Assignment NO. 4 Week12 -Week14

Student Full Name:_____ .

Student ID:_____ .

CRN No:_____ .

Branch: _____.

Total Points

STATISTICS (STAT-101)	

True/False ____/6

MCQ ____/6

Short Answer ____/18

Total ____/30

Good Luck

STATISTICS (STAT-101)

Marks- 30

Answer all the Questions on the same question paper.

Section-I

State whether the following statements are True or False. (6 marks, 1 Mark each)

1. Correlations are always between -1 (perfect negative) and +1 (perfect positive) True
2. Sum of Squares Total (SST) increases as you add more X variables to a regression model. False
3. Positive residuals lie above the regression line. True
4. When conducting a hypothesis test with chi-square analysis, the rejection region in a chi-square distribution is always in the upper or right tail. True
5. ANOVA is the preferred method for finding differences among several population proportions. False
6. If SS (Treatment) =15 and SS (Total) =105 then the value of SS (error) in 1 way ANOVA is 90. True

Section-II

Multiple choice questions. ***(6 marks, 1 Mark Each)***

1. In a regression, the --- that the standard error of the regression is, the greater the accuracy of the prediction will be.
   a) smaller.
   b) larger
   c) we do not know unless we know whether the slope of the regression is positive or negative.
   d) None of the above

2. We measure heights and weights of 100 twenty-year old male college students. Which of the following will have the higher correlation:
   a) corr(height, weight) will be much greater than corr(weight, height)
   b) corr(weight, height) will be much greater than corr(height, weight)
   c) Both will have the same correlation.

d) Both will be about the same, but corr(weight, height) will be a little higher.

3. The regression line is drawn so that:
   a) The line goes through more points than any other possible line, straight or curved
   b) The line goes through more points than any other possible straight line.
   c) The sum of the absolute errors is as small as possible.
   d) The sum of the squared errors is as small as possible

4. What do residuals represent in the simple linear regression model?
   a) The difference between the actual Y values and the mean of Y.
   b) The difference between the actual Y values and the predicted Y values
   c) The square root of the slope.
   d) The predicted value of Y for the average X value

5. Analysis of variance (ANOVA) is a method for testing the hypothesis:

   a) That three or more population means are equal.
   b) Those at most three population means are equal.
   c) That three or more population means are NOT equal.
   d) That two population means are equal.

6. If the equation of regression is given by 5X -10Y +55 = 0, then the values of $b_{0\ (\text{intercept})}$ and $b_{1\ (\text{slope})}$ are respectively given by
   a) 5.5, 0.5
   b) 5, 5.5
   c) 0.5, 55
   d) 10, 55

Section-III

Answer the following Essay Type Question **(18 marks, 3 Mark Each)**

1. A study was conducted to determine the association between the maximum distance at which a highway sign can be read(in feet) and the age of the driver (in years). Fourty drivers of various ages were studied. The summary statistics for distance and age are shown below in table

	N	Minimum	Maximum	Mean	Std Deviation
Distance (y)	40	223.64	720.74	445.86	108.33
Age(x)	40	18	72	46.10	15.82
Valid N	40				

The correlation coefficient between distance and age in this sample is r= -0.5644. Calculate a and b of the least-squares regression equation that would predict the distance at which a highway sign can be read given the age of the driver.

Solution:

Given the values in the problem we calculate the slop b and the y intercept a as follows

$$b = r\frac{S_y}{S_x} = -0.5644\,.\left(\frac{108.33}{15.82}\right)$$

$$b = -3.865$$

$$a = \bar{y} - b(\bar{x}) = 445.86 - (-3.865) * (46.1)$$

$$a = 624.037$$

2. Five pair of shoes print length and height were used to conduct a formal hypothesis test of the claim that there is linear correlation between the two variables. Use the value of r =0.591and find the appropriate test statistics method at 0.05 significance level. Also based on the result compare and conclude whether it is reject hypothesis or fail to reject the null hypothesis. (P-value is 0.2937)

Solution:

Since the value of r is given so we assume that the requirement for the linear correlation are satisfied.

To claim that there is a linear correlation is to claim that the population linear correlation coefficient ρ is different from 0.

