Fr. Conceicao Rodrigues College of Engineering

Department of Artificial Intelligence and Data Science

Subject: Statistics for AI&DS	Total Marks: 20
Semester: V	Date: September 7, 2022

Course Outcomes: Learners will be able to

1. Illustrate exploratory data analysis
2. Describe data and sampling distribution

Q.		Question	Marks	CO	BL	PI
1	A	Describe Numeric and Categorical data type with examples	03	CO1	2	1.6.1
	B	Explain Central Limit theorem	02	CO2	2	1.2.2
	C	Draw the box plot for given data : 100,120,110,150,110,140,130,170,120,220,140,110 Give comments on outliers by analysing the diagram	03	CO1	4	4.6.3
2	A	A factory produces components of which 1% are defective. The components are packed in boxes. (10 components in one box). A box is selected at random. Find the probability that there are at least 2 defective components in the box.	03	CO2	3	1.2.1
	B	Botanist is studying the distribution of daisies in the field. The field is divided into number of equal sized squares. The mean number of daisies per square is assumed to be 3. The daisies are distributed randomly throughout the field. Find the probability that in randomly chosen square there will be more than 2 daisies	03	CO2	3	1.2.1
3	A	Amit earned a score of 940 on a national achievement test. The mean test score was 850 with a sample standard deviation of 100. What proportion of students had a higher score than Amit? (Assume that test scores are normally distributed.)	03	CO2	3	1.2.1
	B	An automotive engineer wants to estimate the cost of repairing a car that experiences a 25 MPH head-on collision. He crashes 24 cars, and the average repair is 11,000. The standard deviation of the 24-car sample is 2,500. Provide a 98% confidence interval for the true mean cost of repair.	03	CO2	3	1.2.1

BL – Bloom's Taxonomy Levels (1- Remembering, 2- Understanding, 3 – Applying, 4 – Analysing, 5 – Evaluating, 6 - Creating)

CO – Course Outcomes PO – Program Outcomes; PI Code – Performance Indicator Code

Vidyavardhini's College of Engineering & Technology, Vasai

Department of Computer Science and Engineering (Data Science)

Academic Year 2022-23

Internal Assessment - 1

Sub: CSDLO5011/Statistics for Artificial Intelligence Data Science Year/Sem:- TE/V
Date: 08/08/2022
Max. Marks: 20 **Duration:- 1Hr**

Q. No.		Questions	Marks
1.	A.	Illustrate the variance and standard deviation of the possibilities when the die is rolled.	2
	B.	Consider a test score for 8 students in a class. Consider the 25th percentile for the 8 numbers. The numbers are given ranks from 1 for the lowest number to 8 for the highest number. Calculate the percentile value.	4

Test Score

Rank	Number
1	3
2	5
3	7
4	8
5	9
6	11
7	13
8	15

OR

A group of customer service surveys were sent out at random. The scores were 90, 50, 70, 80, 70, 60, 20, 30, 80, 90, and 20. Calculate the central tendency.

Q. No.		Questions	Marks
2.	A.	Explain Central limit theorem.	2
	B.	A random sample of 400 members is found to have a mean of 4.45 cms. Summarize it reasonably so it could be regarded as a sample from a large population whose mean is 5 cms and variance is 4 cms ?	5

OR

In 800 families with 4 children each. Classify according to given criteria, how many families would you expect to have

a) 2 boys and 2 girls

b) Atleast 1 boy

c) no girl

<table>
<tr><td>Q3)</td><td>Find 25th &amp;,50th &amp;75th percentile of following Data.

<table>
<tr><th>N</th><th>value</th></tr>
<tr><td>1</td><td>2</td></tr>
<tr><td>2</td><td>2</td></tr>
<tr><td>3</td><td>2</td></tr>
<tr><td>4</td><td>2</td></tr>
<tr><td>5</td><td>3</td></tr>
<tr><td>6</td><td>4</td></tr>
<tr><td>7</td><td>5</td></tr>
<tr><td>8</td><td>6</td></tr>
<tr><td>9</td><td>6</td></tr>
<tr><td>10</td><td>7</td></tr>
</table>

25th Percentile = 2

50th Percentile = 3.5

75th Percentile = 6</td><td>(3)</td><td>CO-1</td></tr>
<tr><td>Q4)</td><td>X is a normally distributed variable with mean $\mu = 28$ and standard deviation $\sigma = 4$. Find a) $P(x < 40)$, b) $P(30 < x < 35)$

a) 0.99865

b) 0.26848</td><td>(2)</td><td>CO-2</td></tr>
<tr><td>Q5)</td><td>The record of weights of the male population follows the normal distribution. Its mean and standard deviations are 70 kg and 15 kg respectively. If a researcher considers the records of 50 males, then what would be the mean and standard deviation of the chosen sample?

