RAMAIAH UNIVERSITY OF APPLIED SCIENCES

ASSIGNMENT

Course Code	19CSC301A
Course Name	Probability and Statistics
Programme	B. Tech.
Department	Mathematics and Statistics
Faculty	FMPS

Name of the Student	Subhendu Maji
Reg. No	18ETCS002121
Semester/Year	5^{TH} SEM / 2018 BATCH
Course Leader/s	Dr Subramanyam T

Declaration Sheet			
Student Name	Subhendu Maji		
Reg. No	18ETCS002121		
Programme	B. Tech.	Semester/Year	
Course Code	19CSC301A		
Course Title	Probability and Statistics		
Course Date		to	
Course Leader	Dr Subramanyam T		

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	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.	**18ETCS002121**	Name of the Student	**SUBHENDU MAJI**

Sections		**Marking Scheme**	Marks		
			Max Marks	Marks Scored	CO
Part -A	1.1	Describe the normal distribution	10		1
		Part-A Max Marks	**10**		
Part-B	2.1	Determine the probabilities	05		2
		Determine the expected value and standard deviation	05		3
	2.2	State the hypotheses	02		3
		Test statistic and calculations	05		4
		Interpretation and Conclusion	03		5
		Part-B Max Marks	**20**		
Part-C	3.1	State the model and fit the data	07		5
		Prediction and plot the graph	03		5
	3.2	Determine the probabilities	10		2
		Part-C Max Marks	**20**		
		Total Assignment Marks	**50**		

Question No. 1

Solution to Question No. 1:

1.1 Describe the normal distribution

Normal Distribution

The normal distribution was first discovered by **De – Movire** and **Laplace** as the limiting form of Binomial distribution. Through a historical error it was credited to Gauss who first made reference to it as the distribution of errors in Astronomy. Gauss used the normal curve to describe theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies.

Definition: A continuous random variable X is said to have a **normal distribution** with parameters $mean\ \mu$ and variance σ^2 if its p. d. f. is given by

$$\boldsymbol{f(x)=\frac{1}{\sigma\sqrt{2\pi}}exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}}\quad -\infty<x<\infty$$

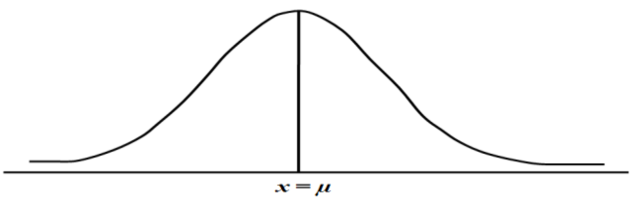
The c. d. f. of X is given by

$$F(x)=P(X\le x)=\int_{-\infty}^{x}f(t)\,dt=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}exp\left\{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right\}dt$$

Notation: $X\sim N(\mu,\sigma^2)$ *Read as* X *follows normal distribution with parameters* μ *and* σ^2.

Note:

1. The graph of $f(x)$ is famous ***bell – shaped*** curve and is symmetric about the line $X=\mu$. The top of the bell is directly above μ. For large values of σ the curve tends to flatten out and for small values of σ it has a sharp peak. The curve of $f(x)$ is given below.



Normal probability curve

2. Whenever the random variable is continuous and the probabilities of it are increasing and then decreasing, in such cases we can think of using normal distribution.

Real life examples:

1) The heights of students.

2) The weights of students.

3) The diameters of bolts manufactured.

4) The lives of electrical bulbs manufactured.

3. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

- Applied to single variable continuous data

  e.g., heights of plants, weights of lambs, lengths of time

- Used to calculate the probability of occurrences less than, more than, between given values

  e.g., “the probability that the plants will be less than 70mm”,

  “the probability that the lambs will be heavier than 70kg”,

  “the probability that the time taken will be between 10 and 12 minutes”

Standard Normal distribution

If $X \sim N(\mu, \sigma^2)$ then $Z = \frac{X-\mu}{\sigma}$ is known as **standard normal distribution** with mean $E(Z) = 0$, with variance $V(Z) = 1$ and we write $Z \sim N(0, 1)$. Its p. d. f. is given by

$$4. \quad \boldsymbol{g(z) = \frac{1}{\sqrt{2\pi}} . exp\left(-\frac{1}{2} z^2\right)}, -\infty < z < \infty$$

and its c. d. f. is given by

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{x} g(t)\,dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} exp\left\{-\frac{1}{2} t^2\right\} dt$$

Characteristics of Normal distribution:

1. All normal curves are bell-shaped with points of inflection at μ±σ.
2. All normal curves are symmetric about the mean μ.
3. The area under an entire normal curve is 1.
4. The height of any normal curve is maximized at x=μ.
5. The shape of any normal curve depends on its mean μ and standard deviation σ.

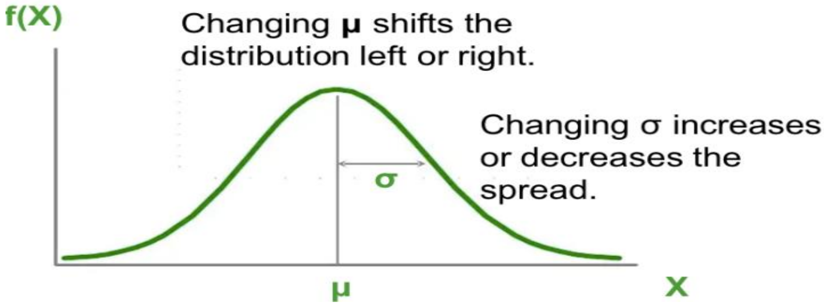
Area Property of Normal Distribution

1. The mean, mode and median are all equal.

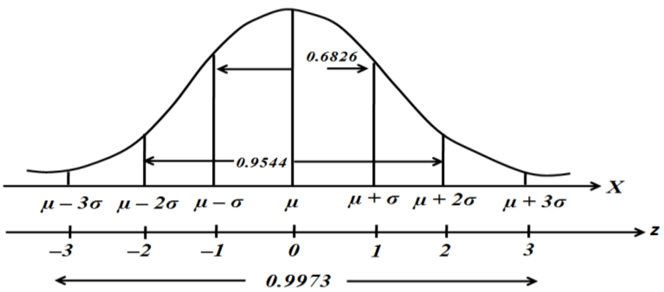
2. The curve is symmetric at the centre (i.e., around the mean, μ).

3. Exactly half of the values are to the left of centre and exactly half the values are to the right.

4. The total area under the curve is 1.



- Symmetric, bell shaped
- Continuous for all values of X between -∞ and ∞ so that each conceivable interval of real numbers has a probability other than zero.
- -∞ ≤ X ≤ ∞ • Two parameters, μ and σ. Note that the normal distribution is actually a family of distributions, since μ and σ determine the shape of the distribution.
- The rule for a normal density function is
- About 2/3 of all cases fall within one standard deviation of the mean, that is $P(\mu - \sigma \leq X \leq \mu + \sigma) = .6826$.
- About 95% of cases lie within 2 standard deviations of the mean, that is $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9544$



Skewness:

Skewness means lack of symmetry. It helps us to study about the shape of the curve which can be drawn with the help of the given data. The distribution is said to be skewed if

