3-6信賴區間與信心水準解讀

常態分配

課本 P.223 始

- 很多資料畫成直方圖後，若將長方形中點連接成折線圖後，發現資料中間高往兩邊下降近似鐘型，則稱此組資料近似常態分配。
- 常態分配：數學家–高斯所創。統計學上最常用也最重要的分配。
- 對稱！

經驗法則

對於常態分配的資料，我們由次數分配圖呈鐘形，知道中間部分所占的比例較大，愈往兩旁所占的比例愈小，但比例大約是多少呢？如果也知道此組資料的平均數 $\overline{x}$，樣本標準差 s，就能估算出：

大約有 68%的資料落在區間 $[\overline{x}-s,\ \overline{x}+s]$ 內，

大約有 95%的資料落在區間 $[\overline{x}-2s,\ \overline{x}+2s]$ 內，

大約有 99.7%的資料落在區間 $[\overline{x}-3s,\ \overline{x}+3s]$ 內，

註：區間 $[a,\ b]=\{x \mid a\leq x\leq b\}$.

此種現象稱之為經驗法則（或稱 68%, 95%, 99.7%規律）.

經驗法則

大約有 68%的資料落在區間$[\bar{x}-s,\ \bar{x}+s]$內,

大約有 95%的資料落在區間$[\bar{x}-2s,\ \bar{x}+2s]$內,

大約有 99.7%的資料落在區間$[\bar{x}-3s,\ \bar{x}+3s]$內,

註:區間$[a,\ b]=\{x \mid a\le x\le b\}$.

