SRN □□□□□□□□□□□□

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE16CS203

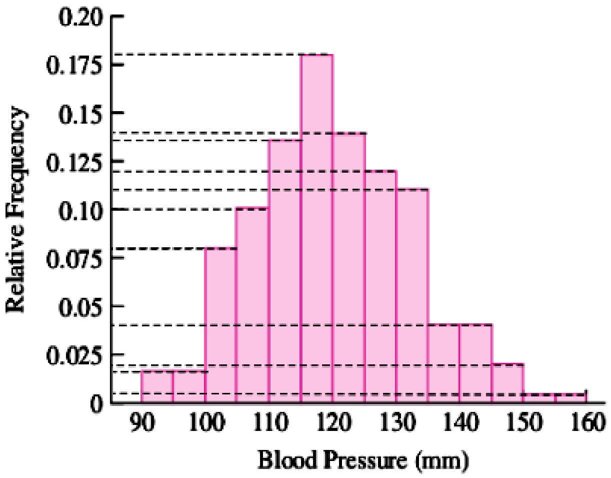
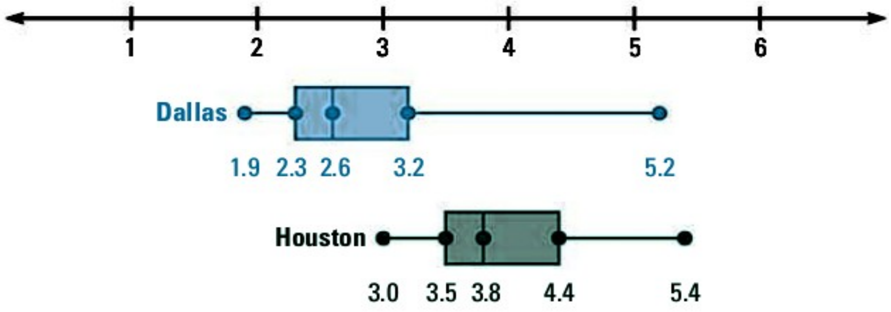
END SEMESTER ASSESSMENT (ESA) B.TECH. III SEMESTER-Dec. 2017

UE16CS203 – Introduction to Data Science

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

Note:

1. **All answers must be precise and to the point.**
2. **IDS Handbook/Formula sheet will be provided to all the Students.**

1.	a)	The given histogram presents the distribution of systolic blood pressure for a sample of women. Use it to answer the following questions. **I. Is the percentage of women with blood pressures above 130 mm closest to 25%, 50%, or 75%?** **II. In which interval are there more women: 130–135 or 140–150 mm?** **III. Comment on the shape of the histogram.** 	5
	b)	What is Data Science? Explain type of statistics with an example each.	5
	c)	The box-and-whisker plots below show the normal precipitation (in inches) each month in Dallas and in Houston, Texas.  **I. For how many months is Houston's precipitation less than 3.5 inches?** **II. Compare the precipitation in Dallas with the precipitation in Houston. (Mention 3 points)** **III. For how many months was the precipitation in Dallas more than 2.6 inches?**	10 (2 + 6 + 2)
2.	a)	A distribution sometimes used to model the largest item in a sample is the extreme value distribution. This distribution has cumulative distribution function, which is given as: $$F(x) = e^{-e^{-x}}$$ **Find P(X > ln 2).**	5

SRN

b) Let X ∼ Geom(p), let n be a non-negative integer, and let Y ∼ Bin(n, p). Show that, **P(X = n) = (1/n)P(Y = 1)**

5

c) Write Python code to construct Sampling distribution of Sample proportion. Run your code 100 times, for each case where, sample size is 50, 100 and 1000 [**that means, define a function called sampling which takes two arguments : sample size and noOfSamples**]. Also print the mean and standard error of such a distribution. You must also specify the output of the code (mainly the diagram and the dummy prints, if any). Interpret your output at the end.

Assume that you are reading data from the file height-weight.csv, aiming to find the proportion of people who are overweight. The file contains two columns : height(in inches) and weight (in Kg's). You must consider those entries as overweight where the weight value is greater than or equal to 70 Kg's and height is less than or equal to 60 inches. Importing required packages is mandatory.

10

3. a) Suppose that X is a discrete random variable with the given probability distribution: where 0 ≤ θ ≤ 1 is a parameter. The following 10 independent observations were taken from such a distribution: (3, 0, 2, 1 ,3, 2, 1, 0, 2 , 1). What is the maximum likelihood estimate of θ.