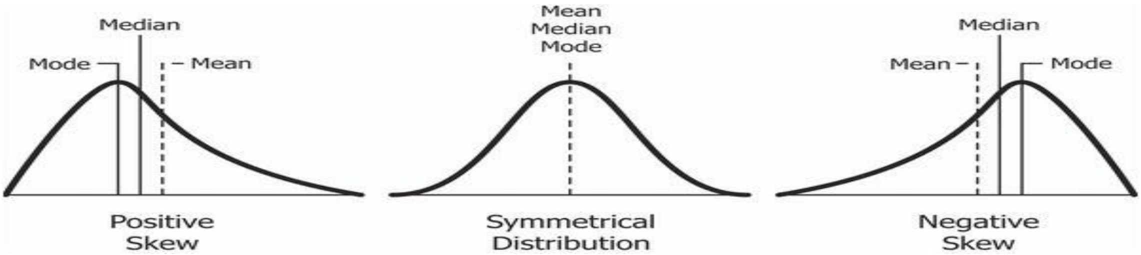
1. Mean, median and Mode fall at different places i.e., Mean ≠ Median ≠ Mode
2. Quartiles are not equidistant from Median and

3. The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

The coefficient of Skewness based on the moments is given by β_1 or γ_1

1. A distribution is said to be symmetric if $\gamma_1 = 0$

   Since $\gamma_1 = \sqrt{\beta_1}$ for γ_1 to be 0 which implies β_1 = 0 and since $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ for β_1 to be 0 which implies $\mu_3 = 0$ i.e. for a symmetrical distribution all odd order moments are zero
2. Positive symmetric if $\gamma_1 > 0$
3. Negative symmetric if $\gamma_1 < 0$



<u>1σ, 2σ and 3σ limits</u>

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

About 68% of the x values lie between –1σ and +1σ of the mean μ (within one standard deviation of the mean).

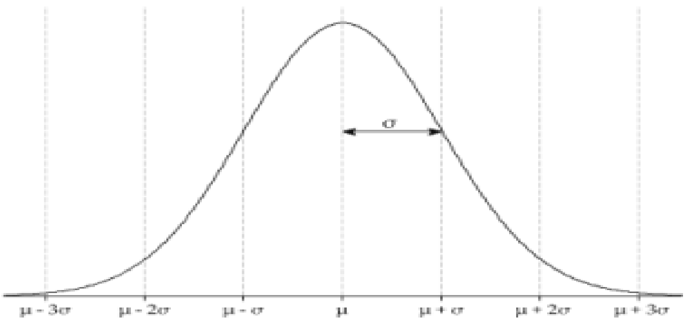
About 95% of the x values lie between –2σ and +2σ of the mean μ (within two standard deviations of the mean).

About 99.7% of the x values lie between –3σ and +3σ of the mean μ(within three standard deviations of the mean). Notice that almost all then 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.



Cumulative distribution function

The cumulative distribution function is the probability that the variable takes a value less than or equal to X

$$F(x) = Pr[X \leq x] = \alpha$$

For a continuous distribution, this can be expressed mathematically as

$$F(x) = \int x - \infty f(\mu) d\mu$$

For a discrete distribution, the cdf can be expressed as

$$F(x) = \sum xi = 0 f(i)$$

The horizontal axis is the allowable domain for given probability function. Since the vertical axis is a probability, it must fall between zero and one. It increases from zero to one as we go from left to right on the horizontal axis.

Probability density function

In probability theory, a probability density function (PDF), or density of a continuous random variable, is a function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample. In other words, while the absolute likelihood for a continuous random variable to take on any particular value is 0 (since there are an infinite set of possible values to begin with), the value of the PDF at two different samples can be used to infer, in any particular draw of the random variable, how much more likely it is that the random variable would equal one sample compared to the other sample.

The terms "probability distribution function" and "probability function" have also sometimes been used to denote the probability density function. However, this use is not standard among probabilists and statisticians. In other sources, "probability distribution function" may be used when the probability distribution is defined as a function over general sets of values or it may refer to the cumulative distribution function, or it may be a probability mass function (PMF) rather than the density. "Density function" itself is also used for the probability mass function, leading to further confusion. In general, though, the PMF is used in the context of discrete random variables (random variables that take values on a countable set), while the PDF is used in the context of continuous random variables.

