# CS6015 : Assignment 2

**Ans 1.** $X_1, X_2, \ldots X_N$ are independent and identically distributed random variables.

Each has mean $\mu$ and variance $\sigma^2$.

Sample Mean $= \overline{X_N} = \dfrac{X_1 + X_2 + \ldots + X_N}{N} = \dfrac{1}{N}\sum_{i=1}^{N} X_i$

Now, $E[\overline{X_N}] = E\left[\dfrac{X_1}{N} + \dfrac{X_2}{N} + \ldots + \dfrac{X_N}{N}\right]$

By linearity of expectation $= \dfrac{1}{N}\Big[E[X_1] + E[X_2] + \ldots E[X_N]\Big]$

$= \dfrac{1}{N}[\mu + \mu + \ldots \mu]$

$E[\overline{X_N}] = \dfrac{1}{N}(N\mu) = \mu.$

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

$Var(\overline{X_N}] = Var\left[\dfrac{X_1}{N} + \dfrac{X_2}{N} + \ldots + \dfrac{X_N}{N}\right]$

Since, $X_1, X_2 \ldots X_N$ are independent r.v.'s.

$Var[\overline{X_N}] = \dfrac{1}{N^2}\Big[Var(X_1) + Var(X_1) + \ldots Var(X_N)\Big]$

$= \dfrac{1}{N^2}\left[\sigma^2 + \sigma^2 + \ldots \sigma^2\right]$

$= \dfrac{1}{N^2} \cdot N\sigma^2$

$\Rightarrow Var[\overline{X_N}] = \dfrac{\sigma^2}{N}$

Ans 2. $\mathbb{P}(\overline{X}_N - \mu \geq \epsilon) \leq e^{\frac{-2N\epsilon^2}{(b-a)^2}}$ —— (1)

and $\mathbb{P}(\overline{X}_N - \mu \leq -\epsilon) \leq e^{\frac{-2N\epsilon^2}{(b-a)^2}}$ —— (2)

from (1) and (2)

$$\mathbb{P}(|\overline{X}_N - \mu| \geq \epsilon) \leq 2e^{\frac{-2N\epsilon^2}{(b-a)^2}} \quad \text{—— (3)}$$

The given form is

$$P\left(\mu \in [\overline{X}_N - \epsilon', \overline{X}_N + \epsilon']\right) \geq 1 - \delta \quad \text{—— (4)}$$

$\delta \in (0,1)$ and $\epsilon' > 0$

$$\Rightarrow P\left(\mu \notin [\overline{X}_N - \epsilon', \overline{X}_N + \epsilon']\right) \leq \delta$$

* $\mu \notin [\overline{X}_N - \epsilon', \overline{X}_N + \epsilon']$ which means

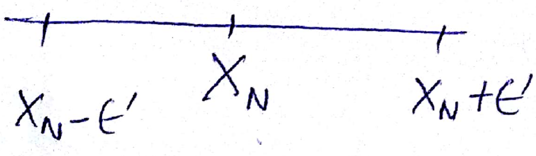
$\mu \leq \overline{X}_N - \epsilon'$ or $\mu \geq \overline{X}_N + \epsilon'$

This is covered in the Hoeffding inequality of eq (3)

Hence, $\mathbb{P}(|\overline{X}_N - \mu| \geq \epsilon') \leq \delta$ —— (5)

From eq (4) and (5);

$$\delta = 2e^{\frac{-2N\epsilon'^2}{(b-a)^2}} \quad \text{—— (6)}$$



Here, in eq (6), '$\delta$' is also called confidence level.

and $[\bar{X}_n - \epsilon', \bar{X}_n + \epsilon']$ is called confidence interval.

Given a value of confidence level '$\delta$' which is valid, we have to set the smallest value of $\epsilon'$ to get a valid confidence interval.

So, from eq (6)

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

$$\Rightarrow \frac{2}{\delta} = e^{\frac{2N\epsilon'^2}{(b-a)^2}}$$

Taking log,

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

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

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

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

Here $\epsilon'$ is a function of $N$ and $\delta$

Ans-3a) Yes, the sample mean is close to the true mean.

For different values of N, N scores were sampled from a uniform distribution 10,000 times and the mean computed each time.

From Central Limit Theorem we know that, for a given distribution with a mean $\mu$ and variance $\sigma^2$, the sampling distribution of the mean approaches a normal distribution with mean $\mu$ and a variance $\sigma^2/N$ (where N= sample size).

And emperically also, we have found that sample mean is close to the true mean.

