

第四章

隨機樣本及其他

4.1 基本性質

在上一章我們討論了多維隨機變數，有一種特別的多維隨機變數，是常用來當作機率模式的，就是所謂**隨機樣本**(random sample)。我們先給一定義。

定義1.1 若隨機變數 $X_1, \dots, X_n$ 為相互獨立，且 X_i 's 之邊際p.d.f. $f(x)$ 皆相同，則稱 $X_1, \dots, X_n$ 為一組從**母體**(population) $f(x)$ 產生之樣本數為 n 之**隨機樣本**(a random sample of size n from the population $f(x)$)。此時亦稱 $X_1, \dots, X_n$ 為**獨立且有共同分佈** (independent and identically distributed, 簡稱i.i.d.)之隨機變數，以 $f(x)$ 為其共同p.d.f.。

在上述定義中，常省略“樣本數為 n 之”數字，此因由 $X_1, \dots, X_n$ 已顯示有 n 個樣本。又p.d.f. $f(x)$ 可換成分佈函數 $F(x)$ ，或明定那一分佈。如可以說 $X_1, \dots, X_n$ 為一組i.i.d.的隨機變數，以 $\mathcal{N}(0, 1)$ 為其共同分佈。諸如， $X_1, \dots, X_n$ 為一組由 $\mathcal{P}(\lambda)$ 分佈所產生的隨機樣本，也都是表示 $X_1, \dots, X_n$ 為i.i.d.的隨機變數，以 $\mathcal{P}(\lambda)$ 分佈為其共同分佈。這些不太一樣的講法，

2 第四章 隨機樣本及其他

其實都是指同一回事。

另外,在上述定義中,我們提到**母體**。早期母體是指對與人有關的探討,所得到的**觀測值**(observation)。今日則不論對動物、機器等,只要是我們有興趣探討的事物,皆可以母體稱之。例如,若想探討猴子之壽命,則母體可取為全世界的猴子。因此所謂**母體**就是觀測值之全體所成的集合。而一組樣本就是母體之一子集合,母體中之一元素就是一個樣本。

由母體得到的一觀測值,為一隨機變數(如上述猴子的壽命),因此便伴隨著機率密度函數,及分佈函數。由於我們是要做統計分析,故雖原始的母體可能是猴子,或飛機等,但我們只關心一母體某一數值特性的機率結構,或說機率模式。而i.i.d.隨機變數,是一種特別、常見,且好處理的樣本。這是定義1.1產生的背景。至於母體與樣本空間之別為何,我們留給讀者自行討論。

由定義1.1,當 $X_1, \dots, X_n$ 為i.i.d.,以 $f(x)$ 為其共同p.d.f.,則 $X_1, \dots, X_n$ 之聯合p.d.f.為

$$(1.1) \quad f(x_1, \dots, x_n) = f(x_1) \cdots f(x_n) = \prod_{i=1}^n f(x_i)。$$

若 $f(x)$ 屬於某一**參數族**(parametric family),即 $f(x)$ 有 $f(x|\theta)$ 的型式, θ 屬於某一集合(在2.1節我們稱此時有一**分佈族**),則 $X_1, \dots, X_n$ 之聯合p.d.f.為

$$(1.2) \quad f(x_1, \dots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)。$$

有時在做某一統計的探討時,由過去的經驗,或由一些合理的假設,我們知道 $f(x|\theta)$ 的型式,但參數 θ 未知,則 $X_1, \dots, X_n$ 之聯合p.d.f.便有(1.2)式所給的型式,只是 θ 未知。

例1.1 設 $X_1, \dots, X_n$ 為一組由 $\mathcal{E}(\lambda)$ 分佈所產生之隨機樣本。則 $X_1, \dots, X_n$ 之聯合p.d.f.為

$$f(x_1, \dots, x_n|\lambda) = \prod_{i=1}^n f(x_i|\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}。$$

$X_1, \dots, X_n$ 可以代表某房間內 n 個燈泡的壽命(單位可取為月)。則兩個月內燈泡皆不壞的機率為

4 第四章 隨機樣本及其他

$$\begin{aligned} & P(X_1 > 2, \dots, X_n > 2) \\ &= P(X_1 > 2) \cdots P(X_n > 2) \\ &= (P(X_1 > 2))^n = (e^{-2\lambda})^n = e^{-2n\lambda}。 \end{aligned}$$

