Proficience – 2017

Basics of Data Analytics

First Test

Duration 1:00 hour | September 23rd, 2017 | Maximum Marks : 50

Part A — **Each question carries 2 marks: 15x2 = 30 Marks**

1. List the three axioms of Probability
2. Define conditional probability
3. Define Random Variable
4. What is a biased sample?
5. In a bivariate data, write the equations for determining the value of the regression coefficients which will minimize the sum of the squared errors.
6. If X is a random variable uniformly distributed over an interval (a,b).

   That is, $f(x) = \frac{1}{b-a}\ ; a \leq x \leq b$

   What is the expected value of X?
7. Write the expression for the total probability of an event.
8. X is a continuous random variable with density function given below.

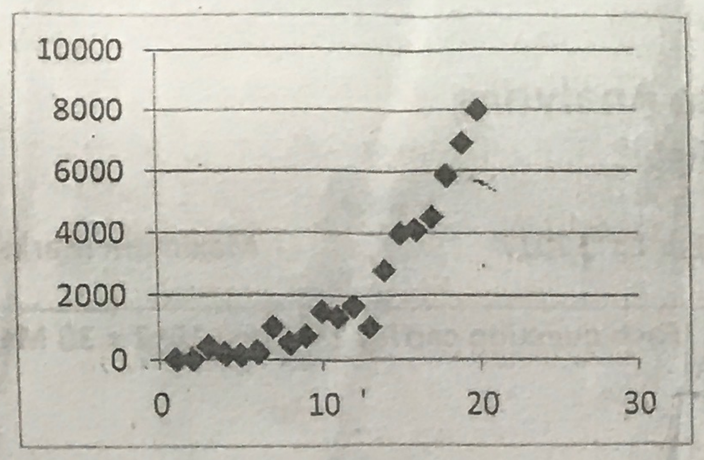
   $f(x) = \frac{a}{x^3}$ $\quad 1500 \leq x \leq 2500$

   $= 0$ $\quad$ else where

   find the value of a
9. 100 observations were made of two variables. Based on this sample, the correlation coefficient was found to be 0.275. Is this a significant value?
10. Write the condition to be satisfied for two random variables to be
    a. mutually exclusive,
    b. independent.
11. A set of observed values of pulp production in metric tons and world pulp price in rupees was analyzed and the following statistics were calculated.
    a. Mean of pulp production
    b. Mean of world pulp price
    c. Standard deviation of pulp production
    d. Correlation coefficient between pulp production and world pulp price

    What are the units of these statistics?

12. If a random variable X has a mean of $\overline{X}$, then what is the mean of the Random Variable $Y = X - \overline{X}$?
13. The scatter plot of a set of observed values is shown below. Write your observation on the relation between the two variables.



14. 10 observations were made of a random variable and the mean was found to be 10. A 11^{th} observation of the variable yielded a value of 21. What is the new value of the mean including the 11^{th} value?
15. Write the density functions of any two standard continuous distributions.

Part B **Each question carries 10 marks: 2x10 = 20 Marks**

1. Given a sample of bivariate data, list the steps to be followed to build a prediction model.
2. Skin cancer rates have been steadily rising over recent years. It is thought that this may be due to ozone depletion. The following data are ozone depletion rates in various localities and the rates of skin cancer.

Ozone depletion (%)	5	7	13	14	17	20	26	30	34	39	44
Skin cancer rate (%)	1	1	3	4	6	5	6	8	7	10	9

a. Fit a straight line regression model to the data
b. What is the rate of skin cancer if ozone depletion is 40%?

10 Marks

Part C OPTIONAL

1. Prove that the value of the correlation coefficient can lie between -1 and +1 only.

Proficience – 2017

Data Analytics

Final Examination

Duration : 2:00 hours | 9th December, 2017 | Maximum Marks : 50

Part A **5 x 6 = 30 Marks**

1. Draw a scatter plot of the following time series data. Identify if the time series is stationary in mean. If not, make the time series stationary.

   1, 2.75, 4, 5.5, 7.25, 8.75, 10.5, 11.5,13 and 14.25

2. The following table has the test scores of various workers and their subsequent production ratings.
   a. Compute the coefficient of regression of Y on X
   b. If test score was 90, what would be your forecast of the production rating?

Test Score (X)	Production Rating (Y)	X * Y	X * X
88	89	7832	7744
84	79	6636	7056
86	84	7224	7396
64	66	4224	4096
45	49	2205	2025
67	76	5092	4489
54	59	3186	2916
73	77	5621	5329
52	51	2652	2704
76	76	5776	5776
32	34	1088	1024
Sum	**721**	**740**	**51536**	**50555**

3. Write the steps for analyzing a time series and building an ARIMA model for forecasting
4. What is stationarity with reference to a time series? Draw illustrative graphs of time series which is stationary in
   a. Mean but not variance
   b. Variance but not mean
   c. Both mean and variance
   d. Neither mean nor variance
5. In a given school there are 100 students. Out of those 55 study Mathematics (event A), 25 study Physics (event B), and 20 study both Mathematics and Physics. Based on Baye's theorem, what is the probability that a student picked at random studies Mathematics GIVEN that we know that the student studies physics?

Part B **5 x 2 = 10 Marks**

1. Analyzing a univariate data in the frequency domain is called_____
2. If two mutually exclusive events A and B are also independent, and P(A) = 0.6, then P(B) = _____
3. What do p, d and q mean in the context of ARIMA modeling?
4. Principal Component Analysis is a method to find what?
5. If ACF has no significant value at any lag, the series is _____

Part C **2 x 5 = 10 Marks**

1. Write a note on Sentiment Analytics.
2. Write a note on Bias and Weights with reference to a Natural Network.

Part D (Optional) **Maximum 10 Marks**

1. Write a note on Yule Walker Equations.
2. Oscar has lost his dog in either forest A (with a priori probability 0.4) or in forest B (with a priori probability 0.6). If the dog is alive and not found by the Nth day of the search, it will die that evening with probability N/(N+2). If the dog is in A (either dead or alive) and Oscar spends a day searching for it in A, the conditional probability that he will find the dog that day is 0.25. Similarly, if the dog is in B and Oscar spends a day looking for it there, he will find the dog that day with probability 0.15.

   The dog cannot go from one forest to the other. Oscar can search only in the daytime, and he can travel from one forest to the other only at night.

   a. In which forest should Oscar look to maximize the probability he finds his dog on the first day of the search?

   b. Given that Oscar looked in A on the first day but didn't find his dog, what is the probability that the dog is in A?