中研院統計所2015統計研習營
2015/07/01~03

(7) 機率分佈
大數法則、中央極限定理
主成份分析、模擬

吳漢銘
淡江大學 數學系
資料科學與數理統計組

http://www.hmwu.idv.tw

本章大綱 & 學習目標

- 機率分佈 (Probability distribution)
  - 統計分配之描述、常見之分配、隨機抽樣、實作QQplot
- 大數法則 (LLN)與 中央極限定理 (CLT)
- 維度縮減: 主成份分析法 (PCA)
- 模擬

統計分配 (Statistical Distributions)

Four fundamental items can be calculated for a statistical distribution:

- 機率密度函數
  - point probability P(X=x) or *probability density function* f(x): **dnorm()**
- 累積機率函數: F(x) = P(X<=x)
  - *cumulative probability distribution function*: **pnorm()**
- 分位數
  - the quantiles of the distribution: **qnorm()**
- 隨機數
  - the random numbers generated from the distribution: **rnorm()**

常用機率分配

以常態分佈normal為例:

機率密度(分配)函數: dnorm()

累積機率(分配)函數: pnorm()

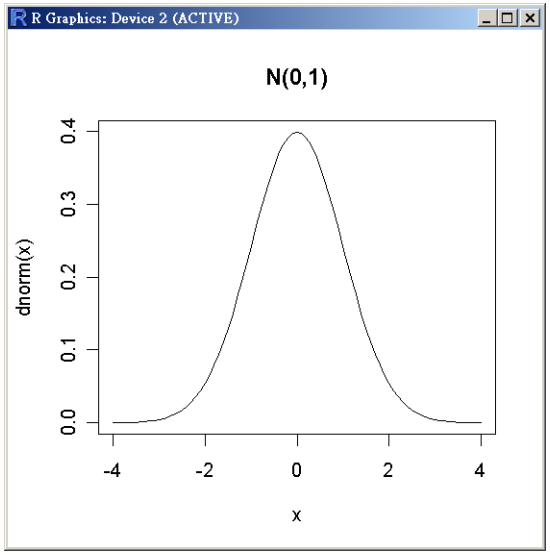
分位數: qnorm()

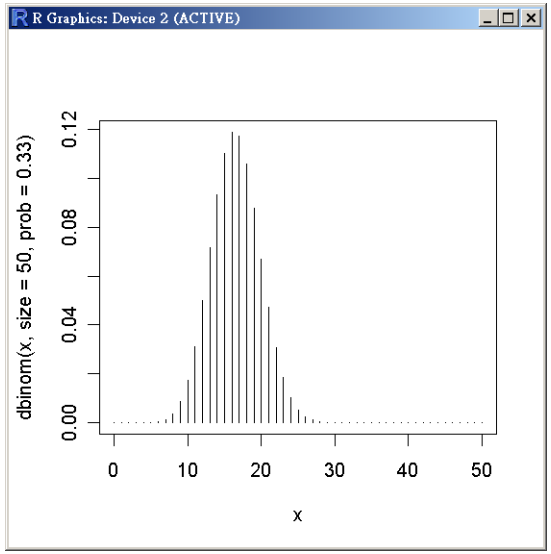
隨機數: rnorm()

Distribution	R name	additional arguments
beta	beta	shape1, shape2, ncp
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-squared	chisq	df, ncp
exponential	exp	rate
F	f	df1, df1, ncp
gamma	gamma	shape, scale
geometric	geom	prob
hypergeometric	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logistic	logis	location, scale
negative binomial	nbinom	size, prob
normal	norm	mean, sd
Poisson	pois	lambda
Student's	t	t df, ncp
uniform	unif	min, max
Weibull	weibull	shape, scale
Wilcoxon	wilcox	m, n

機率密度函數 Density (d)

- The density for a continuous distribution is a measure of the relative probability of "getting a value close to x".
- The probability of getting a value in a particular interval is the area under the corresponding part of the curve.
- For discrete distributions, the term "density" is used for the point probability, the probability of getting exactly the value x.






```
> x <- seq(-4, 4, 0.1)
> plot(x, dnorm(x), type="l", main="N(0,1)")
> curve(dnorm(x), from=-4, to=4)
```

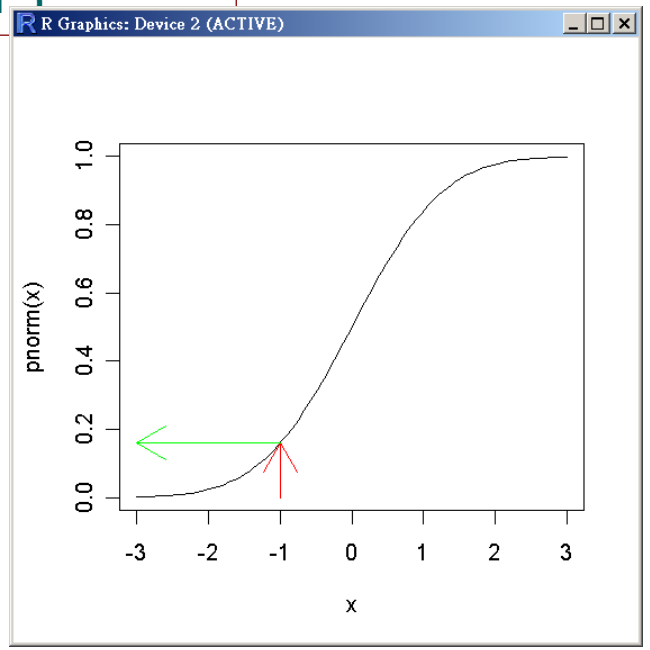
```
x <- 0:50
plot(x, dbinom(x, size=50, prob=0.33), type="h")
#histogram-like
```

累積機率分配函數 CDF (p)

- It is an S-shaped curve showing for any value of x, the probability of obtaining a sample value that is less than or equal to x, P(X<= x).
- The probability density is the slope of this curve (its derivative) of the cumulative probability function.
- It is useful in statistical tests.

```
> curve(pnorm(x), -3, 3)
> arrows(-1, 0, -1, pnorm(-1), col="red")
> arrows(-1, pnorm(-1), -3, pnorm(-1), col="green")
```

The value of (x=-1) leads up to the cumulative probability (red) and the probability associated with obtaining a value of this size (-1) or smaller is on the y-axis (green).

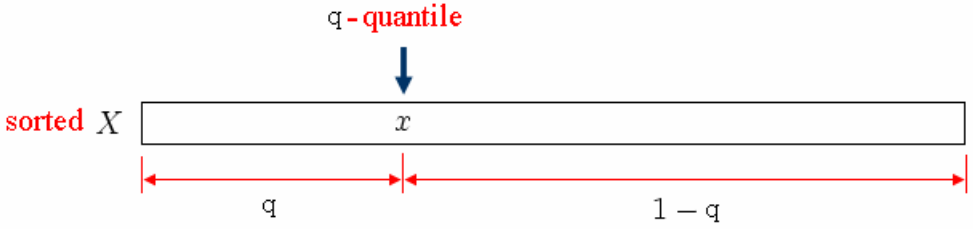



```
> pnorm(-1)
[1] 0.1586553
```

分位數 Quantiles (q)

- The quantile function is the inverse of the cumulative distribution function. F^{-1}(p)=x.
- The q-quantile is the value with the property that there is probability p of getting a value less than or equal to it.

The qth quantile of a data set is defined as that value where a q fraction of the data is below that value and (1-q) fraction of the data is above that value. For example, the 0.5 quantile is the median.

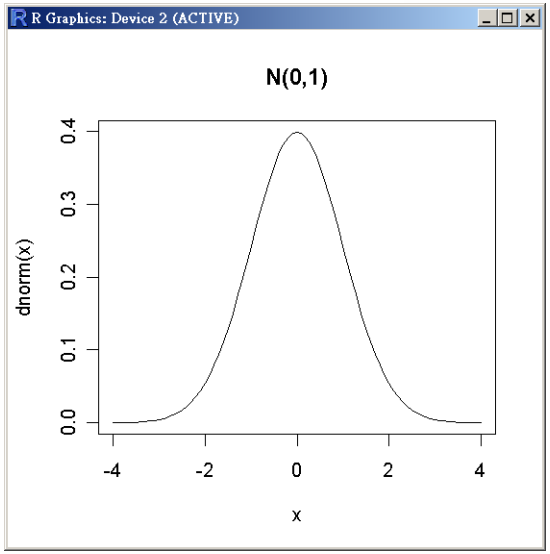


$P(X < x) \leq q$ and $P(X > x) \leq 1 - q$.

2.5% quantile in the normal distribution

```
> qnorm(0.025)
[1] -1.959964

> qnorm(0.975)
[1] 1.959964
```

$\Phi^{-1}(0.975)$



$$\bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.975}\frac{\sigma}{\sqrt{n}}$$

$$P(z_{0.025} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{0.975}) = 0.95$$

隨機數 Random Numbers (r)

- **Definition of a Random Sample (隨機樣本)**
  A random sample of size *n* is an *n*-tuple of identically-distributed independent random variables.

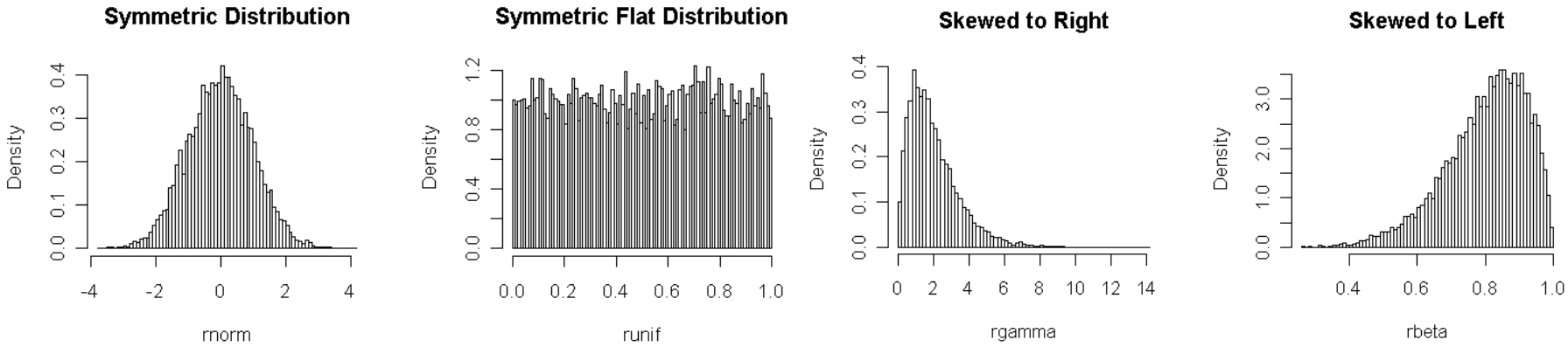
- 隨機樣本(Random Sample): 通常指來自於某個母體分配，樣本大小為*n*，用X1, X2,..., Xn來表示。

- 彼此間相互獨立: 離散型隨機樣本, 連續型隨機樣本。

- In the standard sampling model, Xi is a vector of measurements for the i-th object in the sample, and thus, we think of X1, ..., Xn as independent copies of an underlying measurement vector. In this case, (X1, X2, ..., Xn) is said to be a random sample of size n from the common distribution.

