統計實務

Spring 2024

授課教師：統計系余清祥

日期：2024年5月29日

第十五週：生活化統計

Exercise G.1 Probability and the law

In a celebrated criminal case in California (People versus Collins, 1968), a black male and a white female were found guilty of robbery, partly on the basis of a probability argument. Eyewitnesses testified that the robbery had been committed by a couple consisting of a black man with a beard and a moustache, and a white woman with blond hair in a ponytail. They were seen driving a car which was partly yellow. A couple, who matched these descriptions, were later arrested. In court they denied the offence and could not otherwise be positively identified.

A mathematics lecturer gave evidence that the six main characteristics had probabilities as follows:

negro man with beard	1/10
man with moustache	1/4
girl with ponytail	1/10
girl with blond hair	1/3
partly yellow car	1/10
inter-racial couple in car	1/1000

The witness then testified that the product rule of probability theory could be used to multiply these probabilities together to give a probability of 1/12 000 000 that a couple chosen at random would have all these characteristics. The prosecutor asked the jury to infer that there was only one chance in 12 million of the defendants' innocence, and the couple were subsequently convicted.

Comment on the above probability argument.

法庭裡的有罪、無罪認定

- 某被告被指控搶劫，以統計檢定的角度思考，虛無假設、對立假設分別是什麼？

→虛無假設？

- 問題：舉證的責任在誰身上？

→原告證明被告有罪；

→被告證明自己無罪。

註：哪一個比較合乎常理？

Exercise G.2 Hospital admissions – communication skills

This exercise requires you to reply to the following letter seeking advice:

Dear Statistician,

The data in Table G.1 show the numbers of new patients arriving at a psychiatric hospital during a recent one-year period. We are interested in finding out if there are any systematic variations in arrival rate, especially any that might be relevant in planning future operations, as we are currently reconsidering the running of this unit. We are particularly interested in any evidence of cyclic behaviour. Please analyse the data for us..

Yours sincerely,

The Assistant Planning Officer
Area X Regional Health Authority

The letter which you write in reply should not exceed three sides of paper, and should be written for a numerate layman rather than a statistician. It should be accompanied by tables and graphs as appropriate, covering not more than four sides.

Table G.1 Numbers of daily admissions to a psychiatric hospital in a one-year period, starting on Sunday 1 January. The lines indicate the end of a week and totals are also given

January

Daily admissions	Weekly total
0, 1, 4, 3, 3, 7, 3	21
0, 2, 1, 4, 3, 6, 1	17
0, 1, 5, 3, 6, 4, 0	19
0, 3, 0, 2, 3, 5, 4	17
0, 4, 2	

February

Daily admissions	Weekly total
3, 2, 4, 1	16
2, 0, 2, 3, 6, 2, 1	16
2, 1, 1, 1, 3, 2, 2	12
1, 3, 0, 6, 3, 5, 2	20
1, 2, 3, 3	

March

Daily admissions	Weekly total
1, 3, 1	14
0, 1, 9, 2, 3, 4, 0	19
1, 2, 3, 1, 1, 3, 0	11
0, 1, 3, 1, 2, 1, 2	10
2, 1, 2, 4, 1, 2, 0	12

April

Daily admissions	Weekly total
1, 8, 2, 3, 4, 4, 1	23
1, 3, 4, 1, 3, 2, 1	15
0, 5, 8, 1, 3, 2, 3	22
2, 0, 4, 8, 1, 2, 2	19
2, 2	

May

Daily admissions	Weekly total
3, 2, 3, 5, 2	19
1, 1, 2, 2, 3, 5, 0	14
0, 1, 1, 3, 2, 1, 2	10
1, 2, 4, 1, 3, 7, 2	20
1, 1, 1, 0, 3	

June

Daily admissions	Weekly total
3, 2	11
0, 4, 2, 3, 3, 4, 2	18
2, 4, 0, 5, 6, 4, 2	23
1, 4, 3, 4, 2, 2, 1	17
2, 3, 0, 3, 5, 4, 1	18

July

Daily admissions	Weekly total
1, 3, 2, 1, 0, 1, 4	12
2, 3, 0, 3, 7, 3, 2	20
1, 3, 5, 3, 5, 5, 2	24
3, 2, 3, 2, 5, 2, 0	17
0, 0, 3	

August

Daily admissions	Weekly total
4, 4, 1, 2	14
2, 2, 3, 3, 2, 4, 5	21
1, 2, 3, 6, 3, 5, 4	24
1, 0, 3, 2, 3, 3, 2	14
1, 1, 3, 4, 3, 5	

September

Daily admissions	Weekly total
1	18
0, 5, 1, 2, 7, 2, 0	17
0, 0, 1, 3, 6, 4, 3	17
1, 3, 4, 3, 1, 7, 3	22
0, 2, 1, 3, 2, 4, 0	12
1	

October

Daily admissions	Weekly total
1, 5, 1, 3, 2, 2	15
1, 2, 2, 2, 1, 4, 1	13
1, 1, 1, 3, 2, 3, 2	13
1, 2, 5, 0, 4, 2, 4	18
0, 3, 1, 4	

November

Daily admissions	Weekly total
1, 3, 2	14
0, 4, 0, 3, 3, 7, 1	18
1, 3, 5, 2, 4, 8, 1	24
1, 3, 3, 1, 2, 2, 4	16
0, 3, 4, 2, 3, 7	

December

Daily admissions	Weekly total
1	20
2, 4, 1, 2, 7, 3, 2	21
1, 0, 5, 4, 4, 3, 2	19
0, 1, 3, 3, 4, 1, 2	14
1, 0, 0, 0, 2, 3, 2	8
3, 4	

Exercise G.3 Testing random numbers

In a national lottery, 125 ticket numbers were drawn by computer in one prize draw. The final six digits of each number were supposed to be random. These values are tabulated in Table G.3. Are the numbers random?

Table G.3 The final six digits of the 125 winning prize numbers in a national lottery

535850	842420	655257	469227	885878
603715	863855	754258	883261	571046
075779	048633	111337	346576	051352
724004	089507	552867	476843	025348
355865	001250	095391	934011	094093
771594	616635	992135	473416	021096
726862	768318	218966	928474	538201
593721	318619	908649	198296	122079
081625	046970	477814	516512	738317
461645	489085	619015	627585	222443

何謂「亂數」(Random Numbers)？

- 假設目標是$\{1,2,...,m\}$的亂數，則希望產生的數值能滿足以下三個要求：
  1. Uniformly distributed
  2. Statistically independent
  3. Reproducible

- 問題：如何驗證產生的數值滿足以上三個條件？

- 如何確定亂數具有*U(0,1)*的特性？

  (亂數需具有哪些特質？)

→ 均勻分配(Uniformity)

多半都使用適合度(Goodness-of-fit)的檢定，確定亂數具有要求的特質，常用的方法有卡方適合度檢定(Chi-square goodness-of-fit)及Kolmogorov-Smirnov 檢定。

→ 互相獨立(Independence)

使用的多為無母數方法，例如：Gap test, Up-and-down test, run test.

Exercise G.4 Sentencing policy – two-way tables?

A study was carried out on consecutive groups of men and women who were convicted of theft and shoplifting in an English city. A total of 100 females and 100 males were examined and their sentences were classified as lenient or severe according to the average sentence given for the person's particular crime. The results were as follows:

	Lenient sentence	Severe sentence	Total
Male	40	60	100
Female	60	40	100

Are females treated more leniently than males? If they are, speculate as to why this should be so.

Simpson's Paradox in University admission

UC Berkeley admitted 44% of males and 35% of females who applied in 1973. Data from the six largest departments.

