MTL390 Assignment 1

Pratyush Pandey (2017EE10938)

12^{th} March, 2021

Contents

1 Frequency Histogram

The following histogram is generated using Python (code attached in Appendix ??). We choose number of bins to be 20 - for if the number of bins is too large, the histogram plot doesn't provide much information.

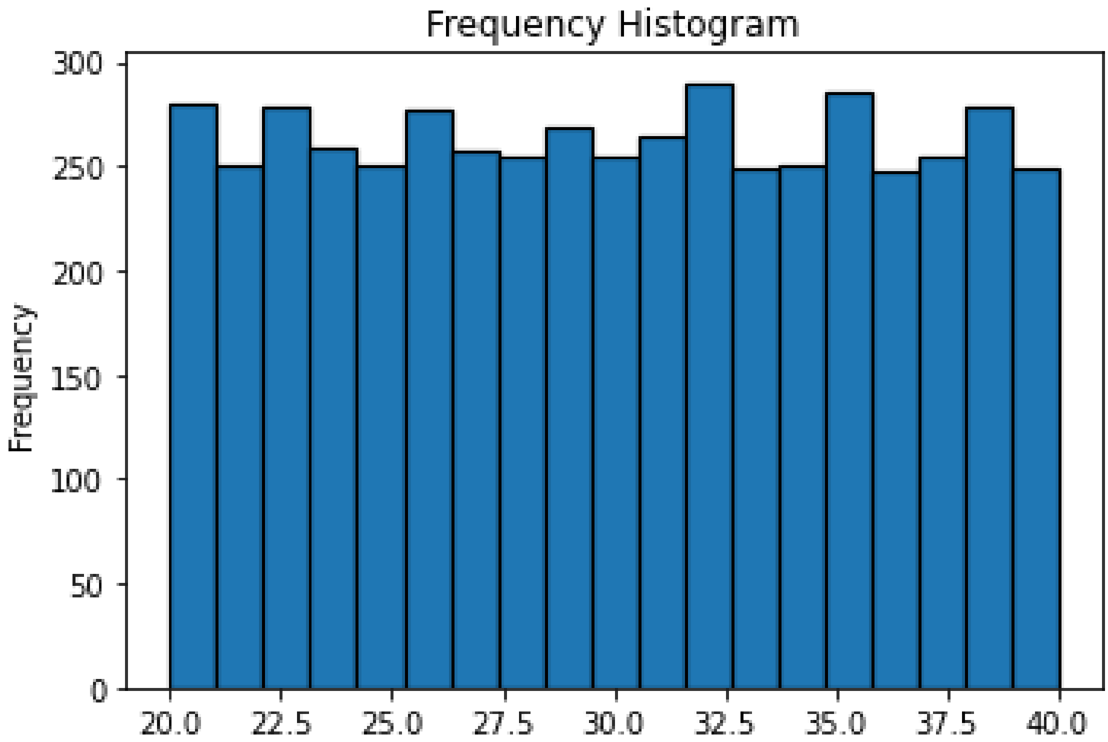


Figure 1: Histogram Plot

2 Bar Plot

- Since we have a continuously distributed data, the bar graph isn't very relevant - specially since the data is ungrouped.
- We take similar intervals as histogram, ie 20 to plot a bar graph.
- If we choose to discretize the values, each data point will have a frequency as 1 (or maybe 2 after rounding). This plot will be very dense and not of much use, hence I am choosing to ignore that plot.

