Q&A

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

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

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

a) Modeling event/time data

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

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

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

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

a) True

b) False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

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

a) 0

b) 5

c) 1

d) 10

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

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

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

Gaussian distribution known as Normal distribution or Bell curve is a probability distribution/function that is characterized to be symmetrical to the mean. Means that data near to mean is more frequent to happen than data further away from the mean. Extreme values are unlikely to happen.
The graph of the normal distribution is characterized by two parameters: the mean (average) what is high point of graph with high response and, as we said, where the graph is symmetric; and the standard deviation, which determines the amount of dispersion away from the mean. A small standard deviation, from the mean, produces a steep graph, and a large standard deviation produces a flat graph.
Distribution shape is said to be perfectly symmetric. Because half of the observations in the data get captured on one side of the middle of the distribution. And the middle point of a normal distribution is the point that has the maximum frequency, means the majority of data points in the middle of the distribution are similar and they occur within a small range of values with fewer outliers on the both ends of the data range. The midpoint of the normal distribution is at which three measures fall: the mean, median, and mode. In a perfectly normal distribution, these three measures are all the same number.
Normal distributions are represented in standard or Z scores, which shows us the distance between an actual and the mean in terms of standard deviations. The standard normal distribution has a mean of 0.0 and a standard deviation of 1.0.
Normal density function:

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

11. How do you handle missing data? What imputation techniques do you recommend?
We can delete it or imputate it. The quick and easy way is to use a mean for numerical features and a mode for categorical ones. Others just insert 0's or discard the missing data and proceed with training if there is no important change in the accuracy, variation and bias the results. But, a mean or mode can reduce the model's accuracy and bias the results.
Or another option is to let the algorithm handle the missing data. Some algorithms can do the best imputation for the missing data based on the training loss reduction (ie. XGBoost). Some others just ignore the missing values (ie. LightGBM). However, other algorithms throw an error when we have missing values (ie. Scikit learn — LinearRegression).
2- Imputation Using (Mean/Median) Values: calculates the mean/median of the non-missing values in a column and then replaces the missing values within each column with the estimated mean or median. Pros: Works well with small numerical datasets. But doesn't work well with the correlations between features. It only works on the column level. Will give poor results on encoded categorical features (do NOT use it on categorical features). Not very accurate.
3- Imputation Using (Most Frequent) or (Zero/Constant) Values: Most Frequent is imputation of missing values even when we have categorical features (strings or numerical representations) by replacing missing data with the most frequent values within each column. Pros: Works well with categorical features. But it doesn't factor the correlations between features. It can introduce bias in the data.
Zero or Constant imputation— it replaces the missing values with either zero or any constant value.
4- Imputation Using k-NN: The k nearest neighbours is an algorithm used for simple classification. The new point is assigned a value based on how closely it resembles the points in the training set.
Pros: Can be much more accurate than the mean, median or most frequent imputation methods.
But it is computationally expensive. KNN works by storing the whole training dataset in memory.
K-NN is quite sensitive to outliers in the data (unlike SVM).
5- Imputation Using Multivariate Imputation by Chained Equation (MICE): This type of imputation works by filling the missing data multiple times. Multiple Imputations (MIs) are much better than a single imputation as it measures the uncertainty of the missing values in a better way.
A new technique for missing data imputation named Single Center Imputation from Multiple Chained Equation(SICE) is a hybrid approach of single and multiple imputation methods.
6- Imputation Using Deep Learning (Datawig): This method works categorical and non-numerical features. It is a library that use Deep Neural Networks and handles categorical data (Feature Encoder). It supports CPUs and GPUs. But as Single Column imputation, it can be quite slow with large datasets.
Other imputation models like Stochastic regression imputation: tries to predict the missing values by calculating it from other related variables in the same dataset plus some random residual value.
Extrapolation and Interpolation: It tries to estimate values from other observations within the range of a discrete set of known data points.

Hot-Deck imputation: Works by choosing the missing value from a set of related and similar variables.

Imputation vs. Removing Data

When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. There are primary methods for deleting data when dealing with missing data: listwise, Pairwise and dropping variables.

