MTL390: Statistical Methods
Assignment 1

Rishit Sanmukhani
(2013CS10255)

Frequency Table

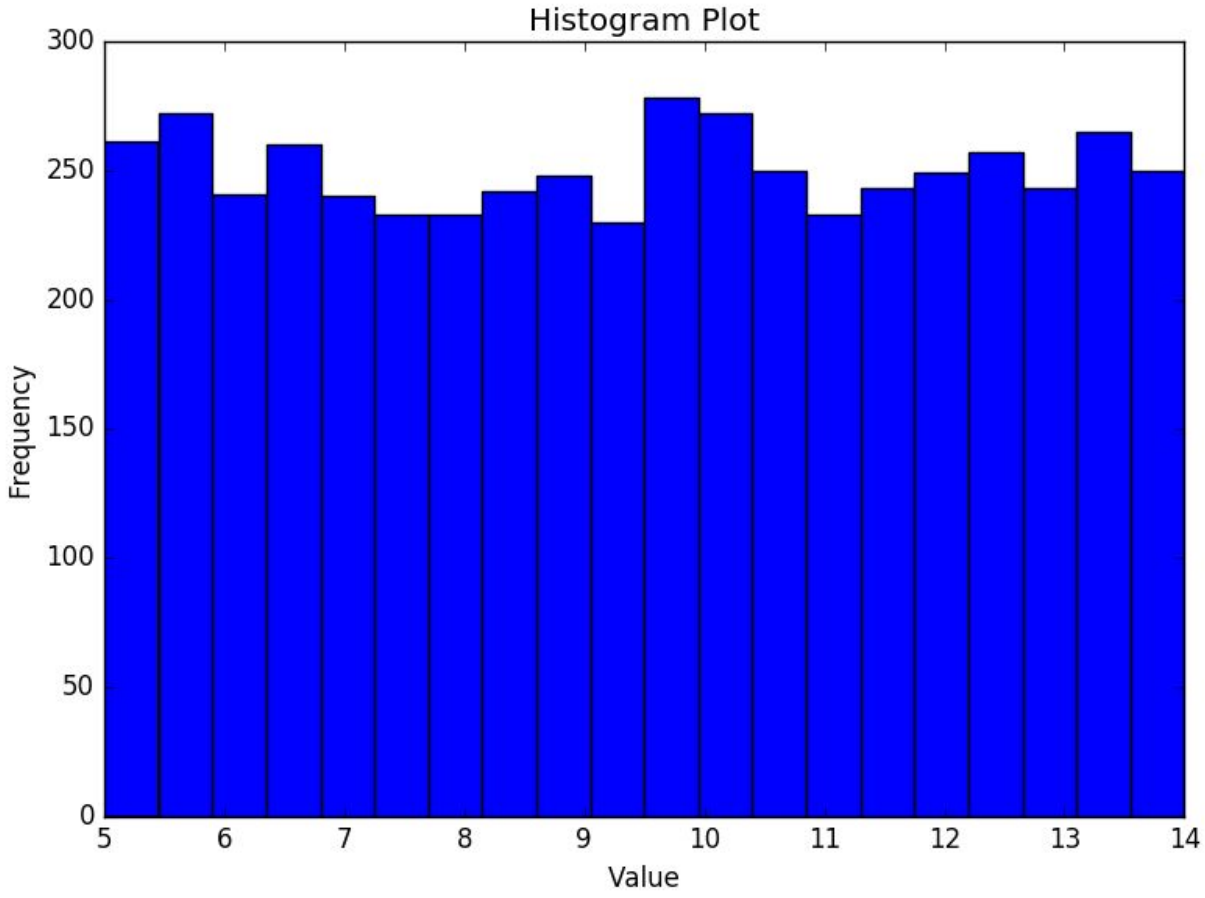
Following is the frequency table with 20 intervals

Interval	Frequency
5.00365259443 - 5.4534048585	261
5.4534048585 - 5.90315712258	272
5.90315712258 - 6.35290938665	241
6.35290938665 - 6.80266165072	260
6.80266165072 - 7.2524139148	240
7.2524139148 - 7.70216617887	233
7.70216617887 - 8.15191844294	233
8.15191844294 - 8.60167070702	242
8.60167070702 - 9.05142297109	248
9.05142297109 - 9.50117523517	230
9.50117523517 - 9.95092749924	278
9.95092749924 - 10.4006797633	272
10.4006797633 - 10.8504320274	250
10.8504320274 - 11.3001842915	233
11.3001842915 - 11.7499365555	243
11.7499365555 - 12.1996888196	249
12.1996888196 - 12.6494410837	257
12.6494410837 - 13.0991933478	243
13.0991933478 - 13.5489456118	265
13.5489456118 - 13.9986978759	250

