Assignment-5

A. We have a dataset of marks given to the post graduate students (of non-engg. and engg. orientation) for a project review by different faculty members of a design department of an institute.

- From your analysis of the dataset, do you find any statistically significant difference between the marks given by different faculty to the students?

- What would you comment on the marks received by the students of engg. vs. non-engg. orientation?

- Is there any difference between the marks given by faculty to engg. vs. non-engg. students? (Along with validation in excel)

- Can you compare two faculty members at a time to find out for which faculty members you find statistically significant differences in grading the students? –

- Please also compute the Power (probability of avoiding a Type II error) of any of these comparisons and find out the sample size for which the Power >= 90%.

- Please share any other insights that you have for this analysis.

- What could be other unknown factors, if any, if considered could have made the analysis more meaningful?

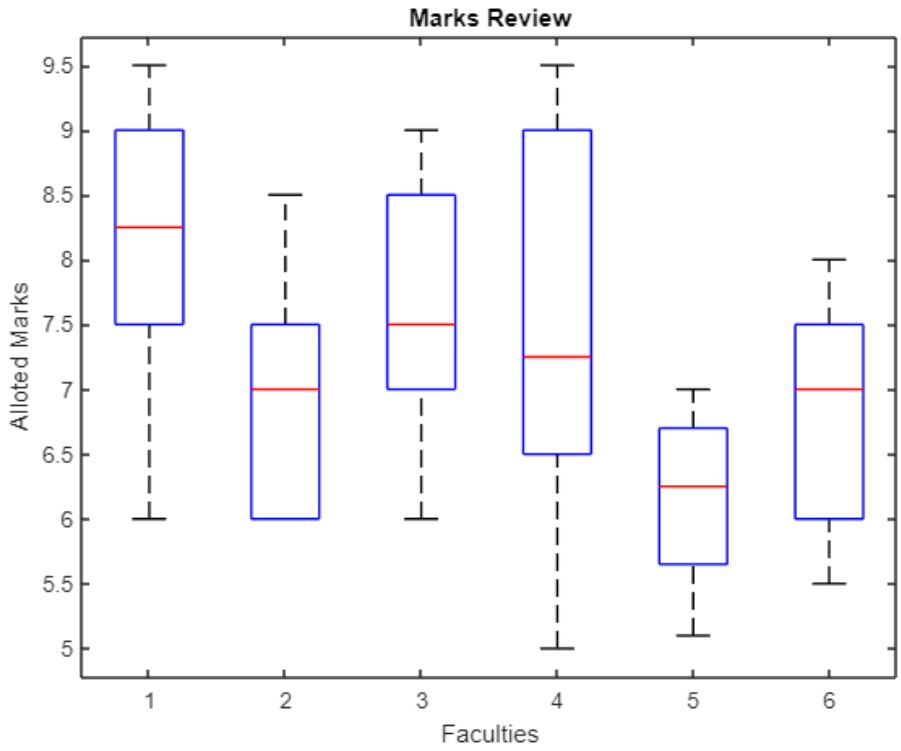
- Please visualize this data using box plots or other means. Please show the calculations/ upload the code and/or the ANOVA tables involved.

Data=

20×7 table

Student	F1	F2	F3	F4	F5	F6
{'Non-Engg.'}	8.5	7	7	7	6.5	8
{'Non-Engg.'}	8.5	8	7	9	6.6	6.5
{'Non-Engg.'}	9	7	9	9.5	6.9	7
{'Non-Engg.'}	8.5	7.5	8	7.5	5.8	7.5
{'Non-Engg.'}	7	7.5	7.5	6.5	5.8	6.5
{'Non-Engg.'}	6	7.5	7.5	5	5.1	5.5
{'Non-Engg.'}	9	7.5	8	8	6.5	7
{'Non-Engg.'}	7.5	7.5	7	6.5	6	6.5
{'Non-Engg.'}	7	6.8	6	9	6.5	6
{'Non-Engg.'}	8	6	8.5	7	5.4	6
{'Engineering'}	9.5	8.5	8.5	9.5	6.8	8
{'Engineering'}	7	6	8	7	5.4	6
{'Engineering'}	8.5	6	7	6.5	5.4	6
{'Engineering'}	7.5	6	7	6	5.5	6
{'Engineering'}	8	6	7.5	6.5	5.9	7.5
{'Engineering'}	8	7	7	6	5.8	7
{'Engineering'}	9	8.5	9	9	6.9	8
{'Engineering'}	9	7	9	9.5	6.9	7
{'Engineering'}	8	6	7	8	7	8
{'Engineering'}	9	8.5	8.5	8	6.5	7

```
%fetching data
data=xlsread('/MATLAB Drive/Published/reviewmarks.xlsx');
%plotting
boxplot(data);
title("Marks Review");
xlabel("Faculties");
ylabel("Alloted Marks");
```



