




8. 探索式資料分析 Exploratory Data Analysis

鄒慶士 (Ching-Shih Tsou)
台北商業技術學院資訊與決策科學所
E-mail : cstsou@mail.ntcb.edu.tw



探索式的資料分析

- 檢視資料集，以決定何統計推論方法較為適合
- 單變量資料
 - 資料分佈是對稱或偏斜？
 - 資料是否近似常態？
 - 資料是長尾或短尾分佈？
 - 資料分佈是單峰、雙峰或多峰？
- 主要使用的工具是電腦繪圖

© Vince Tsou, IDS, NTCB 100年度教育部補助技專校院建立特色典範計畫 2



使用分析工具

- 長條圖：適用類別型資料
- 直方圖、點圖與枝葉圖：看出數值分佈的形狀
- 盒鬚圖：數值分佈的彙總特性，尤其是比較分佈與辨識長尾與短尾分佈
- 常態機率圖：資料是否近似常態？



Example : Homedata (1)



- `str(homedata)`
y1970: a numeric vector; y2000: a numeric vector
包含(y1970)和(y2000)的數值向量資料。
- `attach(homedata)`
可省略homedata\$,直接存取homedata的變數。
- `hist(y1970);hist(y2000)`
做出(y1970)和(y2000)的次數直方圖。

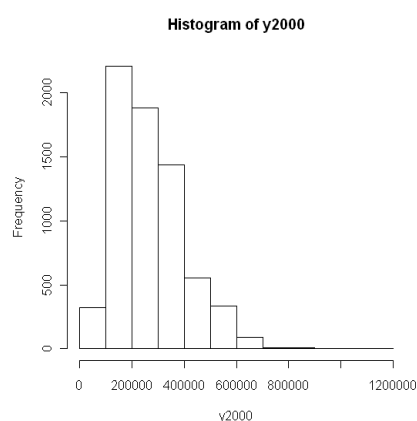
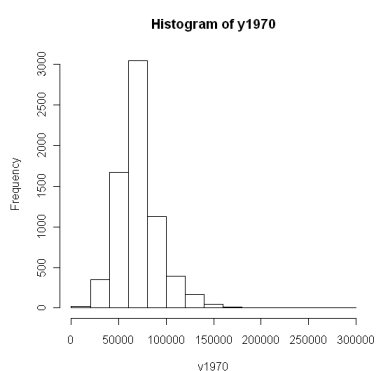
© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建
立特色典範計畫

5



Example : Homedata (1)



- `detach(homedata)`
是attach的解除指令。

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建
立特色典範計畫

6



Example : Homedata (2)

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建
立特色典範計畫

7



- UsingR套件中的simple.eda函數
 - 一次繪出次數直方圖、盒鬚圖(加 ticks)與常態QQ圖
 - `simple.eda(y1970)` ; `simple.eda(y2000)`

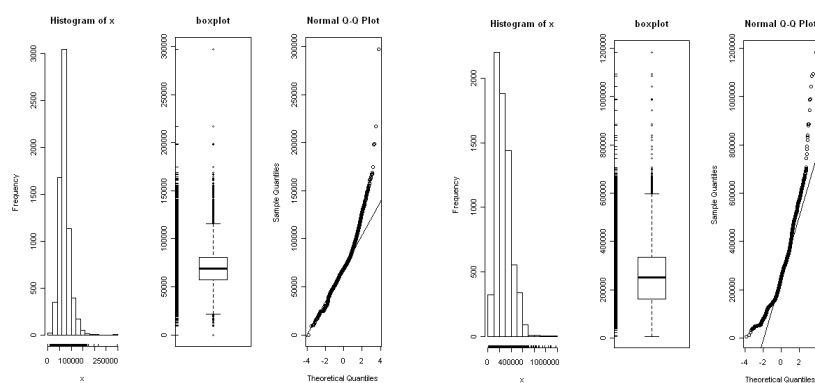
© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建
立特色典範計畫

8



Example : Homedata (2)



