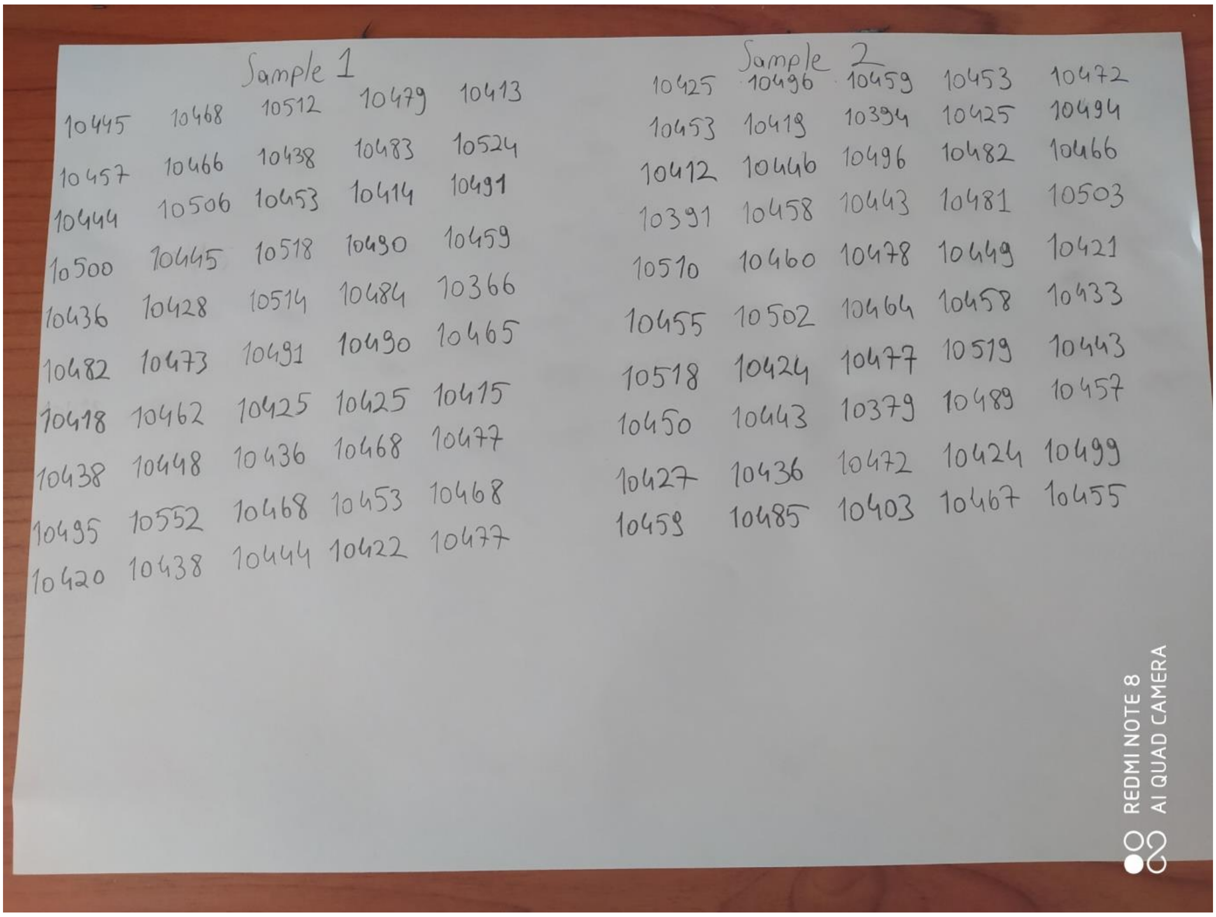
Engineering Statistics Project

Hakan AKTAŞ -18065037
Group : 2

In this project we will analyze the two variable groups that given us.

We don't know what given variables represent.So i assumed that these variables are belong to two factory's weekly produced computer numbers.

Here are the changed values in appropriate way by using my school number.



