RAMAIAH UNIVERSITY OF APPLIED SCIENCES

ASSIGNMENT

Course Code	**19CSC301A**
Course Name	**Probability and Statistics**
Programme	**B. Tech**
Department	**Computer Science and Engineering**
Faculty	**Engineering and Technology**

Name of the Student	**Deepak R**
Reg. No	**18ETCS002041**
Semester/Year	**5th/2020**
Course Leader/s	**Dr Bhargavi Deshpande**

Declaration Sheet				
Student Name	**Deepak R**			
Reg. No	**18ETCS002041**			
Programme	**B. Tech**	**Semester/Year**	**5th/2020**	
Course Code	**19CSC301A**			
Course Title	**Probability and Statistics**			
Course Date		**to**		
Course Leader	**Dr Bhargavi Deshpande**			

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.

Signature of the Student		**Date**	
Submission date stamp **(by Examination & Assessment Section)**			

Signature of the Course Leader and date	Signature of the Reviewer and date

Faculty of Mathematical and Physical Sciences			
Ramaiah University of Applied Sciences			
Department / Faculty	**Mathematics and Statistics / FMPS**	**Programme**	**B. Tech.**
Semester/Batch	**5th / 2018**		
Course Code	**19CSC301A**	**Course Title**	**Probability and Statistics**
Course Leader(s)	**Dr Bhargavi Deshpande**		

Course Assessment			
Reg.No.	**18ETCS002041**	**Name of the Student**	**Deepak R**

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		

Solution for Part A

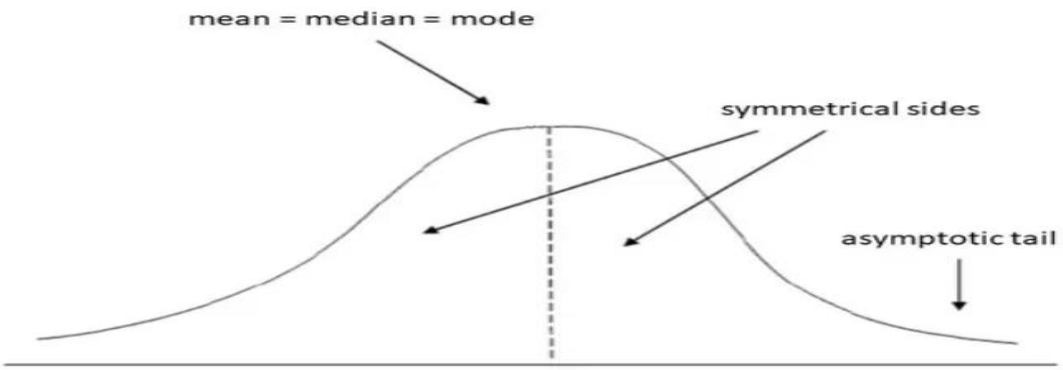
Normal distribution and its characteristics

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.

The area under the normal distribution curve represents probability and the total area under the curve sums to one.

Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon (i.e. x-axis).

For a perfectly normal distribution the mean, median and mode will be the same value, visually represented by the peak of the curve.



The normal distribution is often called the bell curve because the graph of its probability density looks like a bell. It is also known as called Gaussian distribution, after the German mathematician Carl Gauss who first described it.

Probability density function

A normal distribution in a variate X with mean μ and variance σ^2 is a statistic distribution with probability density function

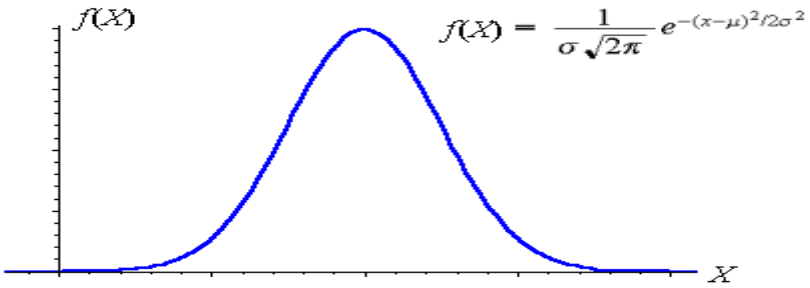
$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \quad \textbf{(1)}$$

on the domain $x \in (-\infty, \infty)$. While statisticians and mathematicians uniformly use the term "normal distribution" for this distribution, physicists sometimes

call it a Gaussian distribution and, because of its curved flaring shape, social scientists refer to it as the "bell curve." Feller (1968) uses the symbol $\varphi(x)$ for $P(x)$ in the above equation, but then switches to $\mathfrak{n}(x)$ in Feller (1971).

