R-Ladies Taipei

R-basic Lesson 9

Descriptive Statistics
敘述統計

黃舒瑜 , 蔡旻均, Ricci Chen

2017.9.25 @Dcard

Outlines

- Correlation 相關性介紹
  - So-called Correlation 統計所謂的相關性
  - Independent or Not , that is not Correlated

    相關係數的觀念釐清(獨立/相依、因果)
- Time Measurement in Different Data Types
  時間變數的各種資料型態
- A Simple Time Series Analysis

  時間序列簡單看

R-Ladies Taipei

Correlation相關性介紹

R-Ladies Taipei

相關性

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

- 如何描述兩變數間的相關性？
- 當其中一個變數上升時，另一個變數會傾向上升還是下降呢？

共變異數Covariance

- 衡量兩變數變動的方向及其程度
- 公式：

$$\mathrm{Cov}(\mathrm{X},\mathrm{Y}) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)]$$

```
> cov(iris[1:4])
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.6856935	-0.0424340	1.2743154	0.5162707
Sepal.Width	-0.0424340	0.1899794	-0.3296564	-0.1216394
Petal.Length	1.2743154	-0.3296564	3.1162779	1.2956094
Petal.Width	0.5162707	-0.1216394	1.2956094	0.5810063

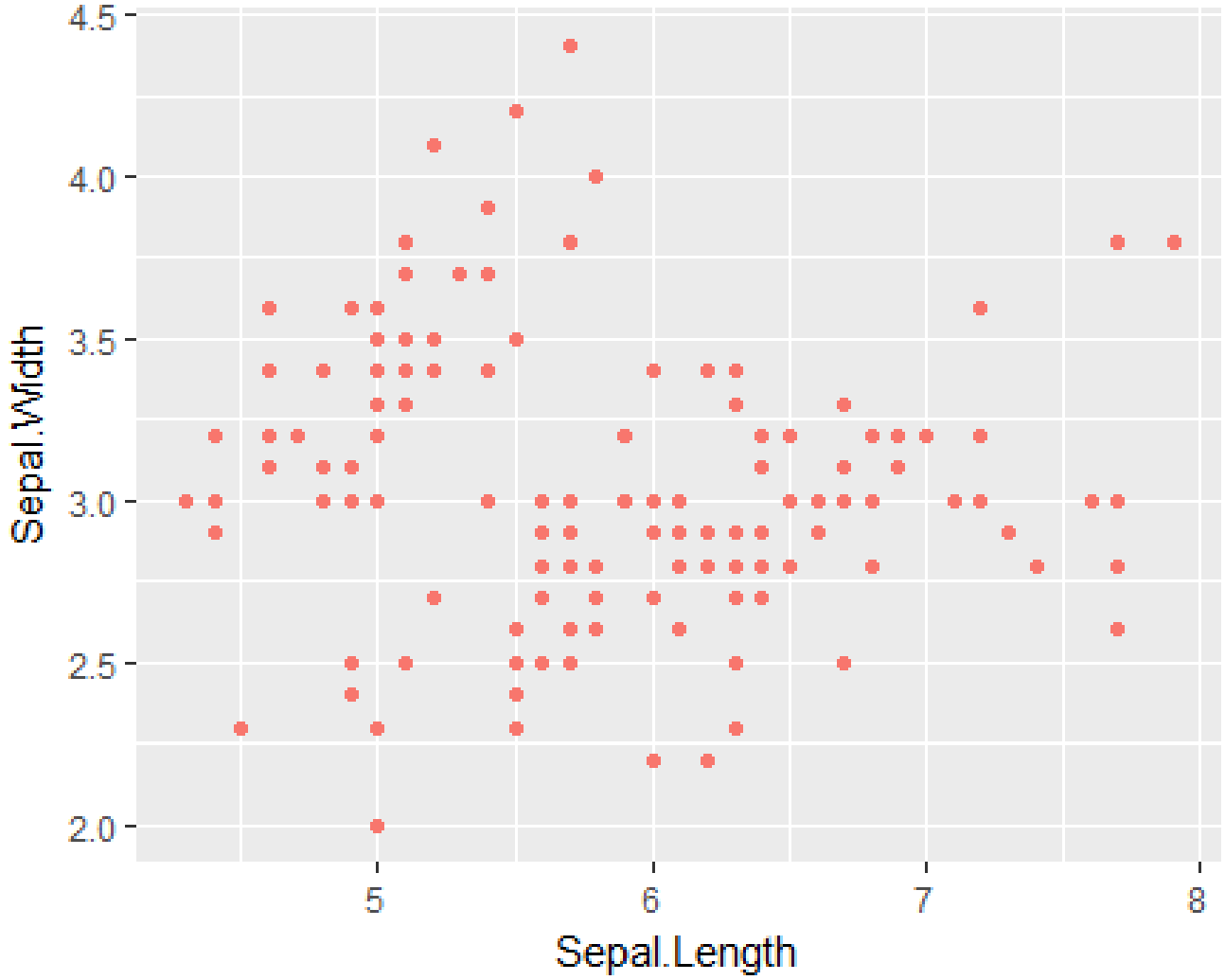
當兩個變數實為同一變數時，共變異數等同於變異數