Mean = 70

SD = 2.12132</td><td>(2)</td><td>CO-2</td></tr>
<tr><td>Q6)</td><td>Average number of accidents at a particular junction is 24. Calculate the probability of that there are exactly 3 accidents in a particular month. (Use Poisson distribution)

Probability for 3 accidents = 0.180447</td><td>(3)</td><td>CO-2</td></tr>
<tr><td>Q7)</td><td>Discuss procedure &amp; key feature of Boot strapping</td><td>(3)</td><td>CO-2</td></tr>
</table>

JNIESTRT'S
SMT. INDIRA GANDHI COLLEGE OF ENGINEERING
GHANSOLI, NAVI MUMBAI – 400 709
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
AIML&IOT DEPARTMENT(2022-23)

Subject:- Statistics for AI&DS | Exam:-Unit Test-I | Date:-29/08/2022
Time:- 11-12 | Sem: V | Max.Marks:-20

Q. No.	Sub Q.No.	Question	Course Outcome	Cognition Level	Marks
1		Attempt any five of following			10
	a)	Explain two types of structured data.	CO1	Remember	02
	b)	The mean of 6, 8, x + 2, 10, 2x - 1, and 2 is 9. Find the value of x and also the value of the observation in the data.	CO1	Understand	02
	c)	The runs scored in a cricket match by 11 players are as follows: 7, 16, 121, 51, 101, 81, 1, 16, 9, 11, 16 .Find the mean, mode, median of this data.	CO1	Analyze	02
	d)	Define Continuous Probability distribution and Probability Distribution Function(PDF)	CO1	Remember	02
	e)	Define Normal distribution.	CO2	Remember	02
	f)	X is a normally distributed variable with mean $\mu = 30$ and standard deviation $\sigma = 4$. Find a) $P(x < 40)$, b) $P(30 < x < 35)$	CO2	Analyze	02
2		Attempt any ONE of following			05
	a)	Discuss Boot strapping Vs re-sampling	CO2	Understand	05
	b)	Find the standard error of the estimate of the mean weight of high school football players using the data given of weights of high school football players from your school.	CO1	Analyze	05
3		Attempt any ONE of following			05
	a)	Find the standard deviation of the average temperatures recorded over a five-day period last winter: 18, 22, 19, 25, 12	CO1	Analyze	05
	b)	An agent sells life insurance policies to five equally aged, healthy people. According to recent data, the probability of a person living in these conditions for 30 years or more is 2/3. Calculate the probability that after 30 years: a.All five people are still living b.Atleast three people are still living c.Exactly two people are still living	CO2	Analyze	05

Table for 2 b):

[illegible]	1	2	3	4	5	6	7	8	9	10
P	156	203	176	196	168	195	189	178	197	172

SARASWATI Education Society's

SARASWATI College of Engineering

Learn Live Achieve and Contribute

Kharghar, Navi Mumbai - 410 210.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING IN DATA SCIENCE

Academic Year 2022-23 (ODD SEM)

Internal Assessment - I

Class/Sem : TE/V **Duration:** 1Hr **Date:** 30 /08 / 2022

Subject: Statistics for Artificial Intelligence Data Science **Marks:** 20

Q No	Question	M	CO	BL	PO	PI
Q.1 A	Explain Histogram and scatter plot.	5	1	1	2.8	2.8.2
	O R					
B.	Construct a frequency distribution table for the following weights (in gm) of 30 oranges using the equal class intervals, one of them is 40-45 (45 not included). The weights are: 31, 41, 46, 33, 44, 51, 56, 63, 71, 71, 62, 63, 54, 53, 51, 43, 36, 38, 54, 56, 66, 71, 74, 75, 46, 47, 59, 60, 61, 63. (a) What is the class mark of the class intervals 50-55? (b) What is the range of the above weights? (c) How many class intervals are there? (d) Which class interval has the lowest frequency?	5	1	5	1.2	1.2.1
Q.2 A	Explain the TypeI and Type II error in detail.	5	3	4	4.6	4.6.4
	O R					
B	The standard deviation calculated from two random samples of sizes 9 and 13 are 1.99 and 1.9. Can the samples be regard as drawn from the normal population with same standard deviation? (Given: F0.025=3.51, dof 8 & 12, F0.025=4.2, dof 12 & 8)	5	3	5	4.6	4.6.4