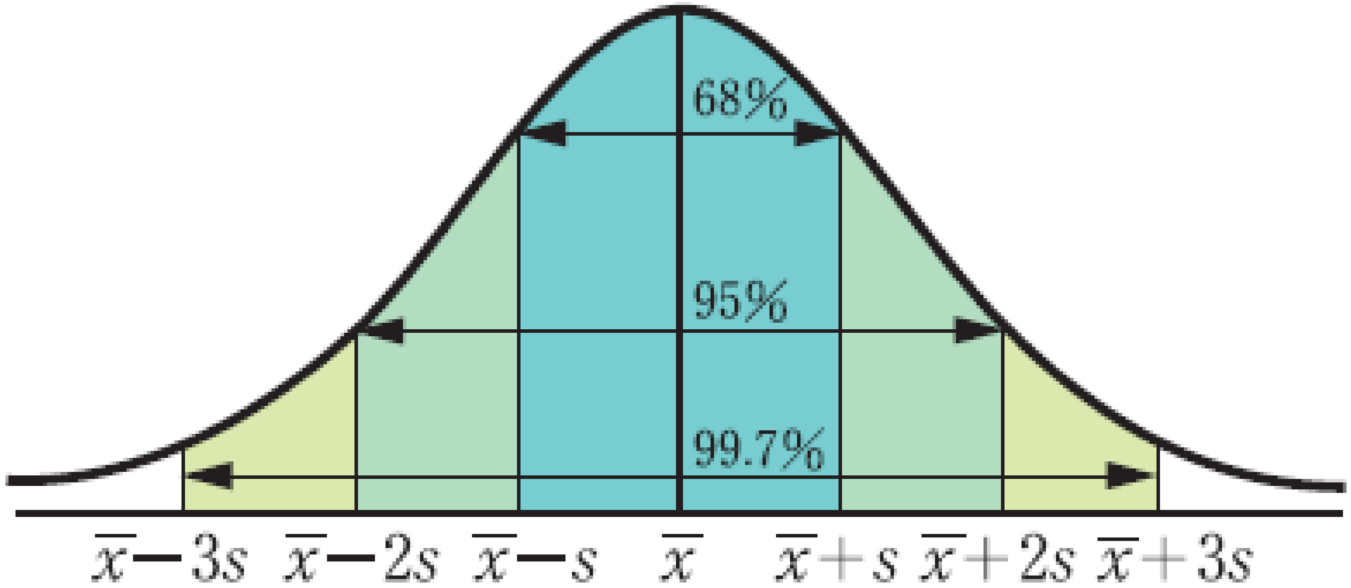


圖 3-11 經驗法則

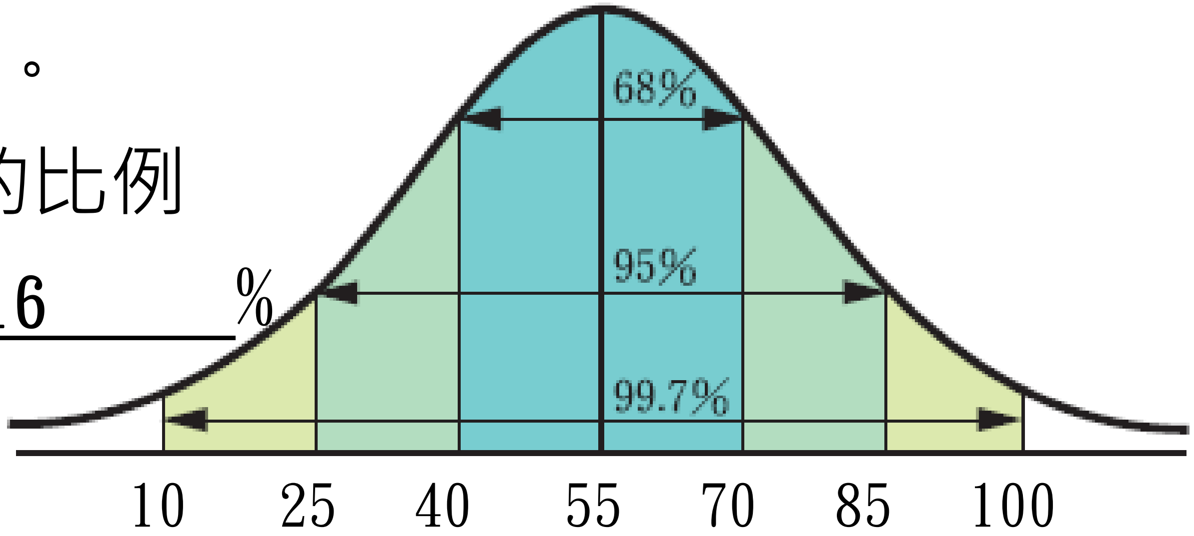
範例一

點線面講義 P.186

- Q ：某高中一年級有學生600人，第一次段考數學成績呈常態分配，平均分數為55分，標準差為15分，請概估下列問題。
- (1)此次段考數學成績40分以下的學生有幾人？
- (2)此次段考成績超過85分的同學有幾位？
- (3)已知大雄考70分，則他在全一年級學生中約排第幾名？

