ASSIGNMENT- 3

Course Code	19CSC301A
Course Name	Probability and Statistics
Programme	B. Tech
Department	CSE
Faculty	FET

Name of the Student	K Srikanth
Reg. No	17ETCS002124
Semester/Year	5th Semester / 3rd Year
Course Leader/s	Dr Bhargavi Deshpande

RAMAIAH UNIVERSITY
OF APPLIED SCIENCES

<table>
<tr><th colspan="6">Declaration Sheet</th></tr>
<tr><td>Student Name</td><td colspan="5">K Srikanth</td></tr>
<tr><td>Reg. No</td><td colspan="5">17ETCS002124</td></tr>
<tr><td>Programme</td><td colspan="3">B. Tech</td><td>Semester/Year</td><td>5th Semester/ 3rd Year</td></tr>
<tr><td>Course Code</td><td colspan="5">19CSC301A</td></tr>
<tr><td>Course Title</td><td colspan="5">Probability and Statistics</td></tr>
<tr><td>Course Date</td><td>14/09/2020</td><td>to</td><td colspan="3">16/02/2021</td></tr>
<tr><td>Course Leader</td><td colspan="5">Dr Bhargavi Deshpande</td></tr>
<tr><td colspan="6">Declaration

The assignment submitted herewith is a result of my own investigations and that I have conformed to the guidelines against plagiarism as laid out in the Student Handbook. All sections of the text and results, which have been obtained from other sources, are fully referenced. I understand that cheating and plagiarism constitute a breach of University regulations and will be dealt with accordingly.</td></tr>
<tr><td>Signature of the Student</td><td colspan="3"></td><td>Date</td><td></td></tr>
<tr><td>Submission date stamp
(by Examination & Assessment Section)</td><td colspan="5"></td></tr>
<tr><td colspan="3">Signature of the Course Leader and date</td><td colspan="3">Signature of the Reviewer and date</td></tr>
<tr><td colspan="3"></td><td colspan="3"></td></tr>
</table>

Name: K Srikanth	**Faculty of Mathematical and Physical Sciences**		Registration Number: 17ETCS002124
Ramaiah University of Applied Sciences			
Department / Faculty	Mathematics and Statistics / FMPS	Programme	B. Tech.
Semester/Batch	5^{th} / 2018		
Course Code	19CSC301A	Course Title	Probability and Statistics
Course Leader(s)	Dr Bhargavi Deshpande and Dr Subramanyam T		

Course Assessment					
Reg.No.	17ETCS002124	Name of the Student	K Srikanth		
Sections		Marking Scheme	Marks: Max Marks	Marks Scored	CO
Part-A	1.1	Describe the normal distribution	07		
	1.2	Determine the probabilities	03		
		Part-A Max Marks	**10**		
Part-B	2.1	Determine the probabilities	05		
		Determine the expected value and standard deviation	05		
	2.2	State the hypotheses	02		
		Test statistic and calculations	05		
		Interpretation and Conclusion	03		
		Part-B Max Marks	**20**		
Part-C	3.1	State the model and Fit the data	07		
		Prediction and Develop the plot	03		
	3.2	Determine the probabilities	10		
		Part-C Max Marks	**20**		
		Total Assignment Marks	**50**		

Assignment - 3

Instructions to students:

1. The assignment consists of 3 parts
2. The assignment has to be neatly word processed as per the prescribed format
3. The maximum number of pages should be restricted to **35**
4. Use only SI units
5. **Submission Date: 16/01/2021**
6. **Submission after the due date is not permitted**
7. Method of evaluation as per the submission and marking scheme
8. At the end, you are required to comment on -
    a. Benefits you have derived by solving this assignment
    b. Whether assignment was able to assess *module learning outcomes* or not?
9. IMPORTANT: It is essential that all the sources used in preparation of the assignment must be suitably referenced in the text.

Preamble:

The module aims to teach elements of Probability Theory, Distributions and Regression that are useful in modelling and analysis of Computer Science and Engineering systems, especially data science, machine learning, simulation, computer networks and operating systems. Probability spaces, random variables, conditioning, distributions, expectations and Probability Laws are discussed. Stochastic Processes are introduced. Statistics, Statistical estimation and Hypothesis Testing are covered.