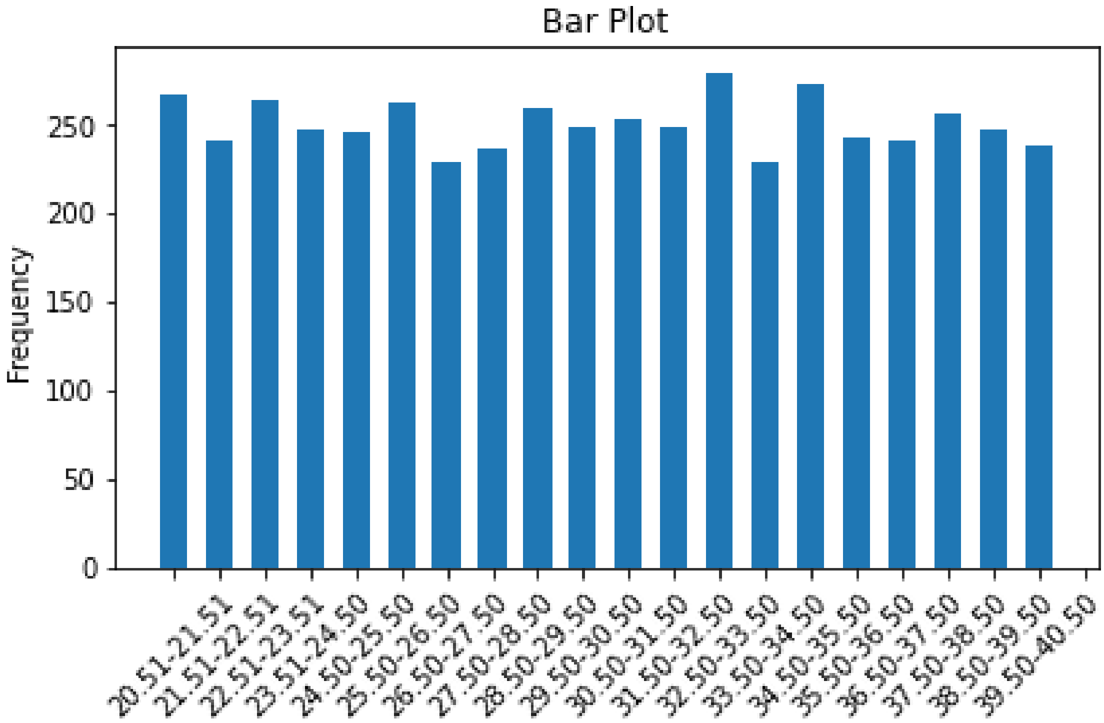


Figure 2: Bar Plot

3 Box Plot

This plot is generated using Python. The plot gives us information about the percentile values. It also clealrly shows the interquartile range at 25% percentile and 75% percentile.

