Q-Q plot: 統計資料分析上常使用 Q-Q plot 圖來檢驗資料是否來自常態分佈。基本上它是一種圖形的檢驗法(graphical test)，用來描述樣本資料與對應常態母體分佈的散佈圖。為了使散佈圖不受測量單位之影響，通常都會先將資料標準化，然後再和標準常態分佈一起繪製散佈圖。Q-Q plot 圖中會加入一條常態直線圖，當 Q-Q plot 散佈圖在此直線附近時，代表樣本資料來自常態分佈。

步驟 假設 $X_1, X_2, \cdots, X_n$ 是一組從(母體)平均數 $-\infty < \mu < \infty$ 和(母體)變異數 $\sigma^2 > 0$ 的隨機變數 X 取出的隨機樣本。

(1) 計算樣本平均數 $\hat{\mu} = \overline{X} = \sum_{i=1}^{n} X_i / n$ 及樣本變異數 $\hat{\sigma}^2 = S^2 = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 / (n-1)$。

(2) 將隨機樣本資料 $X_1, X_2, \cdots, X_n$ 標準化且排序得 $d_1 \le d_2 \le \cdots \le d_n$，其中 $d_i = \left(X_i - \overline{X}\right) / S$。

(3) 依據標準常態分佈計算下列 n 個機率的分位值(quantile): $q_1 = z_{1/2n}$，$q_2 = z_{3/2n}$，…，$q_n = z_{(2n-1)/2n}$，其中 $q_i = z_{(i-1/2)/n}$ 代表標準常態分佈機率為 $(i - 1/2)/n$ 時之分位值。