Ans- 3b) Sample Mean in the interval [9.99, 10.01].

| Number of Samples (N) | Number of times in the interval [9.99, 10.01] | [9.9, 10.1] |
|---|---|---|
| 10 | 383 | 1178 |
| 100 | 394 | 2570 |
| 1000 | 881 | 6893 |
| 10000 | 2432 | 9989 |

Ans-3c)

| Number of Samples (N) | 95% Confidence Interval | # times true mean falls outside the interval |
|---|---|---|
| 10 | 9.388478, 10.599082 | 477 |
| 100 | 9.938316, 10.062068 | 471 |
| 1000 | 9.994077, 10.006398 | 498 |
| 10000 | 9.998954, 10.000195 | 473 |

Ans: 3 d)

- We can't apply Hoeffding's inequality for Poisson r.v.s because the inequality is defined on bounded random-variables whereas Poisson r.v. is unbounded ie $N \to \infty$.

- Poisson Approximation by Binomial:

Let $\lambda = np \Rightarrow \boxed{p = \frac{\lambda}{n}}$

where $n$ = # trials

$p$ = probability of success for each trial

$$\therefore P(X=x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{... Binomial Distribution}$$

$$\text{ie } P(X=x) = \frac{n!}{(n-x)!\,x!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \quad \text{... substituting } p = \lambda/n$$

$$= \frac{\lambda^x}{x!} \frac{n!}{(n-x)!} \left(\frac{1}{n^x}\right) \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-x}$$

$\therefore$ for large $n$, ie, $n \to \infty$:

$$P(X=x) = \frac{\lambda^x}{x!} \lim_{n\to\infty} \frac{n!}{(n-x)!} \left(\frac{1}{n^x}\right) \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-x}$$

$$= \left(\frac{\lambda^x}{x!}\right) (1) (e^{-\lambda}) (1) \quad \text{... using exponential series}$$

$$\boxed{P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!}} \quad \text{... Poisson Distribution}$$

So we have shown that the Poisson distribution is just a special case of the binomial, in which the number of 'n' trials grows to infinity and the chance of success in any particular trial approaches zero.

- Using the formula $\sqrt{\frac{1}{2N}\log\frac{2}{\delta}} = \epsilon$

Here confidence interval is 95%. So, $\delta = 0.05$.

$\epsilon = 0.42946$ (for N = 10)
$\epsilon = 0.13581$ (for N = 100)
$\epsilon = 0.0429469$ (for N = 1000)
$\epsilon = 0.0135810$ (for N = 10,000)

| No. of Samples (N) | 95% confidence Interval (theoretical) | 95% confidence interval (numerical) |
|---|---|---|
| 10 | 9.570, 10.42946 | 10.00, 10.00 |
| 100 | 9.864, 10.135 | 9.939, 10.056 |
| 1000 | 9.9570, 10.0429 | 9.9915, 10.004 |
| 10,000 | 9.9864, 10.0135 | 9.998, 10.000 |

Ans 3e) Suppose we want 95% confidence interval. And it is given that we need an accuracy of 0.1

Then $n \geq \frac{1}{2\epsilon^2} \log \frac{2}{\delta}$ —— (1)

$\delta = 0.05, \quad \epsilon = 0.1$

$n \geq 184.44$

$\therefore$ 185 samples are necessary.

If the accuracy is 0.01, ~~18444~~,
18445 samples ~~are~~ are necessary.

From eq(1), in general if the accuracy is increased by 1 decimal, then the number of samples has to be multiplied by 100.

Ans 4 a) given p.m.f. is:

$$f(k) = \frac{A}{k^2} \quad \text{for } k = \pm 1, \pm 2, \ldots$$

To be a valid p.m.f,

$$\sum_k f(k) = 1$$

$$\therefore \sum_k \frac{A}{k^2} = 1$$

$$\therefore A \sum_{k \neq 0} \frac{1}{k^2} = 1$$

$$\Rightarrow A \sum_{k=1}^{\infty} \frac{2}{k^2} = 1$$

$$\Rightarrow 2A \sum_{k=1}^{\infty} \frac{1}{k^2} = 1$$

$$\Rightarrow 2A \frac{\pi^2}{6} = 1 \qquad \because \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$$

$$\Rightarrow \boxed{A = \frac{3}{\pi^2}}$$

So for $A = \frac{3}{\pi^2}$, $f(x)$ becomes a valid p.m.f.

4) b)

| No of samples (N) | 95% confidence Interval |
|---|---|
| 1000 | −0.028014, 0.013460 |
| 10000 | −0.006830, −0.002668 |

Practically, the sample mean is concentrating towards zero.

But theoretically the mean is ~~tenk~~ undefined for the given p.m.f

because

$$E[k] = \sum_{k \neq 0} k \cdot \frac{3}{\pi^2 k^2} = \frac{3}{\pi^2} \sum_{k \neq 0} \frac{1}{k}$$

is undefined.

because $E[k^-] + E[k^+] = E[k]$

and both $E[k^-]$ and $E[k^+]$ are undefined