# Chapter 6. Convergence of Random Variables and Monte Carlo Methods

## 6.1 Type of convergence

The two main types of convergence are defined as follows.

---

Let $X_1, X_n, \cdots$ be a sequence of r.v.s, and $X$ be another r.v.

1. $X_n$ **converges to** $X$ **in probability**, denoted by $X_n \xrightarrow{P} X$, if for any constant $\epsilon > 0$, $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.

2. $X_n$ **converges to** $X$ **in distribution**, denoted by $X_n \xrightarrow{D} X$, if $\lim_n F_{X_n}(x) = F_X(x)$ for any $x$ (at which $F_X$ is continuous).

---

**Remarks**. (i) $X$ may be a constant (as a constant is a r.v. with probability mass concentrated on a single point.)

(ii) If $X_n \xrightarrow{P} X$, it also holds that $X_n \xrightarrow{D} X$, but not visa versa.

**Example 1**. Let $X \sim N(0,1)$ and $X_n = -X$ for all $n \geq 1$. Then $F_{X_n} \equiv F_X$. Hence $X_n \xrightarrow{D} X$. But $X_n \not\xrightarrow{P} X$, as for any $\epsilon > 0$

$$P(|X_n - X| > \epsilon) = P(2|X| > \epsilon) = P(|X| > \epsilon/2) > 0.$$

However if $X_n \xrightarrow{D} c$ and $c$ is a constant, it holds that $X_n \xrightarrow{P} c$.

**Note**. We need the two types of convergence.

For example, let $\widehat{\theta}_n = h(X_1, \cdots, X_n)$ be an estimator for $\theta$.

Naturally we require $\widehat{\theta}_n \xrightarrow{P} \theta$.

But $\widehat{\theta}_n$ is a random variable, it takes different values with different samples. To consider how good it is as an estimator for $\theta$, we hope that the distribution of $(\widehat{\theta}_n - \theta)$ becomes more concentrated around 0 when $n$ increases.

(iii) It is sometimes more convenient to consider the mean square convergence:

$$E\{(X_n - X)^2\} \to 0 \qquad \text{as } n \to \infty,$$

denoted by $X_n \xrightarrow{m.s.} X$. It follows from Markov's inequality that

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^2 > \epsilon^2) \le \frac{E\{|X_n - X|^2\}}{\epsilon^2}.$$

Hence if $X_n \xrightarrow{m.s.} X$, it also holds that $X_n \xrightarrow{P} X$, but not visa versa.

**Example 2**. Let $U \sim U(0,1)$ and $X_n = nI_{\{U<1/n\}}$. Then $P(|X_n| > \epsilon) \le P(U < 1/n) = 1/n \to 0$, hence $X_n \xrightarrow{P} 0$. However

$$E(X_n^2) = n^2 P(U < 1/n) = n \to \infty.$$

Hence $X_n \stackrel{m.s.}{\not\longrightarrow} 0$.

(iv) $X_n \xrightarrow{P} X$ does not imply $EX_n \to EX$.

**Example 3.** Let $X_n = n$ with probability $1/n$ and 0 with probability $1 - 1/n$. Then $X_n \xrightarrow{P} 0$. However $EX_n = 1 \not\to 0$.

(v) When $X_n \xrightarrow{D} X$, we also write $X_n \xrightarrow{D} F_X$, where $F_X$ is the CDF of $X$.

However the notation $X_n \xrightarrow{P} F_X$ does not make sense!

**Slutsky's Theorem**. Let $X_n, Y_n, X, Y$ be r.v.s, $g$ be a continuous function, and $c$ is a constant.
(i) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$, $X_n Y_n \xrightarrow{P} XY$, and $g(X_n) \xrightarrow{P} g(X)$.
(ii) If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} c$, then $X_n + Y_n \xrightarrow{D} X + c$, $X_n Y_n \xrightarrow{D} cX$, and $g(X_n) \xrightarrow{D} g(X)$.

---

**Note**. $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} Y$ does not in general imply $X_n + Y_n \xrightarrow{D} X + Y$.

Slutzky's theorem is very useful, as it implies, e.g., $\bar{X}_n^2 \xrightarrow{P} \mu^2$, and $\bar{X}_n / S_n \xrightarrow{P} \mu/\sigma$ (see Exercise 4.3).

Recall the limits of sequences of real numbers $x_1, x_2, \cdots$: if $\lim_{n\to\infty} x_n = x$ (or, simply, $x_n \to x$), we mean $|x_n - x| \to 0$ as $n \to \infty$.

For a sequence of r.v.s $X_1, X_2, \cdots$, we say $X$ is the limit of $\{X_n\}$ if $|X_n - X| \to 0$. Now there are some subtle issues here:

(i) $|X_n - X|$ is a r.v., it takes different values in the sample space $\Omega$. Therefore $|X_n - X| \to 0$ should hold (almost) on the entirely sample space. This calls for some probability statement.

(ii) Since r.v.s have distributions, we may also consider $F_{X_n}(x) \to F_X(x)$ for all $x$.

Recall two simple facts: for any r.v.s $Y_1, \cdots, Y_n$ and constants $a_1, \cdots, a_n$,

$$E\Big(\sum_{i=1}^{n} a_i Y_i\Big) = \sum_{i=1}^{n} a_i E Y_i, \qquad (1)$$

and if $Y_1, \cdots, Y_n$ are uncorrelated (i.e. $\mathrm{Cov}(Y_i, Y_j) = 0 \ \forall \ i \neq j$)

$$\mathrm{Var}\Big(\sum_{i=1}^{n} a_i Y_i\Big) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(Y_i). \qquad (2)$$

**Proof for (2)**. First note that for any r.v. $U$, $\text{Var}(U) = \text{Var}(U - EU)$. Because of (1), we may assume $EY_i = 0$ for all $i$. Thus

$$\text{Var}\Big(\sum_{i=1}^{n} a_i Y_i\Big) = E\Big(\sum_{i=1}^{n} a_i Y_i\Big)^2 = E\Big(\sum_{i=1}^{n} a_i^2 Y_i^2 + \sum_{i \neq j} a_i a_j Y_i Y_j\Big)$$

$$= \sum_{i=1}^{n} a_i^2 E(Y_i^2) + \sum_{i \neq j} a_i a_j E(Y_i Y_j) = \sum_{i=1}^{n} a_i^2 \text{Var}(Y_i) + \sum_{i \neq j} a_i a_j (EY_i)(EY_j)$$

$$= \sum_{i=1}^{n} a_i^2 \text{Var}(Y_i).$$

## 6.2 Two important limit theorems: LLN and CLT

Let $X_1, X_2, \cdots$ be IID with mean $\mu$ and variance $\sigma^2 \in (0, \infty)$. Let $\bar{X}_n$ denote the sample mean:

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n), \qquad n = 1, 2, \cdots .$$