Name: K Srikanth Registration Number: 17ETCS002124

Part A

Q-A1.1)

Introduction

The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side and the area under the normal distribution curve represents probability and the total area under the curve sums to one. The normal distribution is often called the bell curve because the graph of its probability density looks like a bell

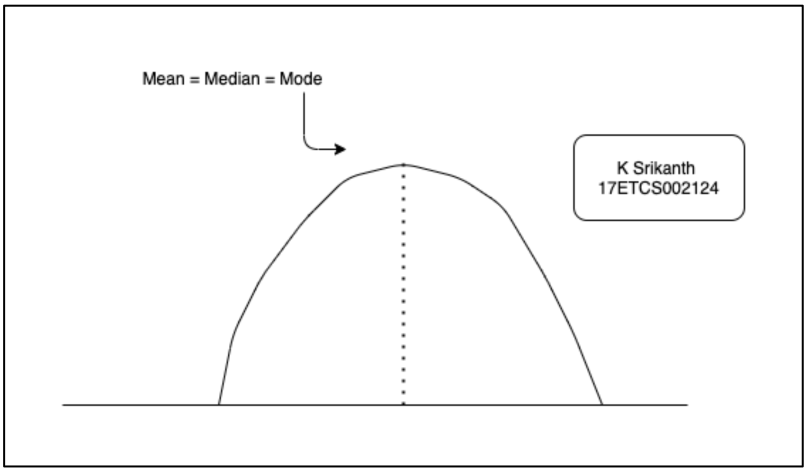

Figure 1 Standard normal model

Properties of a normal distribution

- The mean, mode and median are all equal.
- The curve is symmetric at the centre
- Exactly half of the values are to the left of centre and exactly half the values are to the right.
- The total area under the curve is 1.

1. Cumulative Density Function

2. Probability Density Function

<u>Probability density function</u>

The Probability Density Function(PDF) is the probability function which is represented for the density of a continuous random variable lying between a certain range of values. It is also called a probability distribution function or just a probability function. This function is stated as the function over a general set of values or sometimes it is referred to as cumulative distribution function or sometimes as **Probability Mass Function (PMF).**

Probability Density Function Properties

Let x be the continuous random variable with density function f(x), the probability distribution function should satisfy the following conditions:

- For a continuous random variable that takes some value between certain limits, say a and b, and is calculated by finding the area under its curve and the X-axis, within the lower limit (a) and upper limit (b), then the pdf is given by

$$\boldsymbol{P(x) = \int_a^b f(X)dx}$$

- The probability density function is non-negative for all the possible values, i**.e. f(x)≥ 0,** for all x
- The area between the density curve and horizontal X-axis is equal to 1,

$$\boldsymbol{P(x) = \int_{\infty}^{\infty} f(X)dx = 1}$$

- Due to the property of continuous random variable, the density function curve is continuous for all over the given range which defines itself over a range of continuous values or the domain of the variable.

Cumulative Density Function

The **Cumulative Distribution Function (CDF)**, of a real-valued random variable X, evaluated at x, is the probability function that X will take a value less than or equal to x. It is used to describe the probability distribution of random variables in a table. To determine the probability of a random variable, it is used and also to compare the probability between values under certain conditions. For discrete distribution functions, CDF gives the probability values till what we specify and for continuous distribution functions, it gives the area under the probability density function up to the given value specified.

Cumulative Density Function Properties

The cumulative distribution function X(x) of a random variable has the following important properties:

- Every CDF Fx is non decreasing and right continuous

limx→-∞Fx(x) = limx→+∞Fx(x) = 1

- For all real numbers a and b with continuous random variable X, then the function fx is equal to the derivative of Fx, such that

$$\boldsymbol{Fx(b) - Fx(a) = P(a < X \le b) = \int_a^b fx\,(x)dx}$$

Skew ness

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution.

1σ, 2σ and 3σ limits

- The z-scores for +1σ and –1σ are +1 and –1, respectively and around 68% of the x values lie between –1σ and +1σ of the mean μ,i.e., within one standard deviation of the mean.
- The z-scores for +2σ and –2σ are +2 and –2, respectively and around 95% of the x values lie between –2σ and +2σ of the mean μ ,i.e., within two standard deviations of the mean.
- The z-scores for +3σ and –3σ are +3 and –3 respectively ad around 99.7% of the x values lie between –3σ and +3σ of the mean μ, i.e., within three standard deviations of the mean.

Let X be a normal distribution having the mean μ and variance σ^2.

$$\textbf{Now, } P[\mu < X < x_1] = \int_{\mu}^{x_1} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$P[\mu < X < x_1] = P[0 < Z < z_1] = \int_{0}^{z_1} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}z^2} \sigma dz$$

Area

$$Area = \int_0^{z_1} \phi(z) dz \text{ , where } z = \frac{x-\mu}{\sigma}$$

represents the area under standard normal curve between the ordinates at Z = 0 and Z = z1.