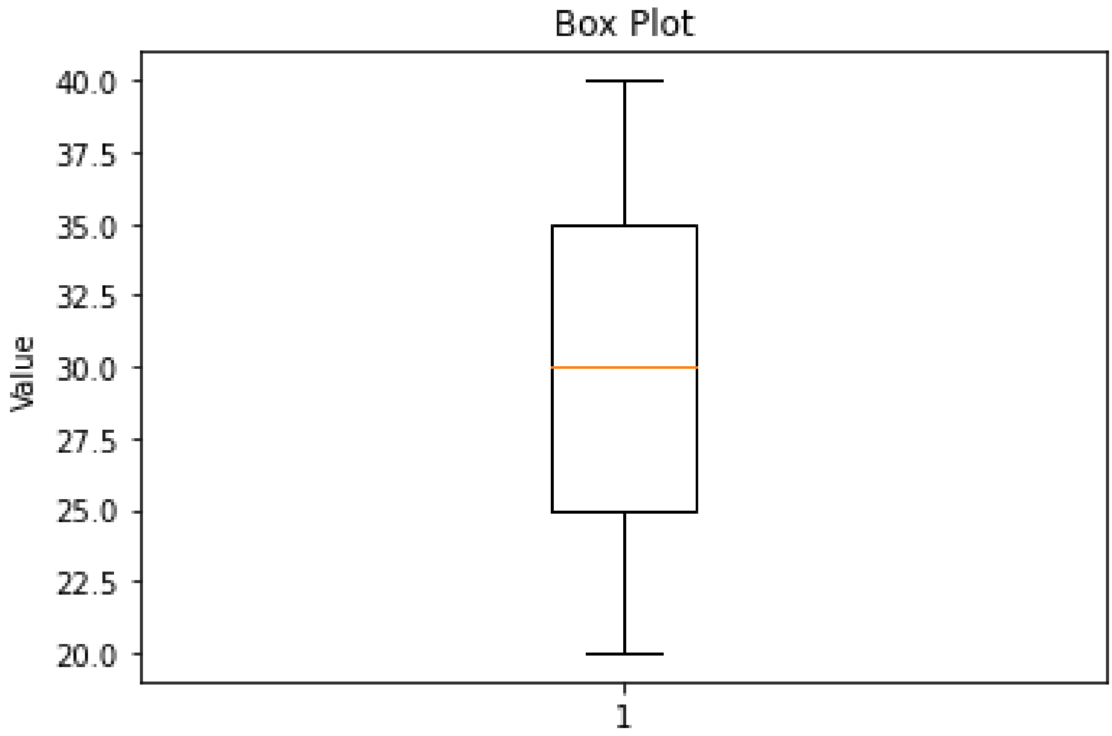


Figure 3: Bar Plot

4 Measures of central tendencies

Measures of Data	Value
Mean	29.961941007654183
Median	30.03107342636215
Mode	30.0123912062
Coefficient of Variation	0.19244327654487922
Coefficient of Skewness	-0.005006017850751467
Coefficient of Kurtosis	-1.1958813543047515
Inter-quartile range	9.964354203548297

Table 1: Measures of central tendencies

4.1 Mode Calculation

Modal Class: 30.002513937652147 - 31.00209352653477
Preceding Class: 29.00293434876952 - 30.002513937652147
Succeeding Class: 31.00209352653477 - 32.0016731154174

$f_m = 253$
$f_m - 1 = 248$
$f_m + 1 = 248$
L = 30.002513937652147
W = 0.99957958888

$$M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} * W = 30.002513937652147 + 0.00987726 = \mathbf{30.0123912062}$$

We use table 5 to find the modal intervals and frequencies.

Now, Mode = 3*Median - 2*Mean = 3*30.03107342636215 - 2*29.961941007654183 = **30.1693382638**

Note: Thus, the relation is not satisfied. 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.

5 Best Fits Distributions

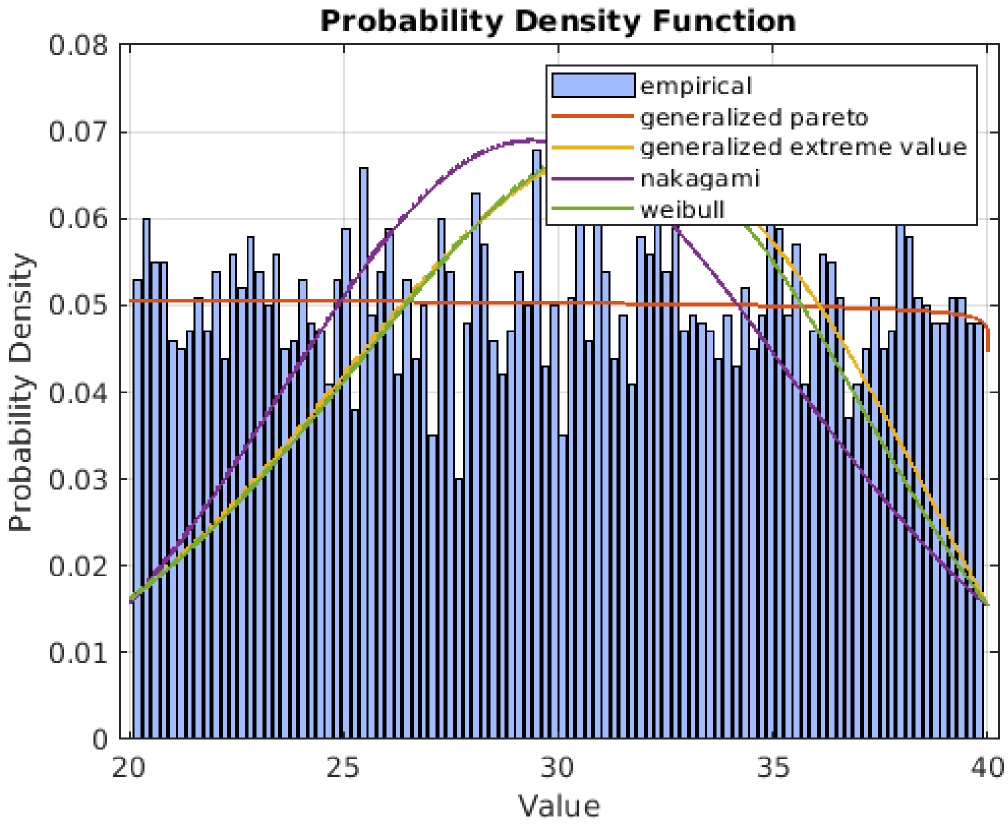
I have used "**allfitdist.m**" package of MATLAB. It will try to fit continuous distributions: Beta, Birnbaum-Saunders, Exponential, Extreme value, Gamma, Generalized extreme value, Generalized Pareto, Inverse Gaussian, Logistic, Log-logistic, Lognormal, Nakagami, Normal, Rayleigh, Rician, t location-scale, Weibull