de Moivre developed the normal distribution as an approximation to the binomial distribution, and it was subsequently used by Laplace in 1783 to study measurement errors and by Gauss in 1809 in the analysis of astronomical data (Havil 2003, p. 157).

The normal distribution is implemented in the Wolfram Language as NormalDistribution[*mu*, *sigma*].



The so-called "standard normal distribution" is given by taking $\mu = 0$ and $\sigma^2 = 1$ in a general normal distribution. An arbitrary normal distribution can be converted to a standard normal distribution by changing variables to $Z \equiv (X - \mu)/\sigma$, so $dz = dx/\sigma$, yielding

$$P(x)\,dx = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz.$$

Cumulative distribution function

The cumulative distribution function (cdf) of a probability distribution, evaluated at a number (lower-case) *x*, is the probability of the event that a random variable (capital) *X* with that distribution is less than or equal to *x*. The cumulative distribution function of the normal distribution is expressed in terms of the density function as follows:

$$\begin{aligned}\Phi_{\mu,\sigma^2}(x) &= \int_{-\infty}^{x} \varphi_{\mu,\sigma^2}(u)\,du \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R},\end{aligned}$$

where the standard normal cdf, Φ, is just the general cdf evaluated with $\mu = 0$ and $\sigma = 1$:

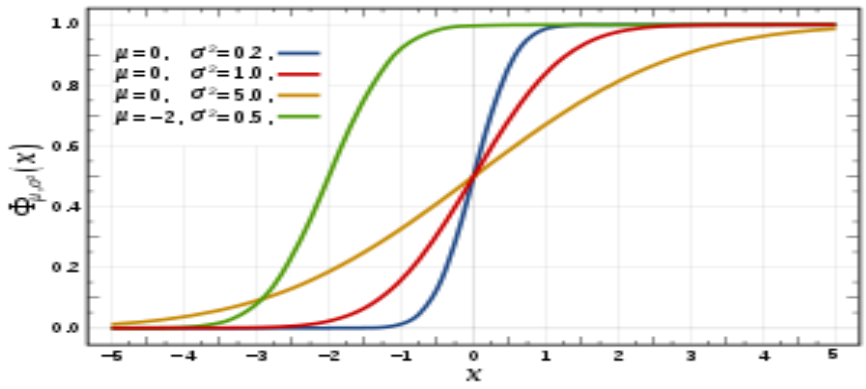
$$\Phi(x) = \Phi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) du, \quad x \in \mathbb{R}.$$

The standard normal cdf can be expressed in terms of a special function called the error function, as

$$\Phi(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right], \quad x \in \mathbb{R},$$

and the cdf itself can hence be expressed as

$$\Phi_{\mu,\sigma^2}(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right], \quad x \in \mathbb{R}.$$



Skewness

For univariate data Y1, Y2, ..., YN, the formula for skewness is:

$$g_1 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3 / N}{s^3}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. Note that in computing the skewness, the s is computed with N in the denominator rather than N - 1.

The above formula for skewness is referred to as the Fisher-Pearson coefficient of skewness. Many software programs actually compute the adjusted Fisher-Pearson coefficient of skewness

$$G_1 = \frac{\sqrt{N(N-1)}}{N-2} \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^3 / N}{s^3}$$

This is an adjustment for sample size. The adjustment approaches 1 as N gets large. For reference, the adjustment factor is 1.49 for N = 5, 1.19 for N = 10, 1.08 for N = 20, 1.05 for N = 30, and 1.02 for N = 100.

The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. If the data are multi-modal, then this may affect the sign of the skewness.