Q.3	a. Explain student t-Distribution in detail.	5	2	5	2.8	2.8.1
	b. The CEO of light bulbs manufacturing company claims that an average light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?	5	1	5	1.2	1.2.1
B	a. Explain Normal and Poisson Distribution.	5	3	5	2.8	2.8.1
	b. Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination. Scores on the GMAT are roughly normally distributed with a mean of 527 and a standard deviation of 112. That is the probability of an individual scoring above 500 on the GMAT? How high must an individual score on the GMAT in order to score in the highest 5%?	5	1	5	1.2	1.2.1

THADOMAL SHAHANI ENGINEERING COLLEGE

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

PERIODIC TEST 1

Year/Sem: TE/V　　　　DATE: 26/08/2022

SUBJECT: Statistics　　　　TIME:2.00 pm – 3.00 pm

<table>
<tr><th colspan="2">(Attempt the following questions)</th><th>Marks (20)</th><th>CO Mapped</th></tr>
<tr><td>Q1)</td><td>Following table shows values of 10 data points of a sample: Find mean, standard deviation , standard error, and 95% confidence interval for the sample.
<table>
<tr><th>Data</th><th>Value</th></tr>
<tr><td>1</td><td>6</td></tr>
<tr><td>2</td><td>7</td></tr>
<tr><td>3</td><td>2</td></tr>
<tr><td>4</td><td>6</td></tr>
<tr><td>5</td><td>2</td></tr>
<tr><td>6</td><td>5</td></tr>
<tr><td>7</td><td>3</td></tr>
<tr><td>8</td><td>2</td></tr>
<tr><td>9</td><td>2</td></tr>
<tr><td>10</td><td>4</td></tr>
</table>
Mean = 3.9
SD = 1.97
SE = 0.62
CI = 1.41</td><td>(4)</td><td>CO-1</td></tr>
<tr><td>Q2)</td><td>Find Correlation coefficient for between variables x & y.
<table>
<tr><th>n</th><th>x</th><th>y</th></tr>
<tr><td>1</td><td>14.2</td><td>215</td></tr>
<tr><td>2</td><td>16.4</td><td>325</td></tr>
<tr><td>3</td><td>11.9</td><td>185</td></tr>
<tr><td>4</td><td>15.2</td><td>332</td></tr>
<tr><td>5</td><td>18.5</td><td>406</td></tr>
<tr><td>6</td><td>22.1</td><td>522</td></tr>
<tr><td>7</td><td>19.4</td><td>412</td></tr>
<tr><td>8</td><td>25.1</td><td>614</td></tr>
<tr><td>9</td><td>23.4</td><td>544</td></tr>
<tr><td>10</td><td>18.1</td><td>421</td></tr>
</table>
Correlation coefficient =
0.97</td><td>(3)</td><td>CO-1</td></tr>
</table>

Paper / Subject Code: 48815 / Statistics for Artificial Intelligence & Data Science (DLOC - I)

TE Sem - V (AIDS) R-19

[Time: 3 Hours] [Marks:80]

N.B. 1. Question No. 1 is compulsory.
2. Attempt any three questions out of remaining five.
3. All questions carry equal marks
4. Assume Suitable data, if required and state it clearly.

Q.1 Attempt any four 20

a) Find the standard deviation of the average temperatures recorded over a five-day period last winter: 19, 21, 18, 24, 12?

b) X is a normally distributed variable with mean $\mu = 30$ and standard deviation $\sigma = 4$. Find:
i) $P(x < 40)$, ii) $P(30 \le x < 35)$?

c) Discuss Boot strapping vs. re-sampling

d) The school principal wants to test if it is true what teachers say – that high school juniors use the computer an average 3.2 hours a day. What are our null and alternative hypotheses?

e) What do you mean by correlation and regression? Explain with example

Q.2 a) Find the value of the correlation coefficient from the data given in the following table: 10

SUBJECT	AGE (X)	GLUCOSE LEVEL(Y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

b) Explain briefly why ANOVA is used? Solve using One-way ANOVA 10

OBSERVATIONS	A	B	C
1	25	31	24
2	30	39	30
3	36	38	28
4	38	42	25
5	31	35	28

method:

Q.3 a) Explain type I & type 2 error in detail. 10
(ii) What is the use of scatter plot and box plot?

