Assignment 3

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

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

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

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)

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

   (a) no
   (b) yes

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

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

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

   (a) no

(b) yes

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

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