Simulation Different Shapes of Distribution

```
> par(mfrow=c(2,2))
> hist.sym <-hist(rnorm(10000),nclas=100,freq=FALSE,
   main="Symmetric Distribution", xlab="rnorm")

> hist.flat <-hist(runif(10000),nclas=100,freq=FALSE,
   main="Symmetric Flat Distribution", xlab="runif")

> hist.skr <-hist(rgamma(10000,shape=2,scale=1),freq=FALSE,
   nclas=100, main="Skewed to Right", xlab="rgamma")

> hist.skl <-hist(rbeta(10000,8,2),nclas=100,freq=FALSE,
   main="Skewed to Left", xlab="rbeta")
```



隨機抽樣 (Random Sampling)

- The concepts of randomness and probability are central to statistics.

```
> sample(x, size, replace = FALSE, prob = NULL)
```

- sampling without replacement

```
> sample(1:40, 5)
```

- sampling with replacement

```
> sample(1:40, 5, replace=TRUE)
```

- Simulate 10 coin tosses (fair coin-tossing)

```
> sample(c("H", "T"), 10, replace=T)
 [1] "T" "T" "T" "H" "H" "H" "T" "H" "T" "H"

> sample(c("succ", "fail"), 10, replace=T, prob=c(0.9, 0.1))
 [1] "succ" "succ" "succ" "fail" "fail" "fail" "succ" "succ"
   "succ" "succ"
```

隨機抽樣 (Random Sampling)

- permutation

```
> x <- 1:5
> sample(x)
[1] 3 1 5 4 2
```

- Clinical trials: randomization: random assign to two groups, total 20 subjects
- random assigning treatment groups

```
> sample(2, size=20, replace=TRUE)
 [1] 2 2 2 1 1 2 2 2 1 2 1 1 2 1 2 1 2 1 1 1
```

- random choose 10 subjects to group 1

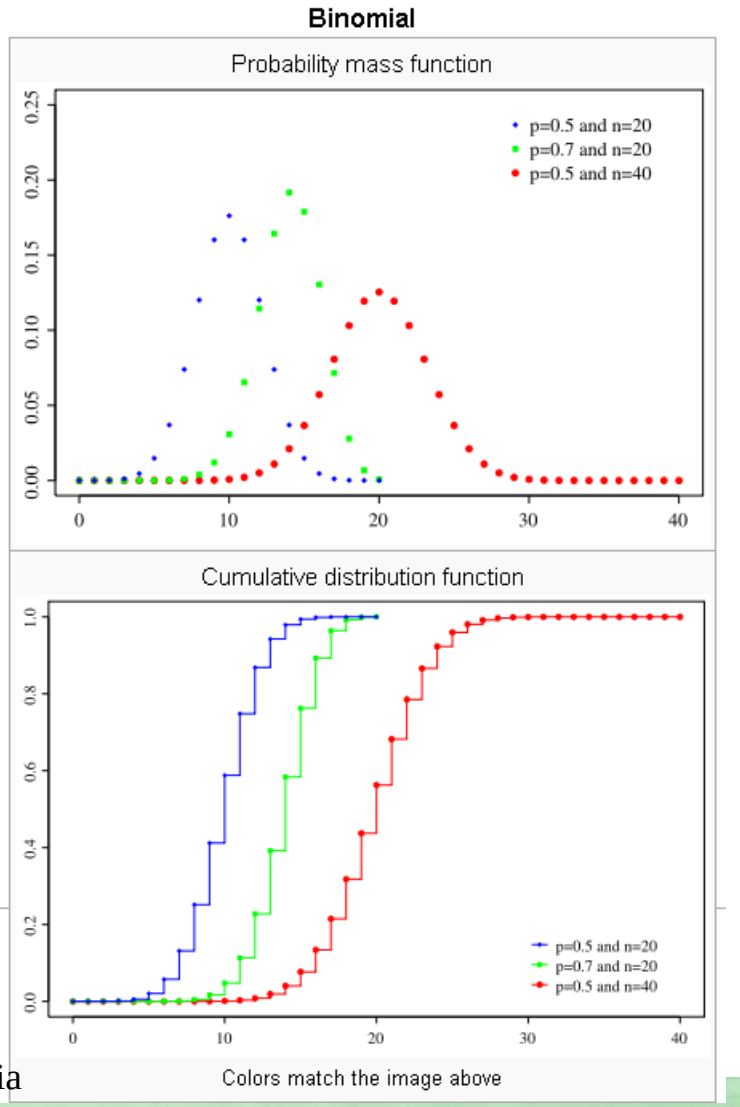
```
> sample(20, size=10, replace=FALSE)
 [1] 10 13 16  8  4 14  7 11  1  5
```

```
> x <- 1:10
> x[x > 8]
[1]  9 10
> sample(x[x > 8]) # length 2
[1] 10  9
> x[x > 9]
[1] 10
> sample(x[x > 9]) # length 10
 [1]  7  9  6  4  8  5  3  2  1 10
```

二項式分佈 (Binomial)

- X~B(n, p)表示n次試驗中(size)，正面結果出現的次數。
- dbinom(x, size, prob) #機率公式值 P(X=x)
- pbinom(q, size, prob) #累加至q的機率值 P(X\leq q)
- qbinom(p, size, prob) #已知累加機率值，對應的機率點
- rbinom(n, size, prob) #隨機樣本數=n的二項隨機變數值

Parameters	$n \geq 0$ number of trials (integer) $0 \leq p \leq 1$ success probability (real)
Support	$k \in \{0, \dots, n\}$
Probability mass function (pmf)	$\binom{n}{k} p^k (1-p)^{n-k}$
Cumulative distribution function (cdf)	$I_{1-p}(n - \lfloor k \rfloor, 1 + \lfloor k \rfloor)$
Mean	np
Median	one of $\{\lfloor np \rfloor - 1, \lfloor np \rfloor, \lfloor np \rfloor + 1\}$
Mode	$\lfloor (n+1)p \rfloor$
Variance	$np(1-p)$
Skewness	$\frac{1-2p}{\sqrt{np(1-p)}}$
Excess kurtosis	$\frac{1-6p(1-p)}{np(1-p)}$
Entropy	$\frac{1}{2}\ln(2\pi nep(1-p)) + O\left(\frac{1}{n}\right)$
Moment-generating function (mgf)	$(1-p+pe^t)^n$
Characteristic function	$(1-p+pe^{it})^n$



Wikipedia

二項式分佈

X~B(10, 0.8)

- 利用二項分配理論公式，計算機率公式值 P(X=3)。

```
> factorial(10)/(factorial(3)*factorial(7))*0.8^3*0.2^7
```

- 利用R函數，計算機率值 P(X=3)。

```
> dbinom(3, 10, 0.8)
```

- 利用R函數，計算累加機率值 P(X<= 2) , P(X<= 3)。

```
> pbinom(3, 10, 0.8)
> pbinom(2, 10, 0.8)
```

- 計算P(X<= 3)- P(X<= 2) ，並和P(X=3)相比較。

```
> pbinom(3, 10, 0.8)- pbinom(2, 10, 0.8)
```

- 已知累加機率值為0.1208，求對應的機率點。

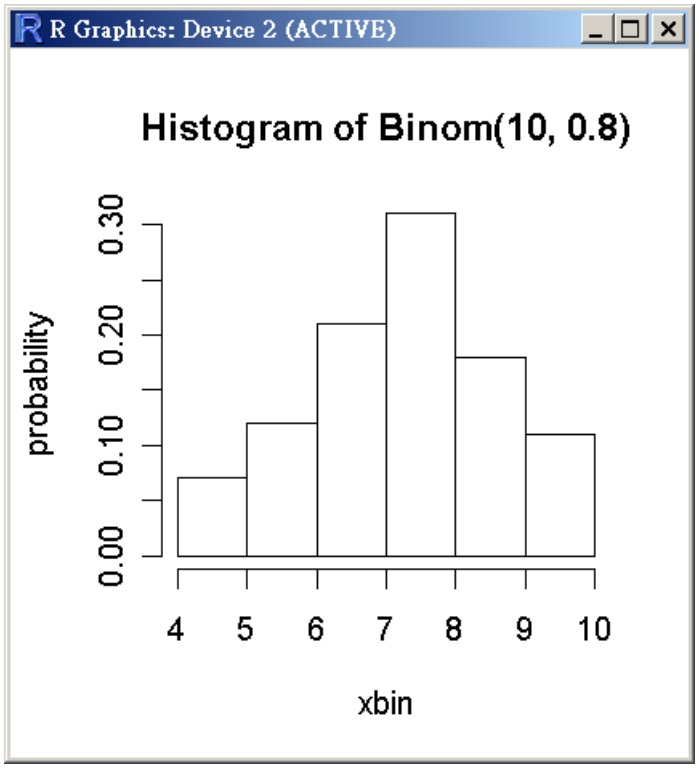
```
> qbinom(0.1208, 10, 0.8)
> pbinom(6, 10, 0.8)
```

二項式分佈

X~B(10, 0.8)

- 產生隨機樣本數100的二項隨機數值，計算其平均數及變異數，並與理論值比較。
- 畫直方圖，x-axis="機率值"，label="probability", title="Histogram of Binom(10, 0.8)"。

```
> n <- 10
> p <- 0.8
> m <- 100
> xbin <- rbinom(m, n, p)
> table(xbin)
xbin
 4  5  6  7  8  9 10
 1  6 12 21 31 18 11
> mu <- n*p; mu
[1] 8
> sigma2 <- n*p*(1-p); sigma2
[1] 1.6
> mean(xbin)
[1] 7.73
> var(xbin)
[1] 1.956667


hist(xbin, ylab="probability", main="Histogram
of Binom(10, 0.8)", prob=T)
```



標準常態分佈 (Standard Normal Distribution)

標準常態分佈: this distribution is central to the theory of parameter statistics.

