Project 6: Statistics and Bootstrapping

EE 511 – Section: Tuesday 5 pm

Name: Darshan Patil

Student ID: 9575227834

I. Problem Statement

The attached datasheet represents a set of 100 independent samples from some population.

a. Compute the sample mean, m, and the sample variance, s^2

b. Use the data to generate a discrete approximation to the Cumulative Distribution Function – the empirical distribution, $F_{X^*}(x)$. Plot this distribution.

c. By splitting the data into equal size intervals (0-5, 6-10, etc.), generate a discrete approximation to the distribution and determine the values of the Probability Mass Function for this discrete approximation.

d. Use the bootstrapping technique to generate M bootstrap samples based on the empirical distribution found in part b) and compute the sample mean and sample variance for each Bootstrap sample. Use M=50 and M=100.

e. The (population) mean of the empirical distribution is m and let m^*(RV based on the empirical distribution) be the mean of a bootstrap sample set. We could compute the MSE of the bootstrap sample means $MSE^* = E_{F^*}((m^* - m)^2)$ by looking at the comprehensive evaluation of all possible bootstrap sample sets. This is impractical, so we use smaller set of bootstrap sample sets as in d) and estimate the MSE by:

$$MSE(m^*) = \frac{1}{M}\sum_{i=1}^{M}(m_i^* - m)^2$$

We take this value to be an estimate of the MSE of the sample mean for the overall population distribution.

f. We can do a similar evaluation of the MSE of the bootstrap sample variance s^{*2}.

$$MSE(s^{*2}) = \frac{1}{M}\sum_{i=1}^{M}\left(s_i^{*2} - s^2\right)^2$$

We take this value to be an estimate of the MSE of the sample variance for the overall population distribution.

II. Theoretical Exploration

Given a set of iid, $\{X_i : i = 1, \dots, n\}$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We may also use m to represent the sample mean,

Then, $E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{n\mu}{n} = \mu$

So $\bar{X}$ is an unbiased estimator of μ.

Computing the mean square error-the expected value of the squared difference between $\bar{X}$ and μ;

$E[(\bar{X} - \mu)^2]$ which is just the variance of $\bar{X}$.

$$\begin{aligned}
MSE(\hat{m}) &= E[(\bar{X} - \mu)^2] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] \\
&= E\left[\frac{\left(\sum X_i\right)^2}{n^2} - 2\mu\frac{\sum X_i}{n} + \mu^2\right] \\
&= E\left[\frac{\sum X_i^2 + \sum\sum_{i \neq j} X_i X_j}{n^2}\right] - 2\mu^2 + \mu^2 \\
&= \frac{nE[X^2] + n(n-1)E[X]E[X]}{n^2} - \mu^2 \\
&= \frac{1}{n}\left(E[X^2] - \mu^2\right) = \frac{\sigma^2}{n}
\end{aligned}$$

Variance of $\bar{X}$ is given by,

$$VAR(\bar{X}) = E[(\bar{X} - E[\bar{X}])^2] = E[(\bar{X} - \mu)^2]$$

Which is the MSE, so the mean square error (or Variance of $\bar{X}$) is:

$$\begin{aligned}
VAR(\bar{X}) &= VAR\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} VAR(X_i) \quad \text{(independence)} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Empirical Distribution

In statistics, an empirical distribution function is the distribution function associated with the empirical measure of a sample. This cumulative distribution function is a step function that jumps up by 1/n at each of the n data points. Its value at any specified value of the measured variable is the fraction of observations of the measured variable that are less than or equal to the specified value.

The empirical distribution function is an estimate of the cumulative distribution function that generated the points in the sample. It converges with probability 1 to that underlying distribution, according to the Glivenko–Cantelli theorem. A number of results exist to quantify the rate of convergence of the empirical distribution function to the underlying cumulative distribution function.

Bootstrapping

Bootstrapping is any test or metric that relies on random sampling with replacement. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods. Generally, it falls in the broader class of resampling methods.

Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution. One standard choice for an approximating distribution is the empirical distribution function of the observed data. In the case where a set of observations can be assumed to be from an independent and identically distributed population, this can be implemented by constructing a number of resamples with replacement, of the observed dataset (and of equal size to the observed dataset).

