SRN □□□□□□□□□□□□□

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE15CS203

END SEMESTER ASSESSMENT (ESA) B.TECH. III SEMESTER-Dec. 2016

UE15CS203 – Introduction to Data Science

Time: 3 Hrs | Answer All Questions | Max Marks: 100

Note: All answers must be precise and to the point. IDS handbook must be provided for reference.

1. a) The diagrams below represent the box plots for the amount of time girls (Fig. 1) and boys(Fig. 2) spend per day on Data Science project. — 5 (2 + 1 + 1 + 1)

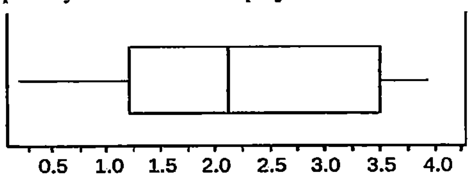


Fig. 1 : Time – girls

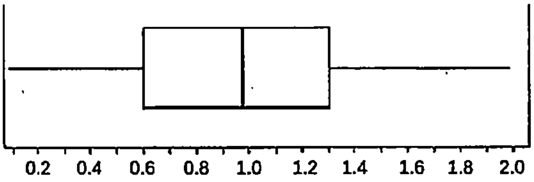


Fig. 2 : Time – boys

Answer the following:

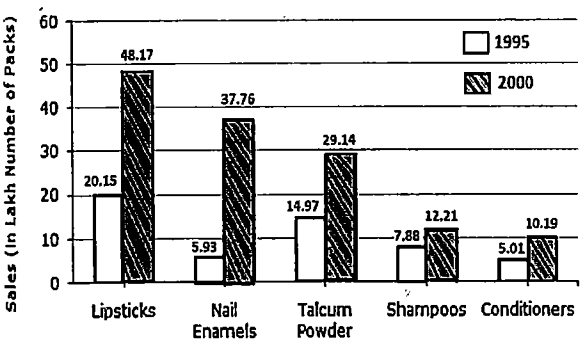
(1) Approximate the girls' IQR and the boys' IQR.

(2) Approximately what percentage of girls spend more than 1.25 hours/day on the project?

(3) Approximately what percentage of boys spend more than 1.35 hours/day on the project?

(4) If one girl spends 6.5 hours/day on the project, would she be considered as an outlier?

b) A cosmetic company provides five different products. The sales of these five products (in lakh number of packs) during 1995 and 2000 are shown in the following bar graph: — 5 (2 + 1 + 2)



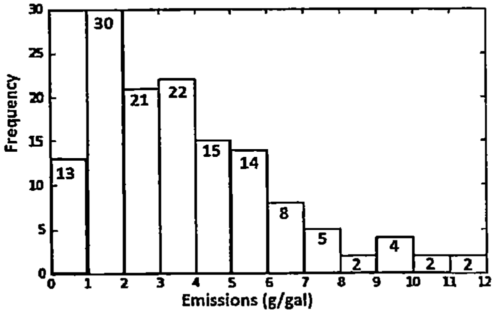
Answer the following:

(1) The sales of lipsticks in 2000 was by what percent more than the sales of nail enamels in 2000?

(2) What is the approximate ratio of the sales of nail enamels in 2000 to the sales of Talcum powders in 1995?

(3) The sales have increased by nearly 55% from 1995 to 2000 in the case of which product?

c) Following is the histogram of Particulate matter (PM) emissions (in g/gal) of vehicles driven at low altitude: — 6 (1+2+1+1+1)



Answer the following:

(1) Estimate the median.

(2) Make a statement about the mean of the data set with respect to the median. Justify your answer.

(3) What are the maximum and minimum values?

(4) Is it a unimodal, bimodal, multimodal or uniform distribution? Give reasons.

(5) How many vehicles have Particulate matter (PM) emissions below 5 g/gal?

d) The number of students in a class who have answered correctly, wrongly, or not attempted each question in an exam, are listed in the table below. The marks for each question are also listed. There is no negative or partial marking. — 4

Q. No.	Marks	Correctly answered	Answered Wrongly	Not Attempted
1	2	21	17	6
2	3	15	27	2
3	1	11	29	4
4	2	23	18	3
5	5	31	12	1

What is the average of the marks obtained by the class in the examination?

