INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR

Mid-Autumn Semester 2018-19

Date of Examination: **25.09.2018** Session: **AN**

Subject No. : **CS40003** Subject: **Data Analytics**

Department: Computer Science & Engineering

Full Marks: 70 Time: 02 hours

Answer to ALL questions

1. Answer briefly to the following questions.

[10×2]

i. What is the smallest and largest units for measuring the size of data?

Ans. Smallest unit : Bit

Largest unit : Quintillon

For example, 1 Quintillon = 10^{18} bits (in UK) or 10^{30} bits (in USA)

NOTE

(Alternatively, if students give answer to this question as Yottabyte (1 YB = 10^{24} bytes).

ii. Write down the expressions for z-value and t-value in terms of the population parameters and sample statistics.

Ans.

z= $(\dot{\bar{x}} - \mu)/(\sigma\ /\ \sqrt{n})$

t = $(\bar{x} - \mu)/(s\ /\ \sqrt{n})$

where, μ = population mean
σ = population standard deviation
s= Sample standard deviation
n= size of sample
x= sample mean

iii. What are the degrees of freedom in the following two cases?

Case 1:

Roll No	. . .	. . .	. . .
Marks	. . .	. . .	. . .

With n data

Case 2:

DOB	. . .	. . .	. . .
Marks	. . .	. . .	. . .

With n data

Ans.

Case 1: degree of freedom = n-1
(Here, Roll No is just a sample number not a data).

Case 2: degree of freedom = n-2
(Degree of freedom of 2-d table is n-2).

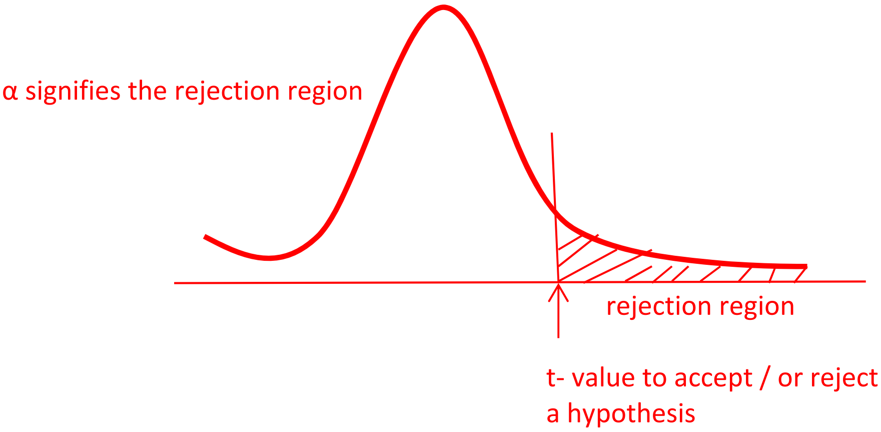
iv. In the following cases of hypothesis testing, which sampling distributions need to be considered?

Case 1: To infer the mean of a normally distributed population when the population variance is *unknown.*

Case 2: To infer the mean of a normally distributed population when the population variance is *known.*

Ans.
Case 1 : t-distribution table
Case 2: z-distribution table

v. For a given degrees of freedom, if α, the confidence level increases, then t-value to reject a hypothesis increases. Justify the statement.
Ans.
The statement is not correct. This is because if α- value increases, the rejection region increases and hence t-value decreases.


α signifies the rejection region
rejection region
t- value to accept / or reject
a hypothesis

vi. State the following statements as correct and incorrect?
a) Karl Pearson's correlation analysis gives accurate result always.
b) Charles Spearman correlation analysis can be applied to both numeric and ordinal data.

Ans.
(a) The statement is not true always. The Pearson's correlation analysis find accurate value of correlation coefficient between two attributes if they are linearly related.
(b) The statement is correct. For both numerical and ordinal data ranks are to be assigned and the rank values then can be used to calculate the rank correlation coefficient.

vii. Clearly state the Central Limit Theorem. What are the underlying assumptions of the theorem? What is its implication?
Ans.
Central Limit Theorem: If random samples each of size n are taken from any distribution with mean μ and variance σ^2, the sample mean $\bar{x}$ will have a distribution approximately normal with mean μ and variance $\frac{\sigma^2}{n}$.

Assumptions:
i. Population with unknown distribution.
ii. Sample distribution is normal.

Implication:
i. Population mean (μ) can be inferred from sample mean ($\bar{x}$).

ii. Sample variance is decreases as n increases and the theorem is reasonably valid for n ≡ 30.