- Z~N(mu, sigma^2)
- dnorm(x, m, sd) #機率密度函數值 f(Z=z)
- pnorm(q, m, sd) #累加至q的機率值P(Z<= q)
- qnorm(p, m, sd) #已知累加機率值p, 對應的機率點
- rnorm(n, m, sd) #隨機樣本數n的標準常態隨機變數

Z ~N(0, 1)

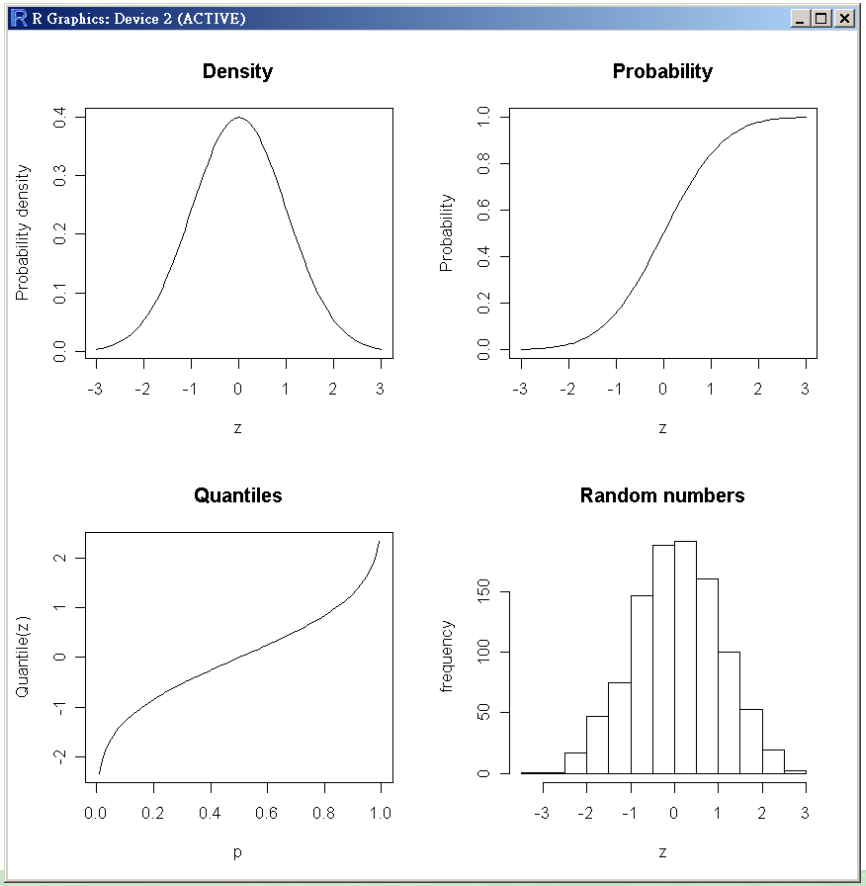
```
> dnorm(0)
[1] 0.4
> pnorm(-1)
[1] 0.16
> qnorm(0.975)
[1] 2.0
```

X~N(10, 4)

```
> dnorm(10, mean=10, sd=2) #f(X=10)
[1] 0.2
> pnorm(1.96, mean=10, sd=2)
[1] 2.9e-05
> qnorm(0.975, mean=10, sd=2)
[1] 14
> rnorm(5, mean=10, sd=2)
[1]  9.4 10.2  9.9  5.0 10.9
> pnorm(15, 10, 2) - pnorm(8, 10, 2)
#P(8<=X<=15)
[1] 0.84
```

常態分佈 (Normal Distribution)

```
> par(mfrow=c(2,2))
> curve(dnorm, -3, 3, xlab="z", ylab="Probability density", main="Density")
> curve(pnorm, -3, 3, xlab="z", ylab="Probability", main="Probability")
> curve(qnorm, 0, 1, xlab="p", ylab="Quantile(z)", main="Quantiles")
> hist(rnorm(1000), xlab="z", ylab="frequency", main="Random numbers")
```

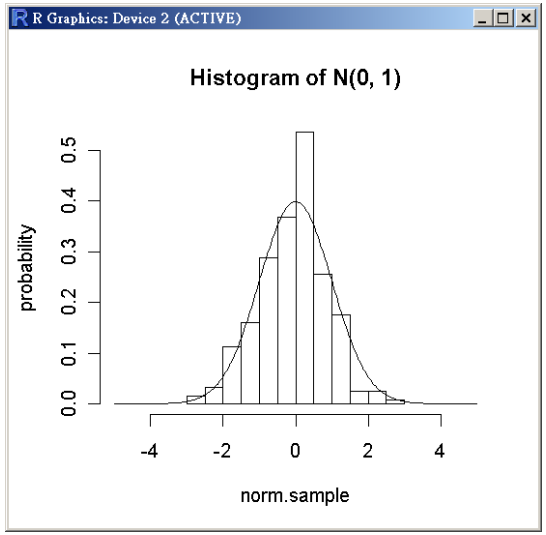


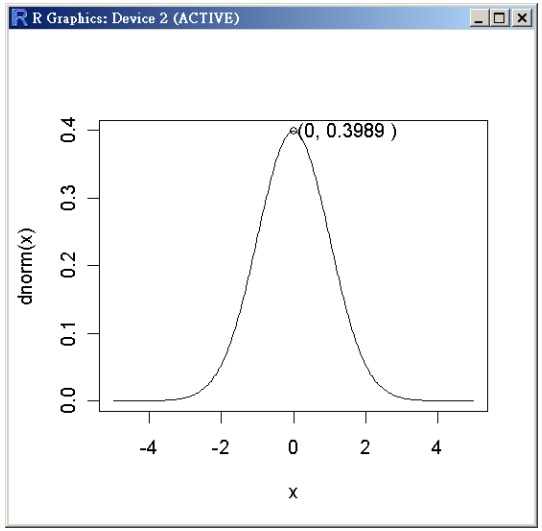
常態分佈

```
> norm.sample <- rnorm(250)
> summary(norm.sample)
> hist(norm.sample, xlim=c(-5, 5), ylab="probability", main="Histogram of N(0, 1)", prob=T)
> x <- seq(from=-5, to=5, length=300)
> lines(x, dnorm(x))
```

標出最頂點的座標

```
> x <- seq(from=-5, to=5, length=300)
> plot(x, dnorm(x), type="l")
> points(0, dnorm(0))
> height <- round(dnorm(0), 4); height
> text(1.5, height, paste("(0,", height, ")"))
```



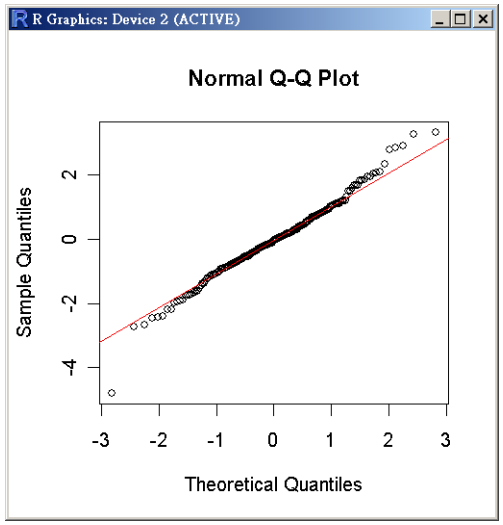


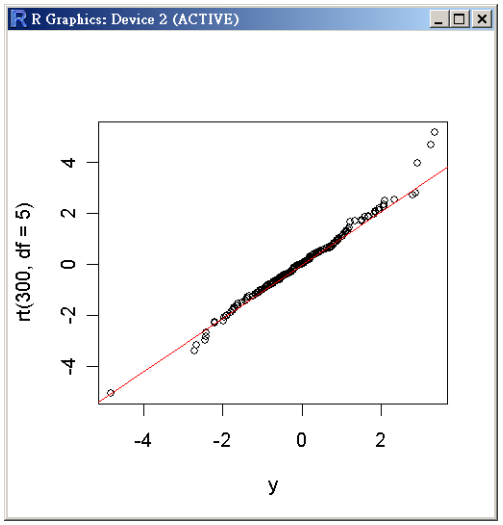
Quantile-Quantile Plots

- QQplot 用來檢驗資料的常態性質。

- **qqnorm()** #produces a normal QQ plot of the values in y.
- **qqline()** #adds a line to a normal quantile-quantile plot which passes through the first and third quartiles.
- **qqplot()** #produces a QQ plot of two datasets.

```
> y <- rt(200, df = 5)
> qqnorm(y)
> qqline(y, col = 2)

> qqplot(y, rt(300, df = 5))
> qqline(y, col = 2)
```