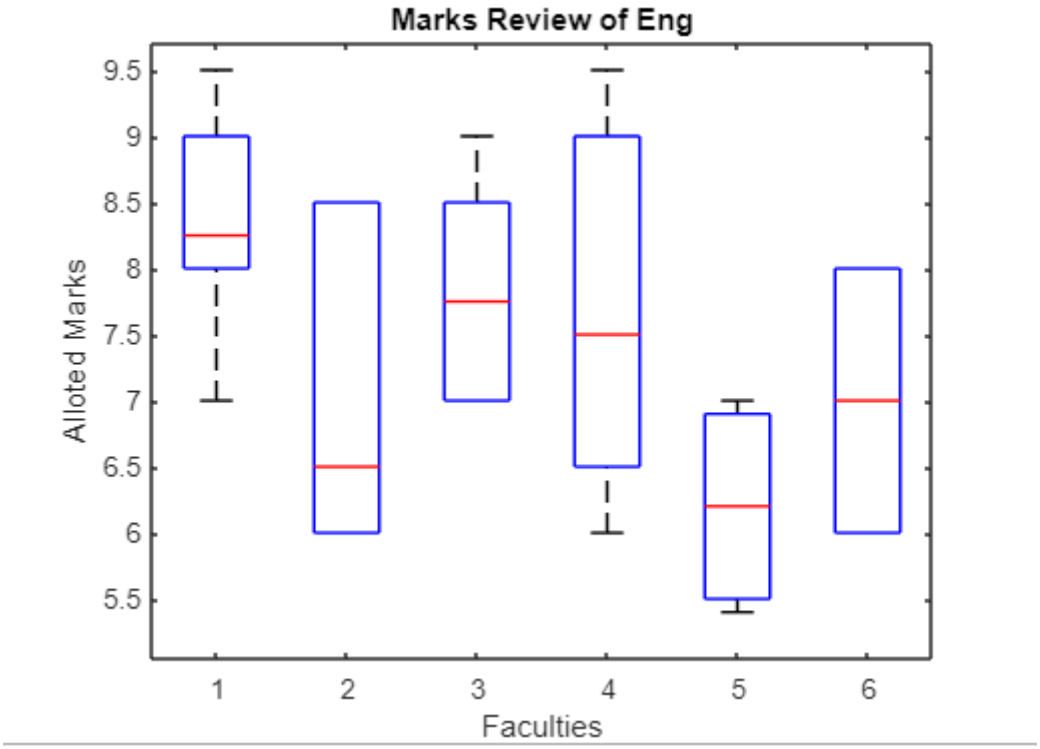
Excel Data

Student	F1	F2	F3	F4	F5	F6
Non-Engg.	8.5	7	7	7	6.5	8
Non-Engg.	8.5	8	7	9	6.6	6.5
Non-Engg.	9	7	9	9.5	6.9	7
Non-Engg.	8.5	7.5	8	7.5	5.8	7.5
Non-Engg.	7	7.5	7.5	6.5	5.8	6.5
Non-Engg.	6	7.5	7.5	5	5.1	5.5
Non-Engg.	9	7.5	8	8	6.5	7
Non-Engg.	7.5	7.5	7	6.5	6	6.5
Non-Engg.	7	6.8	6	9	6.5	6
Non-Engg.	8	6	8.5	7	5.4	6
Engineering	9.5	8.5	8.5	9.5	6.8	8
Engineering	7	6	8	7	5.4	6
Engineering	8.5	6	7	6.5	5.4	6
Engineering	7.5	6	7	6	5.5	6
Engineering	8	6	7.5	6.5	5.9	7.5
Engineering	8	7	7	6	5.8	7
Engineering	9	8.5	9	9	6.9	8
Engineering	9	7	9	9.5	6.9	7
Engineering	8	6	7	8	7	8
Engineering	9	8.5	8.5	8	6.5	7
Size	**20**	**20**	**20**	**20**	**20**	**20**
Mean	**8.125**	**7.09**	**7.7**	**7.55**	**6.16**	**6.85**