2. a) (1) Suppose you were told that scores on an examination were normally distributed with a mean of 500, range of 800, and a standard deviation of 100. A student with a score of 600 has performed better than what percent of the students taking the test? — 5 (3 + 2)

(2) A student with a Z-Score of -2.00 has performed below approximately what percent of the students taking the test?

b) (1) A group of college students are making prank calls. They are dialing numbers randomly and the probability that someone answers the phone on any given call is 0.6. If they make ten calls, what is the probability that exactly seven people will answer the call? — 5 (2 + 1 + 2)

(2) A telephone operator receives calls at a rate of 0.3 per minute. Let X denote the number of calls received in a given 3-minute period.

(a) The distribution of the random variable X is (choose one)

(i) Binomial (ii) Hypergeometric (iii) Negative binomial (iv) Poisson

(b) Find the probability that exactly 1 call arrives in a given 3-minute period.

c) A person arrives at a certain bus stop each morning. The waiting time, in minutes, for a bus to arrive is uniformly distributed on the interval (0, 15). — 6 (2 + 4)

(1) Find the probability that the waiting time is between 5 and 11 minutes.

(2) Suppose that waiting times on different mornings are independent. What is the probability that the waiting time is less than 5 minutes on exactly 4 of 10 mornings?

d) A fair six-sided die is tossed. You win Rs.20 if the result is a "1," you win Rs.10 if the result is a "6," but otherwise you lose Rs.10. Let X represent the amount won or lost. — 4 (1 + 2 + 1)

(1) Is X a discrete random variable or continuous random variable?
(2) Write down the probability distribution of X.
(3) Find E(X).

3. a) Let $X_1, X_2, \ldots X_n$ be n random variables such that, for each i = 1 to n, $X_i \sim$ Geom(p). Find the MLE of p. — 4

b) (1) Two students are doing a statistics project in which they drop toy parachuting soldiers off a building and try to get them to land in a hula-hoop target. They count the number of soldiers that succeed and the number of drops total. In a report analyzing their data, they write the following: — 6 (2 + 2 + 2)

"We constructed a 95% confidence interval estimate of the proportion of jumps in which the soldier land in the target, and we got [0.50, 0.81]. We can be 95% confident that the soldiers land in the target between 50% and 81% of the time. Because the army desires an estimate with greater precision than this (a narrower confidence interval) we would like to repeat the study with a larger sample size, or repeat our calculations with a higher confidence level."

Is there any error in the report?

		(2) **Answer the following:** **a)** Is it appropriate to use Student's t distribution to find Confidence Interval of mean for the following data: 10, 12, 5, 7, 9, 6. Justify your answer. **b)** Make necessary changes in above data if required and write python code to find mean, SD and confidence interval for the mean of the data using t table. **[Max no of lines in the code : 6]**	
	c)	Based on a large sample of 100 capacitors of a certain type, a 95% confidence interval for the mean capacitance, in μF, was computed to be (0.213, 0.241). (1) Find a 90% confidence interval for the mean capacitance of this type of capacitor. (2) How large a sample is needed so that a 95% confidence interval will specify the mean to within ±0.01?	6 (3 + 3)
	d)	An article reports that out of 10,500 surgeries, 850 resulted in complications within six months of surgery. A surgeon claims that the rate of complications is less than 8.5%. With what level of confidence can this claim be made?	4
4.	a)	The manager at Orion mall Hypercity Store assumes the Store's employees are honest. However, there have been many shortages from the cash register lately. There is only one employee who could have taken money during these periods. Realizing that the shortages might have resulted from the employee inadvertently giving incorrect change to customers, the employer does not know whether to forget the situation or accuse the employee of theft. (1) In words, what are the null and alternative hypotheses? Explain. (2) What constitutes a Type I error in this problem? (3) What constitutes a Type II error in this problem? (4) Which do you think is more serious in this problem– Type I or Type II? Explain.	6 (2 + 1 + 1 + 2)
	b)	A reading coordinator in a large public school system suspects that poor readers may test lower in IQ than children whose reading is satisfactory. He draws a random sample of 30 fifth grade students who are poor readers. Historically fifth grade students in the school system have had an average IQ of 105. The sample of 30 has mean 101.5 and standard deviation 1.42. Test the appropriate hypothesis at the 2% level.	4
	c)	**Use Mann–Whitney test to solve the following:** A new post-surgical treatment is being compared with a standard treatment. Seven subjects receive the new treatment, while seven others (the controls) receive the standard treatment. The recovery times, in days, are as follows: Treatment (X) : 12 13 15 19 20 21 27 Control (Y) : 18 23 24 30 32 35 40 Can you conclude that the mean rate differs between the treatment and control? **[State null and alternate hypotheses]**	5
	d)	Write pseudocode or Python code assuming a certain number of equal width intervals, N, to check whether the given data in file "height.csv" is sampled from a normal population, using Chi square goodness-of-fit test. **[State appropriate null and alternate hypotheses]**	5
5.	a)	**Answer the following:** (1) A researcher carefully computes the correlation coefficient between two variables and gets $r = 1.12$. What does this value mean?	5 (1 + 1 + 1 + 2)