實作 Quantile-Quantile Plots

$(X_1, X_2, \cdots, X_n)$

1. 計算樣本平均數及樣本變異數。

$$\bar{X} = \frac{\sum X_i}{n}$$

$$S^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$$

2. 將隨機樣本標準化並排序。 $d_{(1)}, d_{(2)}, \cdots, d_{(n)}$

$$d_{(i)} = \frac{X_i - \bar{X}}{S}$$

3. 查出 n個標準常態值: (將標準態分配，區分成n+1區塊，最左及最右區塊的機率分別為1/2n, 中間的 n-1區塊, 機率分別為1/n)。

$$q_{(1)} = z_{\frac{1}{2n}}, q_{(2)} = z_{\frac{3}{2n}}, \cdots, q_{(n)} = z_{\frac{2n-1}{2n}}$$

$$q_{(i)} = z_{\frac{2i-1}{2n}}$$

$$P(Z < q_{(i)}) = \frac{2i-1}{2n}$$

4. 畫散佈圖: x軸: 排序的標準化樣本，y軸: 標準常態值。

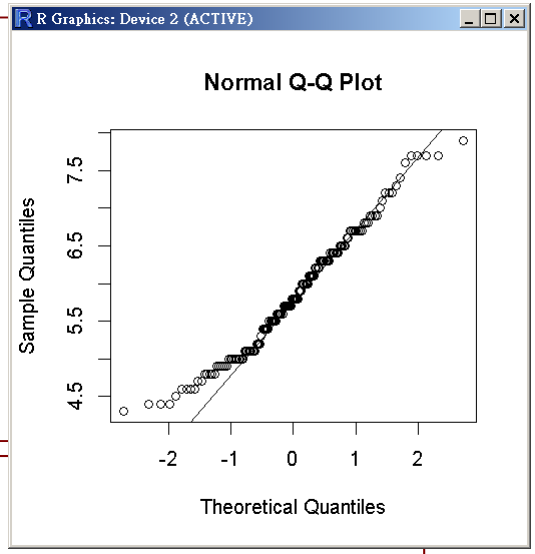
$$(d_{(i)}, q_{(i)})$$

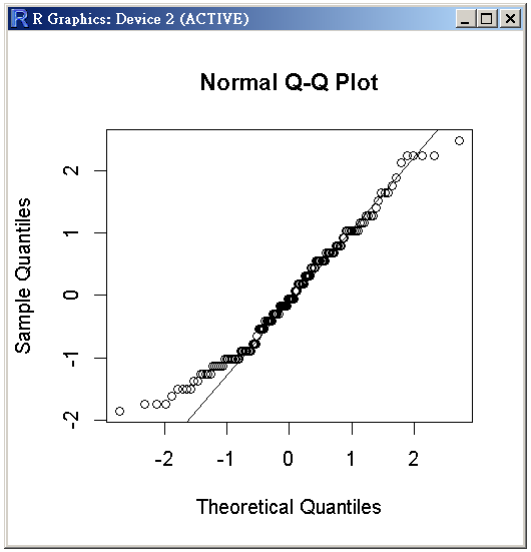
5. 加入一條由(q(i), q(i))產生的標準常態直線。

$$(q_{(i)}, q_{(i)})$$

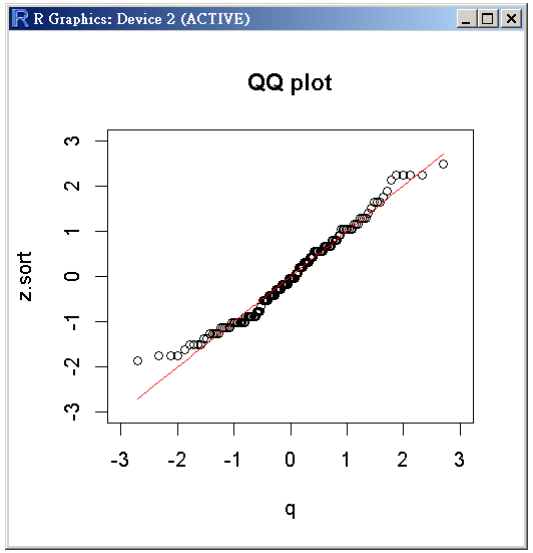
課堂練習

```
> qqnorm(iris[,1])
> qqline(iris[,1])

> qqnorm(scale(iris[,1]))
> qqline(scale(iris[,1]))

> my.qqplot(iris[,1])
```






```
my.qqplot <- function(x){
   x.mean <- mean(x)
   x.var <- var(x)
   n <- length(x)

   z <- (x-x.mean)/sqrt(x.var)
   z.mean <- mean(z)
   z.var <- var(z)
   z.sort <- sort(z)

   k <- 1:n
   p <- (k-0.5)/n
   q <- qnorm(p)

   plot(q, z.sort, xlim=c(-3, 3), ylim=c(-3, 3))
   title("QQ plot")
   lines(q, q, col=2)
}
```



大數法則: The Law of Large Numbers

If $X_1, X_2, \cdots$, an infinite sequence of i.i.d. random variables with finite expected value $E(X_1) = E(X_2) = \cdots = \mu < \infty$, then

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \to \mu \quad \text{as} \quad n \to \infty$$

- 由具有有限(finite)平均數μ的母體隨機抽樣，隨著樣本數*n*的增加，樣本平均數 $\bar{X}_n$ 越接近母體的均數μ 。
- 樣本平均數的這種行為稱為大數法則(law of large numbers)。

特別注意: 本投影片中符號n和m之區別。

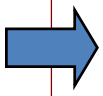
Bernoulli試驗

- Bernoulli試驗(伯努利試驗): 擲一公平硬幣一次，可能出現正面或反面。
- 令*X=1*為出現正面, *X=0*為出現反面。
- *X~Binomial(1, 0.5)*。
- 伯努利分佈的平均數 *p*。

$$X_1, X_2, \cdots, \ Binomial(1, 0.5)$$

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$ ：平均正面次數

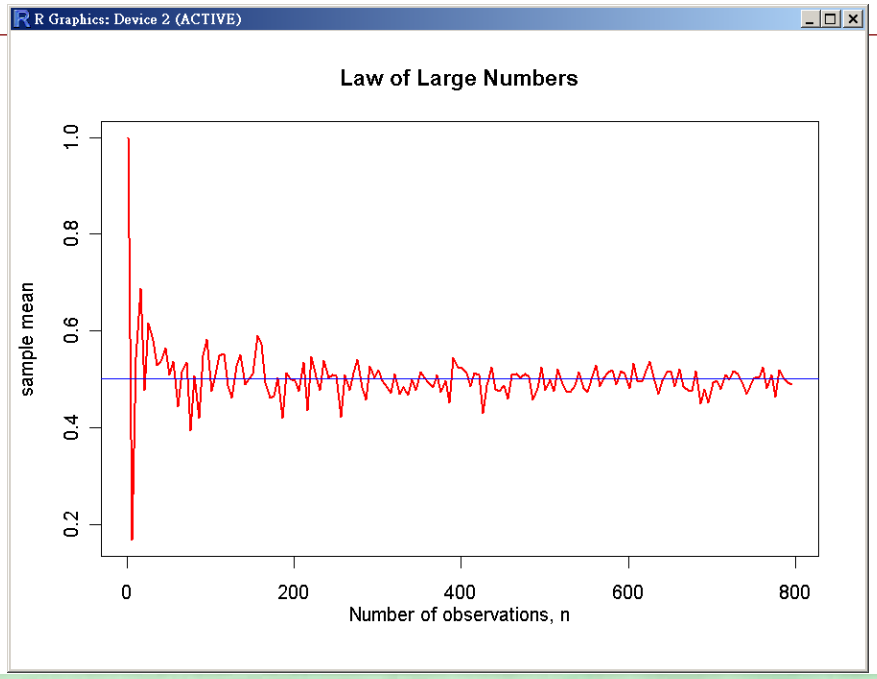
rbinom(m, size=1, prob)
m: number of observations (樣本數)
size=1: number of trials
prob: probability of success on each trial



m Bernoulli random samples:
rbinom(m, 1, 0.5)

利用Bernoulli試驗說明大數法則

```
sample.size <- seq(from=1, to=800, by=5)
m <- length(sample.size)
xbar <- numeric(m)
for(i in 1:m){
  xbar[i] <- mean(rbinom(sample.size[i], 1, 0.5))
}
plot(sample.size, xbar, xlab="Number of observations, n",
ylab="sample mean", main="Law of Large Numbers", type="l",
col="red", lwd=1.5)
abline(h=0.5, col="blue")
```



中央極限定理 (Central Limit Theorem)

- 由一具有平均數μ，標準差σ的母體中抽取樣本大小為*n*的簡單隨機樣本，當樣本大小*n*夠大時，樣本平均數的抽樣分配會近似於常態分配。
- 在一般的統計實務上，大部分的應用中均假設當樣本大小為30(含)以上時， 的抽樣分配即近似於常態分配。
- 當母體為常態分配時，不論樣本大小，樣本平均數的抽樣分配仍為常態分配。

$X_1, X_2, X_3, \cdots$ be a set of n independent and identically distributed random variables having finite values of mean μ and variance $\sigma^2 > 0$.

$$S_n = X_1 + \cdots + X_n$$

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \rightarrow N(0,1) \quad \text{as} \quad n \rightarrow \infty$$

$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$

$$Var(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

算機率

- 於某考試中，考生之通過標準機率為0.7，以隨機變數表示考生之通過與否(X=1表示通過) (X=0表示不通過)，其機率分配為 P(X=1)=0.7, P(X=0)=0.3。
  1. 計算母體平均數及變異數。
  2. 假如有210名考生，計算「平均通過人數」的平均數及變異數。
  3. 計算通過人數> 126的機率。

1.

$$\mu = E(X) = p = 0.7$$

$$\sigma^2 = Var(X) = p(1-p) = 0.21$$

2.

$X_1, X_2, \cdots, X_{210}$:

$X_i = 1$: success

$X_i = 0$: fail

$$\bar{X}_{210} = \frac{X_1 + \cdots + X_{210}}{210}$$

$$\mu_{\bar{X}} = \mu = 0.7$$

$$\sigma_{\bar{X}} = \frac{\sigma^2}{210} = 0.001$$

3.

$$P(X_1 + X_2 + \cdots + X_{210} > 126)$$

$$= P(\bar{X} > \frac{126}{210})$$

$$= P(\bar{X} > 0.6)$$

$$= P(Z > \frac{0.6 - 0.7}{\sqrt{0.001}})$$

$$= P(Z > -3.16228)$$

$$= 0.99922$$

使用R運算

通過人數>126的機率

```
> z <- (126/210 -0.7)/sqrt(0.001)
> -3.16228
[1] -3.16228
> 1-pnorm(z)
[1] 0.9992173
```

寫一「通過人數大於某數的機率」之副程式

- n: 考生總數(210)
- X: 通過考生之人數, X~B(210, 0.7)

```
> pass.prob <- function(x, n, mu, sigma2, digit=m){
    xbar <- x/n
    z <- (xbar-mu)/sqrt(sigma2)
    zvalue <- round(z, digit)
    right.prob <- round(1-pnorm(z), digit)
    list(zvalue=zvalue, prob=right.prob)
  }

> pass.prob(126, 210, 0.7, 0.001, 4)
$zvalue
[1] -3.1623

$prob
[1] 0.9992
```

