110 年特種考試地方政府公務人員考試試題

等 別：三等考試
類 科：統計
科 目：抽樣方法

王瑋老師

一、何謂抽樣誤差？那種樣本可以測量抽樣誤差？有那些方法可以降低抽樣誤差？（10 分）

解題關鍵
1.《考題難易》：★☆☆☆☆ 2.《破題關鍵》：抽樣誤差為抽樣方法乃至於統計學的基本內容，最近在 105 年高考有考過解釋名詞，可參考王瑋 抽樣方法 P.1-13 頁相同範例。

【擬答】

㈠進行抽樣時，由於樣本只是從母體中抽取出來的部分的觀察值，因此由樣本所得到的統計量自不能完整描述出代表母群體性質的母數，會有某種程度的誤差的存在，這些誤差即為抽樣誤差。

㈡抽樣誤差不是由調查過程產生的失誤所引起，而是隨機抽樣所特有的誤差，因此須採用隨機樣本來測量誤差。

㈢欲降低抽樣誤差，可增加抽樣的樣本個案數，並且採用配合研究問題與母體特質的抽樣方法。

二、㈠何謂簡單隨機樣本？（5 分）

㈡考慮簡單隨機抽樣，請證明任一母體元素 $u_i, i=1,...,N$ 被選入樣本的機率為 n/N。（5 分）

㈢請問下列敘述是否正確？「若任一母體元素 $u_i, i=1,...,N$ 被選入樣本的機率皆相等，則此樣本稱為簡單隨機樣本」，若不正確，請舉一反例說明。（10 分）

解題關鍵
1.《考題難易》：★☆☆☆☆ 2.《破題關鍵》：簡單隨機抽樣任一母體元素被選入的機率是抽樣方法基本內容，在 92 年高考考過相同題目，可參考王瑋 抽樣方法 P.2-21 頁完全相同範例。

【擬答】

㈠從母體中抽出若干樣本，其每一單位或元素均有同等被抽出的機會，這種抽樣程序即稱為簡單隨機抽樣法，其抽出的樣本稱為簡單隨機樣本。

㈡若為取後放回抽樣，從母體 N 個元素中抽出 n 個，機率為 $\dfrac{n}{N}$

若為取後不放回抽樣，從母體 N 個元素中抽出 n 個，機率為

$$\frac{C_{n-1}^{N-1}}{C_n^N}=\frac{\dfrac{(N-1)!}{(n-1)!(N-n)!}}{\dfrac{N!}{n!(N-n)!}}=\frac{n}{N}$$

不論取後放回或不放回抽樣，任一母體元素 μ_i 被選入樣本的機率皆為 $\dfrac{n}{N}$

㈢若任一母體元素被選入樣本的機率皆相等，未必一定是簡單隨機樣本。例如考慮群集隨機抽樣，並利用依大小成比例的機率抽樣法抽樣，即比例機率抽樣法，此時任一個母體元素被選入樣本的機率皆相等，但此樣本並非簡單隨機樣本。

三、㈠抽樣時，採分層隨機抽樣方法而不採用簡單隨機抽樣方法的原因有那些？（10 分）

㈡考慮分層隨機抽樣，證明在抽樣總成本固定之下，使樣本平均之變異數最小的各層樣本數最佳配置為 $n_i = n\left(\dfrac{N_i\sigma_i/\sqrt{c_i}}{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{c_k}}\right)$，其中 N_i 是第 i 分層的大小，σ_i^2 是第 i 分層的變異數，c_i 是由第 i 分層獲得一觀察值的成本。（15 分）

㈢考慮分層隨機抽樣，請問在那一種樣本配置下，母體平均估計量與簡單隨機抽樣時的母體平均估計量相同，試證明之。（10 分）

解題關鍵
1.《考題難易》：★★☆☆☆
2.《破題關鍵》：分層隨機抽樣樣本最適配置是考試重要內容，但過去大多以計算題為主，此證明雖僅在 86 年、88 年以及 92 年高考有命題，但此證明為基本內容，應可輕鬆拿分。可參考王瑋 抽樣方法 P.3-67 至 P.3-68 頁內容說明。

【擬答】

㈠

1. 相較於大小相同的簡單隨機樣本而言，分層可產生較小的估計值誤差界限，尤其當層內同質性越高時。
2. 透過將母體元素分層為數個同質性較高的群體或調查上較為接近的族群，調查中每一個觀察值的成本將隨之降低，管理上也較為便利。

㈡抽樣總成本固定之下，期望最小 $\min\ C\cdot V$

其中總成本函數 $C=\sum_{k=1}^{L} c_k n_k$，在此僅考慮各層成本 $c_h n_h$ 總和；

分層變異數 $Var(\bar{y}_{st})=\sum_{k=1}^{L}W_k^2(1-f_k)\frac{\sigma_k^2}{n_k}=\frac{1}{N^2}\sum_{k=1}^{L}\frac{N_k^2\sigma_k^2}{n_k}-\frac{1}{N^2}\sum_{k=1}^{L}N_k\sigma_k^2$