- 比較：兩者均非常態且為偏態與厚尾，考慮中位數與資料轉換。

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建立特色典範計畫

9



Example : CEO salaries

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建立特色典範計畫

10



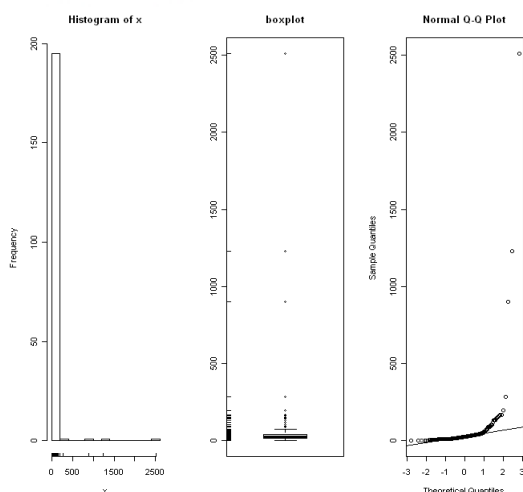
Example : CEO salaries (1)

- 2000年199位美國CEOs薪水
- `simple.eda(exec.pay)`
- 極度右偏的分佈

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建
立特色典範計畫

11



© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建
立特色典範計畫

12



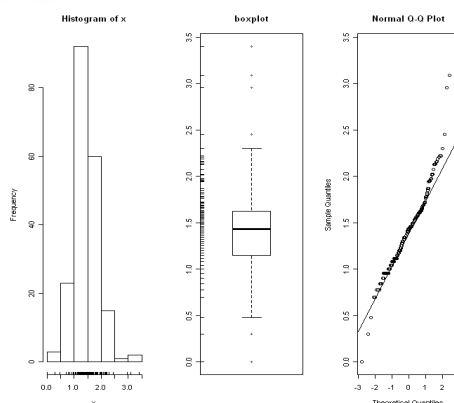
Example : CEO salaries (2)

- 對資料作對數函數轉換
- $\log.\text{exec.pay} = \log(\text{exec.pay}[\text{exec.pay} > 0]) / \log(10)$
 - \log : 以無理數 $e=2.71828..$ 為底
 - \log_{10} 或 \log_2 : 以 10 或 2 為底
 - 一般形式： $\log(x, \text{base})$
 - $\log_{10}(\text{exec.pay}[\text{exec.pay} > 0]) = \log(\text{exec.pay}[\text{exec.pay} > 0]) / \log(10)$
- `simple.eda(log.exec.pay)`

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建立特色典範計畫

13



- 非常對稱
- 資料取對數後呈常態分佈，顯示原資料服從何分配？
- 抗雜訊衡量 or 分析前作資料轉換

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建立特色典範計畫

14



Example : Taxi time at EWR Newark機場飛機滑行時間

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建
立特色典範計畫

15



Example : Taxi time at EWR (1)

□ 用head()看前六筆資料，如何看最後幾筆資料呢？

```
> head(ewr)
  Year Month  AA  CO  DL   HP  NW   TW  UA  US inorout
1 2000   Nov  8.6  8.3  8.6 10.4 8.1   9.1 8.4 7.6      in
2 2000   Oct  8.5  8.0  8.4 11.2 8.2   8.5 8.5 7.8      in
3 2000   Sep  8.1  8.5  8.4 10.2 8.3   8.6 8.2 7.6      in
4 2000   Aug  8.9  9.1  9.2 14.5 9.0  10.3 9.2 8.7      in
5 2000   Jul  8.3  8.9  8.2 11.5 8.8   9.1 9.2 8.2      in
6 2000   Jun  8.8  9.0  8.8 14.9 8.4  10.8 8.9 8.3      in
> |
```

- 共46筆資料

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建
立特色典範計畫

16



Example : Taxi time at EWR (1)

names() 函數檢視欄位(變數)名稱

```
> data(ewr)
> names(ewr)
[1] "Year"      "Month"     "AA"        "CO"        "DL"        "HP"        "NW"
[8] "TW"        "UA"        "US"        "inorout"
```

```
> airnames=names(ewr)
> ewr.actual = ewr[,3:10]
> boxplot(ewr.actual)
> |
```

- 把ewr的資料畫成盒鬚圖(不分taxi in與taxi out)

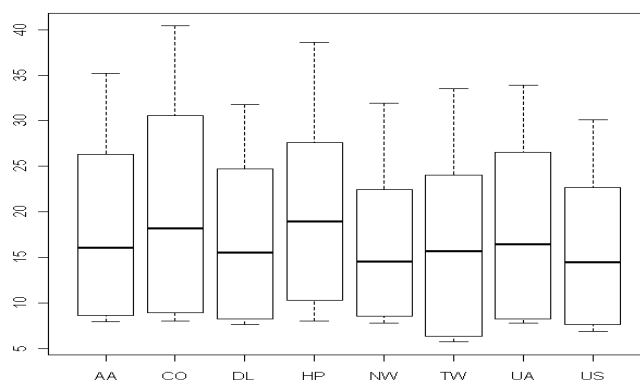
© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建
立特色典範計畫