We recall two simple facts:

$$E\bar{X}_n = \mu, \qquad \mathrm{Var}(\bar{X}_n) = \sigma^2/n.$$

---

**The (weak) Law of Large Numbers (LLN)**:

$$\text{As } n \to \infty,\ \bar{X}_n \xrightarrow{P} \mu.$$

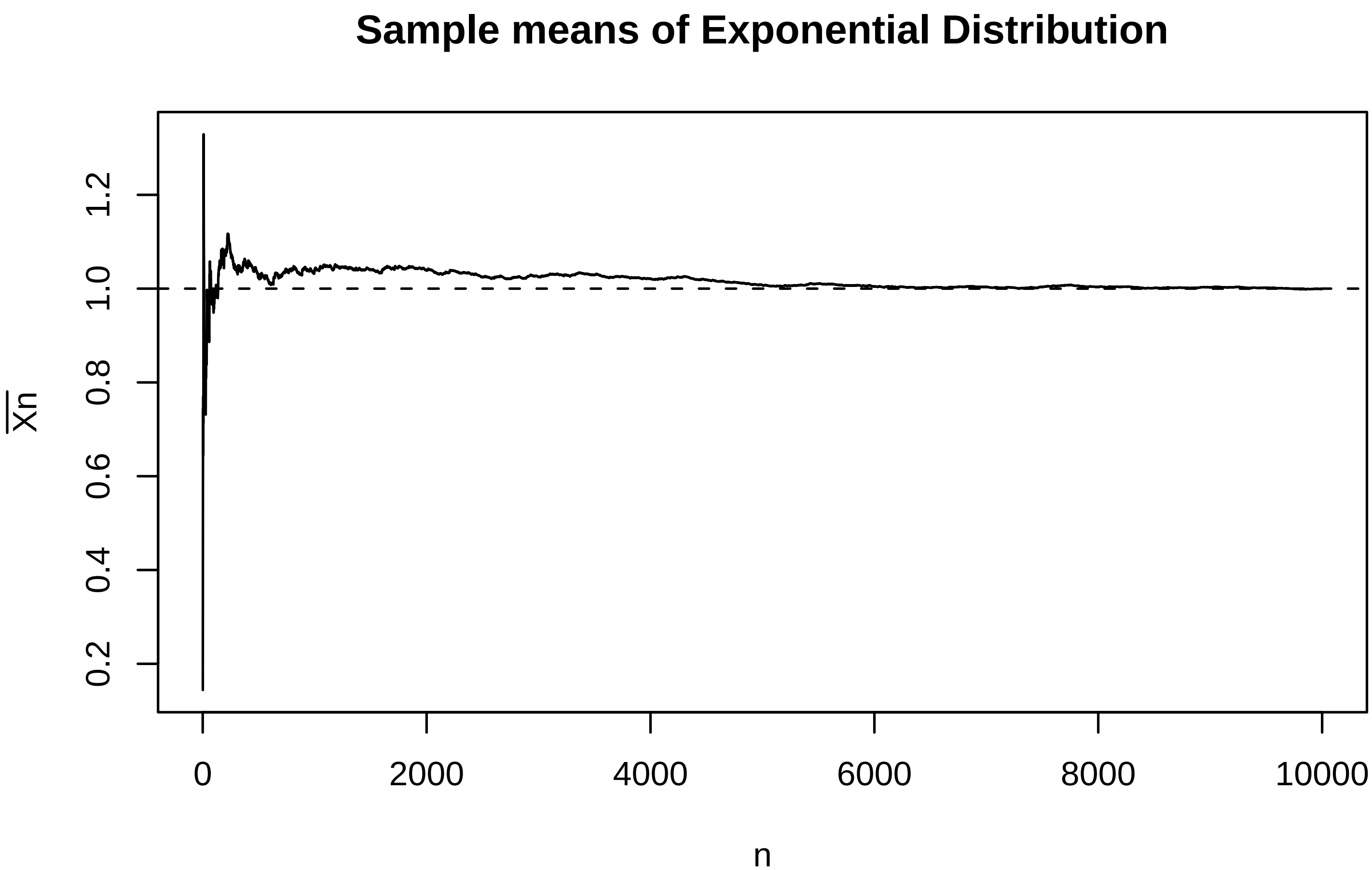
---

The LLN is very natural: When the sample size increases, the sample mean becomes more and more close to the population mean. Furthermore, the distribution of $\bar{X}_n$ degenerates to a single point distribution at $\mu$.

**Proof**. It follows from Chebyshev's inequality directly.

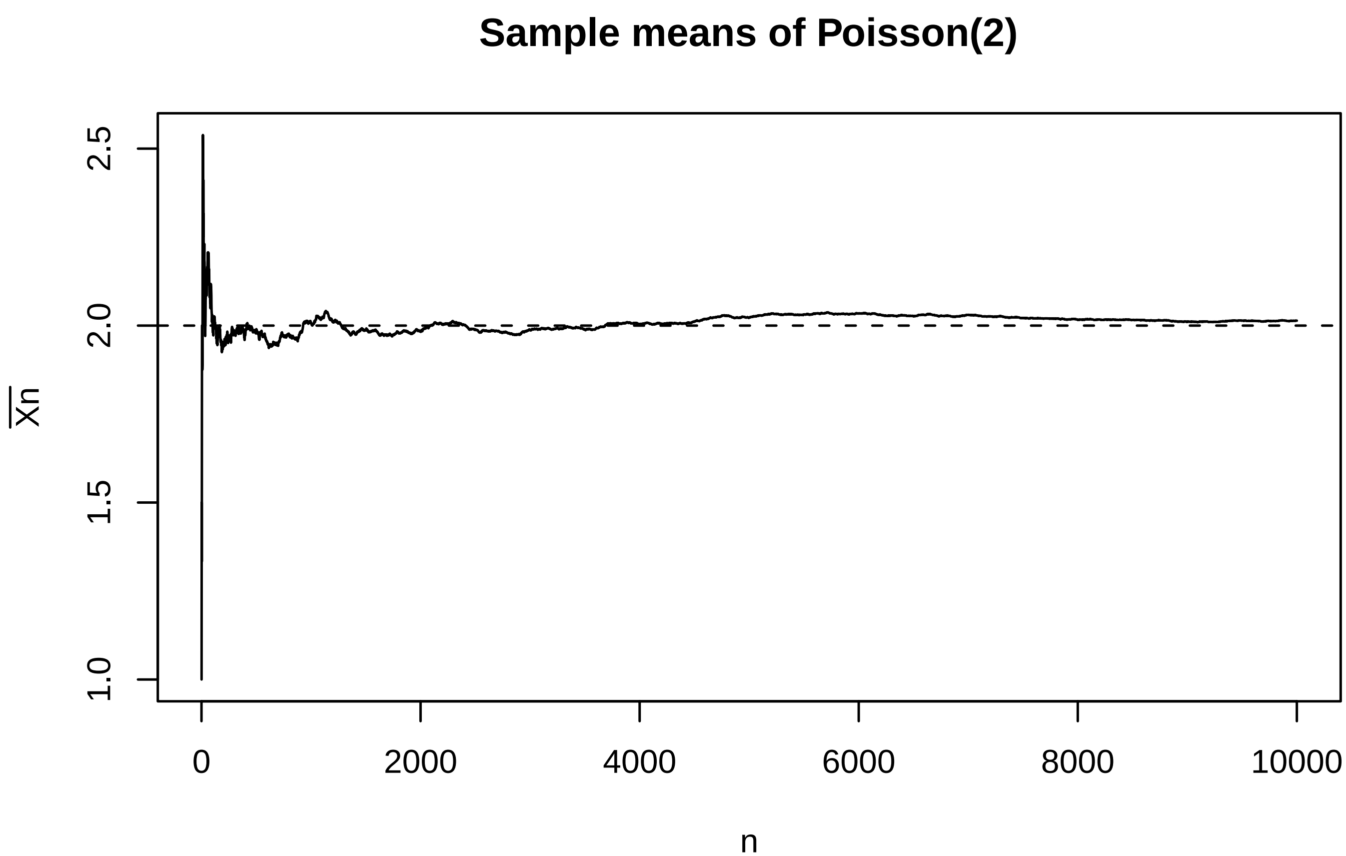
To visualize the LLN, we simulate the sample paths for

```
> x <- rexp(10000)   # generate 10000 random numbers from Exp(1)
> summary(x)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0001666 0.2861000 0.7098000 1.0220000 1.4230000 8.6990000
> n <- 1:10000
> ms <- n
> for(i in 1:10000) ms[i] <- mean(x[1:i])
> plot(n, ms, type='l', ylab=expression(bar(Xn)),
     main='Sample means of Exponential Distribution')
> abline(1,0,lty=2)  # draw a horizontal line at y=1
```



