Test 2 Statistics for Data Science, 4th Semester Course,
RVCE Bangalore Marks 50

Every question given below is a 5 mark question. All the best to you all!!!

1. Recall that the variance of a random variable X is given by $Var[X] = E[X^2] - (E[X])^2$. With this information, justify with a **logical** reason as to which of the two fellows given below are correct?

Batman: There ***DOES*** exist a random variable X with $E[X^2] = 1/9$ and $E[X] = 1/2$.

Spiderman: There ***DOES NOT*** exist a random variable X with $E[X^2] = 1/9$ and $E[X] = 1/2$.

Solution: Spiderman is correct as $Var[X] = E[X^2] - (E[X])^2$ cannot be negative. With the given values Var[X] = 1/9 - 1/4 < 0. Hence, there does not exist a random variable X with the given $E[X^2]$ and E[X].

2. Consider the entire population of undergraduate students in RVCE Bangalore. Identify exactly 4 unique pairs of features/variables (X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4), associated with students, such that the pairs

(a) (X_1, Y_1) are positively correlated

Solution: Number of hours they study and the academic grades they get.

(b) (X_2, Y_2) are negatively correlated

Solution: Number of classes they bunk and their attendance percentage

(c) (X_3, Y_3) have zero correlation

Solution: Number of students in RVCE and the number of buses ply in Bangalore

(d) (X_4, Y_4) such that Y_4 remains the same irrespective of varying X_4.

Solution: Age of students and the number of fingers they have on their hands.

The above examples are only suggestive!

Justify your choice in each case.

For example, X_1 could be the distance traveled by student and Y_1could be the time he/she spends sleeping. Please think of different examples and list such pairs as asked above. Do not give the same example as the one that you just read above.

3. If A and B are two ***independent events*** on a sample space with probabilities P(A) = 0.3 and P(B) = 0.5, find the value of P(A∪B). What is P(A|B)? What is P(B|A)?

Solution: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) = P(A)P(B)$ for independent A and B.

This implies $P(A \cup B) = 0.3 + 0.5 - 0.15 = 0.65$

P(A|B) = P(A) as the events and independent. Similarly P(B|A) = P(B).

4. In a buffet where 5 different dessert items are served, you like each item with probability 0.7 independent of the other. Identify the random variable that you can associate with the number of items that you taste. What is the probability that you taste at most 3 of them?

Solution: It follows binomial distribution. The required probability is $\sum_{k=0}^{3} {}^{5}C_{k}(0.7)^{k}(0.3)^{5-k}$

5. Compute c and $E[X]$ for the following continuous random variable X.

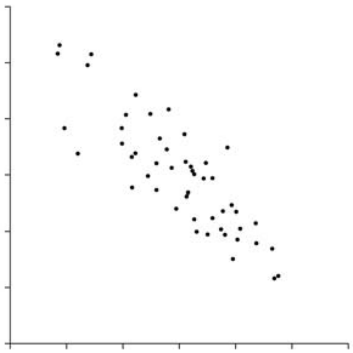
$$f_X(x) = \begin{cases} c(x+1) & 6 \le x \le 8 \\ 0 & \text{otherwise.} \end{cases}$$

Solution: c = 1/16; Integrate the function within the limits 6 to 8 and equate it to 1 to get the value of c = 1/16.

To find E[X]:

E[X] = $\int_{6}^{8} x(x+1)(1/16)\,dx$ = 169/24.

6. Consider the scatter plot below.



The x-axis corresponds to some variable and y-axis another variable (*Do not worry what exact variables they are.*). We are told that exactly one of the following values of correlation coefficient

-0.85, -0.1, 0.1, 0.85, 0

corresponds to the above scatter plot. Identify which among the 5 values given above and justify your choice.

Solution: Correlation of -0.85 corresponds to the plot as we clearly see that as x increases y decreases and the correlation is a bit strong indicating that the value is closer to -1.

7. The following is the probability distribution for a discrete random variable (DRV) X.

$X = x$	-2	-1	0	1	2
$P(X = x)$	c	$2c$	$3c$	c	$3c$

(a) Find c and rewrite the probability distribution table with the exact values.
Solution: c =1/10 . We add the probabilities and equate it to 1. We get 10c = 1 and hence c = 1/10.
(b) Find $P(X \geq 0)$.
Solution: This probability value is the sum of probabilities of all values from 0 to 2. This implies $P(X \geq 0) = 7c = 7/10$
(c) Find $E[X]$.
Solution: $E[X] = \sum_{x=-2}^{2} xp(x)$ = -2c -2c+0c+c+6c = 3c = 3/10 = 0.3.

(d) What will be $E[X]$ if the probability values corresponding to X= 1 is $2c$ and that of $X = 2$ is c i.e, $P(X = 1) = 2c$ and $P(X = 2) = c$?
Solution: The distribution becomes symmetric and hence the expected value becomes 0.