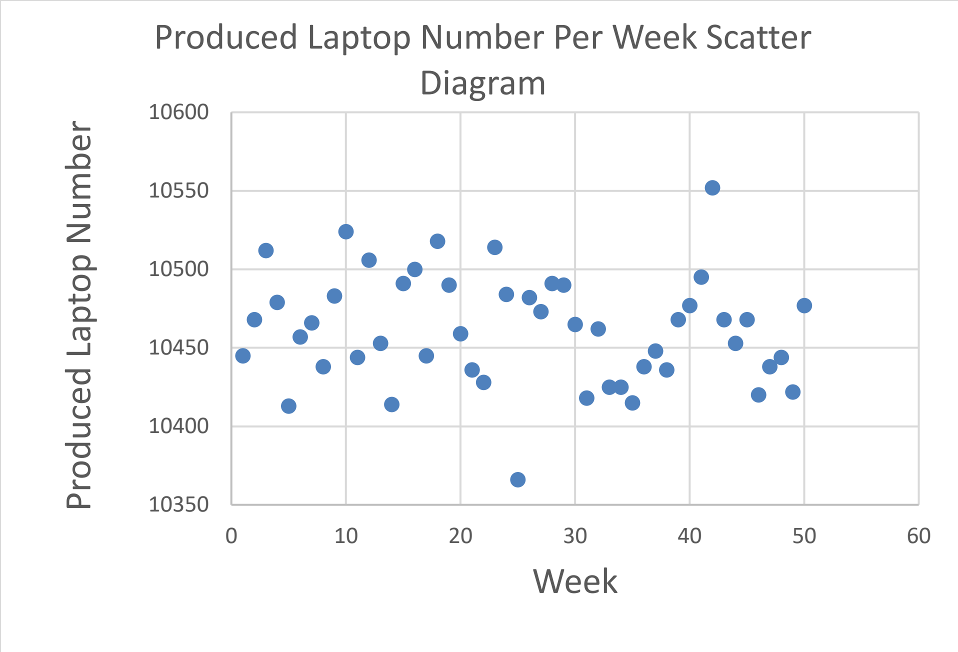
Sample 1

10445	10468	10512	10479	10413
10457	10466	10438	10483	10524
10444	10506	10453	10414	10491
10500	10445	10518	10490	10459
10436	10428	10514	10484	10366
10482	10473	10491	10490	10465
10418	10462	10425	10425	10415
10438	10448	10436	10468	10477
10495	10552	10468	10453	10468
10420	10438	10444	10422	10477

