Probability and Statistics- TAMS11 LAB REPORT

(nikku485)

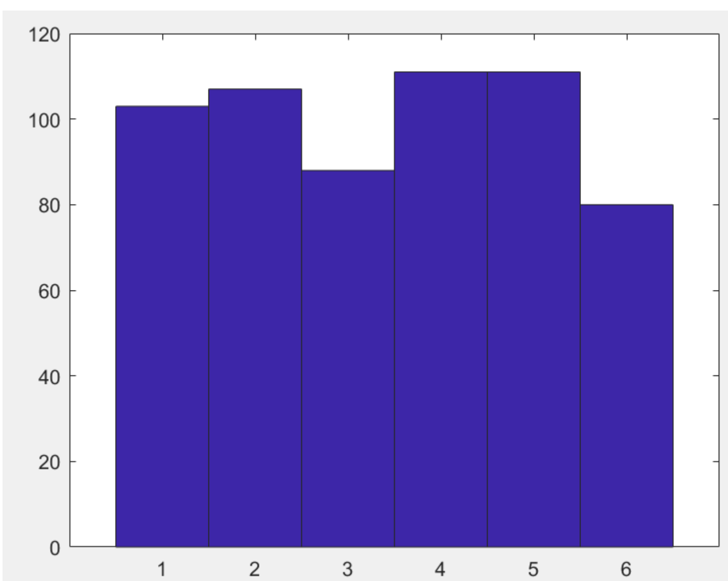
Problem 1

1. Here the function "randi(6,n,1)" does is that it creates a matrix of n*1 which takes random values in the range of 1 to 6. Hence we can see that the function "throw(101:115)" prints the random values from row 101 to 115 in the matrix generated by randi function.

2. When printing the frequency of values from 1:6. We obtain the below output as shown below with the frequencies.

```
Frequency of : 1 is 103
Frequency of : 2 is 107
Frequency of : 3 is 88
Frequency of : 4 is 111
Frequency of : 5 is 111
Frequency of : 6 is 80
```

3. When plotting the histogram, we obtain the below figure with numbers against the corresponding frequencies.



4. When we run the code we get the values of mean and standard deviation as shown below:

```
sample_mean =

    3.4333


sample_standard_deviation =

    1.6754
```

Also, we are computing the values of mean by calculation we get the sum of all the values and divide it by the total number of values 600.

Mean = sum of all the values/ total number of observations
= 2060 / 600
= 3.4333

We can see that the mean value computed using the function and manually calculated are exactly the same.

Standard deviation = $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$

When calculating manually we get the following value for standard deviation.

```
stddev =

    1.6740
```

We can see that there is a slight difference in the value computed using the function and the value we calculated manually.

Problem – 2

CI and HT for the difference of two population means

As specified in the question, we create 10 observations from N(22, 2^2) and 12 observations from N(16, 2^2), using normrnd function. And we use the [H,P,CI,STATS] function for finding the 95% confidence interval of $\mu1 - \mu2$.
When we find the Confidence Intervals by hand we came to know that, the CI intervals coincides when we compute by hand with very small variations as shown below,

Values returned from the function is shown first and the manually computed values shown second.

```
CI =

    4.1958
    7.5224


STATS =

  struct with fields:

    tstat: 7.3480
       df: 20
       sd: 1.8623


c1 =

    4.2905


c2 =

    7.4277
```

ii) When doing the Hypothesis test we get the value H=1 from the function [H,P,CI,STATS] , which means that “indicates that the null hypothesis can be rejected at the 5% level”.

In-order to calculate the Critical region we use the below equation,

$$\begin{cases} H_0: & \mu_X - \mu_Y = c_0, \\ H_a: & \mu_X - \mu_Y \neq c_0, \end{cases}$$

$$C = (-\infty, \ -z_{\alpha/2}) \cup (z_{\alpha/2}, \ \infty)$$

We know the value of α as = .05 and from the Z table we get the critical region as ;
C = $(-\infty$, -1.96) U (1.96, ∞)

When we do the hypothesis testing by hand we need to find $\mu1 - \mu2$. For that we use the below formulas case , where the variances are known.