8. The following is a draw of 20 i.i.d. outcomes from a random variable distributed according to Bernoulli law with an unknown parameter p, which is the probability of the random variable taking the value equal to 1. Think of this as the population.

Draw Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Outcome	1	0	1	1	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0

(a) Can you randomly draw 6 samples of size $N = 4$ and get an estimate of the probability p.
Solution: YES

(b) We know that the definition of standard error, the standard deviation of the sampling distribution, is s.e. = $\frac{\sigma}{\sqrt{N}}$, where σ is the standard deviation of the population and is unknown. What would you do to obtain an "estimate" of the variance of the sampling distribution? *(Hint: Think of the formula for the variance of Bernoulli distribution*).
Solution: p(1-p) = 0.35(0.65) = 0.2275

9. Your friend Christopher Nolan from XYZ Engineering college is on a mission to understand whether people like a particular brand of ice-cream, say Baskin Robbins, available in a chain of stores ***belonging to the brand***. He decides to go to the stores and do a sample survey.

(a) Being a student of Statistics for Data Science course at RVCE, give your opinion to Nolan if his approach is correct or not!.

Solution: It is an incorrect approach as the survey is being done in the stores belonging to the brand.

(b) If you think Nolan's approach is correct, justify in, ***not more than 3 sentences,*** why do you think so? If you think his approach has some issues, ***identify them*** and ***suggest to him, in not more than 3 sentences,*** how to fix those issues!

Solution: Any 3 logical solutions will be considered.

10. Given below are the histograms of popular times of visit at MTR Restaurant, located near Lal Bagh Bangalore.

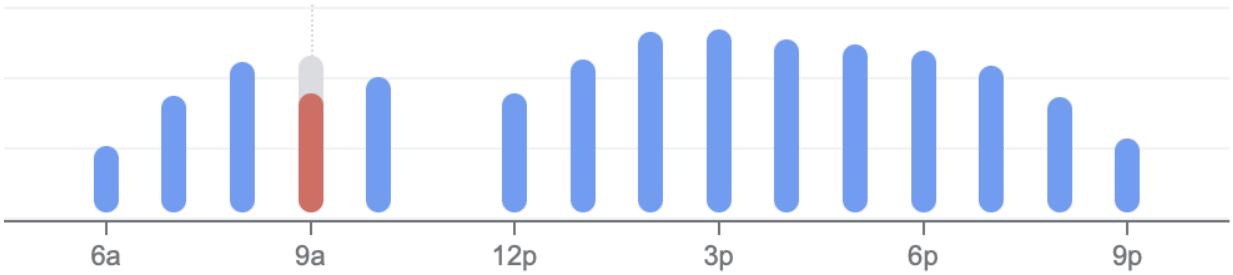


Figure Q10 (a): Histogram of the popular times of visit at MTR Lalbagh on Saturday (*Don't worry about the color red on one of the bars!)*

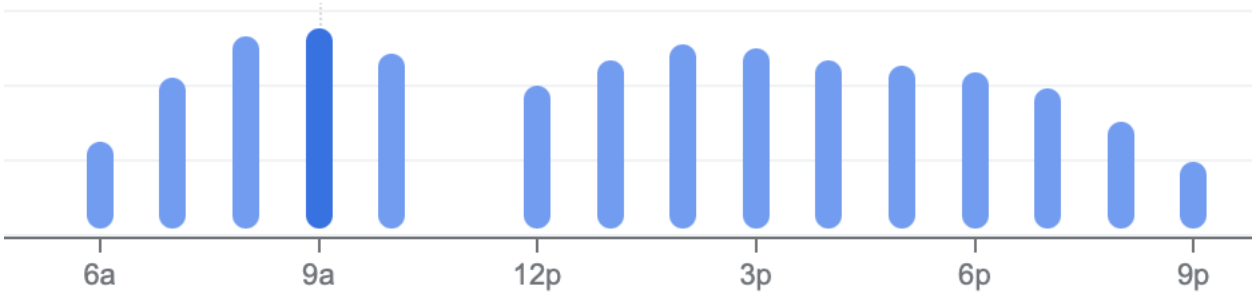


Figure Q10 (b): Histogram of the popular times of visit at MTR Lalbagh on Sunday

(a) What conclusion can you draw about the working hours of MTR Lalbagh?
Solution: It starts at 6am and is open till a little beyond 9pm, as we see crowds at 9pm also.

(b) What is the most popular time band that customers prefer for breakfast at MTR on Saturdays and Sundays? Justify your choice for the above question.
The popular times for breakfast seems to be between 8am to 10am.

(c) On Saturdays, most people prefer to have lunch at 3:00pm. Is this a valid conclusion that you can draw from the figure? Justify your choice.
This cannot be a valid conclusion for the reason that due to more customers arriving, there might be delay in getting seats and hence the lunch time could get pushed beyond. Therefore, we cannot say that people prefer to have lunch at 3pm on saturdays.
Any other valid logical reason will also be considered.