<u>**Question No. 2**</u>

<u>**Solution to Question No. 2:**</u>

2.1Determine the probabilities

Given,

A normal distribution with,

Mean, $\mu = 112$

Standard Deviation, $\sigma = 8$

a. **What is the probability that chemical concentration equals 113? Is less than 105? Is at most 105?**

Let X be a random variable which denotes the chemical concentration $(mmol / L)$

Given the distribution is a normal distribution: $\boldsymbol{X \sim N\ (112, 8)}$

Now,

So as X is a continuous random variable the probability at a fixed point is 0.

a. Therefore, probability that chemical concentration equals 113 is

$$\boldsymbol{P\ (X = 113)\ =\ 0}$$

b. Probability that chemical concentration less than 105 is

$Z = x - \mu / \sigma = 105 - 112/8 = -0.875$

$P(X < 105) = P(Z < -8.875) = 0.190786(table)$

or

$$\mathrm{P(X < 105) = P(}\frac{X - \mu}{\sigma} < \frac{105 - \mu}{\sigma})$$

$$= \mathrm{P}\left(\mathrm{Z} < \frac{105 - 112}{5}\right)$$

$$= \mathrm{P(Z < -0.875)}$$

$$\boldsymbol{P(X < 105)\ =\ 0.190786}$$

c. Probability that chemical concentration less than or equal to 105 is

$$P(X <= 105)\ =\ P((X <\ 105)\ +\ P(X = 105)$$

$$=\ 0.19215 + 0$$

$$\boldsymbol{P(X <= 105)\ =\ 0.190786}$$

b. What is the probability that chemical concentration differs from mean by more than 1 standard deviation? Does this probability depend on the values of μ and σ?

$$|x - \mu| > \sigma$$

$X > \mu + \sigma$

OR, $X < \mu - \sigma$

$$P\ (X > \mu + \sigma)\ or\ P\ (X < \mu - \sigma)$$

$$= P\ (\frac{X - \mu}{\sigma} > 1) + P\ (\frac{X - \mu}{\sigma} - 1)$$

$$Z = x - \frac{\mu}{\sigma}$$

$$= P(Z > 1) + P(Z < -1)$$

$$= [1 - P(Z \leq 1) + P(Z < -1)]$$

$$= [1 - 0.8413] + 0.1587\ (table)$$

$$= \mathbf{0.3174}$$

Therefore, the probability that chemical concentration differs from mean by more than 1 standard deviation is $\mathbf{0.3174}$.

No, this probability is not dependent on the values of μ and σ.

c. How would you characterize the most extreme 0.15% of chemical concentration values?

The most extreme 0.15% of chemical concentration values are the lowest 0.075% and the highest 0.075% of the concentration values, because the normal distribution is symmetric about the mean, i.e.