$$\begin{cases} H_0: & \mu_X - \mu_Y = c_0, \\ H_a: & \mu_X - \mu_Y \neq c_0, \end{cases} \qquad TS = \frac{(\bar{x}-\bar{y})-c_0}{\sqrt{\frac{\sigma_X^2}{n_1}+\frac{\sigma_Y^2}{n_2}}}.$$

We need to find $\mu1 - \mu2$ and so C0 becomes zero. And we compute the TS using the above equation and we get values for test hypothesis which are always greater than 1.96 , which means that the TS value is contained in the Critical region(1.96 to ∞) and hence we can reject our null hypothesis ($\mu1 = \mu2$). Hence we reach to the same conclusion by using both the function and by hand.

Problem – 3

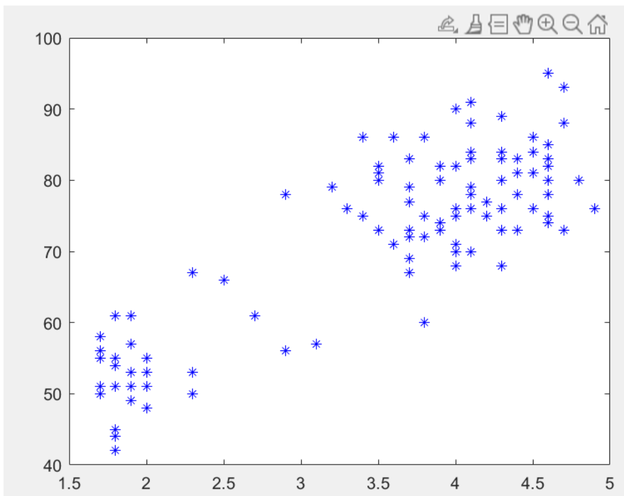
CI using normal approximations

After creating the 95% confidence intervals, when we compute the missing value, which is the number of intervals not containing the p value, we can see that the value is missing is not near to 50 as we get values which are far away from 50. The reason for getting the values which are far way may be because the number of times we are repeating the experiment is small(in our case 16) hence the we get numbers which are not contained in the p value.

We repeat the above code replacing the value of n as 80 that is, Bin(80, 0.3) , now you can see that the value of missing is nearer to 50 since the number of times the experiment is carried out increases which increases in-turn decreases the sample variance than the first case. Hence you get the value of missing nearer to 50.

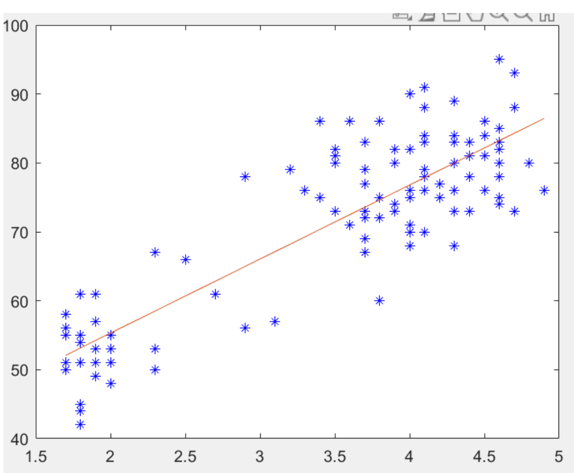
Problem – 4
Simple linear regression

i.) We are plotting the y against x and we obtain the below graph as the output.



We are able to find the correlation value as " 0.8584 " which suggest that there is a linear relation between both x and y since the correlation value that we obtained is near to 1.

ii.) Now we do a full analysis of linear regression using the regstats function and we obtain the below regression line by using the attached code.



iii.)

We have already computed the regression line as shown in the above figure using the code and see how the estimated line fits the points.