平均55分，標準差15分

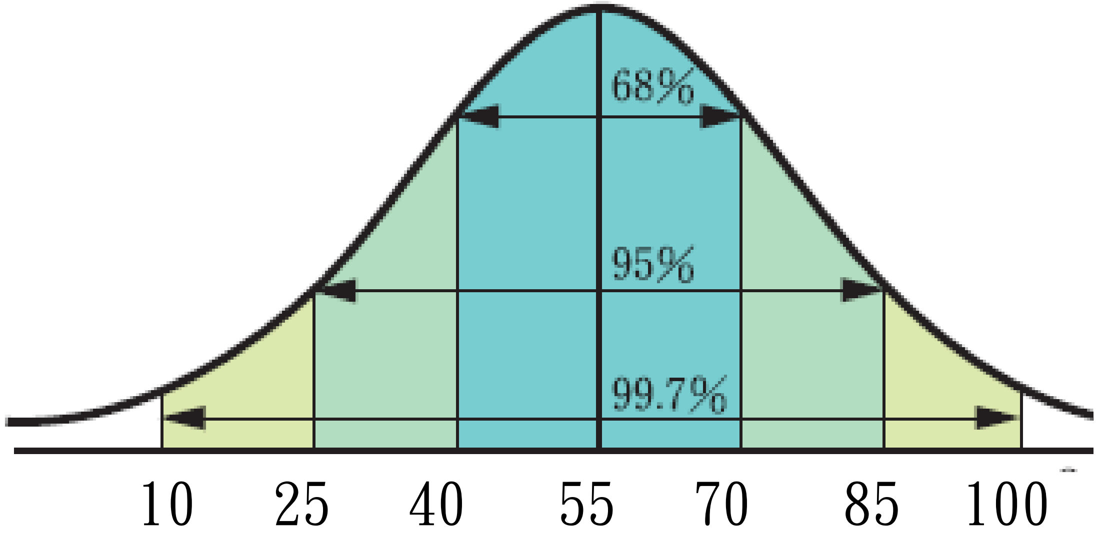
- <Sol> ：畫圖。
- (1)40分以下的比例占全體的___16___%



所以40分以下共有 600×16%=96(人)

平均55分，標準差15分

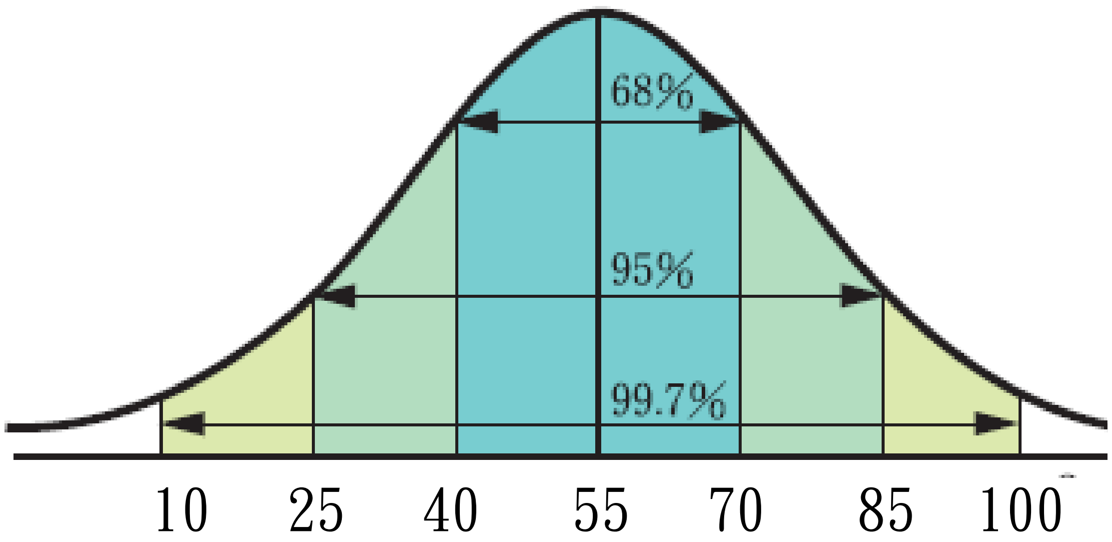
- <Sol>
- (2)85分以上
  所佔的比例
  為 ___2.5___ %



所以85分以上共有 600×2.5%=15(人)

平均55分，標準差15分

- <Sol>
- (3)70分以上
- 約占全體<u>16</u>%



所以約為第600×16%=96（名）

隨機變數

若樣本 $x_1, \cdots, x_n$ 隨機抽樣自此母體，我們會以樣本平均數 $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$ 來估計 μ，但每次抽出的 n 個樣本會不同，因此算出的樣本平均數 $\bar{x}$ 會不一樣，所以稱 $\bar{x}$ 為“**隨機變數**”. 既然每次抽樣算出的 $\bar{x}$ 不同，有必要了解隨機變數 $\bar{x}$ 的抽樣分布長相是如何, 這樣才能知道下決策時，犯錯的機會有多大，而誤差又有多少？

中央極限定理(C. L. T)

- 當n很大時，不管原母體為什麼分配，也不管母體資料是離散型或連續型…，只要樣本數夠大
- (1) $\bar{x}=\dfrac{\sum_{i=1}^{n} x_i}{n}$ 的分配會接近鐘型的常態分配
- (2) $\bar{x}$ 的平均數與原母體平均數相同，都是 μ
- (3)但 $\bar{x}$ 的標準差與原本母體的標準差 σ 不同，變成只有 $\dfrac{\sigma}{\sqrt{n}}$ （證明超出高中範圍）

信賴區間與信心水準

課本 P.229 始

選舉前所做的民調，目的是想在投票前預估某候選人的得票率，當然，我們希望估計的得票率 $\hat{p}$ 與開票後真正的得票率 p 愈接近愈好.