此處用到因 $X_1, \dots, X_n$ 為獨立, 所以對 $\forall x_1, \dots, x_n$,

$$P(X_1 > x_1, \dots, X_n > x_n) = P(X_1 > x_1) \cdots P(X_n > x_n),$$

又因 $X_1, \dots, X_n$ 有相同的分佈, 故

$$P(X_1 > 2) = \cdots = P(X_n > 2)。$$

4.5節之例5.5與本例相關, 可互相對照。

是否存在不是隨機樣本的樣本呢? 答案是肯定的。

自一有 N 張紙牌(分別寫數字1至 N)的袋子中隨機地取紙牌, 每次取一張, 每次取出後放回 (with replacement), 令 $X_1, \dots, X_n$ 表取 n 次後所得之結果。則顯然此為一組隨機樣本, 以 $\mathcal{D}\text{-}\mathcal{U}(1, N)$ 為共同的分佈, p.d.f. 為

$$(1.3) \quad f(x) = P(X = x) = \frac{1}{N}, x = 1, \dots, N。$$

但若每次取出後不放回, 則所得之樣本 $Y_1, \dots, Y_n$ 就不是 i.i.d. 了, 因此非隨機樣本。雖然 $Y_1, \dots, Y_n$ 之邊際分佈仍相同, 還是 $\mathcal{D}\text{-}\mathcal{U}(1, N)$ (習題第2題), 但 $Y_1, \dots, Y_n$ 並不獨立。事實上可看出

$$\begin{aligned} (1.4) \quad P(Y_2 = y | Y_1 = x) &= \begin{cases} \frac{1}{N-1}, & y \neq x, \\ 0, & y = x, \end{cases} \\ &\neq P(Y_2 = y) = \frac{1}{N}, y = 1, \dots, N。 \end{aligned}$$

又對上述每次取出後不放回的模式, 樣本數 n 不能超過 N 。但對每次取出後放回的模式, 樣本數 n 可以任意大。前者母體數為有限, 後者母體數為無限。

自一有限的母體取樣, 每次取出後不放回, 稱為**簡單隨機取樣** (simple random sampling)。這種取樣方式, 在2.2.5節介紹超幾何分佈時, 我們便已引進。諸如民意調查, 便常以簡單隨機取樣的方式, 自母體(如某市的選民)中, 抽出要訪問的對象。要注意的是, 這種樣本不是如定義1.1所描述的隨機樣本(見下例)。在實際情況中, 若母體數 N 比樣本數 n 大很多, 則 $X_1, \dots, X_n$ 便近似於獨立, 亦即若視 $X_1, \dots, X_n$ 為獨立, 所產生的誤差並不會太大(可參考第三章例7.12)。

例1.2 設母體為 $\{1, 2, \dots, 1,000\}$, 即母體數 $N = 1,000$ 。隨機地取出10個樣本, 每次取出後不放回。則所得之樣本 $X_1, \dots, X_{10}$ 皆大於200之機率為何?
解. 令 Y 表 $X_1, \dots, X_{10}$ 中大於200之個數。則因母體中有800個數大於200, 故 Y 有 $\mathcal{H}(1,000, 800, 10)$ 分佈, 因此

$$P(Y = 10) = \frac{\binom{800}{10} \binom{200}{0}}{\binom{1,000}{10}} = \frac{\binom{800}{10}}{\binom{1,000}{10}} \doteq 0.106164。$$

次若視 $X_1, \dots, X_{10}$ 為獨立, 則

$$\begin{aligned} &P(X_1 > 200, \dots, X_{10} > 200) \\ &= P(X_1 > 200) \cdots P(X_{10} > 200) \\ &= (P(X > 200))^{10} = 0.8^{10} \doteq 0.107374。 \end{aligned}$$

