LARP : Assignment-2

Sankaran Vaidyanathan (CS17Z015) and Karthik Thiagarajan (CS16S027)

Problem-1

Each random variable X_i is distributed with mean μ and variance σ^2. Then, the mean of the random variables is also a random variable:

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

Since expectation is a linear operator and given that the N random variables are identically distributed.

$$E(\bar{X}) = \frac{1}{N}\sum_{i=1}^{N} E(X_i) = \mu$$

We make use of the independence assumption to get the variance:

$$\mathrm{Var}(\bar{X}) = \frac{1}{N^2}\sum_{i=1}^{N} \sigma^2 = \frac{\sigma^2}{N}$$

Problem-2

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

$$\mathbb{P}(\bar{X} - \mu \geq \epsilon) \leq \frac{\delta}{2}$$
$$\mathbb{P}(\bar{X} - \mu \leq -\epsilon) \leq \frac{\delta}{2}$$

Since the events $\bar{X} - \mu \geq \epsilon$ and $\bar{X} - \mu \leq -\epsilon$ are mutually exclusive, we get:

$$\mathbb{P}(\bar{X} - \mu \geq \epsilon \text{ and } \bar{X} - \mu \leq -\epsilon) \leq 2\,\frac{\delta}{2} = \delta$$

Rewriting the above inequality, we get

$$\mathbb{P}(|\bar{X} - \mu| \geq \epsilon) \leq \delta$$

The complementary event is $|\bar{X} - \mu| \leq \epsilon$. The corresponding probability is:

$$\mathbb{P}(|\bar{X} - \mu| \leq \epsilon) = 1 - \mathbb{P}(|\bar{X} - \mu| \geq \epsilon) \geq 1 - \delta$$

The event $|\bar{X} - \mu| \leq \epsilon$ can be rewritten as $\mu \in [\bar{X} - \epsilon, \bar{X} + \epsilon]$.

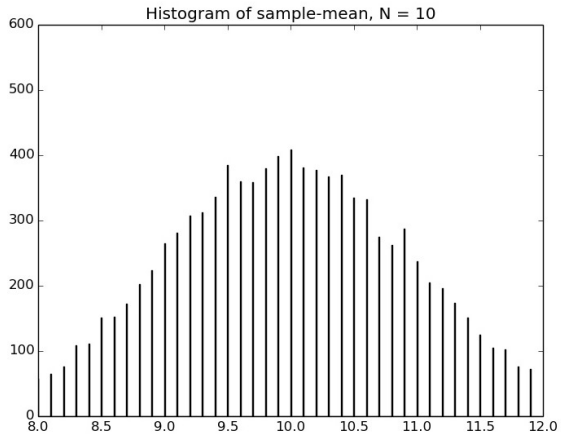
Thus, the required probability becomes:

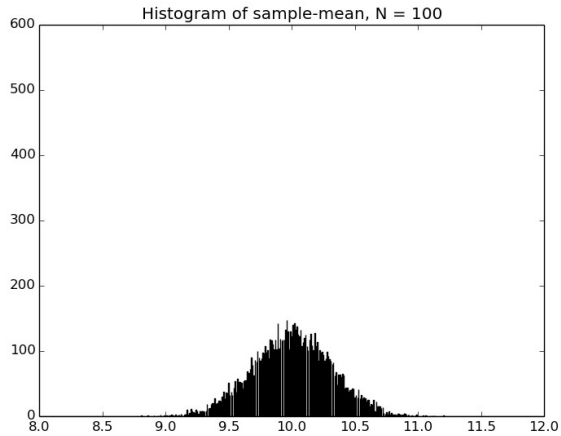
$$\mathbb{P}(\mu \in [\bar{X} - \epsilon, \bar{X} + \epsilon]) \geq 1 - \delta$$