S.D.	**0.90139**	**0.884902611**	**0.84914818**	**1.35627198**	**0.615073808**	**0.796373359**
Variance	**0.8125**	**0.783052632**	**0.721052632**	**1.839473684**	**0.378315789**	**0.634210526**

What would you comment on the marks received by the students of engg. vs. non-engg. orientation? (Along with validation in excel)

```
Eng=data(11:20,1:6);
boxplot(Eng);
title("Marks Review of Eng");
xlabel("Faculties");
ylabel("Alloted Marks");
```

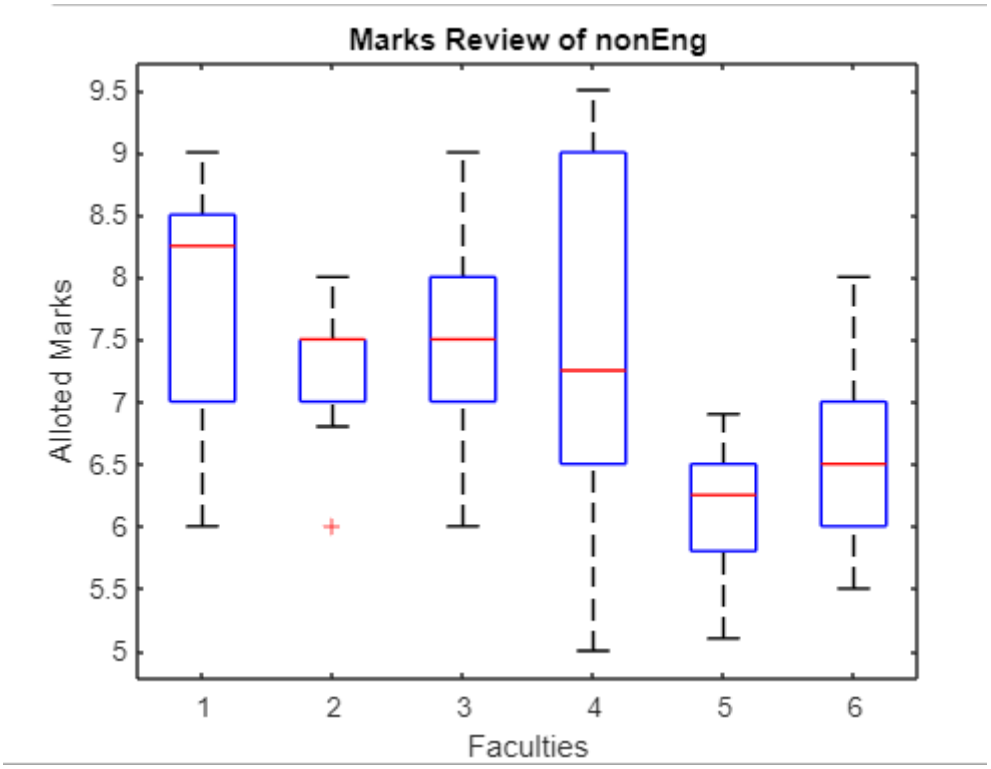


Excel Data

Engineering	9.5	8.5	8.5	9.5	6.8	8
Engineering	7	6	8	7	5.4	6
Engineering	8.5	6	7	6.5	5.4	6
Engineering	7.5	6	7	6	5.5	6
Engineering	8	6	7.5	6.5	5.9	7.5
Engineering	8	7	7	6	5.8	7
Engineering	9	8.5	9	9	6.9	8
Engineering	9	7	9	9.5	6.9	7
Engineering	8	6	7	8	7	8
Engineering	9	8.5	8.5	8	6.5	7
Size	**10**	**10**	**10**	**10**	**10**	**10**

