CONVERGENCE THEOREMS

Suppose that $Y_1, Y_2, \ldots, Y_n$, are independent random variables that have the same distribution, and have expectation μ and variance σ^2.

- **The Weak Law of Large Numbers (WLLN):** For each n, let

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

be the sample mean random variable. Then $\overline{Y}_n \xrightarrow{p} \mu$ as $n \longrightarrow \infty$, that is, the probability distribution of $\overline{Y}_n$ becomes concentrated at μ.

- **The Central Limit Theorem (CLT):** The random variable

$$U_n = \frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma}$$

converges in distribution to a *standard Normal distribution*. We could also write

$$U_n = \frac{(\overline{Y}_n - \mu)}{\sigma/\sqrt{n}} \qquad \text{or} \qquad U_n = \frac{\left(\sum_{i=1}^{n} Y_i - n\mu\right)}{\sqrt{n}\sigma}.$$

Thus, when n is large, we can approximate the distribution of $\overline{Y}_n$, or of

$$S_n = \sum_{i=1}^{n} Y_i$$

using a Normal distribution, irrespective of the actual distribution of $Y_1, \ldots, Y_n$.

EXAMPLE: $Bernoulli(p)$.
If $Y_i \sim Bernoulli(p)$, then $\mathbb{E}[Y_i] = p$ and $\mathbb{V}[Y_i] = p(1-p)$. The WLLN suggests that

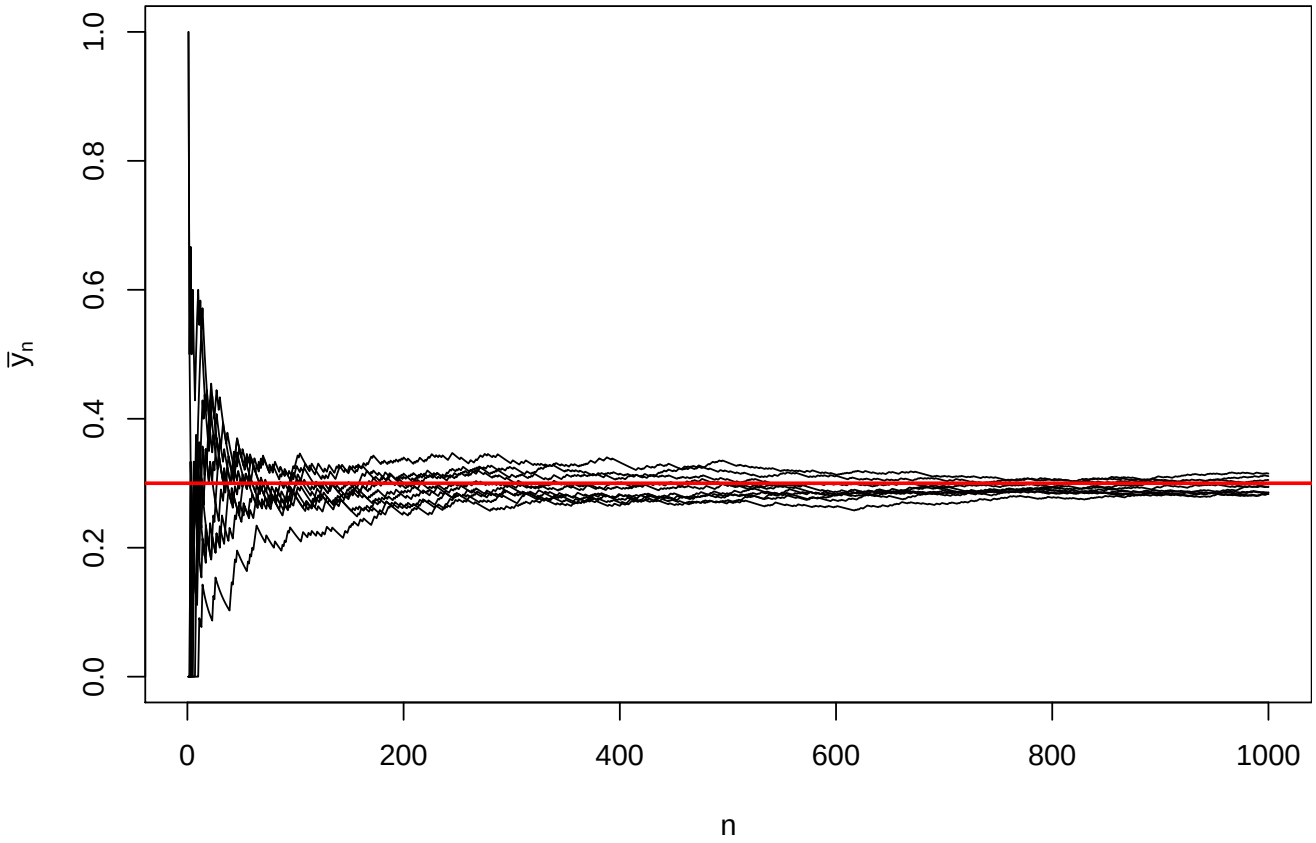
$$\overline{Y}_n \xrightarrow{p} p$$

whereas the CLT suggests that

$$U_n = \frac{\sqrt{n}(\overline{Y}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{d} Normal(0,1).$$

We can check these results in simulation: we generate `M` repeated sequences of length `N`, and study the behaviour of $\overline{Y}_n$ and U_n. In the simulation below, we choose $M = 10$ and $N = 1000$. We first plot the trace plots of $\overline{Y}_n$ for $n = 1, \ldots, N$ for 10 different replicated runs.

```
set.seed(1010)
M<-10
N<-1000
p<-0.3
ybar<-replicate(M,cumsum(rbinom(N,1,p))/c(1:N))
par(mar=c(4,4,1,0))
plot(1:N,ybar[,1],type='l',xlab='n',ylab=expression(bar(y)[n]),ylim=range(0,1))
for(j in 2:M){lines(1:N,ybar[,j])}
abline(h=p,col='red',lwd=2)
```



It is clear that the sample means are converging to $p = 0.3$ (marked as the red line) for each of the repeated runs.