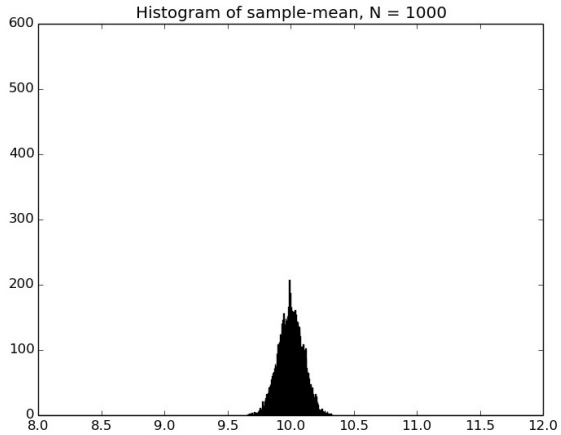
Since $\delta = 2\exp\left(-\frac{2N\epsilon^2}{(b-a)^2}\right)$,

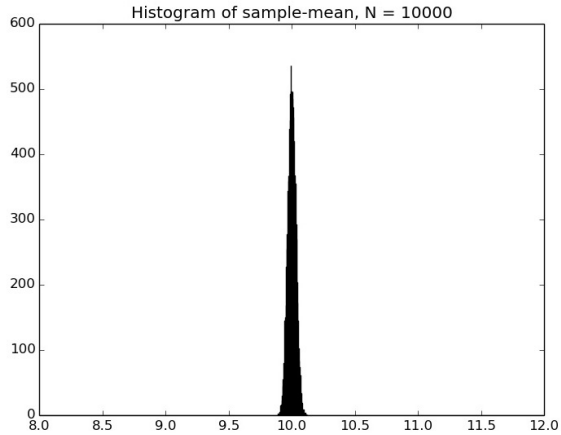
$$\epsilon = \sqrt{\frac{(b-a)^2}{2N} ln\left(\frac{2}{\delta}\right)}$$

Problem-3









(a) The sample mean is close to the true mean. From the first question, we know that the sample mean of N i.i.d random variables, each having mean μ and variance σ^2, is also a random variable with mean μ and variance $\frac{\sigma^2}{N}$. For any given experiment with N samples, their mean is one realisation of this random variable.

(b) This table represents the percentage of experiments where the sample mean falls in the given interval.

N	% of intervals with $\bar{X} \in [9.9, 10.1]$	% of intervals with $\bar{X} \in [9.99, 10.01]$
10	4.09	4.09
100	23.89	1.22
1000	67.96	7.85
10000	99.86	24.38

There are two observations:

1. For a given value of N, wider the interval, greater the number of sample means that it contains. This is quite obvious.
2. As the number of samples are increased, the estimate starts becoming more accurate. This is clearly observed in the increase in the number of sample-means falling in a given interval.

(c) This table represents the percentage of intervals (10000 in all) that contain the true mean.

N	% of intervals with $\mu \in [a, b]$
10	89.97
100	94.69
1000	95.04
10000	95.27

As the sample size grows, the percentage of intervals that contain the true mean moves closer to the 95% mark. This is expected since we are calculating a 95% confidence interval.

(d) The random variables given in problem-1 have a bounded support, $[a, b]$, whereas Poisson random variables have an unbounded support, $[0, \infty)$.

Let $X_i\ Bin(n, \frac{\lambda}{n})$ be the i^{th} random variable with $\mu = \lambda, \sigma^2 = \lambda(1 - \frac{\lambda}{n})$.

The sample-mean is a random variable with $\mu = \lambda, \sigma^2 = \frac{\lambda}{N}(1 - \frac{\lambda}{n})$.

For a 95% confidence interval, we need the $\delta = 1 - 0.95 = 0.05$. Using the equation for ϵ, by substituting the values as follows:

$a = 0, b = n, \delta = 0.05$

$$\epsilon = \sqrt{\frac{(b-a)^2}{2N} ln\left(\frac{2}{\delta}\right)} = 1.36\frac{n}{\sqrt{N}}$$

For a large value of n, the Binomial distribution is quite a good approximation of the Poisson. The numerical estimate of $\epsilon = 0.06$. To get an equivalent theoretical estimate, we need:

$$\frac{n}{\sqrt{N}} = \frac{0.06}{1.36} = 0.04 \implies N = 625n^2$$

If we replace the Poisson with the Binomial for our numerical experiments, we wil have to sample several more examples to replicate the original results.