```
> cov(iris$Sepal.Length, iris$Sepal.Width)
```

[1] -0.042434 算大還是小？

R-Ladies Taipei

Sepal.Length和Sepal.Width的散布圖



相關係數Correlation Coefficient

- 去除自身變異
- 公式：

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = E\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)$$

$Cov(X,Y)$為共變異數，σ_X、σ_Y為標準差

- 相關係數的值介於-1到1之間

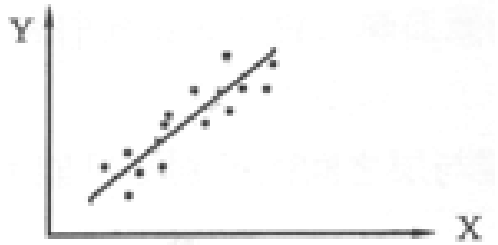
```
> cor(iris[1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

當兩個變數實為同一變數時，相關係數等於1

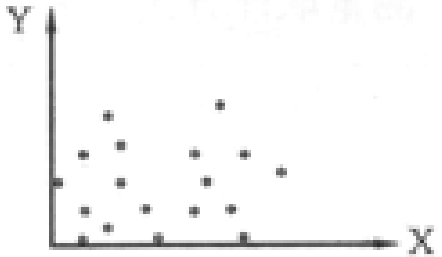
相關係數(續)

- 相關係數的正負號代表相關的方向

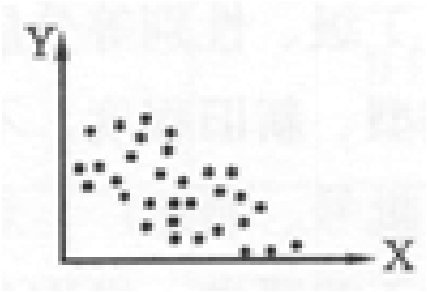
$\rho > 0$ **正相關**



$\rho = 0$ **不相關**



$\rho < 0$ **負相關**



- 相關係數的大小代表相關的程度
  - $0 \leq |\rho| < 0.25$：弱相關
  - $0.25 \leq |\rho| < 0.5$：中度弱相關
  - $0.5 \leq |\rho| < 0.75$：中度強相關
  - $0.75 \leq |\rho| \leq 1$：強相關

R-Ladies Taipei

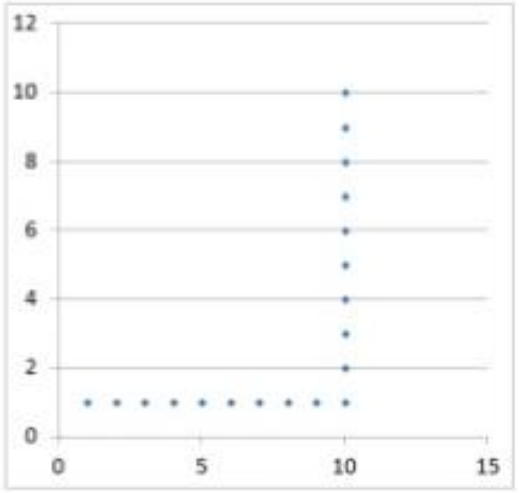
Independent or Not, that is not Correlated

相關係數的觀念釐清

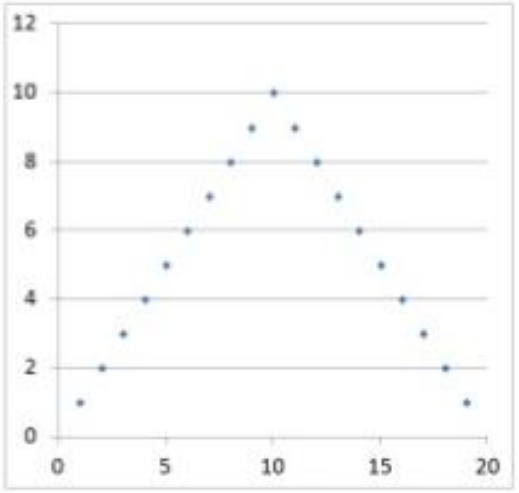
R-Ladies Taipei

相關性 VS 獨立性

- 相關係數衡量的是兩變數間的線性關係
- 相關係數為零代表兩變數無線性相關，不代表兩變數獨立
- 獨立的兩個變數相關係數為零



$(\rho = 0.5973)$



$(\rho = 0)$

R-Ladies Taipei

相關性 VS 因果性

- 相關性不隱含因果性
- 高度相關無法代表兩變數具有因果關係

睡眠7小時 可防老年癡呆症

台灣醒報
劉運 2012年7月18日 下午1:10

留言

【台灣醒報記者劉運綜合報導】科學家指出，將每天將睡眠時間限制在7小時的老人，可以減緩頭腦老化速度2年，而睡眠時間超過或低於7小時都會帶來反效果。美國的研究人員發現，每天睡7小時的年老婦女比起睡9小時的人，可以更集中注意力且記憶力較好。