Department	Male acceptance rate	Female acceptance rate
A	62%	82%
B	63%	68%
C	37%	34%
D	33%	35%
E	28%	24%
F	6%	7%

	Male		Female	
	Applicants	%	Applicants	%
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

- Simpson Paradox的成因：因為女性申請系所的錄取率通常較低。

→Marginal vs. Unconditional probability

	Women		Men	
Dept	Applied	Accepted	Applied	Accepted
A	99	4	1	0
B	1	1	99	10
total	100	5	100	10

決策分析實例

- 萊斯康電子字典

  類似案例：

  → 日本Sony公司的Beta 錄影帶

  → 美國Apple公司的蘋果電腦

表 4.1　消費者購買電子字典的考慮因素

年度	1993	1994	1995
字彙多	47.9%	47.9	46.5
發音準確	37.5	27.2	23.3
功能完整	35.3	21.6	22.3
可插卡	28.0	30.7	33.4
有文法例句	25.1	21.2	22.7
多國會話	24.3	18.7	20.9
多國發音	13.0	8.2	8.7
品牌知名度	8.8	10.0	6.9
國際版權	8.3	7.3	9.2
整句翻譯	-	19.6	20.3
操作容易	-	14.3	22.1
店員介紹	5.1	2.6	5.0
遊戲功能	5.0	4.3	4.8
其他	6.3	-	-

資料來源：ICP 年度調查

表 4.2　電子字典機種銷售結構比

年度	單機型 ($3000-5000)	插卡型 ($6000-9000)
1993	65%	35%
1994	50%	50%
1995	30-40%	60-70%

表 4.3　電子字典市場佔有率

年度	1993	1994	1995
萊思康	31.8%	24.7	21.66
無敵	23.0	37.9	40.09
快譯通	19.2	25.8	27.19
其他	25.8	11.6	11.06

表 4.4　電子字典廣告量統計

廠牌	年度 / 媒體	1993	1994	1995
萊思康	電視	12,302	14,306	13,953
	報紙	9,092	14,056	36,952
	雜誌	3,382	996	910
	合計	24,776	29,358	51,815
無敵	電視	21,799	19,383	18,897
	報紙	27,425	53,619	61,079
	雜誌	3,761	4,058	2,892
	合計	52,985	77,060	82,868
快譯通	電視	21,148	12,852	19,998
	報紙	24,272	18,266	28,100
	雜誌	150	700	1,192
	合計	45,570	31,818	49,290
廣告廠牌總數		13	7	6

資料來源：紅木公司統計　　單位：新台幣千元

— 例題:

1.任何兩位同學屬於同一星座的機會(1/12)

2.任意三位同學都不屬於同一星座的機會

$$\rightarrow P(A)=\frac{12}{12}\times\frac{11}{12}\times\frac{10}{12}=\frac{110}{144}\cong 0.7639$$

→使用排列 $\frac{P(12,3)}{(12)^3}=\frac{12\times 11\times 10}{(12)^3}$

3.任意n位同學都不屬於同一星座的機會大於1/2的最小整數n (5) → $\frac{P(12,5)}{(12)^5}\cong 0.381944$ 與 $\frac{P(12,4)}{(12)^4}\cong 0.572917$

4.任何兩位同學同一天生日的機會(1/365)

5.任何三位同學都不在同一天生日的機會

$$\rightarrow P(A)=\frac{365}{365}\times\frac{364}{365}\times\frac{363}{365}=\frac{132132}{(365)^2}\cong 0.9918$$

6.任意n位同學都不在同一天生日的機會小於1/2的最小整數n (22)

Birthdays

Suppose you have n people. How likely is it that two people have the same birthday?

The probability of n different birthdays is:

$$\left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\cdots\left(1-\frac{n-1}{365}\right)$$

When $n = 23$, this dips below $1/2$.

統計與星座？

- 透過統計分析驗證占星術的推論，以125位十八世紀至二十世紀知名的統計、機率學家為研究對象：

→以1至12的數字將分別代表魔羯座、寶瓶座、南魚座、...、射手座星座。

星座	1	2	3	4	5	6	7	8	9	10	11	12
人數	6	13	7	12	12	16	9	5	12	11	7	6

註：扣除9位出生時間無法判定，列出116人。

→哪些統計方法較為適合？

統計與星座（續）

- 原則上，卡方檢定可使用，分成12個星座的理論個數都大於5，卡方檢定量=13.53，p值約等於0.26，不拒絕均勻分配的假設。

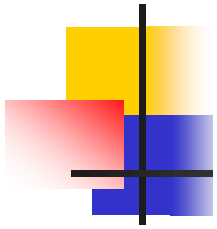
→若分成四個分類：水象(1,4,7)、風象(2,6,10)、水象(3,7,11)、火象(4,8,12)，檢定量=6.69，p值約為0.08。

星座	土象	風象	水象	火象
人數	30	40	23	23

The Car and Goat Game

You're on a game show and you're faced with three doors. Behind one of them is a car. You're asked to pick a door, and you choose door 1. Before the host opens it, she first opens door 2, and you see that there's nothing behind it. She then asks you if you are sure about door 1, or if you want to switch your guess.

Suppose you pick door 1 and that the host opens door 2. The probability that the car is behind 1 is 1/3, so you're probably wrong, but after the host opened door 2, you know that if the car isn't behind 1 (and it probably isn't), then it must be behind 3, so you should switch!



Bus Frequency

Both 97 and 197 have 9 minutes frequency.
23 times with 97 (85.2%) and 4 times with 197.
197 comes 2 min after 97!