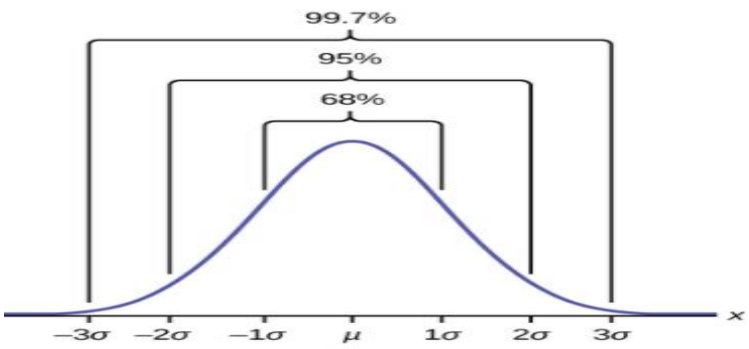
1σ,2σ and 3σ limits

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the Empirical Rule states the following:

- About 68 percent of the x values lie between –1σ and +1σ of the mean μ (within one standard deviation of the mean).
- About 95 percent of the x values lie between –2σ and +2σ of the mean μ (within two standard deviations of the mean).
- About 99.7 percent of the x values lie between –3σ and +3σ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z-scores for +1σ and –1σ are +1 and –1, respectively.
- The z-scores for +2σ and –2σ are +2 and –2, respectively.
- The z-scores for +3σ and –3σ are +3 and –3, respectively.

So, in other words, this is that about 68 percent of the values lie between z-scores of –1 and 1, about 95% of the values lie between z-scores of –2 and 2, and about 99.7 percent of the values lie between z-scores of -3 and 3. These facts can be checked, by looking up the mean to z area in a z-table for each positive z-score and multiplying by 2.

The empirical rule is also known as the *68–95–99.7 rule*.



Area properties

1. The normal curve is symmetrical:

The Normal Probability Curve (N.P.C.) is symmetrical about the ordinate of the central point of the curve. It implies that the size, shape and slope of the curve on one side of the curve is identical to that of the other.

That is, the normal curve has a bilateral symmetry. If the figure is to be folded along its vertical axis, the two halves would coincide. In other words the left and right values to the middle central point are mirror images.

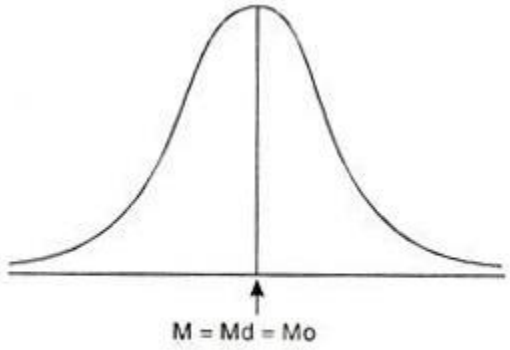


Fig. 6.2 N.P.C., M = Mdn = Mode

2. The normal curve is unimodal:

Since there is only one point in the curve which has maximum frequency, the normal probability curve is unimodal, i.e. it has only one mode.

3. Mean, median and mode coincide:

The mean, median and mode of the normal distribution are the same and they lie at the centre. They are represented by 0 (zero) along the base line. [Mean = Median = Mode]

4. The maximum ordinate occurs at the centre:

The maximum height of the ordinate always occurs at the central point of the curve that is, at the mid-point. The ordinate at the mean is the highest

ordinate and it is denoted by Y_0. (Y_0 is the height of the curve at the mean or mid-point of the base line).

Y_0 is given by $Y_0 = \frac{Ni}{\sigma\sqrt{2\pi}}$ where $\pi = 3.1416 \cdot \sqrt{2\pi} = 2.5066$

5. <u>The normal curve is asymptotic to the X-axis:</u>

The Normal Probability Curve approaches the horizontal axis asymptotically i.e., the curve continues to decrease in height on both ends away from the middle point (the maximum ordinate point); but it never touches the horizontal axis.

It extends infinitely in both directions i.e. from minus infinity (-∞) to plus infinity (+∞) as shown in Figure below. As the distance from the mean increases the curve approaches to the base line more and more closely.

