CS6015 Linear Algebra and Random Processes

Muhammad Arsath K F(CS16S035) and Srihari Maruthachalam (CS16S024)

Programming Assignment-II

Group I

November 6, 2017

Question 1

Expectation, $E\left(\overline{X_N}\right)$

Given:

Sample mean $\overline{X_N} = \frac{1}{N}\sum_{i=1}^{N} X_i$; $E[X_i] = \mu$; $Var(X_i) = \sigma^2$

$$* \; E(\overline{X_N}) = E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right]$$

$$= \frac{1}{N} \cdot E\left[\sum_{i=1}^{N} X_i\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E[X_i] \qquad \left[\because E(X_1+X_2) = E(X_1)+E(X_2)\right]$$

// Linearity Property

$$= \frac{1}{N}\sum_{i=1}^{N} \mu$$

$$= \frac{\mu}{N}[N]$$

$$E[\overline{X_N}] = \mu$$

Variance, $Var\left(\overline{X_N}\right)$

$$* \; var(\overline{X_N}) = var\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)$$

$$= \frac{1}{N^2} var\left(\sum_{i=1}^{N} x_i\right) \qquad \left[\because var(ax) = a^2 var(X)\right]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N} var(x_i) \qquad \left[var(x_1+x_2) = var(x_1)+var(x_2)\right]$$

x_1 & x_2 are independent and uncorrelated.

$$= \frac{1}{N^2}\sum_{i=1}^{N} \sigma^2 \qquad \left[var(x_i) = \sigma^2\right]$$

$$= \frac{\sigma^2}{N^2}\sum_{i=1}^{N} 1$$

$$= \frac{\sigma^2}{N^2} \cdot N$$

$$Var(\overline{X_N}) = \frac{\sigma^2}{N}$$

Question 2

Given:

Let $x_1, x_2 \cdots x_n$ denote a sequence of i.i.d random variables with $x_i \in [a,b], \forall i$ where $-\infty < a \leq b < \infty$

$$\left.\begin{aligned} P(\overline{X_N} - \mu \geq \epsilon) &\leq \exp\left(-\frac{2N\epsilon^2}{(b-a)^2}\right) \\ P(\overline{X_N} - \mu \leq -\epsilon) &\leq \exp\left(-\frac{2N\epsilon^2}{(b-a)^2}\right) \end{aligned}\right\} \longleftarrow (1)$$

From eqn (1), we can rewrite as,

$$P(\overline{X_N} - \mu \geq \epsilon) = P(\overline{X_N} - \epsilon \geq \mu)$$

$$P(\overline{X_N} - \mu \leq -\epsilon) = P(\overline{X_N} + \epsilon \leq \mu)$$

Now,

$$P(\overline{X_N} - \epsilon \geq \mu) + P(\overline{X_N} + \epsilon \leq \mu) \leq 2\exp\left(\frac{-2N\epsilon^2}{(b-a)^2}\right)$$

Let δ be $2\exp\left(\frac{-2N\epsilon^2}{(b-a)^2}\right)$.

ie. $P(\overline{X_N} - \epsilon' \geq \mu) + P(\overline{X_N} + \epsilon' \leq \mu) \leq \delta$

$$1 - \left[P(\overline{X_N} - \epsilon' \geq \mu) + P(\overline{X_N} + \epsilon' \leq \mu)\right] \geq 1 - \delta$$

$$\therefore P(\overline{X_N} - \epsilon' \leq \mu \leq \overline{X_N} + \epsilon') \geq 1 - \delta \quad \text{— (2)}$$

To find ϵ' as a function of δ and N:

WKT,

$$2\exp\left(-\frac{2N\epsilon'^2}{(b-a)^2}\right) = \delta$$

$$\exp\left(-\frac{2N\epsilon'^2}{(b-a)^2}\right) = \frac{\delta}{2}$$

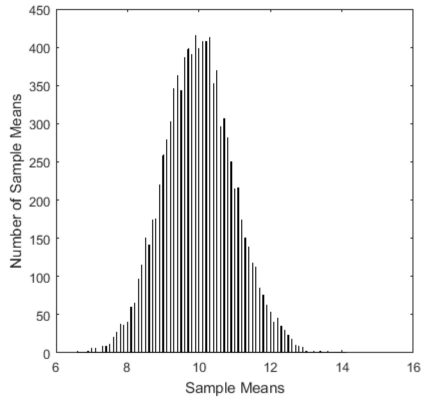
$$-\frac{2N\epsilon'^2}{(b-a)^2} = \log\left(\frac{\delta}{2}\right)$$