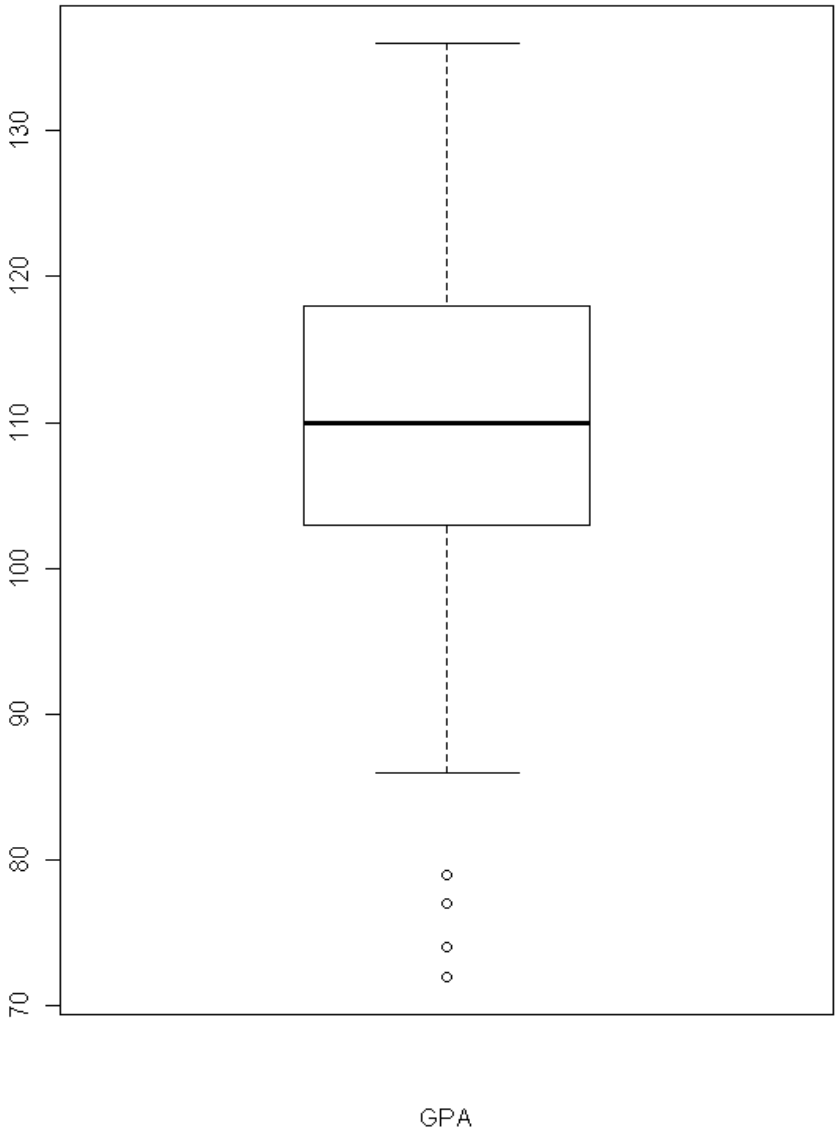
分析範例一（離群值）

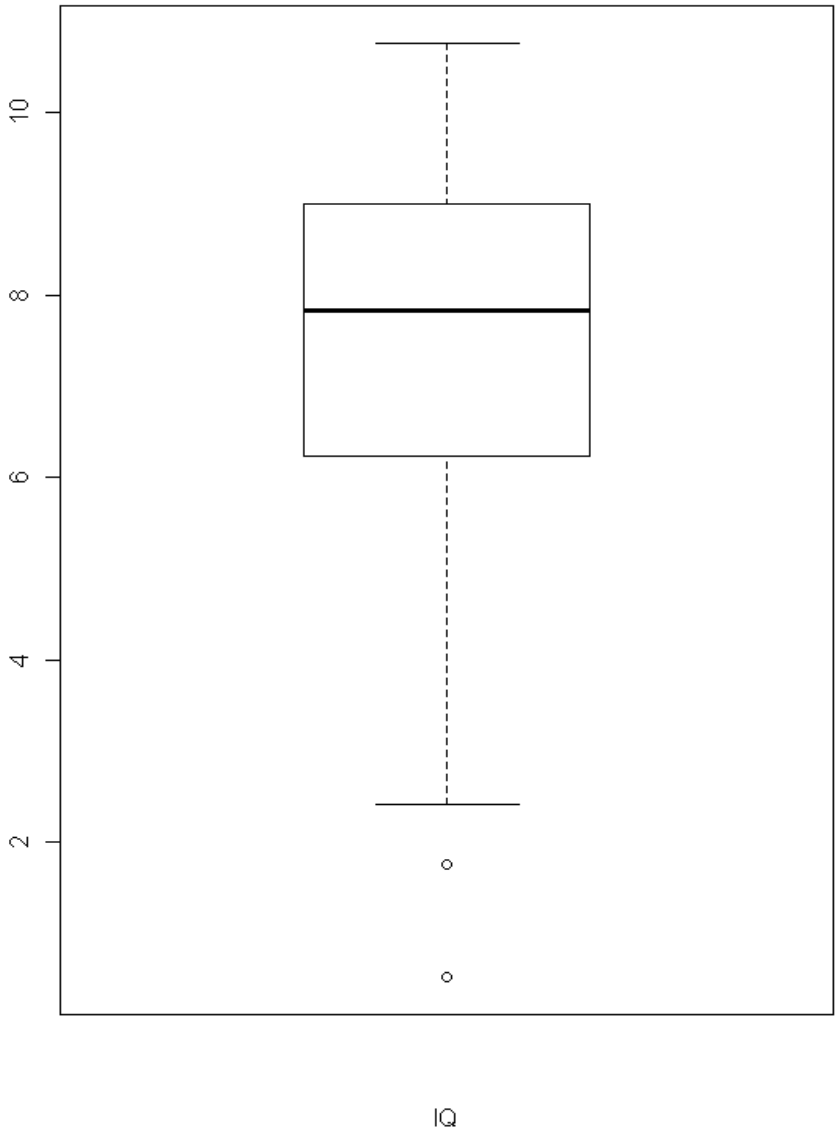
- 離群值與模型假設有非常重要的關聯，不同假設可能有不同的選取結果。
- 以美國中西部國中一年級學生的GPA與IQ為例，找出較為異常的現象。

