資料轉換

吳漢銘

國立臺北大學 統計學系

http://www.hmwu.idv.tw

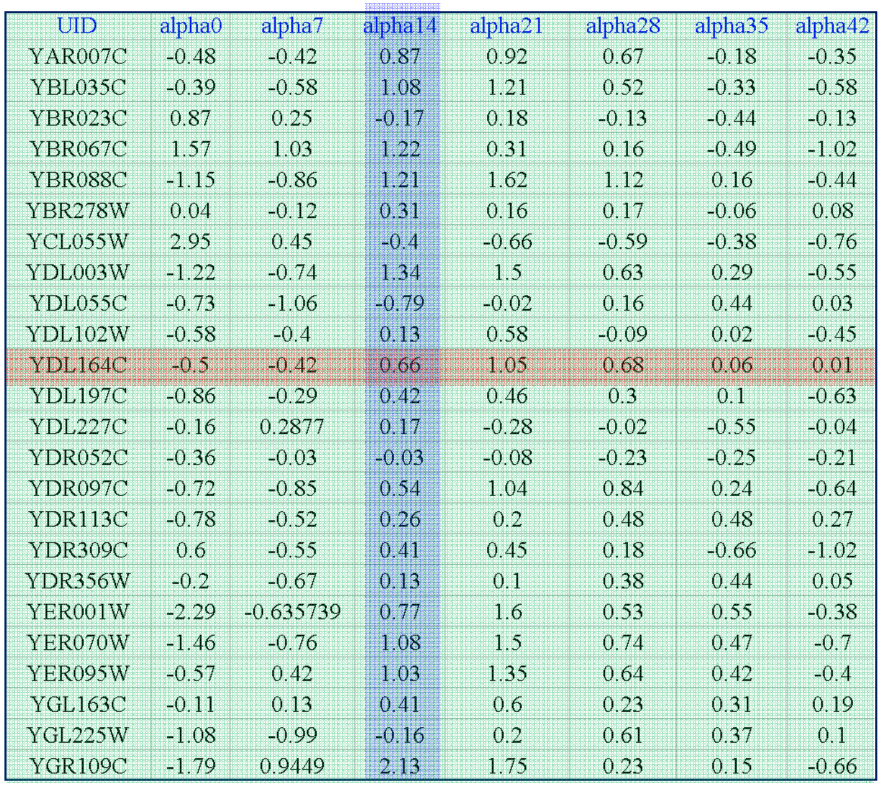
資料轉換 - 大綱

- **主題1**
  - 為什麼要做資料轉換?
  - 常見的資料轉換方式
  - 對數轉換 (Log Transformation)
  - Box-Cox Transformation
  - 標準化 (Standardization)
  - 要使用哪一種資料轉換方式?

Classical (Numerical) Data Table

*j*th variable

*i*th subject
(*i*th sample)



UID	alpha0	alpha7	alpha14	alpha21	alpha28	alpha35	alpha42
YAR007C	-0.48	-0.42	0.87	0.92	0.67	-0.18	-0.35
YBL035C	-0.39	-0.58	1.08	1.21	0.52	-0.33	-0.58
YBR023C	0.87	0.25	-0.17	0.18	-0.13	-0.44	-0.13
YBR067C	1.57	1.03	1.22	0.31	0.16	-0.49	-1.02
YBR088C	-1.15	-0.86	1.21	1.62	1.12	0.16	-0.44
YBR278W	0.04	-0.12	0.31	0.16	0.17	-0.06	0.08
YCL055W	2.95	0.45	-0.4	-0.66	-0.59	-0.38	-0.76
YDL003W	-1.22	-0.74	1.34	1.5	0.63	0.29	-0.55
YDL055C	-0.73	-1.06	-0.79	-0.02	0.16	0.44	0.03
YDL102W	-0.58	-0.4	0.13	0.58	-0.09	0.02	-0.45
YDL164C	-0.5	-0.42	0.66	1.05	0.68	0.06	0.01
YDL197C	-0.86	-0.29	0.42	0.46	0.3	0.1	-0.63
YDL227C	-0.16	0.2877	0.17	-0.28	-0.02	-0.55	-0.04
YDR052C	-0.36	-0.03	-0.03	-0.08	-0.23	-0.25	-0.21
YDR097C	-0.72	-0.85	0.54	1.04	0.84	0.24	-0.64
YDR113C	-0.78	-0.52	0.26	0.2	0.48	0.48	0.27
YDR309C	0.6	-0.55	0.41	0.45	0.18	-0.66	-1.02
YDR356W	-0.2	-0.67	0.13	0.1	0.38	0.44	0.05
YER001W	-2.29	-0.635739	0.77	1.6	0.53	0.55	-0.38
YER070W	-1.46	-0.76	1.08	1.5	0.74	0.47	-0.7
YER095W	-0.57	0.42	1.03	1.35	0.64	0.42	-0.4
YGL163C	-0.11	0.13	0.41	0.6	0.23	0.31	0.19
YGL225W	-1.08	-0.99	-0.16	0.2	0.61	0.37	0.1
YGR109C	-1.79	0.9449	2.13	1.75	0.23	0.15	-0.66

transformation
for each row

transformation
for each column

transformation
for both rows
and columns

為什麼要做資料轉換?

- to make it more closely **the assumptions** of a statistical inference procedure,
- to make it **easier to visualize** (appearance of graphs),
- to improve **interpretability**,
- to make descriptors that have been measured in **different units comparable**,
- to make the relationships among **variables linear**,
- to modify the **weights** of the variables or objects (e.g. give the same length (or norm) to all object vectors)
- to **code** categorical variables into dummy binary variables.

常見的資料轉換方式

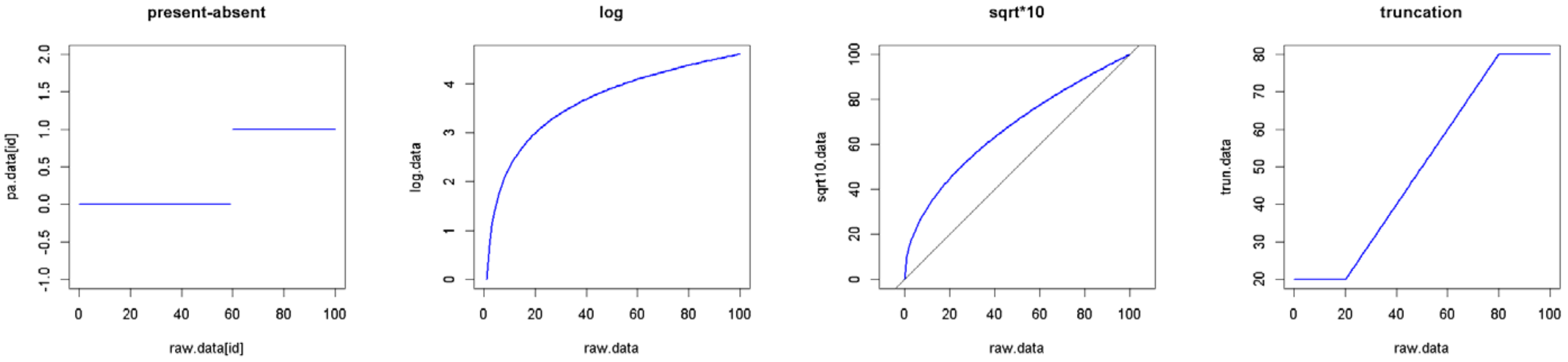
```
par(mfrow=c(1,4))
raw.data <- 0:100
pa.data <- ifelse(raw.data >= 60, 1, 0)
id <- which(pa.data==1)
plot(raw.data[id], pa.data[id], main="present-absent",
+ type="l", lwd=2, col="blue", ylim=c(-1, 2), xlim=c(0, 100))
points(raw.data[-id], pa.data[-id], type="l", lwd=2, col="blue")

log.data <- log(raw.data)
plot(raw.data, log.data, main="log", type="l", lwd=2, col="blue")

sqrt10.data <- sqrt(raw.data)*10
plot(raw.data, sqrt10.data, main="sqrt*10", type="l", lwd=2, col="blue", asp=1)
abline(a=0, b=1)

trun.data <- ifelse(raw.data >= 80, 80, ifelse(raw.data < 20, 20, raw.data))
plot(raw.data, trun.data, main="truncation", type="l", lwd=2, col="blue")
```