誤差並不太大, 相對誤差只有百分之一左右:

$$\frac{0.107374 - 0.106164}{0.107374} \doteq 0.0113 = 1.13\%。$$

但若將 $X_1, \dots, X_{10}$ 視為獨立, 計算上顯然容易很多。

此後, 我們凡是提到隨機樣本, 除非特別聲明, 否則均指符合定義1.1的條件, 即樣本為i.i.d.。

取得隨機樣本後, 我們便要將樣本加以整理及分析, 這是我們下一章起要討論的。如1.3節所述, 在數學裡我們喜歡(或習慣)處理函數。諸如隨機變數、分佈函數、變數代換等, 都是與函數有關的內容。有一些與樣

本有關的函數我們特別有興趣。如**樣本平均** (sample mean 或說**樣本平均數**), 其定義為

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i,$$

及**樣本變異數** (sample variance), 其定義為

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2。$$

至於**樣本標準差** (sample standard deviation) 則為 $S_n = \sqrt{S_n^2}$ 。

有些統計學家喜愛以**樣本變異係數** (sample coefficient of variation) $S_n/\overline{X}_n$ 來量測散佈情況。為什麼標準差要除以樣本平均呢? 可以這樣想, 若考試平均成績為80分, 則標準差5分便不算太大; 但若平均成績為8分, 則標準差5分就算很大了。所以用 $S_n/\overline{X}_n$ 來度量散佈的大小似乎是合理的。當然這是對取正值隨機變數的解釋。若隨機變數可取正值及負值, 則 $\overline{X}_n$ 有可能為0(或很接近0), 此時 $S_n/\overline{X}_n$ 就不易解釋了。

如果樣本數 n 很清楚, 在不會混淆下, 有時可寫成 $\overline{X}$ 、 S^2 及 S 。又大家是否注意到, 既然是函數, 是否應寫成 $\overline{X}_n(X_1, \cdots, X_n)$, 及 $S_n^2(X_1, \cdots, X_n)$ 才較清楚? 習慣上我們省略掉 $(X_1, \cdots, X_n)$ 而只簡單地以 $\overline{X}_n, S_n^2$ 表示。又如同以小寫的 x 表示隨機變數 X 之觀測值, 通常也以小寫的 $\overline{x}_n, s_n^2$, 分別表 $\overline{X}_n$ 及 S_n^2 之觀測值。只是有時並不特別區分。例如, 可寫“得到 $\overline{X}_n = 3.2$ ”。

樣本平均就是隨機樣本的**算術平均** (arithmetic average), 因此大家可能較清楚。樣本變異數是度量樣本的變異, 分別將樣本減去樣本平均, 平方後相加, 那為什麼不除以 n 呢(相對於離散型分佈的變異數, 由於 $X_1, \cdots, X_n$ 可視為有相同的可能性發生, 所以每一 X_i 發生的機率應皆相同, 因此 $(X_i - \overline{X}_n)^2$ 似應各除以 n 再相加)? 這點我們稍後再說明。我們先導出一些關於樣本平均及樣本變異數之性質。

定理1.1 對 $x_1, \dots, x_n \in R$, 令 $\bar{x}_n = \sum_{i=1}^n x_i / n$, $s_n^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2 / (n - 1)$ 。則

$$(i) \min_a \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2;$$

$$(ii) (n-1)s_n^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \sum_{i=1}^n x_i^2 - n\bar{x}_n^2。$$

證明. (i) 首先

$$\begin{aligned} \sum_{i=1}^n (x_i - a)^2 &= \sum_{i=1}^n (x_i - \bar{x}_n + \bar{x}_n - a)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x}_n)^2 + 2 \sum_{i=1}^n (x_i - \bar{x}_n)(\bar{x}_n - a) + \sum_{i=1}^n (\bar{x}_n - a)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - a)^2。 \end{aligned}$$

此處用到

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x}_n)(\bar{x}_n - a) &= (\bar{x}_n - a) \sum_{i=1}^n (x_i - \bar{x}_n) \\ &= (\bar{x}_n - a) \left(\sum_{i=1}^n x_i - n\bar{x}_n \right) = 0。 \end{aligned}$$

因 $(\bar{x}_n - a)^2 \geq 0$, 故若且唯若 $a = \bar{x}_n$ 時, $\sum_{i=1}^n (x_i - a)^2$ 達到極小值。得證(i)。