SRN

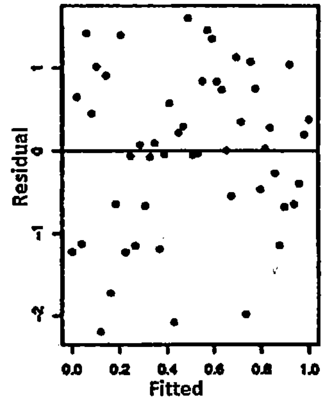
(2) It has been noted that there is a positive correlation between the U.S. economy and the height of women's hemlines (distance from the floor of the bottom of a skirt or dress) with shorter skirts corresponding to economic growth and lower hemlines to periods of economic recession. Comment on the conclusion that economic factors cause hemlines to rise and fall.

(3) Why linear regression is sometimes referred to as least squares?

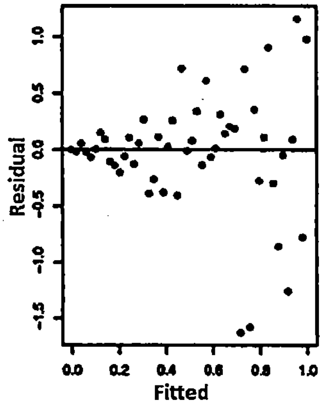
(4) With a short explanation decide whether the statement is true or false: "*We consider the model $y = \beta_o + \beta_1 x + \varepsilon$. Let [-0.01, 1.5] be the 95% Confidence interval for β_1. In this case, a t-Test with significance level 1% rejects the null hypothesis H0: $\beta_1 = 0$.*"

b) Identify Homoscedastic and Heteroscedastic graphs. Justify your answer. 5

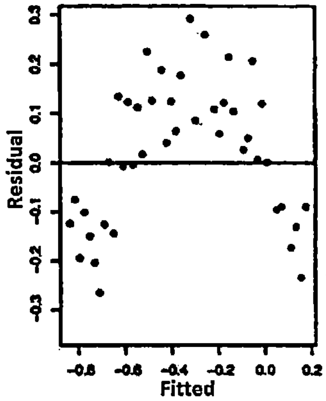
Graph 1



Graph 2

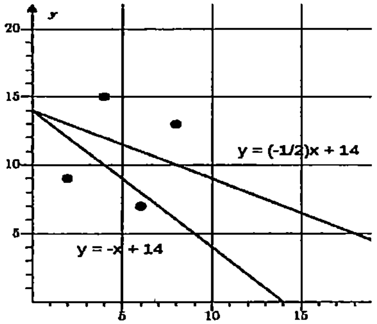


Graph 3



c) The graph shows the data in the table and two lines of fit: 5

x	2	4	6	8
y	9	15	7	13



Which line is a better fit for the data? Justify your answer numerically.

Ⓐ $y = -\frac{x}{2} + 14$ Ⓑ $y = -x + 14$

d) What is Web Scraping? Write python program to scrape all URLs in anchor tag inside all list tags <li> from the following web page: "https://www.smashingmagazine.com/2009/04/from-table-hell-to-div-hell/" . **[Max No. of lines in the code : 10]** The sample source code is given below: 5 (1 + 4)

```
<li class="rss">
    <a href="https://www.smashingmagazine.com/feed/" title="Subscribe to our RSS-feed (120K)">RSS</a>
</li>
<li class="fb">
    <a href="//www.facebook.com/smashmag" title="Join our Facebook page! (267k)">Facebook</a>
</li>
<li class="tw">
    <a href="//twitter.com/smashingmag" title="Follow us on Twitter! (956k)">Twitter</a>
</li>
```