Mean	8.35	6.95	7.85	7.6	6.21	7.05
S.D.	0.78351062	1.14139291	0.85146932	1.39044357	0.67404253	0.831665
Variance	0.61388889	1.30277778	0.725	1.93333333	0.45433333	0.69166667

```
nonEng=data(1:10,1:6);
boxplot(nonEng);
title("Marks Review of nonEng");
xlabel("Faculties");
ylabel("Alloted Marks");
```



Excel Data

Student	F1	F2	F3	F4	F5	F6
Non-Engg.	8.5	7	7	7	6.5	8
Non-Engg.	8.5	8	7	9	6.6	6.5
Non-Engg.	9	7	9	9.5	6.9	7
Non-Engg.	8.5	7.5	8	7.5	5.8	7.5
Non-Engg.	7	7.5	7.5	6.5	5.8	6.5
Non-Engg.	6	7.5	7.5	5	5.1	5.5
Non-Engg.	9	7.5	8	8	6.5	7
Non-Engg.	7.5	7.5	7	6.5	6	6.5
Non-Engg.	7	6.8	6	9	6.5	6
Non-Engg.	8	6	8.5	7	5.4	6
Size	**10**	**10**	**10**	**10**	**10**	**10**
Mean	7.9	7.23	7.55	7.5	6.11	6.65

S.D.	0.9944289 3	0.5538752 4	0.8644201 7	1.3944333 8	0.5820462	0.7472170 6
Variance	0.9888888 9	0.3067777 8	0.7472222 2	1.9444444 4	0.3387777 8	0.5583333 3

Anova Table:

Figure 1: Two-way ANOVA

ANOVA Table

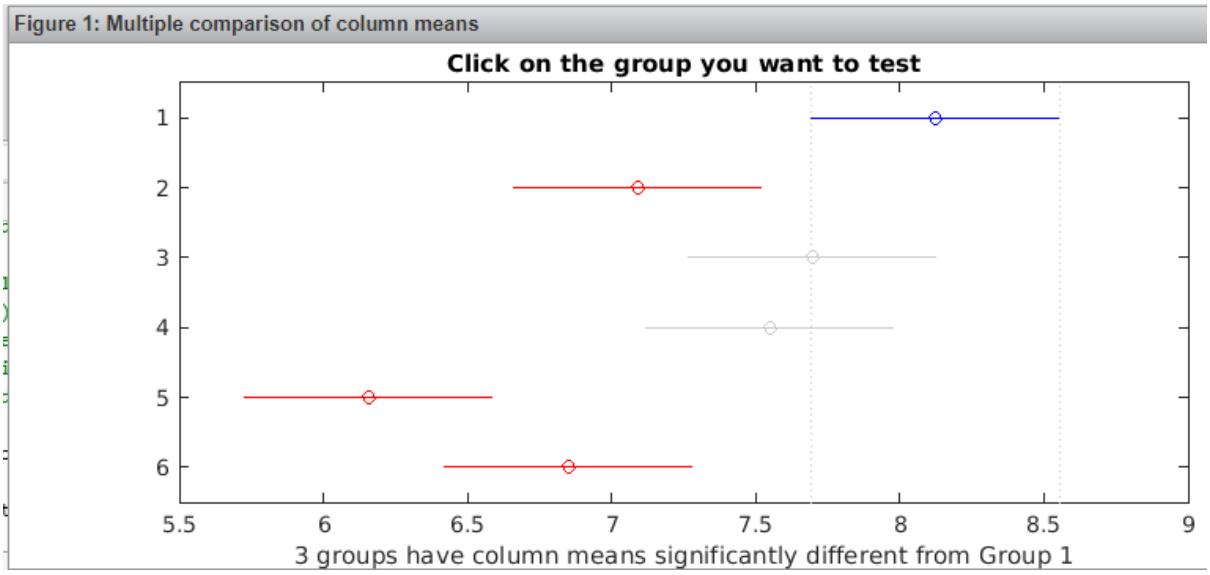
Source	SS	df	MS	F	Prob>F
Columns	48.634	5	9.72688	11.01	0
Rows	0.954	1	0.95408	1.08	0.3011
Interaction	1.8	5	0.36008	0.41	0.8427
Error	95.449	108	0.88379		
Total	146.838	119			