(4) 畫出 (d_i, q_i) 的散佈圖。

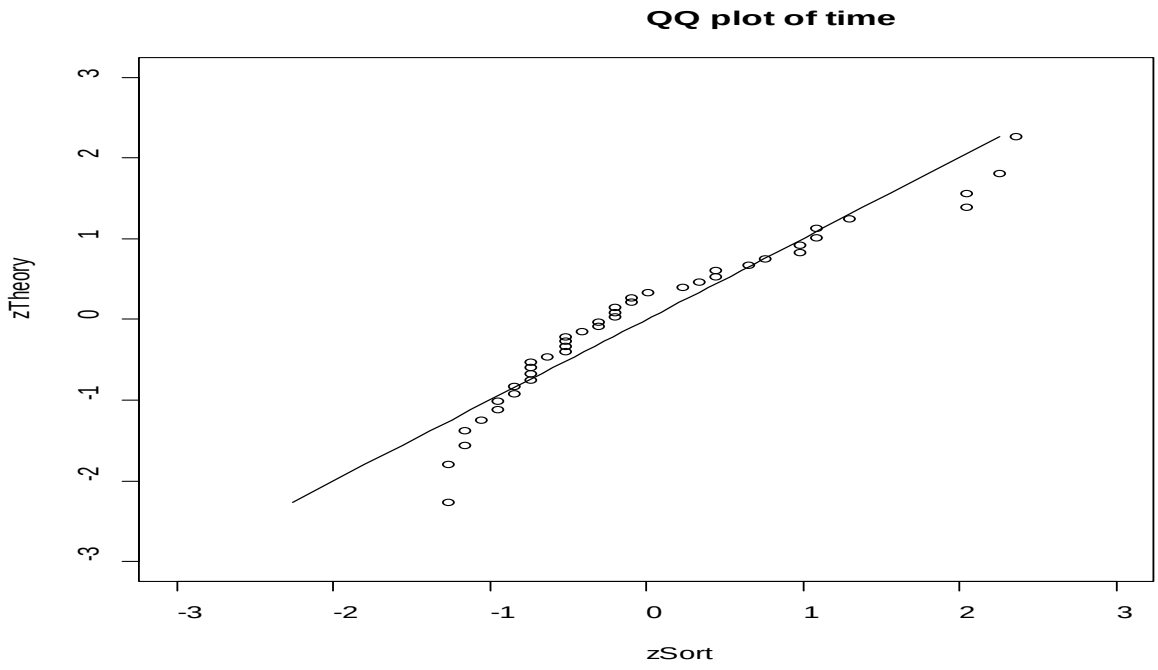
(5) 加入一條由 (q_i, q_i) 所產生的直線。

注意: R 中亦有預設函數 qqnorm 與 qqline，其中橫座標代表排序的標準常態分位值，而縱座標則代表排序的未經標準化樣本值。

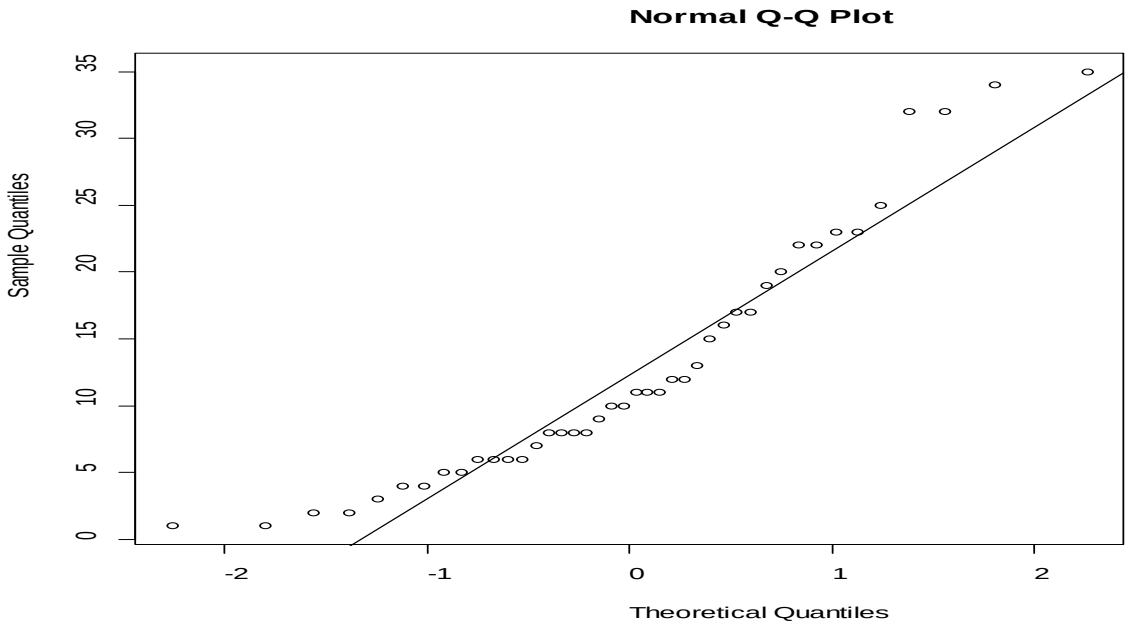
例 繪製資料庫 MASS 中 gehan 的 time Q-Q plot 圖。

```
> library(MASS)
> gehan$time
 [1]  1 10 22  7  3 32 12 23  8 22 17  6  2 16 11 34  8 32 12 25  2 11  5 20
[25]  4 19 15  6  8 17 23 35  5  6 11 13  4  9  1  6  8 10

> QQplot=function(x)
+ {
+ x.mean=mean(x)
+ x.var=var(x)
+ x.n=length(x)
+ zx=(x-x.mean)/sqrt(x.var)
+ zSort=sort(zx)
+ i=1:x.n
+ p=(i-1/2)/x.n
+ zTheory=qnorm(p)
+ plot(zSort,zTheory,xlim=c(-3,3),ylim=c(-3,3))
+ title("QQ plot of time")
+ lines(q,q)
+ }
> QQplot(gehan$time)
```