Histogram Plot

Note: The histogram plot changes as we change the number of bins in the frequency table. If the number of bins is too large, the histogram plot doesn't provide much information. Hence we choose nbins = 20 (neither large, nor small)

Following is the histogram corresponding the above frequency table (Generated using python)

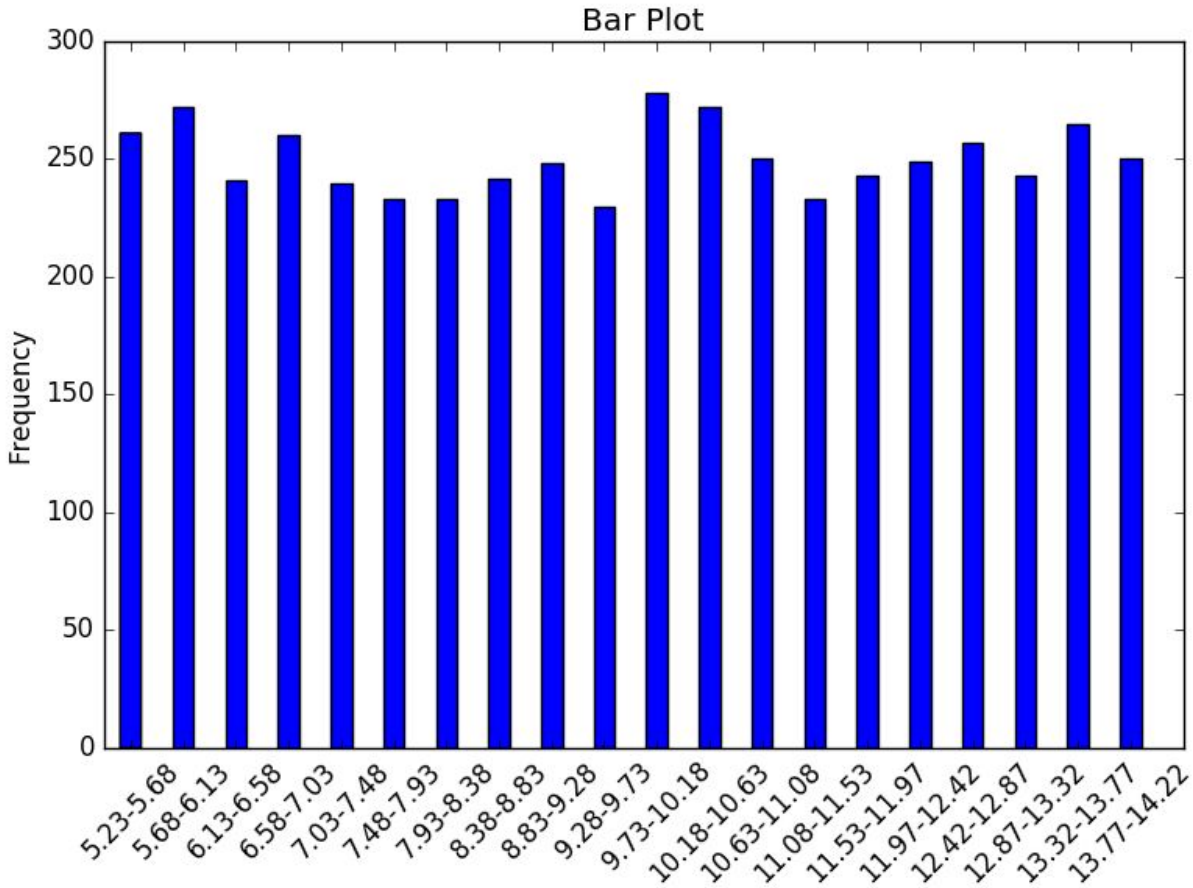


Bar Plot

Note:

