Chapter 6. Convergence of Random Variables

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 θ.

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 θ, 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

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

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 $P(|X_n| > \epsilon) \leq P(U < 1/n) = 1/n \to 0$, hence $X_n \xrightarrow{P} 0$. However

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

Hence $X_n \xrightarrow{m.s.} 0$.

(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 Ω. 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 $a_1, \cdots, a_n$,

$$E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E Y_i, \tag{1}$$

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

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(Y_i). \tag{2}$$

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

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

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

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

6.2 Two important limit theorems: LLN and CLT

Let $X_1, X_2, \cdots$ be IID with mean μ 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):

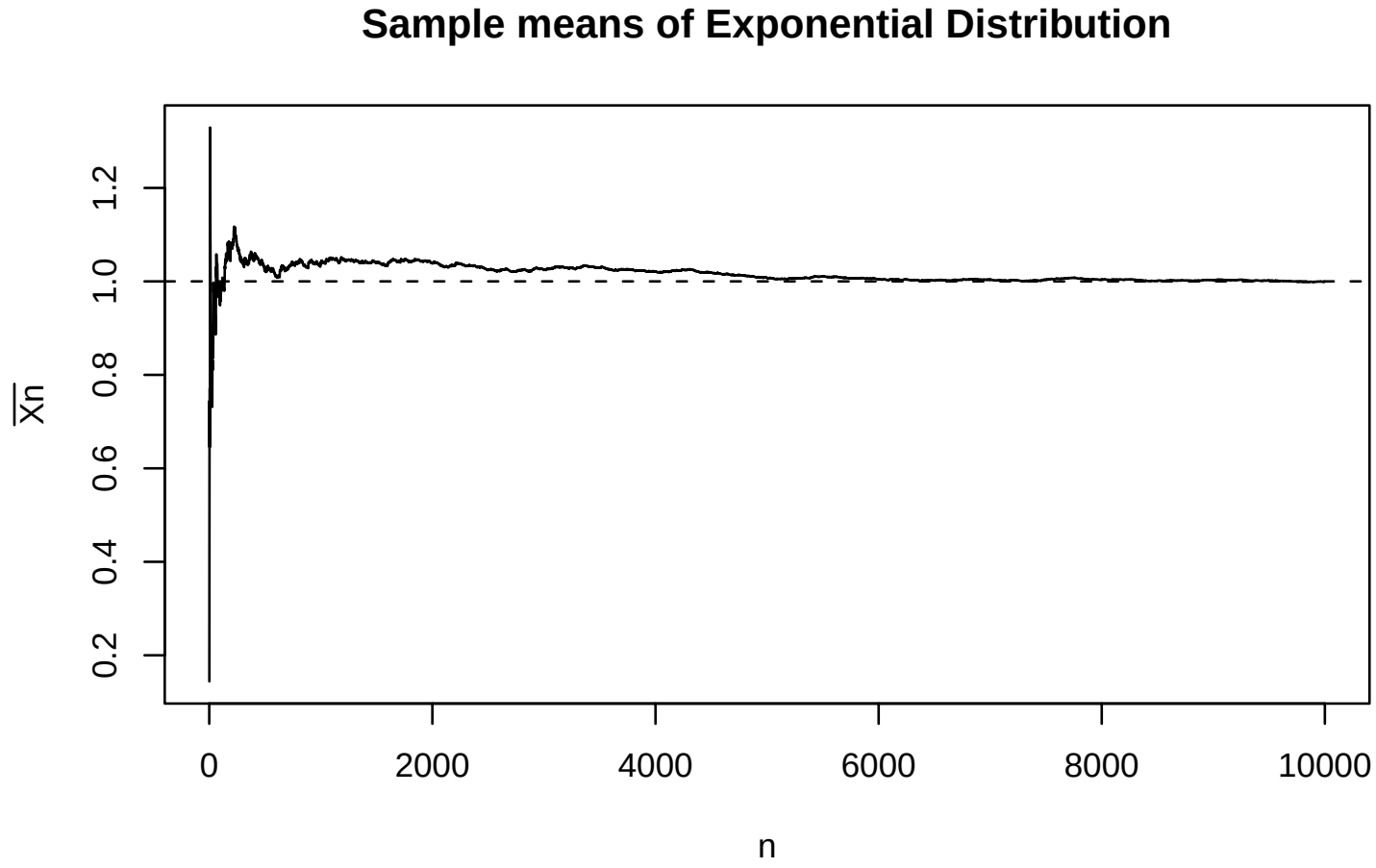
$$\text{As } n \to \infty, \bar{X}_n \xrightarrow{P} \mu.$$

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 μ.

