STATISTIC WORKS

ASSIGNMENT 5

Answer 1 True (the Bernoulli distribution a raise as the result of a binary outcome)

Answer2 A (Central Limit Theorem)

Answer 3 B ( Modeling bounded count data Poisson distribution is used for modeling unbound count data)

Answer4 B

Answer 5 C (Poisson)

Answer 6 B False (Usually replacing the standard error by its estimate value doesn’t change the CLT)

Answer 7 B (Hypothesis)

Answer 8 A (0)

Answer 9 C (outlier cannot conform to the regression relationship)

Answer 10 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.

Answer 10 The Normal Distribution, also known as the Gaussian distribution, is a continuous probability distribution for any dataset, which can be further represented as a Bell Curve. The Normal Distribution is used to analyze data when there is an equal chance for the data to be above and below the average value of the continuous data. It is named after the famous mathematician and physicist Carl Friedrich Gauss.

The 5 common properties of Normal Distribution are:

1.Normal Distribution Curve is symmetric about the mean.

2.Normal Distribution is unimodal in nature, i.e., it has single peak value.

3.Normal Distribution Curve is always bell-shaped.

4.Mean, mode, and median for Normal Distribution is always same.

5.Normal Distribution follows the empirical rule.

We define Normal Distribution as the probability density function of any continuous random variable for any given system. Now for defining Normal Distribution suppose we take f(x) as the probability density function for any random variable X. Also, the function is integrated between the interval, (x, {x + dx}) then,

f(x) ≥ 0 ∀ x ϵ (−∞,+∞),

-∞∫+∞ f(x) = 1

We observe that the curve traced by the upper values of the Normal Distribution is in the shape of a Bell, hence Normal Distribution is also called the “Bell Curve”.

Normal Distribution Examples

We can draw Normal Distribution for various types of data that include,

* Distribution of Height of People
* Distribution of Errors in any Measurement
* Distribution of Blood Pressure of any Patient, etc.

Answer11 Imputation is the process of replacing missing values with substituted data. It is done as a preprocessing step.

* Mean
* Median
* Mode

The common convention is, if you have large amount of data (rows) AND you will still be left with a large enough number of data points, representing the various different patterns and cases, that the ML model can learn from. Then it’s probably a good idea to drop the missing data points. Because, filling in with data like the mean, median etc, is going to introduce some amount of approximation (incorrectness) to it. So if you have large volumes of data, you might wish to drop those records instead of introducing approximate values that could affect the ML model performance instead of enhancing it.An exception of course is when you KNOW for certain what the data should be if the values are missing.

For example:- if the data is about exam scores which contains missing values, because the candidates didn’t give the exams. In that case, you might want to replace the missing with 0.

Answer 12 In statistical terms, A/B testing is a method of two-sample hypothesis testing. This means comparing the outcomes of two different choices (A and B) by running a controlled mini-experiment. This method is also sometimes referred to as split testing.

Answer 13 First, a definition: mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

1. Mean imputation does not preserve the relationships among variablesTrue, 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, but I will get to that in another post.

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

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.Ultimately, because your standard errors are too low, so are your p-values. Now you’re making Type I errors without realizing it.

Answer14 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.

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.

Answer15 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.