Assignment 3 (Sol.)

Introduction to Data Analytics

Prof. Nandan Sudarsanam & Prof. B. Ravindran

1. Based on the data generated from two different approaches to producing tooth paste, you need to decide which one to select. To do this task, you will choose techniques from

   (a) descriptive statistics

   (b) inferential statistics

   (c) predictive analytics

   Sol. (b)

2. In which among the following two sample tests, can the number of data points in the two samples differ?
   (Note: More than one options can be correct)

   (a) two sample z-test

   (b) two sample t-test

   (c) paired t-test

   (d) F-test

   Sol. (a), (b), (d)

3. A study was conducted to test the effect of a special training program over employees. Each employee was given a test twice, both before and after completing the training program. Let ΔX denotes the difference between the first and second test scores of each employee. It means, if mean of ΔX is *zero*, the training program has no effect on the average. The data of 20 employees has been recorded and provided here in the table.

Employee - ID	**Pre-training score**	**Post-training score**	ΔX
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

Your task is to test the hypothesis of no effect against the hypothesis of positive effect. What will be the null and alternative hypothesis?

(a) $H_0 : \mu = 0; H_1 : \mu \neq 0$

(b) $H_0 : \mu = 0; H_1 : \mu < 0$

(c) $H_0 : \mu = \geq; H_1 : \mu < 0$

(d) $H_0 : \mu = 0; H_1 : \mu > 0$

Sol. (d)

4. In the previous question, the value of the test statistic is:

(a) 2.05

(b) 2.837

(c) 3.231

(d) 0.634

Sol. (c)
Mean and standard deviation of the differences are given by $d = \bar{2}.05$ and $s_d = 2.837$.
Standard error: $\frac{s_d}{\sqrt{n}} = \frac{2.837}{\sqrt{20}} = 0.634$
$t = \frac{2.05}{0.634} = 3.231$

5. Using the test statistic calculated in previous question and appropriate degree of freedom, make a decision with $\alpha = 0.05$. The decision is:
(Hint: use the following
z-table: www.stat.ufl.edu/ athienit/Tables/Ztable.pdf or
t-table: www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf or
f-table: www.stat.purdue.edu/ jtroisi/STAT350Spring2015/tables/FTable.pdf as required).

   (a) reject H_0 and conclude that effect is positive

   (b) failed to reject H_0 and conclude that there was no effect

   (c) cannot make any inference as the data is incomplete

   Sol. (a)
   At $t = 3.231$ and $df = 19$, the t-table gives $p = 0.004$. Calculated p value is less than the given α value. So we can reject the null hypothesis and we can conclude that, on average the training program has positive effect on the employees.

6. If the decision you made in the previous question is incorrect, what type of error has been made?

   (a) type I error

   (b) type II error

   (c) type IV error

   (d) both (a) and (b)

   Sol. (a)
   In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (a "false positive").

7. Suppose that a sample (shown in table) of $n = 5$ was selected from the price of properties sold by The New Okhla Industrial Development Authority in the National Capital Region of India.

Year	**Price of 5 samples (Rs./acre)**				
2000	30000	34000	36000	38000	40000
2001	30000	35000	37000	38000	40000
2002	40000	41000	43000	44000	50000

   Your task is to conduct ANOVA over this data to check whether you get evidence that prices over the land were not same for the three years considered. The F-statistic for the given data is:

   (a) 8.96

   (b) 4.312

   (c) 6.834

   (d) 1.337

   (e) none of the above

Sol. (c)

8. By using the data given in previous question, can we reject the null hypothesis (null hypothesis is : $\mu_{2000} = \mu_{2001} = \mu_{2002}$) at 0.01 level? What about 0.05 level?
(Hint: use the following
z-table: www.stat.ufl.edu/ athienit/Tables/Ztable.pdf or
t-table: www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf or
f-table: www.stat.purdue.edu/ jtroisi/STAT350Spring2015/tables/FTable.pdf as required).

   (a) reject the null hypothesis at 0.01 level; reject the null hypothesis at 0.05 level

   (b) reject the null hypothesis at 0.01 level; failed reject the null hypothesis at 0.05 level

   (c) failed to reject the null hypothesis at 0.01 level; reject the null hypothesis at 0.05 level

   (d) failed to reject the null hypothesis at 0.01 level; failed to reject the null hypothesis at 0.05 level

   Sol. (c)
   In the given data, between treatments (numerator) have degree of freedom $= a - 1 = 2$
   Error within treatments have degree of freedom $= N - a = 15 - 3 = 12$
   Following the F-table, we get $P = 0.01044$.
   Since P-value>0.01, so there is not enough evidence to reject the null hypothesis. But at the level of 0.05, P-value<0.05. SO we reject the null hypothesis.

9. What is the purpose of a multiple regression?

   (a) To predict scores on a dependent variable from scores on multiple independent variables.

   (b) To predict scores on an independent variable from scores on multiple dependent variables

   (c) To assess whether there is a significant difference between repeated measures

   (d) To predict scores on a dependent variable from scores on a single independent variable

   (e) To predict scores on an independent variable from scores on a single dependent variable

   (f) To assess whether there is a significant difference between independent groups

   Sol. (a)

10. For a chi-square test, a 4 × 5 contingency table will have how many degrees of freedom?

   (a) 12

   (b) 8

   (c) 9

   (d) 6

   Sol. (a)
   Degree of freedom = (no. of rows - 1) × (no. of columns - 1)
   (4-1) × (5-1) = 12