$$\epsilon'^2 = \frac{(b-a)^2}{2N}(-1)\log\left(\frac{\delta}{2}\right)$$

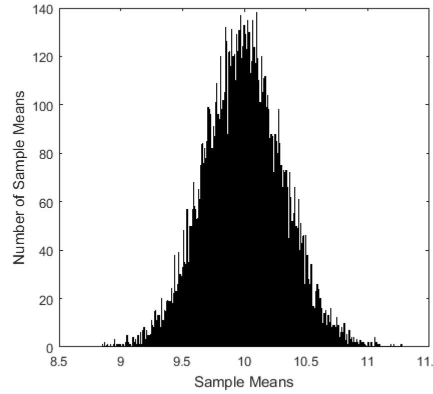
$$\epsilon' = \sqrt{\frac{-(b-a)^2}{2N}\log\left(\frac{\delta}{2}\right)}$$

Question 3

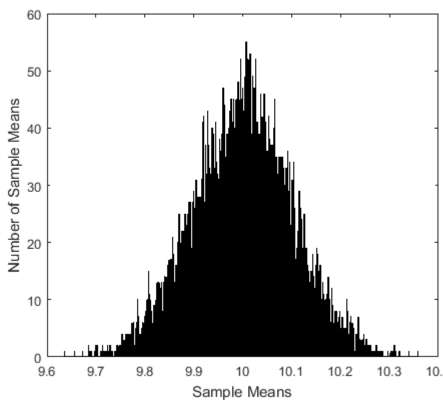
Histogram of the sample means



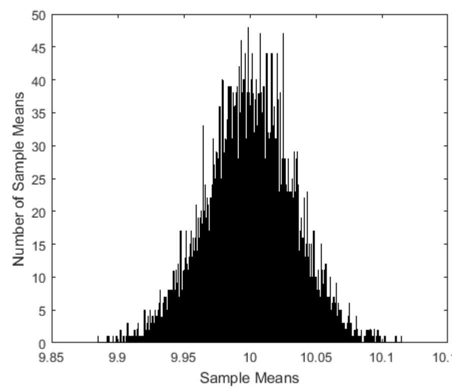
(a) Histogram of 10 samples



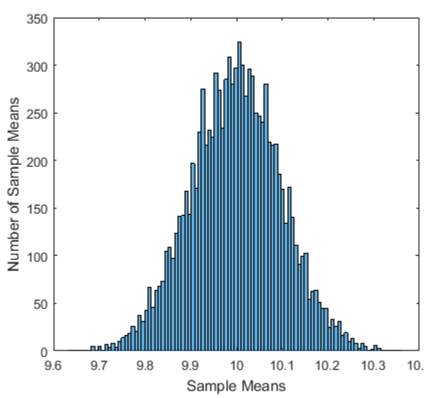
(b) Histogram of 100 samples



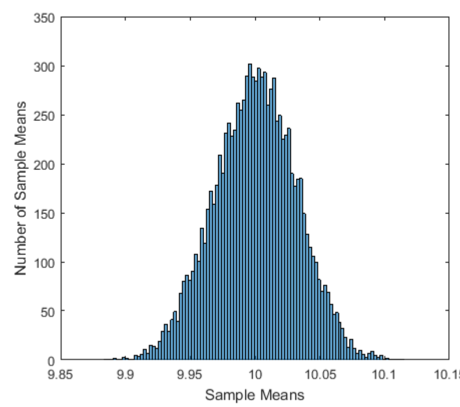
(c) Histogram of 1000 samples



(d) Histogram of 10000 samples



(e) Histogram of 1000 samples(with 100 bins)



(f) Histogram of 10000 samples(with 100 bins)