Before deciding which approach to employ, data scientists must understand why the data is missing (MAR, MCAR, MNAR). Data may be missing due to test design, failure in the observations or failure in recording observations. This type of data is seen as MCAR because the reasons for its absence are external and not related to the value of the observation.

IMPUTATION

When data is missing, it may make sense to delete data, as mentioned above. However, that may not be the most effective option. Instead of deletion, data scientists choose to impute the value of missing data in order to avoid altering the standard deviation or curve of the distribution. However, as in a mean substitution, while a regression imputation substitutes a value that is predicted from other variables, no new information is added, while the sample size has been increased and the standard error is reduced.

Mean, Median and Mode are one of the most common methods of imputing values when dealing with missing data. However, when there are many missing variables, mean or median results can result in a loss of variation in the data. This method does not use time-series characteristics or depend on the relationship between the variables.

Mean substitution: the mean value of a variable is used in place of the missing data value for that same variable. However, with missing values that are not strictly random, the mean substitution method may lead to inconsistent bias. Furthermore, this approach adds no new information but only increases the sample size and conducts to an underestimate of the errors. Thus, mean substitution is not generally accepted.

Single imputation: When using single imputation, missing values are replaced by a value defined by a certain rule . There are many forms of single imputation, for example, last observation carried forward (a participant's missing values are replaced by the participant's last observed value), worst observation carried forward (a participant's missing values are replaced by the participant's worst observed value), and simple mean imputation. In simple mean imputation, missing values are replaced by the mean for that variable. Often result in an underestimation of the variability.

Multiple imputation: Above we discussed about "single imputation": each value in the dataset is filled in exactly once. In general, the limitation with single imputation is that because these techniques find maximally likely values, they do not generate entries which accurately reflect the distribution of the underlying data.

The most advanced methodology for performing missing data imputation is multiple imputation. In multiple imputation we generate missing values from the dataset many times. The individual datasets are then pooled together into the final imputed dataset, with the values chosen to replace the missing data being drawn from the combined results in some way. In other words, multiple imputation breaks imputation out into three steps: imputation (multiple times), analysis (staging how the results should be combined), and pooling (integrating the results into the final imputed matrix). The missing values are exchanged for values that encompass the natural variability and uncertainty of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures, and the multiple analysis results are combined to produce an overall result.

Time-Series Specific Methods: Another option is to use time-series specific methods when appropriate to impute data. There are four types of time-series data: 1) No trend or seasonality. 2) Trend, but no seasonality. 3) Seasonality, but no trend. 4) Both trend and seasonality. The time series methods of imputation assume the adjacent observations will be like the missing data. However, these methods won't always produce reasonable results, particularly in seasonality.

It is not recommend to use LOCF or NOCB : Single imputation methods like last observation carried forward and baseline observation carried forward should not be used as the primary approach to the treatment of missing data unless the assumptions that underlie them are scientifically justified.

Linear Interpolation: is often used to approximate a value of some function by using two known values of that function at other points. This formula can also be understood as a weighted average. When dealing

with missing data, you should use this method in a time series that exhibits a trend line, but it's not appropriate for seasonal data.
Maximum likelihood: There are a number of strategies using the maximum likelihood method to handle the missing data. In these, the assumption that the observed data are a sample drawn from a multivariate normal distribution is relatively easy to understand. After the parameters are estimated using the available data, the missing data are estimated based on the parameters which have just been estimated. When there are missing, the statistics explaining the relationships among the variables may be computed using the maximum likelihood method. That is, the missing data may be estimated by using the conditional distribution of the other variables.
Expectation-Maximization
Expectation-Maximization (EM) is a type of the maximum likelihood method that can be used to create a new data set, in which all missing values are imputed with values estimated by the maximum likelihood methods. This approach can lead to the biased parameter estimates and can underestimate the standard error.

In conclusion, there is the best way to impute the missing values but each strategy can be better for different datasets and type of missing data. Single imputation based missing data handling methods are easy to implement but may provide biased imputations, according to statisticians. On the other hand, multiple imputation based methods consider the uncertainty of a dataset and generate a set of plausible values for each missing data, which are complex to implement. We have rules to follow in order to decide best imputation algorithm but you should experiment and check which model works best for your dataset. We also have 2 imputations algorithm from Sklearn.impute package: SimpleImputer for imputations on univariate datasets; and IterativeImputer for imputations on multivariate datasets. And we also have Missingpy for missing values imputation. It supports K-Nearest Neighbours and MissForest i.e Random Forest-based imputation technique.