To check the CLT, we increase M to 2000, and reduce N to 500. For each replicate run, and each n, we compute

$$U_n = \frac{\sqrt{n}(\overline{Y}_n - p)}{\sqrt{p(1-p)}}$$

which the CLT predicts will converge to a standard Normal random variable as n gets large.

```
M<-2000
N<-500
ybar<-replicate(M,cumsum(rbinom(N,1,p))/c(1:N))
Un<-(ybar-p)/sqrt(p*(1-p))
Un<-apply(Un,2,function(x){return(x*sqrt(1:N))})
```

We can check the claim of Normality by computing mean, variance and *quantiles* of the sampled values; the α quantile of a distribution is the value y_α such that

$$F(y_\alpha) = \alpha.$$

For example, the $\alpha = 0.5$ quantile is the *median*; the $\alpha = 0.25$ quantile is the *lower quartile*; the $\alpha = 0.75$ quantile is the *upper quartile* etc. We compare these summary values with the values of the standard Normal distribution for $n = 100, 200, 300, 400, 500$.

```
Un.sub<-t(Un[100*c(1:5),])
qvec<-c(0.025,0.25,0.5,0.75,0.975)
Un.summary<-apply(Un.sub,2,function(x){return(c(mean(x),var(x),quantile(x,qvec)))})
Un.summary<-cbind(Un.summary,c(0,1,qnorm(qvec)))
colnames(Un.summary)<-c(paste("n=",seq(100,500,by=100),sep=""),"Z")
rownames(Un.summary)[1:2]<-c("mean","var")
round(Un.summary,4)
```

```
+          n=100    n=200    n=300    n=400    n=500        Z
+ mean    0.0001 -0.0382 -0.0361 -0.0477 -0.0453  0.0000
+ var     1.0060  1.0054  1.0111  0.9963  0.9947  1.0000
+ 2.5%   -1.9640 -1.8516 -1.8898 -1.8549 -1.9518 -1.9600
+ 25%    -0.6547 -0.7715 -0.7559 -0.7638 -0.7807 -0.6745
+ 50%     0.0000  0.0000 -0.1260 -0.1091 -0.0976  0.0000
+ 75%     0.6547  0.6172  0.6299  0.6547  0.5855  0.6745
+ 97.5%   1.9640  2.0059  2.0158  1.9640  1.9518  1.9600
```

We can compute the pmf of U_n exactly using transformation methods; we know that

$$S_n = \sum_{i=1}^{n} Y_i \sim Binomial(n, p)$$

– this can be very simply proved using mgf methods – so that

$$p_{S_n}(s) = \binom{n}{s} p^s (1-p)^{n-s} \qquad s \in \{0, 1, 2, \ldots, n\}$$

with $p_{S_n}(s) = 0$ for other values of s. Therefore U_n also has a discrete distribution, with the same probabilities, but restricted to the set of values

$$u = \frac{\sqrt{n}(s/n - p)}{\sqrt{p(1-p)}}$$

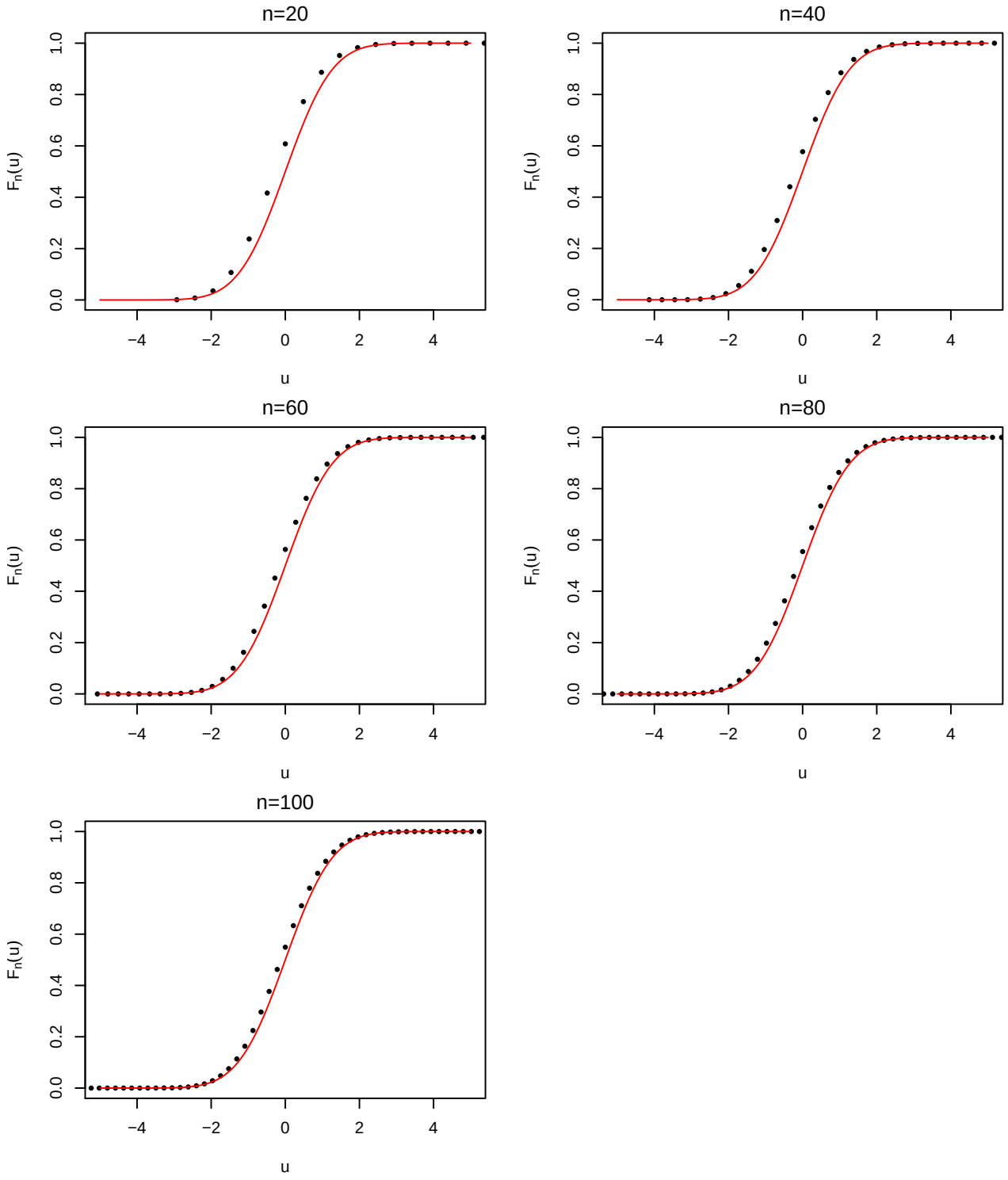
for $s \in \{0, 1, 2, \ldots, n\}$. We compare the *cumulative* probabilities associated with the distribution and its approximation; the CLT tells us that

