STATISTICS QUESTIONS

Step by Step Solutions

www.mathcracker.com

9/24/2016

http://www.mathcracker.com

Problem 1: A researcher is interested in the effects of family size on delinquency for a group of offenders and examines families with one to four children. She obtains a sample of 16 families, four of each size, and identifies the number of arrests per child for delinquency. The data is as follows:

	Group 1 4 children n=4	**Group 2** 3 children n=4	**Group 3** 2 children n=4	**Group 4** 1 child n=4
Family 1	10	8	5	4
Family 2	8	8	6	5
Family 3	9	6	7	2
Family 4	10	9	9	2

a) Calculate the total sum of squares.

b) Calculate the mean square (between groups).

c) Calculate the F-ratio

d) Use the Turkey HSD (alpha=0.05) to test for significance between groups. Which groups differed?

e) Based on your results, write a 1-2 paragraph essay that describes your observations obtained from this sample in regard to the effects of family size on delinquency for a group of offenders.

Solution: (a) The following table with descriptive statistics is obtained from the information provided

Obs.	Group 1	Group 2	Group 3	Group 4
	10	8	5	4
	8	8	6	5
	9	6	7	2
	10	9	9	2
Mean	**9.25**	**7.75**	**6.75**	**3.25**
St. Dev.	**0.957**	**1.258**	**1.708**	**1.5**

We need to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_A : \text{Not all the means are equal}$$

With the data found in the table above, we can compute the following values, which are needed to construct the ANOVA table. We have:

$$SS_{Between} = \sum_{i=1}^{k} n_i \left(\overline{x}_i - \overline{\overline{x}} \right)^2$$

from which we get

$$SS_{Between} = 4(9.25-6.75)^2 + 4(7.75-6.75)^2 + 4(6.75-6.75)^2 + 4(3.25-6.75)^2 = 78$$

Now we also see that,

$$SS_{Within} = \sum_{i=1}^{k} (n_i - 1) s_i^2$$

which implies

$$SS_{Within} = (4-1)\times 0.957^2 + (4-1)\times 1.258^2 + (4-1)\times 1.708^2 + (4-1)\times 1.5^2 = 23$$

Hence, SS_{Total} = 78+23 = 101

(b) Therefore

$$MS_{Between} = \frac{SS_{Between}}{k-1} = \frac{78}{3} = 26$$

Also, we obtain that

$$MS_{Within} = \frac{SS_{Within}}{N-k-1} = \frac{23}{12} = 1.917$$

(c) Therefore, the F-statistics is computed as

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{26}{1.917} = 13.5652$$

The critical value for $\alpha = 0.05$, $df_1 = 3$ and $df_2 = 12$ is given by

$$F_C = 3.4903$$

and the corresponding p-value is

$$p = \Pr\left(F_{3,12} > 13.5652\right) = 0.000$$

Observed that the p-value is less than the significance level $\alpha = 0.05$, then we reject H_0.

(d) The HSD difference is computed as follows:

$$HSD = Q * \sqrt{\frac{MSE}{n}} = 4.20\sqrt{\frac{1.917}{4}} = 2.91$$

The following table is obtained:

Post hoc analysis

Tukey simultaneous comparison t-values (d.f. = 12)

		Group 4	Group 3	Group 2	Group 1
		3.3	6.8	7.8	9.3
Group 4	3.3				
Group 3	6.8	3.58			
Group 2	7.8	4.60	1.02		
Group 1	9.3	6.13	2.55	1.53	

critical values for experimentwise error rate:

0.05	2.97
0.01	3.89

(e) Based on the above results, we have enough evidence to reject the null hypothesis of equal means, at the 0.05 significance level.