12. What is A/B testing?
Split-run testing known as A/B testing is the experiment where two versions of a tested element is shown to different web-site visitor and determine which element version has the maximum impact and uplift the business metrics/ROI.
A/B testing is a technique that involves comparing two versions of a web page or application to check and validate which performs better. These variations, known as A and B, are presented to users randomly. In A/B testing, in general, A refers to the original testing variable and B refers to a new version of the original testing variable. More likely, a half portion of segment of visitors will be sent to the first version, and the rest to the second. A statistical analysis of the reactions determines which version, A or B, performed the best according to predefined indicators, for example, conversion rate (CRO). Running an A/B test can collect and analyze the impact of that change.
A/B testing is also known as split testing, which can be either the same thing as A/B testing or mean split URL testing. In case of A/B test, the two variations are on the same URL. On the other hand, with split URL testing our varied element is on a different URL.

13. Is mean imputation of missing data acceptable practice? NO.
Problem 1: Mean imputation does not preserve the relationships among variables.
It has advantage of being good at preserving unbiased estimates for the mean but it isn't so good to calculate unbiased estimates of relationships of variables. Disfigures relationships between variables by calculating estimates of the correlation toward zero. And it also ignores potential feature correlation.
Problem 2: Mean Imputation Underestimate Standard Errors/deviation
It is right to say you can get the same mean from mean-imputed data that you would have taken without imputing. We can have our mean unbiased. Still, the standard error of that mean will be small. Mean imputation reduces the variance of the data and their mean imputed will change. A smaller variance induce to a small confidence interval in the probability distribution. This introduces a bias to our model. Mean substitution is biased in multivariate estimates such as correlation or regression coefficients. Values that are imputed by a variable's mean have, in general, a correlation of zero with other variables. Thus, relationships between variables are biased toward zero. Because the imputations are themselves estimates, means we have some error associated with them. If our standard errors are too low, p-values also differ and will be low. Which means we're making Type I errors.

Alternative solution would be MICE, KNN, Miss Forest, Fuzzy K-means Clustering from Python and Imputation Libraries from Scikit-Learn library.

14. What is linear regression in statistics?

Simple linear regression is a statistical method that summarize and analyze relationship within two or more continuous variables.

Linear regression is a basic type of predictive analysis.

These regression estimates are employed to explain the relationship between dependent and independent variables. Linear regression modelizes the relationship between variables by fitting a linear equation to considered and observed data in order to check:

If a set of predictor variables do a good job in predicting an outcome.

which variables in particular are significant of the outcome and how?

The simple form of regression with one dependent and one independent variable is described by the formula y = a + b*x, where "y" is a dependent/response/outcome variable, "a" is constant, "b" is the regression coefficient, and "x" is independent/predictor/explanatory variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

Three major uses for regression analysis are determining the strength of predictors, forecasting an effect, and trend forecasting.

If there is no association between the explanatory/independent and dependent variables (let say, scatterplot does not show any increase or decrease trend), then a linear regression will not be a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

15. What are the various branches of statistics?

The two main branches of statistics are descriptive statistics and inferential statistics.

Descriptive statistics works with the collection and presentation of data (first part of statistics). It is based on the explanatory coefficients that summarize a given data set.

It describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of variability like range, quartiles, standard deviation, variance etc. Data is represented in any reasonable chart, tables or graphs.

Inferential statistics draws the conclusions from the statistical analysis that has been obtained and informed using descriptive statistics. Most predictions and generalizations about a population by studying a sample goes under the branch of inferential statistics.

Inferential statistics are techniques that gather information from a sample in order to make inferences, decisions or predictions about a given population. These techniques are used to analyze data, make estimates and conclude the information which is obtained by sampling and testing.. And generalize them to the population.

The calculation used in inferential statistics includes: Regression analysis, Analysis of variance (ANOVA), Analysis of covariance (ANCOVA), Statistical significance (t-test) and Correlation analysis.