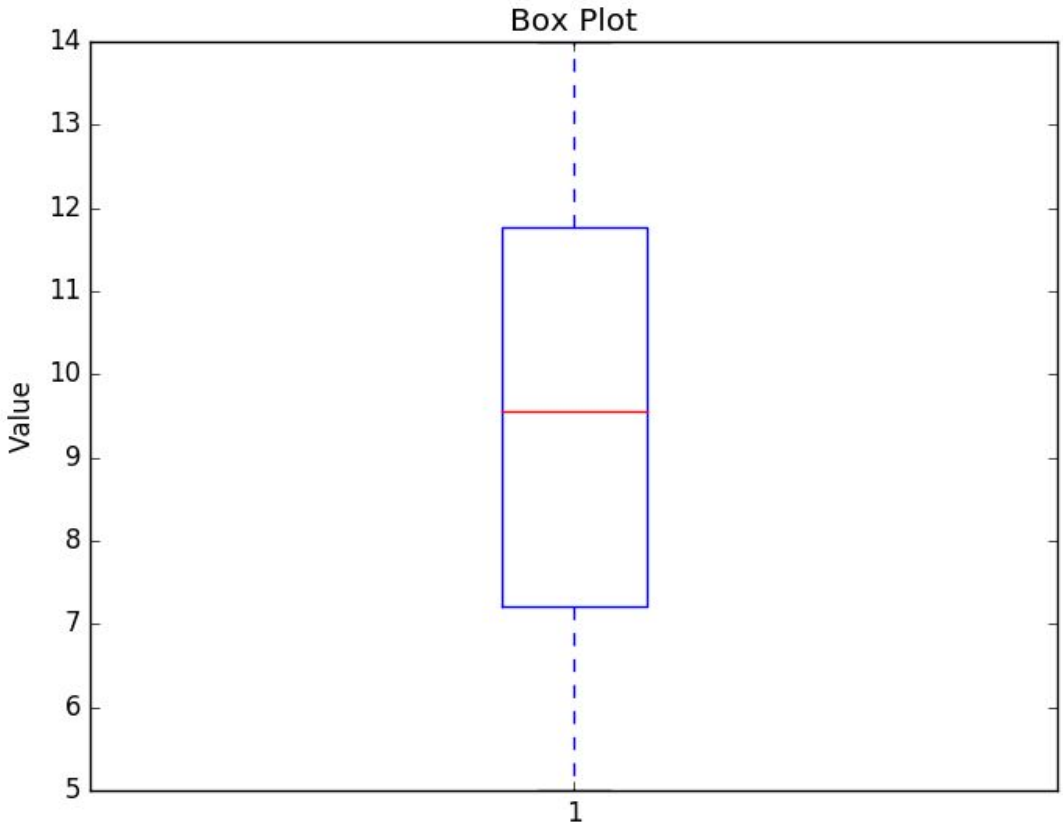
- The bar graph is not much relevant in case of “continuous ungrouped data”. Hence I have takes the similar intervals of histogram to plot the bar graph.
- If we choose to discretize the values, then each data point will have frequency as 1. This isn’t much of use. So I have neglected that plot.

Following is the bar plot corresponding to the above frequency table (generated using python)



Box Plot

Following is the box plot (generated using python). This plot gives us information about the percentile values. It clearly shows the interquartile range, 25%ile, 75%ile.



Measures of Data

Mean	9.50327610136
Median	9.56525912927
Mode	9.90095502545
Coefficient of variation	0.274895816611
Coefficient of skewness	-0.0118070300789
Coefficient of kurtosis	-1.20817393245
Inter-quartile range	4.57103441383

Mode calculation:

Preceding class = 9.05142297109 - 9.50117523517 (230)
Model class = 9.50117523517 - 9.95092749924 (278)
Succeeding class = 9.95092749924 - 10.4006797633 (272)

f_m = 278
f_{m-1} = 230
f_{m+1} = 272
L = 9.50117523517
W = 0.44975226407

mode = 9.90095502545

Now,
3*median - 2*mean = 3*9.56525912927 - 2*9.50327610136 = 9.68922518509
Mode = 9.90095502545

Thus, the relation is not satisfied.