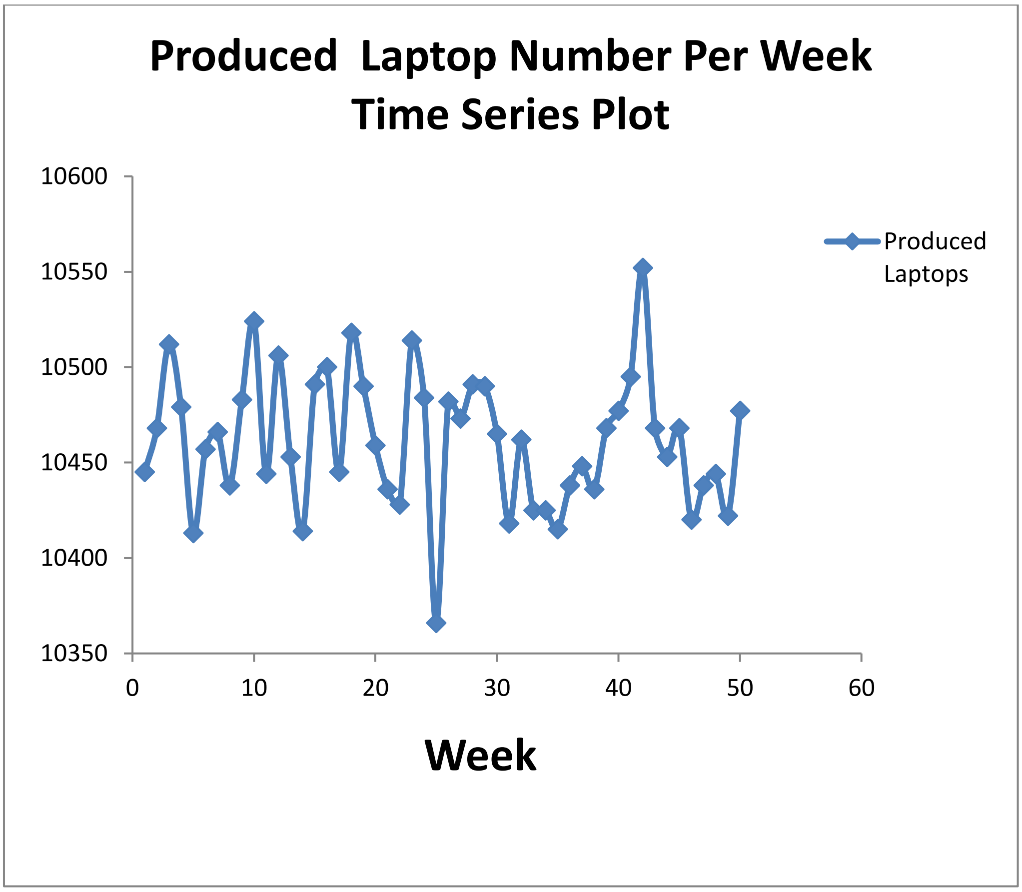
Sample 2

10425	10496	10459	10453	10472
10453	10419	10394	10425	10494
10412	10446	10496	10482	10466
10391	10458	10443	10481	10503
10510	10460	10478	10449	10421
10455	10502	10464	10458	10433
10518	10424	10477	10519	10443
10450	10443	10379	10489	10457
10427	10436	10472	10424	10499
10459	10485	10403	10467	10455

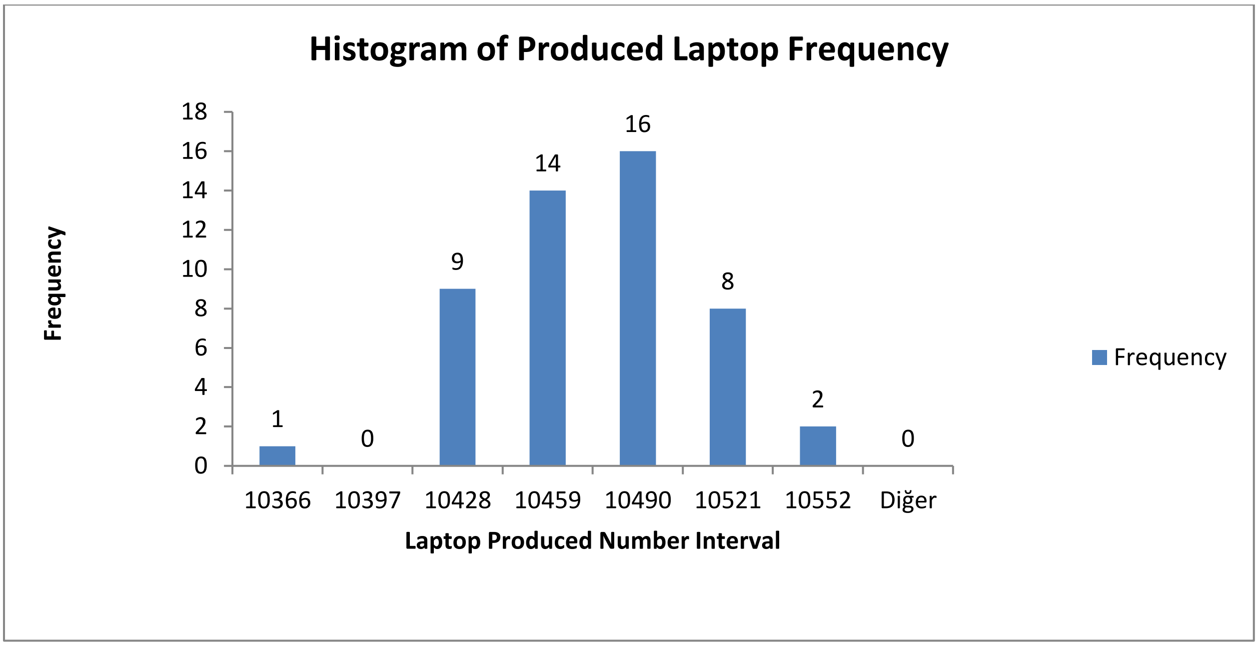
So let’s have a look at the graphics of Sample 1



By using these two graphs it's safe to say that first factory's produced laptop numbers per week are ascending around 10350 and 10550.

