**STATISTICS WORKSHEET-1**

**1. Bernoulli random variables take (only) the values 1 and 0.**

a) True b) False

**ANSWER**---**A) TRUE**

**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) CENTRAL LIMIT THEOREM**

**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

**ANSWER**---**B) BOUNDED COUNT DATA**

**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

**ANSWER**--- **D) ALL OF THE MENTIONED**

**5. \_\_\_\_\_\_ random variables are used to model rates**.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

**ANSWER**--- **C) POISSION**

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

a) True b) False

**ANSWER**---**B) FALSE**

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

**ANSWER**--- **B) HYPOTHESIS**

**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) 0**

**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

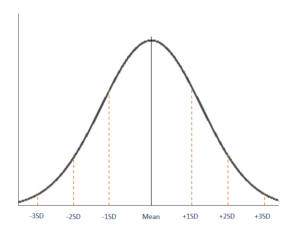
d) None of the mentioned

**ANSWER**--- **C) OUTLINER CANNOT CONFORM TO THE REGRESSION RELATIONSHIP**

**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 Normal distribution will appear as a bell curve 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. For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

The normal distribution model is motivated by the Central Limit Theorem. 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). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.



Shape of Normal Distribution

A normal distribution is symmetric from the peak of the curve, where the mean is. This means that most of the observed data is clustered near the mean, while the data become less frequent when farther away from the mean. The resultant graph appears as bell-shaped where the mean, median, and mode are of the same values and appear at the peak of the curve.

The graph is a perfect symmetry, such that, if you fold it at the middle, you will get two equal halves since one-half of the observable data points fall on each side of the graph.

Parameters of Normal Distribution

The two main parameters of a (normal) distribution are the mean and standard deviation. The parameters determine the shape and probabilities of the distribution. The shape of the distribution changes as the parameter values change.

1. Mean

The mean is used by researchers as a measure of central tendency. It can be used to describe the distribution of variables measured as ratios or intervals. In a normal distribution graph, the mean defines the location of the peak, and most of the data points are clustered around the mean. Any changes made to the value of the mean move the curve either to the left or right along the X-axis.

2. Standard Deviation

The standard deviation measures the dispersion of the data points relative to the mean. It determines how far away from the mean the data points are positioned and represents the distance between the mean and the observations.

On the graph, the standard deviation determines the width of the curve, and it tightens or expands the width of the distribution along the x-axis. Typically, a small standard deviation relative to the mean produces a steep curve, while a large standard deviation relative to the mean produces a flatter curve.

Properties

All forms of (normal) distribution share the following characteristics:

1. It is symmetric

A normal distribution comes with a perfectly symmetrical shape. This means that the distribution curve can be divided in the middle to produce two equal halves. The symmetric shape occurs when one-half of the observations fall on each side of the curve.

2. The mean, median, and mode are equal

The middle point of a normal distribution is the point with the maximum frequency, which means that it possesses the most observations of the variable. The midpoint is also the point where these three measures fall. The measures are usually equal in a perfectly (normal) distribution.

3. Empirical rule

In normally distributed data, there is a constant proportion of distance lying under the curve between the mean and specific number of standard deviations from the mean. For example, 68.25% of all cases fall within +/- one standard deviation from the mean. 95% of all cases fall within +/- two standard deviations from the mean, while 99% of all cases fall within +/- three standard deviations from the mean.

4. Skewness and kurtosis

Skewness and kurtosis are coefficients that measure how different a distribution is from a normal distribution. Skewness measures the symmetry of a normal distribution while kurtosis measures the thickness of the tail ends relative to the tails of a normal distribution.

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

**ANSWER--**Missing data are defined as not available values, and that would be meaningful if observed. Missing data can be anything from missing sequence, incomplete feature, files missing, information incomplete, data entry error etc. Most datasets in the real world contain missing data. Before you can use data with missing data fields, you need to transform those fields to be used for analysis and modelling.

**1. Deleting Rows**

This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set. One has to make sure that after we have deleted the data, there is no addition of bias. Removing the data will lead to loss of information which will not give the expected results while predicting the output.

**2. Replacing With Mean/Median/Mode**

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training. Another way is to approximate it with the deviation of neighbouring values. This works better if the data is linear.

**3. Assigning An Unique Category**

A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values. Here, the features Cabin and Embarked have missing values which can be replaced with a new category, say, U for ‘unknown’. This strategy will add more information into the dataset which will result in the change of variance. Since they are categorical, we need to find one hot encoding to convert it to a numeric form for the algorithm to understand it. Let us look at how it can be done in Python

**4. Predicting The Missing Values**

Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a very high variance. We will be using linear regression to replace the nulls in the feature ‘age’, using other available features. One can experiment with different algorithms and check which gives the best accuracy instead of sticking to a single algorithm

**5. Predictive/Statistical models that impute the missing data**

Linear Regression

In regression imputation, the existing variables are used to predict, and then the predicted value is substituted as if an actually obtained value. This approach has several advantages because the imputation retains a great deal of data over the listwise or pairwise deletion and avoids significantly altering the standard deviation or the shape of the distribution. However, as in a mean substitution, while a regression imputation substitutes a value predicted from other variables, no novel information is added, while the sample size has been increased and the standard error is reduced.

Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that work well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of the bag) imputation error estimates. One caveat is that random forest works best with large datasets, and using random forest on small datasets runs the risk of overfitting.

k-NN (k Nearest Neighbour)

k-NN imputes the missing attribute values based on the nearest K neighbour. Neighbours are determined based on a distance measure. Once K neighbours are determined, the missing value is imputed by taking mean/median or mode of known attribute values of the missing attribute.

Maximum likelihood

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. Several strategies are using the maximum likelihood method to handle the missing data.

Expectation-Maximization

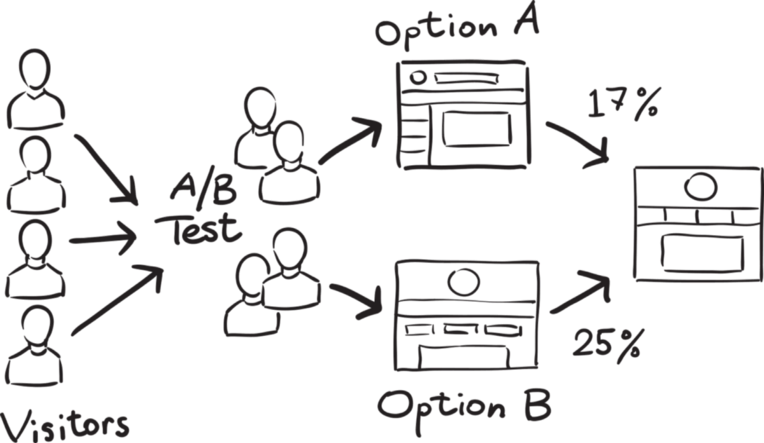
Expectation-Maximization (EM) is the maximum likelihood method used to create a new data set. All missing values are imputed with values estimated by the maximum likelihood methods. This approach begins with the expectation step, during which the parameters (e.g., variances, covariances, and means) are estimated, perhaps using the listwise deletion. Those estimates are then used to create a regression equation to predict the missing data. The maximization step uses those equations to fill in the missing data. The expectation step is then repeated with the new parameters, where the new regression equations are determined to “fill in” the missing data. The expectation and maximization steps are repeated until the system stabilizes.

Sensitivity analysis

Sensitivity analysis is defined as the study which defines how the uncertainty in the output of a model can be allocated to the different sources of uncertainty in its inputs. When analysing the missing data, additional assumptions on the missing data are made, and these assumptions are often applicable to the primary analysis. However, the assumptions cannot be definitively validated for correctness. Therefore, the National Research Council has proposed that the sensitivity analysis be conducted to evaluate the robustness of the results to the deviations from the MAR assumption.

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

**ANSWER---**A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let’s say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.In the scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B’s packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

**Objective**

Our objective here is to check which newsletter brings higher traffic on the website i.e the conversion rate. We will use A/B testing and collect data to analyze which newsletter performs better.

**1. Make a Hypothesis**

A hypothesis is a tentative insight into the natural world; a concept that is not yet verified but if true would explain certain facts or phenomena.

It is an educated guess about something in the world around you. It should be testable, either by experiment or observation. In our example, the hypothesis can be “By making changes in the language of the newsletter, we can get more traffic on the website”.

In hypothesis testing, we have to make two hypotheses i.e Null hypothesis and the alternative hypothesis. Let’s have a look at both.

Null hypothesis or H0:

The null hypothesis is the one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is no difference between the control and variant groups. It states the default position to be tested or the situation as it is now, i.e. the status quo. Here our H0 is ” there is no difference in the conversion rate in customers receiving newsletter A and B”.

Alternative Hypothesis or H0:

The alternative hypothesis challenges the null hypothesis and is basically a hypothesis that the researcher believes to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true.In our example, the Ha is- “the conversion rate of newsletter B is higher than those who receive newsletter A“.

Now, we have to collect enough evidence through our tests to reject the null hypothesis.

**2. Create Control Group and Test Group**

Once we are ready with our null and alternative hypothesis, the next step is to decide the group of customers that will participate in the test. Here we have two groups – The Control group, and the Test (variant) group.

The Control Group is the one that will receive newsletter A and the Test Group is the one that will receive newsletter B.

For this experiment, we randomly select 1000 customers – 500 each for our Control group and Test group.