Sample means of Exponential Distribution
Xn
0.2
0.4
0.6
0.8
1.0
1.2
0
2000
4000
6000
8000
10000
n

We repeat this exercise for Poisson$(2)$:



**The Central Limit Theorem (CLT)**:

$$\text{As } n \to \infty,\ \sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{D} N(0,1).$$

Note the CLT can be expressed as

$$P\left\{\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \le x\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2/2}du = \Phi(x)$$

for any $x$, as $n \to \infty$, i.e. the *standardized* sample mean is approximately standard normal when the sample size is large. Hence in addition to $\sqrt{n}(\bar{X}_n - \mu)/\sigma \approx N(0,1)$, we also see the expressions such as

$$\bar{X}_n \approx N(\mu, \sigma^2/n), \quad \bar{X}_n - \mu \approx N(0, \sigma^2/n), \quad \sqrt{n}(\bar{X}_n - \mu) \approx N(0, \sigma^2).$$

**Note**. The CLT is one of the reasons why normal distribution is the most useful and important distribution in statistics.

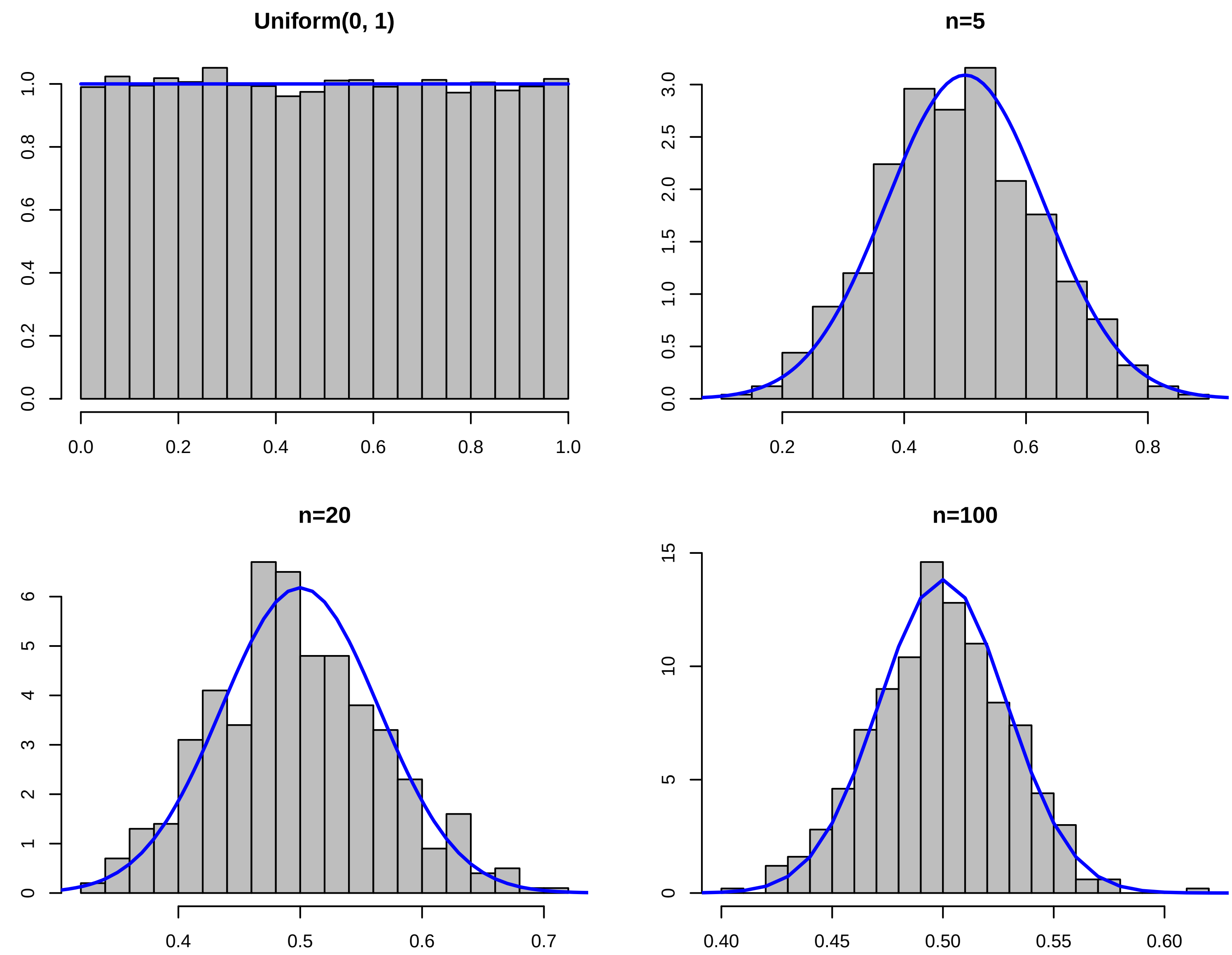
**Example 4**. If we take a sample $X_1, \cdots, X_n$ from $U(0,1)$, the standardized histogram will resemble the density function $f(x) = I_{(0,1)}(x)$, and the sample mean $\bar{X}_n = n^{-1}\sum_i X_i$ will be close to $\mu = EX_i = 0.5$, provided $n$ is sufficiently large.