Mean Square Error (MSE)

We calculate the Mean square error for the population given by:

MSE of Sample mean: $MSE(m^*) = \frac{1}{M}\sum_{i=1}^{M}(m_i^* - m)^2$

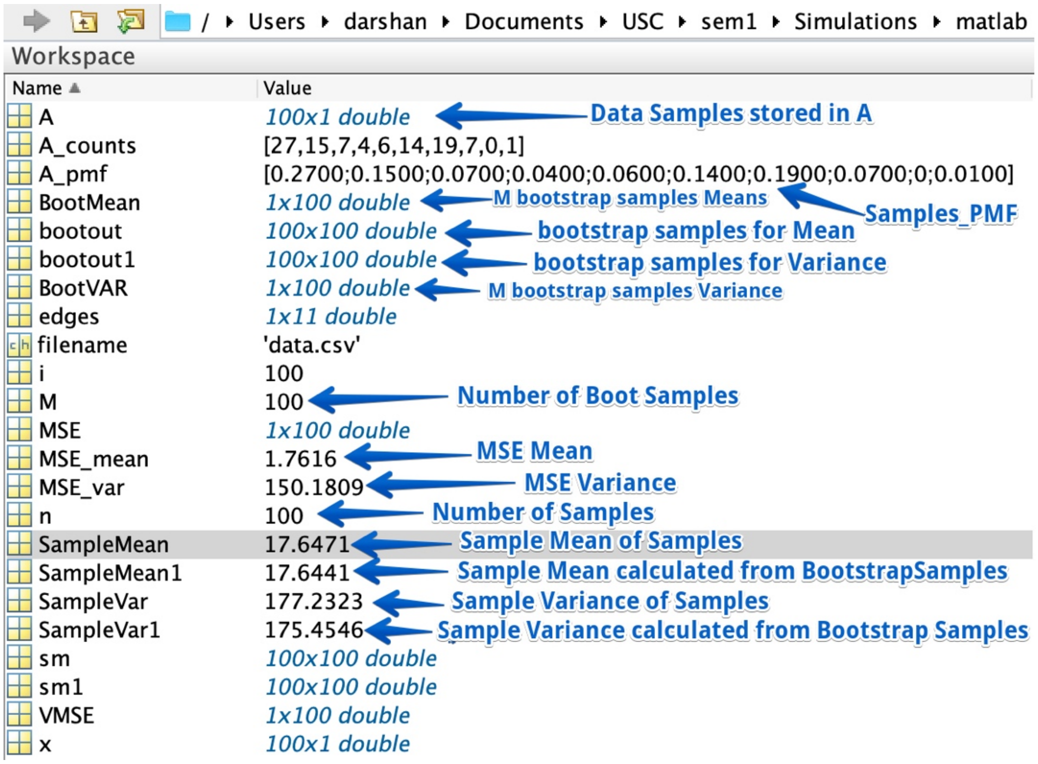
where m^*(RV based on the empirical distribution) be the mean of a bootstrap sample set, and m is sample mean of the given samples.

MSE of Sample Variance: $MSE(s^{*2}) = \frac{1}{M}\sum_{i=1}^{M}\left(s_i^{*2} - s^2\right)^2$

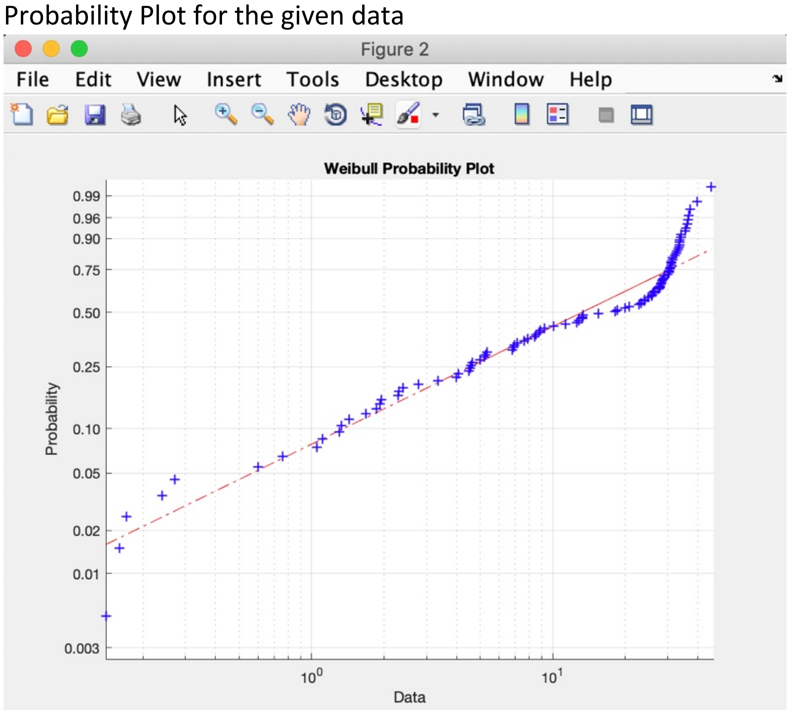
where MSE of the bootstrap sample variance is s^{*2}, and s^2 is sample variance of the given samples.