估計的得票率與真正的得票率的差 $p - \hat{p}$ 稱為**抽樣誤差**. 當抽樣誤差為負數時，表示高估了此候選人的得票率，反之，當抽樣誤差為正數時，表示低估了此候選人的得票率.

民調的意義

我們常在報章上看到下面的一段話：

"此次民意調查有效樣本 1068 人，在信心水準 95% 下，抽樣誤差在 ±3%內".

如由臺北市選民抽出 1068 位有效樣本，訪問他們要投給哪位候選人，結果有 420 位選民表示要投給候選人甲，則候選人甲樣本的得票率為 $\hat{p}=\dfrac{420}{1068}=0.3933$，如果候選人甲真正得票率 $p=0.4$，則抽樣誤差為

$$p-\hat{p}=0.4-0.3933=0.0067.$$

開票前候選人甲真正得票率 p 是未知的，我們想由樣本得票率 $\hat{p}$ 估計真正得票率 p 落在哪個範圍？落在此範圍的機率是多大？

信賴區間
信心水準

通常要求 p 落在哪個範圍是以區間表示，稱為**信賴區間**，而落在此區間的機率稱為**信心水準**（或稱**信賴度**或稱**信賴係數**）.

回來看226頁的問題

在第 226 頁抽樣 1068 位，抽樣誤差在$\pm 3\%$內者約占 95%，真正得票率 p 與樣本得票率 $\hat{p}$ 滿足 $|p-\hat{p}| \leqslant 0.03$，而區間〔$\hat{p}-0.03, \hat{p}+0.03$〕為“此次”調查得票率 p 的 95%信賴區間，例如：抽樣 1068 位樣本中有 420 位表示要投給候選人甲，算出 $\hat{p}=\dfrac{420}{1068}=0.3933$，則區間

〔$\hat{p}-0.03, \hat{p}+0.03$〕=〔0.3633，0.4233〕就是此次民意調查候選人甲得票率 p 的 95%信賴區間。

講義範例二

點線面講義 P.187

- Q：鄉村民調公司在T市對其首長滿意度做民意調查，其中滿意度為5成2，本次調查是在95%的信心水準下，抽樣誤差為±3 個百分點，成功訪問了1150位T市滿20歲以上的民眾。試問：

- (1)本次調查的母體為何？又樣本數為何？
- (2)此次調查的信賴區間為何？

講義範例二

點線面講義 P.187

- <Sol>

- (1) 母體為T市滿20歲以上的民眾
  樣本數為1150

- (2)信賴區間為$[0.52-0.03, \ 0.52+0.03]$
  $=[0.49, \ 0.55]$

接下來將說明此信賴區間如何算出，要求信賴區間主要依據是常態分配的經驗法則，本節也將說明信心水準是95％的意義。

很常見到資料是二分類的，即反應值只有二種可能，例如：每次丟一個硬幣，不是正面就是反面；若袋中只有紅球、藍球二種，則每次拿球，不是拿到紅球就是拿到藍球；選舉時對候選人甲，每位選民不是投給他，就是不投給他；求職者在面試後，不是成功，就是失敗；雜誌訂戶不是續訂，就是不再續訂；折價券不是兌現，就是不兌現.

課本232頁　（重要）

母體比率 p 的 68%，95%，99.7% 信賴區間

最大誤差

母體比率 p 的信賴區間為

$$\left[\, \hat{p}-e,\ \hat{p}+e \,\right].$$

其中，$\hat{p}$ 為樣本比率，e 為**最大誤差**（亦稱**誤差界限**），信心水準 68%，95%，99.7%所對應的最大誤差 e 分別為

(1) $e=\dfrac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$.　(2) $e=\dfrac{2\times\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$.

(3) $e=\dfrac{3\times\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$.

課本例題一

課本 P.232

例題 1

如上述對未婚生子的研究，調查 1068 位，結果有 425 位贊成，求：

(1) 信心水準 95% 的最大誤差.