過去的研究顯示，睡眠超過7小時可能會引發體重上升、心臟相關疾病、或糖尿病。美國最新的研究是第一個連結睡眠時間與大腦專注力的研究，研究人員觀察15,000位70歲以上女性達5年之久。研究人員發現，睡7小時的人比起睡超過或小於7小時的人，在記憶力、集中力及專注持久力上，表現都比較好。

美國波士頓楊百翰女性醫院的研究員伊莉莎白滴佛兒指出，這個研究指出睡眠與心智功能及老年癡呆症的關聯性。她說：「這個研究結果意義重大，因為人們可能在未來找到特別的睡眠方式，來減低患得心智功能喪失及老年癡呆症的風險。」

老年癡呆症協會發言人指出，雖然睡眠與心智健康的關聯性已經在研究中確定，但是還需要進一步的研究，才能確定是睡眠影響心智功能，還是心智功能影響睡眠。發言人指出，良好的睡眠品質、均衡的飲食、定期

錯把相關當因果

R-Ladies Taipei

Time measurement in different data type

時間變數的各種資料型態

時間的組成

時區 → 年月日 → 時分秒

- 格林威治標準時間；當地時間。
- 季節；星期一~日；以年度為準的日數。
- 24小時制；標註上午/下午。

R-Ladies Taipei

時間輸入

時區 → 年月日 → 時分秒

字串；
Sys.Date()；
以1970/1/1起算的正負整數。

字串；
Sys.time()；
以1970/1/1/00:00:00起算的正負整數。

R-Ladies Taipei

系統時間

```
>
> today <- Sys.Date()
> today
[1] "2017-09-25"
> as.numeric(today)
[1] 17434
>
>
> now <- Sys.time()
> now
[1] "2017-09-25 14:16:27 CST"
> as.numeric(now)
[1] 1506320187
>
```

R-Ladies Taipei

處理時間

as.date()

chron()

as.POSIXct()
as.POSIXlt()
ISOdate()

R-Ladies Taipei

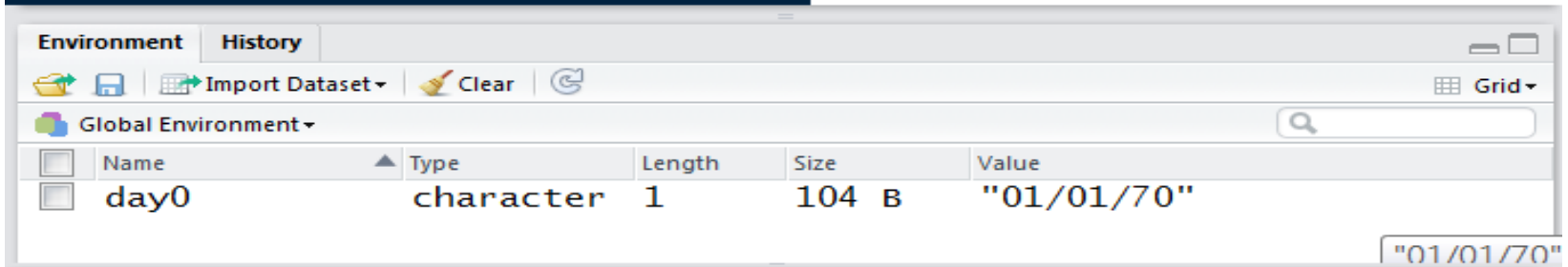
format :指定格式

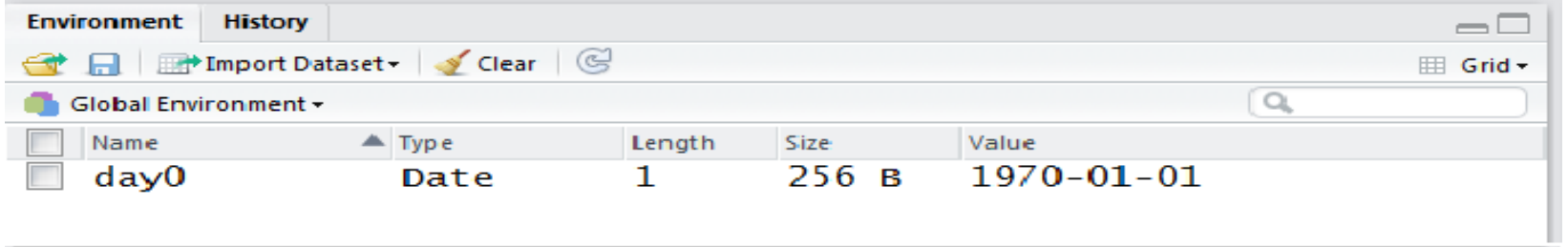
Code	Value
%d	Day of the month (decimal number)
%m	Month (decimal number)
%b	Month (abbreviated)
%B	Month (full name)
%y	Year (2 digit)
%Y	Year (4 digit)

