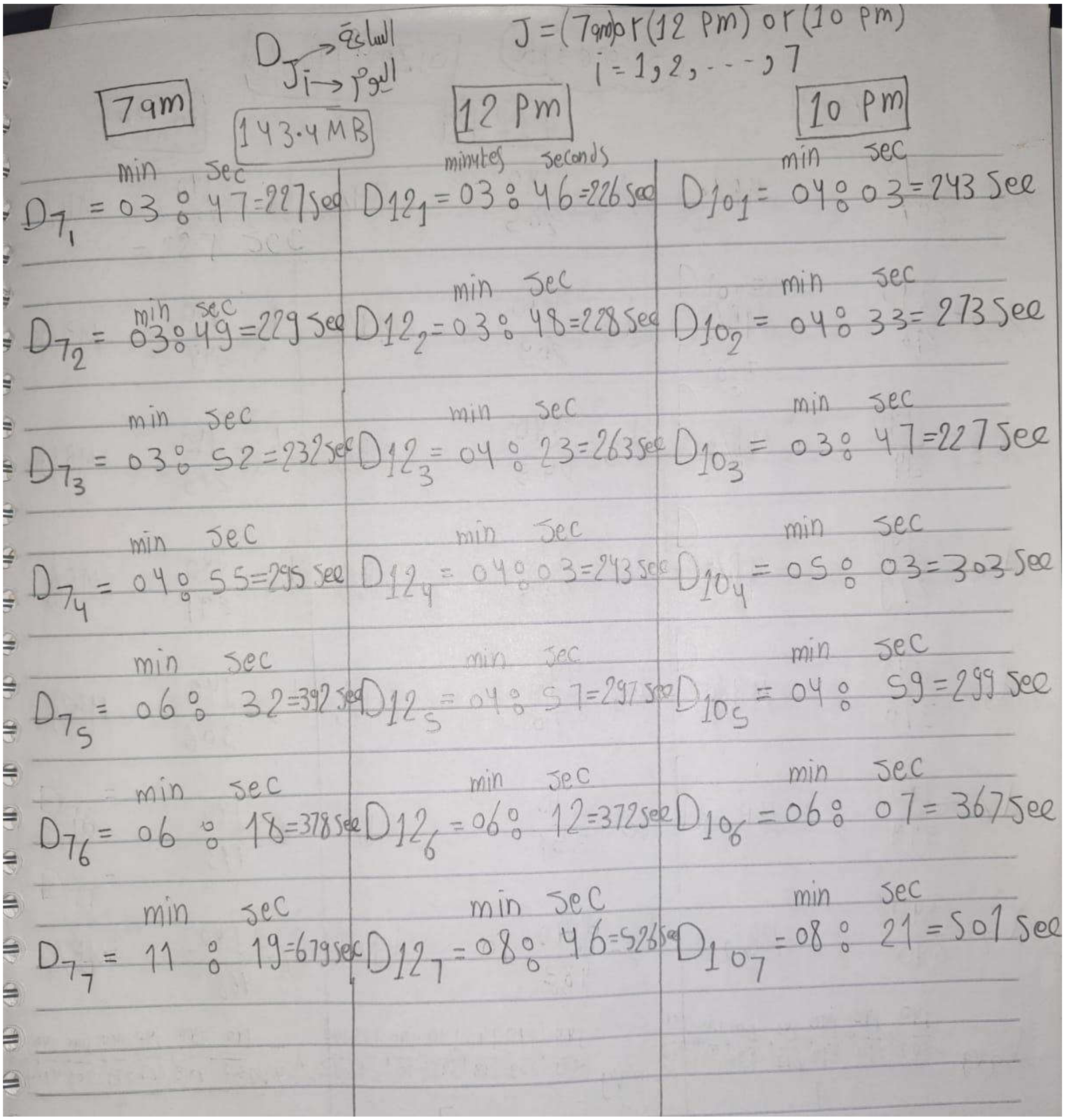
Task 1

1.

(first file)



D_{J_i}: J → الساعة, i → اليوم

J = (7am) or (12 Pm) or (10 pm)

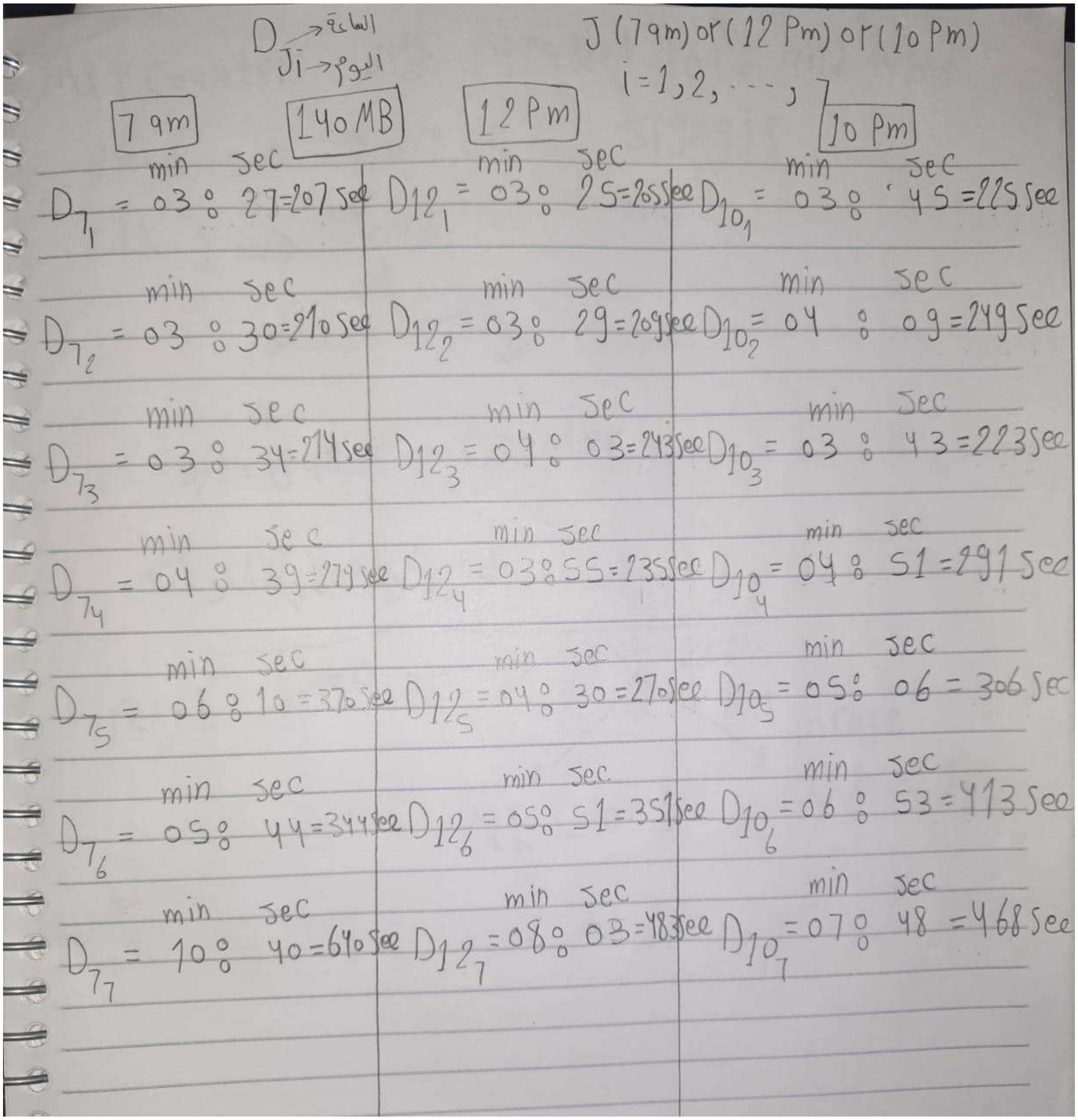
i = 1, 2, ..., 7

143.4 MB

7am	12 Pm	10 Pm
$D7_1$ = 03 : 47 = 227 sec	$D12_1$ = 03 : 46 = 226 sec	$D10_1$ = 04 : 03 = 243 sec
$D7_2$ = 03 : 49 = 229 sec	$D12_2$ = 03 : 48 = 228 sec	$D10_2$ = 04 : 33 = 273 sec
$D7_3$ = 03 : 52 = 232 sec	$D12_3$ = 04 : 23 = 263 sec	$D10_3$ = 03 : 47 = 227 sec
$D7_4$ = 04 : 55 = 295 sec	$D12_4$ = 04 : 03 = 243 sec	$D10_4$ = 05 : 03 = 303 sec
$D7_5$ = 06 : 32 = 392 sec	$D12_5$ = 04 : 57 = 297 sec	$D10_5$ = 04 : 59 = 299 sec
$D7_6$ = 06 : 18 = 378 sec	$D12_6$ = 06 : 12 = 372 sec	$D10_6$ = 06 : 07 = 367 sec
$D7_7$ = 11 : 19 = 679 sec	$D12_7$ = 08 : 46 = 526 sec	$D10_7$ = 08 : 21 = 501 sec

(min : sec)

(Second file)



D → الساعة
J_i → اليوم

J (7am) or (12 Pm) or (10 Pm)
i = 1, 2, ..., 7

7 am	140 MB	12 Pm	10 Pm

7 am (min : sec)	12 Pm (min : sec)	10 Pm (min : sec)
D_{7_1} = 03 : 27 = 207 sec	D_{12_1} = 03 : 25 = 205 sec	D_{10_1} = 03 : 45 = 225 sec
D_{7_2} = 03 : 30 = 210 sec	D_{12_2} = 03 : 29 = 209 sec	D_{10_2} = 04 : 09 = 249 sec
D_{7_3} = 03 : 34 = 214 sec	D_{12_3} = 04 : 03 = 243 sec	D_{10_3} = 03 : 43 = 223 sec
D_{7_4} = 04 : 39 = 279 sec	D_{12_4} = 03 : 55 = 235 sec	D_{10_4} = 04 : 51 = 291 sec
D_{7_5} = 06 : 10 = 370 sec	D_{12_5} = 04 : 30 = 270 sec	D_{10_5} = 05 : 06 = 306 sec
D_{7_6} = 05 : 44 = 344 sec	D_{12_6} = 05 : 51 = 351 sec	D_{10_6} = 06 : 53 = 413 sec
D_{7_7} = 10 : 40 = 640 sec	D_{12_7} = 08 : 03 = 483 sec	D_{10_7} = 07 : 48 = 468 sec

For dealing with data, we convert every one of them to seconds to be easy to deal
EX 3min and 47 sec = 3*60+47=227 sec and so on

1st_file 7 AM	1st_file 12PM	1st_file 10PM	2nd_file 7AM	2nd_file 12PM	2nd_file 10PM
227	226	243	207	205	225
229	228	273	210	209	249
232	263	227	214	243	223
295	243	303	279	235	291
392	297	299	370	270	306
378	372	367	344	351	413
679	526	501	640	483	468

For finding the right distribution we used a program called **Minitab**

Which can detect the right distribution by **the higher P parameter**

2.