(2) 贊成比率 p 的 95% 信賴區間.

$n=1068$，$\hat{p}=0.3979$，所以 95% 最大誤差

$$e=\frac{2\times\sqrt{0.3979\times(1-0.3979)}}{\sqrt{1068}}\doteq 0.03,$$

因此 p 的 95% 信賴區間為

$$[0.3979-0.03,\ 0.3979+0.03]=[0.3679,\ 0.4279].$$

例題 3

某班有 32 位學生，丟一個硬幣做實驗，每位學生各丟此硬幣 1 次，結果如下：

1111，1100，1111，1001，0110，0111，0101，1111

其中 1 表正面、0 表反面，

試求此硬幣出現正面的比率 p 的 95%信賴區間.

32 次實驗中有 23 個 1，9 個 0，所以估計此硬幣出現正面比率 p 為

$$\hat{p}=\frac{23}{32}=0.71875.$$

95%信心水準的最大誤差為

$$e=2\times\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\fallingdotseq 2\times\frac{0.44961}{\sqrt{32}}\fallingdotseq 0.15896.$$

因此，此硬幣出現正面比率 p 的 95% 信賴區間為

$$\begin{aligned}\left[\,\hat{p}-e,\ \hat{p}+e\,\right] &= \left[\,0.71875-0.15896,\ 0.71875+0.15896\,\right] \\ &= \left[\,0.55979,\ 0.87771\,\right].\end{aligned}$$

例題 4

袋中有紅球、藍球各若干個，某人每次從袋中拿出一球看完顏色後又放回袋中，共拿 50 次，結果有 20 次拿出紅球，求袋中紅球所占比率 p 的 95%信賴區間.

50 次實驗中結果拿出紅球有 20 次，所以估此袋中紅球所占比率 p 為

$$\hat{p}=\frac{20}{50}=0.4.$$

95%信心水準的最大誤差為

$$e=2\times\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}=2\times\frac{\sqrt{0.4\times0.6}}{\sqrt{50}}\fallingdotseq 0.13856.$$

因此，此袋中紅球所占比率 p 的 95%信賴區間為

$$\begin{aligned}[\hat{p}-e,\ \hat{p}+e]&=[0.4-0.13856,\ 0.4+0.13856]\\&=[0.26144,\ 0.53856].\end{aligned}$$

根據前面幾個例題，整理出以下幾點注意事項：

(1) 母體比率 p 的信賴區間中點是樣本比率 $\hat{p}$.

(2) 信賴區間愈短愈好，區間長度等於 2 倍最大誤差.

(3) 當抽樣樣本數 n 愈大時，若 $\hat{p}$ 不變，則抽樣誤差界限 e 愈小，所以信賴區間長度變愈短.

(4) 信心水準愈高，則最大誤差 e 愈大，所以信賴區間長度也愈長（參看上面例題 2）.

(5) 信賴區間會隨抽樣資料不同算出不同的樣本比率 $\hat{p}$，而得不同的信賴區間.

範例三

點線面 P.188

- Q ：某市自來水公司以問卷調查市民對”飲用自來水”認同度的支持，回收有效問卷900張，其中認為可飲用自來水有90張. 試求：
- (1)認同可飲用自來水的比率
- (2)在95%的信心水準下，這次調查的正負誤差為多少個百分點？
- (3)95%的信賴區間

回收問卷900張，可飲用有90張

- <Sol>
- （1） 認同比率 $\hat{p}=\dfrac{90}{900}=\dfrac{1}{10}=0.1$
- （2） 在95%信心水準下，

  最大誤差為

$$e=2\cdot\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}=2\cdot\sqrt{\frac{0.1\cdot(1-0.1)}{900}}=0.02=2\%$$

所以此次調查的正負誤差為2%

- <Sol>
- （3） 95%的信賴區間為

$$[\hat{p}-2\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\hat{p}+2\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

$$=[0.1-0.02,0.1+0.02]=[0.08,0.12]$$

信賴區間的意義

課本 P.236

所謂調查樣本數 n 的 95%信賴區間為〔 $\hat{p}-e$, $\hat{p}+e$ 〕, 即 p 落在區間〔 $\hat{p}-e$, $\hat{p}+e$ 〕內的機率是 0.95. 它的真正意義是重複這種抽樣方式共做 k 次調查, 每次調查樣本數都是 n, 每次可得到一個信賴區間, 若 k 很大, 則這 k 個信賴區間中約有 $k\times 95\%$ 個區間包含真正的得票率 p.