$$F_{U_n}(u) = P(U_n \leq u) \backsimeq \Phi(u)$$

where $\Phi(.)$ is the standard Normal cdf. The computation is performed for $n = 20, 40, 60, 80, 100$. In `R`, `pbinom` computes the binomial cdf and `pnorm` computes the Normal cdf.

```
par(mar=c(4,4,2,2),mfrow=c(3,2))
nvec<-seq(20,100,by=20)
xv0<-seq(-2.5,2.5,by=0.5);yv0<-pnorm(xv0)
Fmat<-matrix(0,nrow=length(nvec),ncol=length(xv0))
xv<-seq(-5,5,by=0.01);yv<-pnorm(xv)
for(j in 1:length(nvec)){
    ivec<-0:nvec[j]
    uvec<-sqrt(nvec[j])*(ivec/nvec[j]-p)/sqrt(p*(1-p))
    Pvec<-pbinom(0:nvec[j],nvec[j],p)
    plot(uvec,Pvec,pch=19,cex=0.5,xlim=range(-5,5),
         main=substitute( paste("n=",nval),list(nval = nvec[j])),
         xlab=expression(u),ylab=expression(F[n](u)))
    lines(xv,yv,col='red')
    for(k in 1:length(xv0)){Fmat[j,k]<-ifelse(is.na(Pvec[max(which(uvec<=xv0[k]))]),0,
                                              Pvec[max(which(uvec<=xv0[k]))])}
}
Fmat<-rbind(yv0,Fmat);colnames(Fmat)<-xv0;
rownames(Fmat)<-c("Z",paste("n=",seq(20,100,by=20),sep=""))
round(Fmat,3)
```

```
+        -2.5    -2  -1.5    -1  -0.5     0   0.5     1   1.5     2   2.5
+ Z     0.006 0.023 0.067 0.159 0.309 0.500 0.691 0.841 0.933 0.977 0.994
+ n=20  0.001 0.008 0.035 0.107 0.238 0.608 0.772 0.887 0.952 0.983 0.995
+ n=40  0.003 0.024 0.055 0.196 0.309 0.577 0.703 0.807 0.937 0.968 0.994
+ n=60  0.006 0.014 0.057 0.162 0.342 0.563 0.669 0.838 0.937 0.980 0.990
+ n=80  0.004 0.016 0.053 0.135 0.275 0.555 0.732 0.863 0.941 0.979 0.994
+ n=100 0.005 0.016 0.076 0.163 0.296 0.549 0.711 0.837 0.920 0.979 0.993
```



Here the red line is the plot of the standard normal cdf; clearly the approximation is quite good even for $n = 40$.

EXAMPLE: Continuous $Uniform(0,1)$.
If $Y_i \sim Uniform(0,1)$, then $\mathbb{E}[Y_i] = 1/2$ and $\mathbb{V}[Y_i] = 1/12$. The WLLN suggests that

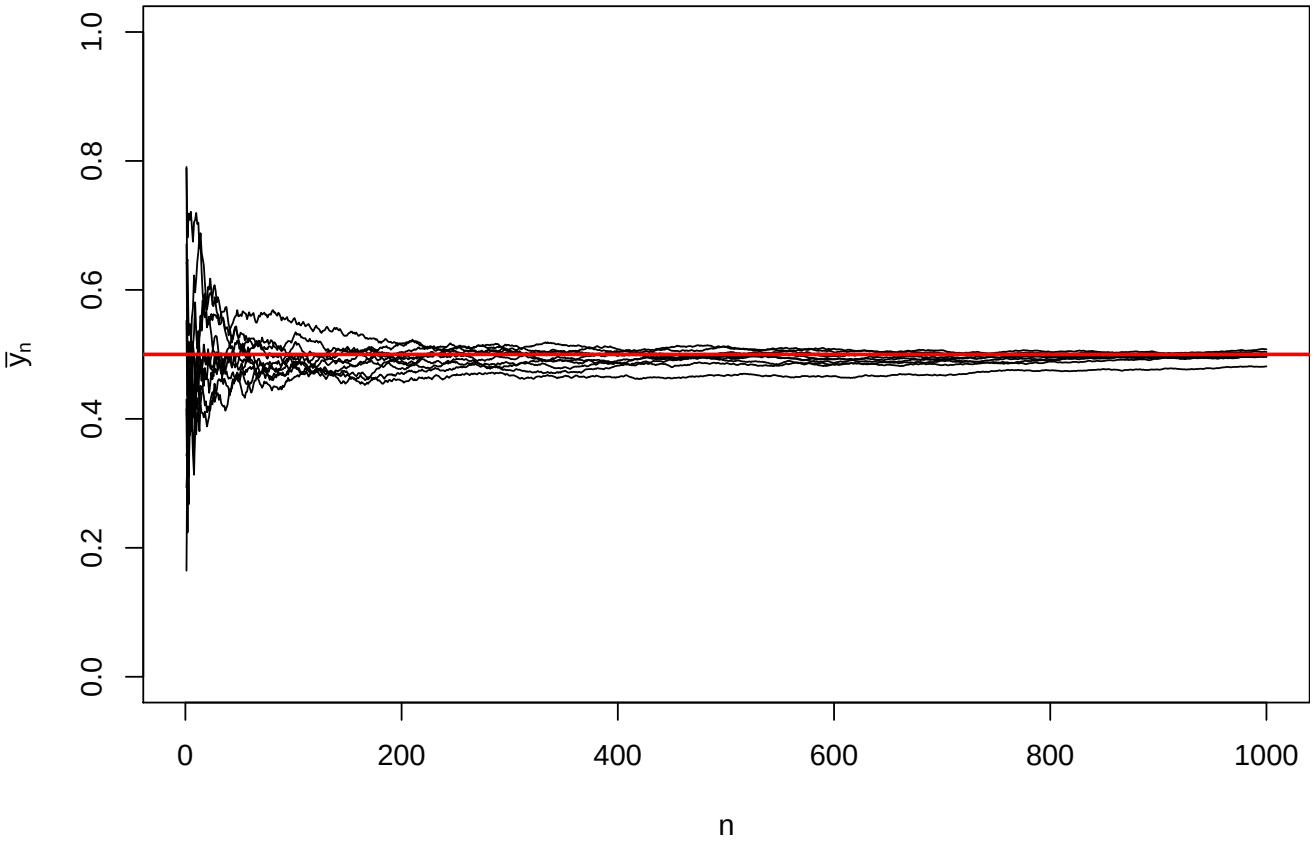
$$\overline{Y}_n \xrightarrow{p} \frac{1}{2}$$