(ii) 左側經展開後得

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x}_n)^2 &= \sum_{i=1}^n (x_i^2 - 2x_i\bar{x}_n + \bar{x}_n^2) \\ &= \sum_{i=1}^n x_i^2 - 2\bar{x}_n \sum_{i=1}^n x_i + n\bar{x}_n^2 \\ &= \sum_{i=1}^n x_i^2 - 2n\bar{x}_n^2 + n\bar{x}_n^2 \\ &= \sum_{i=1}^n x_i^2 - n\bar{x}_n^2。 \end{aligned}$$

證畢。

上述定理之(ii)不論在理論或計算上都很有用。有時表示成 $\sum_{i=1}^n x_i^2 - n\bar{x}_n^2$ 後, 對求 $(n-1)s_n^2$ 較容易。

隨機樣本由於是i.i.d.的隨機變數, 所以期望值及變異數都有簡潔的表示法。下述定理的證明留在習題中(第3題)。

定理1.2 設 $X_1, \dots, X_n$ 為一組隨機樣本, 函數 $g(x)$ 滿足 $E(g(X))$ 及 $\text{Var}(g(X))$ 皆存在。則

- (i) $E(\sum_{i=1}^n g(X_i)) = nE(g(X))$;
- (ii) $\text{Var}(\sum_{i=1}^n g(X_i)) = n\text{Var}(g(X))$ 。

特別地, 若取 $g(x) = x$, 則得

- (iii) $E(\sum_{i=1}^n X_i) = nE(X)$;
- (iv) $\text{Var}(\sum_{i=1}^n X_i) = n\text{Var}(X)$ 。

下述定理為定理1.2之應用。

定理1.3 設 $X_1, \dots, X_n$ 為一組隨機樣本, 且 $E(X) = \mu$, $\text{Var}(X) = \sigma^2$ 皆存在。則

- (i) $E(\bar{X}_n) = \mu$;
- (ii) $\text{Var}(\bar{X}_n) = \sigma^2/n$;
- (iii) $E(S_n^2) = \sigma^2$ 。

證明. (i) 在定理1.2 中, 取 $g(x) = x/n$, 則得

$$E(\bar{X}_n) = E\left(\sum_{i=1}^n \frac{X_i}{n}\right) = nE\left(\frac{X}{n}\right) = n \cdot \frac{1}{n}E(X) = E(X) = \mu。$$

(ii) 同理可得

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\sum_{i=1}^n \frac{1}{n}X_i\right) = n\text{Var}\left(\frac{1}{n}X\right) = n \cdot \frac{1}{n^2}\text{Var}(X) = \frac{1}{n}\text{Var}(X)。$$

(iii) 利用定理1.1 (ii),

$$E(S_n^2) = E\left(\frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\bar{X}_n^2\right)\right)$$

$$\begin{aligned}
&= \frac{1}{n-1}(nE(X^2) - nE(\overline{X}_n^2)) \\
&= \frac{1}{n-1}(n(\mu^2 + \sigma^2) - n(\frac{\sigma^2}{n} + \mu^2)) = \sigma^2。
\end{aligned}$$

證畢。

底下為定理1.3之一立即的推論。

系理1.1 承定理1.3。令

$$(1.5) \quad Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}。$$

則 $E(Z_n) = 0$, $\text{Var}(Z_n) = 1$ 。

在上述二定理中，如同一般的習慣，我們常以 X 代表一組隨機樣本 $X_1, \dots, X_n$ 中的某一個。

樣本平均的期望值等於母體的期望值，這符合我們的直觀，也是我們之所以偏好樣本平均的原因。樣本平均的變異數則隨著樣本數之增加而漸減。至於樣本變異數，如前所述，初見之下除以 $n-1$ 可能不是那麼有道理。現在我們知道了，若照一般求平均的習慣而除以 n ，則 $E(S_n^2)$ 便不等於 σ^2 。