III. Results

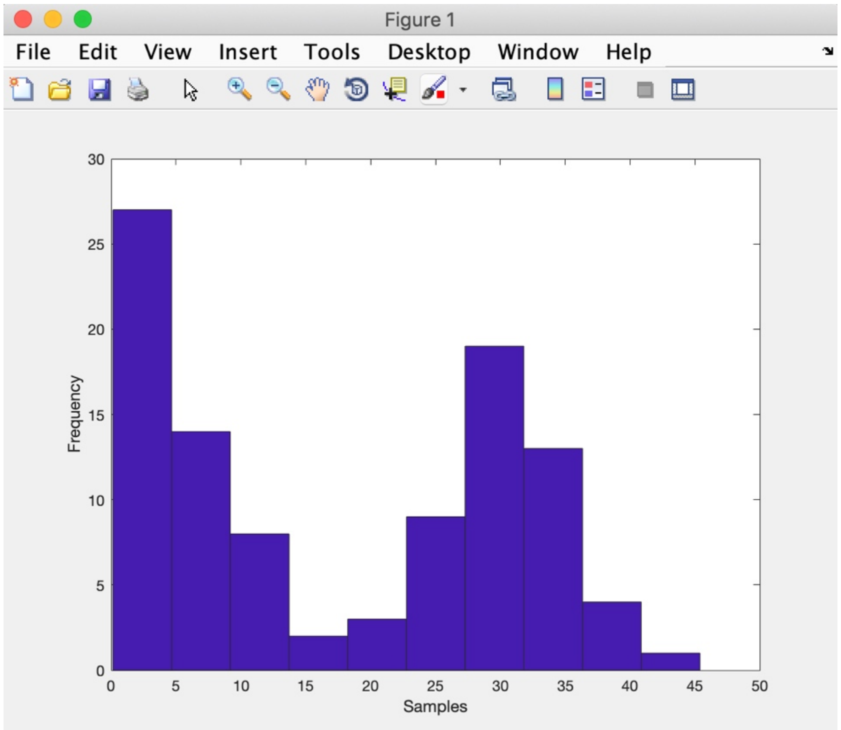
Matlab Result Window



Probability Plot for the given data

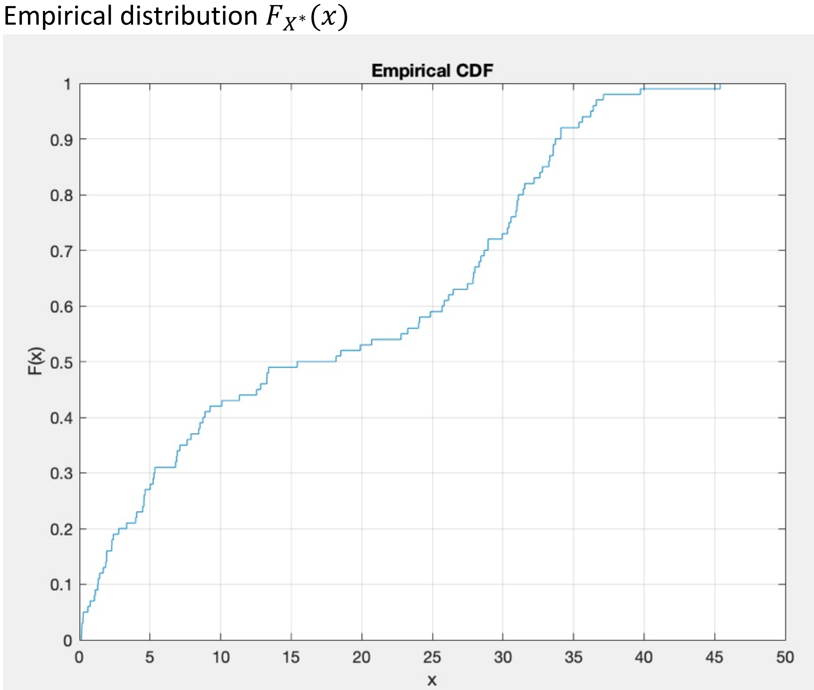


Histogram of given Samples,



a. For the given data samples,
Sample Mean (m)= 17.6471
Sample Variance (s^2)= 177.2323