WORKSHEET 1

Distribution Identification for 1st-file 7am

Goodness of Fit Test

Distribution	AD	P	LRT P
Normal	0.662	0.047	
Box-Cox Transformation	0.394	0.269	
Lognormal	0.454	0.182	
3-Parameter Lognormal	0.455	*	0.037
Exponential	1.412	0.029	
2-Parameter Exponential	0.462	>0.250	0.000
Weibull	0.594	0.099	
3-Parameter Weibull	0.396	0.396	0.008
Smallest Extreme Value	0.866	0.020	
Largest Extreme Value	0.556	0.135	
Gamma	0.547	0.181	
3-Parameter Gamma	0.391	*	0.007
Logistic	0.570	0.086	
Loglogistic	0.443	0.214	

WORKSHEET 1

Distribution Identification for 1st-file 12pm

Goodness of Fit Test

Distribution	AD	P	LRT P
Normal	0.641	0.054	
Box-Cox Transformation	0.261	0.579	
Lognormal	0.460	0.175	
3-Parameter Lognormal	0.271	*	0.055
Exponential	1.751	0.012	
2-Parameter Exponential	0.275	>0.250	0.000
Weibull	0.625	0.085	
3-Parameter Weibull	0.244	>0.500	0.009
Smallest Extreme Value	0.823	0.024	
Largest Extreme Value	0.512	0.179	
Gamma	0.555	0.172	
3-Parameter Gamma	0.271	*	0.023
Logistic	0.568	0.087	
Loglogistic	0.427	0.232	

WORKSHEET 1

Distribution Identification for 1st-file 10pm

Goodness of Fit Test

Distribution	AD	P	LRT P
Normal	0.467	0.167	
Box-Cox Transformation	0.150	0.928	
Lognormal	0.290	0.508	
3-Parameter Lognormal	0.196	*	0.150
Exponential	1.890	0.008	
2-Parameter Exponential	0.215	>0.250	0.000
Weibull	0.493	0.198	
3-Parameter Weibull	0.247	>0.500	0.081
Smallest Extreme Value	0.700	0.052	
Largest Extreme Value	0.270	>0.250	
Gamma	0.364	>0.250	
3-Parameter Gamma	0.353	*	0.756
Logistic	0.384	>0.250	
Loglogistic	0.257	>0.250	

WORKSHEET 1

Distribution Identification for 2nd-file 7am

Goodness of Fit Test

Distribution	AD	P	LRT P
Normal	0.640	0.054	
Box-Cox Transformation	0.360	0.334	
Lognormal	0.420	0.228	
3-Parameter Lognormal	0.406	*	0.066
Exponential	1.369	0.033	
2-Parameter Exponential	0.400	>0.250	0.000
Weibull	0.565	0.127	
3-Parameter Weibull	0.358	0.470	0.012
Smallest Extreme Value	0.853	0.021	
Largest Extreme Value	0.517	0.175	
Gamma	0.513	0.212	
3-Parameter Gamma	0.364	*	0.017
Logistic	0.540	0.104	
Loglogistic	0.405	>0.250	

Distribution Identification f... ⌄ ×

WORKSHEET 1

Distribution Identification for 2nd-file 12pm

Goodness of Fit Test

Distribution	AD	P	LRT P
Normal	0.612	0.065	
Box-Cox Transformation	0.221	0.729	
Lognormal	0.429	0.215	
3-Parameter Lognormal	0.307	*	0.148
Exponential	1.740	0.012	
2-Parameter Exponential	0.223	>0.250	0.000
Weibull	0.598	0.098	
3-Parameter Weibull	0.258	>0.500	0.020
Smallest Extreme Value	0.791	0.030	
Largest Extreme Value	0.471	0.220	
Gamma	0.525	0.200	
3-Parameter Gamma	0.317	*	0.079
Logistic	0.551	0.096	
Loglogistic	0.402	>0.250	

Distribution Identification f... ⌄ ×

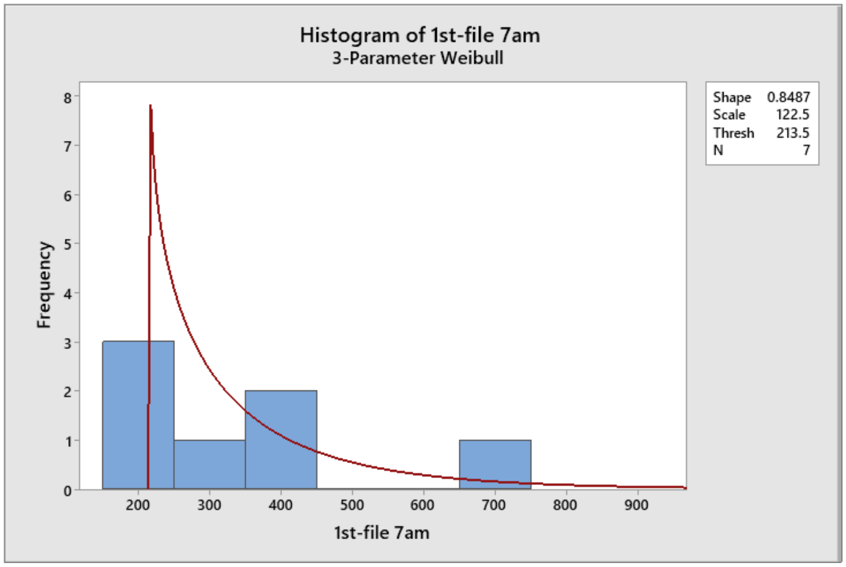
WORKSHEET 1

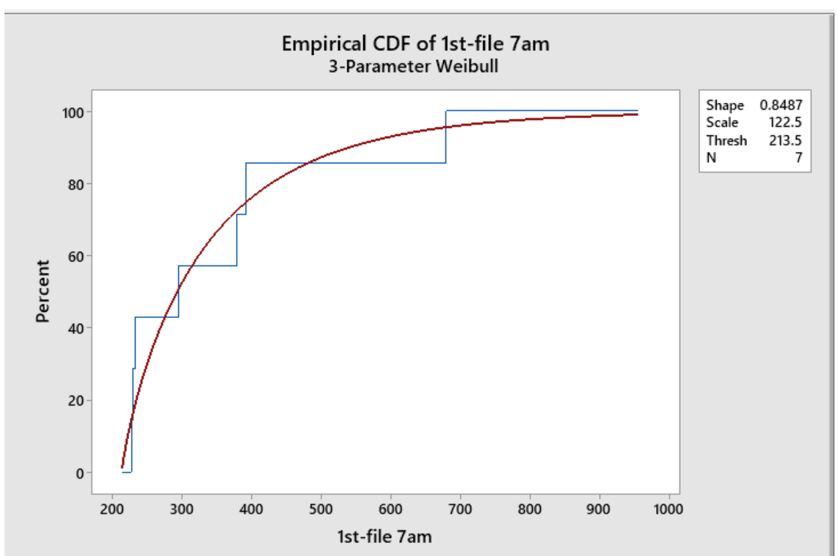
Distribution Identification for 2nd-file 10pm

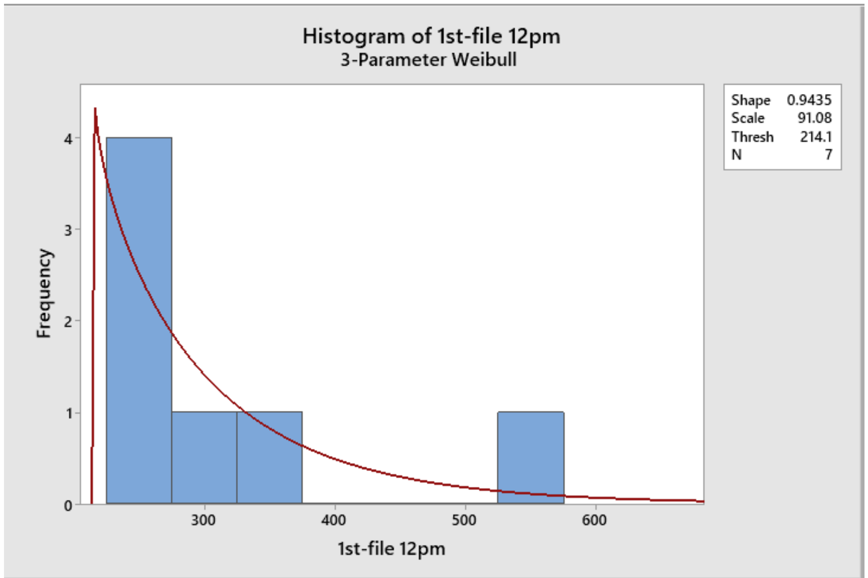
Goodness of Fit Test

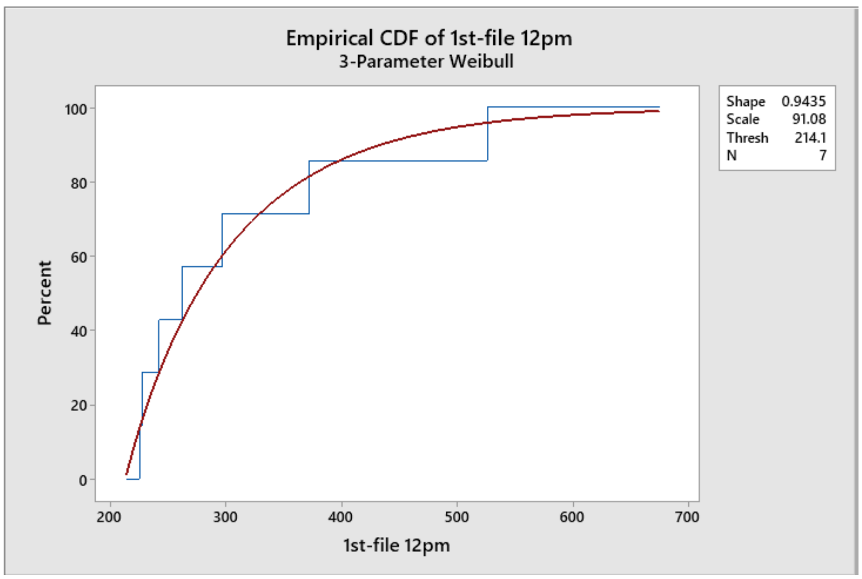
Distribution	AD	P	LRT P
Normal	0.433	0.210	
Box-Cox Transformation	0.277	0.535	
Lognormal	0.334	0.393	
3-Parameter Lognormal	0.341	*	0.427
Exponential	1.779	0.011	
2-Parameter Exponential	0.235	>0.250	0.000
Weibull	0.456	0.235	
3-Parameter Weibull	0.295	>0.500	0.075
Smallest Extreme Value	0.563	0.129	
Largest Extreme Value	0.396	>0.250	
Gamma	0.414	>0.250	
3-Parameter Gamma	0.341	*	0.271
Logistic	0.438	0.220	