whereas the CLT suggests that

$$U_n = \frac{\sqrt{n}(\overline{Y}_n - 1/2)}{\sqrt{11/144}} \xrightarrow{d} Normal(0,1).$$

In the simulation below, we choose $M = 10$ repeated sequences of maximal length $N = 1000$. The trace plots of $\overline{Y}_n$ for $n = 1, \ldots, N$ for 10 different replicated runs. It is clear that the sample means are converging to $1/2$ (marked as the red line) for each of the repeated runs.

```
set.seed(1010)
M<-10
N<-1000
ybar<-replicate(M,cumsum(runif(N))/c(1:N))
par(mar=c(4,4,1,0))
plot(1:N,ybar[,1],type='l',xlab='n',ylab=expression(bar(y)[n]),ylim=range(0,1))
for(j in 2:M){lines(1:N,ybar[,j])}
abline(h=1/2,col='red',lwd=2)
```



To check the CLT, we increase M to 2000, and reduce N to 500. For each replicate run, and each n, we compute

$$U_n = \frac{\sqrt{n}(\overline{Y}_n - 1/2)}{\sqrt{11/144}}$$

which the CLT predicts will converge to a standard Normal random variable as n gets large.

```
M<-2000
N<-500
ybar<-replicate(M,cumsum(runif(N))/c(1:N))
Un<-(ybar-1/2)/sqrt(11/144)
Un<-apply(Un,2,function(x){return(x*sqrt(1:N))})
```

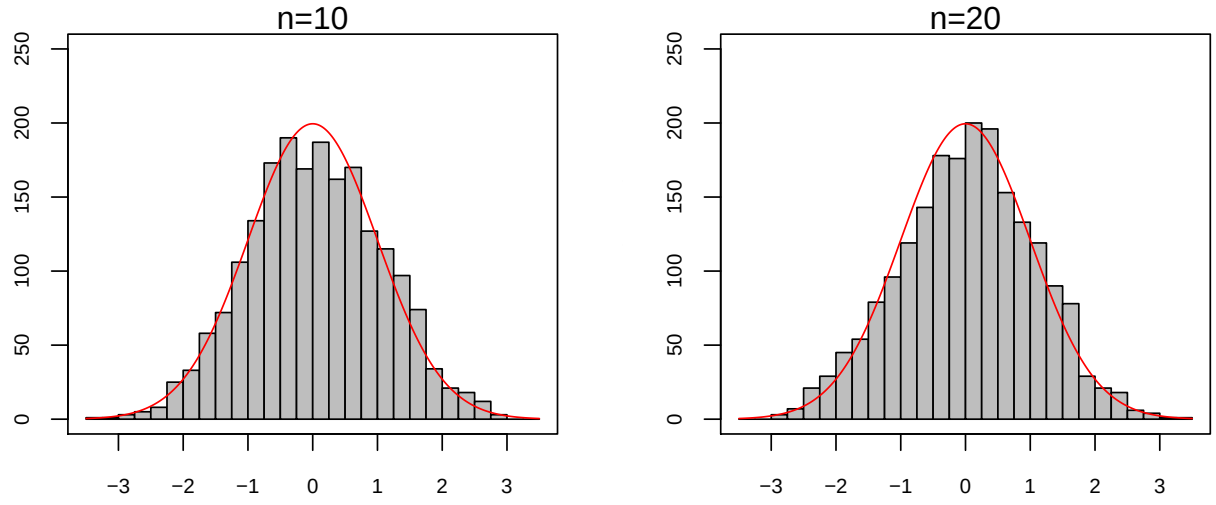
We can check the claim of Normality by computing mean, variance and *quantiles* of the sampled values We compare these summary values with the values of the standard Normal distribution for $n = 100, 200, 300, 400, 500$.

