**Mid-term examination paper**

Name: Sangar

ID:15570

Date:30-Nov-2020

Subject Name: Probability and statistic.

**Question 01:**

1. Compute the least squares regression equation of Yon X for the following data. What is the regression co-efficient? Also find trend values for X= 5,6,8,10,12,13,15,16,17

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 5 | 6 | 8 | 10 | 12 | 13 | 15 | 16 | 17 |
| Y | 16 | 19 | 23 | 28 | 36 | 41 | 44 | 45 | 50 |

Solution:

The estimation regression line of Y on X is

$$Y=a+bx$$

And two normal equations are.

$$\sum\_{}^{}Y=na+b\sum\_{}^{}x$$

$$\sum\_{}^{}XY=a\sum\_{}^{}x+b\sum\_{}^{}x^{2}$$

To compute the necessary summation, we arrange the computations in the below table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | x | Y | XY | X2 |
|  | 5 | 16 | 30 | 25 |
|  | 6 | 19 | 114 | 36 |
|  | 8 | 23 | 184 | 64 |
|  | 10 | 28 | 280 | 100 |
|  | 12 | 36 | 432 | 144 |
|  | 13 | 41 | 533 | 169 |
|  | 15 | 44 | 660 | 225 |
|  | 16 | 45 | 720 | 256 |
|  | 17 | 50 | 850 | 289 |
| Total  | 102 | 302 | 3853 | 1308 |

Xavg=102/9=11.33, Yavg=302/33.56

$$b=\frac{9×3853-102×302}{9×1308-102^{2}}=2.831$$

a=33.56-2.831x11.33=1.47

Hence the desired estimated regression line of X on Y is:

Y=1.47+2.831X

The estimated regression co-efficient, b=2.831 which determine us that the value of Y increase by 2.831 unites for a unit increase in X.

b. Calculate the coefficient of correlation between X and Y from the following data:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 5 | 6 | 8 | 10 | 12 | 13 | 15 | 16 | 17 |
| Y | 16 | 19 | 23 | 28 | 36 | 41 | 44 | 45 | 50 |

Solution:

R1=$\frac{\sum\_{}^{}\left(5-11.33\right)\left(16-33.56\right)}{\sqrt{\left(5-11.33\right)2\left(16-33.56\right)2}}=\frac{-6.33\*-17.56}{-6.33^{2}\*-17.56^{2}}=\frac{111.548}{\sqrt{12365}}=\frac{111.548}{111.2}=1$

R2=$\frac{\sum\_{}^{}\left(6-11.33\right)\left(19-33.56\right)}{\sqrt{\left(6-11.33\right)2\left(19-33.56\right)2}}=\frac{-5.33\*-14.56}{-5.33^{2}\*-14.56^{2}}=\frac{77.6}{\sqrt{6022.5}}=\frac{77.6}{77.6}=1$

R3=$\frac{\sum\_{}^{}\left(8-11.33\right)\left(23-33.56\right)}{\sqrt{\left(8-11.33\right)2\left(23-33.56\right)2}}=\frac{-3.33\*-10.56}{-3.33^{2}\*-10.56^{2}}=\frac{35.165}{\sqrt{1236.56}}=\frac{35.165}{35.165}=1$

So we can say co-efficient of correlation between X and Y is 1

**Question No03:**

a. Discuss the steps of testing of hypothesis with data example.

b. Differentiate the Concept of Regression and Correlation with data results.

Answer:

Definition: Hypothesis is a statement or assumption about the population parameter under the assumption that it is true.

Discuss:

Step 1: State the Null Hypothesis

The null hypothesis can be thought of as the opposite of the "guess" the research made (in this example the biologist thinks the plant height will be different for the fertilizers). So the null would be that there will be no difference among the groups of plants. Specifically in more statistical language the null for an ANOVA is that the means are the same. We state the Null hypothesis as:

H0: µ1=µ2 = …… µk

for k levels of an experimental treatment.

**Step 2: State the Alternative Hypothesis**

HA: treatment level means not all equal

The reason we state the alternative hypothesis this way is that if the Null is rejected, there are many possibilities.

For example, μ1≠μ2=⋯=μk is one possibility, as is μ1=μ2≠μ3=⋯=μk. Many people make the mistake of stating the Alternative Hypothesis as: μ1≠μ2≠⋯≠μk which says that every mean differs from every other mean. This is a possibility, but only one of many possibilities. To cover all alternative outcomes, we resort to a verbal statement of ‘not all equal’ and then follow up with mean comparisons to find out where differences among means exist. In our example, this means that fertilizer 1 may result in plants that are really tall, but fertilizers 2, 3 and the plants with no fertilizers don't differ from one another. A simpler way of thinking about this is that at least one mean is different from all others.

**Step 3: Set α**

If we look at what can happen in a hypothesis test, we can construct the following contingency table:

|  |  |  |
| --- | --- | --- |
|  | **In Reality** |  |
| Decision  | H0 is TRUE |  H0 is FALSE |
| H0 Accept  | OK | Type II Errorβ = probability of Type II Error |
| Reject H0 | Type I Errorα = probability of Type I Error | Ok |

You should be familiar with type I and type II errors from your introductory course. It is important to note that we want to set α before the experiment (a-priori) because the Type I error is the more ‘grevious’ error to make. The typical value of α is 0.05, establishing a 95% confidence level. **For this course we will assume α =0.05.**

**Step 4: Collect Data**

Remember the importance of recognizing whether data is collected through experimental design or observational

Step 5: Calculate a test statistic

For categorical treatment level means, we use an *F* statistic, named after R.A. Fisher. We will explore the mechanics of computing the *F* statistic beginning in Lesson 2. The *F* value we get from the data is labeled F calculated.

**Step 6: Construct Acceptance / Rejection regions**

As with all other test statistics, a threshold (critical) value of *F* is established. This *F* value can be obtained from statistical tables and is referred to as F critical or Fα. As a reminder, this critical value is the minimum value for the test statistic (in this case the *F* test) for us to be able to reject the null.

The F distribution, Fα, and the location of Acceptance / Rejection regions are should be shown in graph.

Step 7: Based on steps 5 and 6, draw a conclusion about H0

If the F calculated from the data is larger than the Fα, then you are in the Rejection region and you can reject the Null Hypothesis with (1−α) level of confidence.

B. Regression and correlation have same fundamental difference that is worth mentioning. In regression analysis there is an asymmetry in the way the dependent and explanatory variables are treated. The dependent variables is assumed to be statistical, random or stochastic that is to have a probability distribution. The explanatory variables, on the other hand, are assumed to have fixed values. In correlation analysis, on the other hand, we treat any (two) variables symmetrically, i.e there is no distinction between the dependent explanatory variables. Or Correlation stipulates the degree to which both of the variables can move together. However, **regression** specifies the effect of the change in the unit, in the known variable(p) on the evaluated variable (q). **Correlation** helps to constitute the connection **between** the two variables.

**correlation** is a statistical technique that **can show** whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. ... An intelligent **correlation** analysis **can** lead to a greater understanding of your **data**.

Regression can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable) and etc.