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

5

b) Comment on the accuracy of Continuity Correction. In a certain large university, 25% of the students are over 21 years of age. In a sample of 400 students, what is the probability that more than 110 of them are over 21?

5
(2 + 3)

c) Leakage from underground fuel tanks has been a source of water pollution. In a random sample of 87 gasoline stations, 13 were found to have at least one leaking underground tank.

I. Find a 95% confidence interval for the proportion of gasoline stations with at least one leaking underground tank.

II. How many stations must be sampled so that a 95% confidence interval specifies the proportion to within ±0.03?

10
(5 + 5)

4. a) Write python code to perform normality check of a given sample using the chi-square test. Specify the particular chi-square test to be used. Mention appropriate null and alternate hypothesis. Write the algorithm first and then the code. Also specify how the output is interpreted at the end.

5

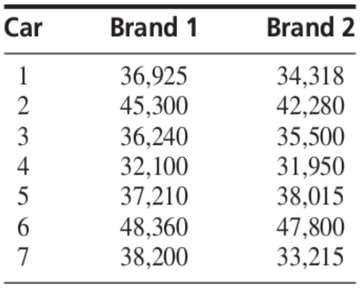
b) A machine manufactures bolts that are supposed to be 3 inches in length. Each day a quality engineer selects a random sample of 50 bolts from the day’s production, measures their lengths, and performs a hypothesis test of H 0 : μ = 3 versus H 1 : μ ≠ 3, where μ is the mean length of all the bolts manufactured that day. Assume that the population standard deviation for bolt lengths is 0.1 in. If H 0 is rejected at the 5% level, the machine is shut down and recalibrated.

If the true mean bolt length on a given day is 3.01 in., find the rejection region. Depict the same information in the form of a diagram with required details.

5

SRN

c) An automobile manufacturer wishes to compare the lifetimes of two brands of tire. She obtains samples of seven tires of each brand. On each of seven cars, she mounts one tire of each brand on each front wheel. The cars are driven until only 20% of the original tread remains. The distances, in miles, for each tire are presented in the given table. Can you conclude that there is a difference between the mean lifetimes of the two brands of tire? — 10



Car	Brand 1	Brand 2
1	36,925	34,318
2	45,300	42,280
3	36,240	35,500
4	32,100	31,950
5	37,210	38,015
6	48,360	47,800
7	38,200	33,215

I. State the appropriate null and alternate hypotheses.

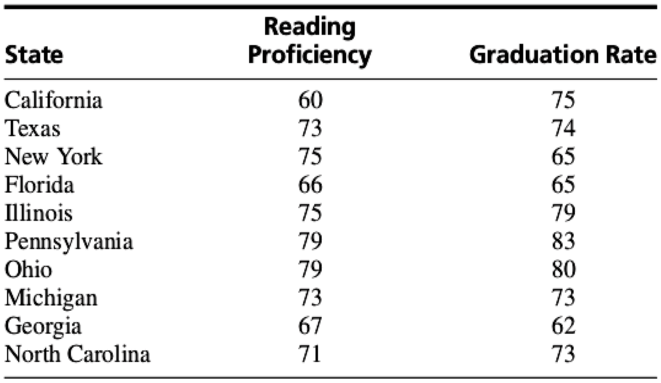
II. Compute the value of the test statistic.

III. Find the P-value and state your conclusion.

5. a) An experiment to determine the effect of load on the drift in signals derived from a piezoelectric force plates is performed. The correlation coefficient between output and time under a load of 588 N was −0.9515. Measurements were taken 100 times per second for 300 seconds, for a total of 30,000 measurements. Find a 95% confidence interval for the population correlation *ρ*. — 5

b) What is a Sitemap? Specify the code to scrape all url's present in the sitemap. Import required packages. — 5

c) The National Assessment for Educational Progress measured the percentage of eighth grade students who were proficient in reading and the percentage of students who graduated from high school in each state in the U.S. The results for the ten most populous states are as follows: — 10 (5 + 2 + 3)



State	Reading Proficiency	Graduation Rate
California	60	75
Texas	73	74
New York	75	65
Florida	66	65
Illinois	75	79
Pennsylvania	79	83
Ohio	79	80
Michigan	73	73
Georgia	67	62
North Carolina	71	73

Reading data from 2005, graduation data from 2007

I. Construct a scatterplot of graduation rate (y) versus reading proficiency (x). Which state is an outlier?

II. Compute the least-squares line for predicting graduation rate from reading proficiency, using the data from all ten states.

III. Remove the outlier and compute the least-squares line, using the data from the other nine states. Is the outlier an influential point? Explain.