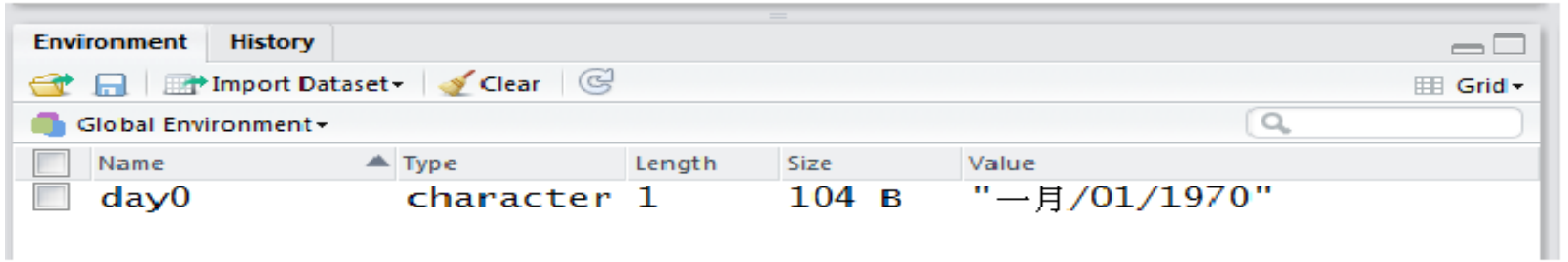
R-Ladies Taipei

format() :指定格式

```
>
> day0 <- c("01/01/70")
>
> day0 <- as.Date(day0 , "%m/%d/%y")
>
> day0 <- format(day0,"%b/%d/%Y")
>
```







R-Ladies Taipei

ISOdate()

- 輸入標準時間字串
- 輸出POSIXct

```
>
> ISOdatetime(2017,09,25,19,20,00,tz="GMT")
[1] "2017-09-25 19:20:00 GMT"
> ISOdatetime(2017,09,25,19,20,00,tz="")
[1] "2017-09-25 19:20:00 CST"
>
```

R-Ladies Taipei

時間的各種資料型態

1.Nominal(名目型態):

```
> bdays = c( tukey=as.Date( '2017-09-18' ) ,
             fisher=as.Date( '2017-09-19' ) ,
             cramer=as.Date( '2017-09-20' ),
             kendall=as.Date( '2017-09-21' ) )

> weekdays(bdays)

tukey      fisher     cramer          kendall
"星期一" "星期二" "星期三" "星期四"
```

R-Ladies Taipei

時間的各種資料型態

2.Ordinal(排序型態):

```
> seq(as.Date('1976-7-4'),by='days',length=10)


[1] "1976-07-04" "1976-07-05" "1976-07-06" "1976-07-07"
[5] "1976-07-08" "1976-07-09" "1976-07-10" "1976-07-11"
[9] "1976-07-12" "1976-07-13"
```

時間的各種資料型態

時間還可以做以下的變數種類

3.Interval (區間型態)

4.Ratio (比率型態)

在這裡先不做討論

R-Ladies Taipei

讀取檔案時間的指令

```
> file.info(dir())[,"mtime",drop=FALSE]

                                mtime
desktop.ini 2017-09-17 08:09:02
```

時間顯示轉換的指令

```
> file_time <- file.info(dir())[, mtime",drop=FALSE]

> file_time

                                mtime
  desktop.ini 2017-09-17 08:09:02

> format(file_time, format="%x %X")

                                  mtime
  desktop.ini 2017/9/17 下午 08:09:02
```

R-Ladies Taipei

A Simple Time Series Analysis

時間序列簡單看

R-Ladies Taipei

時間序列是甚麼？

R-Ladies Taipei

時間序列是甚麼？

包含至少一組以上
以時間為序的資料。

時間	搜尋熱門度
2004/01	12
2004/02	11
2004/03	9
2004/04	8
2004/05	8
2004/06	9
2004/07	9
2004/08	12
2004/09	9
2004/10	8
2004/11	8
2004/12	8
2005/01	7
2005/02	6

R-Ladies Taipei

時間資料分析

1. 和時間相關→ 季節趨勢分解
   (Seasonal Trend Decomposition)
2. 和歷史值相關→ 時間序列分析*
   (Time Series Analysis)
3. 和時間、歷史值皆相關→
   (Panel Analysis)

R-Ladies Taipei

時間資料分析

1. 和時間相關→ 季節趨勢分解
   (Seasonal Trend Decomposition)
2. 和歷史值相關→ 時間序列分析
   (Time Series Analysis)
3. 和時間、歷史值皆相關→
   (Panel Analysis)