NOTE:
```
apply(raw.data.matrix, 2, log)
apply(raw.data.matrix, 2, function(x) sqrt(x)*10)
apply(raw.data.matrix, 2, myfun)
```



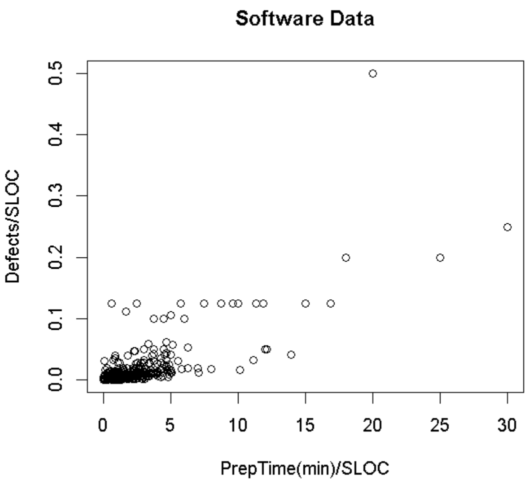
範例: Software Inspection Data

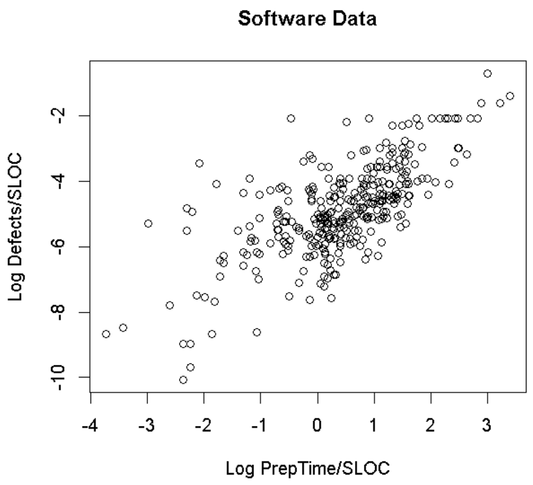
- The data were collected in response to efforts for process improvement in software testing by code inspection.
- The variables are normalized by the size of the inspection (the number of pages or SLOC (single lines of code)):
  - the preparation time in minutes (**`prepage, prepsloc`**),
  - the total work hours in minutes for the meeting (**`mtgsloc`**),
  - and the number of defects found (**`defpage`**, **`defsloc`**).

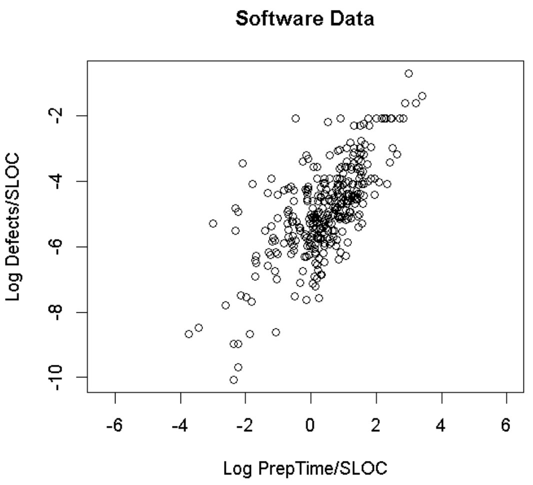
```
> library('R.matlab')
> data <- readMat("software.mat")
> print(data)
...
> str(data)
List of 5
 $ prepsloc: num [1:426, 1] 0.485 0.54 0.54 0.311 0.438 ...
 $ defsloc : num [1:426, 1] 0.005 0.002 0.002 0.00328 0.00278 ...
 $ mtgsloc : num [1:426, 1] 0.075 0.06 0.06 0.2787 0.0417 ...
 $ prepage : num [1:491, 1] 6.15 1.47 1.47 5.06 5.06 ...
 $ defpage : num [1:491, 1] 0.0385 0.0267 0.0133 0.0128 0.0385 ...
```

- **Interested in**: understanding the relationship between the inspection time and the number of defects found.

對數轉換 (Log Transformation)







```
plot(data$prepsloc, data$defsloc, xlab="PrepTime(min)/SLOC", ylab="Defects/SLOC",
main="Software Data")

plot(log(data$prepsloc), log(data$defsloc), xlab="Log PrepTime/SLOC",
ylab="Log Defects/SLOC", main="Software Data")

plot(log(data$prepsloc), log(data$defsloc), xlab="Log PrepTime/SLOC",
ylab="Log Defects/SLOC", main="Software Data", asp=1)
```

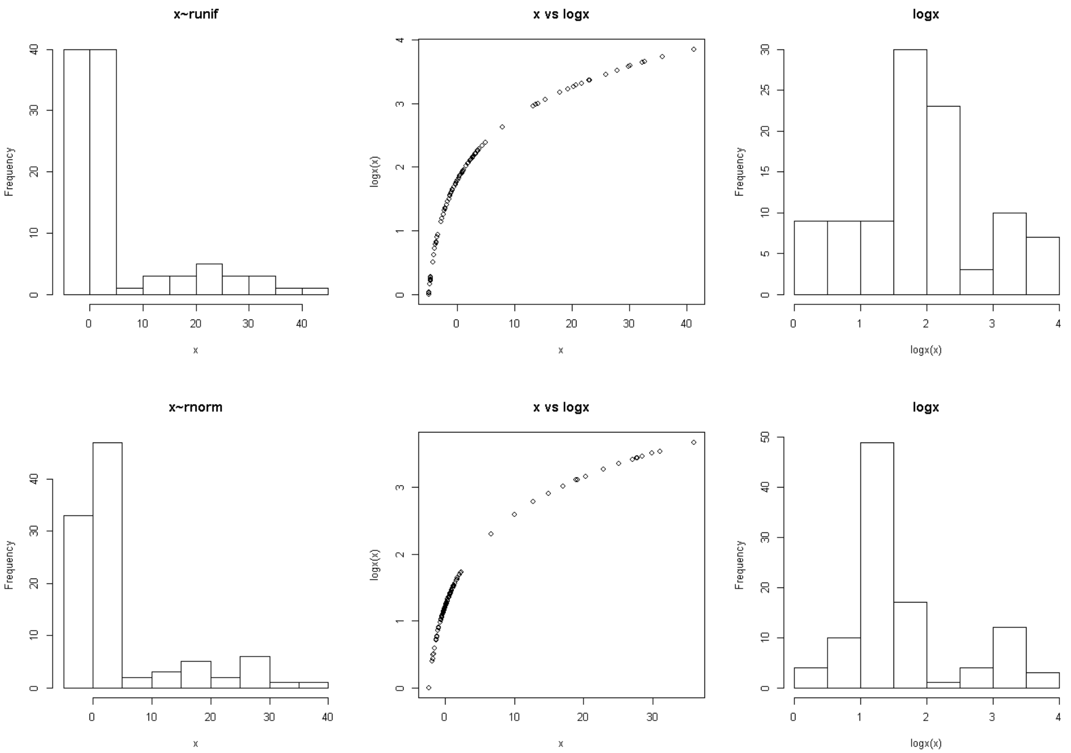
對數轉換: How to Handle Negative Data Values?