100 個信賴區間圖，區間長度不同

圖 3-16　擲 32 個硬幣 100 次所得 95%信賴區間

樣本數的決定

課本 P.238

選舉時，候選人都想知道自己及競爭對手的得票率大約是多少，當然等到開票後就能得到每位候選人真正的得票率，但開票後一切都已成定局，要改變競選策略為時已晚，因此，候選人希望在開票前預估自己的得票率，以便推出新的文宣廣告提高得票率，得票率的民意調查要調查多少人才合適？

一般而言，調查人數愈多，成本愈高，估計也愈準確，調查人數多寡要考慮到成本與估計的準確性.

在 95%信心水準下，如要控制估計誤差在 e 以內（即最大誤差為 e）時，需要調查多少樣本數 n？

(1) 如果能得到 $\hat{p}$ 的估計，則由 $e=\dfrac{2\times\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$，

求得 $$n=\frac{4\times\hat{p}(1-\hat{p})}{e^2}.$$

(2) 但如沒有 $\hat{p}$ 的資訊時，可利用保守的方式求樣本數如下：因 $\sqrt{\hat{p}(1-\hat{p})}\leq 0.5$，所以

$$e=\frac{2\times\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\leq\frac{2\times 0.5}{\sqrt{n}}=\frac{1}{\sqrt{n}},$$

得總數 $n\doteq(\dfrac{1}{e})^2$.

在 95%信心水準下，最大誤差為 e 時，樣本數 n 約為 $(\dfrac{1}{e})^2$

範例四

點線面 P.188

- Q ：某家手機業者想知道在南台灣，其手機占有率p有多少，若在95%信心水準下，而誤差在± 0.04之內，請問需要調查多少人？

解題關鍵：($\hat{p}$ 未知)

- <Sol>
- 在95%信心水準下，抽樣誤差為：

$$e = 2\cdot\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}} \le 2\times\frac{0.5}{\sqrt{n}} = \frac{1}{\sqrt{n}}$$

- 現誤差在 ±0.04 內

$$\therefore \frac{1}{\sqrt{n}} \le 0.04 \Rightarrow \frac{1}{n} \le (0.04)^2$$

$$\therefore n \ge \frac{1}{(0.04)^2} = 625$$

在 95%信心水準下，最大誤差為 e 時，樣本數 n 約為 $\left(\frac{1}{e}\right)^2$

範例五

點線面 P.189

- Q :(本季最後一題XD)據保險公司的調查資料顯示，全市的人口有保壽險的比率為36% ，若想推論資料的準確性，在95%的信心水準及抽樣誤差 $\pm 4\%$ 的條件下，最少應調查有效樣本多少個？

- 解題關鍵：($\hat{p}$ 已知)

- <Sol> 根據題意
- 可得 $\hat{p} = 0.36$
- 95%的信心水準下抽樣誤差

$$e = 2 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \leq 0.04$$

$$\sqrt{\frac{0.36 \cdot (1 - 0.36)}{n}} \leq 0.02$$

$$\Rightarrow n \geq \frac{0.36 \times 0.64}{(0.02)^2} = 576$$

回憶過去，說好的加分呢？☺

- Q1：若 $x+y=100$ ，$x, y \in \mathbb{N}$ ，從解集合中任取一元素，求其積大於1000的機率為多少？

- Q2：桌球一盤比賽先勝3局者贏，贏一盤可以獲得獎金1280元，甲、乙兩人過去勝負的比例為2：6，但今甲已連勝2局，此刻因非人為因素必須停止比賽，則甲應分到多少獎金才公平？

- Q3：已知 $\hat{p}=0.2$ ，$n=100$ 求母體p的95%信賴區間。