However the CLT implies $\bar{X}_n \approx N(0.5, 1/(12n))$ as $\mathrm{Var}(X_i) = 1/12$. What does this mean?

If we take many samples of size $n$ and compute the sample mean for each sample, we then obtain many sample means. The standardized histogram of those samples means resembles the PDF of $N(0.5, 1/(12n))$ provided $n$ is sufficiently large.

```
> x <- runif(50000)  # generate 50,000 random numbers from U(0,1)
> hist(x, probability=T) # plot histogram of the 50,000 data
```

```
> z <- seq(0,1,0.01)
> lines(z,dunif(z)) # superimpose the PDF of U(0,1)
> x <- matrix(x, ncol=500)   # line up x into a 100x500 matrix
           # each column represents a sample of size 100
> par(mar=c(3,3,3,2),mfrow=c(2,2)) # plot 4 figures together
> meanx <- 1:500
> for(i in 1:500) meanx[i] <- mean(x[1:5,i])
        # compute the mean of the first 5 data in each column
> hist(meanx, probability=T, nclass=20, main='n=5')
> lines(z,dnorm(z,1/2,sqrt(1/(12*5))))
        # superimpose the PDF of N(.5, 1/(12*5))
> for(i in 1:500) meanx[i] <- mean(x[1:20,i])
> hist(meanx, probability=T, nclass=20, main='n=20')
> lines(z,dnorm(z,1/2,sqrt(1/(12*20))))
> for(i in 1:500) meanx[i] <- mean(x[1:60,i])
> hist(meanx, probability=T, nclass=20, main='n=60')
> lines(z,dnorm(z,1/2,sqrt(1/(12*60))))
> for(i in 1:500) meanx[i] <- mean(x[,i])
> hist(meanx, probability=T, nclass=20, main='n=100')
> lines(z,dnorm(z,1/2,sqrt(1/(12*100))))
```



Uniform(0, 1)
n=5
n=20
n=100

**Example 5**. Suppose a large box contains 10,000 poker chips distributed as follows

| Values of chips | \$5 | \$10 | \$15 | \$30 |
|---|---|---|---|---|
| No. of chips | 5000 | 3000 | 1000 | 1000 |

Take one chip randomly from the box, let $X$ be its nomination. Then its probability function is

| $X$ | 5 | 10 | 15 | 30 |
|---|---|---|---|---|
| probability | 0.5 | 0.3 | 0.1 | 0.1 |

Furthermore $\mu = EX = 10$ and $\sigma^2 = \text{Var}(X) = 55$.

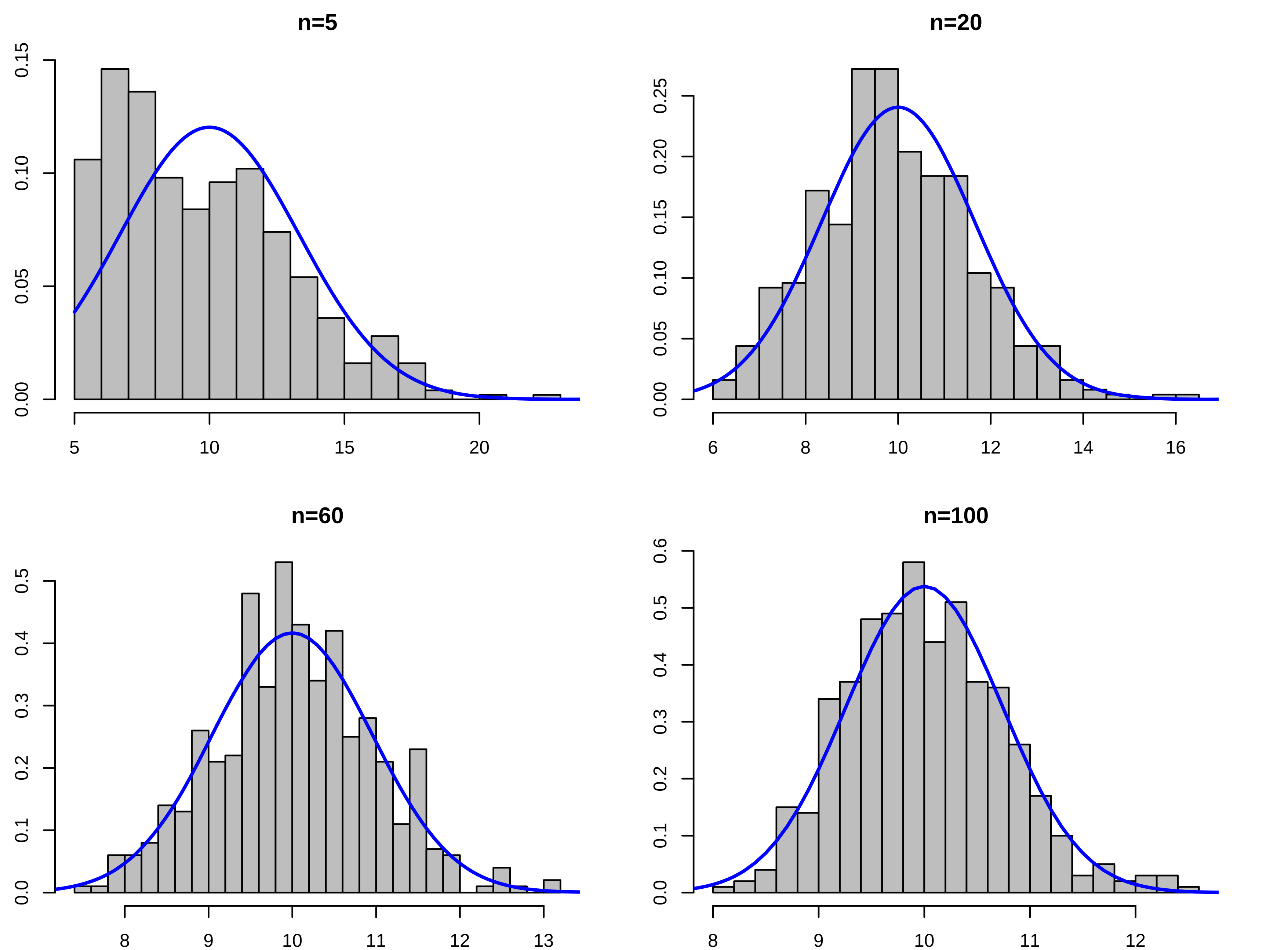
We draw 500 samples from this distribution, compute the sample means $\bar{X}_n$. When $n$ is sufficiently large, we expect $\bar{X}_n \approx N(10, 55/n)$.