The following are the results of the simulation:

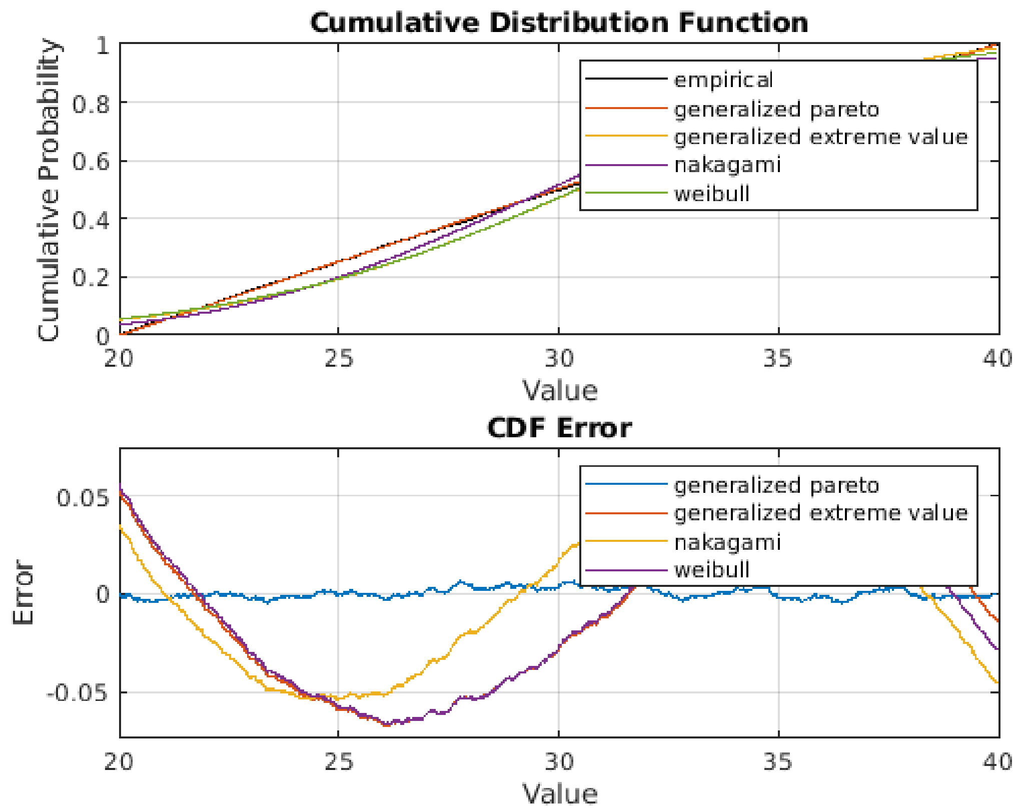
```
data = csvread('2017EE10938.csv',1,1);
data
```

```
data = 5000x1
   32.2681
   22.9795
   21.5283
   24.7889
   32.1239
   25.4768
   25.7775
   37.8503
   35.5771
   36.6851
```

```
[D, PD] = allfitdist(data, 'PDF')
```




```
[D, PD] = allfitdist(data, 'CDF')
```