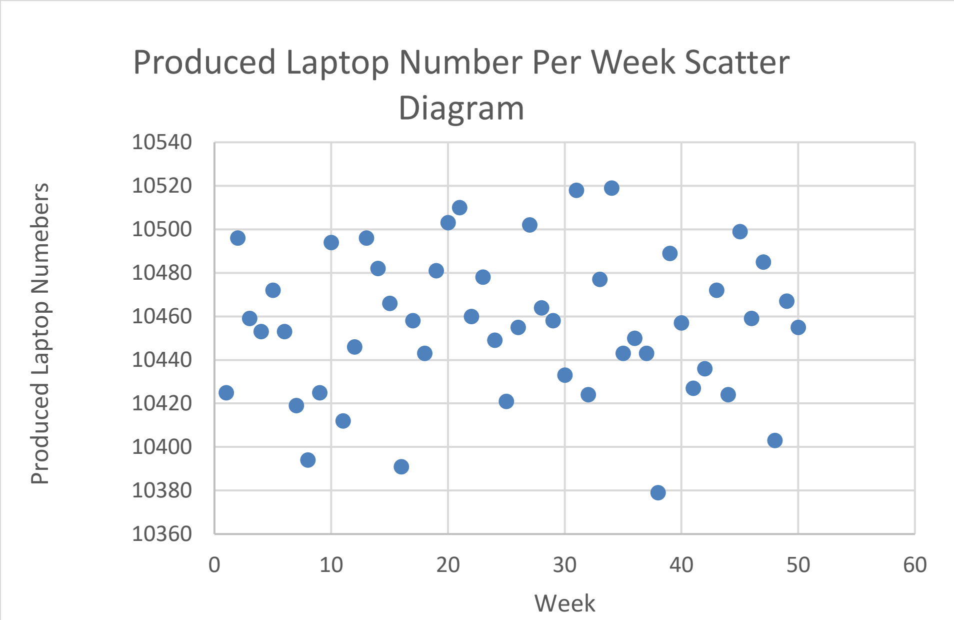


And between the weeks 20-30 we get the lowest producted laptop value. (it's 25th week and number is 10366) , at the beginning of 40's we get the maximum production.(it's 42nd week and number is 10552).