R-Ladies Taipei

季節趨勢分解

將時間資料分解成三個部分

- 趨勢 (trend)
- 周期變化 (seasonality)
- 剩餘部分 (remainder)

在其他條件不變之下，如果把**趨勢**和**周期變化**處理得夠乾淨，**剩餘部分**應該是無法解釋的隨機變數。

R-Ladies Taipei

作一個簡單的季節性資料

- 時間
- 趨勢
- 周期變化
- 剩餘部分=隨機變數

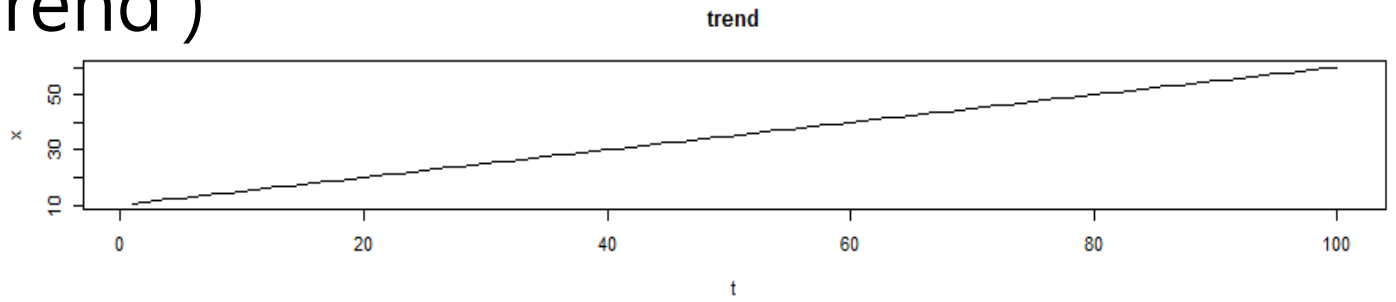
```
// make a simple time data

t=seq(1,100)
x=t/2 +10
s=rep(c(1,2,-2,-1),25)
e=rnorm(100, mean = 0, sd = 1)
y=x+s+e
```

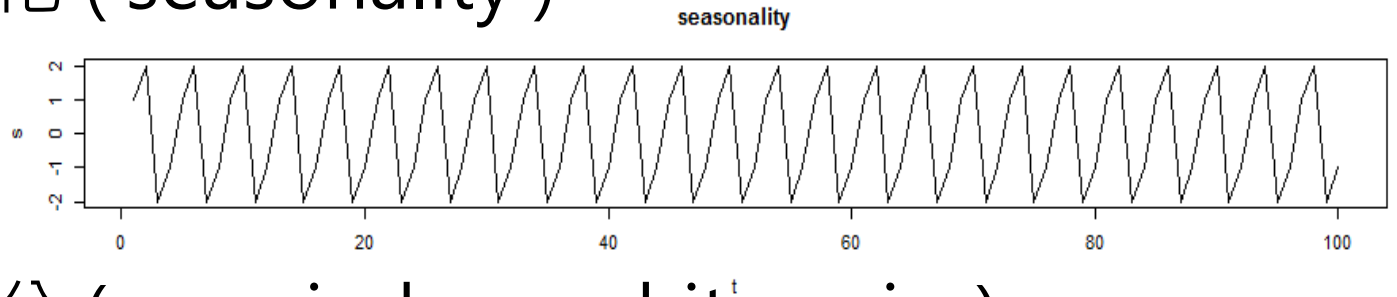
R-Ladies Taipei

由三個部分合成

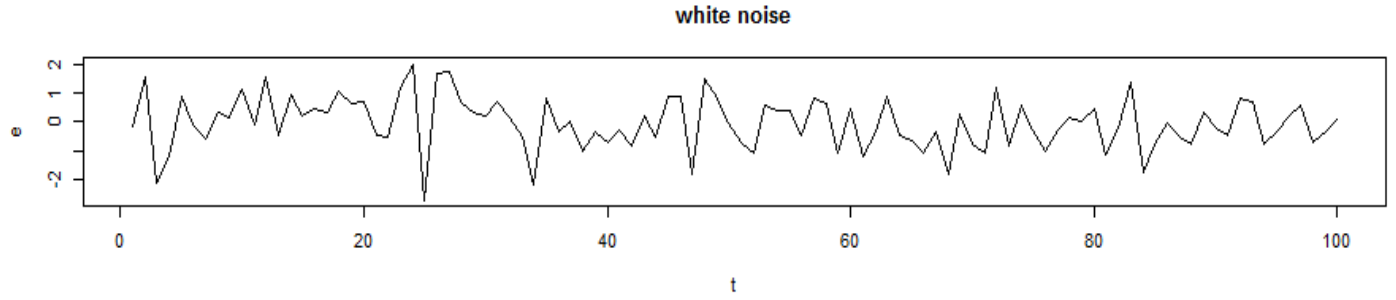
趨勢 (trend)