Part B

Q-B2.1)

Given,

Mean μ = 112
Standard Deviation σ = 8

Let X be a random variable that denotes the chemical concentration (mmol/L). We know that X ~ N (μ, σ)

X ~ N (112, 8)

Probability that chemical concentration equals 113

For a continuous random variable, probability at a fixed point is zero. Which means that:

P (X= 113) = 0

Probability that chemical concentration is less than 105

To find the z for this probability, we use

$$z = \frac{x - \mu}{\sigma} = \frac{105 - 112}{8}$$
$$= -0.875$$

Now**, P (X < 105) = P (z < -0.875) → Equation 1**

Using the negative normal distribution table, we get z at -0.875 as **0.1922**. So, **Equation 1** becomes,

P (z < -0.875) = 0.1922

P (X < 105) = 0.1922 → Equation 2

Probability that chemical concentration is at most 105

Probability that chemical concentration is at most 105 is P(X<=105). It can also be written as:

P (X <= 105) = P (X < 105) + P (X = 105) → Equation 3

From equation (2) we have P(X<105) as 0.1922.

Since P(X=105) is a probability at a fixed point, which means

P (X=105) = 0 → Equation 4

From equations (2) and (4), equation (3) becomes

P (X <= 105) = P (X < 105) + P (X = 105). => P (X <= 105) = 0.1922 + 0 => P (X <= 105) = 0.1922

b)

The chemical concentration is x, the mean is μ and the standard deviation is σ.

If the chemical concentration differs from mean by more than 1 standard deviation, it mean is

$$x - \mu > \sigma \text{ or } x - \mu < -\sigma.$$

So, the probability that chemical concentration differs from mean by more than 1 standard deviation is

$$P(x - \mu > \sigma) + P(x - \mu < -\sigma)$$

Dividing by σ, we get

$$P\left(\frac{x-\mu}{\sigma} > 1\right) + P\left(\frac{x-\mu}{\sigma} < -1\right) \rightarrow \textbf{Equation 5}$$

We know that $\frac{x-\mu}{\sigma}$ is the formula for z.

So, equation 5 becomes,

$$P(z > 1) + P(z < -1)$$

$$(1 - P(z \leq 1)) + P(z < -1) \rightarrow \textbf{Equation 6}$$

Using the positive normal distribution table, we get

$$P(z <= 1) = 0.8413 \rightarrow \textbf{Equation 7}$$

Using the negative normal distribution table, we get

$$P(z < -1) = 0.1587 \rightarrow \textbf{Equation 8}$$

So, using equations (7) and (8), equation (6) becomes,

$$(1 - 0.841) + 0.1587 = 0.3174$$

The probability that chemical concentration differs from mean by more than 1 standard deviation is **0.3174**.

It does not depend on μ and σ values because the probability of 1 is considered, not z.

c)

The extreme 0.15% of the values means, 0.075 % on the extreme left and 0.075% on the extreme right.

So, the probability of z < z1 on the extreme left is 0.075% ⇒ $P(z < z_1) = \frac{0.075}{100}$

$$P(z < z_1) = \mathbf{\mathit{0.00075}} \rightarrow \textbf{Equation 9}$$

Similarly, the z on the extreme right is 1 - 0.075% ⇒ $P(z < z_2) = 1 - \frac{0.075}{100}$

$$P(z < z_2) = \mathbf{\mathit{0.99925}} \rightarrow \textbf{Equation 10}$$

Determining the z values based on the probabilities using the table, we get

⇒ **z = ± 3.15**

$$\textbf{We know that } z = \frac{x - \mu}{\sigma} \Rightarrow x = z\sigma + \mu \rightarrow \textbf{Equation 11}$$

$$\text{So, } x1 = z1 * \sigma + \mu \Rightarrow X_1 = (8 \times 3.15) + 112 \Rightarrow X_1 = 137.2$$

$$x2 = z2 * \sigma + \mu \Rightarrow X_2 = (8 \times -3.15) + 112 \Rightarrow X_2 = 86.8$$

Hence, the most extreme 0.15 % of chemical concentration values are below 86.8 mmol/L and above 137.2 mmol/L.

Q-B2.2)

Given ,

105.6, 90.9, 91.2, 96.9, 96.5, 91.3, 101.1, 105.3, 107.7, 102.6, 98.7, 92.4, 93.7, 104.3,103.5.

Level of significance = 5% ⇒ $\alpha = 0.05$

We are given that n = 15 is the total sample size.