Solution 1: Translate, then Transform

- log(x + 1 - min(x))

```
logx <- function(x){
   log(x + 1 - min(x))
}

x <- runif(80, min = -5, max = 5)
# x <- rnorm(80)
x <- c(x, rnorm(20, mean=20, sd=10))
par(mfrow=c(1, 3))
hist(x, main="x~runif")
plot(x, logx(x), main="x vs logx")
hist(logx(x), main="logx")
```



Solution 2: Use Missing Values

- A criticism of the previous method is that some practicing statisticians don't like to add an arbitrary constant to the data.
- They argue that a better way to handle negative values is to use missing values for the logarithm of a nonpositive number.

Box-Cox Transformations

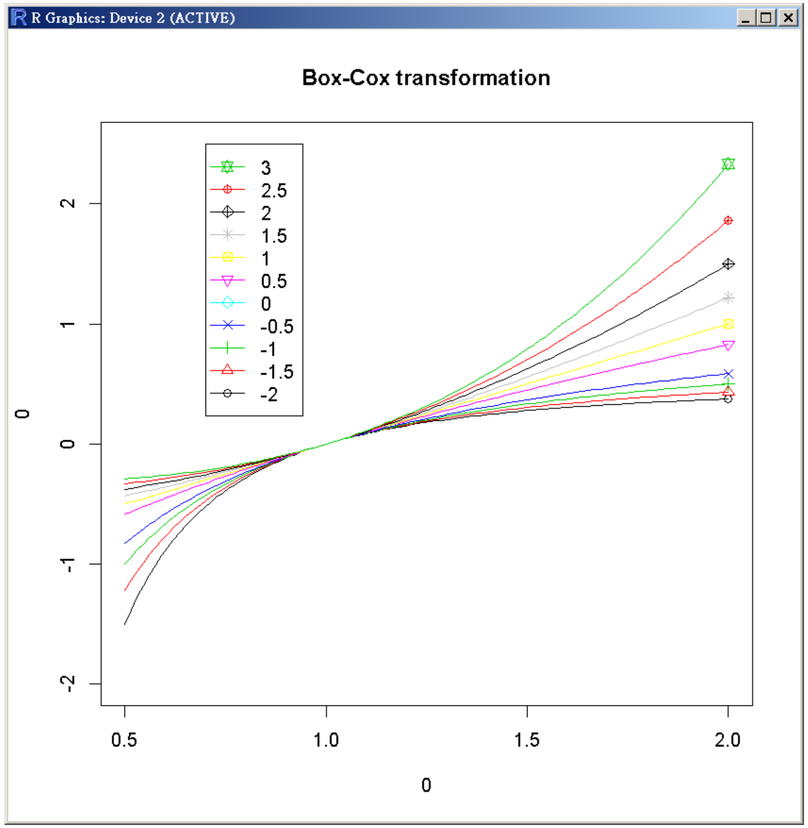
$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

Box and Cox(1964)

- The aim of the Box-Cox transformations is to ensure the usual assumptions for Linear Model hold.

$$\boldsymbol{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$$

```
x <- seq(0.5, 2, length.out=100)
bc <- function(y, lambda){
    (y^lambda -1)/lambda
}
lambda <- seq(-2, 3, 0.5)
plot(0, 0, type="n", xlim=c(0.5, 2),
     ylim=c(-2, 2.5), main="Box-Cox transformation")
for(i in 1:length(lambda)){
   points(x, bc(x, lambda[i]), type="l", col=i)
   points(2, bc(2, lambda[i]), col=i, pch=i)
}
legend(0.7, 2.5, legend=as.character(rev(lambda)),
       lty=1, pch=length(lambda):1,
       col=length(lambda):1)
```



http://onlinestatbook.com/2/transformations/box-cox.html

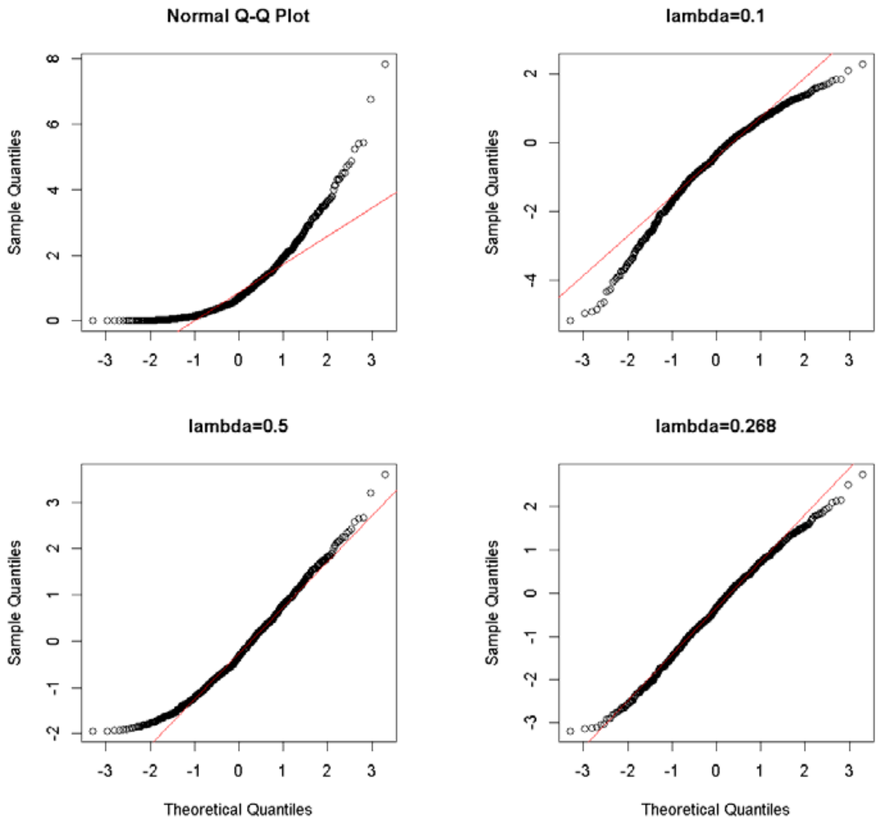
Box-Cox Transformations

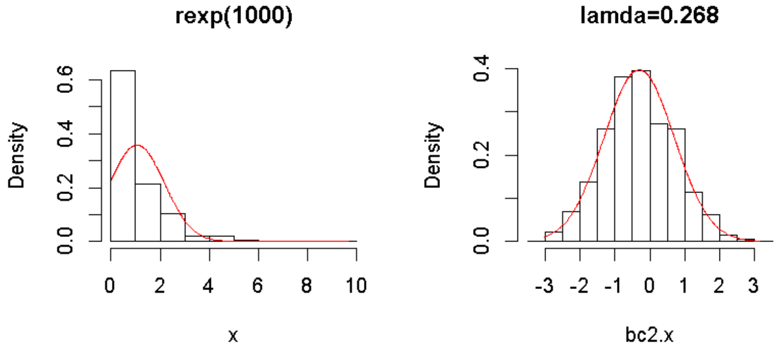
```
x <- rexp(1000)
bc <- function(y, lambda){
    (y^lambda -1)/lambda
}
qqnorm(x); qqline(x, col="red")

bc1.x <- bc(x, 0.1)
qqnorm(bc1.x, main="lambda=0.1")
qqline(bc1.x, col="red")
bc3.x <- bc(x, 0.5)
qqnorm(bc3.x, main="lambda=0.5")
qqline(bc3.x, col="red")

bc2.x <- bc(x, 0.268)
qqnorm(bc2.x, main="lambda=0.268")
qqline(bc2.x, col="red")

hist(x, main="rexp(1000)")
hist(bc2.x, main="lambda=0.268")
```





$$\left(\Phi^{-1}\left(\frac{i-0.5}{n}\right), x_{(i)}\right), \qquad \text{for } i = 1, 2, \ldots, n,$$