周期變化 (seasonality)

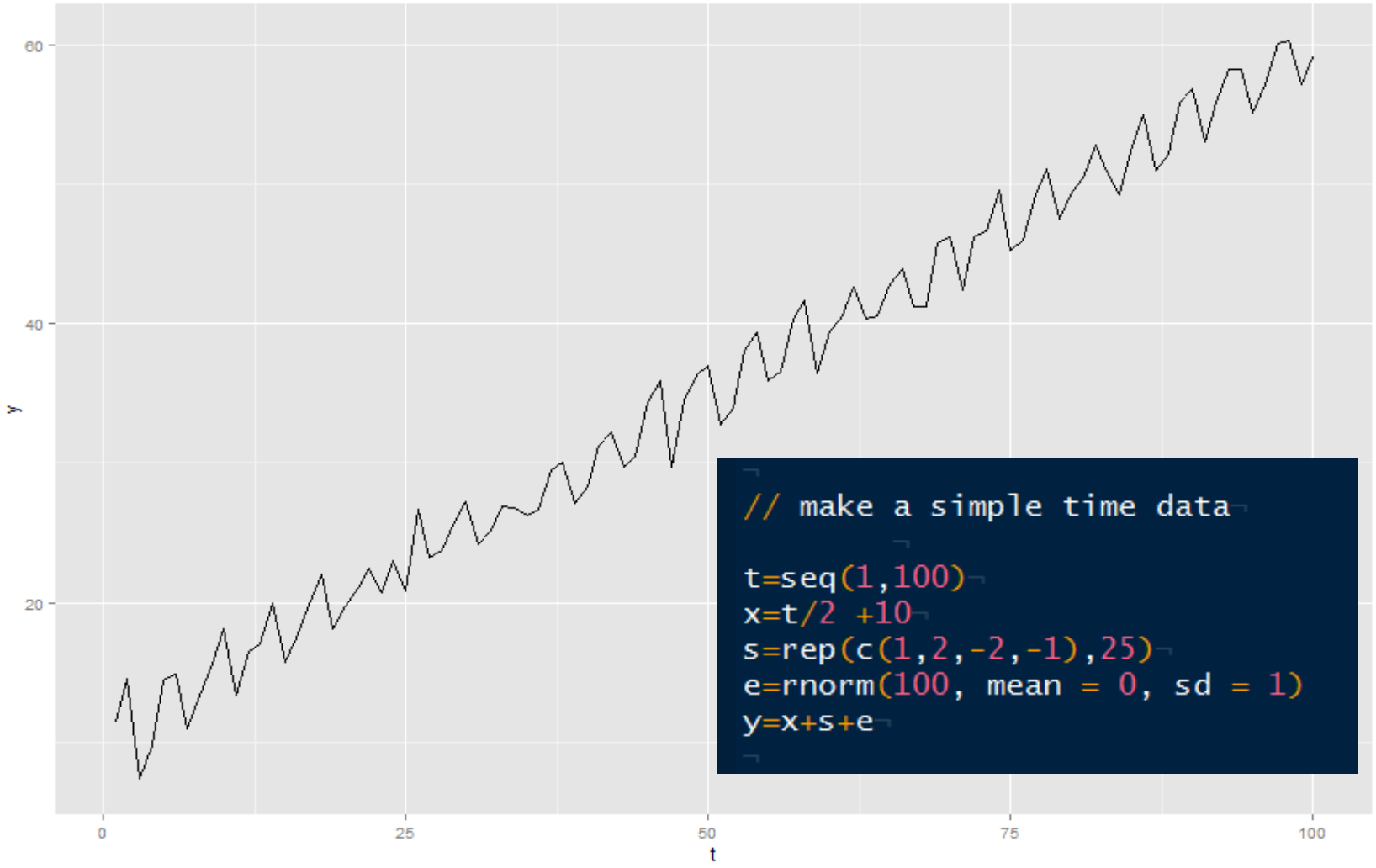


剩餘部分 (remainder = white noise)