b) In a manufacturing unit, four teams of operators were randomly selected and sent to four different facilities for machining techniques training. After the training, the supervisor conducted the exam and recorded the test scores. At 95% confidence level does the scores are same in all four facilities? 10
(Hint: Use Kruskal–Wallis test)

Facility 1	Facility 2	Facility 3	Facility 4
88	77	71	52
82	76	56	65
86	84	64	68
87	59	51	81

Q.4 a) If the sample mean and expected mean value of the marks obtained by 15 students in a class test is 290 and 300 respectively. What is the t-score if the standard deviation of the marks is 50? 10

b) Find out what is the relation between the GPA of a class of students and the number of hours of study and the height of the student 10

GPA	Height	Study Hours
2.9	66	7
3.16	57	7
3.62	64.5	6
2	62	7
3.45	69.5	8
2.8	65	9
3.63	63	6
2.81	68	5
3.33	59.5	4
2.75	64	10
3.86	69	7

Q.5 a) A farmer is trying out a planting technique that he hopes will increase the yield on his pea plants. The average number of pods on one of his pea plants is 145 pods with a standard deviation of 100 pods. This year, after trying his new planting technique, he takes a random sample of his plants and finds the average number of pods to be 147. He wonders whether this is a statistically significant increase. What are his hypotheses and the test statistic? Use a 0.05 significance level. Assume random no of sample n= 144 10

b) Find the simple linear regression equation that fits the given data and coefficient of determination: 10

Hour	Temp
2	21
4	27
6	29
8	86
10	86
12	92

Q.6 a) An agent sells life insurance policies to five equally aged, healthy people. According to recent data, the probability of a person living in these conditions for 30 years or more is 2/3. Calculate the probability that after 30 years if 10

i. All five people are still living.

ii. At least three people are still living.

iii. Exactly two people are still living. (Hint: Binomial Distribution]

b) Write short notes on (any two) 10

i. Confidence Interval

ii. Central Limit Theorem

iii. Standard Error

SAIDS (AIML & DS)

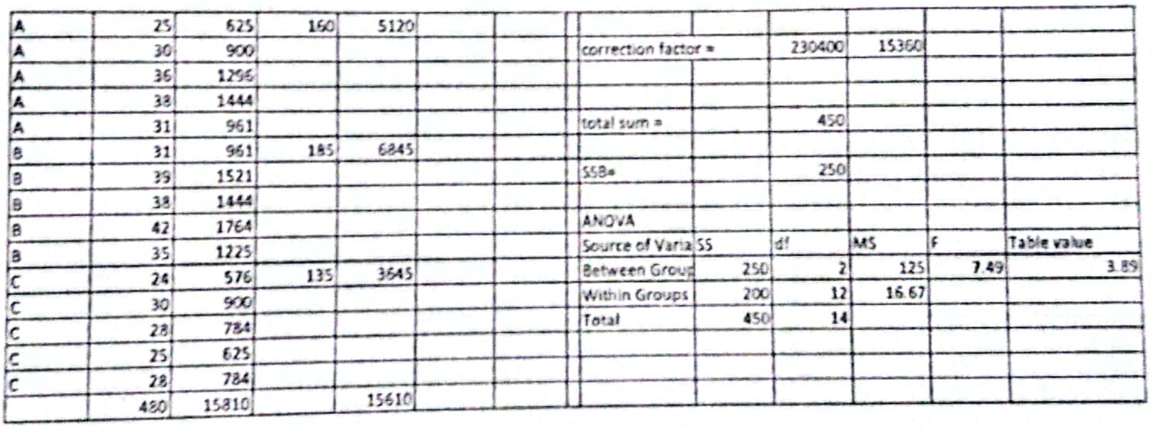
elective Sem V

1	(a)	$\sqrt{19.7}$ = 4.438
	(b)	(i) 0.9938, (ii)=0.3944
	(d)	Ho μ=3.2 hrs, Ha μ≠ 3.2 hrs
2	(a)	0.5298
	(b)	As calculated F=7.5>3.8853 So, H0 is rejected, Hence there is significant differentiation between samples.
3	(b)	While for a right tailed chi-square test with 95% confidence level, and df =3, critical χ^2 value is 7.81. Calculated χ^2 value is greater than the critical value of $\chi^2$for a 0.05 significance level. $\chi^2_{calculated} > \chi^2_{critical}$ hence reject the null hypotheses.
4	(a)	0.7745 or (-.7745)
	(b)	b1=0.038033, b2=-0.10261, a=1.381846, Y=1.38+(0.038*X1)-0.1*X2)
5	(a)	H0: μ ≤ 145 Ha: μ > 145, The critical value will be 1.645. We will reject the null hypothesis if the test statistic is greater than 1.645. The value of the test statistic is 0.24. This is less than 1.645 and so our decision is to fail to reject H0.
	(b)	~~b1=8.1, b0=-3.53, y=-3.53+(8.1*x)~~
		$b1 = 8.414$, $b_0 = -2.066$, $y = -2.066 + 8.414x$, $r = 0.842$
6	(a)	(i) 0.132, (ii) 0.791, (iii) 0.164