樣本平均由於是 i.i.d. 隨機變數的平均，所以它有一些好的性質。如 $\overline{X}_n$ 之動差母函數與 X 的動差母函數有下述關係：

$$\begin{aligned}
(1.6) \quad M_{\overline{X}_n}(t) &= E(e^{t\overline{X}_n}) = E(e^{t(X_1 + \dots + X_n)/n}) \\
&= (E(e^{(t/n)X}))^n = (M_X(t/n))^n。
\end{aligned}$$

例1.3 (i) 設 $X_1, \dots, X_n$ 為由 $\mathcal{N}(\mu, \sigma^2)$ 分佈所產生之隨機樣本。則

$$M_{\overline{X}_n}(t) = (e^{\mu(t/n) + \sigma^2(t/n)^2/2})^n = e^{\mu t + (\sigma^2/n)t^2/2}。$$

因此 $\overline{X}_n$ 有 $\mathcal{N}(\mu, \sigma^2/n)$ 分佈。樣本平均仍有常態分佈，期望值不變，但變異數隨著 n 之增大而變小（見定理1.3）。由此即得 Z_n （定義見(1.5)式）有 $\mathcal{N}(0, 1)$ 分佈（見2.3.6節）。

(ii) 設 $X_1, \dots, X_n$ 為由 $\Gamma(\alpha, \beta)$ 分佈所產生之隨機樣本。則

$$M_{\overline{X}_n}(t) = \left(\frac{1}{(1 - \beta t/n)^\alpha} \right)^n = \frac{1}{(1 - (\beta/n)t)^{n\alpha}}。$$

即 $\overline{X}_n \sim \Gamma(n\alpha, \beta/n)$ 。換句話說, $\overline{X}_n$ 仍有 gamma 分佈, 且仍是

$$\begin{aligned} E(\overline{X}_n) &= n\alpha \cdot \beta/n = \alpha\beta = E(X), \\ \text{Var}(\overline{X}_n) &= n\alpha \cdot (\beta/n)^2 = \alpha\beta^2/n = \text{Var}(X)/n。 \end{aligned}$$

對二獨立的隨機變數 X, Y , 如果其動差母函數(或母函數等轉換), 並無簡單的型式, 則仍可以變數代換的方法, 求和 $Z = X + Y$ 之分佈。如此便引出底下之褶積公式 (convolution formula)。

若 X, Y 皆為取整數值之隨機變數, 則 $Z = X + Y$ 亦然。且對每一整數 z ,

$$\begin{aligned} (1.7) \quad f_Z(z) &= P(Z = z) \\ &= \sum_{x=-\infty}^{\infty} P(X = x, Y = z - x) \\ &= \sum_{x=-\infty}^{\infty} P(X = x)P(Y = z - x) \\ &= \sum_{x=-\infty}^{\infty} f_X(x)f_Y(z - x)。 \end{aligned}$$

上式最後一個和, 便稱為 f_X 與 f_Y 之褶積。故 Z 之 p.d.f. 為 X, Y 之 p.d.f. 的褶積。注意 $f_Z(z)$ 亦可寫成

$$(1.8) \quad f_Z(z) = \sum_{y=-\infty}^{\infty} f_X(z - y)f_Y(y)。$$

例1.4 設 X 與 Y 獨立, 且分別有 $\mathcal{B}(n, p)$ 及 $\mathcal{B}(m, p)$ 分佈。則和 $Z = X + Y$ 之 p.d.f. 為(注意: X, Y 皆只取非負整數值)

$$f_Z(z) = \sum_{x=-\infty}^{\infty} f_X(x)f_Y(z - x)$$

$$\begin{aligned}
&= \sum_{x=0}^z \binom{n}{x} p^x (1-p)^{n-x} \binom{m}{z-x} p^{z-x} (1-p)^{m-z+x} \\
&= p^z (1-p)^{m+n-z} \sum_{x=0}^z \binom{n}{x} \binom{m}{z-x} \\
&= \binom{m+n}{z} p^z (1-p)^{m+n-z}。
\end{aligned}$$

即得 Z 有 $\mathcal{B}(m+n, p)$ 分佈, 與第三章定理2.4所得相同。

若 X, Y 皆為連續型的隨機變數, 則 $Z = X + Y$ 之p.d.f.為

$$\begin{aligned}
(1.9) \quad f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx \\
&= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy, z \in R。
\end{aligned}$$