Randomly selecting the sample from the population is called random sampling. It is a technique where each sample in a population has an equal chance of being chosen. Random sampling is important in hypothesis testing because it eliminates sampling bias, and it’s important to eliminate bias because you want the results of your A/B test to be representative of the entire population rather than the sample itself.

Another important aspect we must take care of is the Sample size. It is required that we determine the minimum sample size for our A/B test before conducting it so that we can eliminate under coverage bias. It is the bias from sampling too few observations.

**3. Conduct the A/B Test and Collect the Data**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%.

**Statistical significance of the Test**

There are two types of errors that may occur in our hypothesis testing:

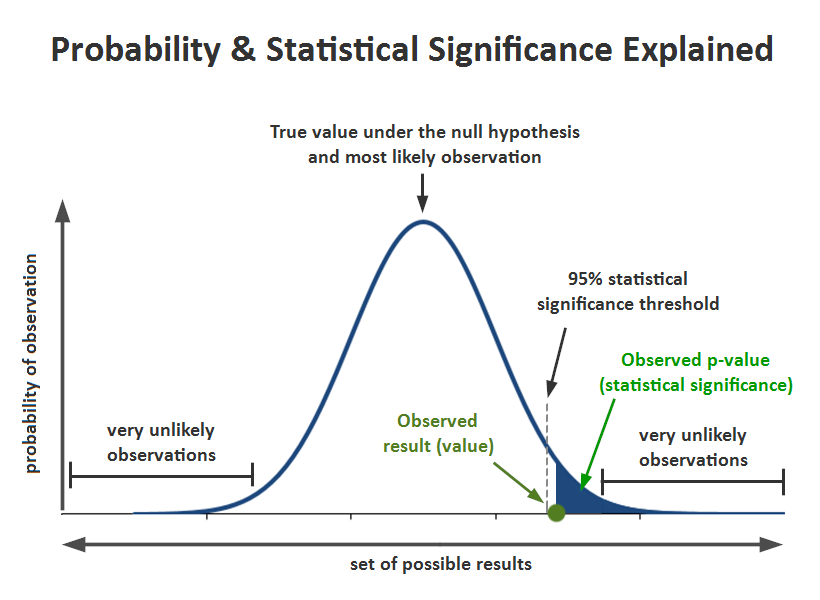
Type I error: We reject the null hypothesis when it is true. That is we accept the variant B when it is not performing better than A

Type II error: We failed to reject the null hypothesis when it is false. It means we conclude variant B is not good when it performs better than A

To avoid these errors we must calculate the statistical significance of our test.

An experiment is considered to be statistically significant when we have enough evidence to prove that the result we see in the sample also exists in the population.

That means the difference between your control version and the test version is not due to some error or random chance. To prove the statistical significance of our experiment we can use a two-sample T-test.The two–sample t–test is one of the most commonly used hypothesis tests. It is applied to compare whether the average difference between the two groups.

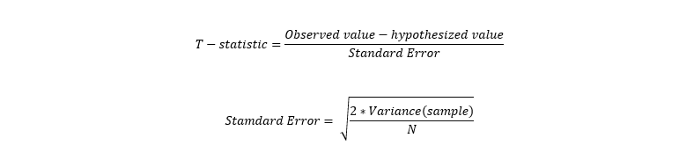


Significance level (alpha): The significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true. Generally, we use the significance value of 0.05

P-Value: It is the probability that the difference between the two values is just because of random chance. P-value is evidence against the null hypothesis. The smaller the p-value stronger the chances to reject the H0. For the significance level of 0.05, if the p-value is lesser than it hence we can reject the null hypothesis

Confidence interval: The confidence interval is an observed range in which a given percentage of test outcomes fall. We manually select our desired confidence level at the beginning of our test. Generally, we take a 95% confidence interval

Next, we can calculate our t statistics using the below formula



**13. Is mean imputation of missing data acceptable practice?**

**ANSWER---**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.

explaining the many reasons not to use mean imputation

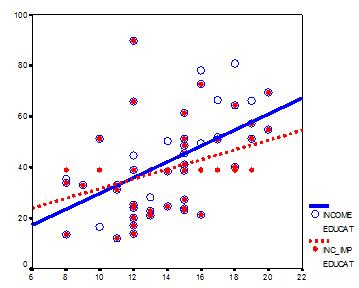
First, a definition: mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

**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, 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.



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.

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

**14. What is linear regression in statistics?**

**ANSWER---**In linear regression we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the criterion variable and is referred to as Y . The variable we are basing our predictions on is called the predictor variable and is referred to as X . When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.The example data in Table 14.1.1 are plotted in Figure 14.1.1 . You can see that there is a positive relationship between X and Y . If you were going to predict Y from X , the higher the value of X , the higher your prediction of Y .