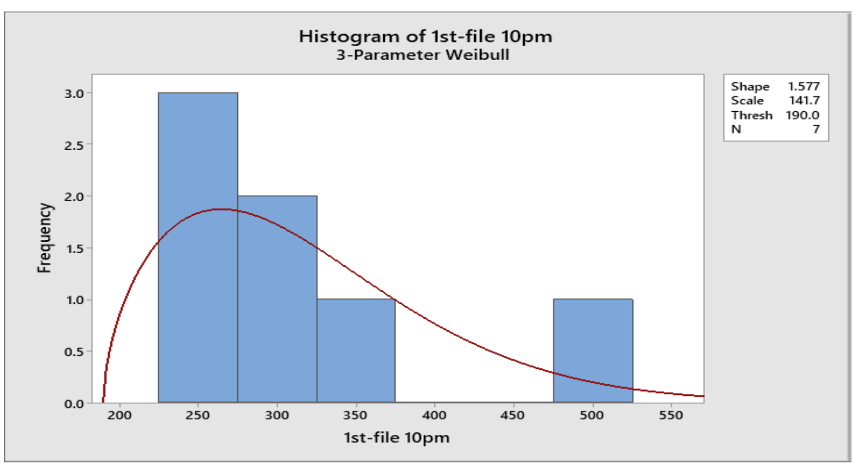
As for high values of p, we have chosen the 3-Parameter Weibull distribution.

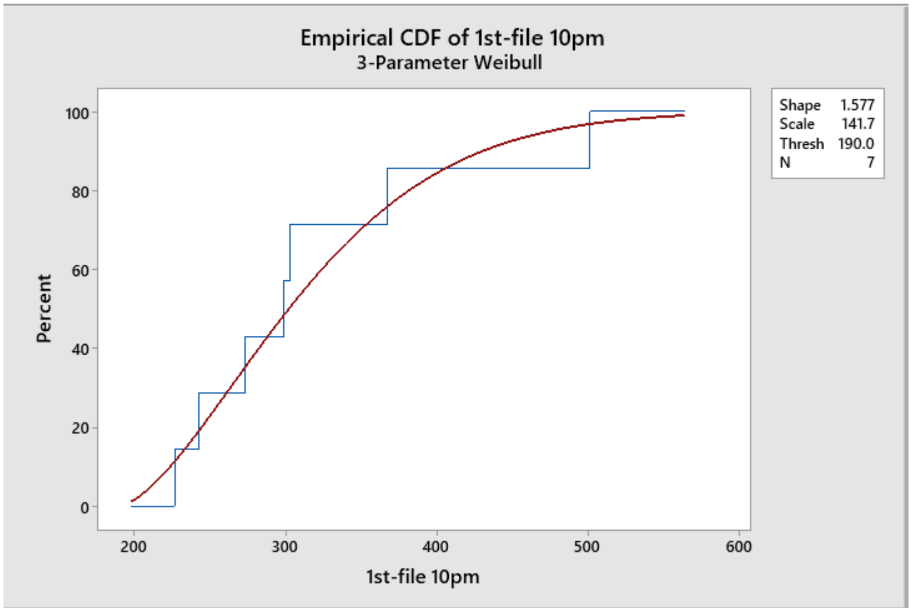

Histogram of 1st-file 7am
3-Parameter Weibull
Shape 0.8487
Scale 122.5
Thresh 213.5
N 7
Frequency
1st-file 7am
0
1
2
3
4
5
6
7
8
200
300
400
500
600
700
800
900

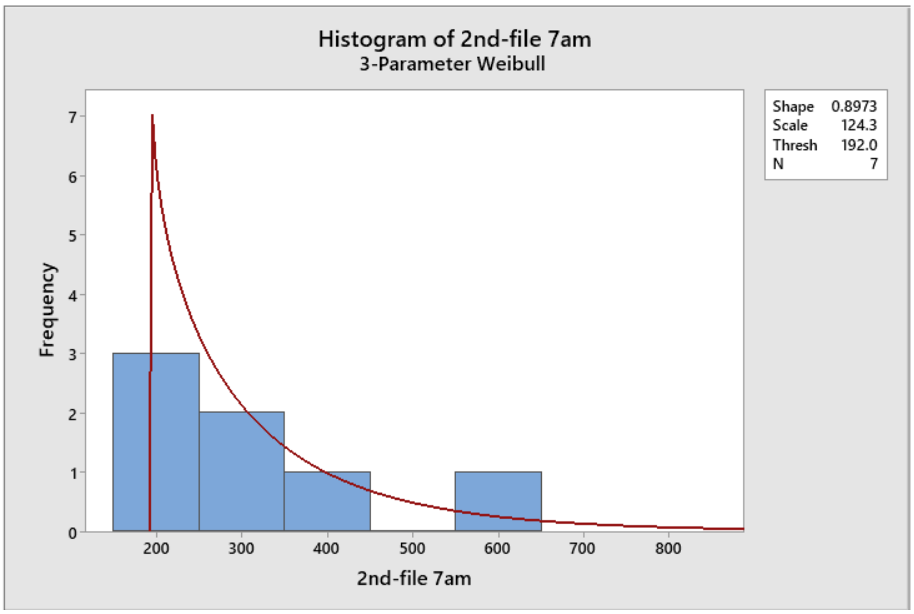

Empirical CDF of 1st-file 7am
3-Parameter Weibull
Shape 0.8487
Scale 122.5
Thresh 213.5
N 7
Percent
1st-file 7am
0
20
40
60
80
100
200
300
400
500
600
700
800
900
1000

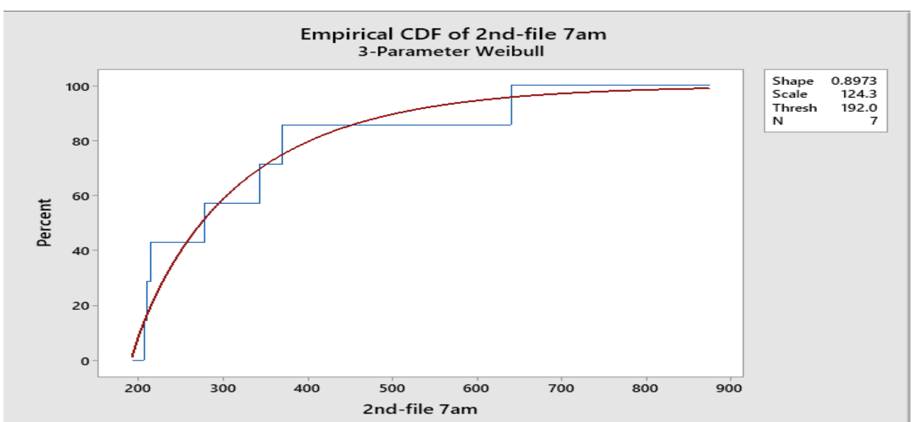

Histogram of 1st-file 12pm
3-Parameter Weibull
Shape 0.9435
Scale 91.08
Thresh 214.1
N 7
Frequency
0
1
2
3
4
300
400
500
600
1st-file 12pm

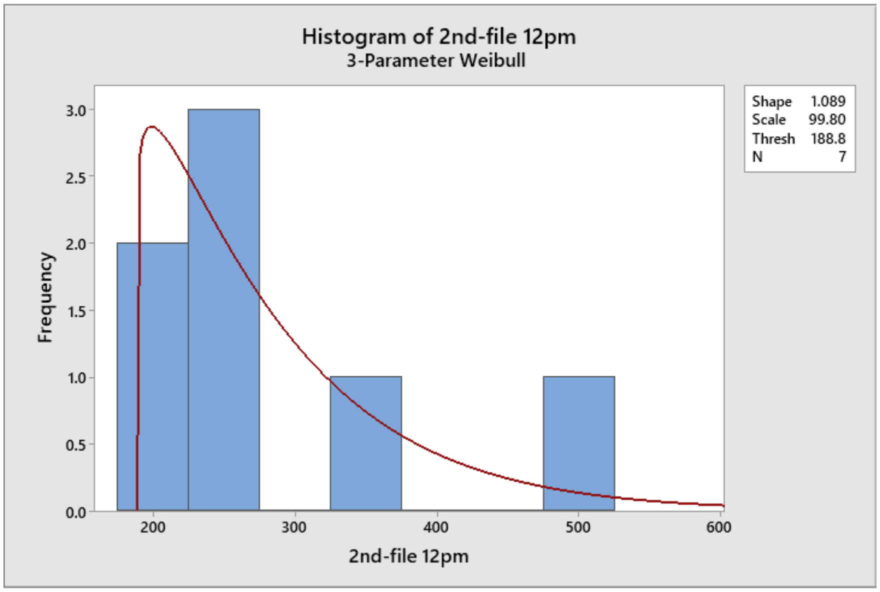

Empirical CDF of 1st-file 12pm
3-Parameter Weibull
Shape 0.9435
Scale 91.08
Thresh 214.1
N 7
Percent
0
20
40
60
80
100
200
300
400
500
600
700
1st-file 12pm