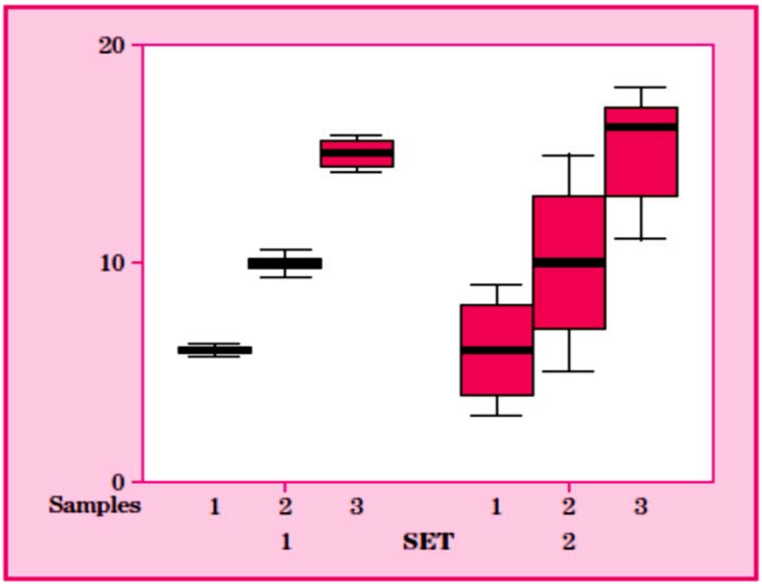
viii. What is the null hypothesis in χ^2 –test? How one can obtain the degrees of freedom given a sample data?

Ans.

Null hypothesis in χ^2 test: Suppose, we are to find the correlation between two attributes A and B. Then the null hypothesis is that A and B are independent.

Degree of freedom in χ^2 test: If the contingency table contains m rows and n columns for a given data set, then the degree of freedom is (m-1) * (n-1)

ix. The box plots obtained from two independent experiments are shown below.



Mark the following statement as true and false.

a) The variances among the means for the two sets are identical.
b) There is a stronger evidence of differences among mean in Set 2 than among mean in Set 1.

Ans.

(a) The statement is TRUE.
(b) The statement is FALSE.

x. What are the five numbers summarized in a box plot? Draw a box plot and locate them precisely.

Ans.

Given a sample, the five numbers are i) minimum value ii) Q_1 (the first quartile) ii) Median iv) Q_3(the third quartile) and maximum value in the sample data.

2. An aptitude test has been conducted to test the aptitude of graduate engineers in the country. The test is conducted so that scores are normally distributed with a standard deviation of 10. A statistical test administered to a random sample of size 500 examinees. The sample yields a mean of 51.07.

a) Test the hypothesis that the population mean is 50. Consider the level of confidence is 5%. You should clearly show all the five steps in your calculation.
b) What should be the least value of confidence, so that the hypothesis will be rejected?

[(1+2+2+1+1)+3]

(a) Ans.

Step 1: Hypothesis for testing and confidence level

$H_0 : \mu = 50$

$H_1 : \mu \neq 50$

Given that $\alpha = 0.05$

Step 2: Determining the test statistics and rejection for H_0

Given that population is normally distributed and standard deviation (hence variance) is known. So, test statistics should follow z-distribution. Further, the degree of freedom is 500-1=499. The z-value for rejecting the hypothesis H_0 at $\alpha = 0.005$ is **1.96** with the two tailed test.

Step 3: Test statistics from the given sample

$Z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

We have $\bar{x} = 51.07$

$\mu = 50.0$

$\sigma = 10$

$n = 500$

Hence, test-statistics, $z = \frac{51.07-50}{10/\sqrt{500}} = \frac{1.07}{0.44721} = \mathbf{2.392}$

Step 4: Decision to make either fail to reject or reject H_0

Since the observed value z = 2.392 is greater than the z-value =1.96 at $\alpha = 0.05$, the hypothesis H_0 is **rejected**.

Step 5: Final comment and interpretation of the result

It is inferred that the mean score of the examinees of the aptitude test is 50.

(b) Given the sample information,

The p-value is $P(z > 2.39) = 0.9916$

Thus, the H_0 hypothesis will not be rejected if $\alpha > 0.01$

3. A survey was conducted among 1500 people (300 literate and 1200 illiterate). In this survey, a person is either a literate or an illiterate and their participation in a poll (is either “cast” or “no cast”) was recorded. It is observed that 250 literate people and 200 illiterate people casted their votes, respectively.

It is required to test (using χ^2 –correlation analysis) if there is any association between literacy and habit to casting vote of the electorate.

Find the following.

a) Give the structure of the contingency table suitable for χ^2 –test in this case.

b) Enter the observed frequencies and expected frequencies in the table.

c) Calculate the χ^2 value from the contingency table.

d) Mention the null hypothesis usually considered for χ^2 –test.

e) Test the hypothesis, if literacy is correlated with voting trend. Assume 0.01 is the confidence level.

[2+2+2+1+3]

Ans.

(a) Contingency table will take the following form

Literacy				
		Literate	Illiterate	Total
Casting habit	Yes			
	No			
	Total			1500

(b) Observed frequencies and expected frequencies are shown in the table.

