Test 3. 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. Given that X and Y are two Random variables, is it always true that Var(X +Y) = Var(X) + Var(Y)? Justify your choice.

Solution: This is not always true, as the covariance and correlation comes into picture when we talk about two random variables. However, this relation holds good when the variables X and Y are independent or uncorrelated.

2. Identify as to when each of the following cases can occur. A and B are events defined on the sample space S.

(a) P(A|B) = P(A) — A and B are Independent
(b) P(A intersection B) = 0 — A and B are disjoint
(c) P(A|B) = 1 — B is a subset of A.
(d)

3. A professor tries to count the number of students attending his lecture. For each student in the audience, the professor either counts the student properly with probability *p* or overlooks with probability 1- *p*. The exact number of students in the course is 70.

(a) The number of students counted by the professor is a random variable *N*. What is the PMF of *N*?

Solution: A student is properly counted with probability p, independent of any other student being counted. Hence, we have 70 Bernoulli trials and N is a binomial (70; p) random variable with PMF

$$P_N(n) = {}^{70}C_n p^n (1-p)^{70-n}$$

(b) Let *U* = 70 – *N* denote the number of uncounted students. What is the PMF of *U* ?

Solution: P(U) = P(70-N) = $P_U(u) = {}^{70}C_u p^{70-u}(1-p)^u$

4. The daily sales total (excepting Saturday) at a small restaurant has a probability distribution that is approximately normal, with a mean **μ** = \$1230 per day and a standard deviation **σ** equal to \$120.

Solution: We are given the mean μ = \$1230 and standard deviation σ = \$120 of the daily sales. Let X be the random variable corresponding to the sales.

a) What is the probability that the sales will exceed \$1400 for a given day?

P(X > 1400) $= P(\frac{X-\mu}{\sigma} > 1400) = P(z > \frac{1400-1230}{120}) = P(z > 1.416) = 1 - F_X(1.416)$ = 1-0.922 = 0.0778.

b) The restaurant must have at least \$1000 in sales per day to break even. What is the probability that on a given day the restaurant will not break even?

For breaking even,

P(X < 1000) =

$P(\frac{X-\mu}{\sigma} < 1000) = P(z < \frac{1000-1230}{120}) = P(z < -1.92) = 0.0274.$

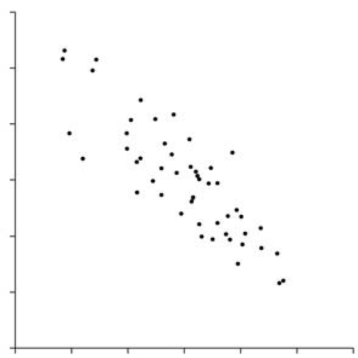
5. Compute c and $E[X]$ for the following continuous random variable X.

$$f_X(x) = cx(1 - x^2) \text{ for } 0 \leqslant x \leqslant 1 \text{ and 0 otherwise.}$$

Solution: $\int_0^1 cx(1 - x^2)\, dx = 1$. This implies c = 4.

E[X} = $\int_0^1 x\,(cx)(1 - x^2)\, dx$ = 1/30

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.*). You figured out in your CIE 2, their correlation was close to -0.85. *Assume that the data was collected by one of your classmates who is known for sleeping, sitting at one of the corners of your classroom.* He wakes up to tell you all that instead of plotting (x, y) he had plotted (y, x). Now, identify among the following, -0.85, -0.1, 0.1, 0.85, 0, which is the correct value of correlation for the correct plot, and justify your choice.

Solution. The covariance/correlation does not change as correlation(X, Y) = Correlation(Y,X).

7. List out 5 important requirements for modeling a random variable as a binomial random variable.

Solution: (a) The number of trails must be finite

(b) The probability of success and failure must be identical across the trials

(c) The outcome of each trial must be independent of each other.

(d) On each trial, the event of interest either occurs or not.

8. There is a new movie of a very famous actor that just got released. You visit the theater and conduct a survey among the people on the first day, first show of the release, to review the movie. Explain what could go wrong in your review about the movie with this survey technique. How would you overcome the problem you just identified? *Please restrict your answer to not more than 3 or 4 sentences together for the two questions.*

Solution: The movie review will be biased since we will take the opinion of the diehard fans of the actor on the first day first show.
Any logical suggestion for overcoming this problem will be considered as answer.

9. Suppose that 70% of the families in your (very large) city have no dogs, 22% have 1 dog and 8% have 2 dogs. (a) Let X be the number of dogs that a randomly chosen family has.
a) Compute E(X) and Var(X).
Solution:

X= x	0	1	2
P(X = x)	0.7	0.22	0.08

$E[X] = 0(0.7) + 1(0.22) + 2(0.08) = 0.3$

$Var[X] = E[X^2] - (E[X])^2 = 0(0.7) + 1(0.22) + 4(0.08) - (0.3)^2 = 0.54 - 0.09 = 0.45$

b) Assuming your 200 family neighborhood constitutes a random sample and that families make their choices about dog ownership independently approximate the probability that there are more than 90 dogs in your neighborhood. *Use the central limit theorem.*

Solution: $P(X > 90) = P(\frac{X-\mu}{\sqrt{np(1-p)}} > 90) = P(z > \frac{90-200(0.3)}{\sqrt{90}}) = P(z > 3.16) = 1-0.9992 = 0.0008$

10. Is it possible to plot 4 dimensional data in 2 dimensions? Justify with not more than 3 reasons, your choice.
Solution: YES. One can use different shapes, colors, intensity maps etc to represent higher dimensional data (typically 3 or 4 or 5) on a 2D plot.