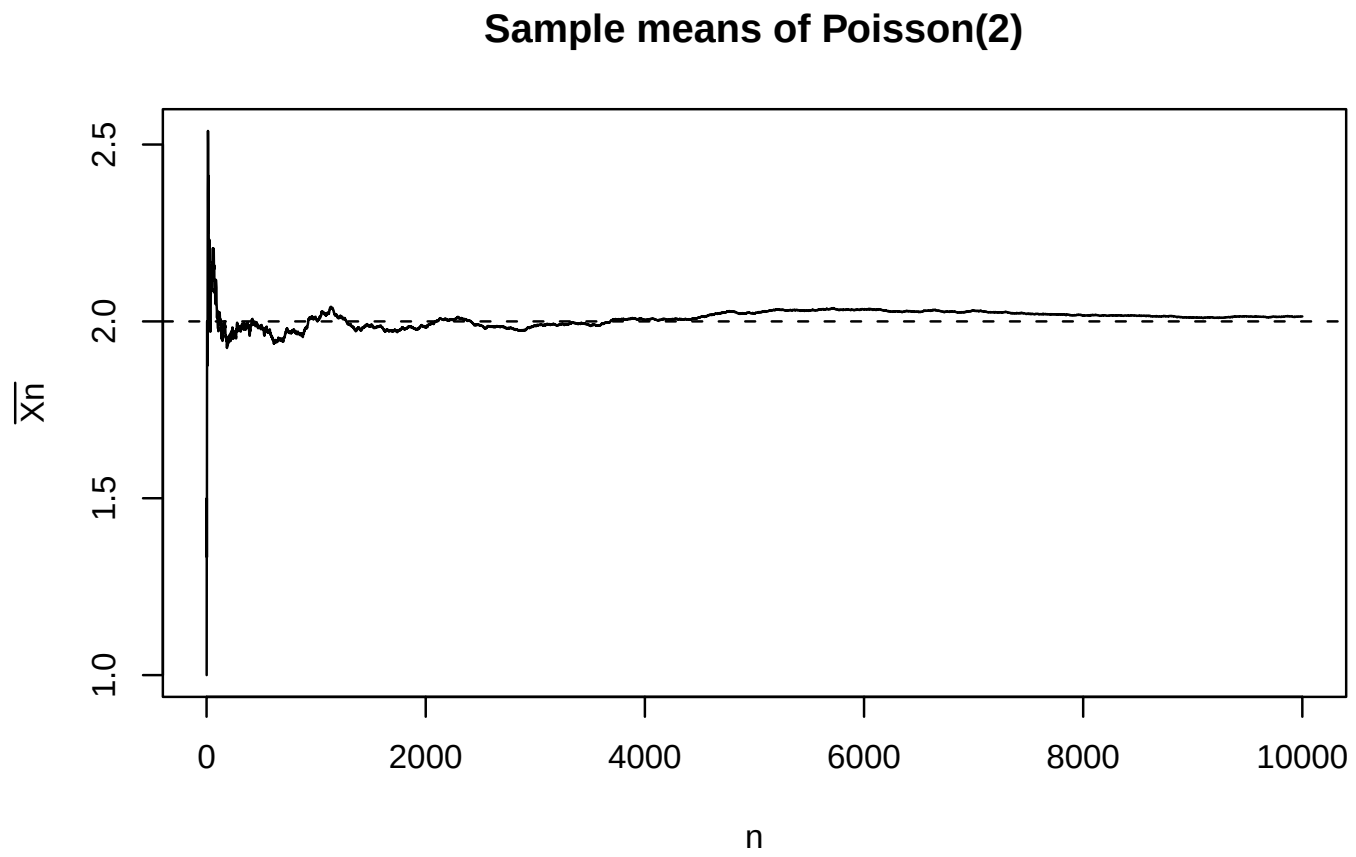
Proof. It follows from Chebyshev's inequality directly.

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
$\overline{Xn}$
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

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

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.