Step 1: $H_0 : \rho = 0$(There is no linear correlation)

Step 2: $H_1 : \rho \neq 0$(There is linear correlation)

Step 3: Significance level is 0.05

Step 4: Given r= 0.591 and n =5 we use the test statistic formula

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.591}{\sqrt{\frac{1-0.591^2}{5-2}}} = 1.269$$

Degree of freedom df = n-2 = 5-2 =3

Step 5: From the given observation P-value = 0.2937 is greater than the significance level 0.05, we fail to reject null hypothesis H_0 .

3. A random sample of 100 weights of Californians is obtained, and the last digits of those weights are summarized in Table 1. When people report weights, they tend to round, so a weight of 197lb might be rounded and reported as a more desirable 170lb. In contrast, if people are actually weighed, the weights tend to have last digits that are uniformly distributed with 0,1,…,9 all occurring with roughly the same frequencies.
Test the claim that the sample is from a population of weights in which the last digit do not occur with the same frequency. Based on the result, what can we conclude about the population used to obtain the weights.

Last digit	Frequency
0	46
1	1
2	2
3	3
4	3
5	30
6	4
7	0
8	8
9	3

Table 1

(From Table A-4 the Critical Value for χ^2 at degree of freedom =9 is 16.919 and p-value=0.000)

Solution:
Requirements are satisfied:
The sample data are assumed to be a random sample. (2) The sample data consist of frequency counts. (3) Each expected frequency is at least 5.
Step 1: The original claim is that the digits do not occur with the same frequency.
Step 2: If the original claim is false, then all of the probabilities are the same.
Step 3: The null hypothesis must contain the condition of equality, so we have
H_0: $p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$.
H_1: At least one of the probabilities is different from the others.
Step 4: The significance level is not specified, so we select $\alpha = 0.05$
Step 5: Because we are testing a claim that the distribution of last digits being a uniform distribution, we use the goodness-of-fit test. The χ^2 distribution is used for the test statistics.

Last digit	Observed Frequency (O)	Expected Frequency (E)=np	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	46	10	36	1296	129.6
1	1	10	-9	81	8.1
2	2	10	-8	64	6.4
3	3	10	-7	49	4.9
4	3	10	-7	49	4.9
5	30	10	20	400	40
6	4	10	-6	36	3.6
7	0	10	-10	100	10
8	8	10	-2	4	0.4
9	3	10	-7	49	4.9
$\chi 2 = \sum \frac{(O-E)^2}{E} = 212.8$					

Table 2

Step 6: Table 2 shows χ^2 test statistics for the digits 0 to 9. The test statistic χ^2 = 212.8 and the degree of freedom (k-1) = 10-1= 9. The critical value is χ^2 = 16.919 (value given in the question which is obtained from the A-4 table at degree of freedom= 9 and significance level =0.05). The p-value is 0.000 (for χ^2 = 212.8 and degree of freedom = 9)

Step 7: The p-value of 0.000 is less than the significance level α =0.05. So we reject the null hypothesis.

Step 8: There is sufficient evidence to support the claim that the last digits do not occur with the same relative frequency.

4. Find Linear correlation coefficient between X and Y and also obtain the Regression equation for the following data:

HOURS(X)	SCORE(Y)
3	2
4	5
6	4
7	5
8	7
8	8

Solution:

HOURS (X)	SCORE (Y)	X^2	Y^2	XY
3	2	9	4	6
4	5	16	25	20
6	4	36	16	24
7	5	49	25	35
8	7	64	49	56
8	8	64	64	64
$\Sigma X = 36$	**$\Sigma Y = 31$**	**$\Sigma X^2 = 238$**	**$\Sigma Y^2 = 183$**	**$\Sigma XY = 205$**

The Linear correlation Coefficient $r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$

$$r = \frac{6(205) - (36)(31)}{\sqrt{(6(238) - (36)^2)(6(183) - (31)^2)}} = \frac{1230 - 1116}{\sqrt{(1428 - 1296)(1098 - 961)}} = \frac{114}{\sqrt{132(137)}}$$
$$= \frac{114}{\sqrt{18084}} = \frac{114}{134.4}$$
$$r = +0.848$$

The Linear correlation Coefficient $(r) = 0.848$
The Regression equation is given by $\hat{y} = b_0 + b_1 x$
Where the value of the slope $b_1 = \frac{n\sum xy - \sum x \sum y}{n(\sum x^2) - (\sum x)^2}$ and y intercept
$b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$