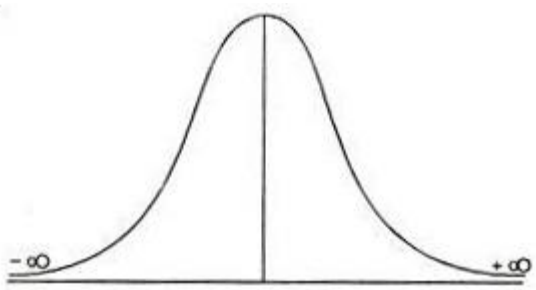


Fig. 6.3 Normal Curve is Asymptotic to the X-axis

6. <u>The total percentage of area of the normal curve within two points of influxation is fixed:</u>

Approximately 68.26% area of the curve falls within the limits of ±1 standard deviation unit from the mean as shown in figure below.

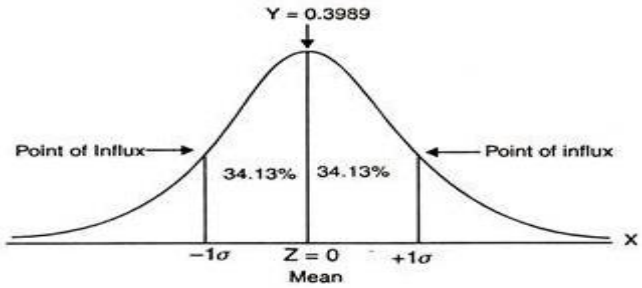


Fig. 6.5 N.P.C., 68.26% area of the curve within the limits of ±1σ

Solution for Question 2

Solution for 2.1

Let X be Random Variable which denotes the chemical concentration (mmol/L).

Since it is given that it is normal distribution form.

X ~ N(112,8)

Here Mean = μ = 112

For part a, we use z scores. The formula for the z score of an individual value is

$$z = \frac{X-\mu}{\sigma}$$

Standard deviation = $\sigma = 8$.

a) P(X = 113) = 0

The probability that x is equal to a given value is 0, regardless of the value, the mean or the standard deviation. This is because the probability in a normal distribution is the area under the curve; this means there must be a range of numbers.

For P(X <105)

z = (105-112)/8 = -7/8 = -0.875 Using a z table, we see that the area under the curve to the right of this is 0.1922.

For P(X ≤ 105)

we use the same probability as P(X < 105). There is no distinction between "less than" and "less than or equal to" So the Answer remains same i.e 0.1922.

b)1 standard deviations above the mean is z = 1. 1 standard deviations below the mean is z = -1.

Using the z table, the area under the curve to the right of z = 1 is 0.8413. The area under the curve to the right of z = -1 is 0.1587. This makes the area between them

0.8413-0.1587 = 0.6826. This means anything farther from the mean than this is

1-0.6826= 0.3174. This does not have anything to do with the value of the mean or the standard deviation; this is because we already have the z score.

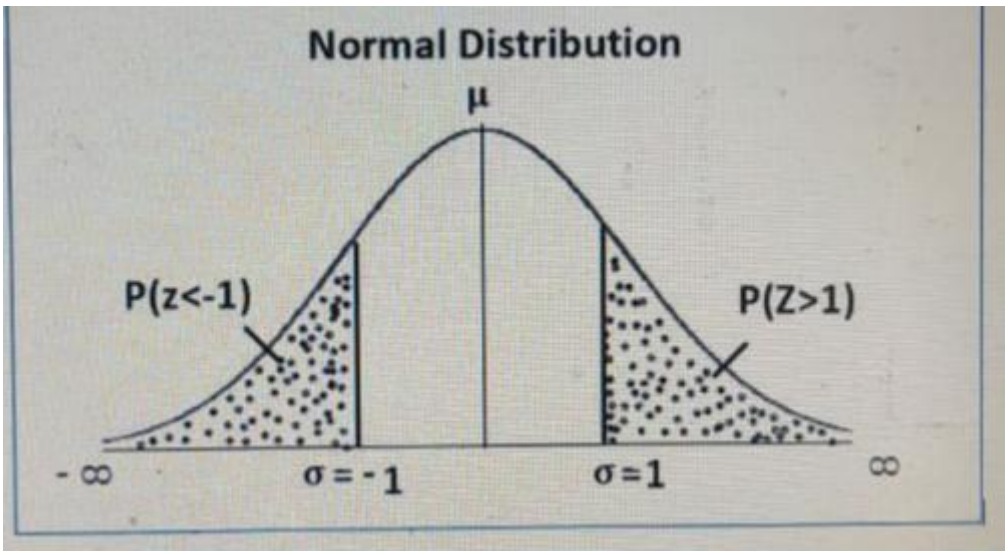
c)From the Z table, $\phi(Z) = 0.15\% = 0.0015$