Source: Box-Cox Transformations: An Overview, Pengfei Li, Department of Statistics, University of Connecticut, Apr 11, 2005

Modified Box-Cox Transformations

Manly(1971)

$$y(\lambda) = \begin{cases} \frac{e^{\lambda y}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ y, & \text{if } \lambda = 0. \end{cases}$$

Negative y's could be allowed. The transformation was reported to be successful in transform unimodal skewed distribution into normal distribution, but is not quite useful for **bimodal** or **U-shaped distribution**.

John and Draper(1980) “Modulus Transformation”

$$y(\lambda) = \begin{cases} \text{Sign}(y)\frac{(|y|+1)^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \text{Sign}(y)\log(|y|+1), & \text{if } \lambda = 0, \end{cases}$$

$$\text{Sign}(y) = \begin{cases} 1, & \text{if } y \geq 0; \\ -1, & \text{if } y < 0. \end{cases}$$

Bickel and Doksum(1981)

$$y(\lambda) = \frac{|y|^{\lambda}\text{Sign}(y) - 1}{\lambda}, \qquad \text{for } \lambda > 0,$$

Yeo and Johnson(2000)

$$y(\lambda) = \begin{cases} \frac{(y+1)^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0,\ y \geq 0; \\ \log(y+1), & \text{if } \lambda = 0,\ y \geq 0; \\ \frac{(1-y)^{2-\lambda}-1}{\lambda-2}, & \text{if } \lambda \neq 2,\ y < 0; \\ -\log(1-y), & \text{if } \lambda = 2,\ y < 0. \end{cases}$$

Source: Box-Cox Transformations: An Overview, Pengfei Li, Department of Statistics, University of Connecticut, Apr 11, 2005

標準化 (Standardization)

- Standardization: (called z-score): the new variate *z* will have a mean of zero and a variance of one. (also called centering and scaling.)

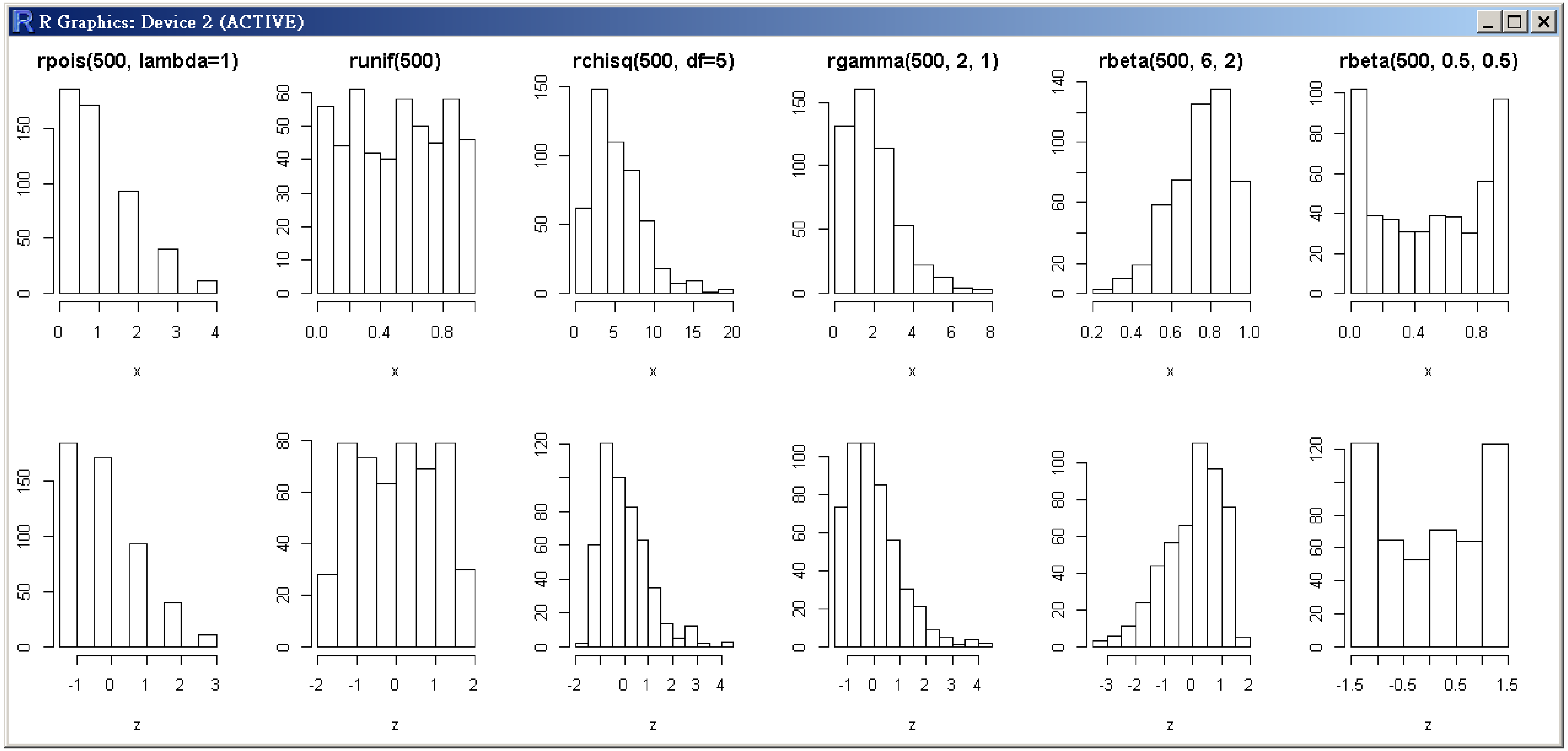
$$z_i = \frac{x_i - \bar{x}}{s}$$

- If the variables are measurements along a **different scale** or if the standard deviations for the variables are different from one another, then one variable might **dominate** the distance (or some other similar calculation) used in the analysis.

- Standardization is useful for comparing variables expressed in different units.