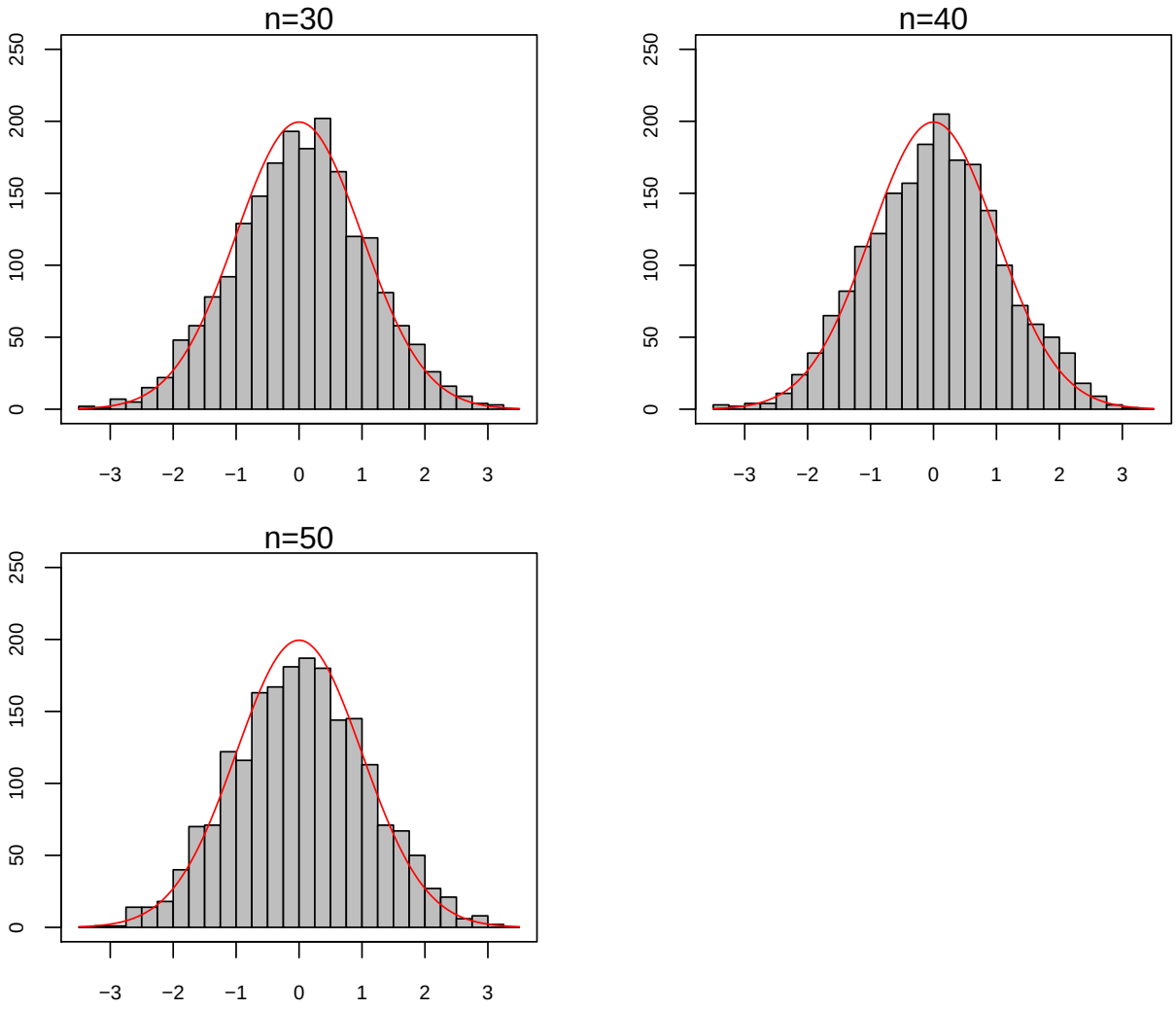
```
Un.sub<-t(Un[100*c(1:5),])
qvec<-c(0.025,0.25,0.5,0.75,0.975)
Un.summary<-apply(Un.sub,2,function(x){return(c(mean(x),var(x),quantile(x,qvec)))})
Un.summary<-cbind(Un.summary,c(0,1,qnorm(qvec)))
colnames(Un.summary)<-c(paste("n=",seq(100,500,by=100),sep=""),"Z")
rownames(Un.summary)[1:2]<-c("mean","var")
round(Un.summary,4)
```

```
+          n=100    n=200    n=300    n=400    n=500       Z
+ mean    0.0152 -0.0242 -0.0435 -0.0454 -0.0433  0.0000
+ var     1.0939  1.0952  1.1300  1.0937  1.0546  1.0000
+ 2.5%   -2.0421 -2.0565 -2.0444 -2.1056 -2.0770 -1.9600
+ 25%    -0.7241 -0.7106 -0.7709 -0.7627 -0.7695 -0.6745
+ 50%    -0.0079 -0.0373 -0.0572 -0.0273 -0.0095  0.0000
+ 75%     0.7311  0.6847  0.6704  0.6830  0.6603  0.6745
+ 97.5%   2.0507  2.1221  2.1162  1.8838  1.8879  1.9600
```

In the uniform case, it is not straightforward to compute the distribution of U_n analytically. We instead compare the histograms of U_n across the M replicates with the standard normal pdf. The computation is performed for $n = 10, 20, 30, 40, 50$.

```
par(mar=c(3,3,1,2),mfrow=c(1,2))
Un.sub<-t(Un[10*c(1:5),])
xv<-seq(-3.5,3.5,by=0.001)
yv<-dnorm(xv)
for(j in 1:5){
    uvec<-Un.sub[,j]
    uvec<-uvec[uvec >= -3.5 & uvec <= 3.5]
    hist(uvec,breaks=seq(-3.5,3.5,by=0.25),col='gray',ylim=range(0,250),cex.axis=0.75,
            main=substitute( paste("n=",nval),list(nval = 10*j)),xlab=expression(u[n]))
    lines(xv,yv*M*0.25,col='red')
    box()
}
```





Here the red line is the plot of the standard normal cdf; the approximation is quite good even for $n = 10$.