$$\phi\left(\frac{x-112}{8}\right) = \phi(Z) = 0.0015$$

Therefore, $Z = -2.9$

$$x = 112 - 2.9(8)$$

$$x = 88.8\ mmol/L$$



Therefore, the smallest 0.15% of chemical concentration values are those less than <u>88.8 mmol/L.</u>

Solution for 2.2

Given Population mean μ = 100 pCi/L

Volume of Radon(pCi/L) in 15 samples,

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

Stating the hypotheses:

Null hypothesis (H_0)

The null hypothesis states that a population parameter (such as the mean, the standard deviation, and so on) is equal to a hypothesized value. The null hypothesis is often an initial claim that is based on previous analyses or specialized knowledge.

Alternative Hypothesis (H_α)

The alternative hypothesis states that a population parameter is smaller, greater, or different than the hypothesized value in the null hypothesis. The alternative hypothesis is what we might believe to be true or hope to prove true.

H_0 : μ = 100

H_α : μ ≠ 100

$$\sum x = 105.6 + 90.9 + 91.2 + 96.9 + 96.5 + 91.3 + 101.1 + 105.3 + 107.7 + 102.6 + 98.7 + 92.4 + 93.7 + 104.3 + 103.5$$

$$\sum x = 1481.7$$

Mean(x) = $\frac{\sum x}{N} = \frac{1481.7}{15} = 98.78$

Sample Volume(x)	$(x_i - x)$	$(x_i - x)^2$
105.6	6.82	46.51
90.9	-7.88	62.09
91.2	-7.78	57.45
96.9	-1.88	3.53
96.5	-2.28	5.198
91.3	-7.48	55.95
101.1	2.32	55.38
105.3	6.52	42.51
107.7	8.92	79.56
102.6	3.82	14.59
98.7	-0.08	0.0064

92.4	**-6.38**	**40.70**
93.7	**-5.08**	**25.80**
104.3	**5.52**	**30.47**
103.5	**4.72**	**22.28**
		$\sum(\mathrm{xi} - \bar{\mathrm{x}})^2 = 492.024$

Standard deviation for sample observations is ,

$$\sigma = \sqrt{\frac{\sum(\mathrm{xi}-\bar{\mathrm{x}})^2}{N-1}} = \sqrt{\frac{492.024}{15-1}} = 5.92852$$

Test Statistics is given by

$$t = (\bar{\mathrm{x}} - \mu) / (\sigma/\sqrt{N}) = (97.78 - 100)/(5.92852/\sqrt{15}) = -0.79700$$

Therefore , the statistic test value is,

t = -0.7970

Now for 5% level of Significance ,α = 0.05 and

Degrees of freedom $df = (N-1) = (15-1) = 14$

From the t – distributed table ,required critical value of $t_{0.05}$ is = 2.145

We can see that,

$t < t_{0.05}$

Null hypothesis is accepted at 5% line of significance since $t < t_{0.05}$ i.e there is no evidence that the mean readings differs from 100 pCi/L

Therefore,the population mean reading of Randon volume for the samples does not differ from 100 pCi/L.