標準化 (Standardization)

Standardization makes no difference to the shape of a distribution.



```
x <- rpois(500, lambda=1)
hist(x, main="rpois(500, lambda=1)"); z <- scale(x); hist(z, main="")
```

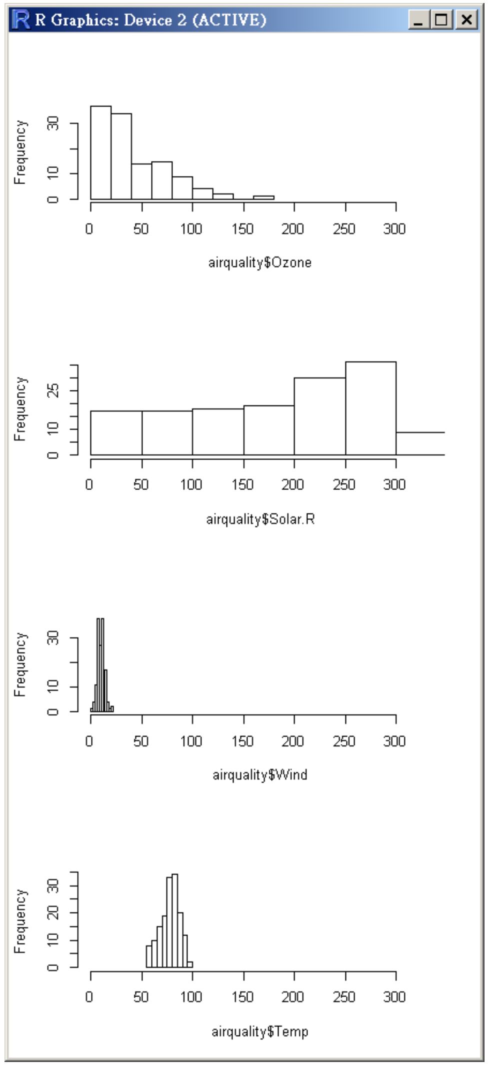
範例: Standardization

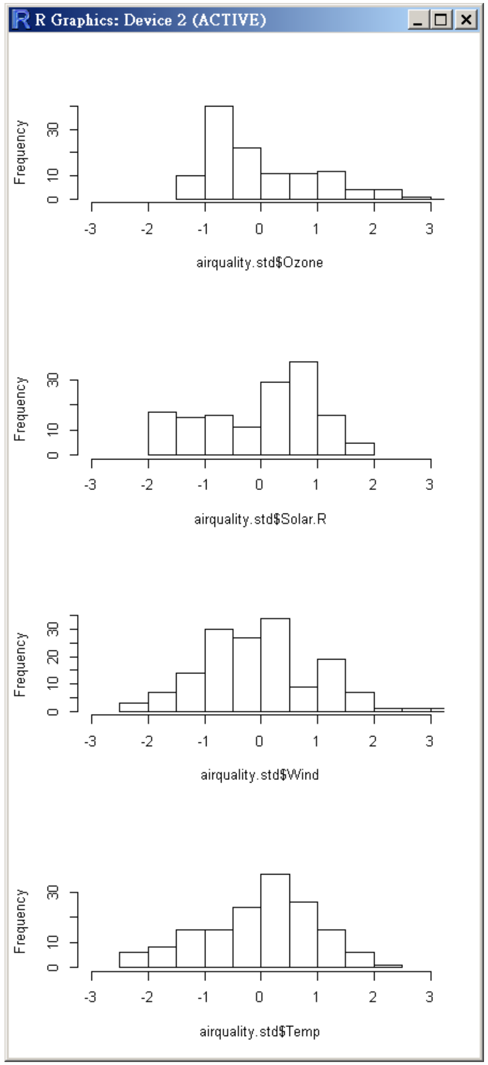
airquality {datasets}

New York Air Quality Measurements: Daily air quality measurements in New York, May to September 1973.
A data frame with 154 observations on 6 variables.

[1] Ozone: Ozone (ppb)
[2] Solar.R: Solar R (lang)
[3] Wind: Wind (mph)
[4] Temp: Temperature (degrees F)
[5] Month: Month (1--12)
[6] Day: Day of month (1--31)

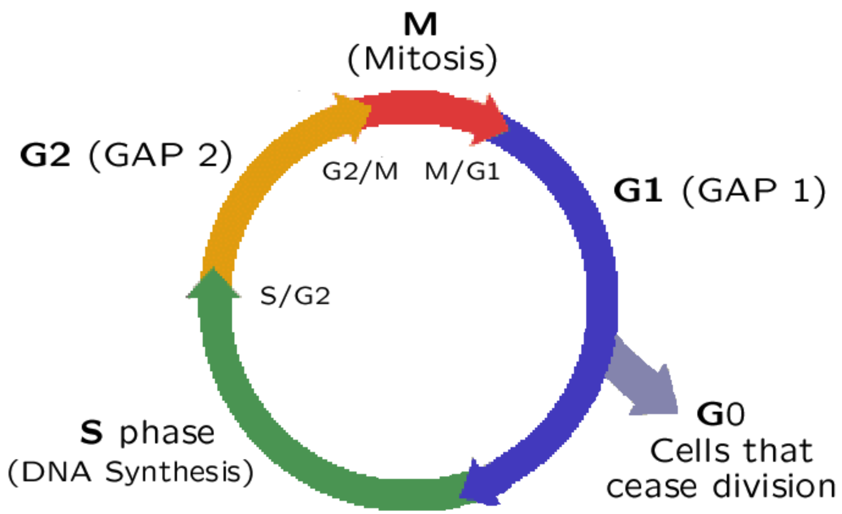
```
> head(airquality )
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
> r <- range(airquality[,1:4], na.rm = T)
> hist(airquality$Ozone , xlim = r)
> hist(airquality$Solar.R, xlim = r)
> hist(airquality$Wind, xlim = r)
> hist(airquality$Temp, xlim = r)
>
> airquality.std <- as.data.frame(
apply(airquality, 2, scale))
> r.std <- c(-3, 3)
> hist(airquality.std$Ozone, xlim = r.std)
> hist(airquality.std$Solar.R, xlim = r.std)
> hist(airquality.std$Wind, xlim = r.std)
> hist(airquality.std$Temp, xlim = r.std)
```

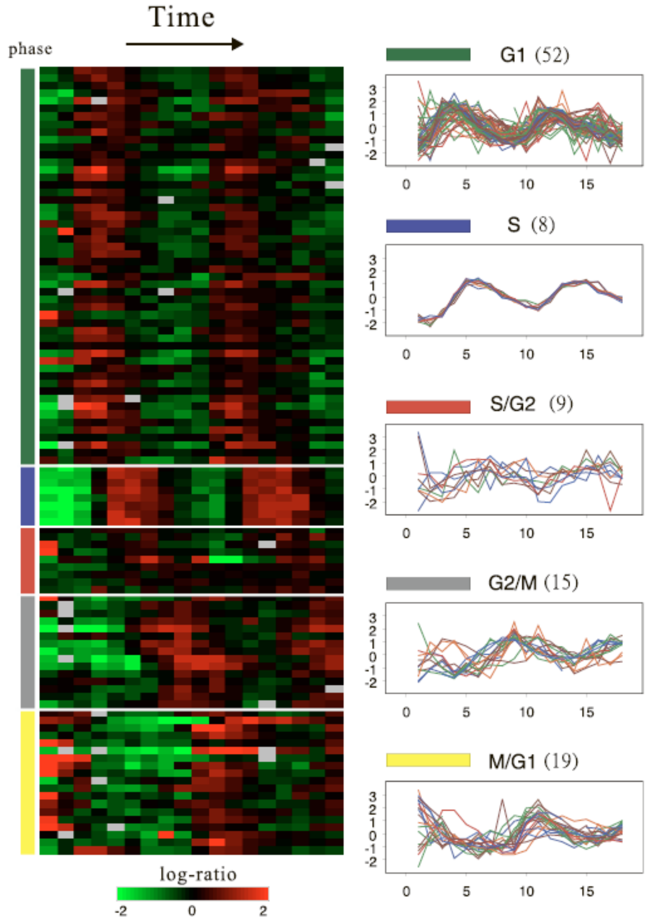




範例: Microarray Data of Yeast Cell Cycle

- Lu and Wu (2010)
  - Time course data: every 7 minutes and totally 18 time points.
  - Known genes: there are 103 cell cycle-regulated genes by traditional method in G1, S, S/G2, G2/M, or M/G1. (Remove NA’ s: 79.)