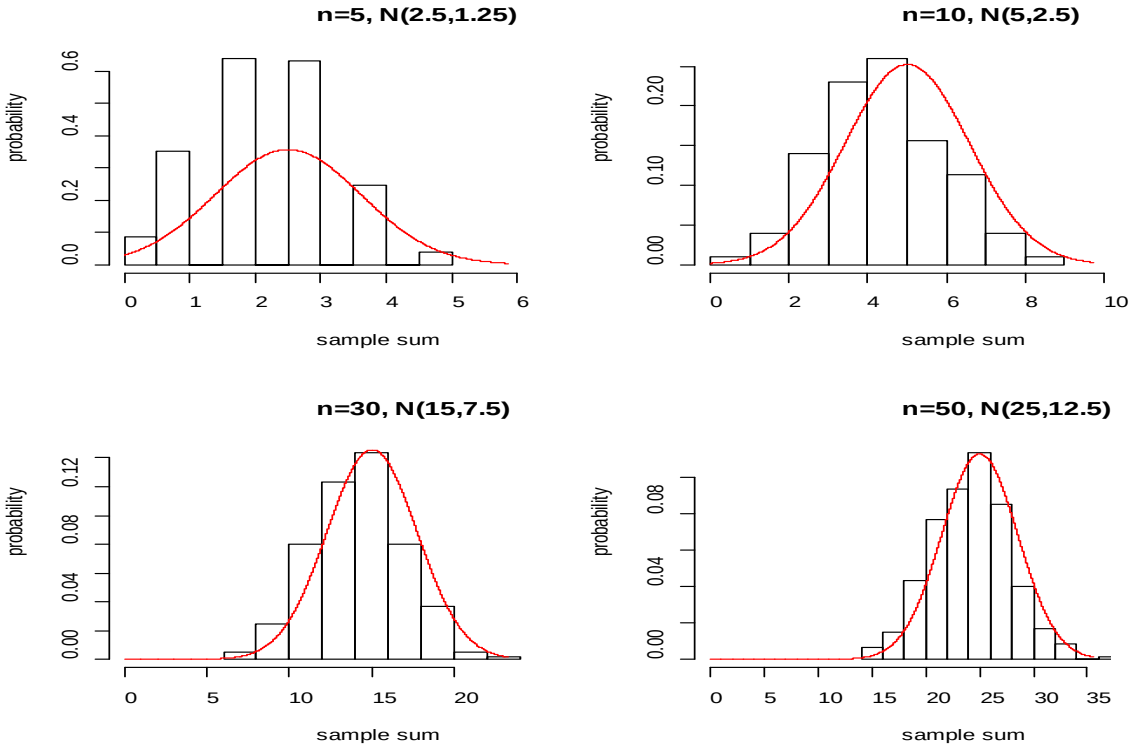

```
> qqnorm(gehan$time)
> qqline(gehan$time)  #由0.25和0.75的標準常態分位值與樣本分位值這兩點
>                     #所形成的直線
```



驗證中央極限定理:

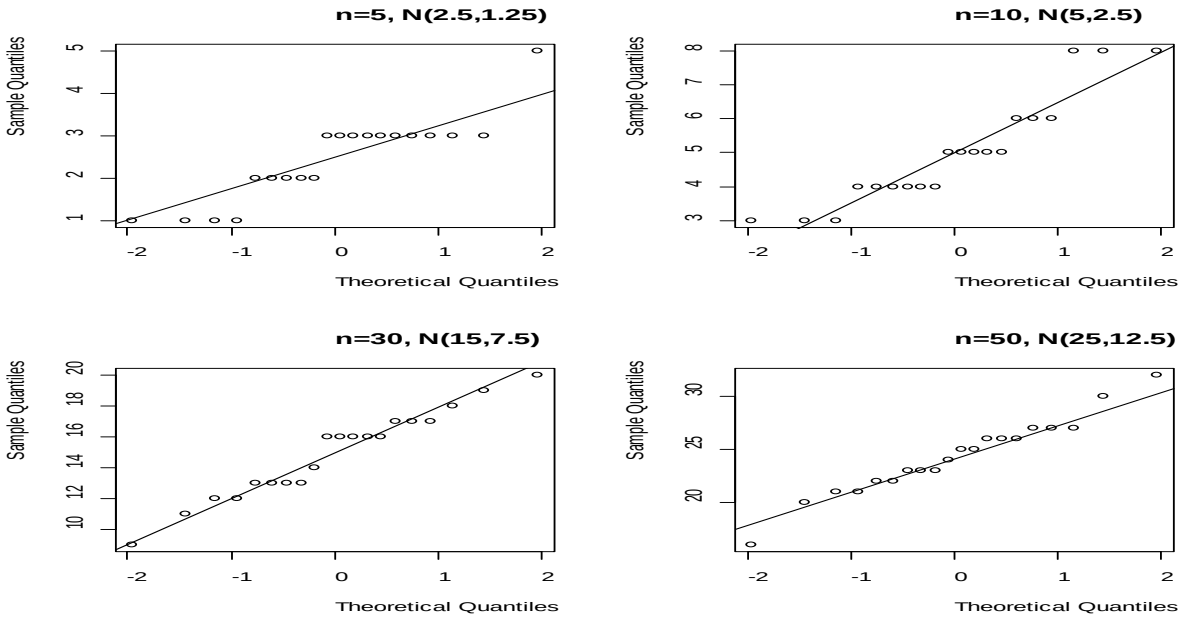
例 1　$X_1, X_2, \cdots, X_n$ 是一組從(母體)平均數 $\mu = p = 0.5$ 和(母體)變異數 $\sigma^2 = p(1-p)$的白奴利分佈 X 取出的隨機樣本，觀察 $\sum_{i=1}^{n} X_i$ 的樣本分佈。

```
> Nsample=300
> Nsize=50
> BernRandomSample=matrix(0,Nsample,Nsize)
> for(i in 1:Nsize)
+ {
+ BernRnumber=rbinom(Nsample,1,1/2)
+ BernRandomSample[,i]=as.matrix(BernRnumber)#將隨機樣本指定為矩陣的第i行
+ }
> rSampleSum=matrix(0,Nsample,4)
> rSampleSum[,1]=apply(BernRandomSample[,1:5],1,sum)
> rSampleSum[,2]=apply(BernRandomSample[,1:10],1,sum)
> rSampleSum[,3]=apply(BernRandomSample[,1:30],1,sum)
> rSampleSum[,4]=apply(BernRandomSample[,1:50],1,sum)
> colnames(rSampleSum)=c("n=5, N(2.5,1.25)","n=10, N(5,2.5)",
+ "n=30, N(15,7.5)","n=50, N(25,12.5)")
> windows()
> par(mfrow=c(2,2))
> mu=c(5*0.5,10*0.5,30*0.5,50*0.5)
> var=c(5*0.5*(1-0.5),10*0.5*(1-0.5),30*0.5*(1-0.5),50*0.5*(1-0.5))
> for(i in 1:4)
+ {
+ int=seq(0,mu[i]+3*sqrt(var[i]),0.001)
+ pdf=dnorm(int,mu[i],sqrt(var[i]))
+ hist(rSampleSum[,i],ylab="probability",xlab="sample sum",pro=T,
+ xlim=c(0,mu[i]+3*sqrt(var[i])),main=colnames(rSampleSum)[i])
+ lines(int,pdf,col="red")
+ }
```