→可以針對GPA、IQ個別分析

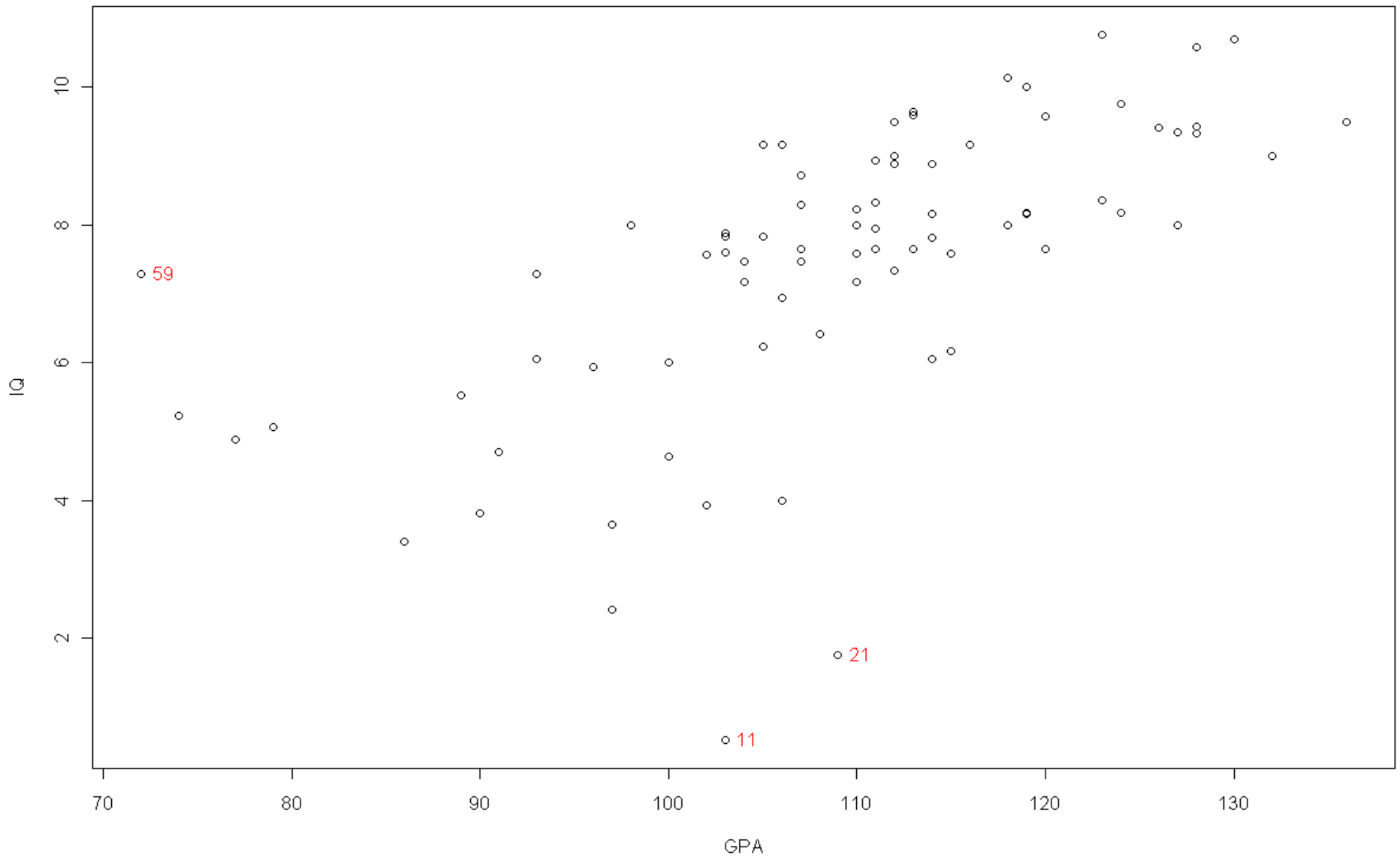
→或是根據加入兩者的關係

註：「Outliers are model-based」！





→GPA及IQ各有4個、2個離群值，IQ有些右偏



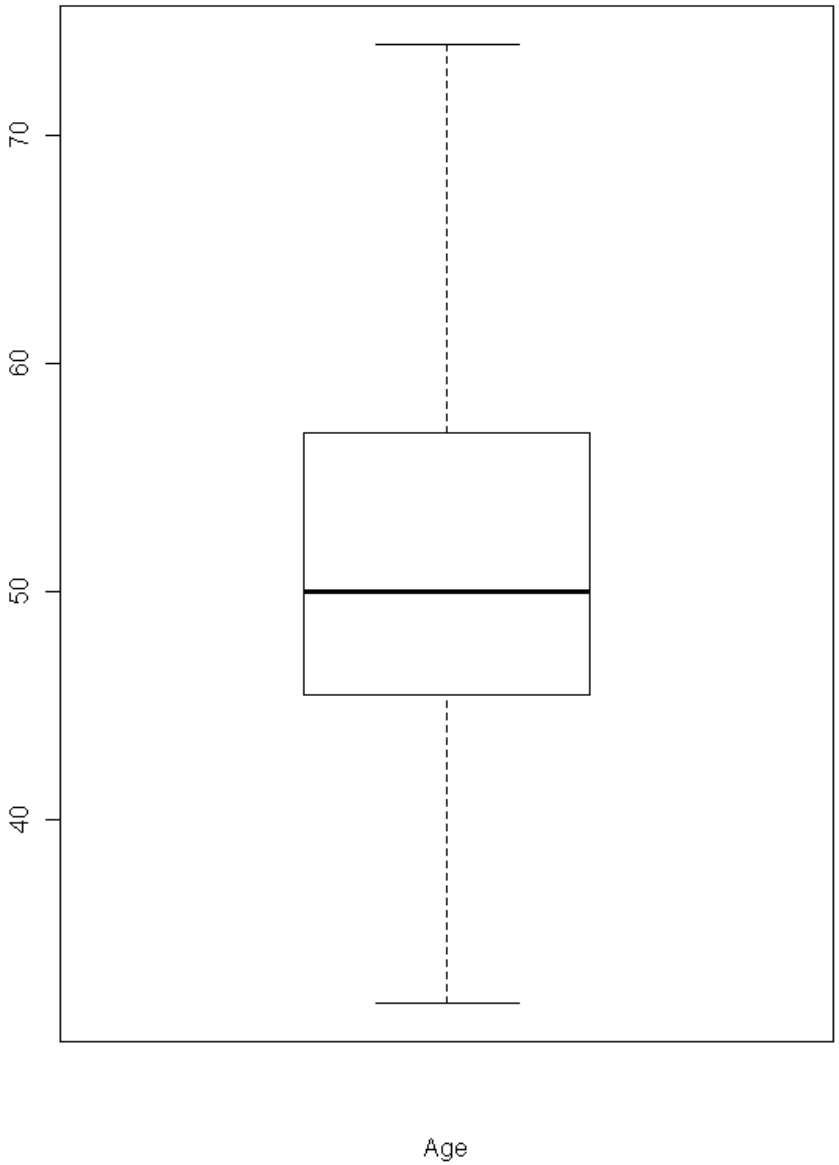
→若考慮GPA及IQ的線性關係，有3個離群值

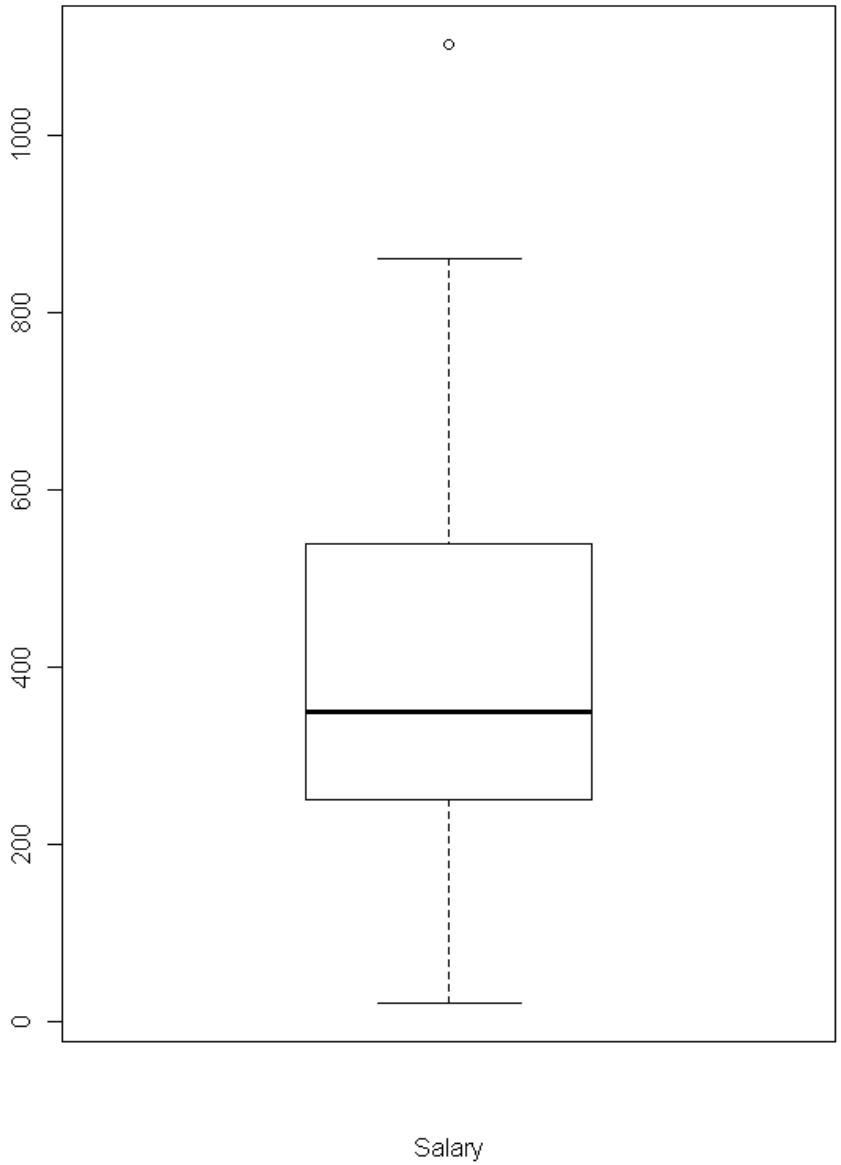
離群值分析範例一（續）

- 一般認為薪水與年齡應該有關聯，但公司經營者是否也如此？1993年美國有一項針對小公司老闆的薪水調查(CEO Salaries)，可用於探討年齡與薪水的關係 。這筆資料是否存在離群值？