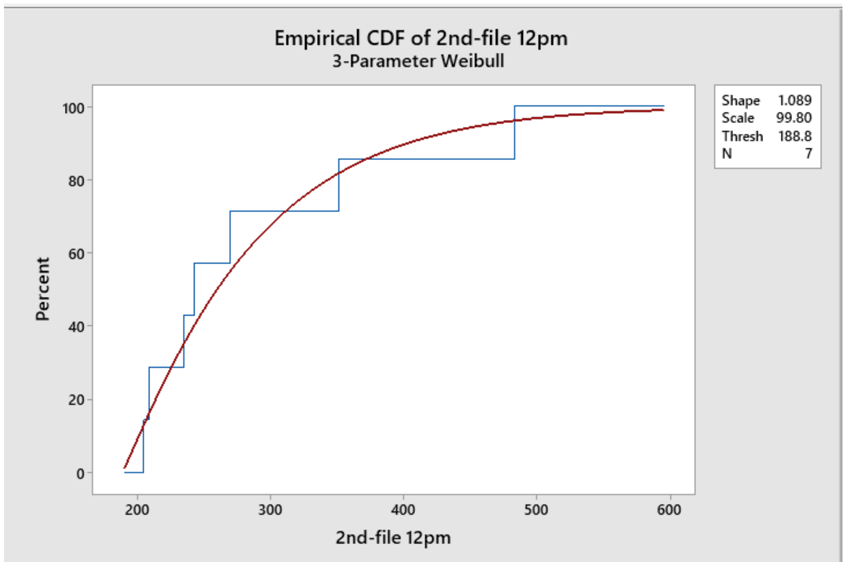

Histogram of 1st-file 10pm
3-Parameter Weibull
Shape 1.577
Scale 141.7
Thresh 190.0
N 7
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
200
250
300
350
400
450
500
550
1st-file 10pm

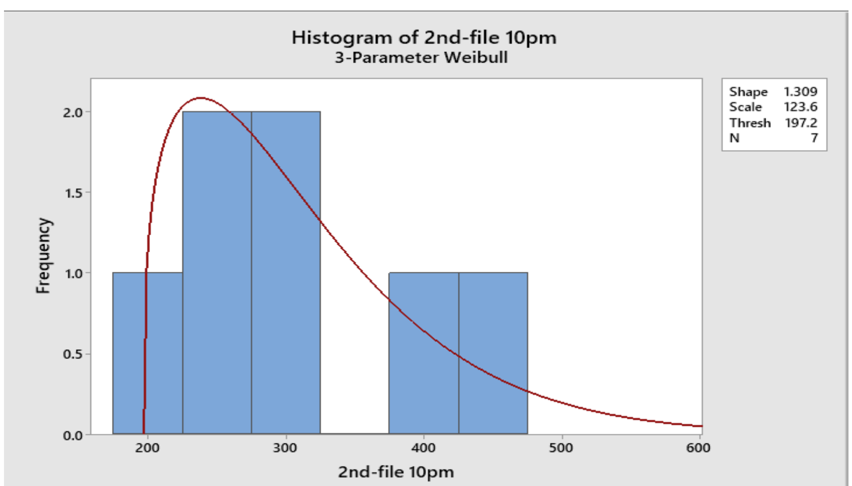

Empirical CDF of 1st-file 10pm
3-Parameter Weibull
Shape 1.577
Scale 141.7
Thresh 190.0
N 7
Percent
0
20
40
60
80
100
200
300
400
500
600
1st-file 10pm

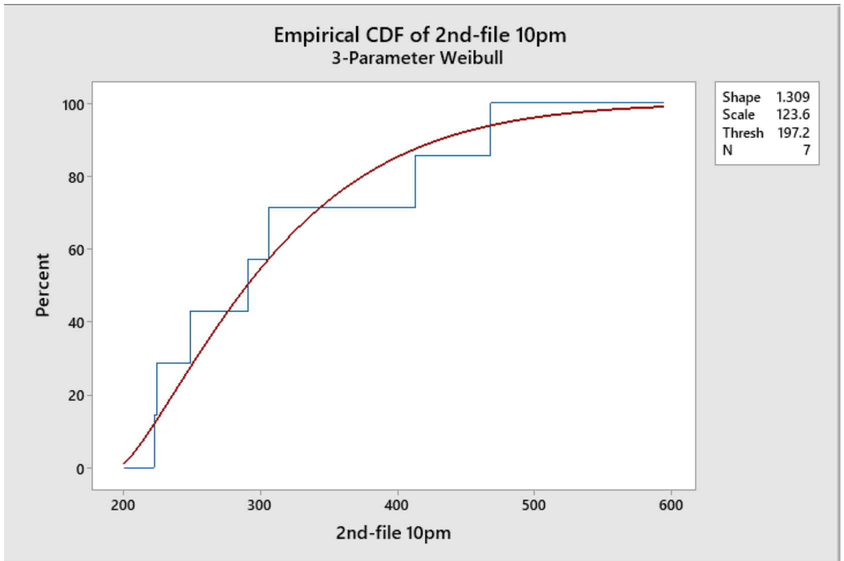

Histogram of 2nd-file 7am
3-Parameter Weibull
Shape 0.8973
Scale 124.3
Thresh 192.0
N 7
Frequency
0
1
2
3
4
5
6
7
200
300
400
500
600
700
800
2nd-file 7am


Empirical CDF of 2nd-file 7am
3-Parameter Weibull
Shape 0.8973
Scale 124.3
Thresh 192.0
N 7
Percent
0
20
40
60
80
100
200
300
400
500
600
700
800
900
2nd-file 7am


Histogram of 2nd-file 12pm
3-Parameter Weibull
Shape 1.089
Scale 99.80
Thresh 188.8
N 7
Frequency
3.0
2.5
2.0
1.5
1.0
0.5
0.0
200
300
400
500
600
2nd-file 12pm


Empirical CDF of 2nd-file 12pm
3-Parameter Weibull
Shape 1.089
Scale 99.80
Thresh 188.8
N 7
Percent
100
80
60
40
20
0
200
300
400
500
600
2nd-file 12pm


Histogram of 2nd-file 10pm
3-Parameter Weibull
Shape 1.309
Scale 123.6
Thresh 197.2
N 7
Frequency
2.0
1.5
1.0
0.5
0.0
200
300
400
500
600
2nd-file 10pm



3.

A) Yes, the distributions are similar. They all follow the 3-Parameter Weibull distribution.

B) The mean of the 1st file at 7am: Mean 347.4 N 7

The mean of the 1st file at 12pm: Mean 307.9 N 7

The mean of the 1st file at 10pm: Mean 316.1 N 7

The mean of the 2nd file at 7am: Mean 323.4 N 7

The mean of the 2nd file at 12pm: Mean 285.1 N 7

The mean of the 2nd file at 10pm: Mean 310.7 N 7

From these results, we recommend to download at 12 pm.

c) we have shown the data of the second file. we have downloaded the first file from Google Drive which has data Centers in countries in North America, South America, Europe and Asia. we have downloaded the second file from GEOFABRIK Site which has a data center in Germany. Germany isn't of the countries which the data centers of Google Drive are located in.

The second file has been downloaded faster from the first file as the mean time of downloading the Second file is less than the first file:

The difference in size between the two files is 3 MB ("first file:143 MB" "second file: 140 MB") so the difference is small so we recommend to download from the second site (GEOFABRIK).

mean of 1^{st} File at 7am = $\frac{227+229+232+295+392+378+679}{7}$ = 347.429 Sec

mean of 1^{st} File at 12 Pm = $\frac{226+228+263+243+297+372+526}{7}$ = 307.857 Sec

mean of 1^{st} File at 10Pm = $\frac{243+273+227+303+299+367+501}{7}$ = 316.143 Sec

mean of 2^{nd} File at 7am = $\frac{207+210+214+279+370+344+640}{7}$ = 323.429 Sec

mean of 2^{nd} File at 12 Pm = $\frac{205+209+243+235+270+357+483}{7}$ = 285.143 Sec

mean of 2^{nd} File at 10Pm = $\frac{225+249+223+291+306+413+468}{7}$ = 310.714 Sec

$\boxed{\text{Saving time}}$

First file

mean at 7am − mean at 12 Pm = 347.429 − 307.857 = 39.572 Sec

mean at 10 Pm − mean at 12 Pm = 316.143 − 307.857 = 8.286 Sec

second File

mean at 7am − mean at 12 Pm = 323.429 − 285.143 = 38.286 Sec

mean at 10 Pm − mean at 12 Pm = 310.714 − 285.143 = 25.571 Sec

$\boxed{\text{Expecting gain}}$

First File

$\frac{\text{mean at 12 Pm}}{\text{mean at 7 am}} = \frac{307.857}{347.429} = 0.8861$

$\frac{\text{mean at 12 Pm}}{\text{mean at 10 Pm}} = \frac{307.857}{316.143} = 0.97379$

Second File

$\frac{\text{mean at 12 Pm}}{\text{mean at 7am}} = \frac{285.143}{323.429} = 0.88162$

$\frac{\text{mean at 12 Pm}}{\text{mean at 10 Pm}} = \frac{285.143}{310.714} = 0.9177$

TASK (2)

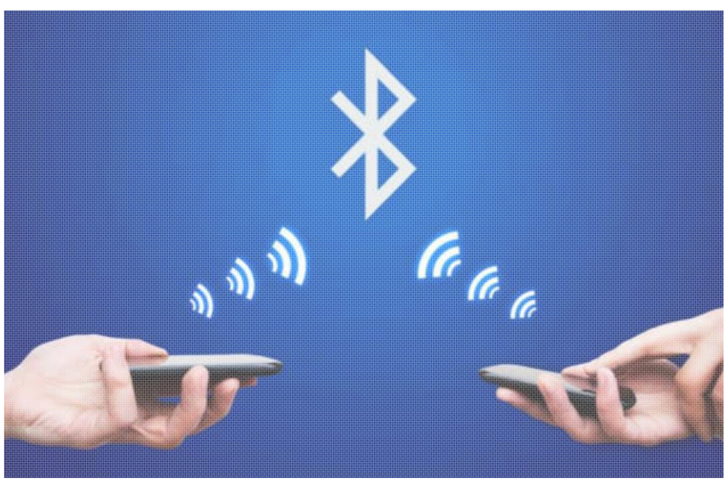
Bluetooth

Brief about the technology :

Bluetooth is a wireless technology standard used for exchanging data between fixed and mobile devices over short distances using UHF radio waves.

Why Bluetooth ?

- Bluetooth removes problem of radio interface by using speed frequency hopping technology.
- low power consumption for the chip almost 0.3 mW so more battery life.
- over comes constrains of line of sight.
- although new technologies nowadays but Bluetooth is a perfect solution for some applications like internet of things (IOT) and connecting some devices wire-less .



we have used in our experiment and all results are depending on a Bluetooth 4.2 version for both the transmitter and the receiver at outdoor .

walking through the experiment:

a. Taking results for sending same file at different time.