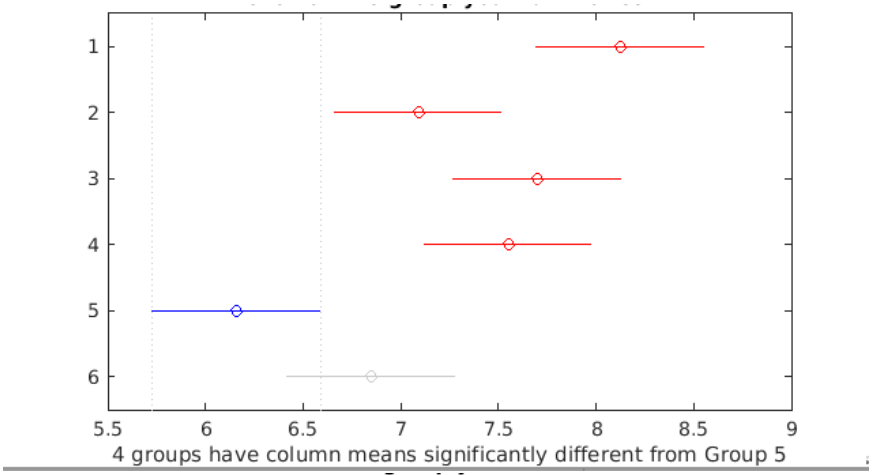
```
struct with fields:

    source: 'anova2'
    sigmasq: 0.8838
  colmeans: [8.1250 7.0900 7.7000 7.5500 6.1600 6.8500]
      coln: 20
  rowmeans: [7.1567 7.3350]
      rown: 60
     inter: 1
      pval: 0.8427
        df: 108
```

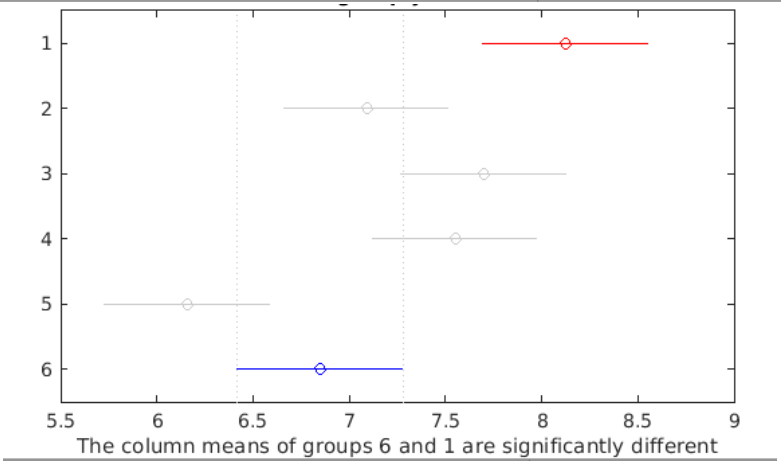
From the above table we can see that the p-value of interaction between faculty and students is 0.842 which is highly different than alpha level 0.05. Thus, we can say that there hasn't been a lot of statistical differences in the grading of engg and non-engg streams.