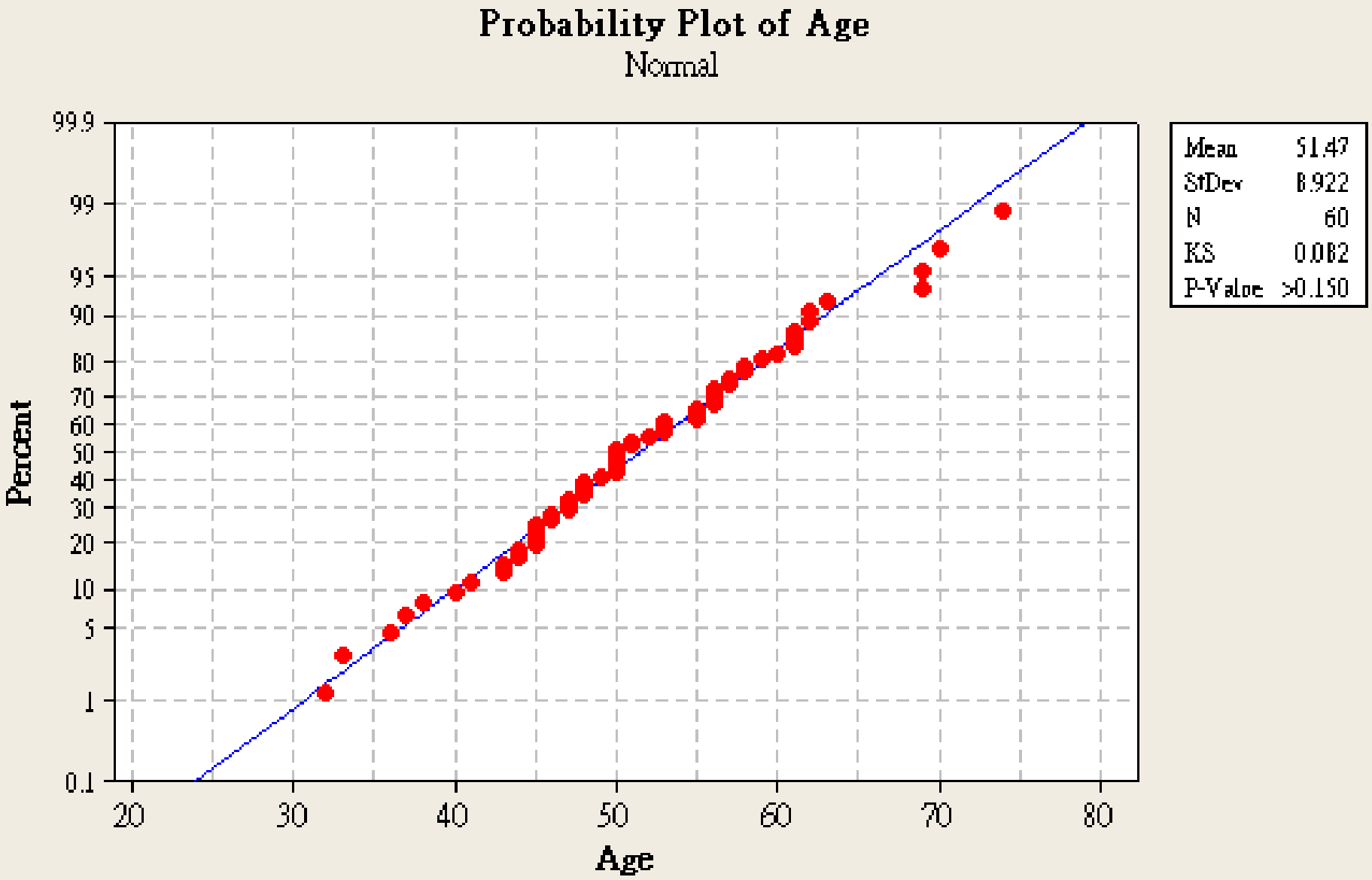
→考量單一變數分析、兩變數關聯分析

→也可考量單一變數的分配，可看出「年齡」接近常態分配，但「薪水」卻不像。

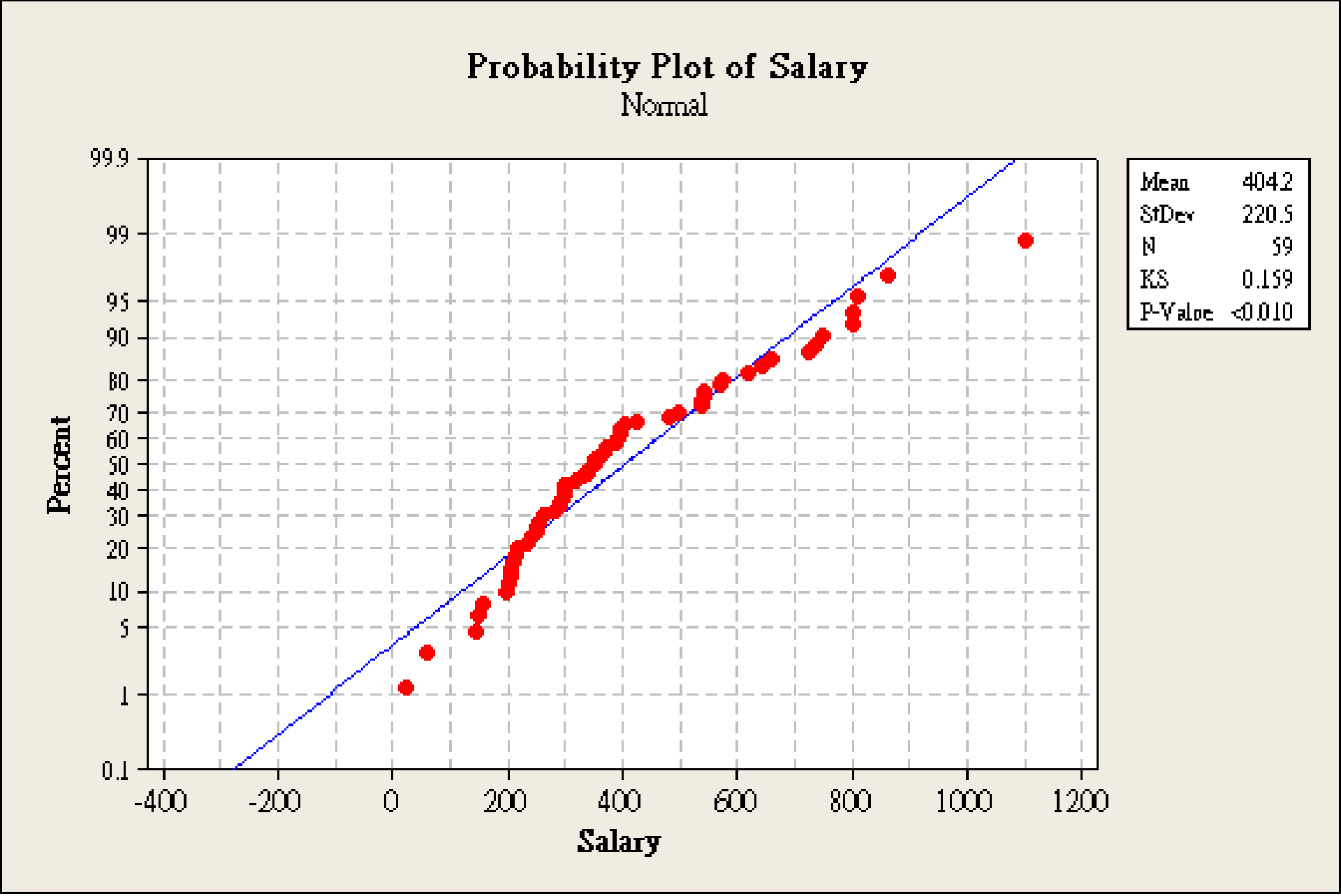




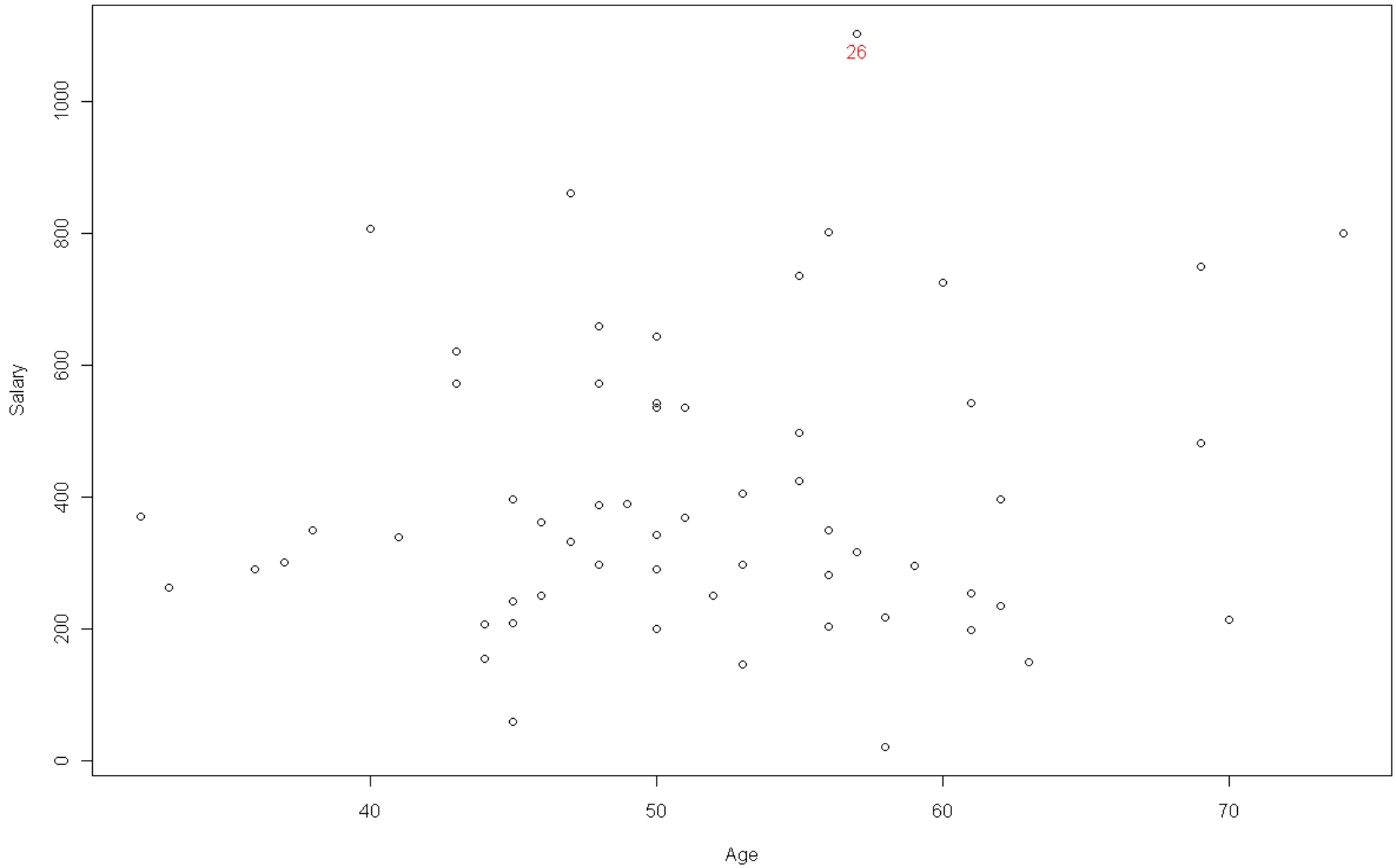
→Age沒有離群值，Salary有1個離群值



→Age常態分配檢定為「不拒絕」，p-value > 0.05



→Salary常態分配檢定為「拒絕」，p-value < 0.05



→Age及Salary沒有明顯的線性關係

資料分析的陷阱

- 如同定義問題，資料分析中也需注意類似陷阱，通常問題有三個主要來源：

(1) Source of Bias（資料偏誤）

→蒐集資料、抽樣、以及資料品質

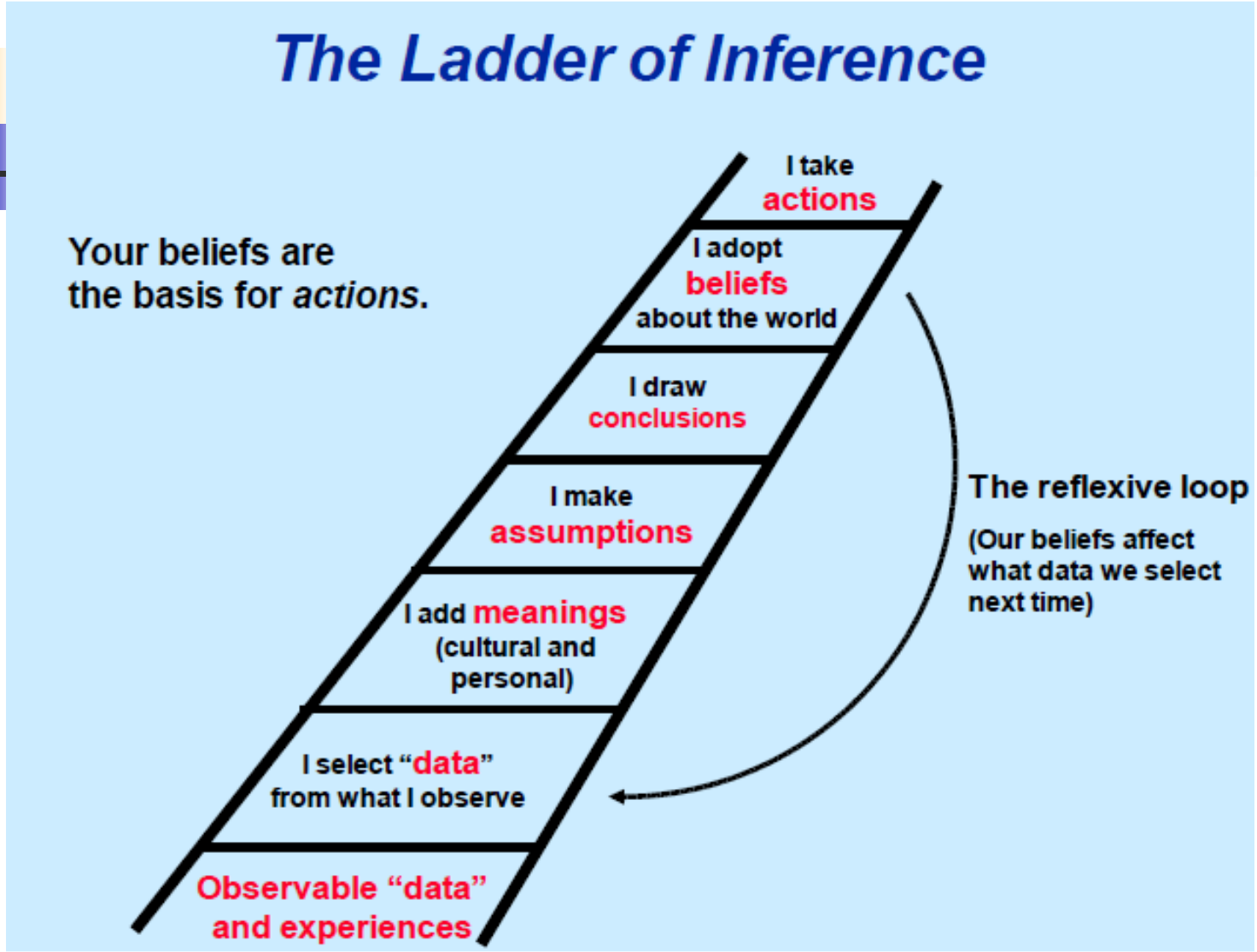
(2) Errors in Methodology（方法瑕疵）

→使用錯誤的分析方法

(3) Problems with Interpretation（過度詮釋）

→不適當地套用分析結果至不同母體

分析推論的步驟



參考資料：Popular Pitfalls of Data Analysis, by Tore Dyba

以下 100 筆資料乃自常態分配 $N(\mu,\sigma^2)$ 抽出的隨機樣本(已排序)：

12.8	42.7	51.2	62.5	73.9
14.8	42.8	51.7	62.7	74.2
21.5	43.5	51.9	62.9	74.8
22.2	44.2	52.4	63.1	75.2
23.8	44.2	53.0	63.8	75.5
30.0	45.1	53.3	63.9	75.8
30.5	45.1	53.6	63.9	76.1
32.2	45.7	56.1	65.5	76.9
32.7	47.2	57.4	65.9	77.2
33.7	47.4	57.5	67.0	78.4
34.1	48.0	57.5	67.1	79.6
34.9	48.3	58.4	67.2	80.1
35.6	49.1	59.1	67.3	80.9
36.1	49.1	59.5	68.6	82.6
36.9	49.2	59.7	69.0	83.7
37.5	49.4	60.1	70.8	83.7
40.8	49.6	60.2	71.2	84.0
41.3	49.7	60.8	72.2	84.3
42.0	50.8	61.4	73.2	84.7
42.5	50.9	61.8	73.8	85.7