17



- 不分Taxi in 和 Taxi out做成的盒鬚圖



© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建
立特色典範計畫

18



Example : Taxi time at EWR (2)

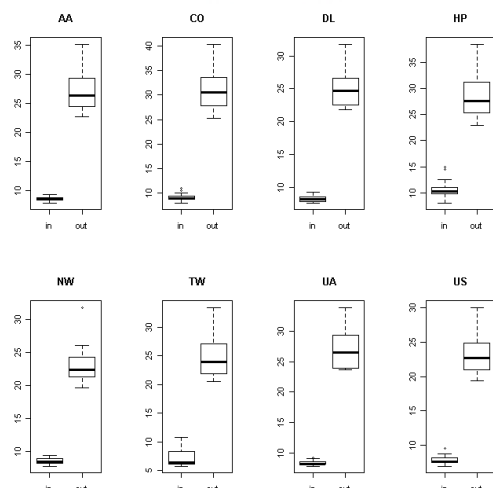
```
> par(mfrow=c(2,4))
> attach(ewr)
> for(i in 3:10) boxplot(ewr[,i]~as.factor(inorout),main=airnames[i])
> |
```

- for表示迴圈，執行八次，依序繪製各家航空公司起飛與降落滑行時間的盒鬚圖，每張圖有兩個盒鬚圖(inorout有兩個因子水準)。
- 其實不需要使用as.factor()函數

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建立特色典範計畫

19



- Taxi in time大致對稱，除了HP
- Taxi out time呈現厚尾，有時要等30分鐘，但NW之taxi out時間較短

© Vince Tsou, IDS, NTCB

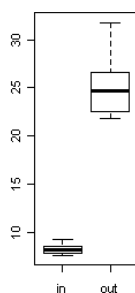
100年度教育部補助技專校院建立特色典範計畫

20



? @ # \$! Try 一下迴圈內的指令吧！

```
> boxplot(ewr[,5]~as.factor(inorout),main=airnames[5])
> |
```



© Vince Tsou, IDS, NTCB

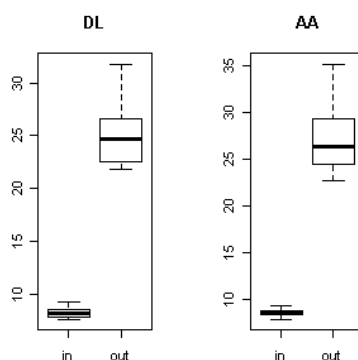
100年度教育部補助技專校院建立特色典範計畫

21



再 Try 一下

```
> boxplot(ewr[,5]~as.factor(inorout),main=airnames[5])
> boxplot(ewr[,3]~as.factor(inorout),main=airnames[3])
> |
```



© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建立特色典範計畫

22



Example :
Symmetric or skewed, long or short ?
對稱或偏態，長尾或短尾

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建
立特色典範計畫

23



對稱性vs.長短尾
單峰資料的六種可能情形

- `runif(n,min=0,max=1)` #均勻分配
- `rnorm(n)` #樣本數為n的常態隨機變數值
- `rt(n,df)` #樣本數為n的t分配隨機變數值
- `boxplot(X)` #盒鬚圖