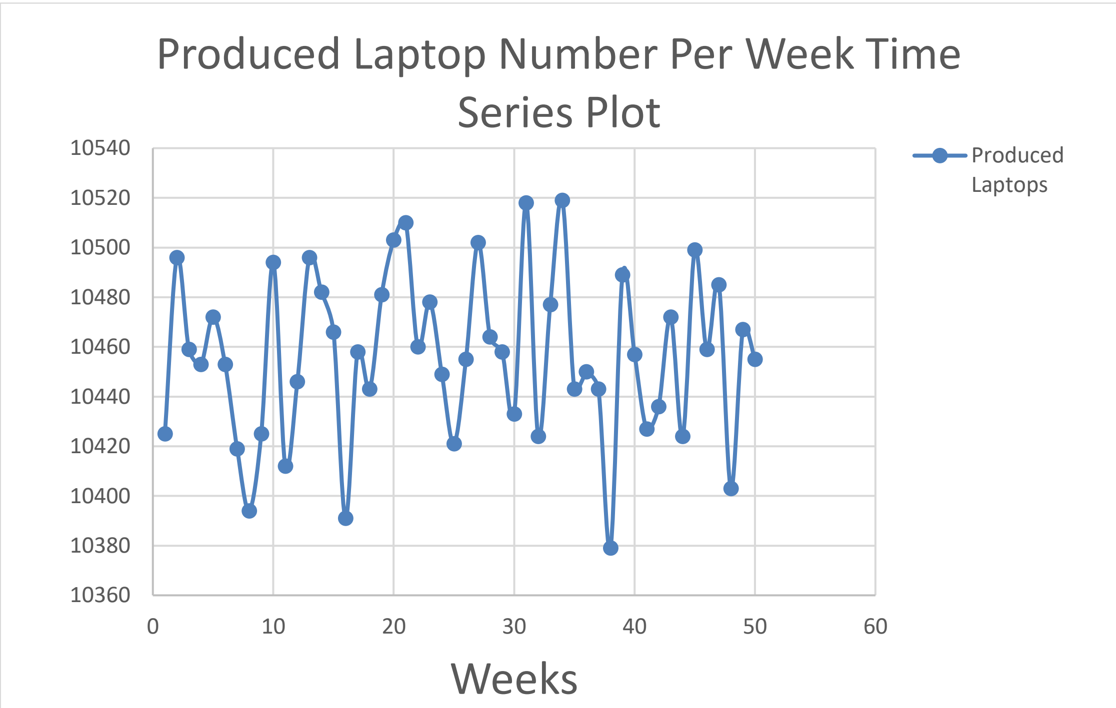


Most of the variables are being in 10459-10490 (16) and there are no any variables between 10366-10397 also we don't have any variables bigger than 10552 and lower than 10366.(There is no certain proof that we don't have a value which is lover than 10366 but if we add all frequencies, their sum equals to 50.Thus that means we don't have any variables which are beyond our limits.)

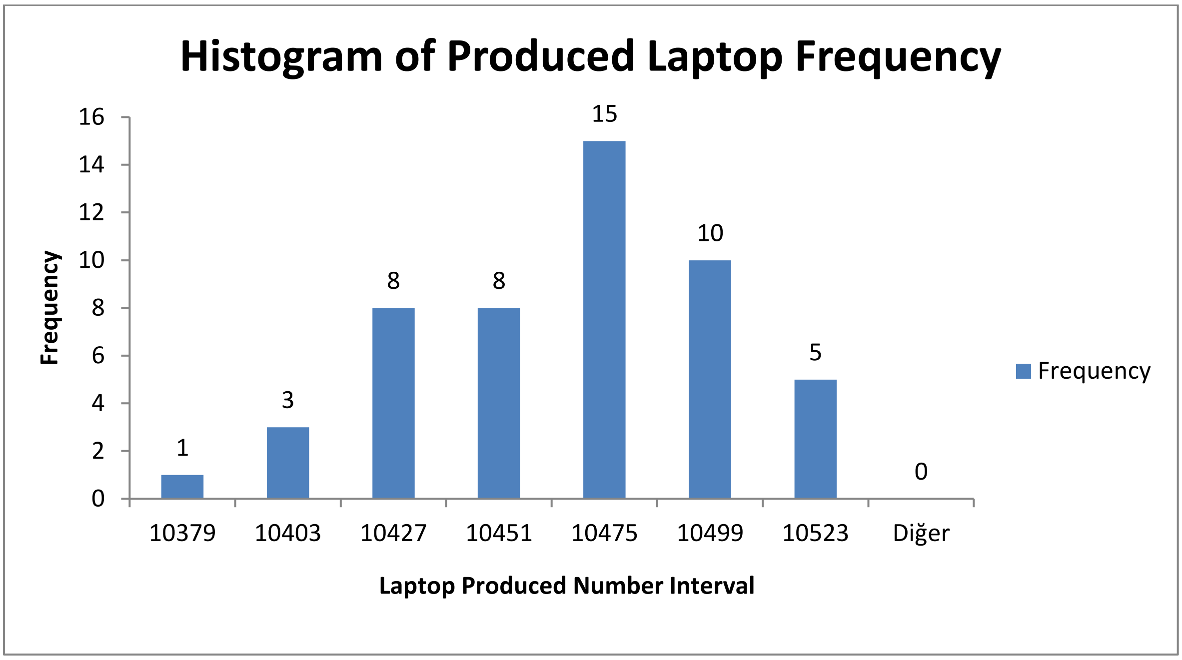
For the Sample 2



By using these two graphs it’s safe to say that first factory’s produced laptop numbers per week are ascending around 10360 and 10520.