Let $c1$ and $c2$ be two points such that, in $\boldsymbol{X \sim ND(112, 8)}$

$P(x > c1) = 0.00075$

And, $P(x < c2) = 0.00075$

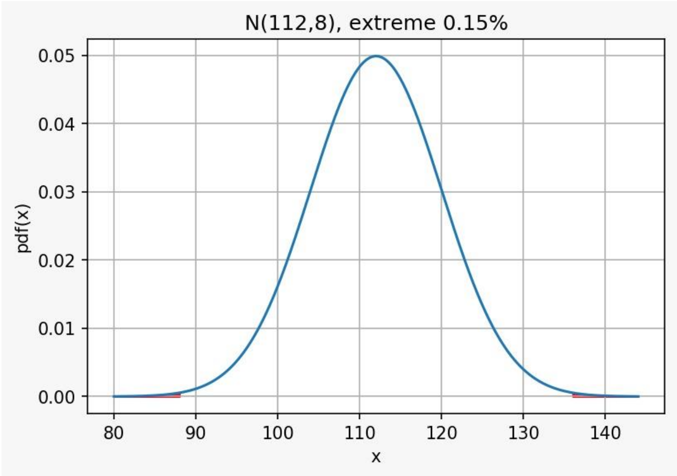


Figure 1 N(112,8), extreme 0.15%

In the standard normal distribution,

We know,

$$\text{If, } P(z > z1) = 0.00075$$

$$P(z \le z1) = 0.99925$$

$$P(z < z2) = 0.00075$$

Then, $z1 = 3.175$

So, if $z1 = 3.175$ then $z2 = -3.175$

We know,

$$Z1 = \frac{c1-\mu}{\sigma}$$

$$3.175 = \frac{c1-112}{8}$$

$$C1 = 3.175 \times 8 + 112$$

$$\boldsymbol{c1 = 137.36} \textbf{ mmol/L}$$

$$Z2 = \frac{c2-\mu}{\sigma}$$

$$-3.175 = \frac{c2-112}{8}$$

$$C2 = -3.175 \times 8 + 112$$

$$\boldsymbol{C2 = 86.64\, mmol/L}$$

Hence, the most extreme 0.15% of chemical concentration values are all concentrations greater than 137.4 $mmol/L$ and all concentrations less than 86.6 $mmol/L$.

2.2 State the hypotheses

Given,

$n = 15$

$$mean, \mu_0 = 100$$

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

5% level of significance, i.e.

$\alpha = 0.05$

$\mu = \dfrac{1481.7}{15} = 98.78$

2.2.1 State the hypotheses

- **Null Hypotheses**

$$H_0: \mu = \mu_0$$

- **Alternative Hypotheses**

$$H_a: \mu \neq \mu_0$$

We know, $\alpha = 0.05$

degree of freedom $n - 1 = 14$

Rejection rule:

$t \leq -t(\frac{\alpha}{2}, \mathrm{n} - 1)$

$t \geq t(\frac{\alpha}{2}, \mathrm{n} - 1)$

From t-table:

$t\left(\frac{\alpha}{2}, \mathrm{n} - 1\right) = \mathbf{2.145}$

Therefore, the decision rule is:

$$\boldsymbol{If\ t\ is\ less\ than\ -2.145, or\ greater\ than\ 2.145, reject\ the\ null\ hypothesis}\ H_0.$$

2.2.2 Test statistic and calculations

X	$X-\bar{x}$	$(X-\bar{x})^2$
105.6	`6.82	46.5124
90.9	−7.88	62.0944
91.2	−7.58	57.4564
96.9	−1.88	3.5344
96.5	−2.28	5.1984
91.3	−7.48	55.9504
101.1	2.32	5.3824
105.3	6.52	42.5104
107.7	8.92	79.5664
102.6	3.82	14.5924
98.7	−0.08	0.0064
92.4	−6.38	40.7044
93.7	−5.08	25.8064
104.3	5.52	30.4704
103.5	4.72	22.2784
1481.7		**492.064**

We know,

$$t = \frac{\mu - \mu 0}{s/\sqrt{n}}$$

where s is the standard deviation using Bessel's correction.

$$s = \sqrt{\frac{(\ \mathrm{x} - \bar{\mathrm{x}})^2}{n-1}} = \sqrt{\frac{492.064}{15-1}} = \mathbf{5.92852}$$

$$\Rightarrow t = \frac{\mu - \mu 0}{s/\sqrt{n}} = \frac{98.78 - 100}{5.92852/\sqrt{15}} = \mathbf{-0.7970}$$

2.2.3 Interpretation and Conclusion

The value of t, is neither less than -2.145, nor greater than 2.145. Therefore, the null Hypotheses is true

This data does not suggest that the population mean reading under these conditions differ from 100. Or the mean reading of the sample of 15 radon detectors is 100.