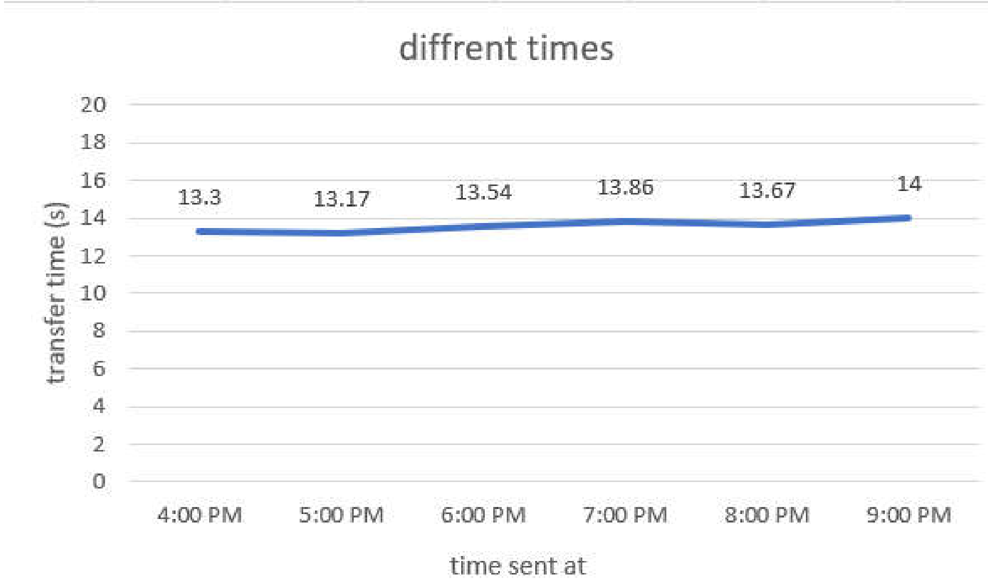
a file size 2.3 MB is used .

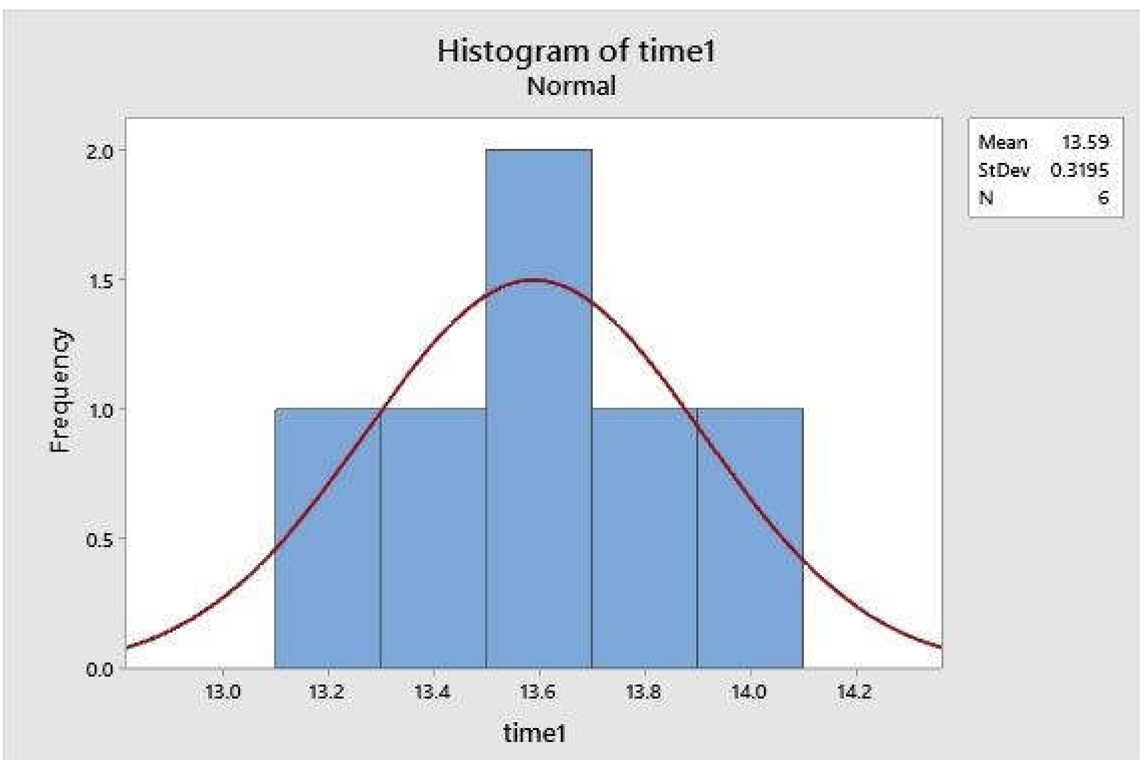
starting the experiment at 4 pm and repeating for every hour.

Time sent	Transfer time(S)
4:00 pm	13.3
5:00 pm	13.17
6:00 pm	13.54
7:00 pm	13.86
8:00 pm	13.67
9:00 pm	14

From the previous results we can see clearly that transfer time for same file at different times is almost the same during whole experiment So , transfer time is **Almost** independent of the time we send at .

The time we send at doesn't really affect the results .





Our results can be expressed as normal distribution.

b. repeating the measurement for different distances between the wireless devices.

A 1.56 MB file is used through this part .

Taking in consider that Bluetooth 4.2 range out door is different from indoor

Our results based on outdoor readings .

Distance between devices (m)	Transfer time (s)
0	9.09
2.5	9.14
5	9.61
10	10.17
15	10.94
20	13.07
25	13.79
30	16
35	26.5
40	30.825
45	34.5
50	Connection failed

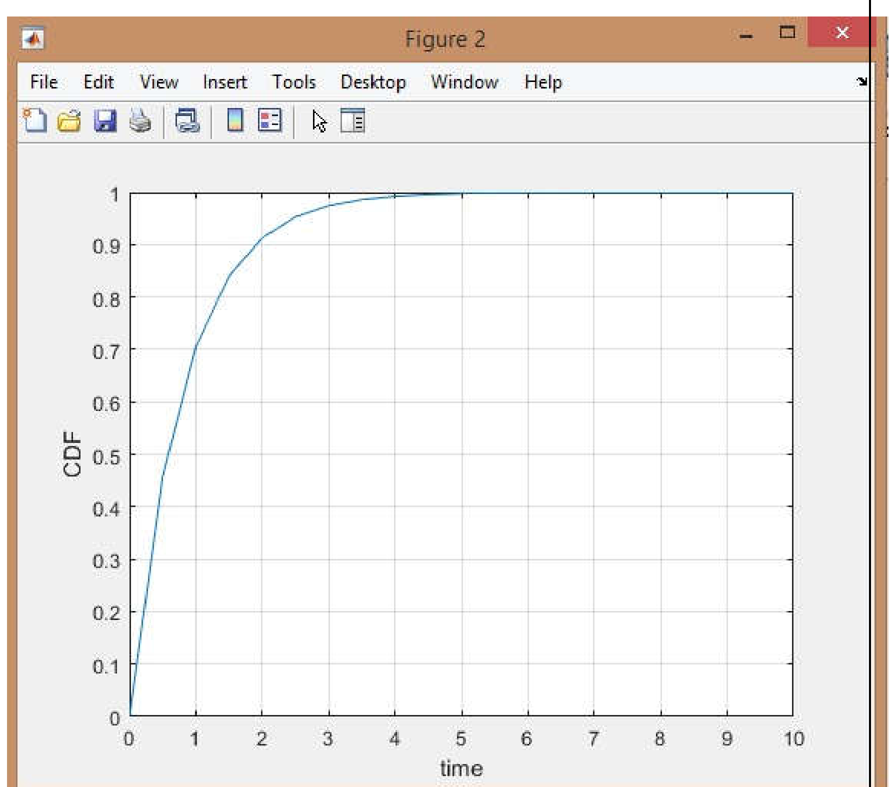
We have noted that time required to transfer the file depends on other factors rather the distance like the crowd in street so the readings is almost all in a little crowded street .

Notes from results :

- As long as we increase the distance the time required to transfer the same file will increase .
- At 50m the connection between devices have failed So we can say that :
- Maximum range to transfer is < 50M , beyond this range connection failed .
- Our recommended range is < 25M ,connection was so good and take the shortest times to transfer at this range .
  The time needed in a range less than 20m is better by almost 35% the maximum range .

-of course all ranges are for outdoor .

- Probability same file to be sent decreases as long as we increase the distance we can say that it is almost exponential distributed .



The figure represent a C.D.F for previous results considering it follows an exponential distribution .

c. by repeating for different file sizes.

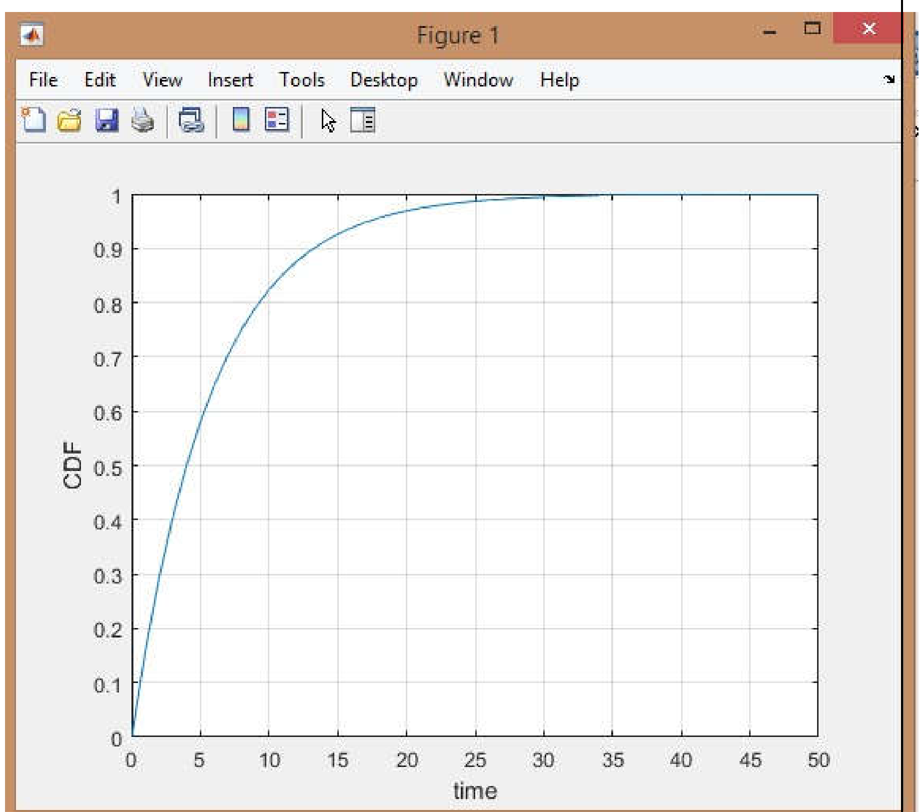
We have used a variety of file sizes up to 12.5 MB all the operation at a constant distance between the devices .

