Assignment 3 (Sol.)

Introduction to Data Analytics

Prof. Nandan Sudarsanam & Prof. B. Ravindran

1. A company wants to determine whether a particular training module improves the efficiency of its employees on a particular task. A random sample of 25 employees are tested and rated on the relevant task. Once the training is imparted to them, they are again tested and rated on the same task. To determine whether the average performance improved or not, which of the following would you consider the most suitable test?

   (a) one sample z-test

   (b) one sample t-test

   (c) two sample t-test

   (d) paired t-test

   Sol. (d)
   Since each employee is being tested twice, i.e., there is a logical pairing between the data points in the two samples, we would prefer to use the paired t-test.

2. In conducting the two sample t-test, we find that the variances of the two samples are 13 and 15, with the first sample having 21 data points and the second sample consisting of 25 data points. What is the value of the degrees of freedom to be used in performing the t-test? Suppose that the variances of the two samples were equal. What is the value of the degrees of freedom in this case?

   (a) 43.946, 44

   (b) 43.502, 44

   (c) 43.502, 45

   (d) 0.082, 44

   Sol. (b)
   In case the variances are unequal, we use the following formula to calculate the degrees of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Substituting the values of $s_1^2 = 13$, $s_2^2 = 15$, $n_1 = 21$, and $n_2 = 25$, we have

$$df = \frac{\left(\frac{13}{21} + \frac{15}{25}\right)^2}{\frac{\left(\frac{13}{21}\right)^2}{21-1} + \frac{\left(\frac{15}{25}\right)^2}{25-1}} = 43.502$$

In case the two sample variances were equal, the degrees of freedom would be

$$df = n_1 + n_2 - 2 = 21 + 25 - 2 = 44.$$

3. Suppose that in a hypothesis testing problem, we negate the null hypothesis. What relation do the new type I and type II errors have to the previous type I and type II errors?

   (a) in both cases, the different types of errors stay the same

   (b) the errors switch roles, the new type I errors are the same as the old type II errors and vice versa

   (c) it is problem dependent and cannot be predicted in general

   Sol. (b)
   Suppose the original null hypothesis was $H_0 : \mu <= c$, and the significance level for the test is α. In this case, a type I error corresponds to the case where the null hypothesis is actually true, but the calculated p-value, i.e., $P(z > z_{stat})$ considering the z-test, is less than the value of α and hence, the null hypothesis is rejected. A type II error occurs if the null hypothesis is actually false, but we fail to reject the null hypothesis, which happens if $P(z > z_{stat}) > \alpha$.

   Now consider the case where the null hypothesis has been negated (i.e., $\mu > c$). For a type I error to occur in this scenario, the null hypothesis must be true and we have to reject the null hypothesis, i.e., $P(z <= z_{stat}) < \alpha$. But this corresponds to the type II error case in the first scenario, since $P(z > z_{stat}) > \alpha => P(z <= z_{stat}) < \alpha$. Similarly, it can be seen that a type II error in the second scenario matches the case where a type I error occurs with the original null hypothesis.

4. Assume that the marks obtained by students in a test follows a normal distribution. The teacher randomly selects 20 papers for correction, and from this sample, finds an average score of 63 with a standard deviation of 8. Set up a 95% confidence interval estimate for the average score of all students in the test. (Hint: use the following z-table:

   http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf or t-table:

   http://www.stat.ufl.edu/~athienit/Tables/Ttable.pdf as required).

   (a) (59.907, 66.093)

   (b) (59.256, 66.744)

   (c) (62.07, 63.93)

   (d) (62.091, 63.909)

   Sol. (b)
   From the question, we have $\bar{x} = 63$, sample standard deviation, $s = 8$, sample size, $n = 20$,

and confidence level = 95%. Since the standard deviation is computed from the sample, the estimate will be a t-interval and not a z-interval. The corresponding formula is

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}$$

Here, $\alpha/2 = 0.025$, hence from the t-table, we have

$$t_{\alpha/2,n-1} = t_{0.025,19} = 2.093$$

Thus, the required confidence interval is

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}} = 63 \pm 2.093\frac{8}{\sqrt{20}} = 63 \pm 3.744 = (59.256, 66.744)$$

5. Is it possible for the F statistic calculated in ANOVA to be negative?

   (a) no

   (b) yes

   Sol. (a)
   Both the numerator and denominator are variances which are by definition positive. Hence, we cannot have a negative value of the F statistic.

6. In a particular class, the students are split into three groups with a mentor being assigned to each group. A test was conducted for the entire class. The following table shows the scores of a sample of the students from the three groups.

Group 1	Group 2	Group 3
1200	1000	890
1000	1100	650
980	730	1100
880	800	900
750	500	400
800	700	380

   According to the ANOVA method, what are the respective values of MSB and MSE?