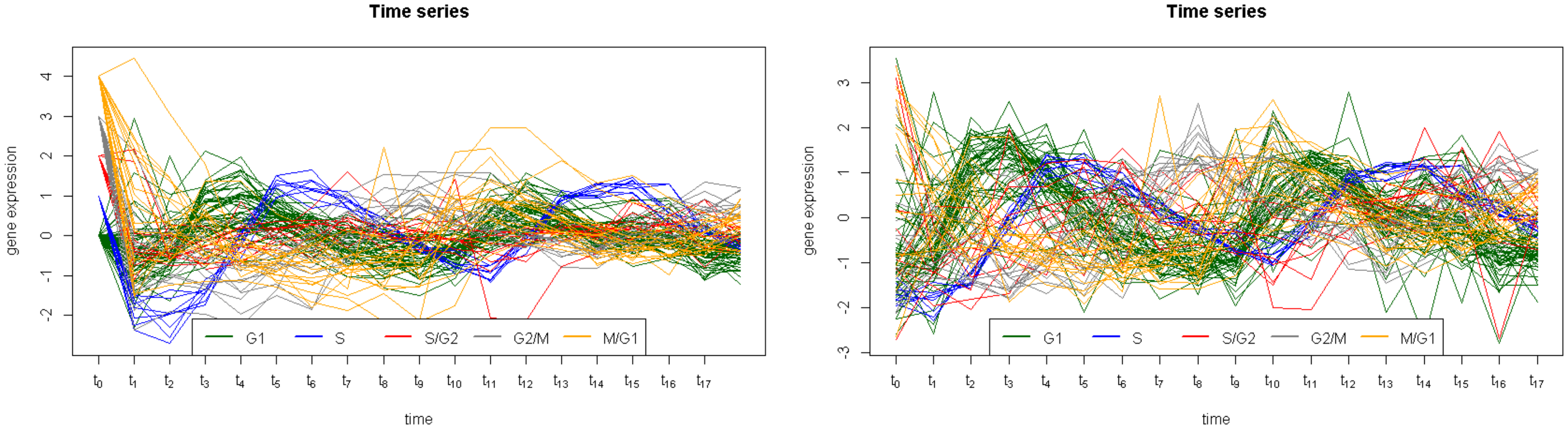




See also: Using R to draw a Heatmap from Microarray Data
http://www2.warwick.ac.uk/fac/sci/moac/people/students/peter_cock/r/heatmap/

Standardization in Time Series Microarray Gene Expression Experiments

```r
cell.raw <- read.table("trad_alpha103.txt", row.names=1, header=T)
head(cell.raw)
cell.xdata <- t(scale(t(cell.raw[,2:19]), center=T, scale=T))
y.C <-  as.integer(cell.raw[,1])
table(y.C)
no.cluster <- length(unique(y.C))
cellcycle.color <- c("darkgreen", "blue", "red", "gray50", "orange")
p <- ncol(cell.raw) -1
ycolors <- cellcycle.color[y.C+1]
my.pch <- c(1:no.cluster)[y.C+1]
phase <- c("G1", "S", "S/G2", "G2/M", "M/G1")
matplot(t(cell.xdata), pch = 1:p, lty=1, type = "l", ylab="gene expression",
            col=ycolors, xlab="time", main="Time series", xaxt="n")
time.label <- parse(text=paste("t[",0:p,"]",sep=""))
axis(1, 1:(p+1), time.label)
legend("bottom", legend=phase, col=cellcycle.color, lty=1, horiz = T, lwd=2)
```



The data map for 103 cell cycle-regulated genes and the plots of time courses for each phase. Each expression profile is normalized as mean equals zero and variance 1.

範例: Crab Data

crabs {MASS}

Morphological Measurements on Leptograpsus Crabs

Description: The crabs data frame has 200 rows and 8 columns, describing 5 morphological measurements on **50 crabs each of two colour forms and both sexes**, of the species Leptograpsus variegatus (紫岩蟹) collected at Fremantle, W. Australia.

This data frame contains the following columns:

sp: species - "B" or "O" for blue or orange.

sex: “M” or “F” for male or female

index: 1:50 within each of the four groups.

FL: carapace frontal lobe (lip) size (mm).

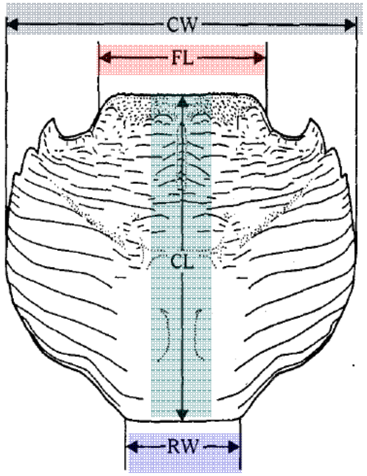
RW: carapace rear width (mm).

CL: carapace length (mm).

CW: carapace width (mm).