(e)

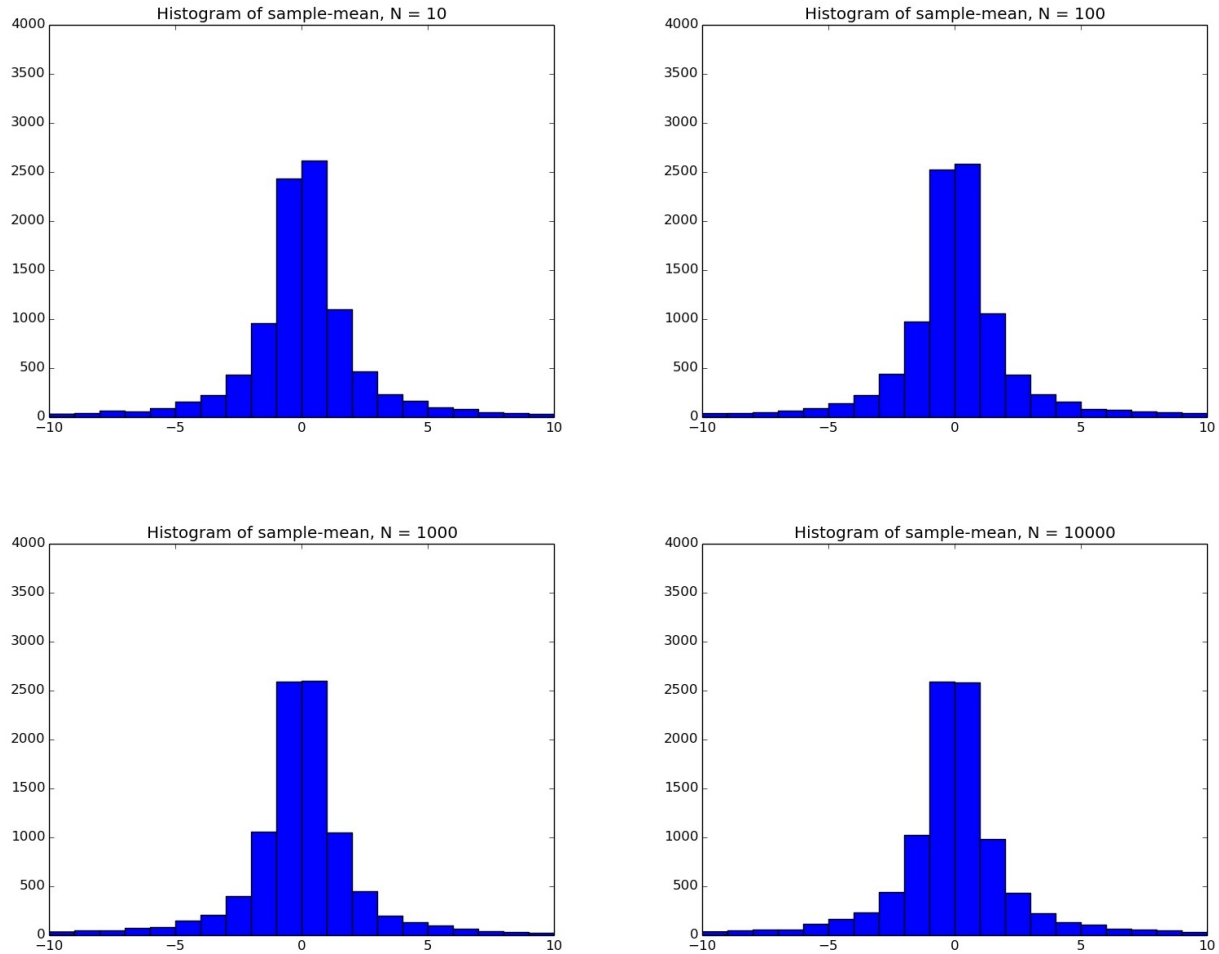
$\lvert\bar{X} - \mu\rvert$	**N @ % confidence**
0.1	2700 @ 90%
0.01	677000 @ 99%
0.001	109000000 @ 99.9%

$$N = \left(\frac{z\sigma}{|\bar{X} - \mu|}\right)^2$$

where, $z = -F^{-1}\left(\frac{|\bar{X} - \mu|}{2}\right)$. F^{-1} is the inverse of the Gaussian c.d.f.