We want to test:

$$H_0: \mu = 100$$

$$H_a: \mu \neq 100$$

at significance level 0.05.

The test statistic to test this hypothesis is

$$T = \frac{\bar{X} - \mu}{S} \sqrt{n} \rightarrow \textbf{Equation 12}$$

where $\bar{X}$ is the mean and S is the standard deviation?

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} xi$$

$$\bar{X} = \frac{105.6 + 90.9 + 91.2 + 96.9 + 96.5 + 91.3 + 101.1 + 105.3 + 107.7 + 102.6 + 98.7 + 92.4 + 93.7 + 104.3 + 103.5}{15}$$

$$\bar{X} = 98.78$$

Using the formula of standard and substituting the values in them, we get,

$$S = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^{n} (xi - x)^2 \right)}$$

$$\mathbf{S = 5.9285}$$

Substituting values of $\bar{X}$ and S in equation (12), we get

$$\boldsymbol{T} = \frac{\mathbf{98.78 - 100}}{\mathbf{5.9285}} \sqrt{\mathbf{15}} \Rightarrow \boldsymbol{T} = -\mathbf{0.2057} * \sqrt{\mathbf{15}} \Rightarrow \mathbf{T = -0.797}$$

Since, the test is two sided, the rejection region is

$$\mathbf{C = \{-\infty, - t_{\alpha/2, df}\} \cup \{t_{\alpha/2, df}, \infty\}} \quad \rightarrow \textbf{Equation 13}$$

We have, α/2 = 0.025

df is the degrees of freedom which is 14.

The corresponding value in the table for $t_{0.025,\ 14}$ *is 2.145.*

So, now we have the rejection region as **C = {-∞, - 2.145} U {2.145, ∞}.**

And since the value of test statistic T = -0.797 doesn't fall into the rejection region, we **accept the hypothesis** because we do not have enough evidence to reject the claim that the population mean reading is 100.

Q-C3.1)

Given,

Number Ordered (X)	Price (Y)	XY
90	120	10800
115	106	12190
121	95	11495
138	70	9660
155	65	10075
182	58	10556
801	**514**	**64776**

We consider Price as Y and Number Ordered as X.

Fit a linear regression model to the data and interpret the coefficients

The linear regression graph is fit into the formula:

$$Y_i = \beta_0 + \beta_1 x \rightarrow \textbf{Equation 14}$$

To find β_1, we use,

$$\beta 1 = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}}$$

$$\beta 1 = \frac{64776 - \frac{(801*514)}{6}}{112159 - \frac{264196}{6}}$$

$$\boldsymbol{\beta 1 = -0.735} \rightarrow \textbf{Equation 15}$$

To find β_0, we use,

$$\beta 0 = \frac{\sum Y}{N} - \beta 1 \frac{\sum X}{N}$$

$$\beta 0 = \frac{514}{6} - (-0.735)\frac{801}{6}$$

$$\boldsymbol{\beta 0 = 183.789} \rightarrow \textbf{Equation 16}$$

Substituting values of β_1 and β_0 in equation 14, we have,

$$Y_i = \beta_0 + \beta_1 x$$

$$\mathbf{Y_i = 183.789 - 0.735\, x} \rightarrow \textbf{Equation 17}$$

This is the linear regression model for the data.

How many units do you think would be ordered if the price were 60?

The price is 60, Y = 60.

Substituting value of 5 in equation 17, we get

$$60 = 183.789 - 0.735\,x \Rightarrow \mathbf{x = 168.42}$$

So, for the price of 60, we can get 168.42 units.

c. Draw a scatter diagram and impose the fitted line of regression.

Code in R

```
> data <- data.frame (Number_Ordered=c(90, 115 ,121, 138, 155, 182),Price=c(120, 106 ,95 ,70 ,65, 58))
> data
  Number_Ordered Price
1             90   120
2            115   106
3            121    95
4            138    70
5            155    65
6            182    58
> linmod = lm(data$Price ~ data$Number_Ordered)
> abline(data$Number_Ordered ~ data$Price)
> abline(linmod, col="red")
> Name <- "K Srikanth"
> Registration_Number <- "17ETCS002124"
```