   (a) 70350, 44183

   (b) 70350, 53020

   (c) 46900, 44183

   (d) 46900, 53020

   Sol. (b)
   We have $n = 6, a = 3$, and $N = 18$ From the table, we can calculate the following averages:
   $\bar{y}_{G1.} = 935$, $\bar{y}_{G2.} = 805$, $\bar{y}_{G3.} = 720$, $\bar{\bar{y}}_{..} = 820$
   Now, the SSB is given by

$$n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{\bar{y}}_{..})^2 = 140700$$

Hence, the MSB = SSB/DoF = SSB/(a-1) = SSB/2 = 70350.

Similarly, we calculate SSE which is given by

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{i,j}-\bar{y}_{i.})^2 = 131950 + 234750 + 428600 = 795300$$

Hence, the MSE = SSE/DoF = SSE/(N-a) = SSE/15 = 53020

7. In a clinical trial of two groups of participants with controlled diets, the following observations were made.

	Symptom1	Symptom 2	Symptom 3
Diet 1	200	150	50
Diet 2	250	300	50

Do the two diets significantly affect the symptoms observed in the two groups of participants? Use a 0.05 level of significance. (Hint: use the following chi-square table:

http://sites.stat.psu.edu/˜mga/401/tables/Chi-square-table.pdf)

(a) no

(b) yes

Sol. (b)

We start with the null hypothesis that the diets and symptoms are independent.

We have $r = 2$ and $c = 3$ and total frequency, $n = 1000$. Calculating individual expectations from the table, we have

$E_{1,1} = 400 * 450/1000 = 180$

$E_{1,2} = 400 * 450/1000 = 180$

$E_{1,3} = 400 * 100/1000 = 40$

$E_{2,1} = 600 * 450/1000 = 270$

$E_{2,2} = 600 * 450/1000 = 270$

$E_{2,3} = 600 * 100/1000 = 60$

The test statistic is calculated as

$$\chi^2 = \sum_{i=1}^{r}\sum_{j-1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$$

$$\chi^2 = \frac{(200-180)^2}{180} + \frac{(150-180)^2}{180} + \frac{(50-40)^2}{40} + \frac{(250-270)^2}{270} + \frac{(300-270)^2}{270} + \frac{(50-60)^2}{60}$$

$$\chi^2 = 16.20$$

For a degree of freedom of $(r-1)*(c-1) = 2$ and a significance level of 0.05 we find that the value of the test statistic for which the corresponding p-value is 0.05 is 5.991. Since the value of $\chi^2 = 16.20$ exceeds the value of 5.991, the resultant probability will be less than 0.05. Hence the null hypothesis can be rejected, i.e, the two diets do significantly affect the symptoms observed in the two groups of participants. (Note that using a Chi-square calculator, the p-value observed for the calculated test statistic is 0.0003).

8. From the solution to the previous question, is it possible to conclude that certain diets cause certain symptoms?

   (a) no

   (b) yes

   Sol. (a)
   Even though in the last question we have seen that diets and symptoms are not independent, the test that was performed only comments on the independence of the variables. It does not allow us to infer causality.

9. Suppose that we have two variables, X, the independent variable and Y, the dependent variable. We wish to find the relation between them. An expert tells us that relation between the two has the form $Y = mX^2 + c$. Available to us are samples of the variables X and Y. Is it possible to apply linear regression to this data to estimate the values of m and c?

   (a) no

   (b) yes

   Sol. (b)
   Instead of considering the dependent variable directly, we can transform the independent variable by considering the square of each value. Thus, on the X-axis, we can plot values of X^2 and on the Y-axis, we can plot values of Y. The relation between the dependent and the transformed independent variable is linear and the value of slope and intercept can be estimated using linear regression.

10. Recall the graph between explanatory variable (X) and response variable (Y) in the introduction to regression lesson. We mentioned that, given some data, one way to fix the parameters of the line modelling the relation between the two variables is to find that line which minimises the distance each point has to the line.

    Suppose that using advanced regression techniques, for the same data, we come up with a non-linear model (i.e., a curve instead of a straight line) which fits each training data point (i.e., the data available to us to build the model - essentially the data that is visible in the graph) perfectly - the curve passes through each data point and hence the cumulative distance between the data points and the curve is zero. For making general predictions about the value of Y (i.e., we may want to make predictions for points not in the training data), do you think this non-linear model is preferable to the linear model?

    (a) no

    (b) yes

    Sol. (a)
    From the graph, we can see that the relation between the two variables is essentially a straight line. However, the reason that the straight line model does not result in zero error is essentially due to noise in the training data. Using a highly non-linear model to achieve zero error on the training data will result in overfitting, i.e., the curve will model the noise in the data set. The result of such overfitting is generally poor performance on unseen data points, i.e., poor generalisation.