Now $b_1 = \frac{n\sum xy - \sum x \sum y}{n(\sum x^2) - (\sum x)^2} = \frac{6(205) - (36)(31)}{6(238) - (36)^2} = \frac{1230 - 1116}{1428 - 1296} = \frac{114}{132} = 0.863$
$b_1 = 0.863$

$$b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} = \frac{(31)(238) - (36)(205)}{6(238) - (36)^2} = \frac{7378 - 7380}{1428 - 1296} = -\frac{2}{132}$$

$$b_0 = -0.015$$

Therefore, Regression equation is given by $\hat{y} = b_0 + b_1 x = -0.015 + 0.863x$

5. A one-way ANOVA for one of the experiment that measured the amount of time it took 51 different subject and there are three groups to complete the task. Check the value of F and complete the ANOVA table?

Source	df	SS	MS	F
Between groups	x	224891	x	3.486
Within groups	x	x	32253.54	
Total	50	1773061		

Solution:

Number of groups (g) =3, Number of total subjects (N) =51,

The missing data are:

Numerator: $df(between)$= g-1 =3-1 =2

Denominator: $df(\text{within}) = N - g$ =51-3=48

$$\text{MS}(between) = \frac{\text{SS(between)}}{df(between)} \Longrightarrow \frac{224891}{2} = 112445.5$$

$$\text{SS(within)} = \text{SS(Total)} - \text{SS}(between) \Longrightarrow 1773061 - 224891 = 1548170$$

$$\text{F} = \frac{\text{MS(between)}}{\text{MS}(within)} = \frac{112445.5}{32253.54} = 3.486$$

$$\text{F} = 3.486$$

Hence, F-statistics is correct

Source	df	SS	MS	F
Between groups	2	224891	112445.5	3.486
Within groups	48	1548170	32253.54	
Total	50	1773061		

6. The air travel industry conducted a survey to determine if the air travel anxiety depends upon the frequency of air travel. The number of passengers self -reporting air travel anxiety based on their frequent fliers status is as follows.

	Anxiety	No Anxiety	Total
Not frequent fliers	12	28	40
Frequent Fliers	3	37	40
total	15	65	80

Test at level $\alpha = 0.05$ and include the null and alternative hypothesis, the χ^2 -test statistics and a statement whether or not you reject the null hypothesis.

Solution:

The "grand total" is the sum of all frequencies in the table, which is 80.

At the observed frequency 12, Row total =40 and Column total =15.

The expected frequency $E = \frac{(Row\ total)\times(Column\ total)}{(Grand\ total)} = \frac{(40)\times(15)}{(80)} = 7.5$

At the observed frequency 37, Row total =40 and Column total =65.

The expected frequency $E = \frac{(Row\ total)\times(Column\ total)}{(Grand\ total)} = \frac{(40)\times(65)}{(80)} = 32.5$

There is a discrepancy between $O = 12$ and $E = 7.5$, and $O = 37$ and $E = 32.5$ such discrepancies are key components of the test statistic.

Requirements are satisfied: randomly assigned to frequency counts, expected frequencies are all at least 5.

Step 1:

H_0: air travel anxiety is independent of frequent flier status

H_1: air travel anxiety is dependent on frequent flier status

Step 2: Significance level is $\alpha = 0.05$.

Step 3: Contingency table: χ^2 distribution is used

$$\chi 2 = \sum \frac{(O - E)^2}{E} = \frac{(12 - 7.5)^2}{7.5} + \frac{(28 - 32.5)^2}{32.5} + \frac{(3 - 7.5)^2}{7.5} + \frac{(37 - 32.5)^2}{32.5}$$

$$\chi 2 = 6.646$$

The number of degrees of freedom given by $(r - 1).(c - 1) = (2 - 1).(2 - 1) = 1$.

Step 4: The critical value for df=1 found from χ^2 Table with $\alpha = 0.05$ is 3.841 less than

Calculated χ^2 $=6.646$ we reject the null hypothesis

Step 5: Since p- value is less than $\alpha = 0.05$, we reject the null hypothesis(air travel anxiety is independent of frequent flier status) . It appears that air travel anxiety is related to frequent flier status.