We create a plain text file ‘porkerChip.r’ as below, which illustrate the central limiting phenomenon for the samples from this simple distribution.

```
y<- runif(50000) # generate 50,000 U(0,1) random numbers
x<- y
for(i in 1:50000)
    if(y[i]<0.5) x[i]<-5 else {
        if(y[i]<0.8) x[i]<-10 else {
            ifelse(y[i]<0.9, x[i]<-15, x[i]<-30)
        }
    }   # By now x are random numbers from the required distribution
        # of the poker chips
cat("mean", mean(x), "\n")
cat("variance", var(x), "\n")

x <- matrix(x, ncol=500)  # line up x into a 100x500 matrix
                          # each column represents a sample of size 100
par(mar=c(3,3,3,2),mfrow=c(2,2)) # plot 4 figures together

meanx <- 1:500
```

```r
z<-seq(5,25,0.1)

for(i in 1:500) meanx[i] <- mean(x[1:5,i])
        # compute the mean of the first 5 data in each column
hist(meanx, probability=T, main='n=5')
lines(z,dnorm(z,10,sqrt(55/5)))
       # draw N(10, 55/n) together with the histogram

for(i in 1:500) meanx[i] <- mean(x[1:20,i])
        # compute the mean of the first 20 data in each column
hist(meanx, probability=T, main='n=20')
lines(z,dnorm(z,10,sqrt(55/20)))

for(i in 1:500) meanx[i] <- mean(x[1:60,i])
        # compute the mean of the first 60 data in each column
hist(meanx, probability=T, main='n=60')
lines(z,dnorm(z,10,sqrt(55/60)))

for(i in 1:500) meanx[i] <- mean(x[,i])
        # compute the mean of the whole 100 data in each column
hist(meanx, probability=T, main='n=100')
```

```
lines(z,dnorm(z,10,sqrt(55/100)))
```


n=5
n=20
n=60
n=100

**Example 6**. Suppose $X_1, \cdots, X_n$ is an IID sample. A natural estimator for the population mean $\mu = EX_i$ is the sample mean $\bar{X}_n$. By the CLT, we can easily gauge the error of this estimation as follows:

$$
\begin{aligned}
P(|\bar{X}_n - \mu| > \epsilon) &= P\big(\sqrt{n}|\bar{X}_n - \mu|/\sigma > \sqrt{n}\epsilon/\sigma\big) \approx P\{|N(0,1)| > \sqrt{n}\epsilon/\sigma\} \\
&= 2P\{N(0,1) > \sqrt{n}\epsilon/\sigma\} = 2\{1 - \Phi(\sqrt{n}\epsilon/\sigma)\}.
\end{aligned}
$$

With $\epsilon$, $n$ given, we can find the value $\Phi(\sqrt{n}\epsilon/\sigma)$ from the table for standard normal distribution, *if we know* $\sigma$.

**Remarks.** (i) Let $\epsilon = 2\sigma/\sqrt{n} = 2 \times \mathrm{STD}(\bar{X}_n)$ (as $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$), $P(|\bar{X}_n - \mu| < 2\sigma/\sqrt{n}) \approx 2\Phi(2) - 1 = 0.954$. Hence

*If one estimates $\mu$ by $\bar{X}_n$ and repeats it a large number times, about the 95% of times $\mu$ is within $2 \times \mathrm{STD}(\bar{X}_n)$-distance from $\bar{X}_n$.*

(ii) Typically $\sigma^2 = \mathrm{Var}(X_i)$ is unknown in practice. We estimate it using the sample variance

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

In fact it still holds that

$$\sqrt{n}(\bar{X}_n - \mu)/S_n \xrightarrow{D} N(0,1), \qquad \text{as } n \to \infty.$$

Similar to Example 6 above, we have now

$$P(|\bar{X}_n - \mu| > \epsilon) \approx 2\{1 - \Phi(\sqrt{n}\epsilon/S_n)\}$$

Let $\epsilon = S_n/\sqrt{n}$, $P(|\bar{X}_n - \mu| > \epsilon) \approx 2\{1 - \Phi(1)\} = 0.317$, or $P(|\bar{X}_n - \mu| < S_n/\sqrt{n}) \approx 1 - 0.317 = 0.683$.

Let $\epsilon = 2S_n/\sqrt{n}$, we obtain:

$$P(|\bar{X}_n - \mu| < 2S_n/\sqrt{n}) \approx 0.954.$$

Hence

> *If one estimates $\mu$ by $\bar{X}_n$ and repeats it a large number times, about the 95% of times the true value is within $(2S_n/\sqrt{n})$-distance from $\bar{X}_n$.*

**Standard Error**: $\mathrm{SE}(\bar{X}_n) \equiv S_n/\sqrt{n}$ is called the standard error of the sample mean. Note

$$\mathrm{SE}(\bar{X}_n) = \left\{\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right\}^{1/2}.$$

**The Delta Method**. If $\sqrt{n}(Y_n - \mu)/\sigma \xrightarrow{D} N(0,1)$ and $g$ is a differentiable function and $g'(\mu) \neq 0$. Then

$$\frac{\sqrt{n}\{g(Y_n) - g(\mu)\}}{|g'(\mu)|\sigma} \xrightarrow{D} N(0,1).$$

Hence if $Y_n \approx N(\mu, \sigma^2/n)$, then $g(Y_n) \approx N(g(\mu), (g'(\mu))^2\sigma^2/n)$.

**Example 7**. Suppose $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{D} N(0,1)$ and $W_n = e^{\bar{X}_n} = g(\bar{X}_n)$ with $g(x) = e^x$. Since $g'(x) = e^x$, the Delta method implies $W_n \approx N(e^\mu, e^{2\mu}\sigma^2/n)$.

## 6.3 Monte Carlo methods

### 6.3.1 Basic Monte Carlo integration

The LLN may be interpreted as

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \int x f(x) dx$$

if $\{X_1, \cdots, X_n\}$ is a sample from the distribution with PDF $f$.

In general, for any function $h$, we apply the LLN to the sample $H_i \equiv h(X_i)$ $(i = 1, \cdots, n)$, leading to

$$\bar{H}_n \equiv \frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{P} E\{h(X_1)\} = \int h(x) f(x) dx. \tag{3}$$

**Monte Carlo integration method**: generate a sample $\{X_1, \cdots, X_n\}$ from PDF $f$, then the integral on the RHS of (3) may be approximated by the mean $\bar{H}_n$.