b. Empirical distribution $F_{X^*}(x)$



Sample X_i = [0.14, 0.16, 0.17, 0.24, 0.27, 0.6, 0.76, 1.05, 1.11, 1.3, 1.33, 1.43, 1.68, 1.86, 1.92, 1.94, 2.28, 2.3, 2.4, 2.78, 3.35, 3.98, 4.06, 4.48, 4.56, 4.57, 4.65, 5.01, 5.21, 5.27, 5.34, 6.81, 6.89, 6.93, 7.12, 7.63, 7.9, 8.45, 8.53, 8.75, 8.9, 9.25, 10.08, 11.33, 12.55, 12.83, 13.27, 13.29, 13.4, 15.43, 18.18, 18.52, 19.91, 20.71, 22.79, 23.26, 24.03, 24.1, 24.87, 25.69, 25.84, 26.16, 26.48, 27.49, 27.88, 27.92, 28.0, 28.32, 28.44, 28.67, 28.95, 28.96, 29.95, 30.33, 30.43, 30.57, 30.92, 30.99, 31.03, 31.1, 31.44, 31.54, 32.2, 32.62, 32.79, 33.25, 33.32, 33.55, 33.56, 33.72, 34.1, 34.1, 35.38, 35.62, 36.22, 36.4, 36.62, 37.12, 39.74, 45.39, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54]

Empirical CDF = [0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.2, 0.21, 0.22, 0.23, 0.24, 0.25, 0.26, 0.27, 0.28, 0.29, 0.3, 0.31, 0.32, 0.33, 0.34, 0.35, 0.36, 0.37, 0.38, 0.39, 0.4, 0.41, 0.42, 0.43, 0.44, 0.45, 0.46, 0.47, 0.48, 0.49, 0.5, 0.51, 0.52, 0.53, 0.54, 0.55, 0.56, 0.57, 0.58, 0.59, 0.6, 0.61, 0.62, 0.63, 0.64, 0.65, 0.66, 0.67, 0.68, 0.69, 0.7, 0.71, 0.72, 0.73, 0.74, 0.75, 0.76, 0.77, 0.78, 0.79, 0.8, 0.81, 0.82, 0.83, 0.84, 0.85, 0.86, 0.87, 0.88, 0.89, 0.9, 0.92, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99, 1.0, 1, 1, 1, 1, 1]

c. Splitting the data into equal size intervals (0-5, 6-10, etc), the values of Probability Mass Function for the discrete approximation is found.

Interval	Number of values in each bin	PMF
0-5	27	0.27
6-10	15	0.15
11-15	7	0.07
16-20	4	0.04
21-25	6	0.06
26-30	14	0.14
31-35	19	0.19
36-40	7	0.07
41-45	0	0
46-50	1	0.01

d. Sample Mean and Sample Variance for M bootstrap Samples,
For M=50,

Sample Mean of bootstrap samples,

17.0352	18.7956	19.7026	16.9792	13.3122
17.0242	20.986	16.0586	12.5292	22.0196
18.5548	17.777	14.041	17.5414	16.4034
16.0664	21.4642	22.5586	15.091	18.8572
16.7358	19.0192	14.7866	14.6486	16.2558
19.4714	20.0314	19.0896	17.7822	18.3002
16.1738	19.9844	18.7466	18.2254	15.8666
18.8978	20.703	15.8796	18.4248	17.0334
17.9816	20.764	19.8498	20.0038	14.4566
17.2226	17.1974	14.3282	15.5212	16.1764

We can observe that the bootstrap sample means are close to the empirical sample mean.

Sample Variance of each M bootstrap sample,

177.207019	171.81892	176.715899	172.503047	176.470974
170.551128	170.31717	173.868865	175.955851	177.318051
175.32806	171.963251	176.446762	168.492907	174.823745
170.836007	171.283851	173.960301	176.073799	172.707272
173.857773	174.253546	172.735833	170.732687	170.526207
171.845983	175.419409	167.42841	176.322063	176.442345
173.706682	174.737978	172.227842	171.88975	166.786575
170.949376	172.773302	175.650881	174.610853	171.655276
176.859877	175.760637	175.632591	174.109818	178.318292
178.48018	177.866107	170.077966	177.933283	178.556862