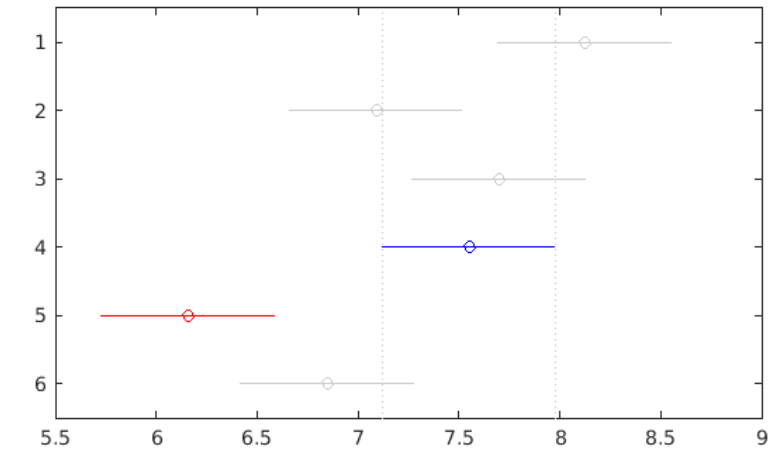
The p-value (of rows) of students' background is 0.301 which is also not statistically significant at alpha level 0.05. The p-value (of columns) of faculty is 0 which means there are statistically genuine differences in the grades given by faculty as it is significant to alpha level 0.05.

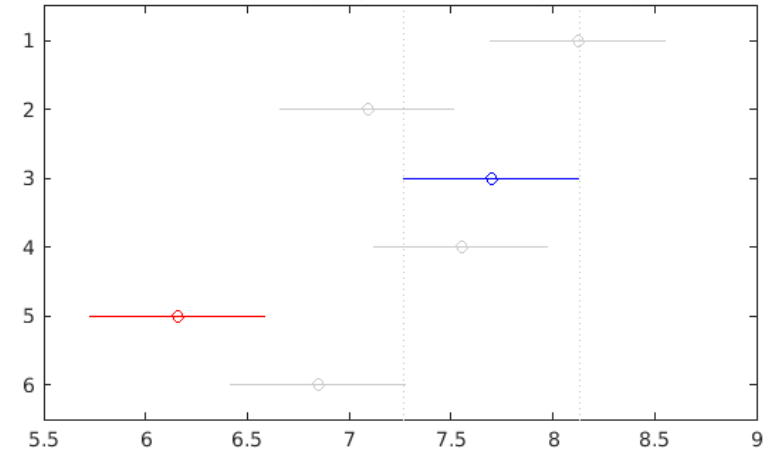
Using *multcompare* to see the statistically significant differences in grades given by faculty:

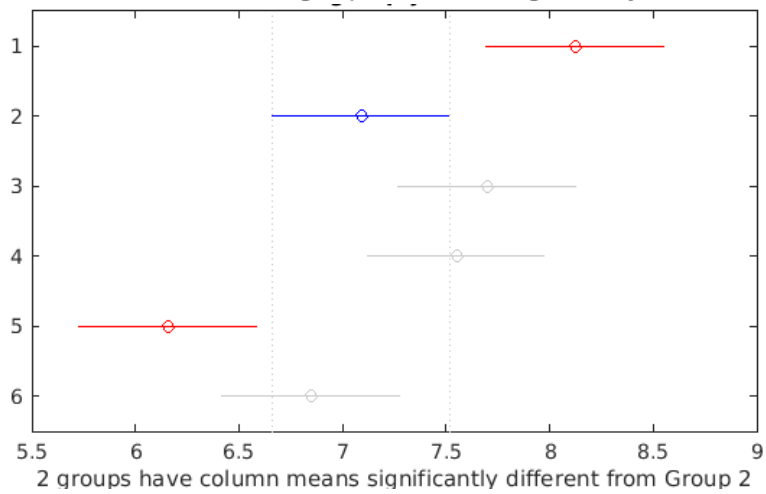












From the above 6 plots, it is seen that significant variations are there in F1-F5, F1-F6, F1-F2, F2-F5, F3-F5 and F4-F5. Now, seeing two faculties 1 and 5 for Type II error.