Figure 5: Histogram for various number of sample means for 10000 repetition

Closeness of sample mean to the true mean

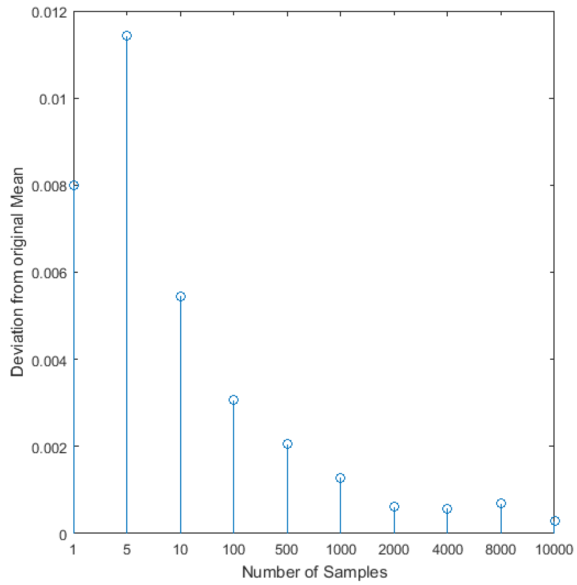


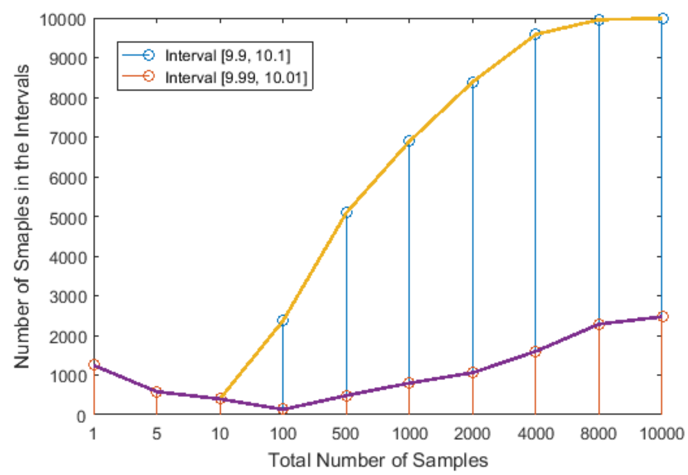
Figure 6: Absolute difference between sample mean and true mean

- As we can see from the above plot, the sample mean is getting closer to the true mean as we increase the number of samples.
- **Reason :** *Law of Large numbers* says as the number of experiments increases, the actual ratio of outcomes will converge on the theoretical, or expected, the ratio of outcomes.

Sample mean in the intervals $[9.9, 10.1]$ and $[9.99, 10.01]$

Table 1: My caption

-2*Number of Samples	Sample mean in the interval [9.9, 10.1]	[9.99, 10.01]
1	1248	1248
5	582	582
10	398	398
100	2370	133
500	5086	483
1000	6891	799
2000	8386	1059
4000	9585	1598
8000	9957	2284
10000	9988	2471



95% confidence interval for the sample mean

Number of Samples	Confidence Interval		Number of times True mean falls outside 95 % Confidence interval
1	3.99999	16.01601	**411**
5	7.2	12.77716	**494**
10	8.0891	11.9	**490**
100	8.99694	10.99694	**31**
500	8.9979524	10.9979524	**0**
1000	8.9987134	10.9987134	**0**
2000	8.9993856	10.9993856	**0**
4000	9.000577625	11.00057763	**0**
8000	9.000703138	11.00070314	**0**
10000	9.0002905	11.0002905	**0**

Table 2: Confidence Intervals using the numerical results

- To get the 95% confidence interval, we initiated from the sample mean and extended the ϵ value on both sides of the distribution.
- As we can clearly see from the above table, as we increase the number of samples, confidence interval shrinking and comes closer to true mean.
- As we can see from the above table, as we increase the number of samples, the number of times true mean fall outside the confidence interval tends to zero.
- To infer more, we attempted to estimate confidence interval using Gaussian and Poisson Standard Deviations. The formula used to estimate and the numerical obtained are shown below.

Formula to estimate confidence interval using Gaussian Standard Deviations

$$Confidence\ Interval = [\hat{\mu} - 2\sigma, \hat{\mu} + 2\sigma]$$

where $\hat{\mu}$ is the sample mean, and σ is Gaussian Standard Deviation.

Formula to estimate confidence interval using Poisson Standard Deviations

$$Confidence\ Interval = \left[\hat{\lambda} - 1.96\sqrt{\frac{\hat{\lambda}}{n}}, \hat{\lambda} + 1.96\sqrt{\frac{\hat{\lambda}}{n}}\right]$$

where $\hat{\lambda}$ is the sample mean, and n is number of samples.