BD: body depth (mm).

```
> library(MASS)
> data(crabs)
```

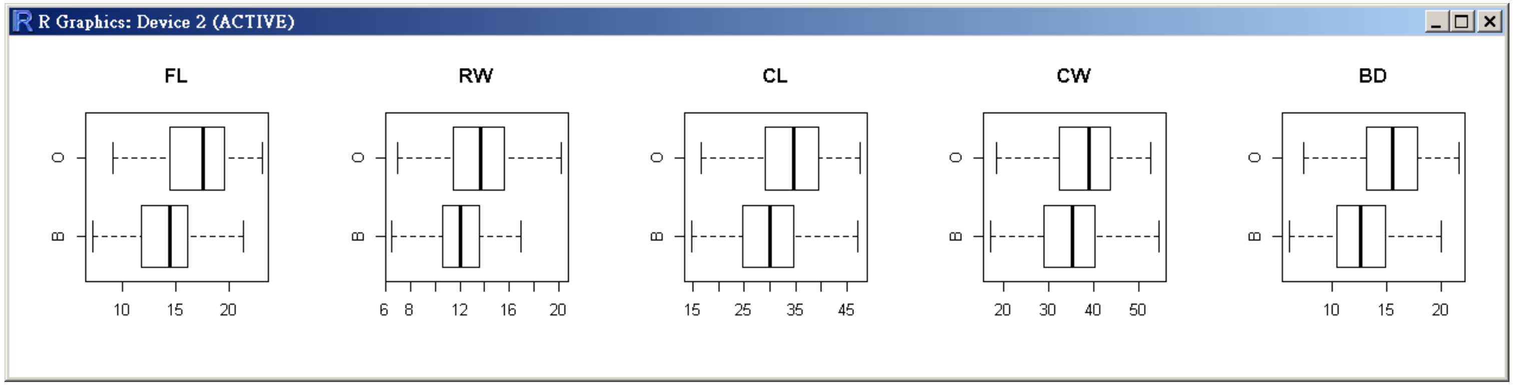


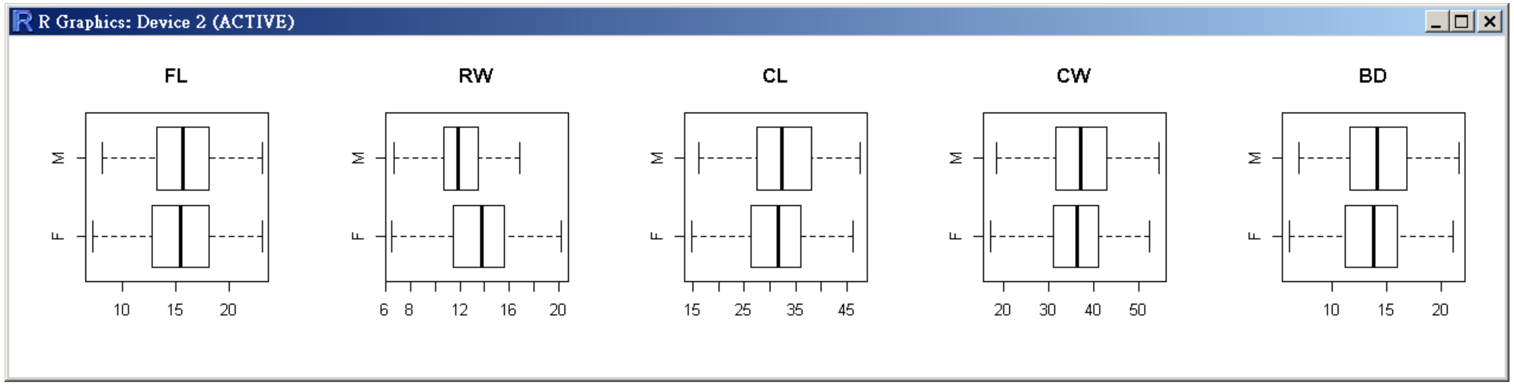
Aust. J. Zool. 1974, 22, 417-25

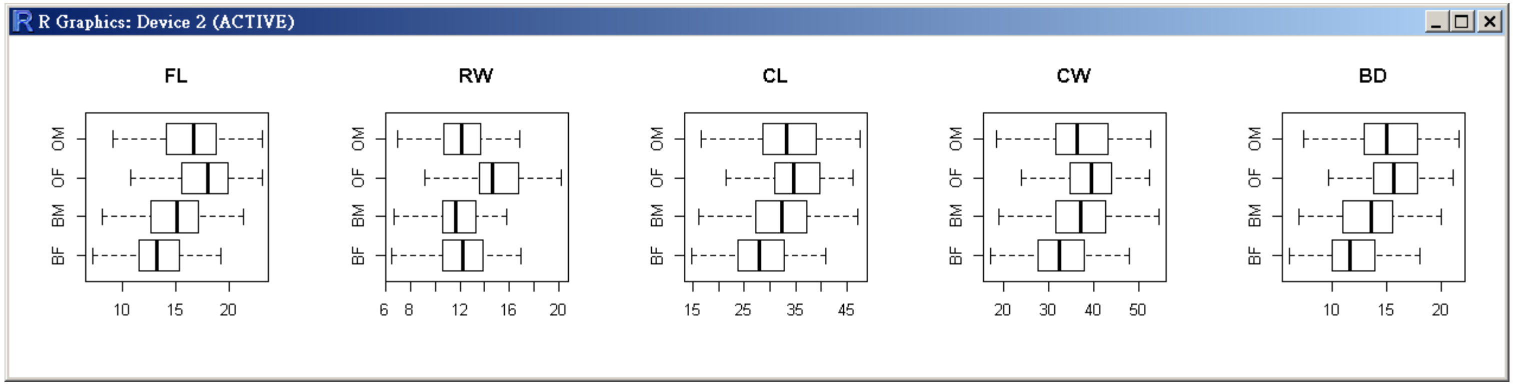
http://www.qm.qld.gov.au/Find+out+about/Animals+of+Queensland/Crustaceans/Common+marine+crustaceans/Crabs/Purple+Swift-footed+Shore+Crab#.VhPWYiurFhs

範例: Crab Data

```
boxplot(crabs$FL~crabs$sp, main="FL", horizontal=T)
```







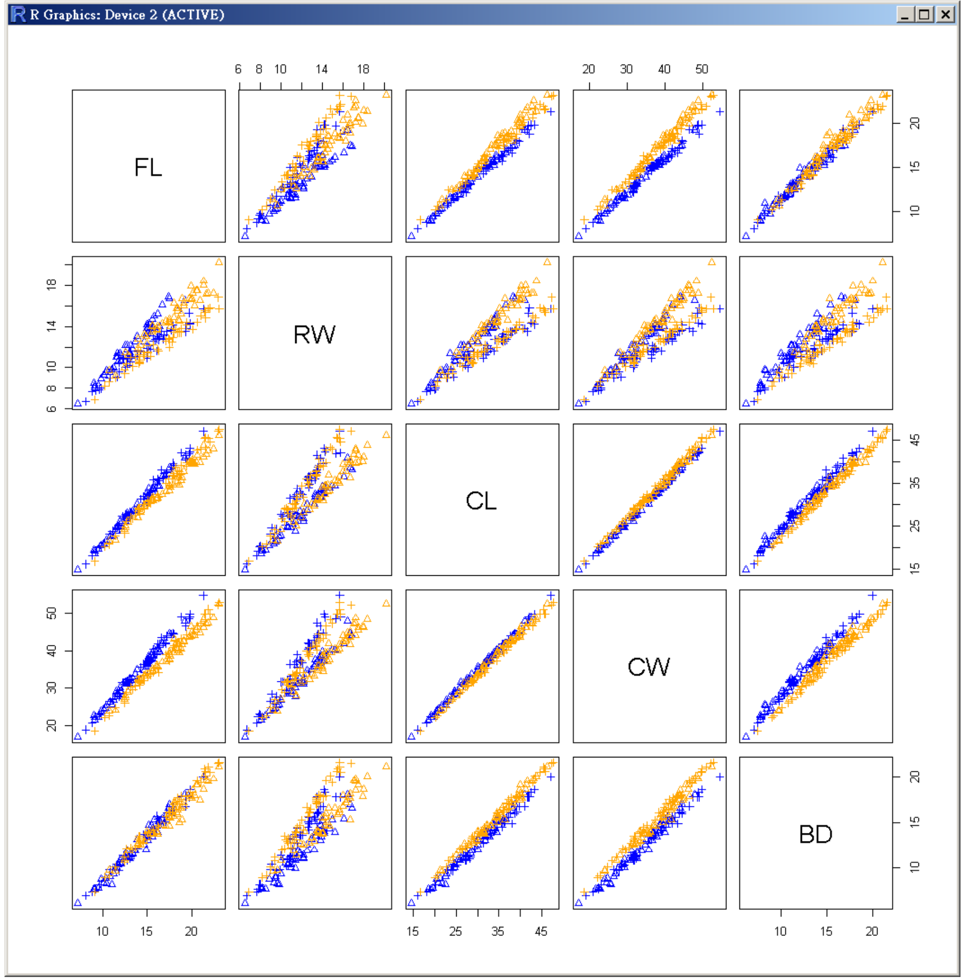
範例: Crab Data

```
# tri: F,  cross: M
pairs(crabs[,4:8],
pch=as.integer(crabs$sex)+1,
col=c("blue","orange")[as.integer(crabs$sp)])
```

- The analysis of ratios of body measurements is deeply ingrained in the taxonomic literature.