Question No. 3

Solution to Question No. 3:

3.1.1 State the model and fit the data

$\boldsymbol{Number\ ordered\ (Y)}$	$\boldsymbol{Price\ (X)}$	$\boldsymbol{X2}$	$\boldsymbol{XY}$
90	120	14400	10800
115	106	11236	12190
121	95	9025	11495
138	70	4900	9660
155	65	4225	10075
182	58	3364	10556
$\boldsymbol{\Sigma\, x\ =\ 801}$	$\boldsymbol{\Sigma\, y\ =\ 514}$	$\boldsymbol{\Sigma\, x2\ =\ 47150}$	$\boldsymbol{\Sigma\, xy\ =\ 64776}$

$$n = 6$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon$$

$Here,$

$Y_i \rightarrow dependent\ variable$

$x_i \rightarrow independent\ variable$ 90

β_0 and β_1 $\rightarrow regressive\ coefficient, i.e., intercept\ and\ slope, respectively$

$\varepsilon \rightarrow residual\ or\ error\ term.$

$$\Sigma\, xy = \beta_0 \sum x + \beta_1 \Sigma\, x^2$$

$$\sum y = n\beta_0 + \beta_1 \sum x$$

$$6\beta_0 + 514\ \beta_1 = 801$$

$$514\beta_0 + 47150\,\beta_1 = 64776$$

Solving the above simultaneous equations of 1 and 2 we get answers as $\boldsymbol{\beta_1 = -1.23278}$ and $\boldsymbol{\beta_0 = 239.10853}$

Or use the below formulae

$$\beta_1 = \frac{\Sigma xy - \Sigma x . \Sigma y}{\Sigma(x^2) - \Sigma(x^2)/n}$$

$$\beta_1 = \frac{64776 - (801 * 514)}{(47150) - (\frac{47150}{6})}$$

$$\beta_1 = -1.23278$$

From $Y_i = \beta_0 + \beta_1 x_i + \varepsilon$

We get $\beta_0 = Y_i - \beta_1 x_i + \varepsilon$

Substituting the values of $x_i, \beta_1, Y_i, \varepsilon$

$$\beta_0 = \frac{\Sigma y}{n} - \beta_1 * (\frac{\Sigma x}{n})$$

$$\beta_0 = \frac{801}{6} - (-1.23278) * (\frac{514}{6})$$

We get $\beta_0 = 133.5 - ((-1.23278) \times 85.6667$

$= 239.10853$

Regressive coefficients $\boldsymbol{\beta_1 = -1.23278}$ **and** $\boldsymbol{\beta_0 = 239.10853}$

Regression equation $\boldsymbol{Y_{i=} - 1.23278 * X + 239.10853}$

3.1.2 Prediction and plot the graph

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

```
import scipy.stats as sc
import numpy as np
import matplotlib.pyplot as plt

[2] y = np.array([90,115,121,138,155,182])
    x = np.array([120,106,95,70,65,58])

[3] res = sc.linregress(x,y)
    print("the slope and intercepts are: ",res.slope,res.intercept)

    the slope and intercepts are:  -1.2327844311377245 239.10853293413174

[4] res.slope*60+res.intercept

    165.14146706586826
```

Figure 2 python implementation of linear regression

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

$$Y_{i=} - 1.23278X + 239.10853$$

$$Y_{i=} - 1.23278 * 60 + 239.10853$$

$$Y_{i=} - 73.9668 + 239.10853$$

$$\boldsymbol{Y_{i=} 165.14173}$$

Therefore, 165.14173 units would be ordered if the price were 60

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

```
[5] plt.plot(x, y, 'o', label='original data')
    plt.plot(x, res.intercept + res.slope*x, 'r', label='fitted line')
    plt.legend()
    plt.show()
```