File size (MB)	Transfer time (s)	Rate (MB/S)
1.2	6.92	0.173
1.6	9.31	0.171
1.9	11.41	0.166
3	17.68	0.169
3.3	18.65	0.176
5.2	30	0.173
7.2	41.22	0.174
8.7	48.53	0.180
10.7	59.85	0.178
12.4	68.88	0.180

Notes from results :

- as long as the file size increase the time needed to transfer the file also increased .
- data rate is almost constant for variety of sizes .
- from our results we can say that bluetooth is a good choice for transferring small sizes of files but for large files it won't be the smartest choice .

the figure represent a C.D.F if it follows an exponential distribution .



Figures in this task is done by matlab with following code for the C.D.F's .

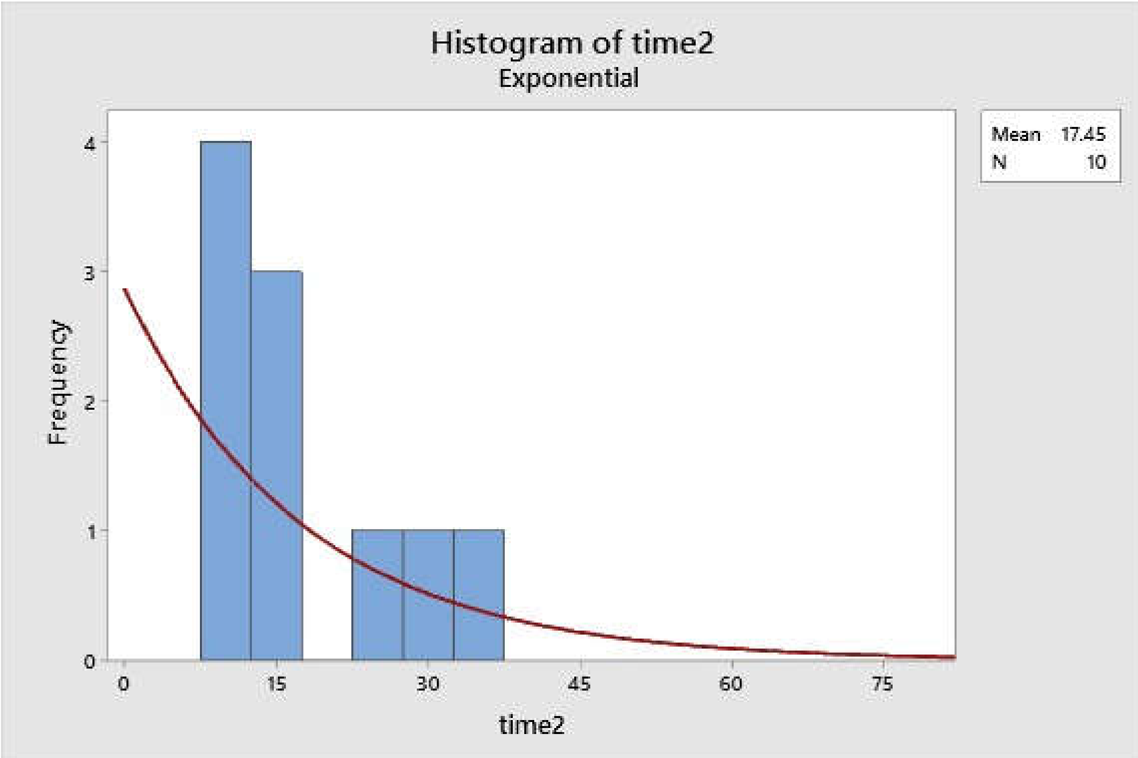
Note : the average used in the equations is the average of our results from previous tables .

Code :

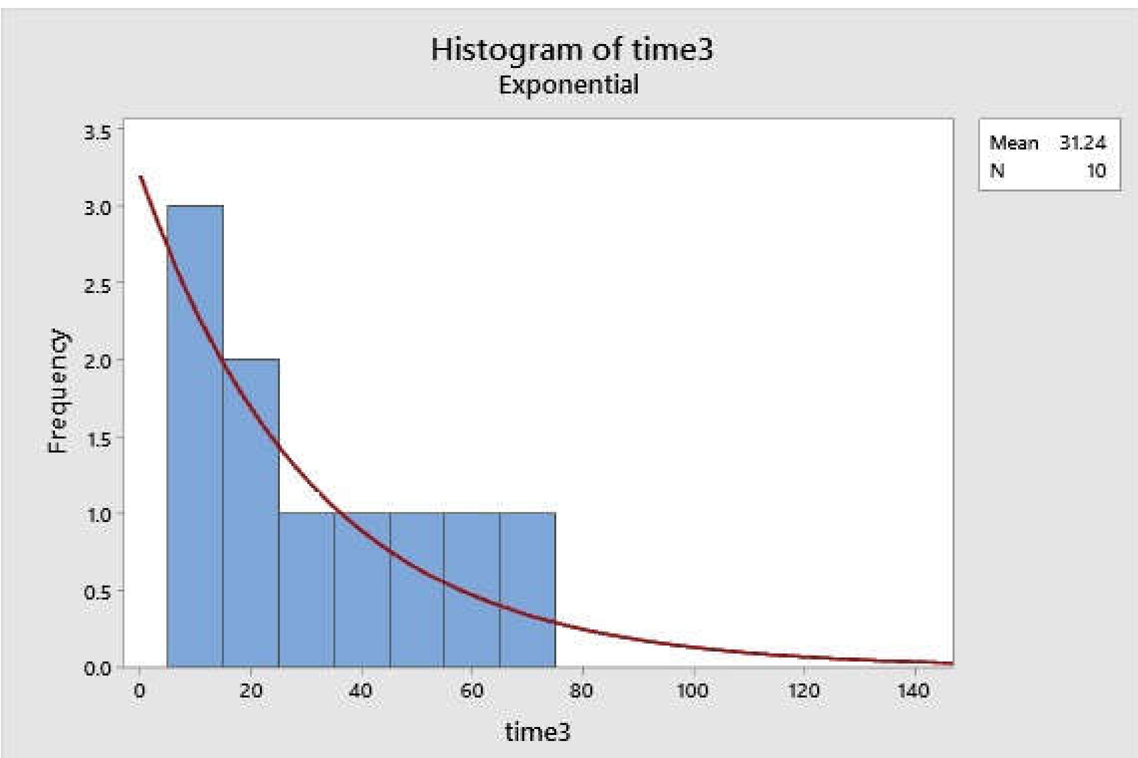
```
%Plotting CDF of the file size
x1=[0:1:50];
e=2.71828182845904523536028747135 27;
y1=1-(e.^(-.1739*x1));
figure(1);
plot(x1,y1), xlabel('time'), ylabel('CDF');
grid on;

%Plotting CDF of ditance
x2=[0:.5:10];
y2=1-(e.^(-1.223*x2));
figure(2);
plot(x2,y2),xlabel('time'), ylabel('CDF');
grid on;
```

Histograms and P.D.F using Minitab



distance



File size

Task 3

Explanation about task 3

In this task we chose two different data:

1) Data about the 20 years of Olympic history(athletes): their age, height and weight for almost 27,000 athletes

the link for this data is

https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

2) Data about the population of the cities in India (population of men, population of women, men who graduated, women who graduated for almost 500 different cities in India

the link for this data is

https://www.kaggle.com/zed9941/top-500-indian-cities

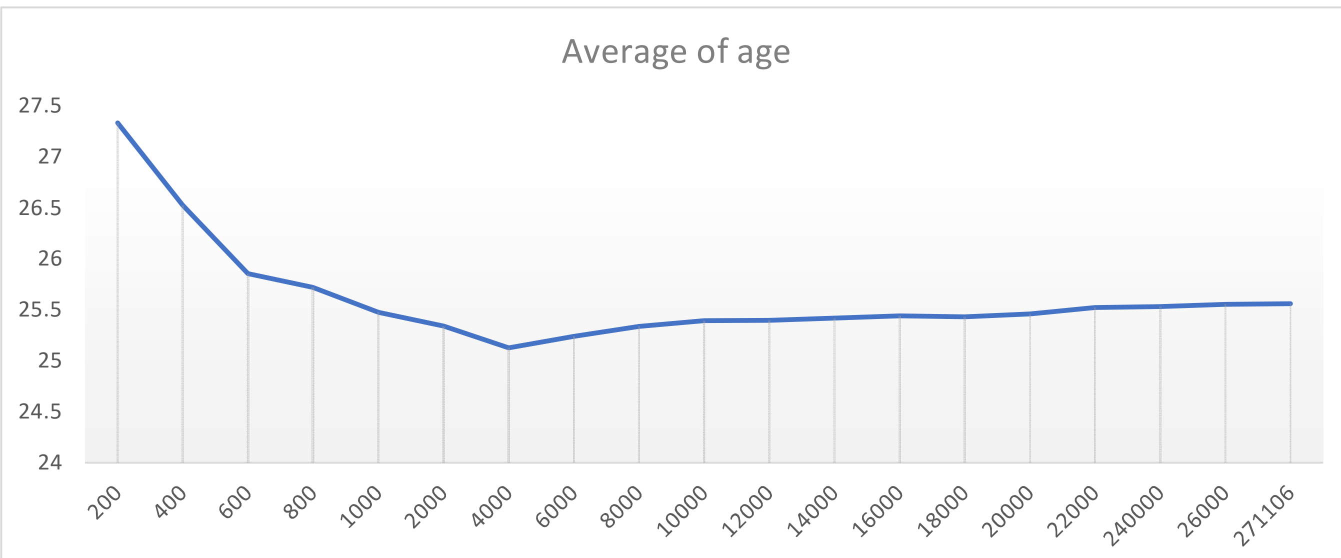
these two data are from Kaggle website and there are real as many data scientists and machine learning developer take this data to use it in their work

we use every Random variable in this data, take samples, take their average and each time we increase the size of the sample until the mean you compute is independent from the sample size