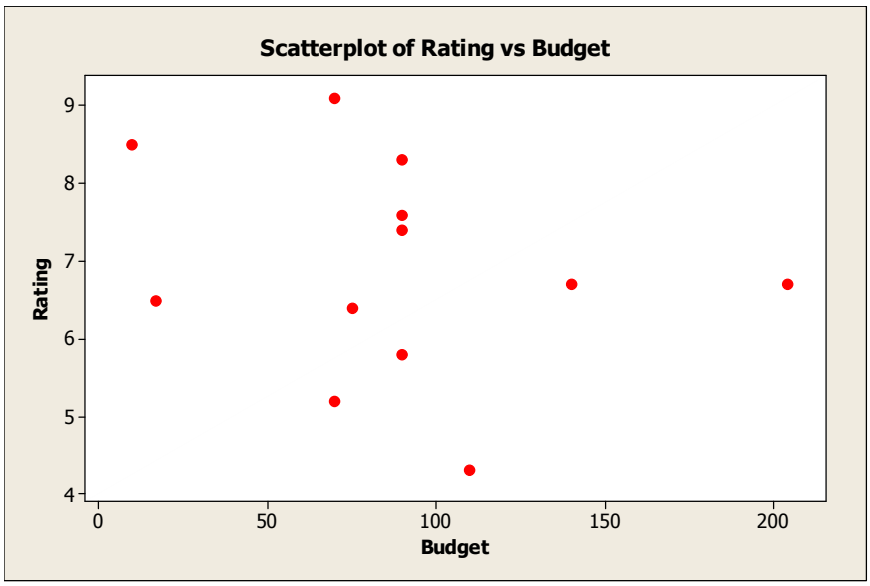
Summarizing, we have the following ANOVA table:

Source	**SS**	**df**	**MS**	**F**	**p-value**	**Crit. F**
Between Groups	78	3	26	13.5652	0.000	3.4903
Within Groups	23	12	1.917			
Total	101	15				

The pairwise differences that are significant are between Group 1 and Group 4, Group 2 and Group, and Group 3 and Group 4. In fact, the mean for Group 4 is significantly lower when compared to the means for groups 1, 2 and 3, respectively.

Problem 2: Movie Success. Using the data in Table 7.2, make a scatter diagram for the relationship between production budget and viewer rating of movies. Estimate the correlation coefficient. Based on these data, do you think a large production budget is likely to result in a movie with a high viewer rating? Explain.

Solution: The scatter plot is shown below.



It seems like there's a mild negative linear relationship between Budget and Rating. The actual correlation coefficient is computed as

Correlations: Budget, Rating

```
Pearson correlation of Budget and Rating = -0.238
P-Value = 0.456
```

As predicted by the visual trend, the correlation is negative, but since it's very small, the relationship is fairly weak. This means that is not certain that a larger budget will produce a higher rating, as it's not certain that a larger budget will produce a lower rating, but there a inclination to have lower rating with higher budgets.

Problem 3: Which of these models is a better representation of the relationship between students' age and starting salary? Explain your decision.

Solution: As mentioned in the previous part, the model obtained once the outlier was eliminated is relatively similar to the model with n=25 cases, as the regression coefficients don't change dramatically. But still this relatively small difference in coefficients makes a relatively large difference in R^2. In fact, for the model with n = 25 we get R2 = 0.334, and for the model with n = 25 we get R2 = 0.447. This makes the second model (with n = 24) the preferred one. The preferred model is

Starting Salary^ = -67,941.7485 + 3,635.6857* Age

Problem 4:

Compute an imprisonment rate per 1000 population for 2000. Introduce this incarceration rate as an independent variable into the model run in Part B.

- Test the hypothesis that the R squared =0.
- Does this model fit the data better than the model in Part B above? Explain.
- Does each of the independent variables have a statistically significant effect on homicide? Explain.
- How strong is the effect of each of the independent variables? Explain.
- Which of the independent variables has the stronger effect on the homicide rate? Explain.

Solution: The new variable is computed as

ImprPer1000 = Prison20/pop20

(let us recall that pop20 is already given in 1000's).

The following is obtained with Excel:

Regression Analysis							
	R^2	0.499					
	Adjusted R^2	0.466	n	49			
	R	0.707	k	3			

	Std. Error	1.639	Dep. Var.	**homrt20**			
ANOVA table							
Source	*SS*	*df*	*MS*	*F*	*p-value*		
Regression	120.4481	3	40.1494	14.95	6.87E-07		
Residual	120.8453	45	2.6855				
Total	241.2935	48					
Regression output					*confidence interval*		
variables	*coefficients*	*std. error*	*t (df=45)*	*p-value*	*95% lower*	*95% upper*	*std. coeff.*
Intercept	-1.9451	1.1721	-1.660	.1040	-4.3059	0.4156	0.000
ImprPer1000	0.2430	0.1707	1.423	.1616	-0.1009	0.5869	0.175
sglmom80	24.3703	5.6975	4.277	.0001	12.8949	35.8458	0.539
unempl20	36.9631	27.2997	1.354	.1825	-18.0215	91.9476	0.150