- Whether for plants or animals, certain ratios are commonly indicated in identification keys, diagnoses, and descriptions.

(Hannes Baur and Christoph Leuenberger, Analysis of Ratios in Multivariate Morphometry, Systematic Biology 60(6), 813-825.)



範例: Crab Data

- The use of **ratios of measurements** (i.e., of body proportions), has a long tradition and is deeply ingrained in morphometric taxonomy.

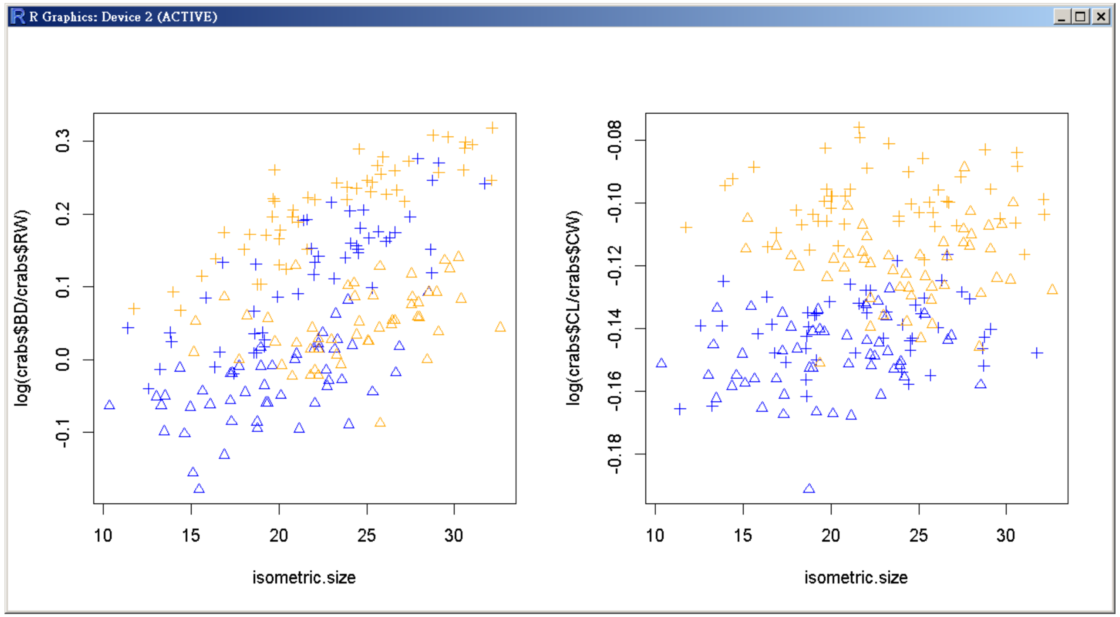
Three size vectors have been commonly proposed in the literature:

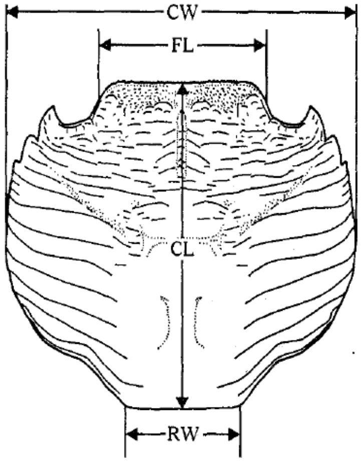
(a) isometric size

(the arithmetic mean of x),

(b) allometric size,

(c) shape-uncorrelated size.

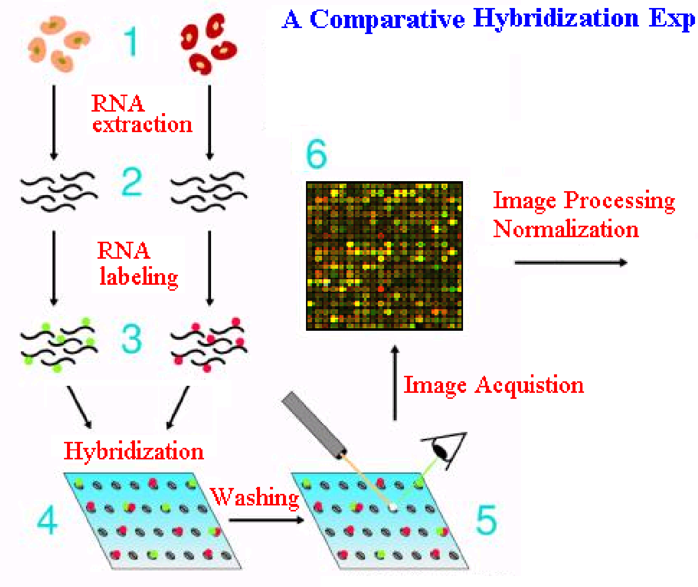


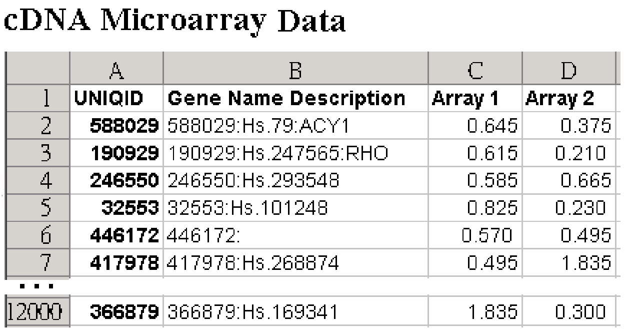


```
par(mfrow=c(1,2))
mp <- as.integer(crabs$sex)+1
mc <- c("blue","orange")[as.integer(crabs$sp)]
isometric.size <- apply(crabs[,4:8], 1, mean)
plot(isometric.size,  log(crabs$BD/crabs$RW), pch=mp, col=mc)
plot(isometric.size, log(crabs$CL/crabs$CW), pch=mp, col=mc)
```

範例: cDNA Microarray Gene Expression Data

微陣列資料統計分析 Statistical Microarray Data Analysis
http://www.hmwu.idv.tw/index.php/mada




cDNA Microarray Data

	A	B	C	D
1	UNIQID	Gene Name Description	Array 1	Array 2
2	**588029**	588029:Hs.79:ACY1	0.645	0.375
3	**190929**	190929:Hs.247565:RHO	0.615	0.210
4	**246550**	246550:Hs.293548	0.585	0.665
5	**32553**	32553:Hs.101248	0.825	0.230
6	**446172**	446172:	0.570	0.495
7	**417978**	417978:Hs.268874	0.495	1.835
...				
12000	**366879**	366879:Hs.169341	1.835	0.300

Matrix of genes (rows) and samples (columns)

Why Normalization?

Non-biological factor can contribute to the variability of data, in order to reliably compare data from multiple probe arrays, differences of non-biological origin must be minimized.

(Remove the systematic bias in the data).

- Within-Array Normalization
- Between-Array Normalization
- Paired-slides Normalization
- ...

要使用哪一種資料轉換方式?

- Use a transformation that other researchers **commonly use in your field**.

- Guidance for how data should be transformed, or whether a transformation should be applied at all, should come from the particular statistical analysis to be performed.

- The main criterions in choosing a transformation:
    - what works with the data?
    - what makes physical (biological, economic, whatever) sense.

- If you have a **large** number of observations, compare the effects of different transformations on the normality and the homoscedasticity of the variable.

http://www.biostathandbook.com/transformation.html