Solution for Question 3
Solution for Question 3.1

a. Find a linear regression model to the data:

Let X = number of spare parts ordered

Y = price of the order

We know that, X is an independent variable and Y is a dependent variable. The suitable regression model is:

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \text{.....(1)}$$

To estimate the above equation(1), we use OLS

The two normal equations are:

$$\sum Y = n\beta_0 + \beta_1 \sum X$$

$$\sum XY = \beta_0 \sum X + \beta_1 \sum X^2 \qquad \text{.....(2)}$$

Tabulating:

(X) Number ordered	(Y) Price	XY	X^2
90	120	10,800	8,100
115	106	12,190	13,225
121	95	11,495	14,641
138	70	9,660	19,044
155	65	10,075	24,025
182	58	10,556	33,124
$\sum X = 801$	$\sum Y = 514$	$\sum XY = 64{,}776$	$\sum X^2 = 1{,}12{,}159$

Therefore from the table,

$$n = 6$$

$$\sum X = 801 \sum Y = 514 \sum XY = 64,776 \sum X^2 = 1,12,159$$

Substituting the above values in the equation(2), we get

$$514 = 6\beta_0 + 801\beta_1 \qquad \text{.....(3)}$$

$$64,776 = 801\beta_0 + 1,12,159\beta_1 \qquad \text{.....(4)}$$

Solving these equations we get the values for β_0 and β_1

(Or)

Using the formula,

$$\beta_1 = \left[\frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}\right] \qquad \text{.....(\#)}$$

Substituting the values in the above equation(#), we get

$$\beta_1 = [\frac{64{,}776 - \frac{801*514}{6}}{1{,}12{,}159 - \frac{801^2}{6}}]$$

$$\beta_1 = [\frac{-3,843}{5,225.5}]$$

$$\beta_1 = -0.735$$

$$\beta_0 = \frac{\sum Y}{n} - \beta_1 \frac{\sum X}{n} \qquad \text{.....(5)}$$

Substituting β_1, $\sum X$ and $\sum Y$ in equation(5), we get β_0

$$\beta_0 = \frac{514}{6} + (0.735)\frac{801}{6}$$

$$\beta_0 = 183.847$$

Therefore,

$$\beta_0 = 183.847$$

$$\beta_1 = -0.735$$

Therefore, the linear regression model is:

$$Y = 183.847 - 0.735X$$

b) Number of units that would be ordered if the price were 60

i.e., $Y = 60$

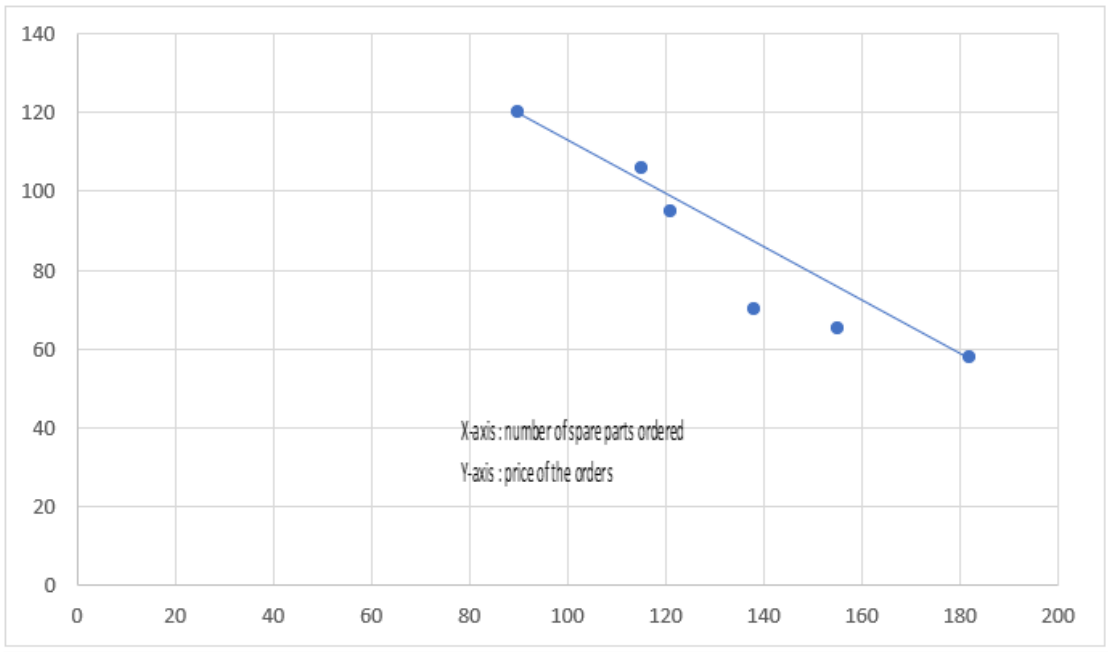
Substituting Y in regression equation, we get