EXAMPLE: $Exponential(1)$.
If $Y_i \sim Exponential(1)$, then $\mathbb{E}[Y_i] = 1$ and $\mathbb{V}[Y_i] = 1$. The WLLN suggests that

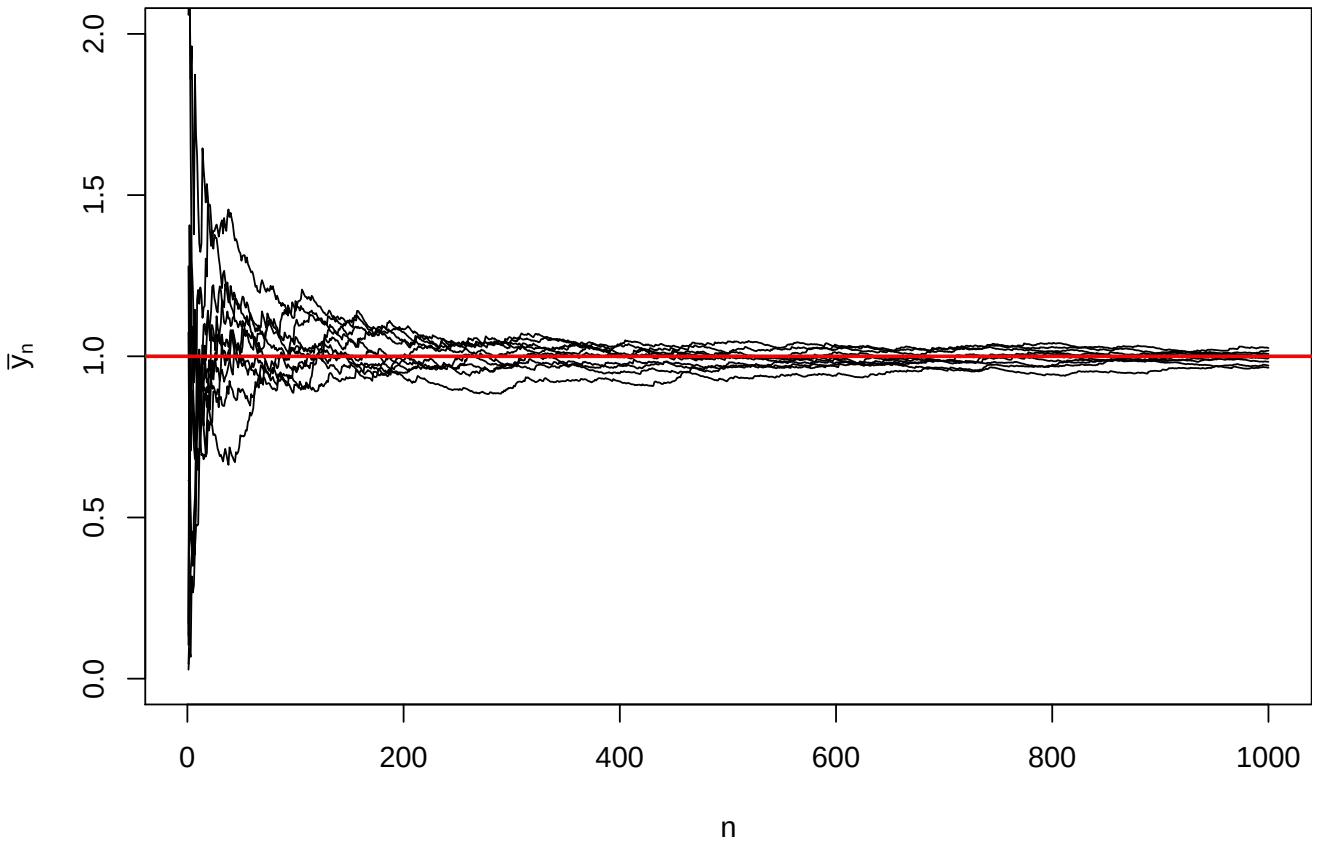
$$\overline{Y}_n \xrightarrow{p} 1$$

whereas the CLT suggests that

$$U_n = \frac{\sqrt{n}(\overline{Y}_n - 1)}{\sqrt{1}} \xrightarrow{d} Normal(0, 1).$$

In the simulation below, we choose $M = 10$ repeated sequences of maximal length $N = 1000$. The trace plots of $\overline{Y}_n$ for $n = 1, \ldots, N$ for 10 different replicated runs. It is clear that the sample means are converging to 1 (marked as the red line) for each of the repeated runs.

```
set.seed(1010)
M<-10
N<-1000
ybar<-replicate(M,cumsum(rexp(N))/c(1:N))
par(mar=c(4,4,1,0))
plot(1:N,ybar[,1],type='l',xlab='n',ylab=expression(bar(y)[n]),ylim=range(0,2))
for(j in 2:M){lines(1:N,ybar[,j])}
abline(h=1,col='red',lwd=2)
```



To check the CLT, we increase M to 2000, and reduce N to 500. For each replicate run, and each n, we compute

$$U_n = \sqrt{n}(\overline{Y}_n - 1)$$

which the CLT predicts will converge to a standard Normal random variable as n gets large.

```
M<-2000;N<-500
ybar<-replicate(M,cumsum(rexp(N))/c(1:N))
Un<-(ybar-1)/sqrt(1)
Un<-apply(Un,2,function(x){return(x*sqrt(1:N))})
```

We can check the claim of Normality by computing mean, variance and *quantiles* of the sampled values We compare these summary values with the values of the standard Normal distribution for $n = 100, 200, 300, 400, 500$.

```
Un.sub<-t(Un[100*c(1:5),])
qvec<-c(0.025,0.25,0.5,0.75,0.975)
Un.summary<-apply(Un.sub,2,function(x){return(c(mean(x),var(x),quantile(x,qvec)))})
Un.summary<-cbind(Un.summary,c(0,1,qnorm(qvec)))
colnames(Un.summary)<-c(paste("n=",seq(100,500,by=100),sep=""),"Z")
rownames(Un.summary)[1:2]<-c("mean","var")
round(Un.summary,4)

+          n=100    n=200    n=300    n=400    n=500       Z
+ mean   -0.0183 -0.0131 -0.0031   0.0071   0.0399  0.0000
+ var     1.0232   1.0160   0.9749   0.9907   0.9947  1.0000
+ 2.5%   -1.9907 -1.9022 -1.8869 -1.8845 -1.9160 -1.9600
+ 25%    -0.7410 -0.6931 -0.6797 -0.6855 -0.6287 -0.6745
+ 50%    -0.0377 -0.0594 -0.0227   0.0077   0.0031  0.0000
+ 75%     0.6480   0.6573   0.6376   0.7153   0.7174  0.6745
+ 97.5%   2.0436   2.0394   1.9368   2.0287   2.0006  1.9600
```

