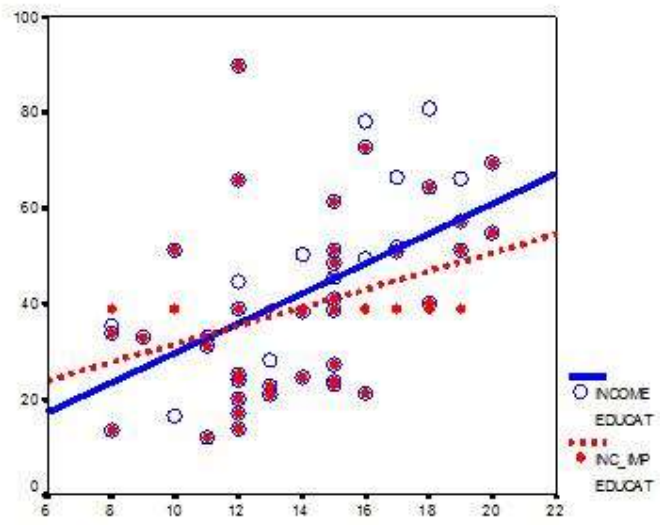
Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too.

This is the original logic involved in mean imputation.

If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate.

It will still bias your standard error.

Since most research studies are interested in the relationship among variables, mean imputation is not a good solution. The following graph illustrates this well:



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.

WORKSHEET-1_STATISTICS

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 than .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.

Problem 2: Mean Imputation Leads to An Underestimate of Standard Errors

A second reason is applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low.

In other words, yes, you get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small.

Because the imputations are themselves estimates, there is some error associated with them. But your statistical software doesn't know that. It treats it as real data.

WORKSHEET-1_STATISTICS

Ultimately, because your standard errors are too low, so are your p-values. Now you're making Type I errors without realizing it.That's not good.

14. What is linear regression in statistics?

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.

15. What are the various branches of statistics?

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

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.

WORKSHEET-1_STATISTICS

- 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.