Number of Samples	Confidence Interval	
1	3.807457056	16.20854294
5	3.794055892	16.18310411
10	3.798174989	16.19092501
100	3.799824162	16.19405584
500	3.800522776	16.19538202
1000	3.80104792	16.19637888
2000	3.801511794	16.19725941
4000	3.802334406	16.19882084
8000	3.802421023	16.19898525
10000	3.80213626	16.19844474

(a) Confidence Intervals using Gaussian Standard Deviation

Number of Samples	Confidence Interval	
1	3.629739091	16.38626091
5	7.138925567	12.83823443
10	8.027007291	11.96209271
100	9.354763556	10.63911644
500	9.713106338	10.28279846
1000	9.801875769	10.19555103
2000	9.857072964	10.14169824
4000	9.902588406	10.09856684
8000	9.931359073	10.0700472
10000	9.936878809	10.06370219

(b) Confidence Intervals using Poisson Standard Deviation

Table 3: Confidence Intervals

Theorem 1 not applicable for Poisson Random Variables

- **Reason :** Theorem 1 is applicable for distributions which follow symmetricity along mean. But Poisson distribution doesn't follow symmetricity along mean.

Given:

Let $x_1, x_2 \cdots x_n$ denote a sequence of i.i.d random variables with $x_i \in [a,b], \forall i$ where $-\infty < a \leq b < \infty$

Here, $X_i \sim \text{Binomial}(n, \lambda/n)$

WKT,
$$\epsilon' = \sqrt{\frac{-(b-a)^2}{2N} \log(\delta/2)}$$

We need 95% confidence interval. So $\delta = 0.05$

$$\epsilon' = \sqrt{\frac{-(b-a)^2}{2N} \log(0.025)}$$

Using Theorem2,

$$P(\overline{X_N} - \epsilon' \leq \mu \leq \overline{X_N} + \epsilon') \geq 1 - \delta$$

i.e. $$P\left(\overline{X_N} - \left(\sqrt{\frac{-(b-a)^2}{2N} \log(0.025)}\right) \leq \mu \leq \overline{X_N} + \left(\sqrt{\frac{-(b-a)^2}{2N} \log(0.025)}\right)\right) \geq 0.95$$

Samples needed to attain various accuracy

Accuracy Level (α)	Number of samples necessary
0.1	23
0.01	110
0.001	1048
0.0001	10359
0.00001	100298

- As we can see from the above table, as accuracy level increase, number of samples needed to attain the same also increases.
- To generalize, the number of samples needed to attain accuracy(i.e. α), N = $\left(\frac{1}{\alpha}\right)$
- To infer more, we attempted to check the accuracy of the sample mean and true mean with the various number of samples. We have shown the corresponding table below.

Number of Samples	**Sample Mean for 10K iterations**	**Absolute difference from true mean**	**Accuracy precision of 0.1**	**Accuracy precision of 0.01**	**Accuracy precision of 0.001**	**Accuracy precision of 0.0001**
1	10.008	0.008	✔	✔	✘	✘
5	9.98858	0.01142	✔	✘	✘	✘
10	9.99455	0.00545	✔	✔	✘	✘
100	9.99694	0.00306	✔	✔	✘	✘
500	9.9979524	0.0020476	✔	✔	✘	✘
1000	9.9987134	0.0012866	✔	✔	✘	✘
2000	9.9993856	0.0006144	✔	✔	✔	✘
4000	10.00057763	0.000577625	✔	✔	✔	✘
8000	10.00070314	0.000703138	✔	✔	✔	✘
10000	10.0002905	0.0002905	✔	✔	✔	✘