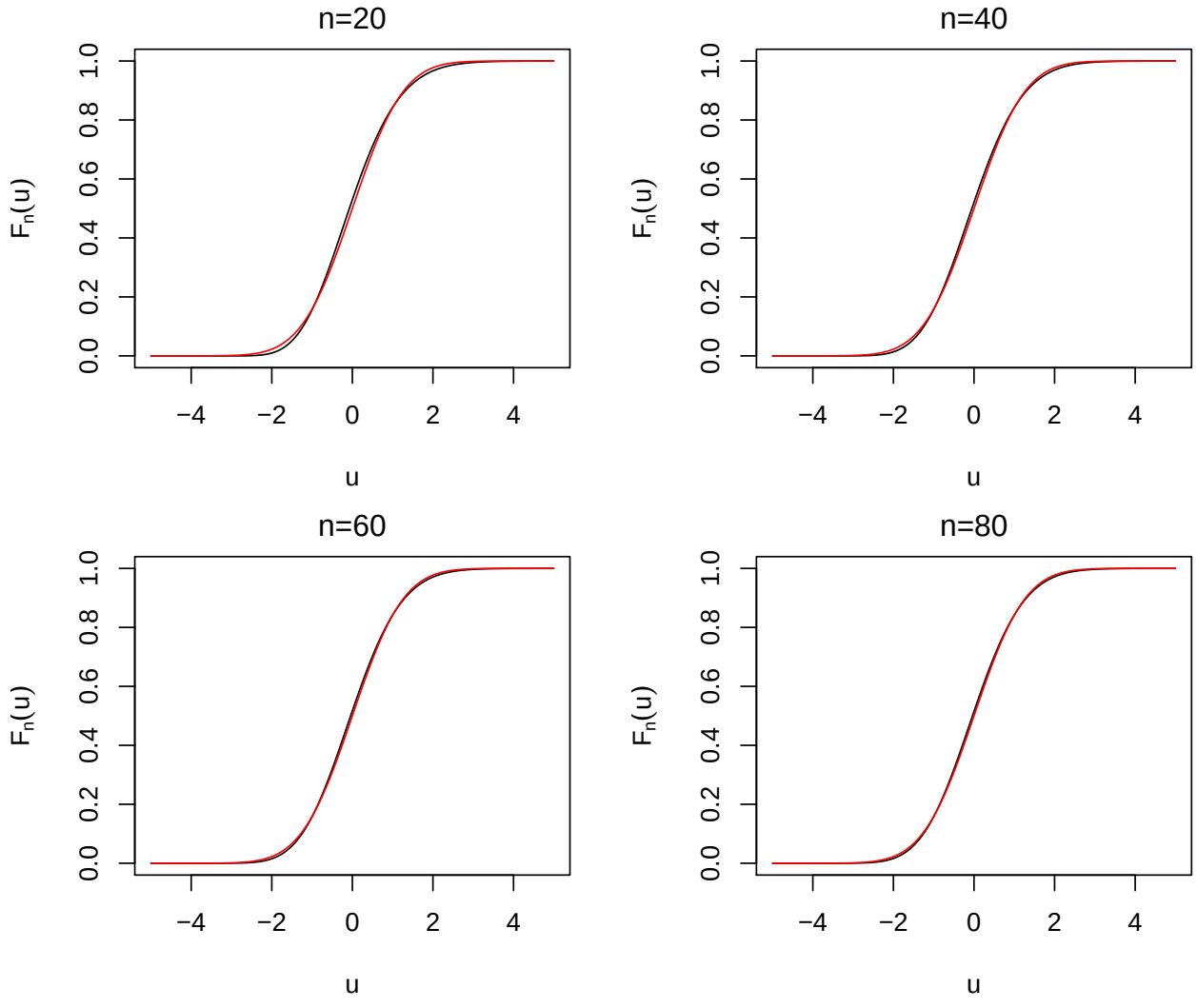
In the Exponential case, we can show using mgfs that

$$S_n = \sum_{i=1}^{n} Y_i \sim Gamma(n, 1)$$

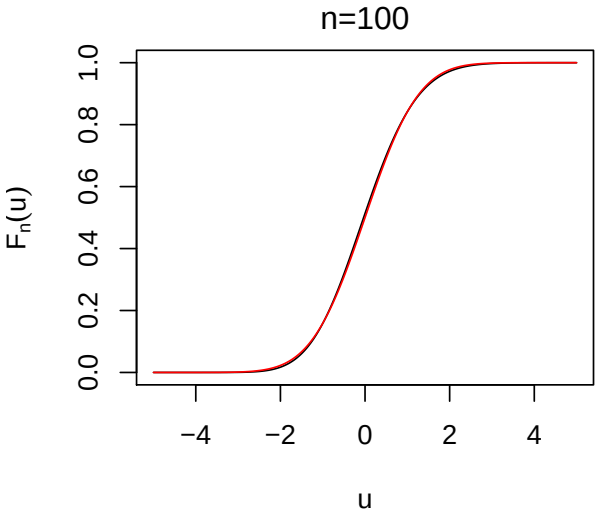
and therefore we can compute the cdf of U_n analytically from first principles:

$$F_{U_n}(u) = P(U_n \leq u) = P\left(\sqrt{n}(\overline{Y}_n - 1) \leq u\right) = P(S_n \leq \sqrt{n}u + n) = F_{S_n}(\sqrt{n}u + n)$$

```
par(mar=c(4,4,2,2),mfrow=c(1,2))
nvec<-seq(20,100,by=20)
xv0<-seq(-2.5,2.5,by=0.5);yv0<-pnorm(xv0)
Fmat<-matrix(0,nrow=length(nvec),ncol=length(xv0))
xv<-seq(-5,5,by=0.01);yv<-pnorm(xv)
for(j in 1:length(nvec)){
    ivec<-0:nvec[j]
    svec<-sqrt(nvec[j])*xv+nvec[j]
    Fvec<-pgamma(svec,nvec[j],1)
    plot(xv,Fvec,pch=19,cex=0.5,xlim=range(-5,5),type='l',
         main=substitute( paste("n=",nval),list(nval = nvec[j])),
         xlab=expression(u),ylab=expression(F[n](u)))
    lines(xv,yv,col='red')
    Fmat[j,]<-pgamma(sqrt(nvec[j])*xv0+nvec[j],nvec[j],1)
}
```