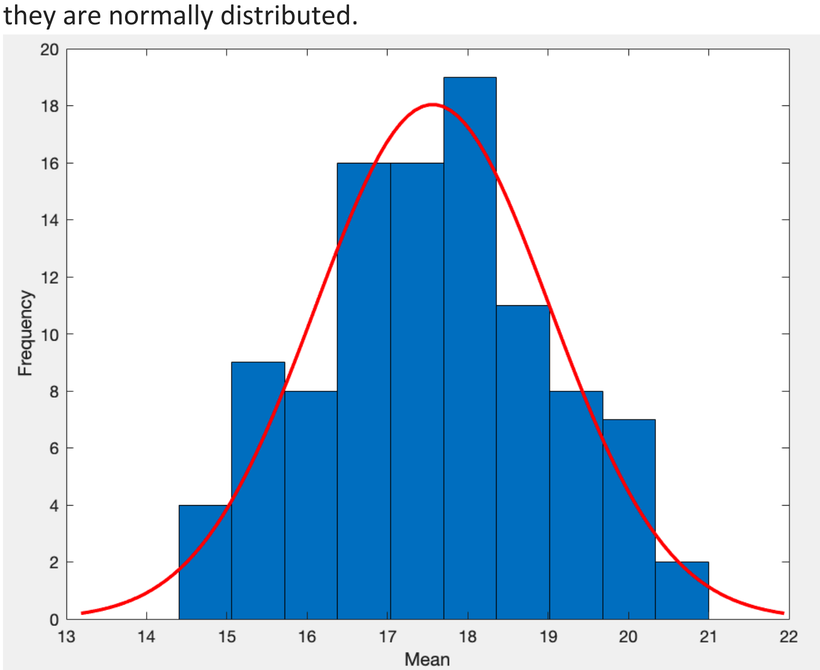
We can observe that the bootstrap sample variance is close to the sample variance.

When M increases the distribution tends to be normally distributed. i.e. Bootstrap Sample Mean tends to the true mean of the population, and Bootstrap Sample Variance tends to the true variance of the population.

For M=100,
Sample Mean of each M bootstrap sample,

16.8442	19.2832	18.2816	18.1606	17.3659
18.9737	18.1951	16.4784	18.7036	14.7557
17.7776	17.5464	16.2799	16.325	16.7132
16.7109	16.7323	17.0473	16.6005	19.7247
17.1714	17.1022	17.2878	18.2017	17.5834
17.6894	18.3923	17.6157	19.0465	17.2947
20.7622	16.6874	15.1867	17.0017	16.5918
16.4418	18.7892	16.9035	18.517	15.4832
20.8072	17.2171	17.2499	15.9516	16.568
18.0108	18.5081	18.5817	16.2039	17.6315
17.2813	16.9713	16.4927	20.6971	19.2295
17.9692	17.6382	18.7155	19.5119	17.2578
18.2414	17.8312	17.4322	17.747	21.1038
18.5202	18.12	16.96	19.3177	17.6014
18.7443	17.6506	17.0151	16.6116	17.7937
15.2059	16.9643	18.1227	18.8459	18.976
17.0882	17.2996	18.0994	18.7083	19.7625
20.3679	17.023	16.0955	18.3536	16.3656
17.0642	17.3551	17.5117	16.407	18.8028
17.506	16.9754	18.9831	16.8661	18.3137

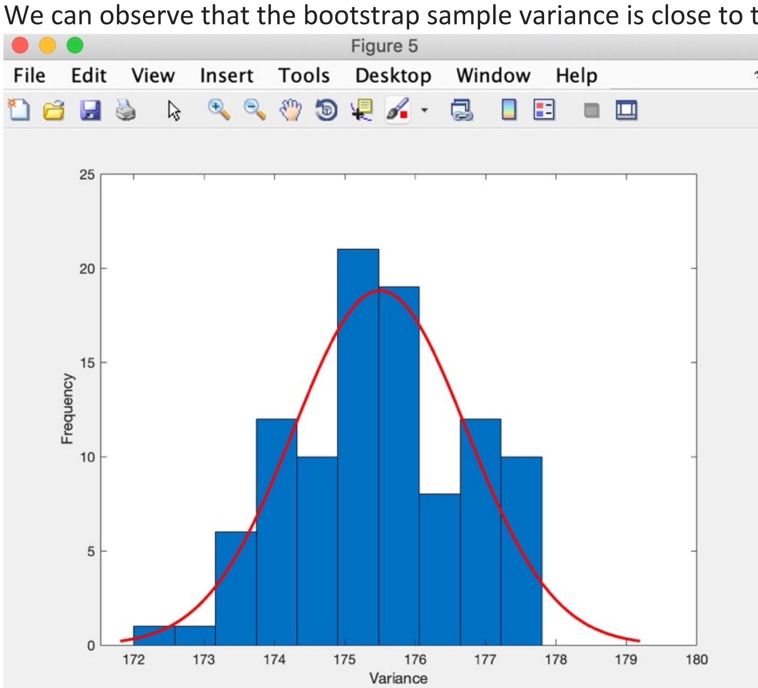
We can observe that the bootstrap sample means are close to the empirical sample mean, and they are normally distributed.