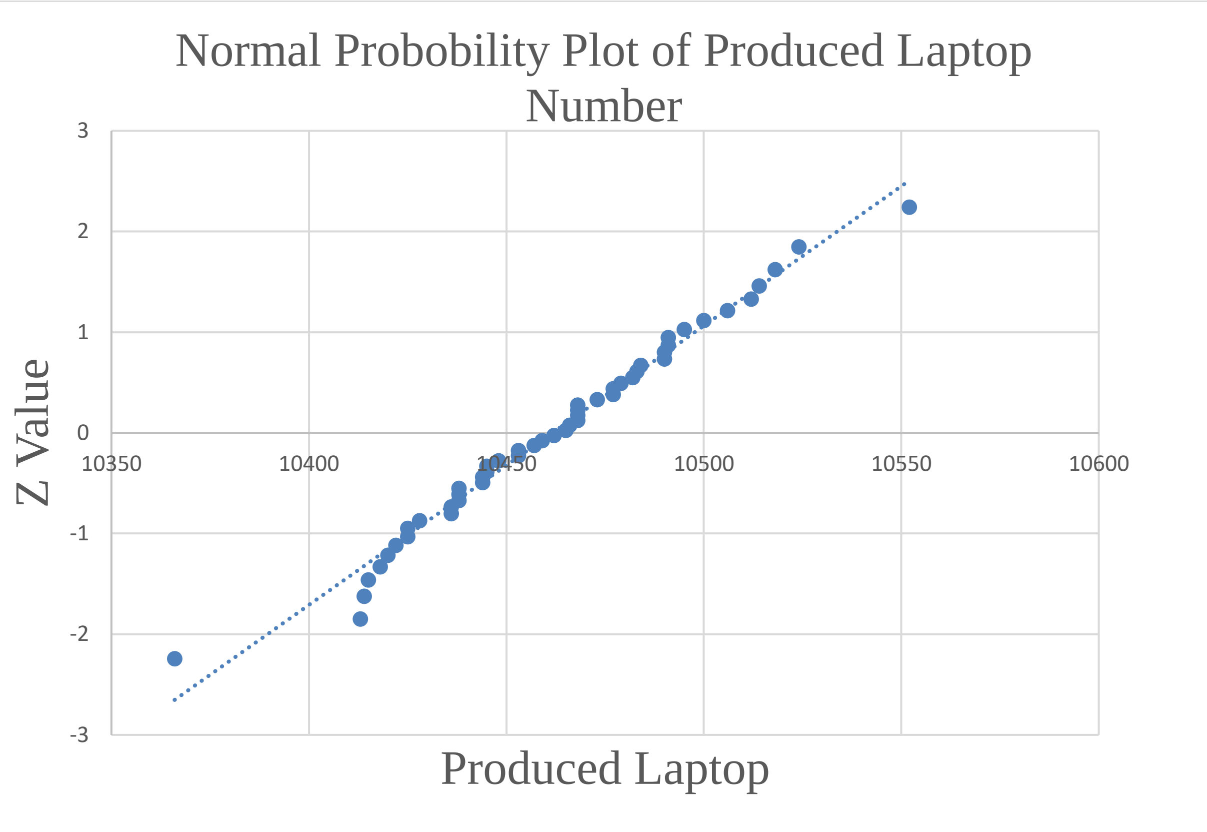


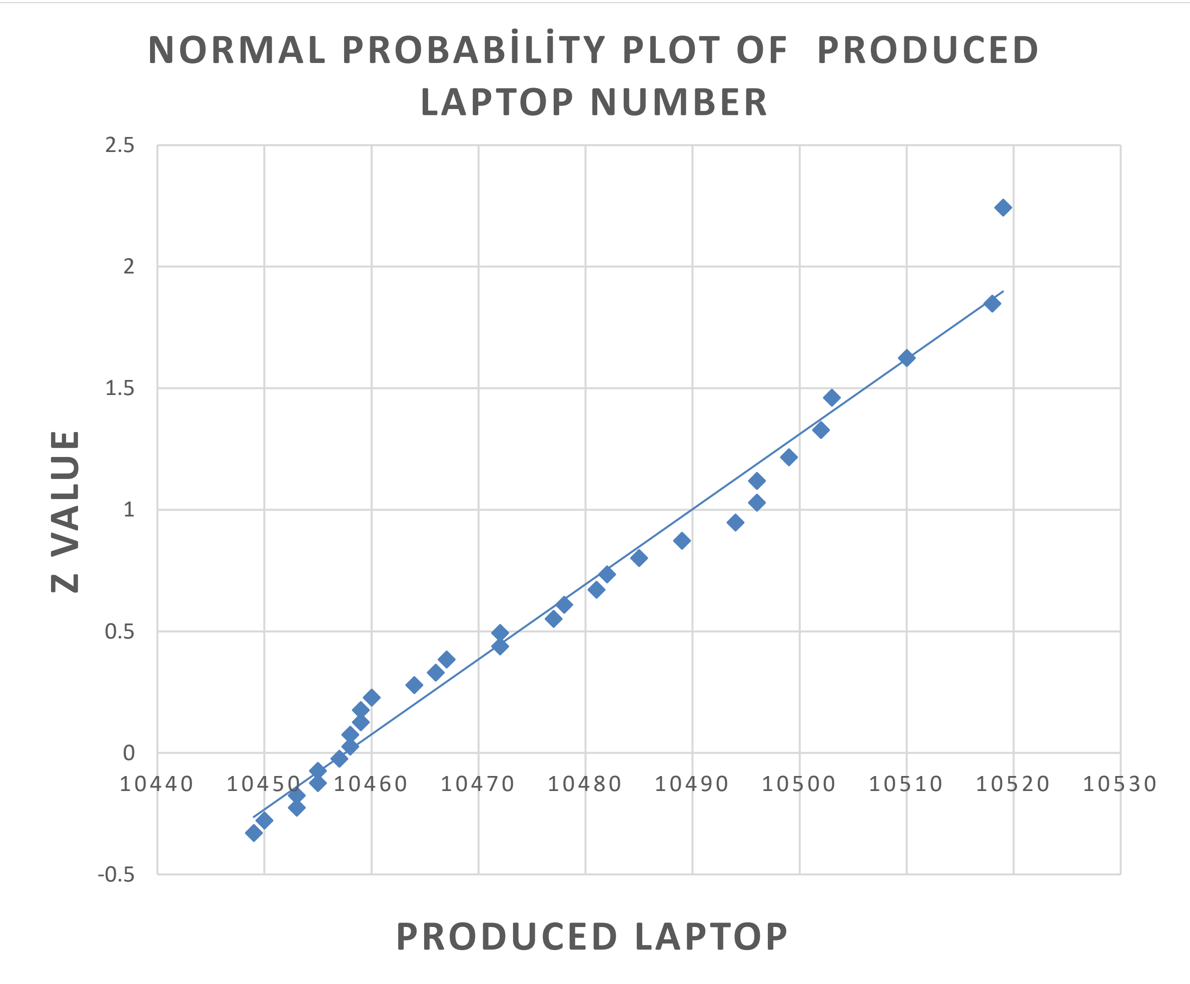
At the towards the end of 30’s we get the lowest producted laptop value. (it’s 38th week and number is 10379) and at the middle of 30’s we get the maximum production.(it’s 34th week and number is 10519).



Most of the variables are being in 10451-10475 (15) .Also we don't have any variables bigger than 10523 and lower than 10366.(There is no certain proof that we don't have a value which is lover than 10379 but if we add all frequencies, their sum equals to 50.Thus that means we don't have any variables which are beyond our limits.)

Next step we will discuss are these Samples fitting to normal distrubition.Here are the normal probability diagrams of Sample 1 and Sample 2 respectively.


Normal Probobility Plot of Produced Laptop Number
Z Value
3
2
1
0
-1
-2
-3
10350
10400
10450
10500
10550
10600
Produced Laptop


NORMAL PROBABİLİTY PLOT OF PRODUCED LAPTOP NUMBER
Z VALUE
2.5
2
1.5
1
0.5
0
-0.5
10440
10450
10460
10470
10480
10490
10500
10510
10520
10530
PRODUCED LAPTOP

As we can see we don’t have much curve shape in graphics and trend line is highlighting most of the points for both Sample 1 and Sample 2.These graphics are providing that we have normal distrubition.

We test the hypothesis that there is really no difference between the two population means.

$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$

Here Ho is implying that the two populations that the samples were taken from are, in effect, a single population since there is no difference between them.

As we know, the null hypothesis, Ho is a "straw man." We set it up and then see whether or not we can knock it down based on the sample evidence.