Fields	DistName	NLogL	BIC	AIC	AICc	ParamNames	ParamDe...	Params	Paramci	ParamCov	Support
1	'generalized pareto'	1.4976e+04	2.9978e+04	2.9959e+04	2.9959e+04	1×3 cell	1×3 cell	[-0.9903,19.7973,20.0067]	[-1.0181,19.2491,20.0067;-0.9625,20.3611,20.0067]	[0.0002,-0.0040,0;-0.0040,0.0805,0;0,0,0]	1×1 struct
2	'generalized extreme value'	1.5725e+04	3.1476e+04	3.1456e+04	3.1456e+04	1×3 cell	1×3 cell	[-0.4403,6.0702,28.3747]	[-0.4711,5.9091,28.1771;-0.4096,6.2357,28.5723]	[2.4612e-04,-0.0010,-0.0008;-9.7560e-04,0.0069,0.0004;-8.1014e-04,0.0004,0.0102]	1×1 struct
3	'nakagami'	1.5842e+04	3.1702e+04	3.1689e+04	3.1689e+04	1×2 cell	1×2 cell	[6.8530,930.9644]	[6.5956,921.1591;7.1204,940.8740]	[0.0179,0.0000;0.0000,25.2941]	1×1 struct
4	'weibull'	1.5845e+04	3.1708e+04	3.1695e+04	3.1695e+04	1×2 cell	1×2 cell	[32.3537,5.9416]	[32.1946,5.8125;32.5135,6.0736]	[0.0066,0.0017;0.0017,0.0044]	1×1 struct
5	'rician'	1.5853e+04	3.1724e+04	3.1711e+04	3.1711e+04	1×2 cell	1×2 cell	[29.3782,5.8260]	[29.2133,5.7105;29.5431,5.9439]	[0.0071,-0.0007;-0.0007,0.0035]	1×1 struct
6	'normal'	1.5855e+04	3.1726e+04	3.1713e+04	3.1713e+04	1×2 cell	1×2 cell	[29.9619,5.7666]	[29.8021,5.6557;30.1218,5.8819]	[0.0067,-0.0000;-0.0000,0.0033]	1×1 struct
7	'tlocationscale'	1.5855e+04	3.1735e+04	3.1715e+04	3.1715e+04	1×3 cell	1×3 cell	[29.9623,5.7661,4.2957e+06]	[-Inf,-Inf,-Inf;Inf,Inf,Inf]	[NaN,NaN,NaN;NaN,NaN,NaN;NaN,NaN,NaN]	1×1 struct
8	'gamma'	1.5863e+04	3.1743e+04	3.1730e+04	3.1730e+04	1×2 cell	1×2 cell	[26.2298,1.1423]	[25.2277,1.0982;27.2716,1.1881]	[0.2717,-0.0118;-0.0118,0.0005]	1×1 struct
9	'birnbaumsaunders'	1.5888e+04	3.1792e+04	3.1779e+04	3.1779e+04	1×2 cell	1×2 cell	[29.3833,0.1984]	[29.2225,0.1945;29.5441,0.2023]	[0.0067,-5.3528e-09;-0.0000,3.9376e-06]	1×1 struct
10	'inverse gaussian'	1.5888e+04	3.1794e+04	3.1781e+04	3.1781e+04	1×2 cell	1×2 cell	[29.9619,753.5083]	[29.7963,723.9713;30.1275,783.0452]	[0.0071,0.0000;0.0000,227.1098]	1×1 struct
11	'lognormal'	1.5897e+04	3.1810e+04	3.1797e+04	3.1797e+04	1×2 cell	1×2 cell	[3.3807,0.1978]	[3.3753,0.1940;3.3862,0.2018]	[7.8285e-06,-2.9112e-20;-2.9112e-20,3.9154e-06]	1×1 struct
12	'extreme value'	1.6006e+04	3.2029e+04	3.2016e+04	3.2016e+04	1×2 cell	1×2 cell	[32.8298,5.2117]	[32.6769,5.1010;32.9828,5.3248]	[0.0061,-0.0015;-0.0015,0.0033]	1×1 struct
13	'logistic'	1.6080e+04	3.2176e+04	3.2163e+04	3.2163e+04	1×2 cell	1×2 cell	[29.9685,3.5038]	[29.7948,3.4263;30.1422,3.5830]	[0.0079,-0.0000;-0.0000,0.0016]	1×1 struct
14	'loglogistic'	1.6107e+04	3.2231e+04	3.2218e+04	3.2218e+04	1×2 cell	1×2 cell	[3.3881,0.1195]	[3.3822,0.1169;3.3940,0.1223]	[9.1119e-06,-1.5281e-07;-1.5281e-07,1.8718e-06]	1×1 struct
15	'rayleigh'	1.8812e+04	3.7632e+04	3.7625e+04	3.7625e+04	1×1 cell	1×1 cell	21.5750	[21.2801;21.8783]	0.0233	1×1 struct
16	'exponential'	2.2000e+04	4.4008e+04	4.4001e+04	4.4001e+04	1×1 cell	1×1 cell	29.9619	[29.1485;30.8101]	0.1795	1×1 struct