Sample Variance of each M bootstrap sample,

178.143254	171.819273	176.308721	175.419247	177.536664
173.016636	174.698457	174.354186	177.179804	176.456874
174.781834	176.411388	175.755534	176.598573	175.087175
175.022673	174.214282	176.683123	174.93587	175.922977
175.402796	176.784294	174.157475	175.375479	175.218419
175.884127	174.733818	175.378622	174.588616	174.811396
175.802528	174.791448	173.916823	176.561229	175.203919
175.826426	175.771163	177.073383	175.363646	172.858957
177.209853	178.059187	177.98651	174.916602	175.261016
176.044053	174.56259	174.633926	175.041784	173.98957
174.708035	178.720363	172.657635	176.925382	176.2093
175.461776	173.884798	173.321246	173.868524	176.544033
175.390545	175.644075	175.982394	175.493334	176.79365
176.326007	174.235401	174.635038	173.696392	175.164854
174.5948	175.805558	176.742049	175.001508	175.359962
175.681278	174.371514	174.74678	174.079373	175.340975
175.519419	172.80107	176.152556	176.848792	176.213126
177.136212	175.521141	174.947917	175.147724	174.772915
176.336431	178.768616	175.897398	175.617297	175.943407
175.343321	175.366093	176.034647	174.556719	175.595017

We can observe that the bootstrap sample variance is close to the sample variance.



e. MSE of Mean,
For N= 50, $MSE_F(m) = 1.8931$
For N= 100, $MSE_F(m) = 1.7616$

f. MSE of Variance,
Considering the set of bootstrap samples and using the sample variance found in part d) the MSE of population variance is estimated:
For N=50, $MSE_{F^*}(s^2) = 136.8376$
For N=100, $MSE_{F^*}(s^2) = 150.1812$

The mean squared error indicates how close a regression line is to a set of points.
MSE calculates this by considering the distances from the points to the regression line (these distances are the “errors”) and squaring them.

IV. Reference

- https://en.wikipedia.org/wiki/
- Lecture Notes by professor Silvester

V. Source Code

```
clc;
clear all;
close all;
filename = 'data.csv';
A = csvread(filename);  % read the samples from .csv file
M=100;              % M number of bootstrap samples
n=100;
edges=0:5:50;
A_counts=histcounts(A,edges);   % Determine data in each interval

A_pmf=A_counts'/n;          % Calculated the pmf for the given samples
for i=1:M
   x=randsample(A,M);
   bootout(i,:)=bootstrp(M,@mean,x);   % Using Bootstrap technique to
                                        generate M bootstrap samples
   bootout1(i,:)=bootstrp(M,@var,x);
end

SampleMean=mean(A);          % Calculate the sample mean
SampleVar=var(A);            % Calculate the sample variance
BootMean=mean(bootout);
SampleMean1=mean(BootMean);    % Calculate the sample mean of bootstrap samples
BootVAR=mean(bootout1);
SampleVar1=mean(BootVAR);      % Calculate the sample variance of bootstrap samples
sm=(bootout-SampleMean).^2;
MSE=sum(sm)/M;
MSE_mean=mean(MSE);           %Calculate MSE for Mean
sm1=(bootout1-SampleVar).^2;
VMSE=sum(sm1)/M;
MSE_var=mean(VMSE);           %Calculate MSE for Variance
figure(1);
hist(A);                  %Plot the data
xlabel('Samples');
ylabel('Frequency');
figure(2);
wblplot(A);                  % Plot the probability
figure(3);
cdfplot(A);                  % empirical cdf plot
figure(4);
histfit(BootMean);           % Plot Bootstrap mean values
xlabel('Mean');
ylabel('Frequency');
figure(5);
histfit(BootVAR);            % plot bootstrap variance values
xlabel('Variance');
ylabel('Frequency');
```