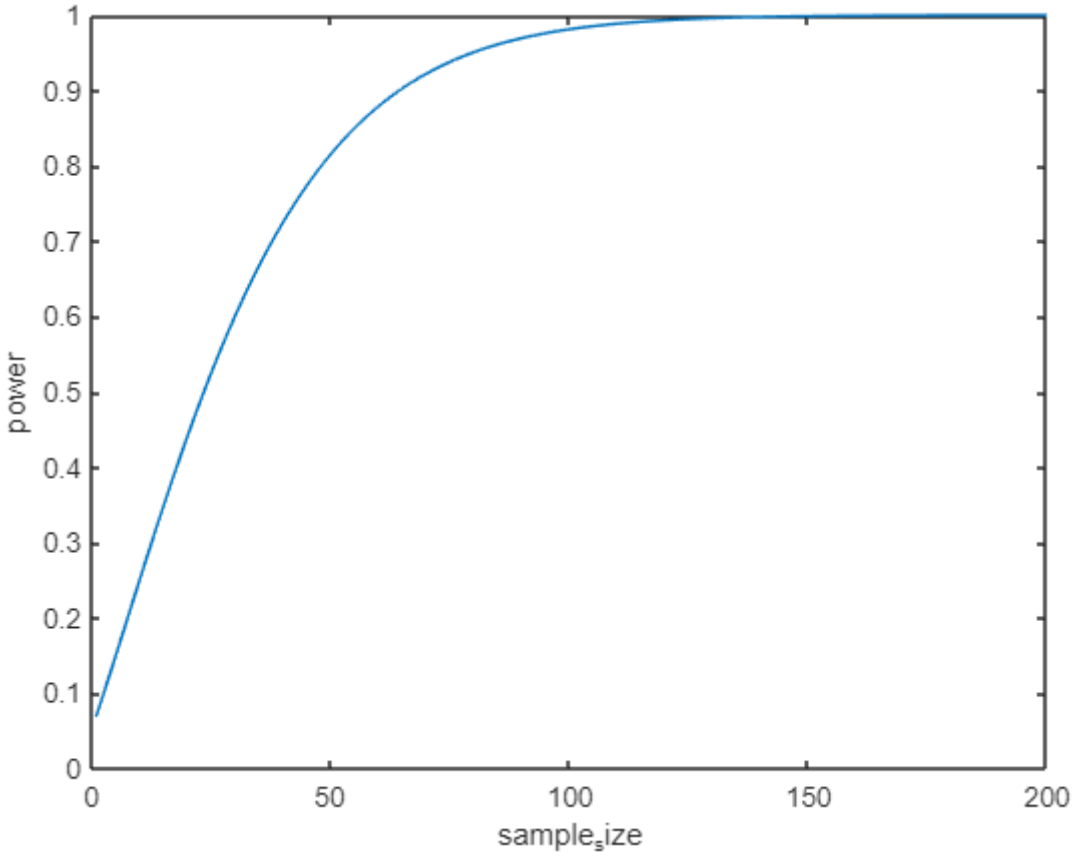
Power and sample size with t and z:

Noofsamples is at power=90%

```
pow=sampsizepwr('t',[0 4.88],1.97,[],20);

noofsamples= sampsizepwr('z',[0 4.88],1.97, 0.9,[]);
n=1:200;
pwr=sampsizepwr("z",[0 4.88],1.97, [],n);
plot(n,pwr);
xlabel('sample_size');
ylabel('power');
```

```
pow =

    0.4030

>> noofsamples

noofsamples =

    65
```