Figure 4: Best Fit Rankings

As evident from the Rankings in figure 4, the distribution resembles Generalised Pareto distributions.
The parameters are:

- k = -0.9903
- σ = 19.9973
- θ = 20.0067

Placing the above parameters in a generalised pareto equation, we get:

$$y = f(x) = \frac{1}{\sigma} = \frac{1}{19.9973}$$

This is nothing but a uniform distribution.
Thus, the best fit is a uniform distribution ie **U(20,40)**

Validation: You can construct a frequency table with 20 bins and see the relative frequency of elements in each bin.

Interval Start	Interval End	Frequency
20.0067180488259	21.006297637708524	267
21.006297637708524	22.00587722659115	240
22.00587722659115	23.005456815473774	263
23.005456815473774	24.0050364043564	247
24.0050364043564	25.004615993239025	245
25.004615993239025	26.004195582121646	262
26.004195582121646	27.00377517100427	229
27.00377517100427	28.003354759886896	236
28.003354759886896	29.00293434876952	259
29.00293434876952	30.002513937652147	248
30.002513937652147	31.00209352653477	253
31.00209352653477	32.0016731154174	248
32.0016731154174	33.00125270430002	279
33.00125270430002	34.00083229318265	229
34.00083229318265	35.00041188206527	273
35.00041188206527	35.9999914709479	242
35.9999914709479	36.99957105983052	240
36.99957105983052	37.99915064871315	255
37.99915064871315	38.99873023759577	247
38.99873023759577	39.9983098264784	238

Table 2: Frequency Table

Hence, A uniform distribution is a good approximation for the given data.

6 Estimators

Let X_1, X_2, ... X_n be random sample drawn from given distribution $U(20, 40)$, represented as $U(A, B)$.