R-Ladies Taipei

季節性資料

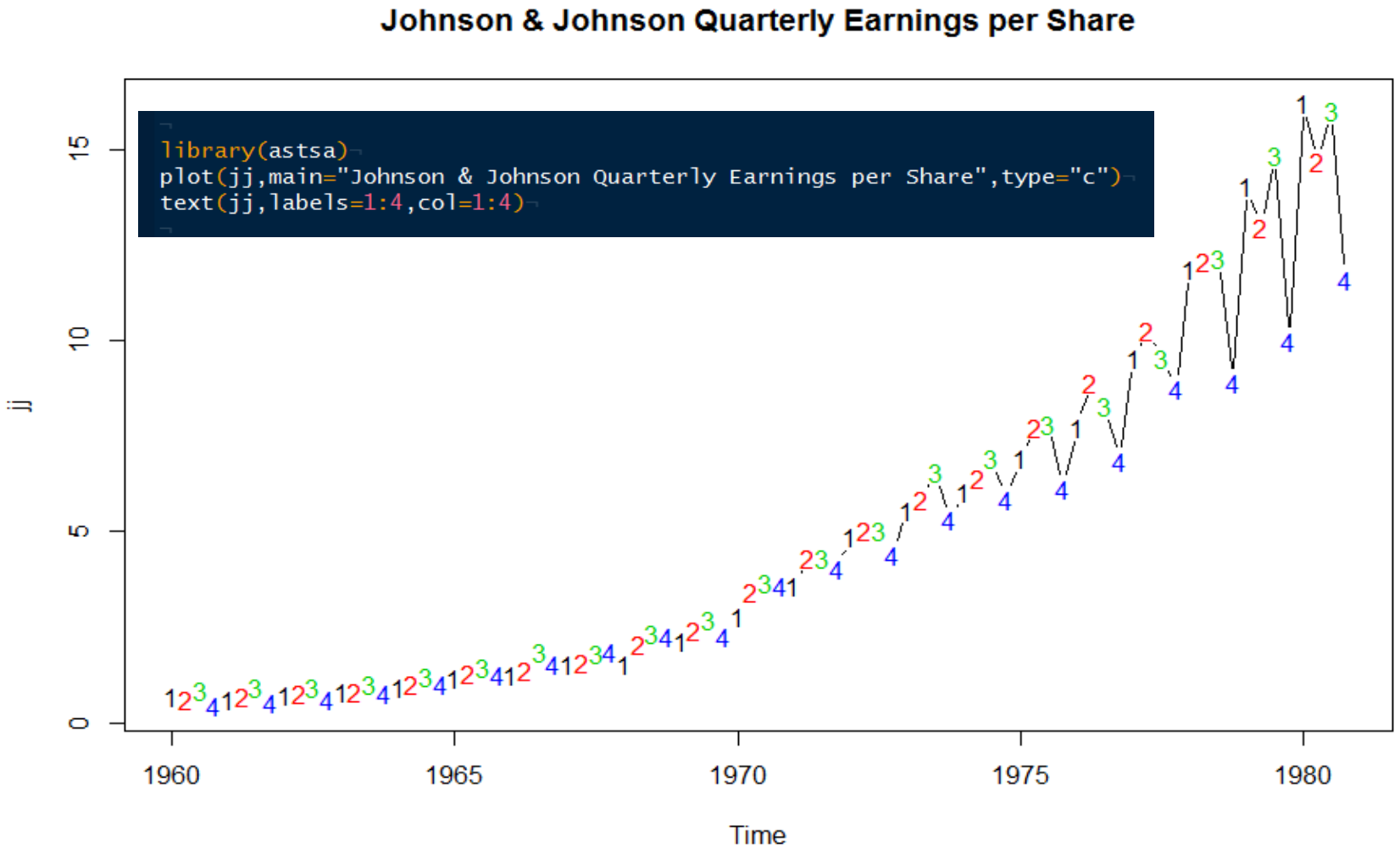



```
// make a simple time data

t=seq(1,100)
x=t/2 +10
s=rep(c(1,2,-2,-1),25)
e=rnorm(100, mean = 0, sd = 1)
y=x+s+e
```