驗証中央極限定理

1. 先做隨機樣本的取樣。

$X \sim D(\cdot)$ $X_1, X_2, \cdots, X_{m_0} \sim D(\cdot)$ $m = m_0$

2. 計算樣本平均。

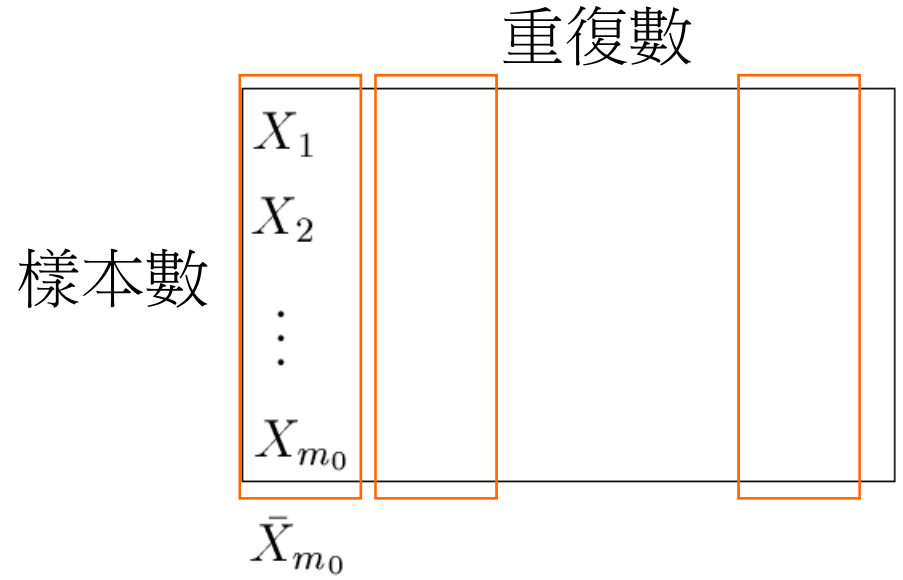
$$\bar{X}_{m_0} = \frac{1}{m_0}(X_1 + X_2 + \cdots + X_{m_0})$$

3. 重復上述動作數百或數仟次，得到抽樣平均的分佈。
4. 描繪出抽樣平均之抽樣分配直方圖。
5. 畫出相對應的qqplot。
6. 再做各種不同樣本數(m0=1,5,15,30,...)的抽樣計算。

範例: Uniform Distribution

$$X_1, X_2, \cdots \sim U(5, 80)$$

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$



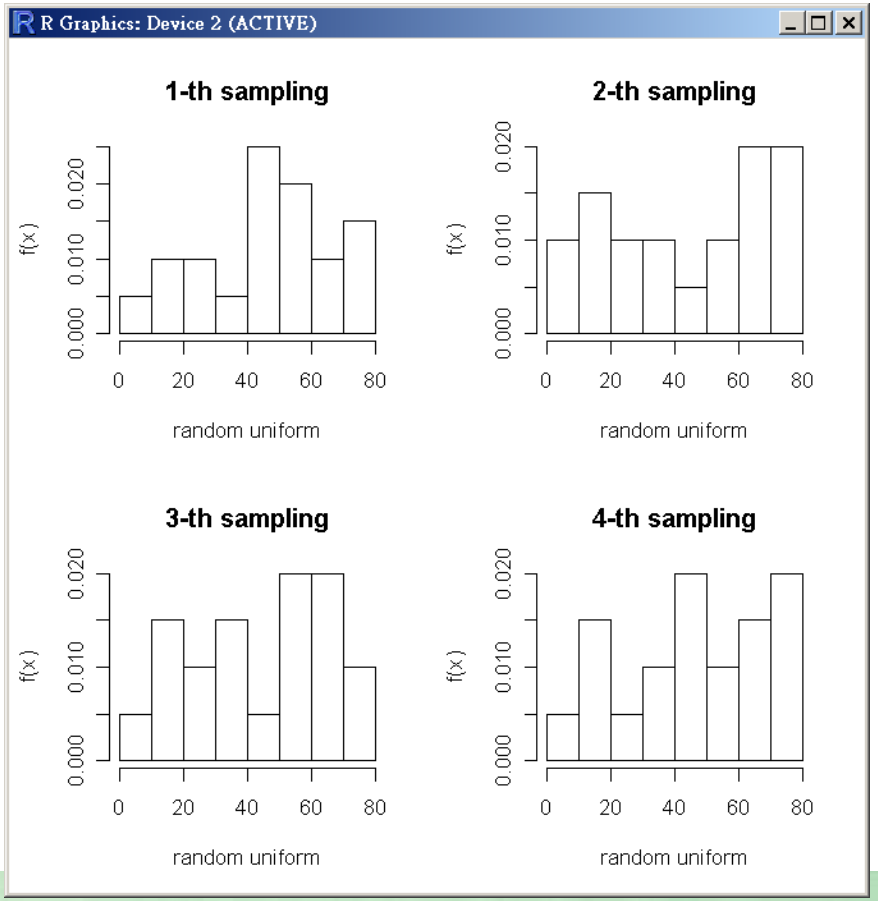

```
umin <- 5
umax <- 80
n.sample <- 20
n.repeated <- 500

RandomSample <- matrix(0, n.sample, n.repeated)
for(i in 1:n.repeated){
   rnumber <- runif(n.sample, umin, umax)
   RandomSample[,i] <- as.matrix(rnumber)

}
dim(RandomSample)
```

抽樣樣本之直方圖

```
par(mfrow=c(2,2))
for(i in 1:4){
  title <- paste(i,"-th sampling", sep="")
  hist(RandomSample[,i], ylab="f(x)", xlab="random uniform", pro=T, main=title)
}
```

$$X_1, X_2, \cdots \sim U(5, 80)$$

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

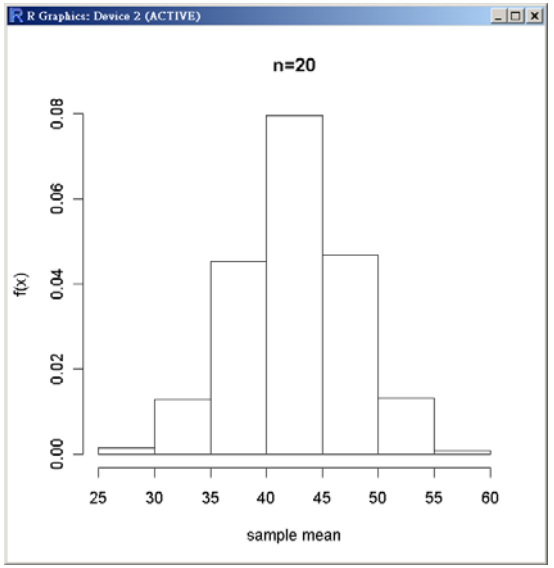


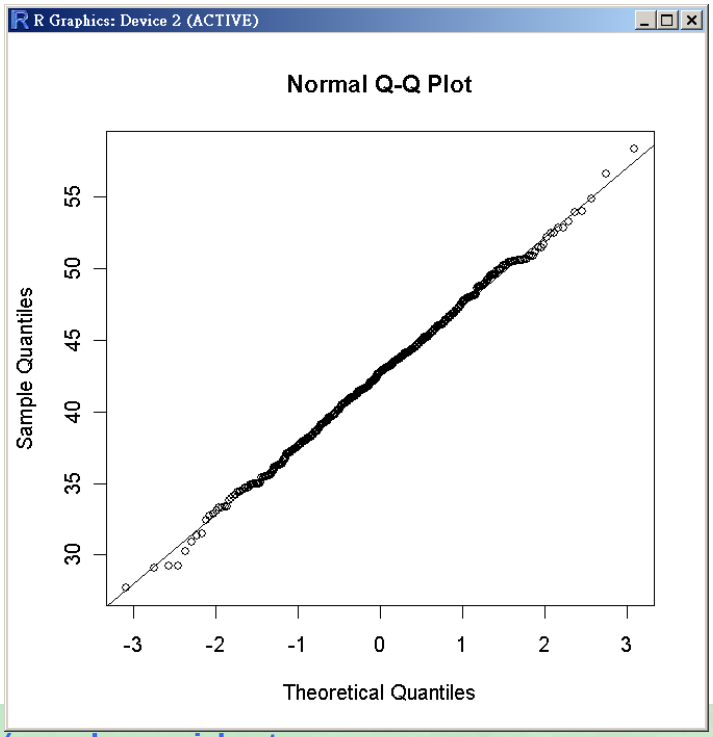
抽樣樣本平均之直方圖&QQplot

```
> SampleMean <- apply(RandomSample, 2, mean)
> hist(SampleMean, ylab="f(x)", xlab="sample mean", pro=T, main="n=20")
```

$$X_1, X_2, \cdots \sim U(5, 80)$$

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

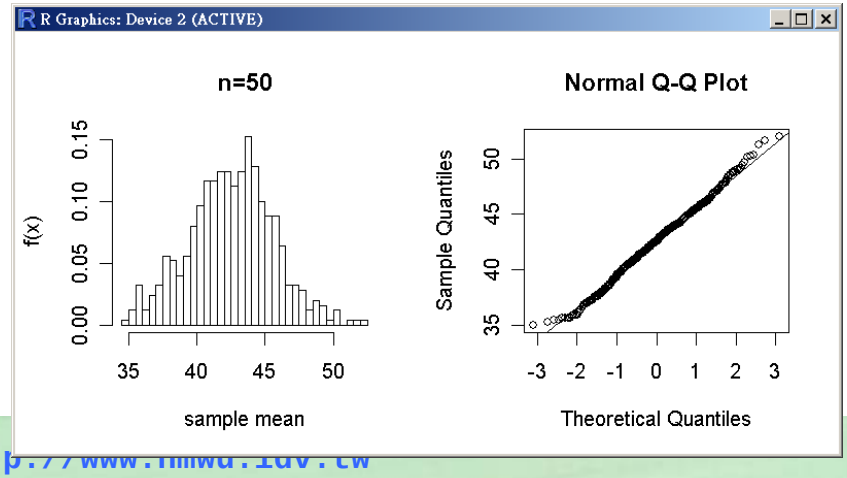





```
> qqnorm(SampleMean)
> qqline(SampleMean)
```