© Vince Tsou, IDS, NTCB

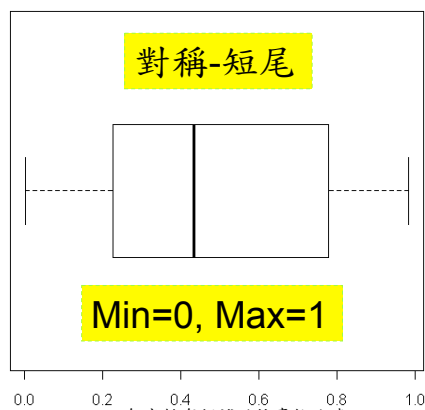
100年度教育部補助技專校院建
立特色典範計畫

24



對稱－短尾

- `X=runif(100); boxplot(X, horizontal=T, bty="n")`
#均勻分配



© Vince Tsou, IDS, NTCB

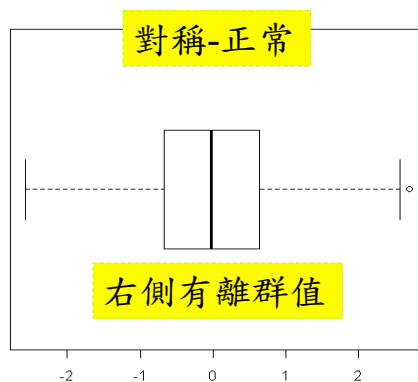
100年度教育部補助技專校院建立特色典範計畫

25



對稱－正常

- `X=rnorm(100); boxplot(X, horizontal=T, bty="n")`
#常態分配



© Vince Tsou, IDS, NTCB

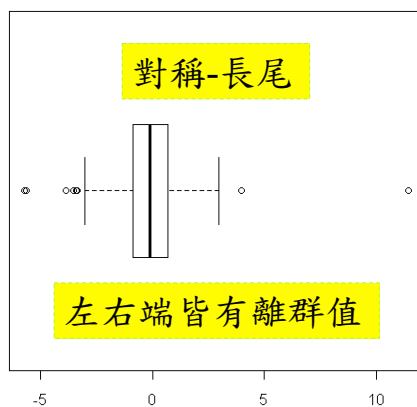
100年度教育部補助技專校院建立特色典範計畫

26



對稱－長尾

- `X=rt(100,2);boxplot(X,horizontal=T,bty="n")`
#學生t分配



© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建立特色典範計畫

27



函數意義

- `sample(X,size,p,replace)` #從X中無置回抽取size個樣本，p給定各出象抽取機率，選項`replace=TRUE`表示有放回抽樣。
- `abs(X)` #計算X的絕對值 $|X|$ ，即`abs(-3)=3`
- `rexp(n,rate)` #樣本數為n的指數分配隨機變數值

© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建立特色典範計畫

28



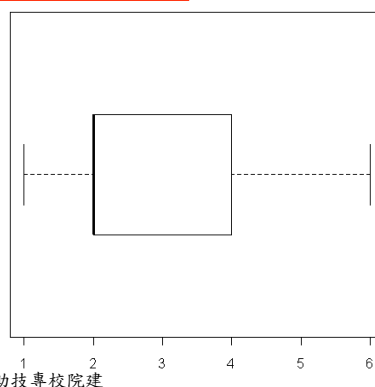
不對稱－短尾

- # 三角分配
- $X = \text{sample}(1:6, 100, p = 7 - (1:6), \text{replace} = T)$;
 $\text{boxplot}(X, \text{horizontal} = T, \text{bty} = "n")$

從1到6中，以置回抽樣方式抽出100個樣本，各數字比例分別為6,5,4,3,2,1(此即為 $p = 7 - (1:6)$ 的設定)。

不對稱(右偏)-短尾

Min=1, Max=6



© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建立特色典範計畫

29



不對稱－正常

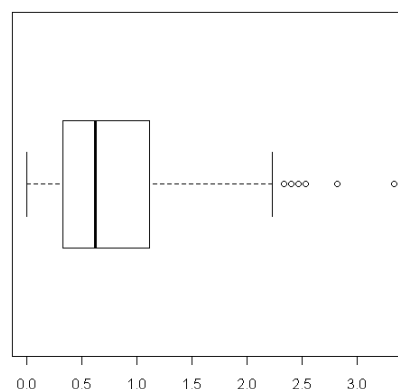
- $X = \text{abs}(\text{rnorm}(200))$; $\text{boxplot}(X, \text{horizontal} = T, \text{bty} = "n")$

#非負常態分配

隨機樣本數=200的常態隨機變數值取絕對值儲存為X，並繪製X的盒鬚圖。

不對稱(右偏)-正常

右側有離群值



© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建立特色典範計畫

30

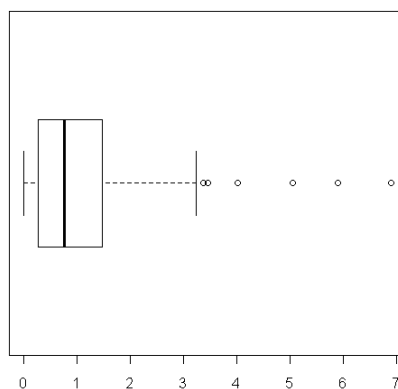


不對稱－長尾

- `X=rexp(200);boxplot(X,horizontal=T,bty="n")`
#指數分配

不對稱(右偏)-長尾

右側有離群值



© Vince Tsou, IDS, NTCB

100年度教育部補助技專校院建立特色典範計畫

31

The End