Every addition of a decimal place contributes to at least an order of magnitude increase in the number of samples required. This could have been calculated at a fixed confidence, say at 95%, in which case the value of z will remain fixed. But such an estimate of N will give us the required accuracy only 95% of the time, i.e., if we perform the 10000 experiments, for 95% of them, the absolute difference between the true mean and sample mean will be under 0.1. Ideally we want this difference to be under 0.1 almost all the time (100%). But a 100% accuracy corresponds to a z value of ∞. This is one reason we have computed z for each increase in the decimal place.

Problem-4



(a)

$$f_X(k) = \frac{A}{k^2}, k \in \mathbb{Z} - \{0\}$$

For f to be a valid p.m.f, its sum over the support should be one.

$$\sum_{k=-\infty}^{\infty} \frac{A}{k^2} = 1 \implies A\frac{\pi^2}{3} = 1 \implies A = \frac{3}{\pi^2}$$

We have used the fact that the infinite sum $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$ (Basel problem).

(b)

Method : To sample from this p.m.f, we consider a modified distribution of a random variable Y, with a positive support:

$$f_Y(k) = \frac{A^+}{k^2}, k \in \mathbb{Z}^+$$

where $A^+ = \frac{6}{\pi^2}$. We note that $A^+ = 2A$.

f_Y is a valid p.m.f. Let S be a Bernoulli random variable with $p = \frac{1}{2}$. Now define $Z = (1 - 2S)Y$. The claim is that Z has the same p.m.f as X. To see why this is the case, we can first write the p.m.f of Z as a piecemeal function:

$$\begin{aligned} f_Z(k) &= \mathbb{P}(S = 0 \text{ and } Y = k), k > 0 \\ &= \mathbb{P}(S = 1 \text{ and } Y = k), k < 0 \end{aligned}$$

Since S and Y are independent, this further simplifies as:

$$\begin{aligned} f_Z(k) &= \frac{1}{2} \cdot f_Y(k), k \in \mathbb{Z} - \{0\} \\ &= \frac{A}{k^2}, k \in \mathbb{Z} - \{0\} \\ &= f_X(k), k \in \mathbb{Z} - \{0\} \end{aligned}$$

We first sample from Y. Generate a uniform random number u_1, return the smallest positive integer, y, for which the c.d.f of Y exceeds u_1. Now generate another uniform random number u_2. If $u_2 > 0.5$, return y, else return $-y$.

The sample mean concentrates around 0. This is because the true mean of X is 0. A few confidence intervals for the sample mean:

N = 1000	N = 10000
(-51, 17)	(-14, 5)
(-22, 75)	(-3, 1)
(-1, 5)	(-13, 3)
(-14, 4)	(-0.5, 1)
97.55%	98.06%

The last row contains the percentage of intervals containing the true mean. One observation is that the variance in the endpoints of the interval is very high. Let us look at the formula for the confidence interval:

$$[\bar{X} - z\frac{\sigma}{\sqrt{N}}, \bar{X} + z\frac{\sigma}{\sqrt{N}}]$$

The erratic behavior of the endpoints is due to the high variance, σ. Some of the sample variances are 23, 58, 281 and 606. Why is σ so high? Let us look at the p.m.f values of both Poisson and X around their respective means.

Poisson:

k	$\mathbb{P}(k)$
0	4e-5
5	3e-2
10	0.12
15	3e-2
20	1e-4
25	4e-12

For X in this problem:

k	$\mathbb{P}(k)$
±50	1e-4
±45	1e-4
±25	4e-4
±5	1e-2
±1	0.6

Note that the probability mass is decays rapidly on either side of the mean for Poisson. But the decay is more controlled for X. This is because the square in the denominator grows much slower than the factorical. As a consequence the probability mass is more evenly distributed in X. This explains the high value of the sample variance.

END