Figure 2 R Code to Plot Linear Regression using in-built Function

Plot

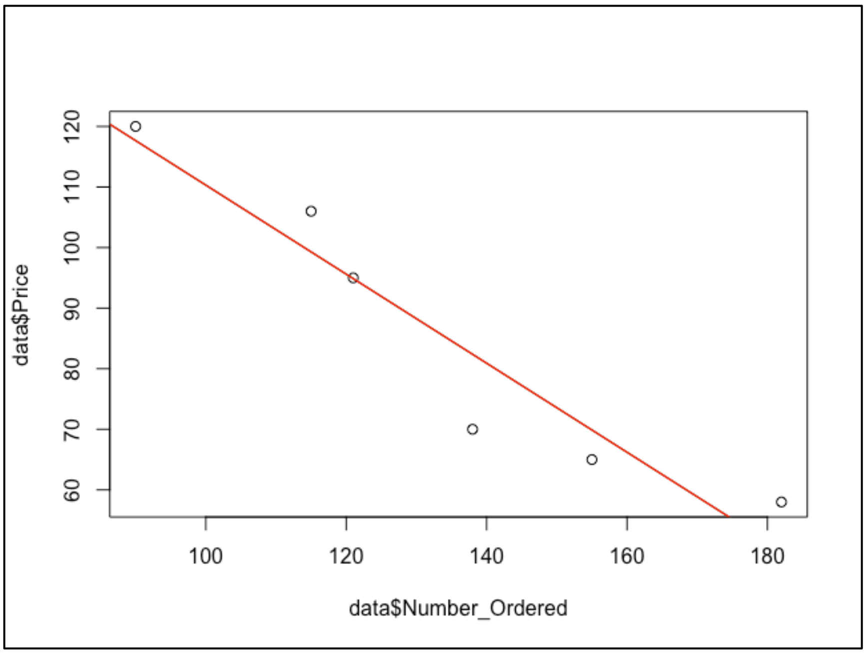

Figure 3 Plotted fitted line of regression

Q-C3.2)

Given,

λ = 0.08 flaws / sq. foot

A = 10 sq. feet

Total μ = 0.8 flaws → **Equation 18**

a)

Using the Poisson distribution formula, we have,

$$\mathbf{P\ (X=0)} = \frac{e^{-\lambda}\ \mu^{\lambda}}{x\,!}$$

Here we have $\boldsymbol{\lambda}$ = 0.8

Since there are no flaws, x = 0

$$\text{P (X =0)} = \frac{e^{-0.8}\ 0.8^{0}}{0\,!} \Rightarrow \text{P (X =0)} = e^{-0.8}$$

$$\text{P (X = 0)} = 0.4493 \quad \rightarrow \textbf{Equation 19}$$

b)

Let X be the number of boilers that have surface flaws in a fleet of 10 boilers.

In equation (19), we have P (X=0) = 0.4493.

Therefore, the probability that a car has any surface flaws

p = (1- 0.4493) = 0.5507

If we treat the boilers as sequence of 10 Bernoulli trials, then X is a binomial random variable with *n* = 10 and *p* = 0.5507 and q = 0.4493.

Probability of at least 2 is, P (X >= 2) = 1 – (P (0) + P (1)) → **Equation 20**

Finding, P (0) and P (1)

P (0) = $^{10}C_0 * p^0 * q^{10}$

P (0) = 1 * $0.5507^0 * 0.4493^{10}$

P (0) = 0.00035

Similarly,

P (1) = $^{10}C_1 * p^1 * q^9$

P (1) = 10 * $0.5507^1 * 0.4493^9$

P (1) = 0.0041035

Substituting values of P (0) and P (1) in equation (20), we have

P(X >= 2) = 1 – (0.00035 - 0.0041035)

P(X >= 2) = 0.9955

c)

Let X be the number of boilers that have surface flaws in a fleet of 12 boilers.

In equation (19), we have P (X=0) = 0.4493.

Therefore, the probability that a car has any surface flaws

p = (1- 0.4493) = 0.5507

If we treat the boilers as sequence of 10 Bernoulli trials, then X is a binomial random variable with *n* = 12 and *p* = 0.5507 and q = 0.4493.

Probability of at most 1 is, P (X <= 1) = P (0) + P (1) → Equation 21

Finding, P (0) and P (1)

$$P(0) = {}^{12}C_0 * p^0 * q^{12}$$

$$P(0) = 1 * 0.5507^0 * 0.4493^{12}$$

P (0) = 0.0000676

Similarly,

$$P(1) = {}^{12}C_1 * p^1 * q^{11}$$

$$P(1) = 12 * 0.5507^1 * 0.4493^{11}$$

P (1) = 0.000995

Substituting values of P (0) and P (1) in equation (20), we have

P (X <= 1) = 0.0000676 + 0.000995

P (X <= 1) = 0.00106