→請問這些資料有什麼特性？

(a) 在不借助於樣本平均數及樣本變異數，估計期望值 μ 、變異數 σ^2 。(註：也就是純粹藉由目視來判斷！)

(b) 由你/妳從(a)估計出的期望值及變異數，驗證這 100 筆資料是否服從常態分配。(註：不能使用圖表，建議使用平均數與樣本標準差！)

(c) 以這些資料繪製 Boxplot，並以繪出的圖形驗證資料是否具有常態分配的特性。

(d) 假設這些資料在記錄時，意外地將 90 筆來自 $N(\mu_1,\sigma^2)$ 分配、10 筆來自 $N(\mu_2,\sigma^2)$ 分配混在一起，其中 $\mu_1 \neq \mu_2$ 。請說明如何判斷 μ_1,μ_2 兩者何者較大，同時大略估計 μ_1,μ_2 兩者的差異大小。

Note: 上述資料是由 90 筆來自 $N(60,15^2)$ 分配、10 筆來自 $N(30,15^2)$ 分配組成。

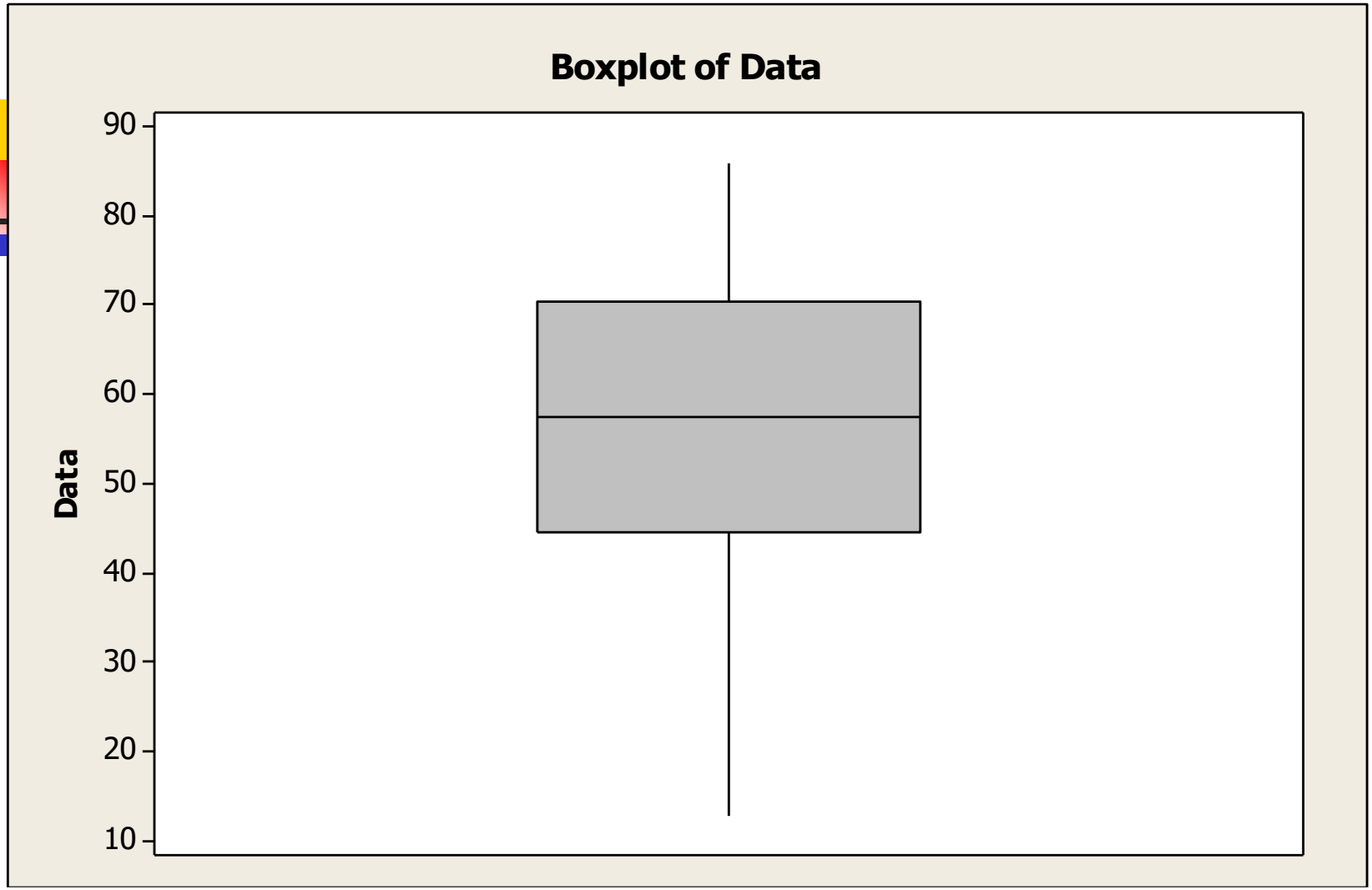
Variable	N	Mean	Median	TrMean	StDev	SE Mean
第一組	90	59.45	59.94	59.53	14.75	1.55
第二組	10	29.76	26.93	28.40	13.92	4.40
合併	100	56.49	57.56	57.01	17.13	1.71

Variable	Minimum	Maximum	Q1	Q3
第一組	30.55	85.79	48.26	72.52
第二組	12.89	57.54	19.90	38.56
合併	12.89	85.79	44.51	70.38

(a) 平均數(期望值)可由中位數代替，標準差可由全距/4 或全距/4 近似。本題中的樣本平均數為56.49與中位數57.56很接近；樣本標準差為17.13與全距/4 ＝18.23，但與全距/6 ＝12.15較為接近，而與相去較遠。