To measure the accuracy of this Monte Carlo approximation, we may use the standard deviation $\sigma/\sqrt{n}$ (if we know $\sigma^2 = \mathrm{Var}(H_1)$), or the standard error:

$$\Big(\frac{1}{n(n-1)} \sum_{i=1}^{n} \{h(X_i) - \bar{H}_n\}^2\Big)^{1/2}.$$

**Example 8**. (*Area of the quarter circle*) The area of a quarter of the unit circle is $\pi/4 = 0.7854$.

Suppose we do not know the answer. It can be written as

$$J \equiv \int_0^1 \sqrt{1-x^2}dx.$$

However it is not obvious how to solve this integral. We provide a Monte Carlo solution. Let

$$h(x) = \sqrt{1-x^2}, \quad f(x) = I_{(0,1)}(x).$$

Then $f$ is the PDF of $U(0,1)$ and

$$J = \int h(x)f(x)dx = E\{h(X)\},$$

where $X \sim U(0,1)$. Hence we generate a sample from $U(0,1)$ and estimate $J$ by

$$\widehat{J} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{1-X_i^2}, \quad \mathrm{SE} = \left\{\frac{1}{n(n-1)}\sum_{i=1}^{n}(\sqrt{1-X_i^2}-\widehat{J})^2\right\}^{1/2}.$$

The STD of $\widehat{J}$ is $\sigma/\sqrt{n}$, where

$$\sigma^2 = \mathrm{Var}(\sqrt{1-X_1^2}) = E(1-X_1^2) - (\frac{\pi}{4})^2 = \frac{2}{3} - (\frac{\pi}{4})^2 = 0.0498.$$

The R-function ‘quartercircle.r’ below perform this Monte Carlo calculation. It is used to produce the table

| $n$ | 1000 | 2000 | 4000 | 8000 |
|---|---|---|---|---|
| $\widehat{J}$ | .7950 | .7834 | .7841 | .7858 |
| STD | .0071 | .0050 | .0035 | .0025 |
| SE | .0072 | .0050 | .0036 | .0025 |

R-function ‘quartercircle.r’:

```r
quartercircle<-function(n)
        # This function calculates the area of the quarter circle
        # using Monte Carlo method
        # The true value is \pi/4 = 0.7854
        # n is the sample size
{
        x <- runif(n)
        h <- sqrt(1-x*x)
        list(quarterarea=mean(h), STD=sqrt(.0498/n),
             SE=sqrt(var(h)/n), SampleSize=n)
        # use 'list' to keep more than one outputs
}
```

You may call the function to perform the simulation:

```
> source("quartercircle.r")
> t=quartercircle(2000)
> summary(t)
            Length Class  Mode   
quarterarea 1      -none- numeric
STD         1      -none- numeric
SE          1      -none- numeric
SampleSize  1      -none- numeric
> t
$quarterarea
[1] 0.7913048
$STD
[1] 0.00498999
$SE
[1] 0.004946009
$SampleSize
[1] 2000
> t$quarterarea
[1] 0.7913048
```

### 6.3.2 Composition (Sequential sampling)

Let $X \sim f_X(\cdot)$, $Y|X \sim f_{Y|X}(\cdot|X)$. To obtain

$$Y_1, \cdots, Y_n \sim_{iid} f_Y(\cdot) \equiv \int f_{Y|X}(\cdot|x) f_X(x) dx,$$

we may repeat the composition below $n$ times:

> Step 1. Draw $X_i$ from $f_X(\cdot)$,
> Step 2. Draw $Y_i$ from $f_{Y|X}(\cdot|X_i)$.

Then $\{(X_i, Y_i), 1 \leq i \leq n\}$ are i.i.d. from the joint density

$$f_{X,Y}(x, y) = f_{Y|X}(y|x) f_X(x).$$

Hence $Y_1, \cdots, Y_n$ are i.i.d. from its marginal density $f_Y(\cdot)$.

**Remarks**.

(a) This method is applied when it is difficult to sample directly from $f_Y(\cdot)$.

(b) With $Y_1, \cdots, Y_n \sim_{iid} f_Y(y)$, we may estimate $E(Y)$ by $n^{-1} \sum_i \mathbf{Y}_i$. In general we estimate $E\{\psi(Y)\}$, for a known $\psi(\cdot)$, by

$$\bar{\psi} \equiv \frac{1}{n} \sum_{i=1}^{n} \psi(Y_i),$$

with the standard error

$$\frac{1}{\sqrt{n(n-1)}} \Big[ \sum_{i=1}^{n} \{\psi(Y_i) - \bar{\psi}\}^2 \Big]^{1/2}.$$

(c) The density function $f_Y(\cdot)$ may be estimated by

$$\widehat{f}_Y(y) = \frac{1}{n} \sum_{i=1}^{n} f_{Y|X}(y|X_i).$$

It also provides an estimate for $EY$: $\int y \widehat{f}_Y(y) dy$.

**Example 9**. Let $Y = X_1 + \cdots + X_T$, where $X_1, X_2, \cdots$ are IID Bernoulli($p$), $T \sim$ Poisson($\lambda$), and $T$ and $X_i$'s are independent. Then a sample from the distribution of $Y$ can be drawn as follows:

(i) Draw $T_1, \cdots, T_n$ independently from Poisson($\lambda$),
(ii) Draw $Y_i \sim$ Bin($T_i, p$), $i = 1, \cdots, n$, independently.

**Example 10**. Mixture of Normal distributions:

$$p\,N(\mu_1, \sigma_1^2) + (1-p)\,N(\mu_0, \sigma_0^2), \quad p \in (0,1),$$

(i.e. with PDF $\frac{p}{\sigma_1}\varphi(\frac{x-\mu_1}{\sigma_1}) + \frac{1-p}{\sigma_0}\varphi(\frac{x-\mu_0}{\sigma_0})$.)

A sample $X_1, \cdots, X_n$ can be drawn as follows:

(i) $I_1, \cdots, I_n \sim$ Bernoulli($p$) independently,

(ii) $X_i \sim N(\mu_{I_i}, \sigma_{I_i}^2)$, $i = 1, \cdots, n$, independently.

**Example 11**. The lifetime $X$ of a product follows the exponential distribution with mean $e^{1+U/4}$, where $U$ is a quality index of the raw materials used in producing the product and $U \sim N(\mu, \sigma^2)$. Find the mean, variance and the PDF of $X$ when $\mu = 1$ and $\sigma^2 = 2$.