1. Let B_1 be an estimator of B, i.e. B_1 is the largest sample or $B_1 = max(X_1, X_2, ...X_n) = max(X_i)$
2. Let A_1 be an estimator of A, i.e. A_1 is the smallest sample or $A_1 = min(X_1, X_2, ...X_n) = min(X_i)$

Proof:
For Uniform (A,B) the likelihood function is $L(x_1, ..., x_n|A, B) = (\frac{1}{B-A})^n$ for any sample. To maximize this we must minimize the value of (BA) (interval length), yet we must keep all samples with in the range, i.e. $\forall x_i, x_i \in (A, B)$.
An MLE for A and B would then be $A_1 = min(X_i)$, $B_1 = max(X_i)$.
These values yield the minimal length since it's the smallest interval to include all sampled points.

7 Classification of Estimators

7.1 Part a

To check if the MLE is biased or not we must first find its Expected value. For this we must first find the p.d.f. of $max(x_i)$. Let our uniform distribution be $U(0, B)$. Denote $Y = max(x_i)$:

$$F_Y(y) = P(Y \leq y) = P(max(x_i) \leq y) = P(x_i \leq y, ..., x_n \leq y) =$$
$$P(x_1 \leq y)...P(x_n \leq y) = P^n(x \leq y) = \frac{y^n}{B}$$

So $f_Y(y) = F_Y^{'}(y) = n(\frac{1}{B})^n y^{n-1}$. Now we can find E($max(x_i)$):

$$E(max(x_i)) = E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_0^B yn(\frac{1}{B})^n y^{n-1} dy = \int_0^B n(\frac{y}{B})^n dy$$
$$n(\frac{1}{B})^n(\frac{y^{n+1}}{n+1})|_0^n = (\frac{n}{n+1})(\frac{1}{B})^n(B)^{n+1} = (\frac{n}{n+1})B < B$$

So the MLE $max(x_i)$ is biased since $E(max(x_i)) \neq B$.
Similarly we prove that the MLE $min(x_i)$ is biased.
Finally for the general uniform distribution $U(A, B)$ we have:
Let $X_i's$ be iid U(A,B) variables, $Y_i = \frac{X_i - A}{B-A}$ are iid U(0,1) variables where $1 \leq i \leq n$.
Now it can be shown that $Y_{(1)} \sim Beta(1, n)$ and $Y_{(n)} \sim Beta(n, 1)$, implying $E(Y_{(1)} = \frac{1}{n+1}$ and $E(Y_{(n)} = \frac{n}{n+1}$.

$$E(X_{(1)}) = \frac{B-A}{n+1} + A \geq A$$
$$E(X_{(n)}) = \frac{(B-A)n}{n+1} + B \leq B$$

Now finally, checking for consistency:

$$MSE(B_1) = Var(B_1) + Bias(B_1)^2$$

It can be shown that $Var(B_1) = \frac{n}{(n+1)^2(n+2)}B^2$ and $Bias(B_1) = -\frac{B}{n+1}$

$$\lim_{n\to\infty} MSE(B_1) = \lim_{n\to\infty} \frac{2B^2}{(n+2)(n+1)} = 0$$

We can prove that $\lim_{n\to\infty} MSE(B_1) = 0$ means that B_1 is a consistent estimator of B.

- B_1 is a biased, consistent estimator since it will never overestimate B and will underestimate B unless the value of the largest sample equals B.
- A_1 is a biased, consistent estimator since it will never underestimate A and will overestimate A unless the value of the smallest sample equals A.