Table 14.1.1 : Example data

X Y

1.00 1.00

2.00 2.00

3.00 1.30

4.00 3.75

5.00 2.25

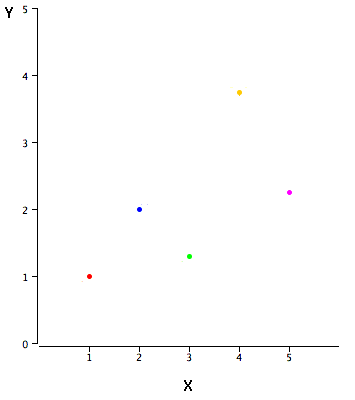


Figure 14.1.1 : A scatter plot of the example data

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. The black diagonal line in Figure 14.1.2 is the regression line and consists of the predicted score on Y for each possible value of X . The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.

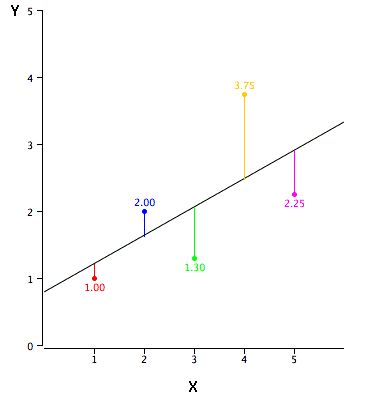


Figure 14.1.2 : A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction

The error of prediction for a point is the value of the point minus the predicted value (the value on the line). Table 14.1.2 shows the predicted values ( Y′ ) and the errors of prediction ( Y−Y′ ). For example, the first point has a Y of 1.00 and a predicted Y (called Y′ ) of 1.21 . Therefore, its error of prediction is −0.21 .

Table 14.1.2 : Example data

X Y Y' Y-Y' (Y-Y')2

1.00 1.00 1.210 -0.210 0.044

2.00 2.00 1.635 0.365 0.133

3.00 1.30 2.060 -0.760 0.578

4.00 3.75 2.485 1.265 1.600

5.00 2.25 2.910 -0.660 0.436

You may have noticed that we did not specify what is meant by "best-fitting line." By far, the most commonly-used criterion for the best-fitting line is the line that minimizes the sum of the squared errors of prediction. That is the criterion that was used to find the line in Figure 14.1.2 . The last column in Table 14.1.2 shows the squared errors of prediction. The sum of the squared errors of prediction shown in Table 14.1.2 is lower than it would be for any other regression line.

The formula for a regression line is

Y′=bX+A(14.1.1)

where Y′ is the predicted score, b is the slope of the line, and A is the Y intercept. The equation for the line in Figure 14.1.2 is

Y′=0.425X+0.785(14.1.2)

For X=1 ,

Y′=(0.425)(1)+0.785=1.21(14.1.3)

For X=2 ,

Y′=(0.425)(2)+0.785=1.64(14.1.4)

**Computing the Regression Line**

the regression line is typically computed with statistical software. However, the calculations are relatively easy, and are given here for anyone who is interested. The calculations are based on the statistics shown in Table 14.1.3 . MX is the mean of X , MY is the mean of Y , sX is the standard deviation of X , sY is the standard deviation of Y , and r is the correlation between X and Y .

Formula for standard deviation

Formula for correlation

Table 14.1.3 : Statistics for computing the regression line

MX MY sX sY r

3 2.06 1.581 1.072 0.627

The slope ( b ) can be calculated as follows: b=rsYsX (14.1.5)

and the intercept ( A ) can be calculated as A=MY−bMX (14.1.6)

For these data,

b=(0.627)(1.072)/1.581=0.425(14.1.7)

A=2.06−(0.425)(3)=0.785(14.1.8)

Note that the calculations have all been shown in terms of sample statistics rather than population parameters. The formulas are the same; simply use the parameter values for means, standard deviations, and the correlation.

**Standardized Variables**

The regression equation is simpler if variables are standardized so that their means are equal to 0 and standard deviations are equal to 1 , for then b=r and A=0 . This makes the regression line:

ZY′=(r)(ZX)(14.1.9)

where ZY′ is the predicted standard score for Y , r is the correlation, and ZX is the standardized score for X . Note that the slope of the regression equation for standardized variables is r .

**15.What are the various branches of statistics?**

**ANSWER---**the two broad branches of statistics. Often, the types of work we do hide many aspects of statistics. However, it is essential to understand the whole idea of statistical analysis for you to feel the beauty of it. The two branches of statistics are descriptive statistics and inferential statistics. All these branches of statistics follow a specific scientific approach which makes them equally essential to every statistics student.

**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

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.

**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

At homeworkdoer.org, we duly understand that statistics can be challenging especially due to the complex formulae involved. However, that should not stress you anymore with unreliable statistics math help. We are here to offer the most affordable and reliable "pay someone to do my homework" services.