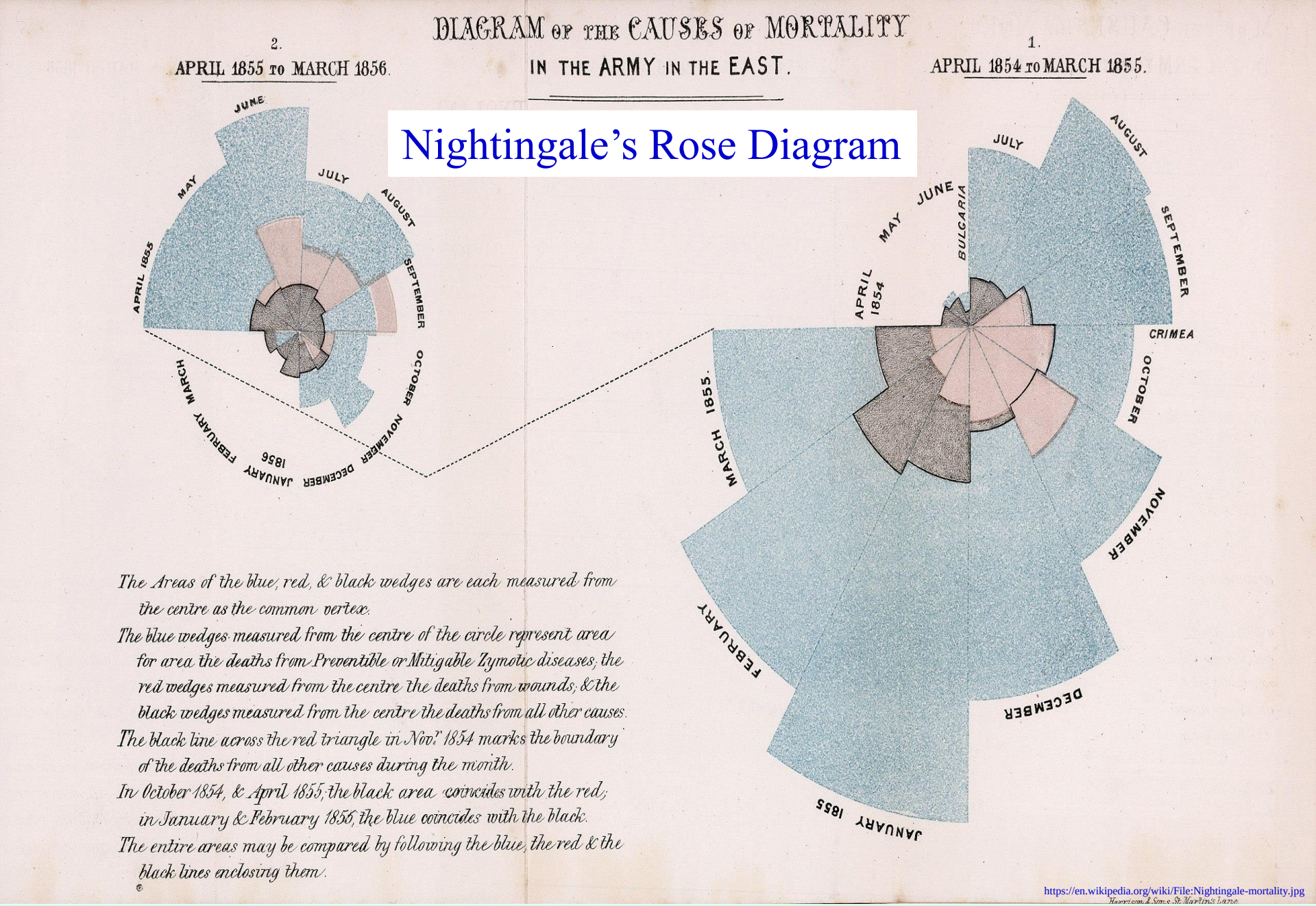
(b) 樣本平均數加減一倍標準差約可涵蓋68%的觀察值，樣本平均數加減兩倍標準差約可涵蓋95%的觀察值。依此想法檢查，我們發現：各有48個、84個觀察值落入各區間，與預期有一段相當大的差距。(若代入18.23為標準差則有較佳的結果，分別有69、98個點落在區間內。)

(c) 由下圖可知資料數值較小的散佈較為分散，雖然沒有明顯的離群值、中間的Box較為集中，常態分配可能有問題。（QQ plot及常態分配檢定不拒絕！）



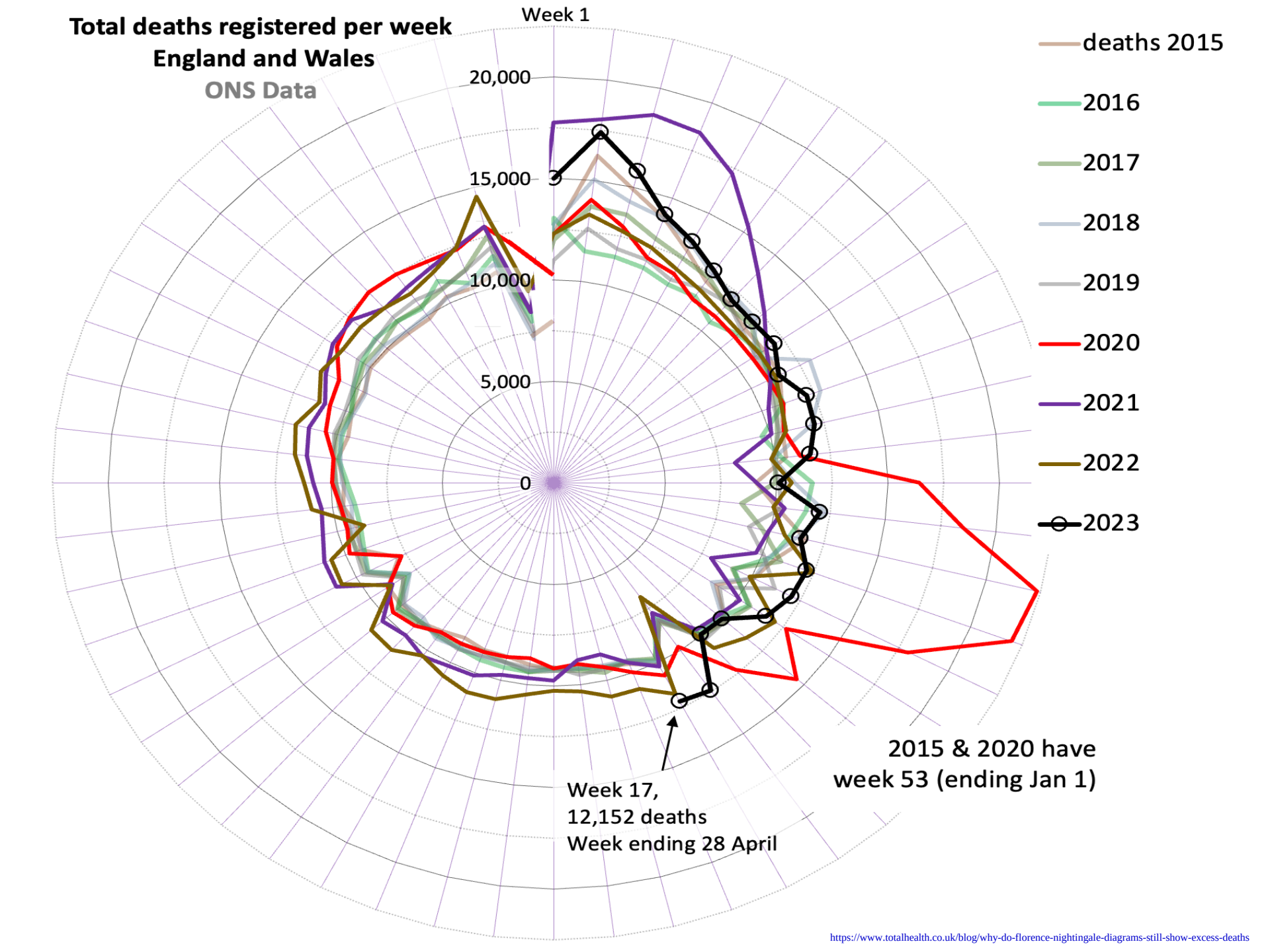
除了資料較多者多一些外，無法確定是否不為常態分配

Nightingale's Rose Diagram



https://en.wikipedia.org/wiki/File:Nightingale-mortality.jpg

Florence Nightingale, Mortality and Health Diagrams



https://www.totalhealth.co.uk/blog/why-do-florence-nightingale-diagrams-still-show-excess-deaths