As $X|U \sim Exp(e^{1+U/4})$ and $U \sim N(\mu, \sigma^2)$, we have

$$f_X(x) = \int f_{X|U}(x|u) f_U(u) du,$$

$$f_{X|U}(x|u) = e^{-(1+u/4)} \exp\{-xe^{-(1+u/4)}\} \quad \text{for } x > 0.$$

We use Monte Carlo simulation as follows:

1. Draw $U_1, \cdots, U_n$ from $N(\mu, \sigma^2)$
2. Draw $X_i$ from $Exp(e^{1+U_i/4})$, $i = 1, \cdots, n$.

Then the estimated mean for $X$ is $\bar{X}_n = n^{-1} \sum_i X_i$ with the standard error $\widehat{\sigma}/\sqrt{n}$, where

$$\widehat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

is an estimator for the variance of $X$. The estimated PDF is

$$\widehat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} f_{X|U}(x|U_i) = \frac{1}{n} \sum_{i=1}^{n} e^{-(1+U_i/4)} \exp\{-x e^{-(1+U_i/4)}\}$$

We write $R$-function `lifetimeMeanVar` to simulate $EX$ and $\mathrm{Var}(X)$, and `lifetimePDF` to produce the PDF $f_X$ and also $EX$.

```
lifetimeMeanVar <- function(n, mu, sigma2) {
       u <- rnorm(n, mu, sqrt(sigma2))
             # generate n random numbers from N(mu, sigma2)
       x <- u
       for(i in 1:n) x[i] <- rexp(1, 1/exp(1+u[i]/4))
         # x[i] is a random number from Exponential
         # distribution with mean e^{1+u[i]/4}
       vx <- var(x)
       list(Mean=mean(x), Min=min(x), Max=max(x),
                   StandardError=sqrt(vx/n), Var=vx)
}
```

The function is saved in the file 'lifetimeMeanVar.r', we source it into R and produce the required results:

```
> source("lifetimeMeanVar.r")
> outcome <- lifetimeMeanVar(500,1,2)
> outcome$Mean
[1] 3.763913
> outcome$Min
[1] 0.02139847
> outcome$Max
[1] 50.12281
> outcome$StandardError
[1] 0.1906219
> outcome$Var
[1] 18.16836
```

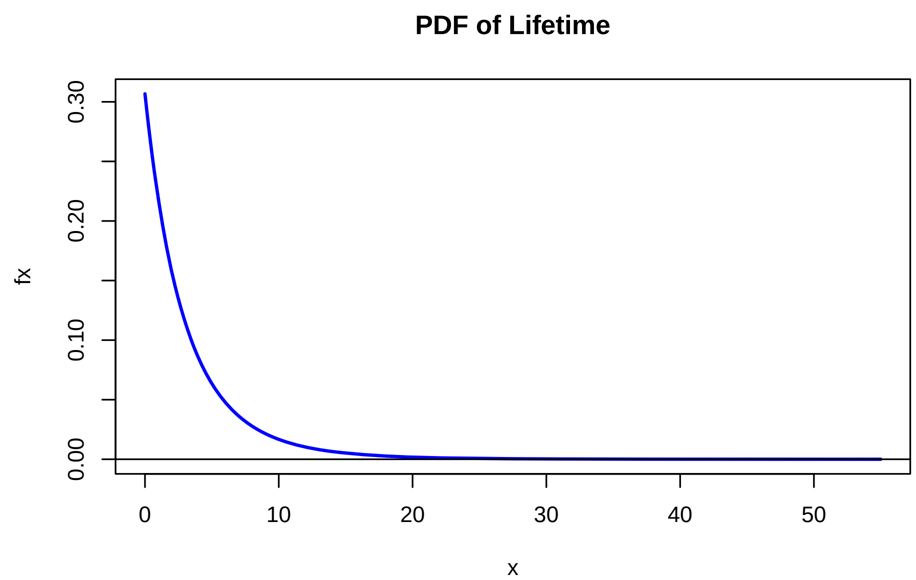
You may also try `summary(outcome)`.

The function `lifetimePDF` produces the PDF curve of $X$ in the given range `(xmin, xmax)`. It also computes $EX$ according to the estimated PDF.

```
lifetimePDF <- function(n,xmin,xmax,mu,sigma2) {
    u <- rnorm(n, mu, sqrt(sigma2))
    eu <- exp(-1-u/4)
    h <- (xmax-xmin)/400
    x <- seq(xmin, xmax, h)
    fx <- x
    for(i in 1:401) fx[i] <- mean(eu*exp(-x[i]*eu))
    m <- sum(x*fx*h)  # calculate the mean
    plot(x, fx, type='l', main="PDF of Lifetime")
    abline(0,0)   # abline(a,b) draw the straight line y=a+bx
    cat("Mean", m, "\n") # print out the mean
}  # Definition of function lifetimePDF' ends here
```

Source it into R to produce the required results:

```
> source("lifetimePDF.r")
> lifetimePDF(500,0,55,1,2)
> Mean 3.779971
```



### 6.3.3 Importance sampling

Let us consider the composition method discussed in section 6.3.2: To obtain an estimate for

$$f_Y(\cdot) = \int f_{Y|X}(\cdot|x) f_X(x) dx$$

or to obtain a sample from $f_Y(\cdot)$, we need to draw a sample $\{X_1, \cdots, X_n\}$ from $f_X(\cdot)$.

However sometimes we cannot directly sample from $f_X(\cdot)$. Importance sampling offers an indirect way to achieve this goal via an appropriately selected PDF $p(\cdot)$.

Let $p(\cdot)$ be a density satisfying:

(a) the support of $p$ contains the support of $f_X$,
i.e. $p(\mathbf{x}) = 0$ implies $f_X(\mathbf{x}) = 0$, and
(b) it is easy to sample from $p(\cdot)$.

**Importance sampling method** for approximating

$$J \equiv E\{h(X)\} = \int h(x) f_X(x) dx$$

(i) Draw $X_1, \cdots, X_n \sim_{i.i.d.} p(\cdot)$
(ii) Compute the estimator

$$\widehat{J} = \sum_{i=1}^{n} w_i h(X_i) \Big/ \sum_{i=1}^{n} w_i,$$

where $w_i = f_X(X_i)/p(X_i)$.

*Importance sampling* places weights greater than 1 on the regions where $f_X(x) > p(x)$, and downweights the regions where $f_X(x) < p(x)$.

**Choice of** $p(\cdot)$: as close to $f_X(\cdot)$ as possible among all PDF satisfying (a) and (b) in the previous page.

The standard error of $\widehat{J}$ is

$$\Big[\sum_{i=1}^{n}\{h(X_i) - \widehat{J}\}^2 w_i^2\Big]^{1/2} \Big/ \sum_{i=1}^{n} w_i.$$

which is inflated when $p(\cdot)$ poorly approximates $f_X(\cdot)$.

**Note**. $\sum_{i=1}^{n} w_i$ can be viewed as a version of the effective sample size in the importance sampling. When $p(\cdot)$ differs substantially from $f_X(\cdot)$, all $w_i$ are small. Hence the sampling is inefficient.

**Remark**. In the above calculation, we may *replace the PDF $f_X(\cdot)$ by $g(\cdot) \equiv C_0 f_X(\cdot)$*, where $C_0 > 0$ is an unknown constant. The algorithm stays the same but with the weights

$$w_i = g(X_i)/p(X_i).$$

For example, $f_X(x) = C_0^{-1} e^{-x^2/(|x|+2)}$, where the normalised constant $C_0 = \int e^{-x^2/(|x|+2)} dx$ is not easy to compute. In this case we may use $g(x) = e^{-x^2/(|x|+2)}$ instead of $f_X(x)$ in importance sampling.