上述積分亦稱為 f_X 及 f_Y 之褶積。至於 Z 之分佈函數則為

$$\begin{aligned}
(1.10) \quad F_Z(z) &= P(X + Y \leq z) \\
&= \int_{-\infty}^{\infty} f_X(x) F_Y(z-x) dx \\
&= \int_{-\infty}^{\infty} f_Y(y) F_X(z-y) dy, z \in R。
\end{aligned}$$

若 X, Y 皆為取正值之隨機變數, 則

$$\begin{aligned}
(1.11) \quad f_Z(z) &= \int_0^z f_X(x) f_Y(z-x) dx \\
&= \int_0^z f_X(z-y) f_Y(y) dy, z \in R,
\end{aligned}$$

且

$$\begin{aligned}
(1.12) \quad F_Z(z) &= \int_0^z f_X(x) F_Y(z-x) dx \\
&= \int_0^z f_Y(y) F_X(z-y) dy, z \in R。
\end{aligned}$$

一般而言, 若以 F_X 及 F_Y 分別表隨機變數 X 及 Y 之分佈函數, 則 $Z = X + Y$ 之分佈函數可寫成 $F_Z = F_X * F_Y$, 稱為 F_X 與 F_Y 之褶積。又以

F_2 表 $F * F$, 以 F_n 表 $F * F_{n-1}, n \geq 2$ 。 F_n 稱為 F 與其自身之 n 層褶積。 F_n 即為 n 個i.i.d.隨機變數, 以 F 為其共同分佈函數, 其和之分佈函數。至於p.d.f.亦有褶積符號。例如, $f_Z = f_X * f_Y$ 。

例1.5 設 X, Y 為獨立的隨機變數, 皆有 $\mathcal{E}(\lambda)$ 分佈。則和 $Z = X + Y$ 之p.d.f.為

$$\begin{aligned} f_Z(z) &= \int_0^z \lambda e^{-\lambda x} \lambda e^{-\lambda(z-x)} dx \\ &= \int_0^z \lambda^2 e^{-\lambda z} dx \\ &= \lambda^2 z e^{-\lambda z}, z > 0. \end{aligned}$$

即 Z 有 $\Gamma(2, \lambda^{-1})$ 分佈。

我們再來介紹幾個定義。首先對兩組隨機樣本 $X_1, \dots, X_n$ 及 $Y_1, \dots, Y_n$, 令

$$\begin{aligned} S_{XY} &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}), \\ R_{XY} &= \frac{S_{XY}}{S_1 \cdot S_2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{((\sum_{i=1}^n (X_i - \bar{X})^2)(\sum_{i=1}^n (Y_i - \bar{Y})^2))^{1/2}}, \end{aligned}$$

分別表其樣本共變異數(sample covariance)及樣本相關係數(sample correlation coefficient), 後者亦可稱為樣本相關(sample correlation), 其中 S_1 及 S_2 分別為 $X_1, \dots, X_n$ 及 $Y_1, \dots, Y_n$ 之樣本標準差。可看出 $X_1, \dots, X_n$ 與 $X_1, \dots, X_n$ 之樣本共變異數, 即為 $X_1, \dots, X_n$ 之樣本變異數。 S_{XY} 及 R_{XY} 之定義, 可分別與隨機變數 X 與 Y 之共變異數及相關係數對照。

最後我們要指出, 一般而言, 在簡單隨機取樣裡(即取出後不放回), 對應的樣本平均之期望值及變異數, 與隨機取樣裡是不同的。我們以下例來說明。

例1.6 自整數 $1, \dots, N$ 中, 簡單隨機取樣 n 個數。設依序取出 $X_1, \dots, X_n$ 。令 $E(X) = \mu$, $\text{Var}(X) = \sigma^2$, $\bar{X}_n = (X_1 + \dots + X_n)/n$ 。試證

- (i) $X_1, \dots, X_n$ 之邊際分佈相同,
(ii) $\mu = \frac{N+1}{2}, \quad \sigma^2 = \frac{N^2-1}{12},$

$$E(\overline{X}_n) = \frac{N+1}{2}, \quad \text{Var}(\overline{X}_n) = \frac{(N+1)(N-n)}{12n} = \frac{N-n}{N-1} \frac{\sigma^2}{n}.$$

