Recall the following from an earlier Homework. Suppose that the number of earthquakes that occur in a year, anywhere in the world, is a Poisson random variable with mean $\lambda$. Suppose that the probability that any given earthquake has magnitude at least 5 on the Richter scale is $p$. We showed that the probability that there are $n$ earthquakes with magnitude at least 5 in a year is $\frac{e^{-\lambda p}(\lambda p)^n}{n!}$.

This is an example of a *Thinning of Poisson points.* More precisely suppose we generate Poisson (1) data, $X$. Say the sample is 16 points. Then for each of sixteen points we toss a coin with probability $\frac{1}{2}$ and decide to keep them. Then let $Y$ be the number of points kept. As in the above problem one can observe the following facts:

$$X \sim \text{Poisson}(1)$$
$$Y \mid X = n \sim \text{Binomial}\left(n, \frac{1}{2}\right)$$
$$Y \sim \text{Poisson}\left(\frac{1}{2}\right)$$

In this worksheet we will verify this via simulation in `R`.

1. **Simulation:** The below code will generate 1000 samples of a thinned poisson points from Poisson(20) with the probability of keeping a point being 0.25.

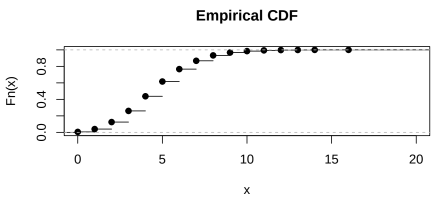
```
> lambda=20
> prob=0.25
> func=function(){sum(runif(rpois(1,lambda))<prob)}
> a=replicate(1000,func())
```

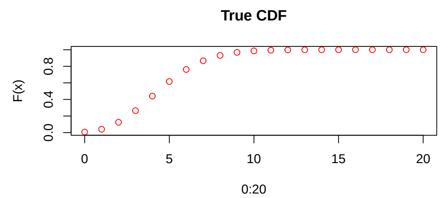
   Explain how the function `func()` generates one sample of a thinned Poisson points from Poisson(20) with the probability of keeping a point being 0.25. Explain what the function `replicate` does.

2. **Empirical CDF versus True CDF:** Suppose $Y_1, Y_2, \ldots Y_{1000}$ were the simulated sample. Then the empirical CDF is given by

$$F_{1000}(x) = \sum_{i=1}^{1000} 1(Y_i \leq x) \text{ as } x \text{ varies}$$

   The below code plots the empirical cumulative distribution function of the thinned process and the true cummulative distribution function of Poisson(5) (i.e. the distribution of the thinned Poisson process).






```
> plot(ecdf(a),main="Empirical CDF", xlim=c(0,20))
> #__(plotting the empirical CDF from the samples)
> plot(0:20,ppois(0:20,lambda*prob),col='red',
+      ylab="F(x)",main="True CDF", xlim=c(0,20))
> #__(plotting the true CDF using ppois)
```

Compare the two plots to verify that the two distributions match.

Modify the above R-codes to generate 5000 samples of a thinned poisson points from Poisson(21) with the probability of keeping a point being 0.65. Do the same verification via the empirical CDF and the true CDF.

Let the random variables $X$ and $Y$ be the results of rolling two fair dice. We will assume the rolls are independent. Define $Z = \min\{X, Y\}$. It is easy to check that

| $z$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $P(Z = z)$ | $\frac{11}{36}$ | $\frac{9}{36}$ | $\frac{7}{36}$ | $\frac{5}{36}$ | $\frac{3}{36}$ | $\frac{1}{36}$ |

In this worksheet we shall verify the above formulae by generating a sample in R and comparing it with the sample distribution.

1. **Simulation:** This is the same as done in Enrichment Worksheet 3. We leave it as an exercise to understand the code.

```
> func=function(){min(sample(1:6,1),sample(1:6,1))}
> #__(what is the above command doing?)
>
> a=sample(1:100,1)
> #__(Use R help for more information on the function 'sample')
>
> set.seed(a)
> #__(Refer to https://r-coder.com/set-seed-r/ for more information)
>
> table(replicate(10000, func()))/10000
> #__(calculating the empirical probabilities)
>
> c(11/36, 9/36, 7/36, 5/36, 3/36, 1/36)
> #__(vector of true probabilities)
```

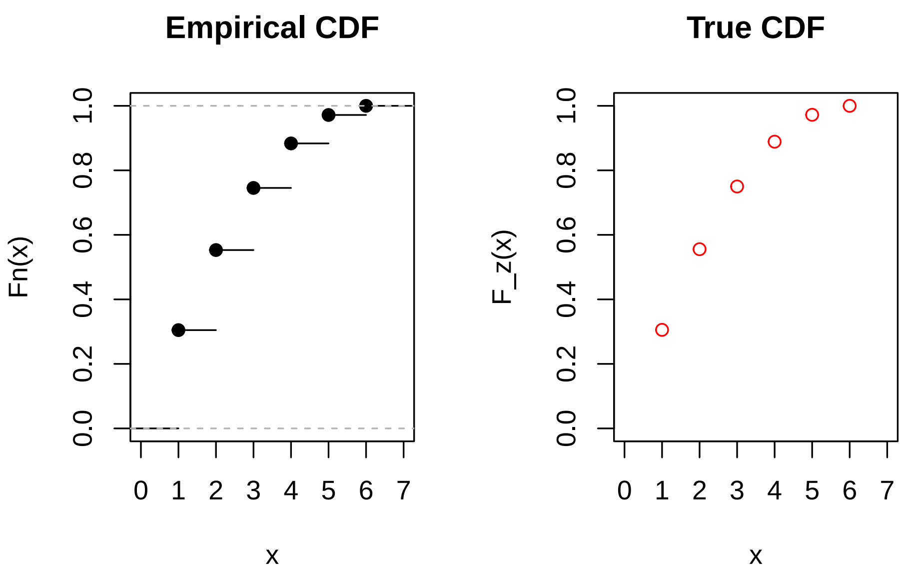
2. **Sample versus True: (Mean and Variance)** This is the same as done in Enrichment Worksheet 3. We leave it as an exercise to understand the code.

```
> set.seed(a)
> mean(replicate(10000, func()))
> #__(sample mean)
>
> sum(c(11/36, 9/36, 7/36, 5/36, 3/36, 1/36)*1:6)
> #__(true mean)
>
> var(replicate(10000, func()))
> #__(sample variance)
```

```
>
> sum(c(11/36, 9/36, 7/36, 5/36, 3/36, 1/36)*(1:6)^2)-
+   (sum(c(11/36, 9/36, 7/36, 5/36, 3/36, 1/36)*1:6))^2
> #__(true variance)
```

3. **Empirical CDF versus True CDF:** This is the same as done in Enrichment Worksheet 5. We leave it as an exercise to understand the code [1].

```
> set.seed(a)
> par(mfrow=c(1,2))
> plot(ecdf(replicate(10000, func())),main="Empirical CDF")
> plot(1:6,c(11/36, 20/36, 27/36, 32/36, 35/36, 36/36),col='red',
+      main="True CDF",ylab="F_z(x)",ylim=c(0,1),xlab="x",xlim=c(0,7))
```




[1] We can also use a Chi squared test to see if the given sample belongs to a particular population.