The null hypothesis that $\mu_1 = \mu_2$ is equivalent to stating that the difference between the two population means is 0. So Ho could be stated as : $(\mu_1 - \mu_2) = 0$.

Note that H_0 is always about population parameters, in this case the difference between the two population means.

The random variable for this test is $(\bar{x}_1 - \bar{x}_2)$, and, as always, takes its value from the sample data.

If the n observation in a sample are denoted by $x_1, x_2, \ldots x_n$, the sample mean is

$$\bar{x} = (x_1 + x_2 + \ldots x_n)/n,$$

$$= \frac{\sum_{i=1}^{n} x_i}{n},$$ by using this formula we find $\bar{x}_1 = 10461,66$ and $\bar{x}_2 = 10456,48$

If the n observations in a sample are denoted by $x_1, x_2, \ldots x_n$, then the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1},$$ by using this formula we find $s_1^2 = 1232,11$ and $s_2^2 = 1108,58$

We can use these values for our calculations since we have numerous observations at least much more than 30)

If the samples are large, random, and independent, then $(\bar{x}_1-\bar{x}_2)$, which is a random variable, has approximately a normal distribution, with

$$E(\bar{x}_1-\bar{x}_2)=\mu_1-\mu_2 \quad \text{and} \quad \sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

(σ means standard deviation of population)

If the sample sizes are large enough, or if population standard deviations are known, we use a two-sample Z-test. But if we have a small sample and also don't know the standart deviations of populations. We have to use a two-sample t-test.

As for our problem we have 50 observations for each sample so we can use Z-test.

to calculate Z, we use ;

$$Z=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \quad \text{or} \quad \frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

→ as long as n_1+n_2 is large enough

Sample 1

Sample size : $n_1=50$

Sample mean : $\bar{x}_1=10461{,}66$ productions per week

Sample standart deviation : $S_1=35{,}10$ productions per week

Sample 2

Sample size : $n_2=50$

Sample mean : $x_2=10456{,}48$ productions per week

Sample standart deviation : $S_2=33{,}30$ productions per week

We will investigate whether $\mu_1=\mu_2$ claim is true or not and we will test null hypothesis at $a=0{,}05$

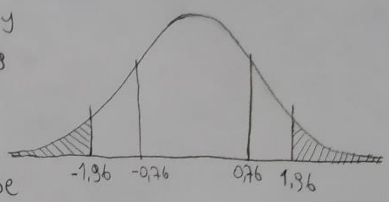
→ $\mu_1-\mu_2=0$ Because we're testing H_0 *

$$Z=\frac{10461{,}66-10456{,}48-(0)}{\sqrt{\frac{(1232{,}11)}{50}+\frac{1108{,}58}{50}}}=0{,}76$$

I should denote that this is a two tail test cause variables can be greater or lower than μ" And if you open up "Z table" you will see plenty of numbers. We are looking for (because of two tail) $(1-a/2)=0{,}97500$ appropriate Z value is 1,96. In Z distribution graph we will select $\mp$ 1,96 points. Because of symetry a error was distributed equally beyond each boundary.

$\frac{a}{2}=0{,}025$ *

Since our z interval betwen $\mp 1{,}96$ we don't have enough proof to reject H_0 hypothesis. Otherwise we could say resecting H_0 contains $\leq 0{,}05$ marging error. Neglecting this risk and rejecting H_0 would be a smart move. But this situation didn't happen in our exomple.



It's time to define %95 confidence interval we use

$(\bar{X}_1-\bar{X}_2) \mp Z\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$, (I have showed $Z=\mp 1{,}96$)

$$(10461{,}66-10456{,}48) \mp 1{,}96\sqrt{\frac{1232{,}11}{50}+\frac{1108{,}58}{50}}$$

As result we get $-8{,}23 \longleftrightarrow 18{,}59$ %95 CIE (%95 possibility $(\mu_1-\mu_2)$ between this interval)

You should notice that 0 is between these two values. So that means it's possible to say factory1's laptop producing mean may equal to factory2's laptop producing mean. But it's not certain it's just possible. We have to understand the difference.

Reference Books and Programmes Used

Engineering Statistics (5th ed.), by Douglas C. Montgomery, George C. Runger, Norma F. Hubele, John Wiley, New York, NY (2012).

Microsoft Excel