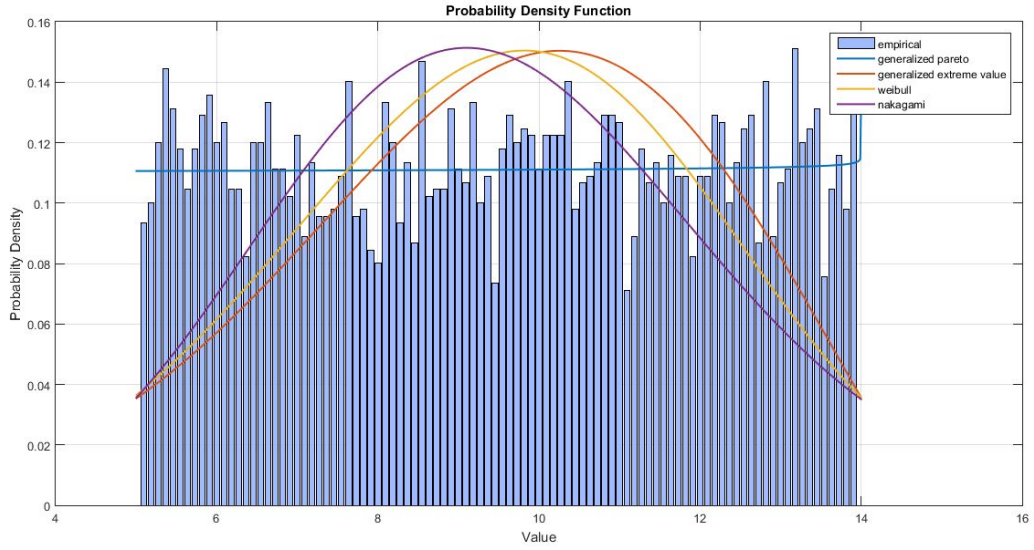
Note: If we go according to the distribution (i.e uniform), we won't have any single mode. Mode of uniform distribution is not defined. But here we solve without assuming any distribution and strictly from the given data.

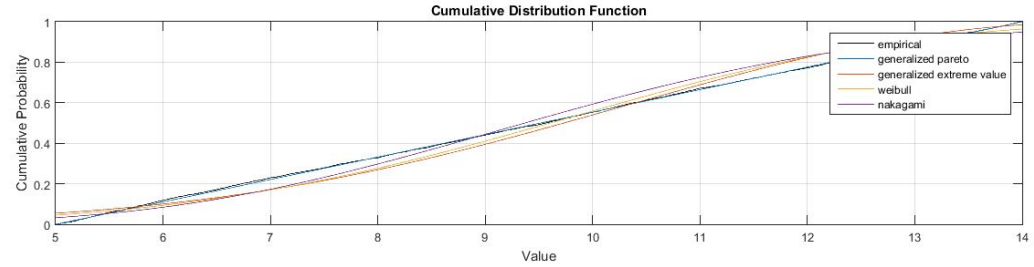
Best-fit distribution

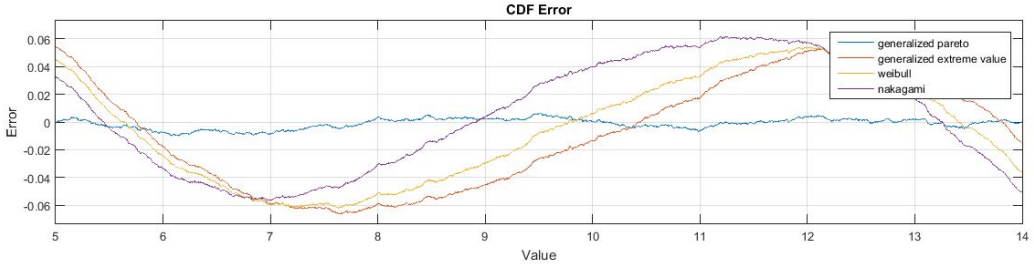
I have used "**allfitdist.m**" package of MATLAB.

List of continuous distributions it will try to fit are:
Beta, Birnbaum-Saunders, Exponential, Extreme value, Gamma, Generalized extreme value, Generalized Pareto, Inverse Gaussian, Logistic, Log-logistic, Lognormal, Nakagami, Normal, Rayleigh, Rician, t location-scale, Weibull

Following are the result:







As evident from the above figure, the distribution resembles “Generalized Pareto” distribution.
The parameters are:
k = -1.00518947041166
sigma = 9.04172480280973
theta = 5.00365259443000

But placing the above parameters in the generalized pareto equation, we get:
y = f(x) = 1/sigma = 1/9.04172480280973
This is nothing but a uniform distribution.

Thus, the best fit distribution is **uniform distribution i.e U (5, 14)**