證明. (i) 由總機率法則,

$$\begin{aligned} P(X_{k+1} = x) &= \sum_{j_1, \dots, j_k \neq x} P(X_{k+1} = x | X_1 = j_1, \dots, X_k = j_k) \\ &\quad \cdot P(X_1 = j_1, \dots, X_k = j_k) \\ &= \sum_{j_1, \dots, j_k \neq x} \frac{1}{N-k} \cdot \frac{1}{N(N-1) \cdots (N-k+1)} \\ &= (N-1) \cdots (N-1-k+1) \cdot \frac{1}{N(N-1) \cdots (N-k)} \\ &= \frac{1}{N} = P(X_1 = x), k = 1, \dots, n-1, \end{aligned}$$

故得證 $X_1, \dots, X_n$ 有相同之邊際分佈。

(ii) 首先

$$\begin{aligned} \mu = E(X) &= \sum_{i=1}^N i \cdot \frac{1}{N} = \frac{N(N+1)}{2} \cdot \frac{1}{N} = \frac{N+1}{2}, \\ \sigma^2 = \text{Var}(X) &= E(X^2) - (E(X))^2 \\ &= \sum_{i=1}^N i^2 \cdot \frac{1}{N} - \left(\frac{N+1}{2}\right)^2 \\ &= \frac{N(N+1)(2N+1)}{6N} - \frac{(N+1)^2}{4} \\ &= \frac{N^2-1}{12}. \end{aligned}$$

故

$$E(\overline{X}_n) = E(X) = \frac{N+1}{2}.$$

其次利用第三章定理7.5,

$$\text{Var}(\overline{X}_n) = \frac{1}{n^2} (n \text{Var}(X) + n(n-1) \text{Cov}(X_1, X_2)).$$

而

$$\begin{aligned}
\text{Cov}(X_1, X_2) &= E(X_1 X_2) - (E(X))^2 \\
&= \sum_{i \neq j} ij \cdot \frac{1}{N(N-1)} - \left(\frac{N+1}{2}\right)^2 \\
&= \frac{1}{N(N-1)} \left((1 + \cdots + N)(1 + \cdots + N) - \sum_{i=1}^N i^2 \right) - \frac{(N+1)^2}{4} \\
&= \frac{1}{N(N-1)} \left(\left(\frac{N(N+1)}{2}\right)^2 - \frac{N(N+1)(2N+1)}{6} \right) - \frac{(N+1)^2}{4} \\
&= \frac{1}{N(N-1)} \frac{N(N+1)(3N(N+1) - 2(2N+1))}{12} - \frac{(N+1)^2}{4} \\
&= \frac{(N+1)(N-1)(3N+2)}{12(N-1)} - \frac{(N+1)^2}{4} \\
&= \frac{(N+1)(3N+2-3N-3)}{12} = -\frac{N+1}{12}。
\end{aligned}$$

故得

$$\begin{aligned}
\text{Var}(\bar{X}_n) &= \frac{1}{n^2} \left(\frac{n(N^2-1)}{12} - \frac{n(n-1)(N+1)}{12} \right) \\
&= \frac{(N+1)(N-n)}{12n} = \frac{N-n}{N-1} \frac{\sigma^2}{n}。
\end{aligned}$$

證畢。

如果是隨機取樣, 則 $E(\bar{X}_n) = (N+1)/2$, 而 $\text{Var}(\bar{X}_n) = \sigma^2/n$ 。故取出後放回與取出後不放回, 樣本平均不變, 但後者之樣本變異數較小, 為前者之 $(N-n)/(N-1)$ 倍。減小是合理的, 這點留給讀者自行討論。

一般民調裡的取樣, 往往屬於取出後不放回。但若 n 與 N 相比很大, 則 $(N-n)/(N-1)$ 很接近 1, 此時 $\text{Var}(\bar{X}_n)$ 近似於取出後放回的 $\bar{X}_n$ 之變異數 σ^2/n 。