7.2 Part b

For the given dataset, n = 5000 and our estimators have the following values:

1. $B_1 = max(X_i) = 39.9983098264784$
2. $A_1 = min(X_i) = 20.0067180488259$

8 Method of Moments

Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on (A, B), $-\infty < A < B < \infty$
Note that

$$E(X_i) = (A + B)/2 = \mu_1$$

and

$$E(X_i^2) = (A^2 + B^2 + AB)/3 = \mu_2$$

Substituting, we get

$$(2\mu_1 - B)^2 + B^2 + (2\mu_1 - B)B = 3\mu_2$$

which is the same as

$$(B - \mu_1)^2 = 3(\mu_2 - \mu_1^2)$$

Since $B > E(X)$ we obtain that

$$B = \mu_1 - \sqrt{3(\mu_2 - \mu_1^2)} = \overline{X} + \sqrt{\frac{3(n-1)}{n} S^2}$$

and

$$A = \mu_1 - \sqrt{3(\mu_2 - \mu_1^2)} = \overline{X} - \sqrt{\frac{3(n-1)}{n} S^2}$$

S^2 is sample variance and $\overline{X}$ is sample mean.
These estimators are not functions of the sufficient and complete statistic $(X_{(1)}, X_{(n)})$

9 UMVUE

Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on (A, B), $-\infty < A < B < \infty$
We need to find UMVUE for the parameters A, B.
Using factorization theorem we can show that $T(X) = (min X_1, X_2, .., X_n, max X_1, X_2, .., X_n) = (X_{(1)}, X_{(n)})$, where $T = (X_{(1)}, X_{(n)})$ is a complete sufficient statistic and $X_{(k)}$ is the k^{th} order statistic.
Now, since the $X_i's$ are iid U(a,b) variables, $Y_i = \frac{X_i - a}{b-a}$ are iid U(0,1) variables where $1 \leq i \leq n$.
Now it can be shown that $Y_{(1)} \sim Beta(1, n)$ and $Y_{(n)} \sim Beta(n, 1)$, implying $E(Y_{(1)} = \frac{1}{n+1}$ and $E(Y_{(n)} = \frac{n}{n+1}$. So we can now simply solve for a and b from the equations

$$E(X_{(1)}) = \frac{b-a}{n+1} + a$$

$$E(X_{(n)}) = \frac{(b-a)n}{n+1} + a$$

a and b are unbiased estimators of some function T, and by Lehmann-Scheffe theorem, those will be the corresponding UMVUEs.
For the given data sample we have $n = 5000$, solving for a, b we have:

$$b = (1 + n^{-1})X_{(n)} = 1.0002 * 39.9983098264784 = 40.00630948844369568$$

$$a = (1 - n^{-1})X_{(1)} = 0.9998 * 20.0067180488259 = 20.00271670521613482$$

10 Interval Estimator

The given distribution is $U(20, 40)$
We know that $P(Z > Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$
Also, $P(Z \leq Z_{\frac{\alpha}{2}}) = \frac{1}{20}(Z - 20)$, which is the CDF
From the above two equations we have

$$\frac{1}{20}(Z - 20) = 1 - \frac{\alpha}{2} \implies Z = 20(2 - \frac{\alpha}{2})$$

We will find the interval estimate of the mean of the given distribution $U(20, 40)$ with $n = 5000$ using the following formula:

$$\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Where σ is the standard deviation, μ is the population mean, $\bar{x}$ is the sample mean and α represents the $(1 - \alpha)100\%$ confidence interval of the population mean.
Now we know that $\bar{x} = 29.961941007654183$ and $\sigma = \sqrt{\frac{(B-A)^2}{12}} = 5.7735027$

1. $\alpha = 0.01$
   Here the interval is $29.961941007654183 \pm 20(2 - \frac{0.01}{2})\frac{5.7735027}{\sqrt{5000}}$
   Hence $26.7041196 \leq \mu \leq 33.2197624$

2. $\alpha = 0.05$
   Here the interval is $29.961941007654183 \pm 20(2 - \frac{0.05}{2})\frac{5.7735027}{\sqrt{5000}}$
   Hence $26.7367795 \leq \mu \leq 33.2197624$

3. $\alpha = 0.1$
   Here the interval is $29.961941007654183 \pm 20(2 - \frac{0.1}{2})\frac{5.7735027}{\sqrt{5000}}$
   Hence $26.777604 \leq \mu \leq 33.146278$