3.74

13.37

2(b)



A	25	625	160	5120							
A	30	900					correction factor =		230400	15360	
A	36	1296									
A	38	1444									
A	31	961					total sum =		450		
B	31	961	185	6845							
B	39	1521					SSB=		250		
B	38	1444									
B	42	1764					ANOVA				
B	35	1225					Source of Varia	SS	df	MS	F	Table value
C	24	576	135	3645			Between Grou	250	2	125	7.49	3.89
C	30	900					Within Groups	200	12	16.67		
C	28	784					Total	450	14			
C	25	625										
C	28	784										
	480	15810		15610								

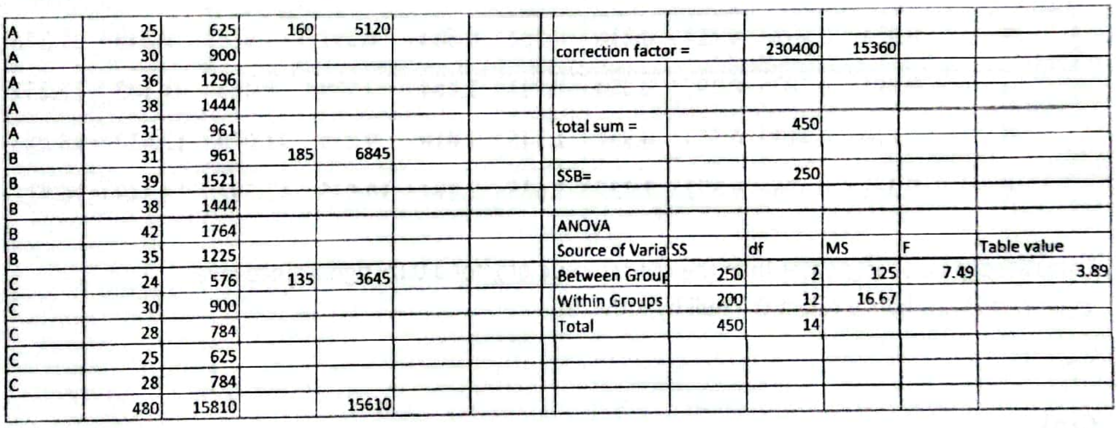
3(b)

- Null Hypothesis H_0: The distribution of operator scores are same
- Alternative Hypothesis H_1: The scores may vary in four facilities

Rank the score in all the facilities

1	(a)	19.7
	(b)	(i) 0.9938, (ii)=0.3944
	(d)	Ho μ=3.2 hrs, Ha $\mu \neq$ 3.2 hrs
2	(a)	0.5298
	(b)	As calculated F=7.5>3.8853 So, H0 is rejected, Hence there is significant differentiation between samples.
3	(b)	While for a right tailed chi-square test with 95% confidence level, and df =3, critical χ^2 value is 7.81. Calculated χ^2 value is greater than the critical value of $\chi^2$for a 0.05 significance level. $\chi^2_{calculated} > \chi^2_{critical}$ hence reject the null hypotheses.
4	(a)	0.7745 or (-.7745)
	(b)	b1=0.038033, b2=-0.10261, a=1.381846, Y=1.38+(0.038*X1)-0.1*X2)
5	(a)	H0: $\mu \leq 145$ Ha: $\mu > 145$, The critical value will be 1.645. We will reject the null hypothesis if the test statistic is greater than 1.645. The value of the test statistic is 0.24. This is less than 1.645 and so our decision is to fail to reject H0.
	(b)	b1=8.1, b0=-3.53, y=-3.53+(8.1*x)
6	(a)	(i) 0.132, (ii) 0.791, (iii) 0.164

2(b)