**Proof of Remark**. By the LLN, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} w_i \xrightarrow{P} \int \frac{g(x)}{p(x)} p(x)dx = \int g(x)dx = C_0 \int f_X(x)dx = C_0,$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} w_i h(X_i) &\xrightarrow{P} \int \frac{g(x)}{p(x)} h(x)p(x)dx \\ = &\int g(x)h(x)dx = C_0 \int f_X(x)h(x)dx = C_0 E\{h(X)\}.\end{aligned}$$

Hence, by Slutzky's theorem,

$$\sum_{i=1}^{n} w_i h(X_i) \Big/ \sum_{i=1}^{n} w_i \xrightarrow{P} E\{h(X)\}.$$

**Application to sequential sampling**: $f_Y(\cdot) = \int f_{Y|X}(\cdot|x) f_X(x) dx$

(i) Draw $X_1, \cdots, X_N \sim_{i.i.d.} p(\cdot)$,
(ii) Draw $Y_i \sim f_{Y|X}(\cdot|X_i)$, $i = 1, \cdots, n$, independently.

Let $w_i = g(X_i)/p(X_i)$ and $\mu_y = E(Y)$, then

$$\widehat{f}_Y(y) = \sum_{i=1}^{n} w_i f_{Y|X}(y|X_i) \Big/ \sum_{i=1}^{n} w_i,$$

$$\widehat{\mu}_y = \sum_{i=1}^{n} w_i Y_i \Big/ \sum_{i=1}^{n} w_i,$$

which is guaranteed by the fact $(X_i, Y_i) \sim_{i.i.d.} p(x) f_{Y|X}(y|x)$.

**Note**. Importance sampling does not yield correct samples, as

$$X_i \nsim f_X(\cdot), \qquad Y_i \nsim f_Y(\cdot)$$

**Example 11** (Continue). Suppose now the quality index of the raw materials $U$ follows a generalised normal distribution with PDF

$$f_U(u) \propto \exp\Big\{-\frac{1}{2}\Big|\frac{u-\mu}{\sigma}\Big|^{\nu}\Big\} \equiv g(u)$$

where $\nu > 0$ is a constant. Recall

$$f_{X|U}(x|u) = e^{-(1+u/4)} \exp\{-xe^{-(1+u/4)}\} \quad \text{for } x > 0.$$

We adopt an importance sampling scheme as follows:

1. Draw $U_1, \cdots, U_n$ from $N(\mu, \sigma^2)$, compute the weight $w_i = g(U_i)/\phi(\frac{U_i-\mu}{\sigma})$, where $\phi$ denotes the PDF of $N(0,1)$.
2. Draw $X_i$ from $Exp(e^{1+U_i/4})$, $i = 1, \cdots, n$.

Then the estimated mean for $X$ is

$$\bar{X}_n = \sum_{i=1}^{n} w_i X_i \Big/ \sum_{i=1}^{n} w_i.$$

The estimated PDF is

$$\widehat{f}_X(x) = \frac{\sum_{i=1}^{n} w_i f_{X|U}(x|U_i)}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} w_i e^{-(1+U_i/4)} \exp\{-x e^{-(1+U_i/4)}\}}{\sum_{i=1}^{n} w_i}.$$

The R-function `lifetimeMeanIS` implements the above scheme for calculating $EX$:

```
lifetimeMeanIS <- function(n, mu, sigma2, nu) {
u=rnorm(n, mu, sqrt(sigma2)) #generate n numbers from N(mu, sigam2)
w=exp(-0.5*abs((u-mu)/sqrt(sigma2))^nu)/dnorm((u-mu)/sqrt(sigma2))
            # compute the weights w_i
x<-u
for(i in 1:n) x[i]<-rexp(1, 1/exp(1+u[i]/4))
list(Mean=sum(x*w)/sum(w), Min=min(x), Max=max(x))
}
```

The results for $\mu = 1$, $\sigma^2 = 2$ and $\nu = 0.5$ or 3 are as follows:

```
> source("lifetimeMeanIS.r")
> lifetimeMeanIS(5000,1,2,0.5)
$Mean
[1] 0.8827147
$Min
[1] 0.0003652474
$Max
[1] 57.21467
> lifetimeMeanIS(10000,1,2,3)
$Mean
[1] 1.616474
$Min
[1] 0.00125402
$Max
[1] 56.77547
```

The R-function `lifetimePDF.IS` implements the above scheme for estimating PDF $f_X$ and $E(X)$:

```
lifetimePDF.IS <- function(n,xmin,xmax,mu,sigma2,nu) {
  u <- rnorm(n, mu, sqrt(sigma2))
  Eu <- exp(-(1+u/4))   # Eu=e^{-(1+u/4)}
  w=exp(-0.5*abs((u-mu)/sqrt(sigma2))^nu)/dnorm((u-mu)/sqrt(sigma2))
          # compute the weights w_i
  sumw <- sum(w)
  h <- (xmax-xmin)/400
  x <- seq(xmin, xmax, h)
  fx <- x
  t <- 1:n
  m <- 0
  for(i in 1:401) {
      t <- Eu*exp(-x[i]*Eu)
    # t = PDF of  Exp(1/e^(1+u/4)) at x=x[i] --- THIS IS MORE
      fx[i] <- sum(t*w)/sumw
      m <- m+x[i]*fx[i]*h  # calculate the mean
  }
  plot(x, fx, type='l', main="PDF of Lifetime")
```

```
  abline(0,0)   # abline(a,b) draw the straight line y=a+bx
  cat("Mean", m, "\n")  # print out the mean
}
```

You may source it in, and try `lifetimePDF.IS(5000,0,60,1,2,0.5)` etc.

## Importance of using appropriate sampling distributions

An alternative measure for the effective sample size (ESS) is defined as $n/\{1+cv(w)\}$, where $cv(w)$ is the sample coefficient of variation of the weights

$$cv(w) = \Big\{\frac{1}{n-1}\sum_{i=1}^n (w_i - \bar{w})^2\Big\}^{1/2}\Big/\bar{w}, \qquad \bar{w} = \frac{1}{n}\sum_{i=1}^n w_i.$$

We illustrate the importance of choosing 'correct' $p(\cdot)$ in the example below.

**Example 12**. Estimate $\mu$ for $N(\mu,1)$ based on the importance sampling method using $N(0,1)$ as the sampling distribution $p(\cdot)$. The table below is produced by R-function `effectN` with $n=1000$.

| $\mu$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Estimated $\mu$ | -0.022 | 1.026 | 1.756 | 2.806 | 2.873 | 3.325 |
| ESS | 1000 | 448.9 | 246.1 | 113.4 | 65.7 | 33.8 |

```r
effectN=function(n, mu) {
x=rnorm(n)
w=dnorm(x,mu,1)/dnorm(x)  # sampling weights
muhat=mean(w*x)/mean(w)  # estimate for mu by importance sampling
ess=n/(1+sqrt(var(w))/mean(w)) # effective sample size
list(SampleSize=n, Mean=mu, EstimatedMean=muhat, ESS=ess)
}
```