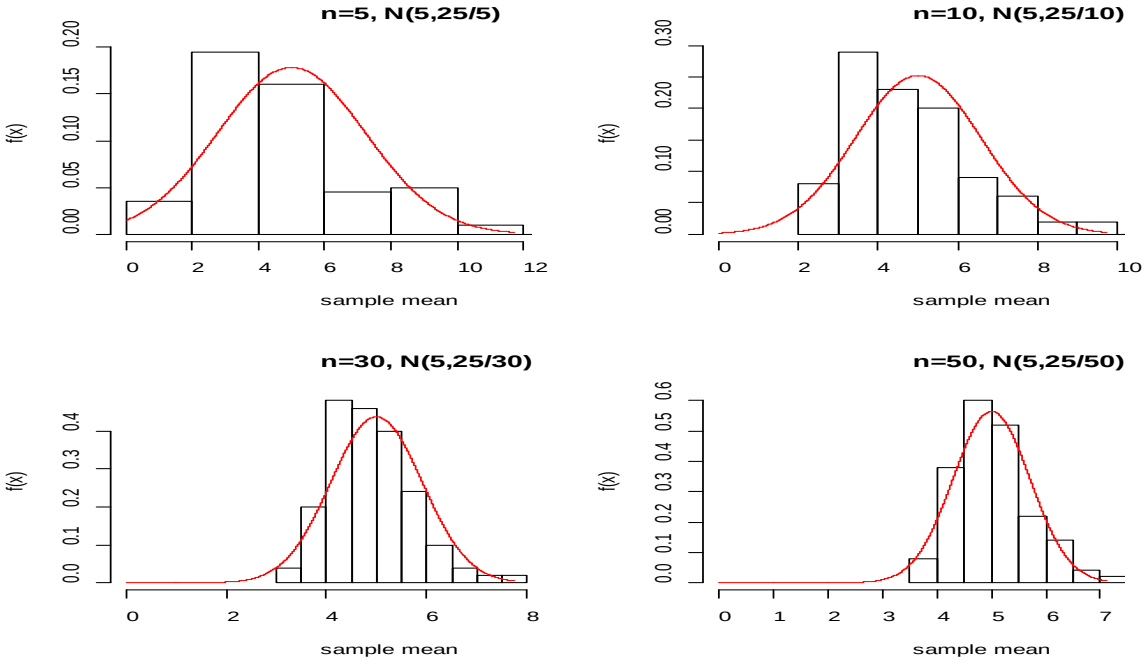
例 1(續) 對於不同樣本數 *n*=5, 10, 30, 50 的 20 個觀測值畫 Q-Q plot。

```
> windows()
> par(mfrow=c(2,2))
> for(i in 1:4)
+ {
+ qqnorm(rSampleSum[1:20,i],main=colnames(rSampleSum)[i])
+ qqline(rSampleSum[1:20,i])
+ }
```