We have got the below values from the Regstat function.

```
betahat =

   33.8282
   10.7410


se =

    2.2618
    0.6263


t =

   14.9562
   17.1489


s2 =

   44.6573
```

We have got the values of standard errors sβˆ0 and sβˆ1 as 2.2618 and 0.6263 from the above figure.

iv.) We have the a significance level α = 0.01 and we test the hypothesis ,

H0 : β1 = 0 against H1 : β1 != 0

Below is shows the intervals we need to choose from the t table and we came to know the value of C as (−∞, −2.58) ∪ (2.58,∞).

$$\begin{cases} H_0: & \beta_i = c_0, \\ H_a: & \beta_i \neq c_0, \end{cases}$$

$$C = (-\infty,\ -t_{\alpha/2}(n-2)) \\ \cup\,(t_{\alpha/2}(n-2),\ \infty).$$

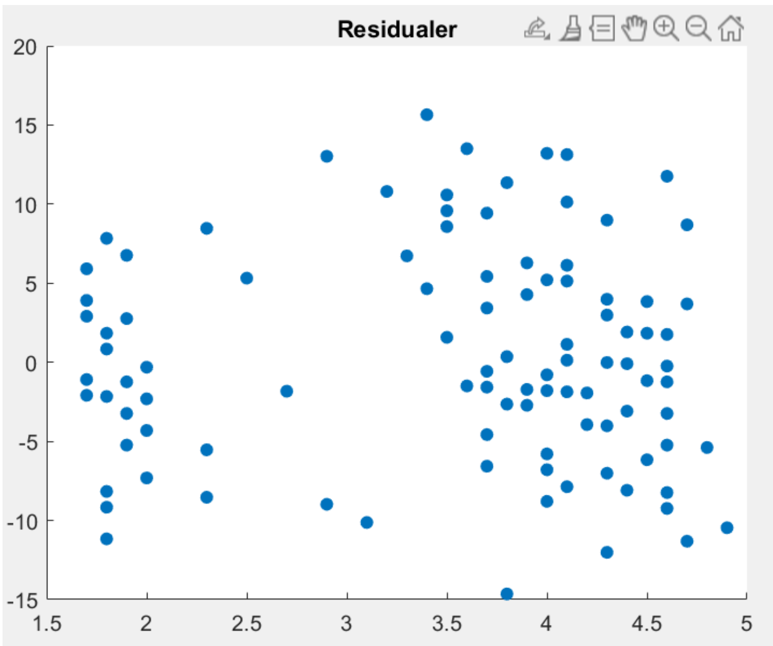
Since you have the value in the range 2.58 to infinity from the significance level from the Z table.

When we calculated the Test hypothesis from the below formula,

$$TS = \frac{\hat{\beta}_i - c_0}{s_{\hat{\beta}_i}},$$

And we get the output as : 17.1489 , Thus we can see that the test hypothesis value is contained in the critical region and hence we have enough evidence to reject the null hypothesis(H0).

v.) We plot the residuals to know whether the error terms $\varepsilon j \sim N(0, \sigma 2)$. And we obtain the below graph,



From the above graph, there is no pattern in the plot of residuals so as to say that the error terms are not normal.

Problem – 5

Logistic regression

i.) We use MATLAB to find the estimated logit functions and we get the values of the Estimated Coefficients β̂0 and β̂1 as 16.328 and -0.25083 respectively. We get the following output while running the function.

```
Generalized linear regression model:
    logit(y) ~ 1 + x1
    Distribution = Binomial

Estimated Coefficients:
                   Estimate      SE        tStat      pValue
                   ________    _______    ______    ________

    (Intercept)      16.328     6.8229    2.3931    0.016708
    x1             -0.25083    0.10029    -2.501    0.012385


32 observations, 30 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 21.8, p-value = 3.01e-06
```

ii.) We test the hypothesis with significance level α = 0.05, with
H0 : β1 = 0 & Ha : β1 != 0

Plugging in the values in the below equation for test hypothesis we get,

$$TS = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}}$$

= -0.25083 / 0.10029
= -2.5010

We thus calculate the critical region using the formula,

$$C = (-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, +\infty).$$

And from the z table with α = 0.05, we get the values of critical region as,

C = (−∞, −1.96) ∪ (1.96,∞)

We can see that the TS lies in the critical region and hence we can reject the Null hypothesis which is β1 = 0.

iii.) We need to find the probability when the value of x= 65. We find the probability by using the following equation,

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}.$$

When we input the values we obtain the probability value as 0.50601. And we can see that the value is higher than 0.5 and we classify Y(65)= 1 as failure. Thus we conclude that, when the temperature is 65 then the launch will fail.