重複不同的樣本數

```
CLT.unif <- function(umin, umax, n.sample, n.repeated){
    RandomSample <- matrix(0, n.sample, n.repeated)
    for(i in 1:n.repeated){
       rnumber <- runif(n.sample, umin, umax)
       RandomSample[,i] <- as.matrix(rnumber)

    }
    SampleMean <- apply(RandomSample, 2, mean)
    par(mfrow=c(1,2))
    title <- paste("n=",n.sample, sep="")
    hist(SampleMean, breaks=30, ylab="f(x)", xlab="sample mean", pro=T,
main=title)
    qqnorm(SampleMean)
    qqline(SampleMean)
}
```

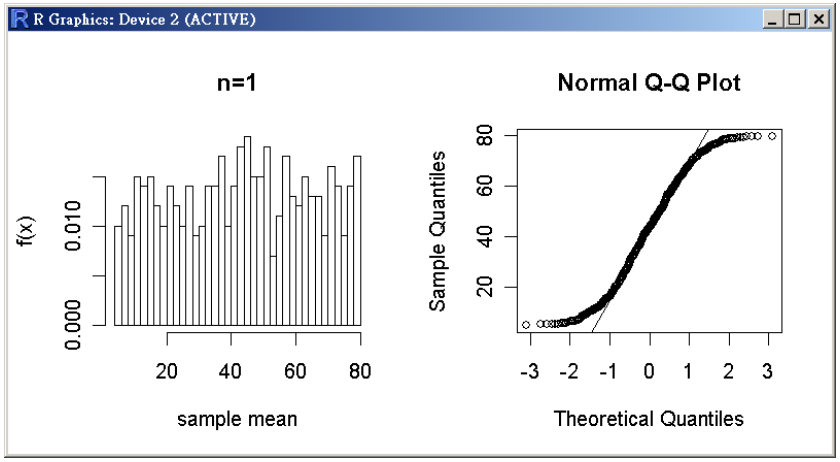



```
CLT.unif(5, 80, 50, 500)
```

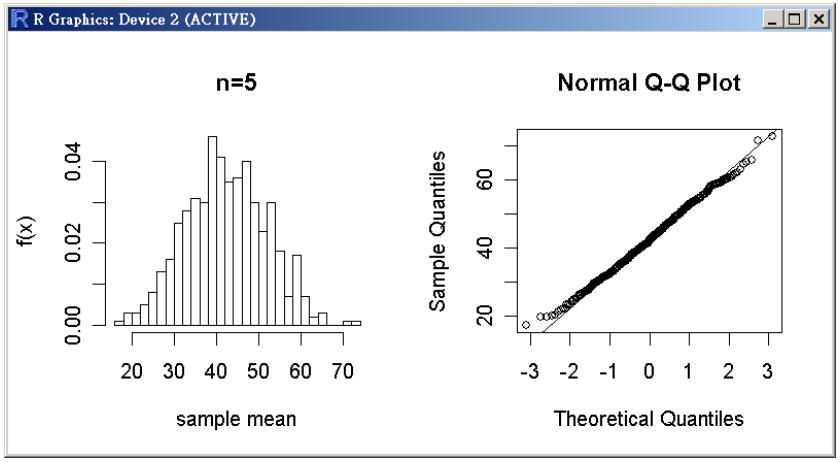
當樣本數*n*愈大時，從樣本平均數的抽樣分配可以得到「中央極限定理」的主要結論。

CLT.unif(umin, umax, n.sample, n.repeated

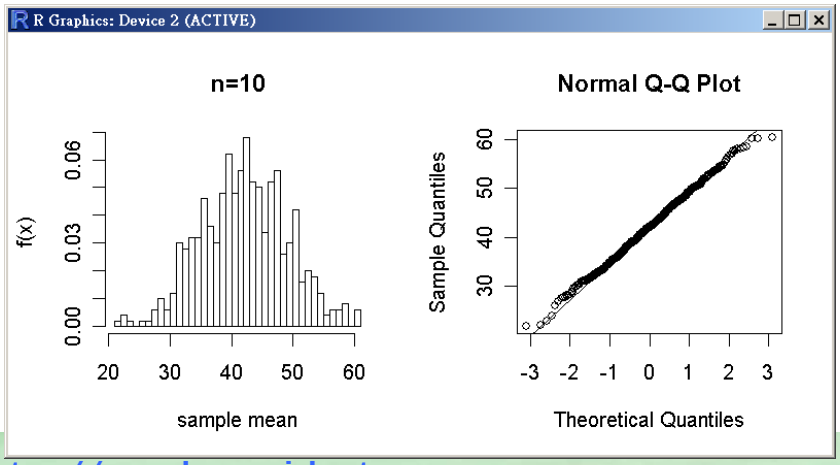
CLT.unif(5, 80, 1, 500)



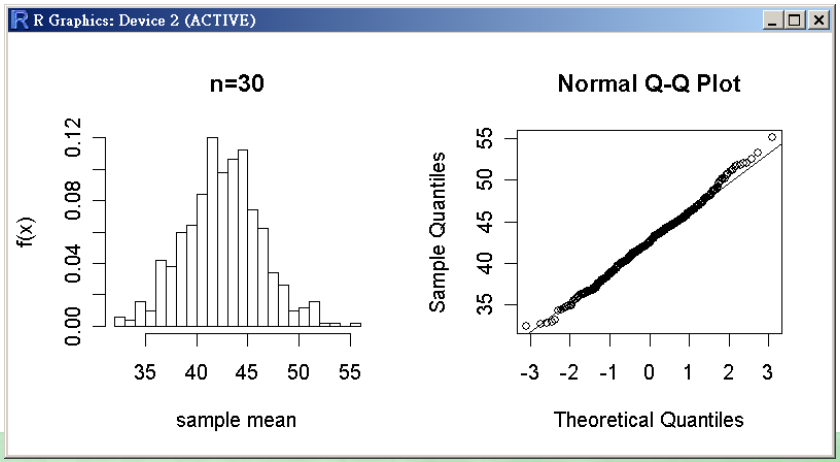
CLT.unif(5, 80, 5, 500)



CLT.unif(5, 80, 10, 500)



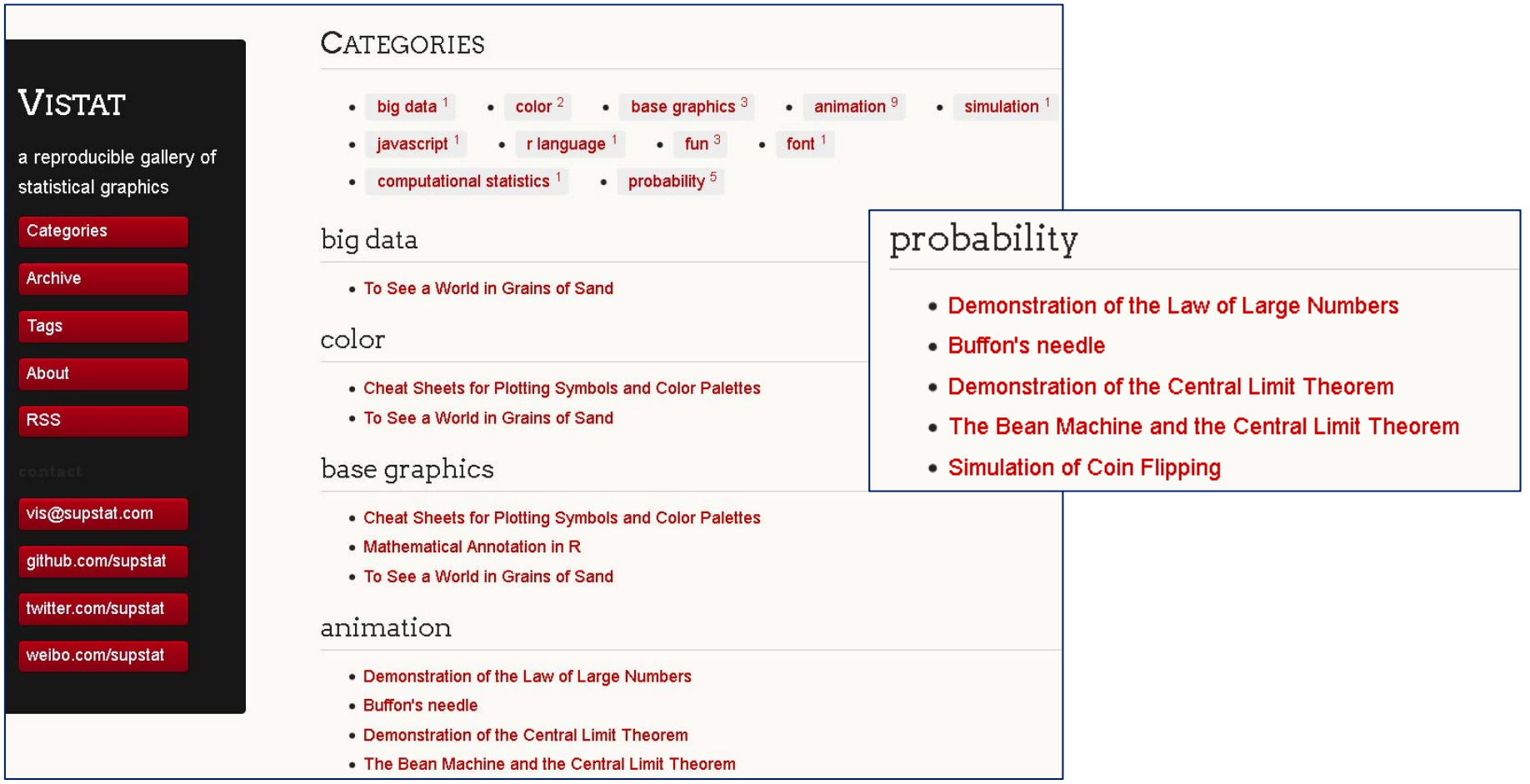
CLT.unif(5, 80, 30, 500)



R package: animation

http://cran.r-project.org/web/packages/animation/index.html
http://yihui.name/animation/

A Gallery of Animations in Statistics and Utilities to Create Animations



維度縮減 (Dimension Reduction)

keep information as much as possible without loss of information.