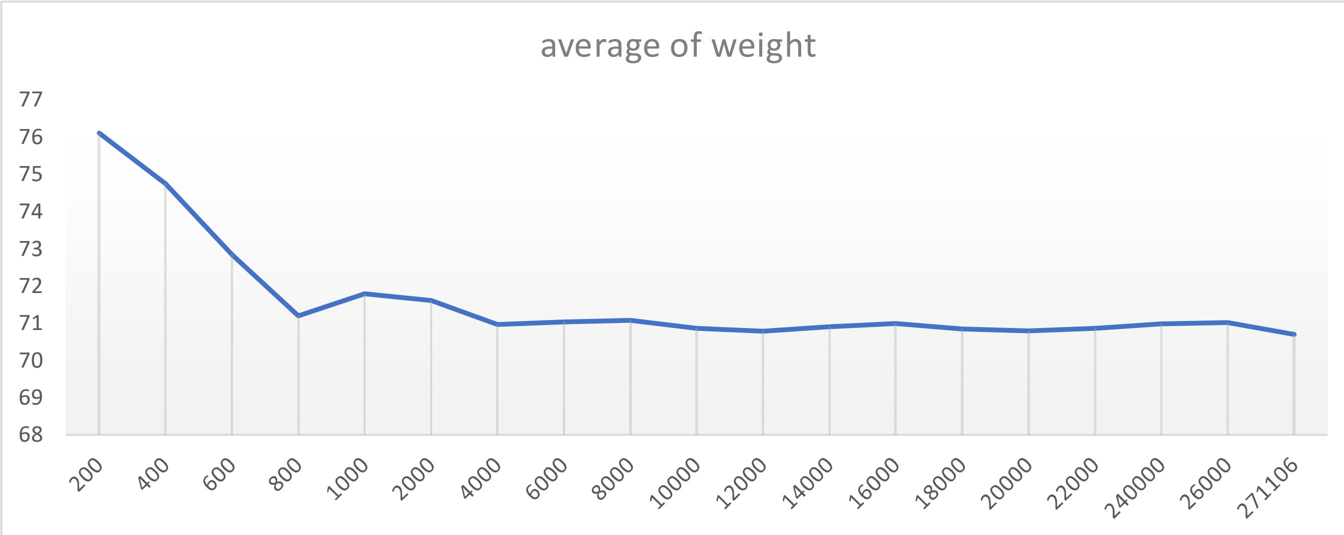
we also include the excel data sheets in our submission and we include the table of the number of samples, the mean of them and the plots which explain the results in this report

Data about the 20 years of Olympic history(athletes)

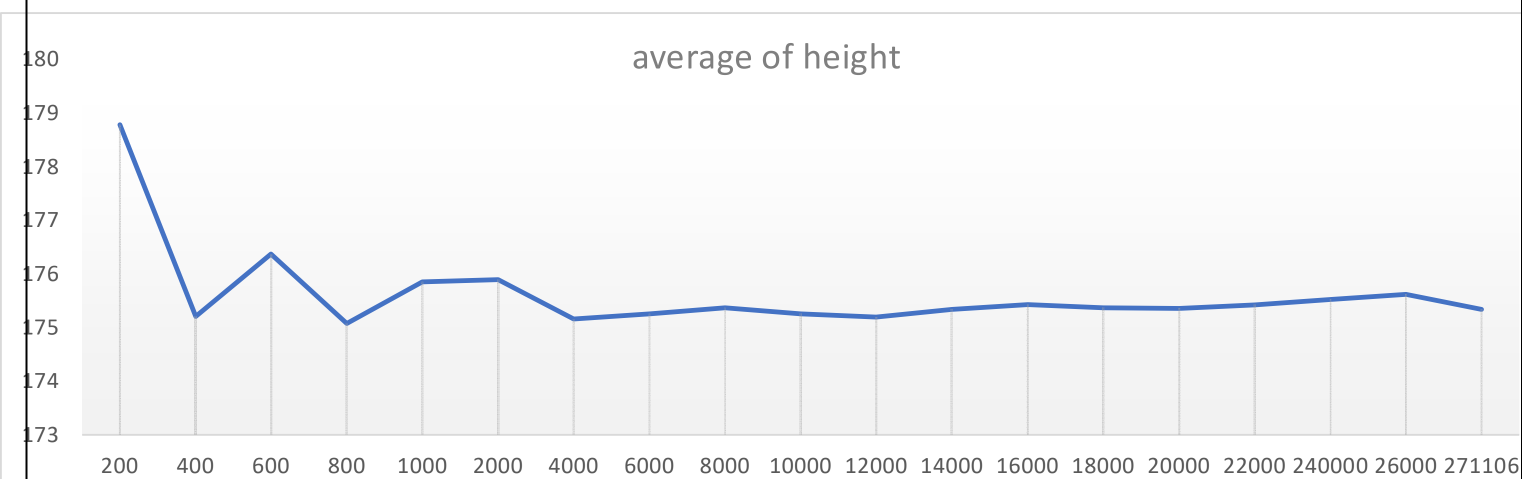
Number of sample	average of age	average of height	average of weight
200	27.33163265	178.7831325	76.10542169
400	26.52432432	175.2069173	74.74914676
600	25.85424354	176.3673469	72.84451902
800	25.72027027	175.0793388	71.20343137
1000	25.47435897	175.8503937	71.79335072
2000	25.34032512	175.8928571	71.6120801
4000	25.12734975	175.1595711	70.97222222
6000	25.2383629	175.2569809	71.03729456
8000	25.33596215	175.3727668	71.08134824
10000	25.39120395	175.2568819	70.86644147
12000	25.39523102	175.1957311	70.78490627
14000	25.41817502	175.3400329	70.9071356
16000	25.43783077	175.4330414	70.99428295



for **age variable**, we recommended to take a sample size **more than or equal 22000** because it’s obvious that the mean started to be constant from that size



for **weight variable**, we recommended to take a sample size **more than or equal 10000** because it's obvious that the mean started to be constant from that size

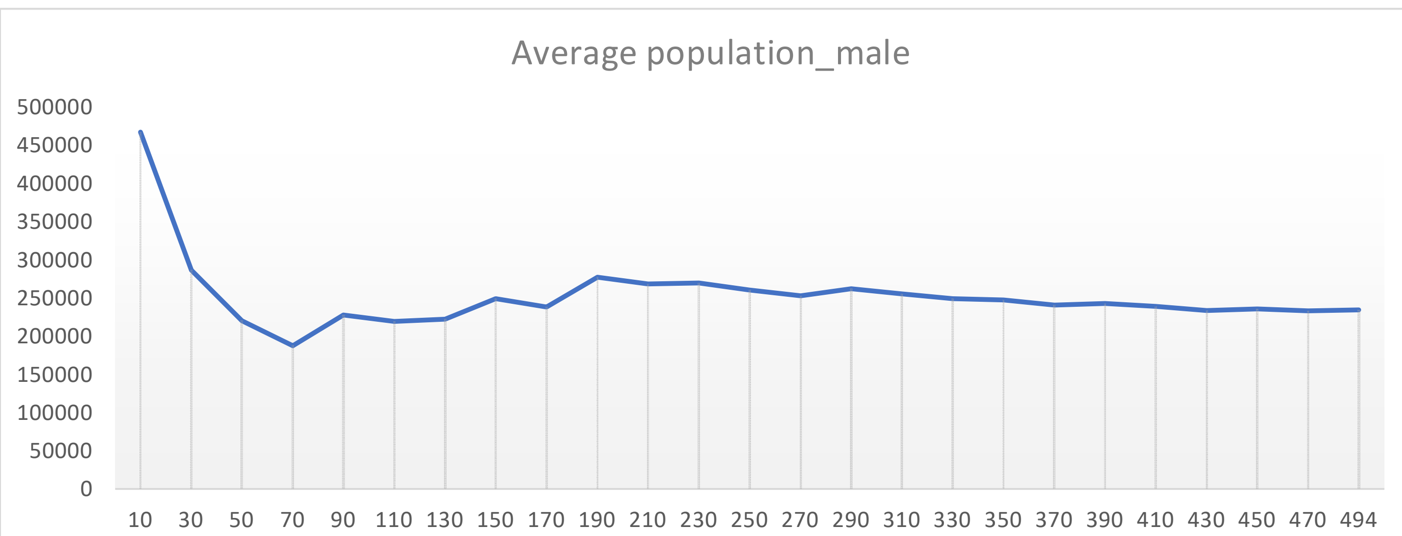


for **height variable**, we recommended to take a sample size **more than or equal 10000** because it's obvious that the mean started to be constant from that size