Figure 3 python code for plotting scatter graph

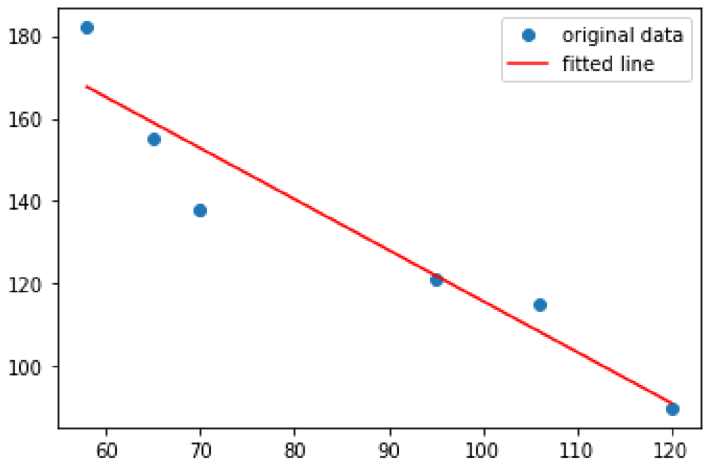


Figure 4 scatter diagram

3.2.1 Determine the probabilities

a. What is the probability that there are no surface flaws in a boiler?

$Mean = \lambda = 0.08$

For 10 $sq\ feet$ panel $\lambda = 10 * 0.08$

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

Let the boiler contain 10ft of metal surface.

Therefore, $\lambda = 0.08/ft$

$$\Rightarrow \lambda = 0.08 * 10 = 0.8 \; for \; 10ft \; of \; metal \; surface$$

$$P(X = 0) = \frac{e^{(-0.8)} . \lambda^0}{0!} = \frac{e^{(-0.8)} . 1}{1} = 0.4493$$

$$\boldsymbol{P(X = 0) = 0.449328}$$

The probability that there are no surface flaws in a boiler is 0.449328

b. If 10 boilers are sold to a company, what is the probability that at least two of the 10 boilers have any surface flaws?

let Y be number of boilers with surface flaws

Y is in binomial distribution with $n = 10$

$$P = 1 - 0.4493$$

$$= 0.5507$$

$$\Rightarrow \boldsymbol{Y \sim BD(10, 0.5507)}$$

$$\boldsymbol{P(at\ least\ 2\ of\ the\ 10\ boilers\ have\ surface\ flaws)} = P(X \geq 2)$$

$$= 1 - P(X = 0) - P(X = 1)$$

$$= 1 - C_0^{10} p^0 q^{10-0} + C_0^{10} p^1 q^{10-1}$$

$$= 1 - C_0^{10} (0.5507)^0 (0.4493)^{10} + C_0^{10} (0.5507)^1 (0.4493)^9$$

$$= 1 - [3.35247 \times 10^{-4} + 4.109 \times 10^{-3}]$$

$$= 1 - 4.4442 \times 10^{-3} = 0.9956$$

$$\boldsymbol{P(X >= 2) = 0.99555556 = 0.9956}$$

c. If 12 boilers are sold to a company, what is the probability that at most one boiler has any surface flaws?

$n = 12$

Probability that at most 1 boiler have surface flaws is, $P(X <= 1)$

$$P(X <= 1) = P(0) + P(1)$$

Probability that at most one boiler has any surface flaws will be $P\ (X <= 1)$

$$P(X \leq 1) = C_0^{12} p^0 q^{12-0} + C_1^{12} p^1 q^{12-1}$$

$$= C_0^{12} (0.5507)^0\ (0.4493)^{10} +\ C_0^{12} (0.55007)^1\ (0.4493)^{11}$$

$$= 6.7676 \times 10^{-5} + 9.9539 \times 10^{-4}$$

$$= 1.0630 \times 10^{-3}$$

$$\boldsymbol{P(X <= 1)\ =\ 1.0630 * 10\hat{}-3}$$

the probability that at most one boiler has any surface flaws is 1.0630×10^{-3}

Bibliography

1. https://en.wikipedia.org/wiki/Normal_distribution#Standard_normal_distribution
2. 2. https://www.slader.com/discussion/question/the-number-of-surface-flaws-in-plastic-panels-used/
3. 3. https://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm
4. 4. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.t.html