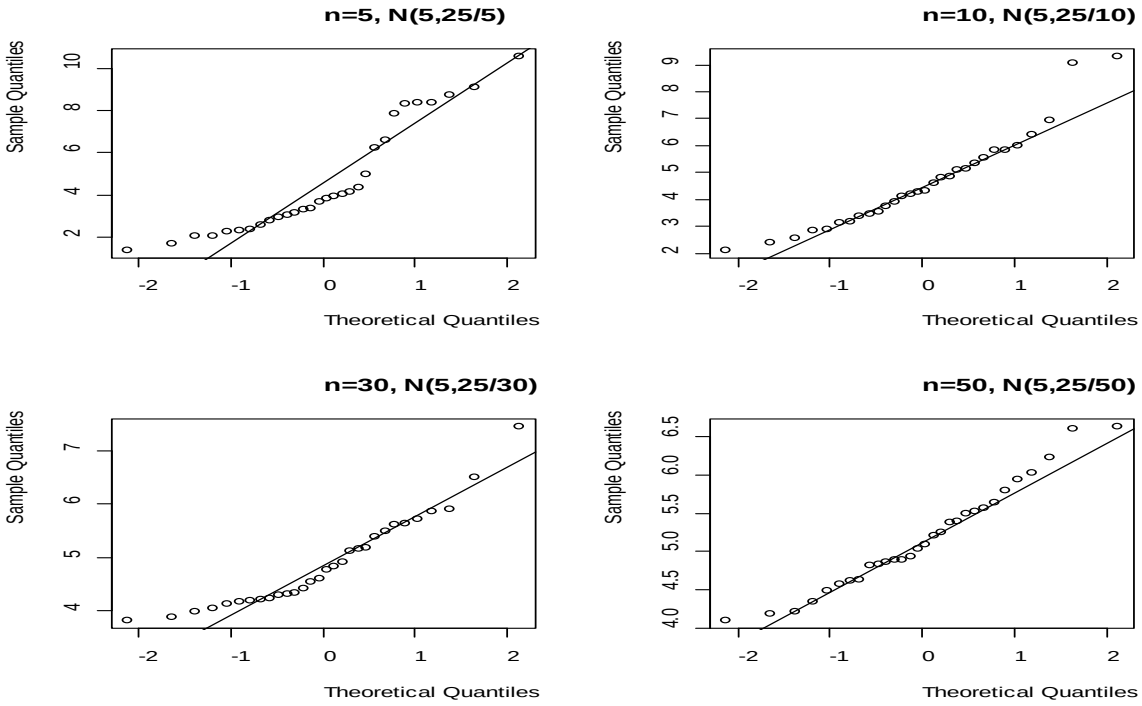


例 2　$X_1, X_2, \cdots, X_n$ 是一組從平均數 μ =1/λ=5 和變異數 σ^2=1/λ^2 = 25的指數分佈 X 取出的隨機樣本，觀察 $\overline{X} = \sum_{i=1}^{n} X_i / n$ 的樣本分佈。

```
> Nsample=100
> Nsize=50
> expRandomSample=matrix(0,Nsample,Nsize)
> for(i in 1:Nsize)
+ {
+ expRnumber=rexp(Nsample,0.2)
+ expRandomSample[,i]=as.matrix(expRnumber)#將隨機樣本指定為矩陣的第i行
+ }
> rSampleMean=matrix(0,Nsample,4)
> rSampleMean[,1]=apply(expRandomSample[,1:5],1,mean)
> rSampleMean[,2]=apply(expRandomSample[,1:10],1,mean)
> rSampleMean[,3]=apply(expRandomSample[,1:30],1,mean)
> rSampleMean[,4]=apply(expRandomSample[,1:50],1,mean)
> colnames(rSampleMean)=c("n=5, N(5,25/5)","n=10, N(5,25/10)",
+ "n=30, N(5,25/30)","n=50, N(5,25/50)")
> windows()
> par(mfrow=c(2,2))
> mean.unif=1/0.2
> var.unif=1/0.2^2
> mu=rep(mean.unif,4)
> var=c(var.unif/5,var.unif/10,var.unif/30,var.unif/50)
> for(i in 1:4)
+ {
+ int=seq(0,mu[i]+3*sqrt(var[i]),0.001)
+ pdf=dnorm(int,mu[i],sqrt(var[i]))
+ hist(rSampleMean[,i],ylab="f(x)",xlab="sample mean",pro=T,
+ xlim=c(0,mu[i]+3*sqrt(var[i])),main=colnames(rSampleMean)[i])
+ lines(int,pdf,col="red")
+ }
```