Table 4: Accuracy attained with various number of samples

Question 4

Finding A

a) Given: $f(k) = \frac{A}{k^2}$ for $k = \pm 1, \pm 2, \pm 3 \ldots$

To be a valid PMF,

$$\sum_{k \neq 0} f(k) = 1$$

i.e. $$\sum_{k \neq 0} \frac{A}{k^2} = 1.$$

we can rewrite as,

$$\sum_{k=1}^{\infty} \frac{2A}{k^2} = 1 \qquad [\because f(k) \text{ is an even function}]$$

$$2A \sum_{k=1}^{\infty} \frac{1}{k^2} = 1 \quad \text{--- } (1)$$

From Basel's approximation,

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \simeq 1.644934$$

sub in eqn (1),

$$2A(1.644934) = 1$$

$$A = \frac{1}{3.289868}$$

$$\therefore A = 0.3039635$$

b) Expectation :

$$E[f(k)] = \sum_{k \neq 0} k \cdot \frac{A}{k^2} \quad , \quad k = \pm 1, \pm 2, \ldots$$

$$= \sum_{k \neq 0} \frac{A}{k} \quad , \quad k = \pm 1, \pm 2 \ldots$$

$$= A \sum_{k \neq 0} \frac{1}{k} \quad , \quad k = \pm 1, \pm 2 \ldots \quad \text{—(1)}$$

Here, $f(k)$ is odd function

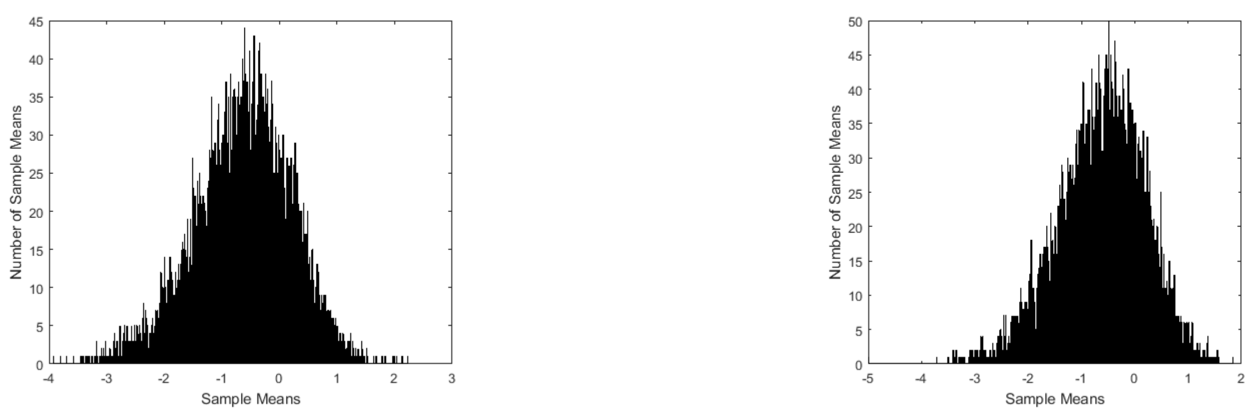
i.e. $\sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} f(k) = 0$, when $f(k)$ is odd.

Apply the result in eqn (1)

$$E[f(k)] = A(0)$$

$$E[f(k)] = 0$$

Histogram of the sample means



(a) Histogram of 1000 samples (b) Histogram of 10000 samples

Figure 10: Histogram for various number of sample means for 10000 repetition

- We observed that the sample mean is -0.62155489999 when N=1000 and sample mean is -0.607428530000 when N=10000
- We can see that sample mean concentrates around -0.6(approx) for various choices of sample counts.
- From theoretical results, we expected sample mean to be zero. But, sample mean concentrates on -0.6(approx).

95% confidence interval for the sample mean

Number of Samples	Confidence Interval		Number of times True mean falls outside 95 % Confidence interval
2	-1.7982	1	**0**
10	-2.01818	0.9	**444**
100	-2.113953	0.930007	**495**
500	-2.108631	0.910009	**499**
1000	-2.1690049	0.9258951	**499**
2000	-2.1505077	0.9011523	**499**
4000	-2.100323025	0.891756975	**498**
8000	-2.145130088	0.939529913	**499**
10000	-2.14610853	0.93125147	**499**

Table 5: Confidence Intervals using the numerical results

- To get the 95% confidence interval, we initiated from the sample mean and extended the ϵ value on both sides of the distribution.
- As we can clearly see from the above table, as we increase the number of samples, confidence interval shrinking and comes closer to true mean.

Reference

- Probability and Random Processes, 3rd Edition by Geoffrey R. Grimmett, David R. Stirzaker.
- Law of large numbers(Wikipedia)
- MANJUNATH KRISHNAPUR Lecture Notes on PROBABILITY AND STATISTICS