R-Ladies Taipei

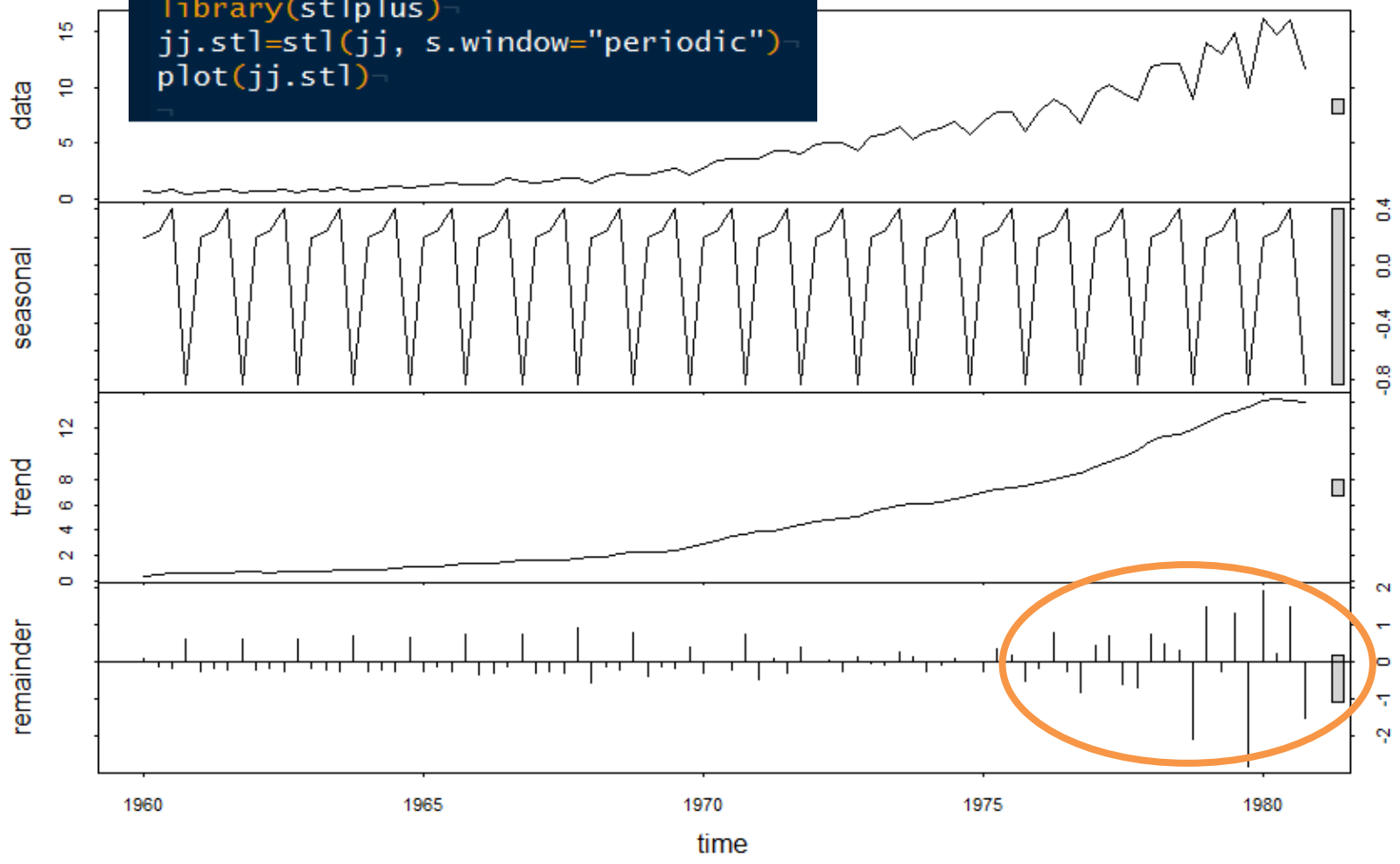
舉例：嬌生公司季報EPS



R-Ladies Taipei

Package : "stlplus"

```
library(stlplus)
jj.stl=stl(jj, s.window="periodic")
plot(jj.stl)
```



R-Ladies Taipei

Reference & Further Readings

- Wush筆記 R: DateTime格式的心得
  http://wush978.github.io/blog/2012/02/29/rdatetime/#localtime
- Dates and Times in R
  https://www.stat.berkeley.edu/~s133/dates.html
- Data Camp : Introduction to Time Series
  https://www.datacamp.com/courses/introduction-to-time-series-analysis

R-Ladies Taipei

Introduction of ARMA

時間序列：自回歸移動平均模型簡介

R-Ladies Taipei

時間序列分析

由線性迴歸模型概念延伸，

$$Y = \beta X + \varepsilon$$

認為**現在**受到兩種歷史因素影響：

- 前期的資料(X是過去的Y)
- 前期的隨機因素(white noise)

R-Ladies Taipei

時間序列分析

認為**現在**受到兩種歷史因素影響：

$$Y_t = \beta X_t + \varepsilon$$

- 前期的自變數→ 自回歸 Auto Regression

$$X_t = \phi X_{t-1} + \varepsilon_t$$

- 前期的隨機因素→移動平均 Moving Average

$$\varepsilon_t = W_t + \theta W_{t-1}$$

合稱為ARMA

$$X_t = \phi X_{t-1} + W_t + \theta W_{t-1}$$

R-Ladies Taipei

時間序列分析

單一變數ARMA模型一般化：

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + W_t + \sum_{i=1}^{q} \theta_i W_{t-i}$$

1. 期數不只前一期，可能是不連續的好幾期。
2. AR期數(p)和MA期數(q)理論上無關。
3. W_t 必須符合 white noise要求：
   - $E(W_t)=0$, $var(W_t)$ =固定常數 , for all t;
   - $cov(W_t\ W_{t-k})=cov(W_{t-j}\ W_{t-k-j})=0$, for all j,k, $j \neq k$

R-Ladies Taipei

ARMA模型判斷步驟

1. 以自我相關函數ACF和偏自我相關PACF來判斷AR和MA的期數（p,q）；
2. 以簡單線性迴歸估計的係數顯著性，作為刪除變數的依據；
3. 以LM或Q統計量檢定殘差含有ARMA型態資料；
4. 以JB統計量檢定殘差是否符合常態性；
5. 若有多組(p,q)則以AIC , BIC(=SBC)準則擇一