input data matrix: X

Group	Data	X1	X2	X3	•••	Xp
1	subject01	0.81	-1.29	-0.50		1.13
1	subject02	0.64	2.16	-1.51		0.00
2	subject03	0.13	0.60	1.10		0.11
2	subject04	-0.17	0.31	-0.37		-0.50
3	subject05	-1.01	0.99	0.70		-0.08
3	subject06	0.95	0.75	-0.83		0.60
3	subject07	0.72	1.12	-1.35		-1.22
1	subject08	0.77	1.24	-0.04		1.03
1	subject09	-0.49	0.02	-1.73		1.61
1	subject10	1.93	0.45	-0.01		0.03
3	subject11	-0.15	-1.36	1.05		0.50
3	subject12	-1.16	0.11	-0.57		-0.80
3	subject13	-0.02	2.05	-1.18		0.45
3	subject14	-0.05	0.79	1.33		0.81
2	subject15	-0.21	-0.38	0.72		-0.61
1	subject16	-0.28	0.57	1.02		-0.01
⋮	⋮					
2	subjectN	0.33	0.01	1.19		-0.33

SVD
PCA
FA
MDS

transformed data matrix: Z

Data	Z1	Z2	•••	Zk
subject01	-1.55	-0.66		0.60
subject02	0.57	-0.51		-1.03
subject03	1.99	1.44		-0.60
subject04	0.10	0.20		-1.21
subject05	-0.20	-0.64		0.24
subject06	2.85	1.32		-0.61
subject07	-0.34	-0.35		0.15
subject08	0.66	0.44		0.28
subject09	8.44	1.66		2.12
subject10	0.37	-0.17		-1.73
subject11	-1.14	0.01		0.61
subject12	-1.73	-1.13		0.81
subject13	1.53	0.67		0.48
subject14	-0.21	-0.14		-0.29
subject15	-0.03	0.66		0.17
subject16	2.56	-2.25		0.26
⋮				
subjectN	2.04	0.71		0.76

Visualization
Clustering
Classification
....

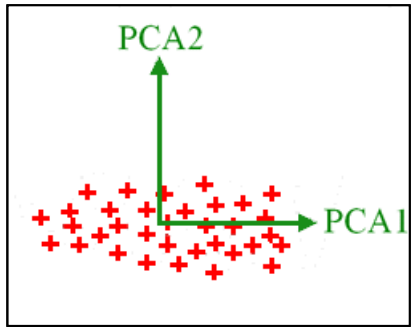
Y

Methods using additional information *y:*
LDA, Sufficient Dimension Reduction (SIR, SAVE, pHd, IRE,...)

主成份分析

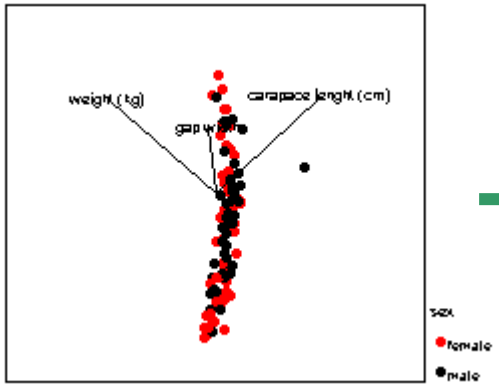
Principal Component Analysis (PCA)

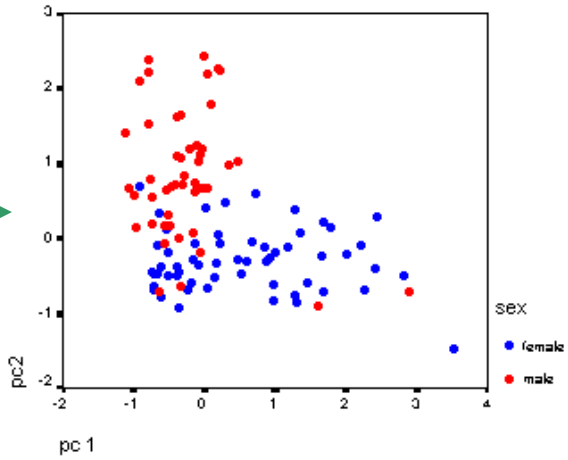
PCA is a method that reduces data dimensionality by finding the new variables (major axes, principal components).



$$\mathrm{PCA}_1 = a_1\mathbf{X} + b_1\mathbf{Y}$$

$$\mathrm{PCA}_2 = a_2\mathbf{X} + b_2\mathbf{Y}$$





$$\mathrm{PCA}_1 = a_1\mathbf{X} + b_1\mathbf{Y} + c_1\mathbf{Z}$$

$$\mathrm{PCA}_2 = a_2\mathbf{X} + b_2\mathbf{Y} + c_2\mathbf{Z}$$

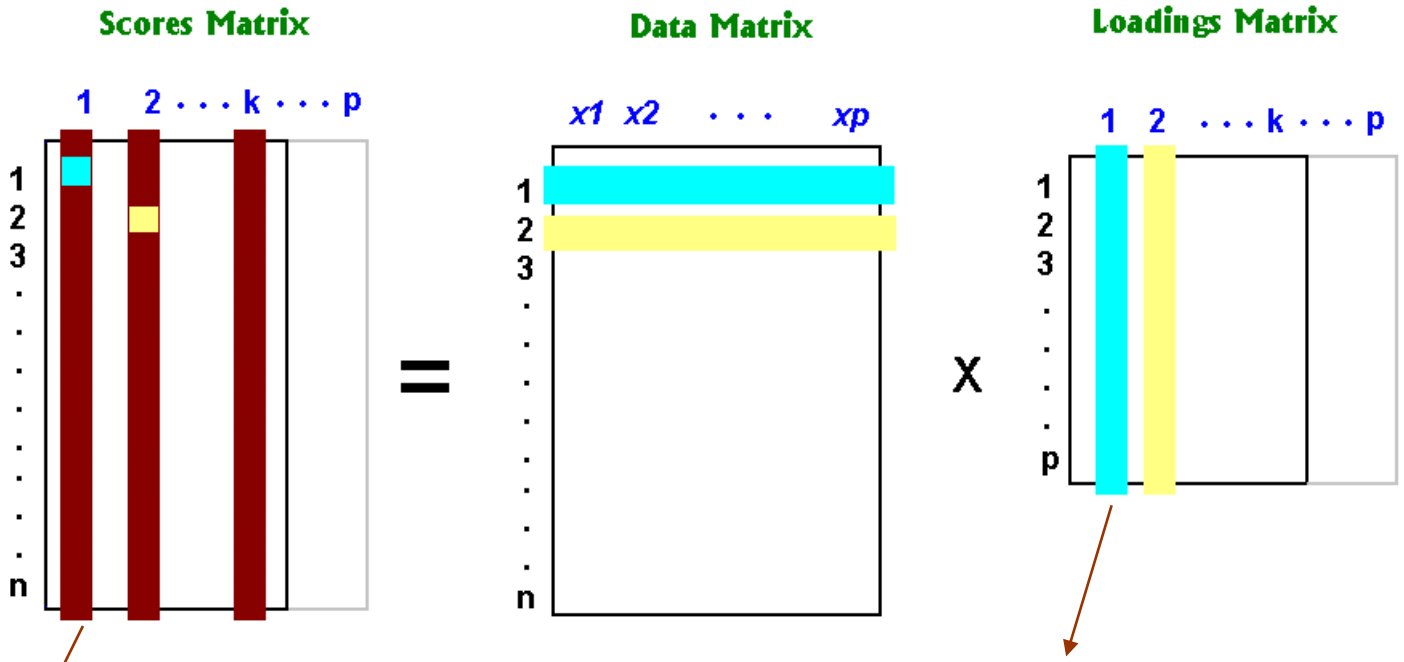
$$\mathrm{PCA}_1 = a_{11}\mathbf{X}_1 + a_{12}\mathbf{X}_2 + \cdots + a_{1p}\mathbf{X}_p$$

$$\mathrm{PCA}_2 = a_{21}\mathbf{X}_1 + a_{22}\mathbf{X}_2 + \cdots + a_{2p}\mathbf{X}_p$$

Amongst all possible projections, PCA finds the projections so that the maximum amount of information, measured in terms of variability, is retained in the smallest number of dimensions.

PCA: Loadings and Scores

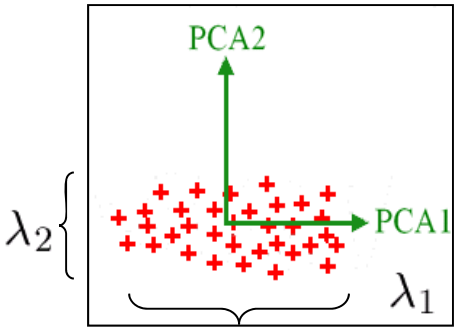
$$\mathbf{Z} = \mathbf{X}\,\mathbf{W}$$



The ith principal component of $\mathbf{X}$ is $\mathbf{X}\mathbf{w}_i$, where $\mathbf{w}_i$ is the ith normalized eigenvector of $\Sigma_{\mathbf{x}}$ corresponding to the ith largest eigenvaules.

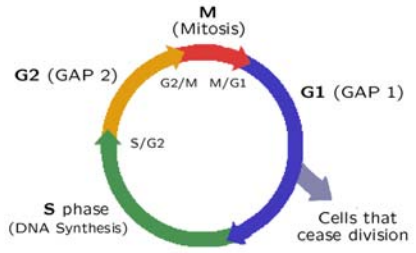
Eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$

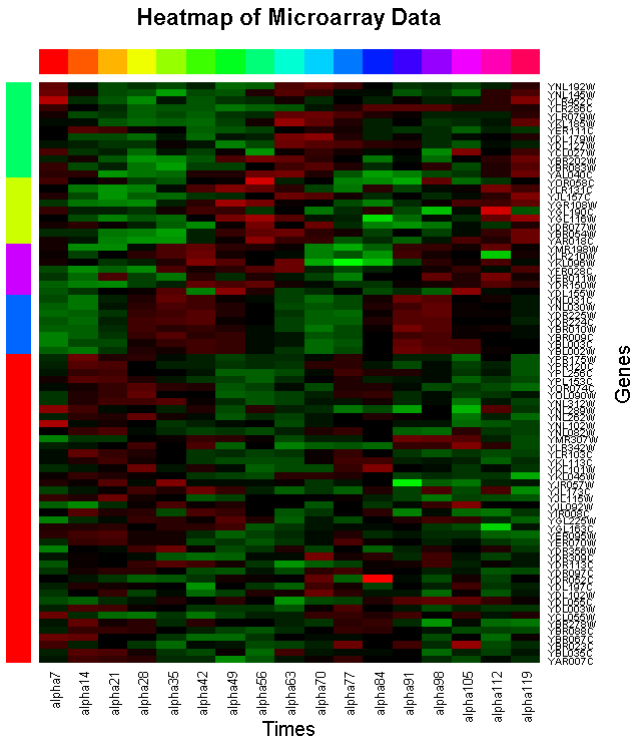
$$\text{proportion} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