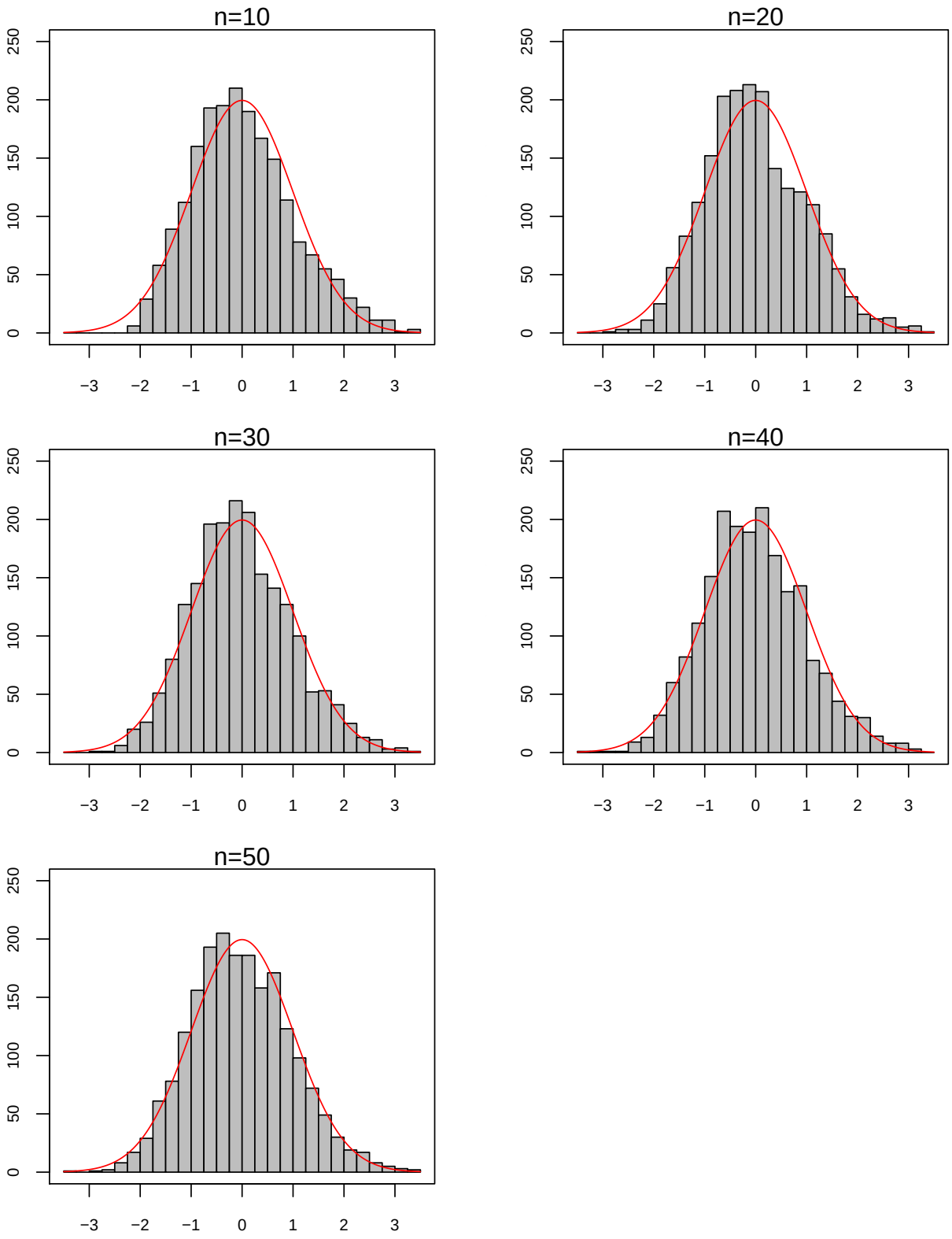
```
Fmat<-rbind(yv0,Fmat);colnames(Fmat)<-xv0;
rownames(Fmat)<-c("Z",paste("n=",seq(20,100,by=20),sep=""))
round(Fmat,3)

+        -2.5    -2  -1.5    -1  -0.5     0   0.5     1   1.5     2   2.5
+ Z     0.006 0.023 0.067 0.159 0.309 0.500 0.691 0.841 0.933 0.977 0.994
+ n=20  0.001 0.010 0.051 0.156 0.328 0.530 0.711 0.843 0.924 0.967 0.987
+ n=40  0.002 0.014 0.057 0.158 0.322 0.521 0.705 0.842 0.926 0.969 0.989
+ n=60  0.003 0.015 0.059 0.158 0.320 0.517 0.703 0.842 0.927 0.971 0.990
+ n=80  0.003 0.016 0.060 0.158 0.318 0.515 0.701 0.842 0.928 0.972 0.990
+ n=100 0.003 0.017 0.061 0.158 0.317 0.513 0.700 0.842 0.928 0.972 0.991
```



In these plots, the black (exact cdf of U_n) and red (standard Normal approximation to the distribution) are almost identical.

We also compare the histograms of U_n across the M replicates with the standard normal pdf. The computation is performed for $n = 10, 20, 30, 40, 50$.

```
par(mar=c(3,3,1,2),mfrow=c(1,2))
Un.sub<-t(Un[10*c(1:5),])
xv<-seq(-3.5,3.5,by=0.001)
yv<-dnorm(xv)
for(j in 1:5){
    uvec<-Un.sub[,j]
    uvec<-uvec[uvec >= -3.5 & uvec <= 3.5]
    hist(uvec,breaks=seq(-3.5,3.5,by=0.25),col='gray',ylim=range(0,250),cex.axis=0.75,
           main=substitute( paste("n=",nval),list(nval = 10*j)),xlab=expression(u[n]))
    lines(xv,yv*M*0.25,col='red');box()
}
```



Here the red line is the plot of the standard normal cdf; the approximation is good for $n = 50$, but not so good for smaller n.