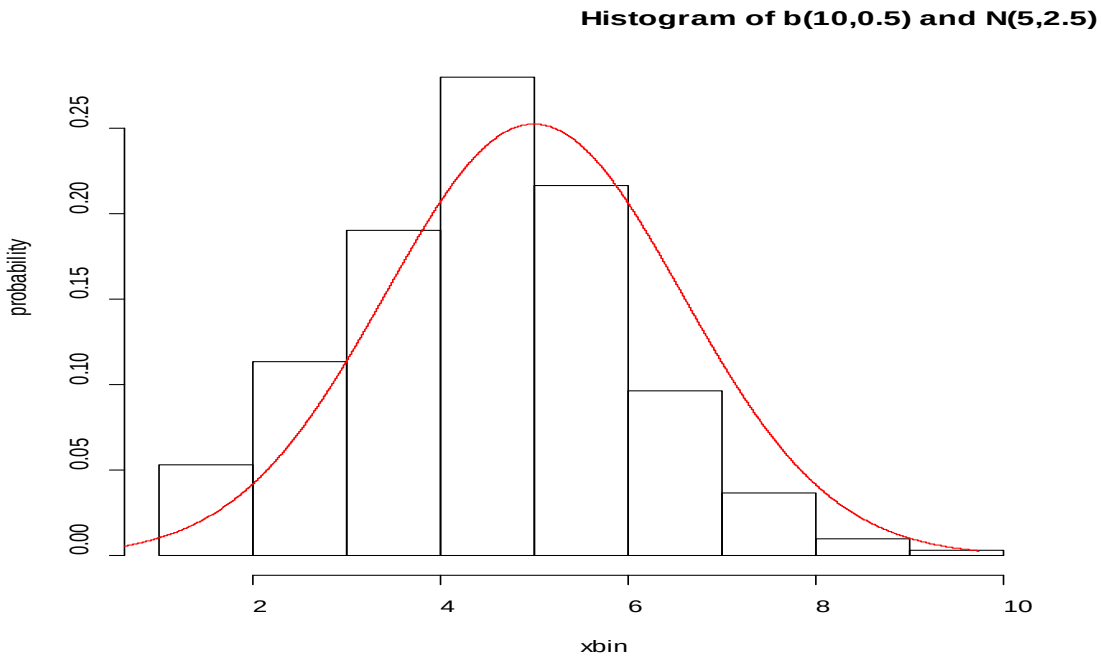
```
> windows()
> par(mfrow=c(2,2))
> for(i in 1:4)
+ {
+ qqnorm(rSampleMean[1:30,i],main=colnames(rSampleMean)[i])
+ qqline(rSampleMean[1:30,i])
+ }
>
```



連續的修正(continuity correction，以連續型常態分佈來近似離散型分佈的機率時使用):

例 3 隨機變數 $Y \sim b(10, 0.5)$，則 $P(Y \leq 5)$=?

```
> windows()
> xbin=rbinom(300,10,0.5)
> mean=10*0.5
> var=10*0.5*(1-0.5)
> int=seq(0,mean+3*sqrt(var),0.001)
> pdf=dnorm(int,mean,sqrt(var))
> hist(xbin,ylab="probability", main="Histogram of b(10,0.5) and N(5,2.5)",
+ pro=T)
> lines(int,pdf,col="red")
```




```
> pbinom(5,10,0.5)
[1] 0.6230469
> z=(5-10*0.5)/sqrt((10*0.5*(1-0.5)))
> pnorm(z)
[1] 0.5
> z1=(5+0.5-10*0.5)/sqrt((10*0.5*(1-0.5)))
> pnorm(z1)
[1] 0.6240852
```

$P(Y > 5)$=? $P(3 < Y \leq 7)$=? $P(Y \leq 3)$=?

例 4 隨機變數 $Y \sim \chi^2_{25}$，則 $P(15 < Y \leq 35)$=?

```
> pchisq(35,25)-pchisq(15,25)
[1] 0.8531794
> a=(15-25)/sqrt(2*25)
> b=(35-25)/sqrt(2*25)
> pnorm(b)-pnorm(a)
[1] 0.8427008
```

$P(Y > 5)$=?

```
> 1-pchisq(5,25)
[1] 0.9999945
> z=(5-25)/sqrt(2*25)
> 1-pnorm(z)
[1] 0.9976611
```