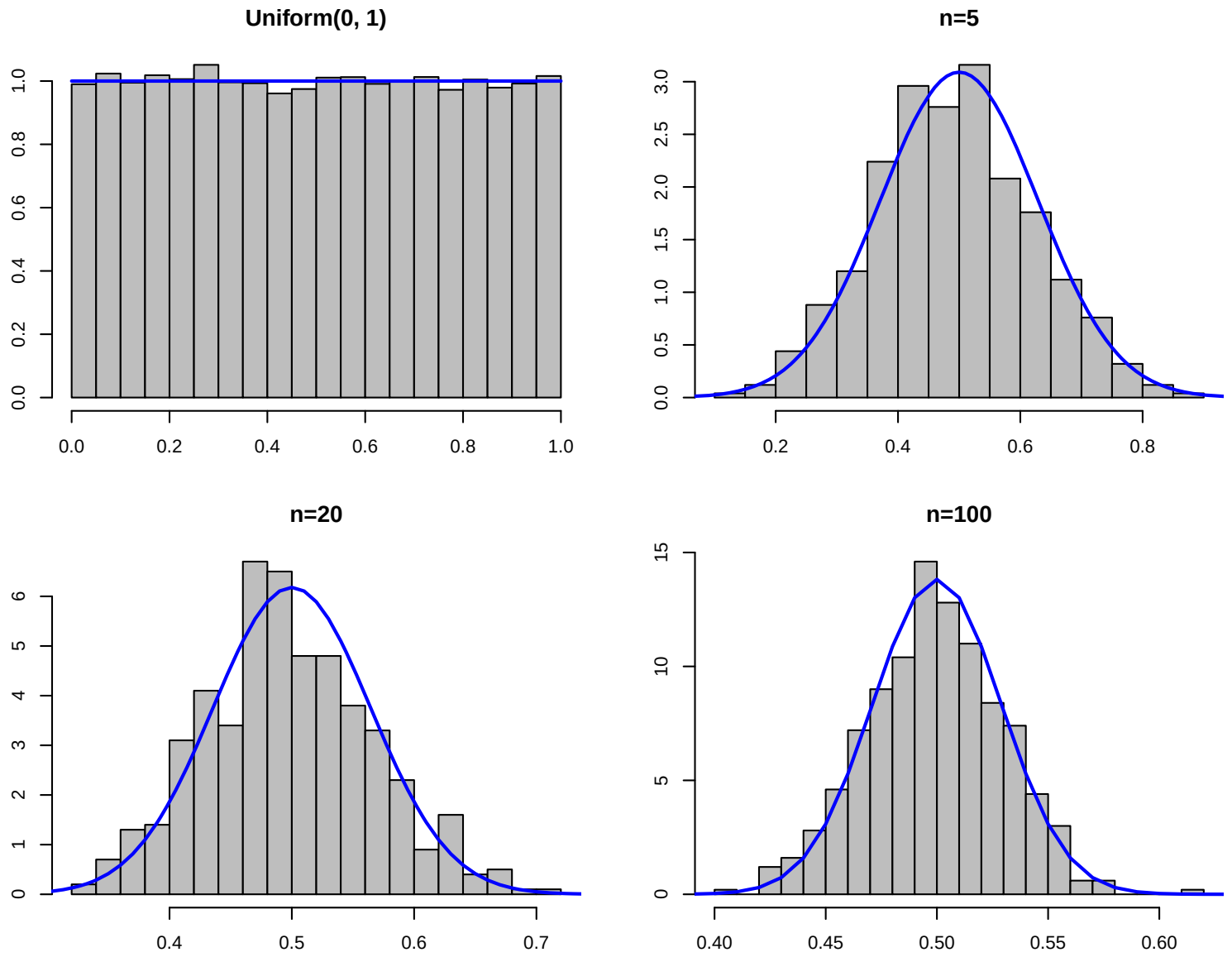
Example 4. If we take a sample $X_1, \cdots, X_n$ from $U(0,1)$, the standardized histogram will resemble the density function $f(x) = I_{(0,1)}(x)$, and the sample mean $\bar{X}_n = n^{-1} \sum_i X_i$ will be close to $\mu = EX_i = 0.5$, provided n is sufficiently large.

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
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
2.5
3.0
1
3
4
5
6
7
10
15
0.40
0.45
0.50
0.55
0.60

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 $\sigma^2 = \text{Var}(X) = 55$.