n=1 to 200

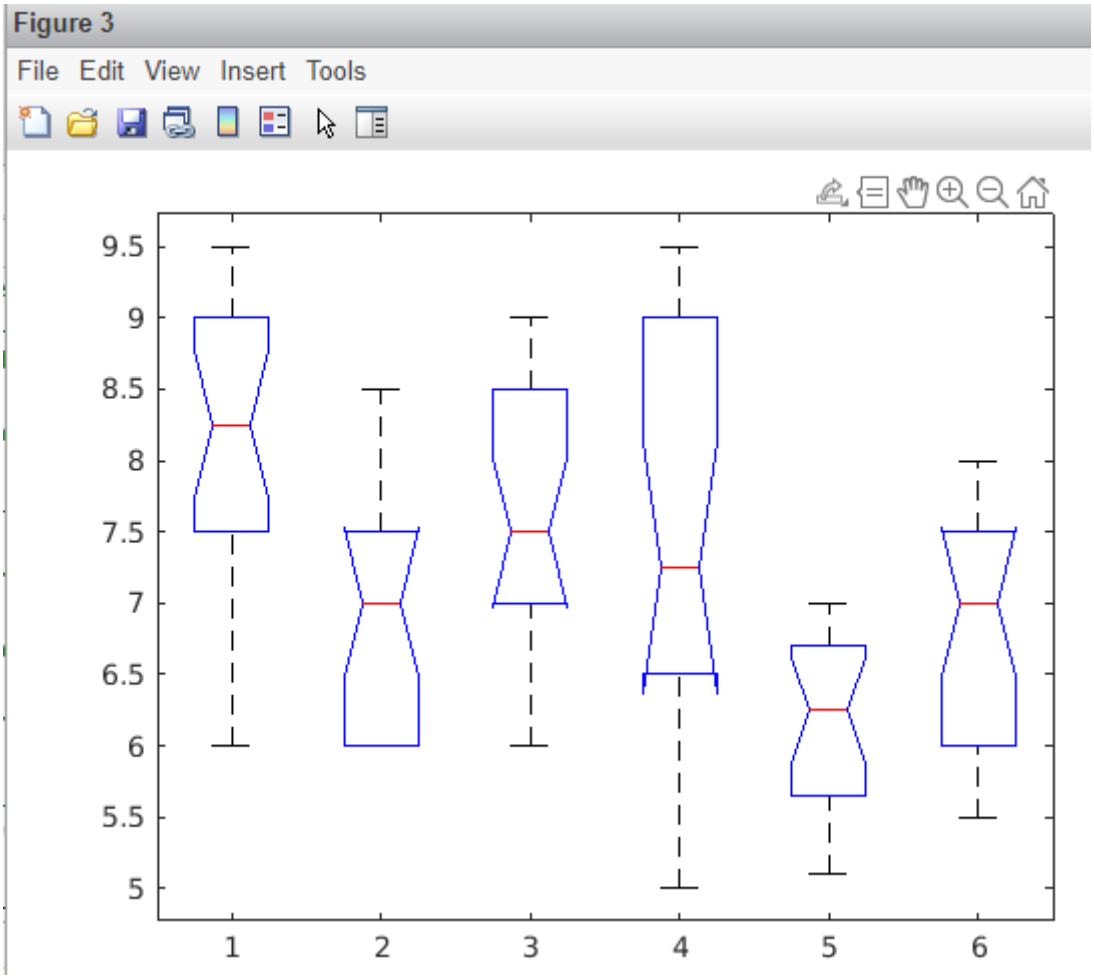


Calculated power= 40.3% and 65 no of samples should be increased for power of 90%. When normality functions are not valid, Kruskal-walls method is used:

Figure 2: Kruskal-Wallis One-way ANOVA

Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	51561.8	5	10312.4	43.25	3.28439e-08
Error	90299.7	114	792.1		
Total	141861.5	119			




```
kw_table =

    3.2844e-08
```

Chi-Sq value is 43.25; Prob>Chi-sq value is very very less.

B. Can you think of other real-world applications of the Statistical Methods learnt in the class? Please outline any five of them.

1) Data Mining and Data Compression: Data mining is performed using functions to find irregularities or inconsistency within data

   Programs like WinZip are able to reduce the size of file by applying data compression techniques opted from data science. The data compression uses statistical algorithms to compress the data.

2) Weather Forecasts:
   The working of weather forecast models giving information ahead of time. These computer models are built using statistical methods that compare prior weather conditions with current weather to predict future weather.

3) Clinical Trial Design: Key statistical concepts are essential for the design and construction of clinical trials. Principles, Trials, and Designs presents a timely and view of the central statistical concepts used to build clinical trials that obtain the best results. The modern approaches of statistics are vital for understanding, creating, and evaluating data obtained throughout the various stages of clinical trial design and analysis. A variety of statistical concepts and principles such as longitudinal data, missing data, covariates, biased-coin randomization, repeated measurements, and simple randomization is used in different phases of trials.

4) Drug Testing: Statistics is an important tool in pharmacological research that is used to summarize experimental data in terms of central tendency (mean or median) and variance (standard error of the mean, confidence interval or range) but more importantly it enables us to conduct hypothesis testing. This is great importance when we are attempting to determine whether the pharmacological effect of one drug is superior to another.