The model is

Homicide Rate in 2000 = -1.9451 + 0.2430* ImprPer1000 + 24.3703* sglmom80 + 36.9631* unempl20

Notice that the model is significant overall, since F(3, 45) = 14.95, p = 0.000000687 < 0.05, so then R^2 is significantly greater than zero.

This model fits only slightly better than the previous one, since now Adj. R^2 = 0.466, which means that in this case the amount of explained variation in the response variable by this model is 46.6%.

Notice that in this model, the variable *sglmom80* is individually significant, with t = 4.277 and p = 0.0001 < 0.05, but the variable *uempl20* is not individually significant, *t = 1.354*, *p = 0.1825 > 0.05*. The variable ImprPer1000 is not significant either, since *t = 1.423, p = 0.1616 > 0.05*.

The effect of *ImprPer1000* and *uempl20* is quite moderate since the standardized coefficients associated to them are less than 0.2 (this is, an increase in one standard deviation in either of the variables brings a change of less than 0.2 standard deviations in the response variable). The variable with the strongest effect is *sglmom80*, with a standardized coefficient of 0.539.

Problem 5: Using the data below, answer the following questions using a table format.

X_i	4	6	3	7
y_i	5	2	-1	4

a. $\sum_{i=1}^{4} x_i$ b. $\sum_{i=1}^{4} y_i$ c. $\sum_{i=1}^{4} x_i y_i$ d. Show that $\sum_{i=1}^{4} x_i \cdot \sum_{i=1}^{4} y_i \neq \sum_{i=1}^{4} x_i y_i$

e. $\sum_{i=1}^{4} x_i^2$ f. $\sum_{i=1}^{4} y_i^2$ g. $\left(\sum_{i=1}^{4} x_i y_i\right)^2$ h. Show that $\sum_{i=1}^{4} x_i^2 \neq \left(\sum_{i=1}^{4} x_i\right)^2$

i. Show that $\sum_{i=1}^{4} (x_i - \bar{x}) = \sum_{i=1}^{4} (y_i - \bar{y}) = 0$

Solution: We have:

	X	Y	X^2	Y^2	X*Y
	4	5	16	25	20
	6	2	36	4	12
	3	-1	9	1	-3
	7	4	49	16	28
Sum =	20	10	110	46	57

(a) $\sum_{i=1}^{4} x_i = 20$

(b) $\sum_{i=1}^{4} y_i = 10$

(c) $\sum_{i=1}^{4} x_i y_i = 57$

(d) Notice that $\sum_{i=1}^{4} x_i y_i = 57$, and $\left(\sum_{i=1}^{4} x_i\right)\left(\sum_{i=1}^{4} y_i\right) = 20 \times 10 = 200$, which means that $\sum_{i=1}^{4} x_i y_i \neq \left(\sum_{i=1}^{4} x_i\right)\left(\sum_{i=1}^{4} y_i\right)$ in this case.

(e) $\sum_{i=1}^{4} x_i^2 = 110$

(f) $\sum_{i=1}^{4} y_i^2 = 46$

(g) $\left(\sum_{i=1}^{4} x_i y_i\right)^2 = 57^2 = 3249$

(h) $\left(\sum_{i=1}^{4} x_i\right)^2 = 20^2 = 400$, and $\sum_{i=1}^{4} x_i^2 = 110$, so then $\sum_{i=1}^{4} x_i^2 \neq \sum_{i=1}^{4} x_i$

(i) we get that $\overline{X} = 5, \overline{Y} = 2.5$. Observe that

X	**Y**	**X-Xbar**	**Y-Ybar**
4	5	-1	2.5
6	2	1	-0.5
3	-1	-2	-3.5
7	4	2	1.5
	Sum =	0	0

so then

$$\sum_{i=1}^{4}(x_i - \overline{x}) = \sum_{i=1}^{4}(y_i - \overline{y}) = 0$$