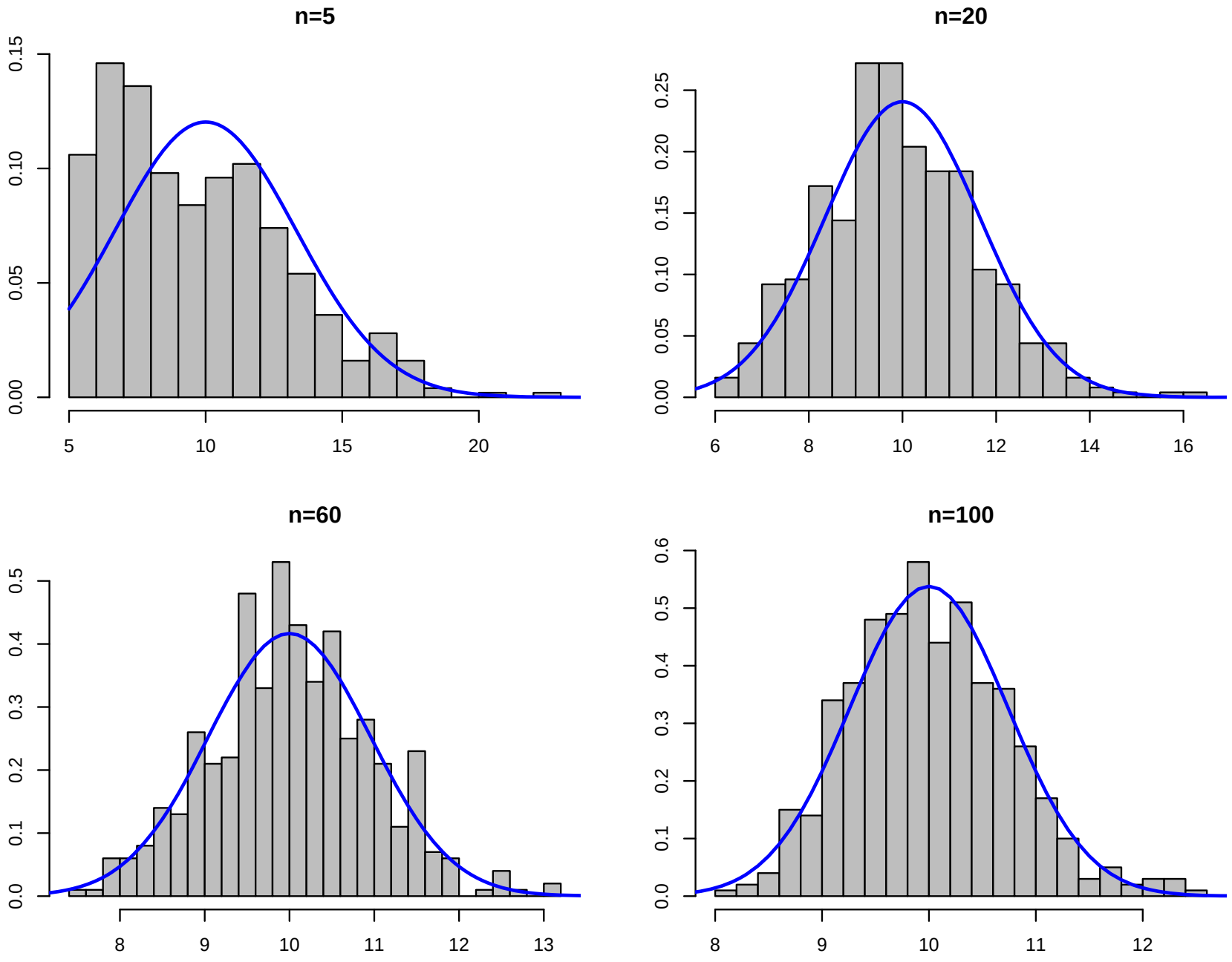
We draw 500 samples from this distribution, compute the sample means $\bar{X}_n$. When n is sufficiently large, we expect $\bar{X}_n \approx N(10, 55/n)$.

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
0.00
0.05
0.10
0.15
5
10
15
20
n=20
0.00
0.05
0.10
0.15
0.20
0.25
6
8
10
12
14
16
n=60
0.0
0.1
0.2
0.3
0.4
0.5
8
9
10
11
12
13
n=100
0.0
0.1
0.2
0.3
0.4
0.5
0.6
8
9
10
11
12

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:

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

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

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 μ by $\bar{X}_n$ and repeats it a large number times, about the 95% of times μ 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 μ 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 $W_n \approx N(e^\mu, e^{2\mu}\sigma^2/n)$.