A	25	625	160	5120									
A	30	900						correction factor =		230400	15360		
A	36	1296											
A	38	1444											
A	31	961						total sum =		450			
B	31	961	185	6845									
B	39	1521						SSB=		250			
B	38	1444											
B	42	1764						ANOVA					
B	35	1225						Source of Varia	SS	df	MS	F	Table value
C	24	576	135	3645				Between Group	250	2	125	7.49	3.89
C	30	900						Within Groups	200	12	16.67		
C	28	784						Total	450	14			
C	25	625											
C	28	784											
	480	15810		15610									

3(b)

- Null Hypothesis H_0: The distribution of operator scores are same
- Alternative Hypothesis H_1: The scores may vary in four facilities

Rank the score in all the facilities

	Facility 1	Facility 2	Facility 3	Facility 4
	88(16)	77(10)	71(8)	52(2)
	82(12)	76(9)	56(3)	65(6)
	86(14)	84(13)	64(5)	68(7)
	87 (15)	59 (4)	51 (1)	81 (11)
T_i	57	36	17	26

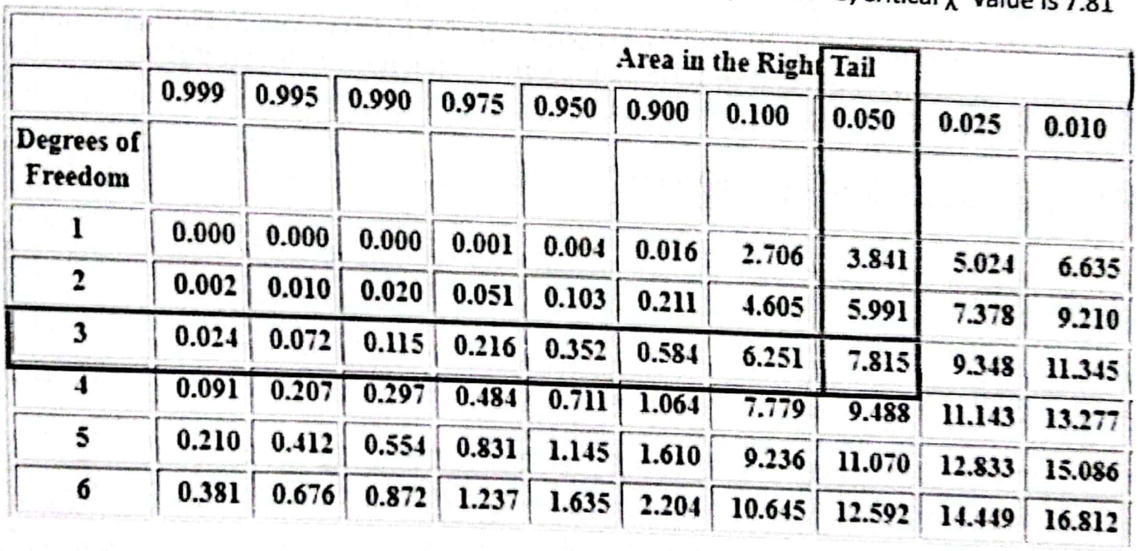
N=16

$$H = \frac{12}{N(N+1)} \sum \frac{T_i^2}{N_i} - 3(N+1)$$

$$H = \frac{12}{16(17)} (\frac{57^2 + 36^2 + 17^2 + 26^2}{4}) - 3(17)$$

$$H = \frac{12}{16(17)} \left(\frac{5510}{4}\right) - 3(17) = 9.77$$

While for a right tailed chi-square test with 95% confidence level, and df =3, critical χ^2 value is 7.81



						Area in the Right Tail				
	0.999	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010
Degrees of Freedom										
1	0.000	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.002	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.024	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.091	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.210	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.381	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812

Calculated χ^2 value is greater than the critical value of $\chi^2$for a 0.05 significance level.
$\chi^2_{calculated} > \chi^2_{critical}$ hence reject the null hypotheses.

6(a)

1. All five people are still living

$B(5, \frac{2}{3}) \qquad p = \frac{2}{3} \qquad 1 - p = \frac{1}{3}$

$p(X = 5) = \binom{5}{5}(\frac{2}{3})^5 = 0.132$

2. At least three people are still living

$$p(X \geq 3) = p(X = 3) + p(X = 4) + p(X = 5)$$

$$= \binom{5}{3}\left(\frac{2}{3}\right)^3\left(\frac{1}{3}\right)^2 + \binom{5}{4}\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right) + \binom{5}{5}\left(\frac{2}{3}\right)^5 = 0.791$$

3. Exactly two people are still living.

$$p(X = 2) = \binom{5}{2}\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^3 = 0.164$$