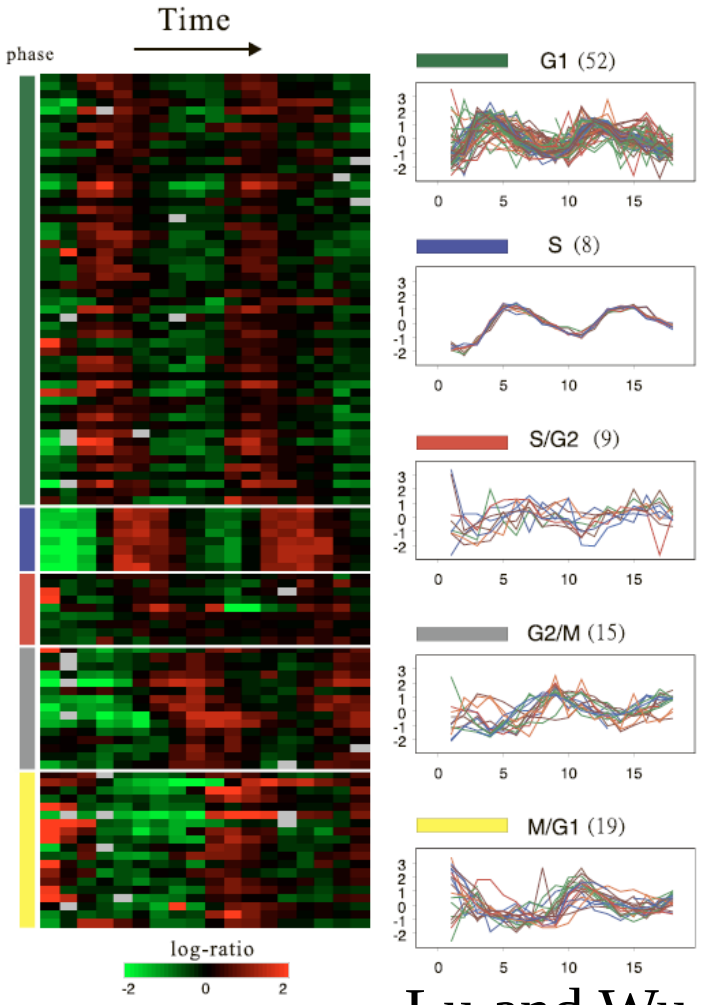


Microarray Data of Yeast Cell Cycle

- Synchronized by alpha factor arrest method (Spellman et al. 1998; Chu et al. 1998)
- 103 known genes: every 7 minutes and totally 18 time points. (remove NA's: 79 genes)

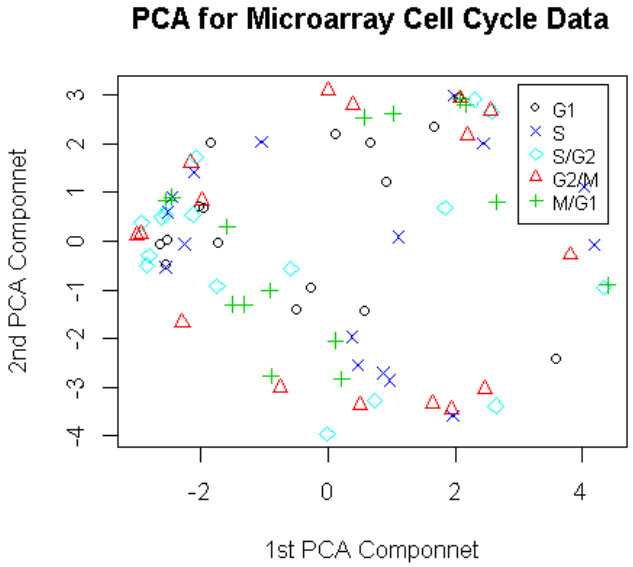


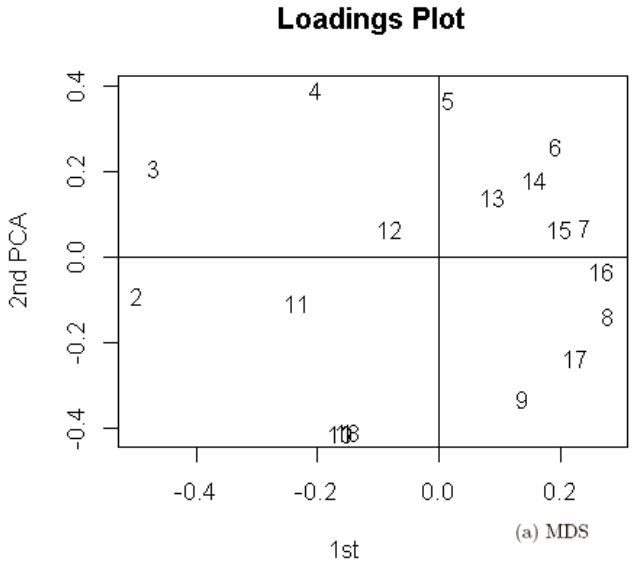


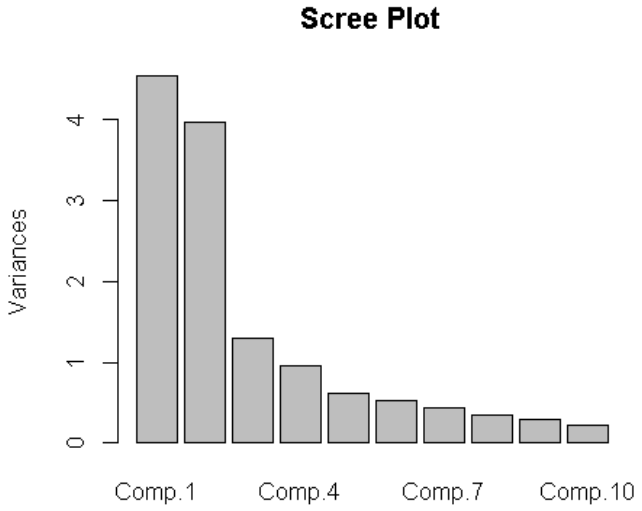


Lu and Wu (2010)

PCA Results







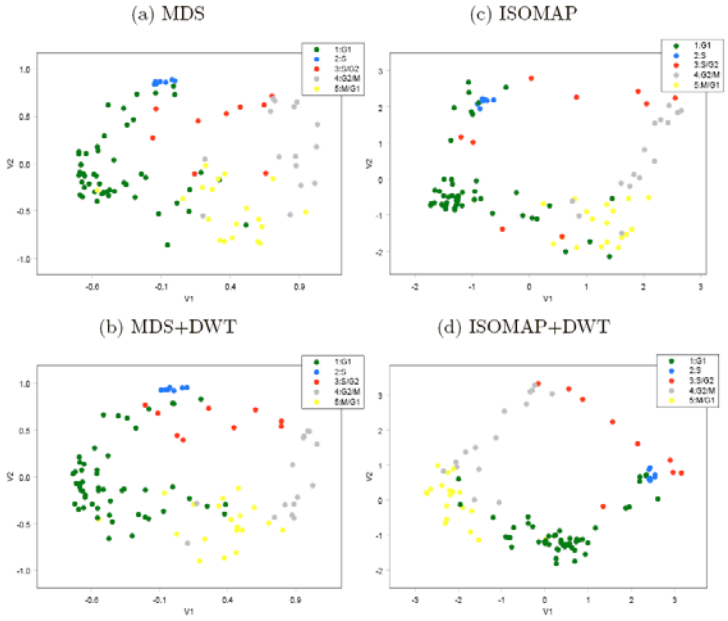


Figure 2: The 2D representation of 103 cell cycle-regulated genes are displayed.

R Code

```
cell.matrix  <- read.table("YeastCellCycle_alpha.txt", header=TRUE, row.names=1)
n <- dim(cell.matrix)[1]
p <- dim(cell.matrix)[2]-1
cell.data <- cell.matrix[,2:p+1]
gene.phase <- cell.matrix[,1]
phase <- unique(gene.phase)
phase.name <- c("G1", "S", "S/G2", "G2/M", "M/G1")
#cell.sdata <- scale(cell.data)
cell.sdata <- (cell.data – apply(cell.data, 1, mean))/sqrt(apply(cell.data, 1, var))
cell.pca <- princomp(cell.sdata, scores=TRUE)

# 2D plot for first two components
pca.dim1 <- cell.pca$scores[,1]
pca.dim2 <- cell.pca$scores[,2]
plot(pca.dim1, pca.dim2, main="PCA for Microarray Cell Cycle Data", xlab="1st PCA Componnet",
     ylab="2nd PCA Componnet",col=c(phase), pch=c(phase))
legend(3, 3.2, phase.name, pch=c(phase), col=c(phase), cex=0.8)

# shows a screeplot
plot(cell.pca, main="Scree Plot")

## loadings plot
plot(loadings(cell.pca)[,1], loadings(cell.pca)[,2], xlab="1st PCA", ylab="2nd PCA",
main="Loadings Plot", type="n")
text(loadings(cell.pca)[,1], loadings(cell.pca)[,2], labels=paste(1:p))
abline(h=0)
abline(v=0)

# print loadings
loadings(cell.pca)
summary(cell.pca)
```

```
> summary(cell.pca)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6      C
Standard deviation     2.133482 1.9948646 1.14254479 0.97400844 0.78526679 0.73128905 0.663
Proportion of Variance 0.325943 0.2849644 0.09347819 0.06793432 0.04415687 0.03829499 0.031
Cumulative Proportion  0.325943 0.6109074 0.70438562 0.77231994 0.81647681 0.85477180 0.886
```

模擬算機率

- 一對夫婦計劃生孩子生到有女兒才停，或生了三個就停止。他們會擁有女兒的機率是多少？
- 腎臟移植的病人資料：撐過移植手術的占90%，另外10%會死亡。在手術後存活的人中有60%移植成功，另外的40%還是得回去洗腎。五年存活率對於換了腎的人來說是70% ，對於回去洗腎的人來說是50%。計算能活過五年的機率。