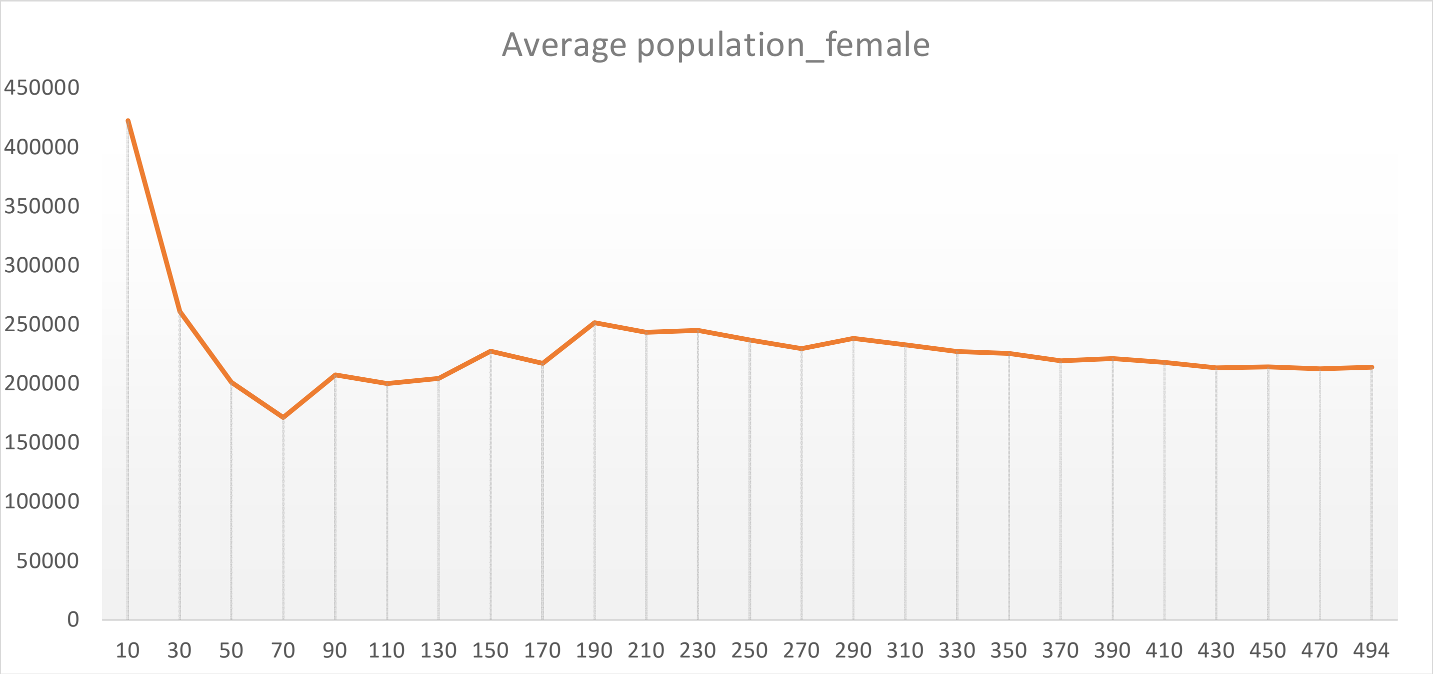
so, as a result: while increasing the sample size the mean will be orbit around a specific value until we reach a constant value for the mean which don't depend on the sample size

Data about the population of the cities in India

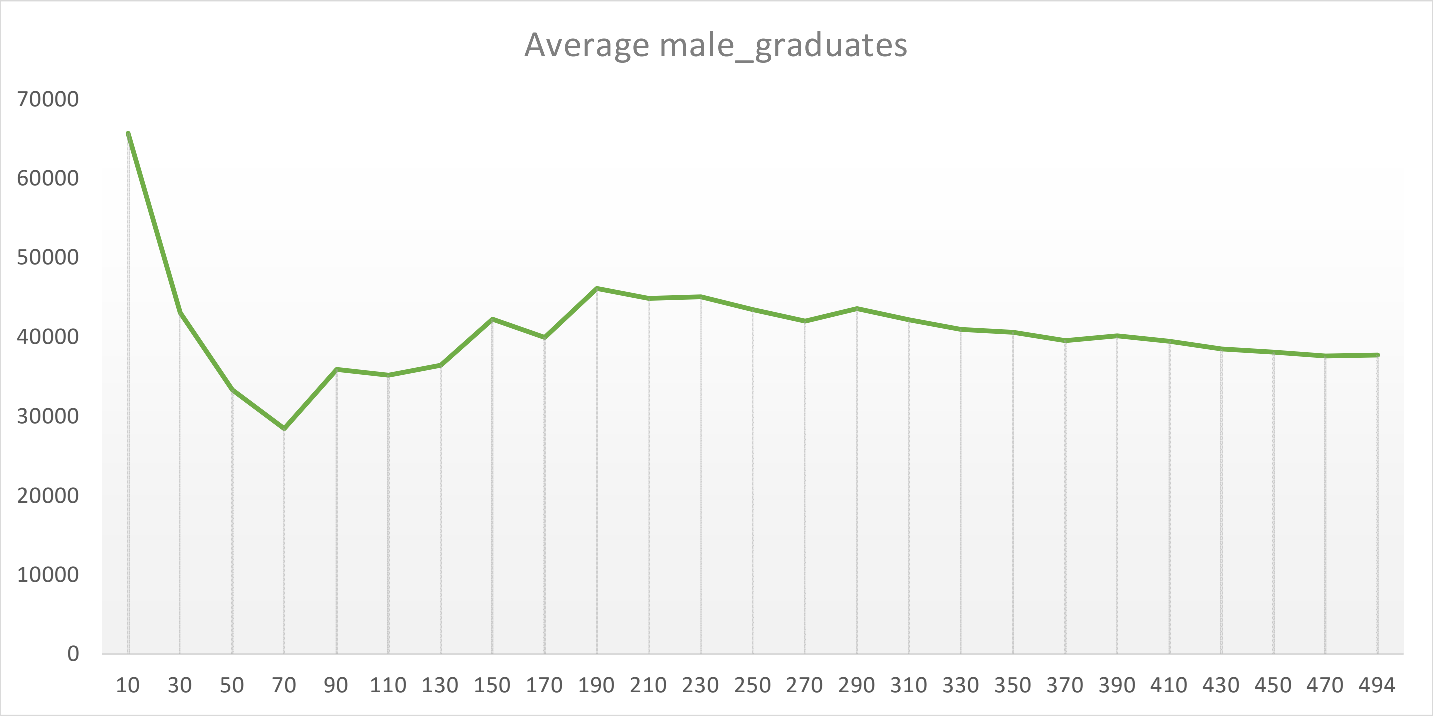
number of samples	Average population_male	Average population_fe	Average male_gradua	Average Female_graduat
10	467738	422560.6	65703.4	49668.3
30	286739.7667	260961.2667	43040.96667	33282
50	220414.32	201037.94	33301.72	24821.1
70	187602.0143	171184.7286	28454.58571	21062.1
90	227922.9222	207372.4667	35911.77778	26391.48889
110	219693.0455	199955.5818	35184.88182	25557.90909
130	222286.2615	204096.9923	36433.5	26897.34615
150	249553.0533	227356.3133	42270.37333	31744.87333
170	238262.5059	217084.4353	39907.08235	29899.39412
190	277606.0526	251481.1263	46104.81579	34793.64737
210	268621.6905	243346.5905	44872.73333	33750.66667
230	269901.213	244897.8348	45082.43043	33928.51304
250	260717.16	236741.708	43444.132	32607.292
270	253127.1778	229531.6	41999.66667	31543.94815
290	262345.0517	238211.0724	43583.2069	33025.84828
310	255959.1613	232851.5774	42188.75161	31962.58065
330	249505.0818	227006.5879	40942.49091	30976.20909
350	247780.7086	225573.0657	40565.88857	30710.76
370	240860.1676	219407.2459	39536.98919	29911.69189
390	242803.6256	221207.8154	40128.13077	30327.29487
410	238945.9268	217797.3585	39432.29024	29779.01707
430	233848.3023	213250.3372	38473.43256	29033.44651
450	235774.9111	214255.1356	38088.6	28705.24
470	233376.8574	212541.5574	37606.27021	28441.22128
494	234346.789	213765.5842	37715.56187	28486.79513



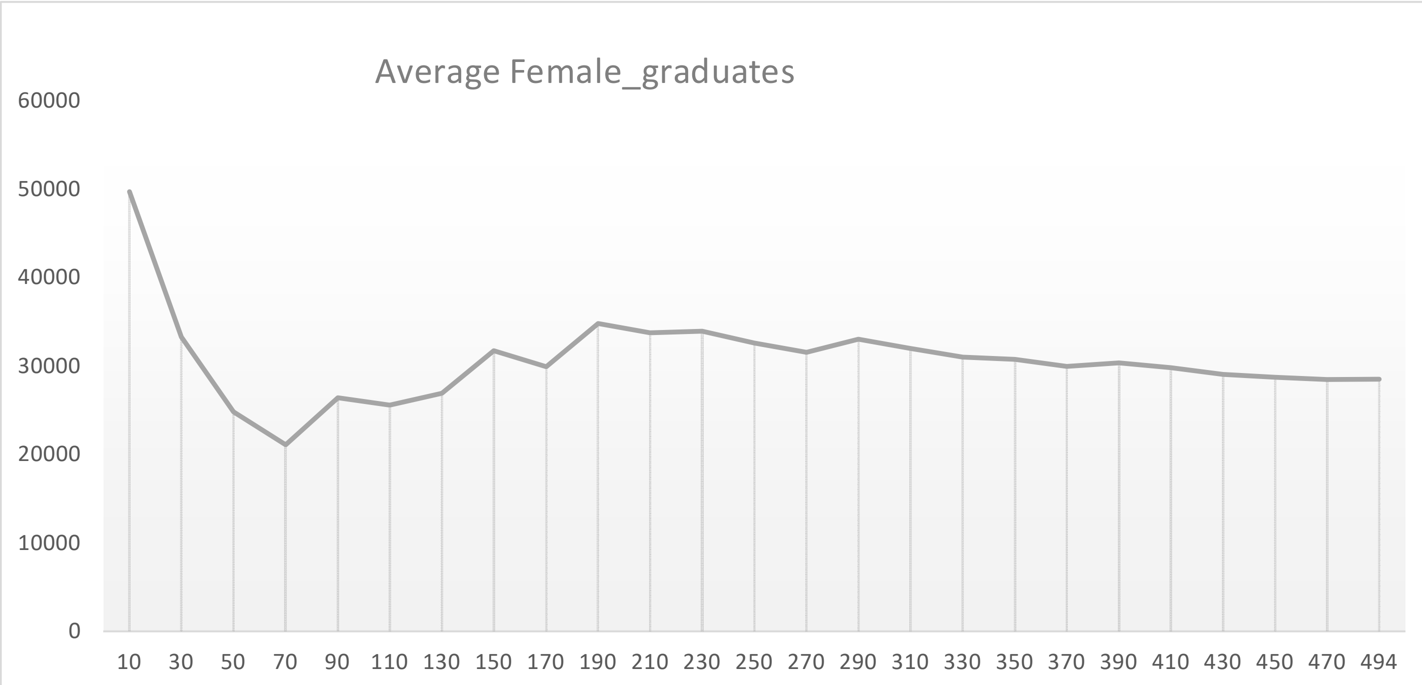
for **population male variable**, we recommended to take a sample size **more than or equal 430** because it's obvious that the mean started to be constant from that size



for **population female variable**, we recommended to take a sample size **more than or equal 430** because it's obvious that the mean started to be constant from that size



for **male graduate's variable**, we recommended to take a sample size **more than or equal 450** because it's obvious that the mean started to be constant from that size



for **female graduate's variable**, we recommended to take a sample size **more than or equal 450** because it's obvious that the mean started to be constant from that size

so, **as a result**: increasing the sample lead to a different value of mean until a specific number of sample (which depend on the variable) where the mean is independent from the sample size