$$60 = 183.847 - 0.735X$$

$$X = 168.499$$

Therefore , 168.499 units would be ordered if the price were 60.

c)Scatter Diagram



Solution for Question 3.2

Let X be the Poisson random variable with the mean equal to $E(X) = \lambda T$.

The probability mass function of X:

$$f(k) = P(X = k) = \frac{(\lambda T)^k e^{-\lambda T}}{k!}, k \in N_0$$

If the mean is 0.08 flaw per 1 square foot, we can set the parameter λ to be:

λ = 0.08 flaw/square foot

and we measure T in units of square foot. Substitute $T = 10$ for the following calculations because one boiler is said to have 10 square foot.

a) $X \sim P_0(\lambda)$ then $P(X = r) = \frac{e^{-\lambda}.\ \lambda^r}{r!}$

Let the boiler contain 10ft of metal surface.

Therefore, λ=0.08/Sq ft

λ=0.08*10 = 0.8 for 10 sq ft of metal surface

$$P(X = 0) = \frac{e^{(-0.8)}.\lambda^0}{0!} = \frac{e^{(-0.8)}.1}{1} = 0.449$$

The probability that there are no surface flaws in a boiler is 0.449.

b) Probability that at least 2 of the 10 boilers has any surface flaws is,

The probability distribution is given by

$P(x) = {}^nC_x\ P^x\ q^{(n-x)}$

x be the no. of surface flaws

P is probability of success of surface flaws,q=1-p is failure .

n=No. of samples =10

q=0.4493 . q=1-p

p=1-0.4493 = 0.5507

$P(x) = {}^{10}C_x\ (0.4493)^{10-x}\ (0.5507)^x$

WE know that

$P(x \geq 2) = 1 - (\ P(x = 0)^{\square} + P(x = 1)^{\square})$

$P(x=0) = {}^{10}C_0\ (0.4493)^{10}\ (0.5507)^0$

$P(x=0) = (0.4493)^{10}$

P(x=0)=0.0034

P(x=1) = $^{10}C_1$ $(0.4493)^9$ $(0.5507)^1$
P(x=1) = 0.00411

$P(x \geq 2)$**)=1-(0.0034+0.00411)**

$$P(x \geq 2) = 0.9955$$

Probability that at least 2 of the 10 boilers has any surface flaws is 0.9955**.**

c)The probability of either 0 or exactly 1 boiler being flawed is (using independence):

Here the total number of boilers sold to the company is 12,
Therefore,
The probability distribution is given by

$P(x) = {}^nC_x\, P^x\, q^{(n-x)}$
x be the no. of surface flaws
P is probability of success of surface flaws
q=1-p is failure .
n=No. of samples =12

q=0.4493 . q=1-p

p=1-0.4493 = 0.5507

$P(x) = {}^{12}C_x\, (0.4493)^{12-x}\, (0.5507)^x$
WE know that
$P(x \leq 1) = (P(x=0) + P(x=1))$
$P(x=0) = {}^{12}C_0\, (0.4493)^{12}\, (0.5507)^0$
$P(x=0) = (0.4493)^{12}$
P(x=0)=0.0000678

$P(x=1) = {}^{12}C_1\, (0.4493)^{11}\, (0.5507)^1$
P(x=1) = 0.0009954

$P(x \leq 1) =$ **0.0000678 + 0.0009954**

$$P(x \leq 1) = 0.00106$$

Therefore, that at most one boiler has any surface flaws in 12 is 0.00106

References

Introduction to Probability, Statistics, and Random Processes by Hossein Pishro-Nik.