Literacy				
		Literate	Illiterate	Total
Casting habit	Yes	250 (90)	200 (360)	450
	No	50 (210)	1000 (840)	1050
	Total	300	1200	1500

(c) Calculation of χ^2 – value

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(200-360)^2}{360} + \frac{(50-210)^2}{210} + \frac{(1000-840)^2}{840}$$
$$= 507.93$$

(d) The hypothesis for testing is
H_0: Literacy and voting habit is independent to each other.
H_1: Otherwise.

(e) Testing the hypothesis
The degree of freedom for the data is (2-1) * (2-1) = 1
With 1 degree of freedom and $\alpha = 0.01$, the χ^2 value needed to reject the hypothesis is 10.828.
Since, the test statistic value of $\chi^2 = 507.93$ greater than critical χ^2–value, the null hypothesis is rejected.
That is, Literacy and casting habit is highly correlated.

4. Marks obtained by 10 students in two subjects Physics and Chemistry in an examination is given in Table 1.

Find the following.

a) Coefficient of variances in cases of the results in both the subjects.
b) Calculate the covariance between the scores in the two subjects.
c) Critically comment on the results observed in the calculations 4(a) and 4(b).

Table 1 (Q. No. 4)

Student	Physics	Chemistry
1	70	60
2	80	90
3	65	75
4	75	75

5	80	70
6	90	85
7	85	90
8	40	80
9	70	30
10	95	90

[4+4+2]

Ans.

(a) Coefficient of variance (Physics) = $\frac{15.45}{75}$18.47 * 100 = 20.6
Coefficient of variance (Chemistry) = $\frac{18.47}{74.5}$ * 100 = 24.8

(b) Covariance (σ_{xy}) = $\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu y)$ = 86.11

(c) Coefficient of variations signifie that STD is 20.65% (in Physics) and 24.6% (in Chemistry) of the mean values. On the other hand, +ve value of covariance implies that there is postice correlation.

5. The table below shows the lifetimes under controlled conditions, in hours in excess of 1000 hours of samples of 60W electric light bulbs of three different brands.

Table 2 (Q. No. 5)

Brand		
1	2	3
16	18	26
15	22	31
13	20	24
21	16	30
15	24	22

a) Identify the factor(s) and level(s) in the above-mentioned problem statement.

b) Calculate the sample mean and variance for each level.

c) Calculate the pooled estimate of variance.

d) Calculate the variability between samples.

e) Assuming all lifetime to be normally distributed with common variance, test, at the 1% significance level, the hypothesis that there is no difference between the brands with respect to mean lifetime.

[1+2+3+2+2]

Ans.

(a) Hence, there is one factor (i.e. brand) at three levels (i.e. 1, 2 and 3). There are three samples each of size 5.

(b)

	Brand		
	1	2	3
Sample size	5	5	5
Sum	80	100	135
Sum of squares	1316	2040	3689
Mean	16	20	27
Variance	9	10	11

(c) $\widehat{\sigma_w}^2 = \frac{(5-1)*9 +((5-1)*10+(5-1)*11}{(5-1)+(5-1)+(5-1)}$ = 10

(d)

	Brand		
	1	2	3
Sample mean	16	20	27
Sum		63	
Sum of squares		1385	
Mean		21	
Variance		31	

(e) H_0 : $\mu_1 = \mu_2 = \mu_3 = 20$
H_1 : $\mu 1 \neq \mu_2 \neq \mu_3 \neq 20$

Significance level, α =0.01
Degree of freedom v_1=2, v_2= 12
(k-1) (n-k)
Critical region is F > 6.926
Test statistics is F > $\frac{5.\widehat{\sigma_B}^2}{\widehat{\sigma_w}^2}$ = 155/10 = 15.5
Since, the value lies in the critical region, thus, there is evidence that, at the 1% significance level, the true mean lifetime of three bulbs from three brands do differ.

6. Write down the formula for calculating the following.

a) Probability distribution function according to Poisson distribution.
b) Probability distribution function according to Binomial distribution.
c) i) Arithmetic mean, ii) Geometric mean and iii) Harmonic mean of a sample of size n.

Your answer should include the precise statements on each symbol in each expression.

[2+2+(3×2)]

Ans.

(a) Poisson distribution : P(x; λt) = $\frac{e^{-\lambda t}(\lambda t)^x}{x!}$, x= 0, 1, 2,...
x = number of outcomes in a given time interval.
λ = average number of outcomes per unit time.

(b) Binomial distribution : b(x; n, p) = $\left(nc_p\right) p^x q^{n-x}$ x= 0,1,2,...,n
x= number of success
n = independent trials
p = probability of success
q = probability of failure
q = 1- p

(c) Arithmetic mean : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Geometric mean : $\bar{x} = (\prod_{i=1}^{n} x_i)^{1/n}$

Harmonic mean: $\bar{x} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$

--- * ---