在此僅關心各分層的樣本大小，所以變異數的部分只考慮 $V=\sum_{k=1}^{L}\frac{N_k^2\sigma_k^2}{n_k}$

$$C\cdot V=(\sum_{k=1}^{L}c_kn_k)\cdot(\sum_{k=1}^{L}\frac{N_k^2\sigma_h^2}{n_k})$$

$$=(c_1n_1+c_2n_2+\cdots+c_Ln_L)\left(\frac{N_1^2\sigma_1^2}{n_1}+\frac{N_2^2\sigma_2^2}{n_2}+\cdots+\frac{N_L^2\sigma_L^2}{n_L}\right)$$

根據科西不等式

$$(c_1n_1+c_2n_2+\cdots+c_Ln_L)\left(\frac{N_1^2\sigma_1^2}{n_1}+\frac{N_2^2\sigma_2^2}{n_2}+\cdots+\frac{N_L^2\sigma_L^2}{n_L}\right)$$

$$\geq\left(\sqrt{c_1n_1}\cdot\frac{N_1\sigma_1}{\sqrt{n_1}}+\sqrt{c_2n_2}\cdot\frac{N_2\sigma_2}{\sqrt{n_2}}+\cdots+c_Ln_L\cdot\frac{N_L\sigma_L}{\sqrt{n_L}}\right)^2$$

極值出現在 $\frac{\sqrt{c_1n_1}}{\frac{N_1\sigma_1}{\sqrt{n_1}}}=\frac{\sqrt{c_2n_2}}{\frac{N_2\sigma_2}{\sqrt{n_2}}}=\cdots=\frac{\sqrt{c_Ln_L}}{\frac{N_L\sigma_L}{\sqrt{n_L}}}\Rightarrow\frac{n_1}{\frac{N_1\sigma_1}{\sqrt{c_1}}}=\frac{n_2}{\frac{N_2\sigma_2}{\sqrt{c_2}}}=\cdots=\frac{n_L}{\frac{N_L\sigma_L}{\sqrt{c_L}}}$

$$\Rightarrow\frac{n_i}{\frac{N_i\sigma_i}{\sqrt{c_i}}}=\frac{n_1+n_2+\cdots+n_L}{\frac{N_1\sigma_1}{\sqrt{c_1}}+\frac{N_2\sigma_2}{\sqrt{c_2}}+\cdots+\frac{N_L\sigma_L}{\sqrt{c_L}}}=\frac{n}{\sum_{i=1}^{L}\frac{N_i\sigma_i}{\sqrt{c_i}}}$$

可得 $n_i=\frac{N_i\sigma_i/\sqrt{c_i}}{\sum_{k=1}^{L}N_k\sigma_k/\sqrt{c_k}}\times n$

(三)在比例配置下，不考慮每一層內成本 c_i 與變異 σ_i^2 差異，即假設 $c_1=c_2=\cdots=c_L$ 與 $\sigma_1^2=\sigma_2^2=\cdots=\sigma_L^2$，

因此樣本數配置 $n_i=\frac{N_i}{\sum_{i=1}^{L}N_i}\times n\Rightarrow w_i=\frac{N_i}{\sum_{i=1}^{L}N_i}=W_i$

分層隨機抽樣母體平均數估計值為 $\bar{y}_{st}=\sum W_i\bar{y}_i$

簡單隨機抽樣母體平均數估計值為 $\bar{y}=\sum w_i\bar{y}_i$

所以比例配置下，$w_i=W_i$，則分層隨機抽樣母體平均數估計值與簡單隨機抽樣估計量相同。

四、考慮系統抽樣，請導出 $\rho=\dfrac{(k-1)nMSB-SST}{(n-1)SST}$，其中 $N=nk$，ρ 是系統樣本（集群）內任 2 元素的相關係數，$MSB=\dfrac{n}{k-1}\sum_{i=1}^{k}(\bar{y}_i-\bar{\bar{y}})^2$，$MSW=\dfrac{n}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2$，$SST=\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2$（10 分）

解題關鍵
1.《考題難易》：★★★★★ 2.《破題關鍵》：系統樣本內任 2 元素的相關係數為過去未曾命題的內容，若能清楚此相關係數的定義下，後續才能透過系統抽樣變異數的展開，若無事先準備此內容，一般考生在考場中無法輕易導出結果。

【擬答】

系統樣本內任 2 元素的相關係數為 $\rho=\dfrac{E(y_{ij}-\bar{\bar{y}})(y_{iu}-\bar{\bar{y}})}{E(y_{ij}-\bar{\bar{y}})^2}$

其中取後不放回共有 $n(n-1)$ 個不同的雙變量，而 k 個系統樣本中，則有 $kn(n-1)$ 個不同的雙變量，故 $E(y_{ij}-\bar{\bar{y}})(y_{iu}-\bar{\bar{y}})=\dfrac{\sum_{i=1}^{k}\sum_{j\neq u}^{n}(y_{ij}-\bar{\bar{y}})(y_{iu}-\bar{\bar{y}})}{kn(n-1)}$

而 $E(y_{ij}-\bar{\bar{y}})^2=\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2}{N}=\dfrac{SST}{nk}$

所以 $\rho=\dfrac{E(y_{ij}-\bar{\bar{y}})(y_{iu}-\bar{\bar{y}})}{E(y_{ij}-\bar{\bar{y}})^2}=\dfrac{\sum_{i=1}^{k}\sum_{j\neq u}^{n}(y_{ij}-\bar{\bar{y}})(y_{iu}-\bar{\bar{y}})}{(n-1)SST}$

又
$$
\begin{aligned}
Var(\bar{y}_{sy})&=\frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i-\bar{\bar{y}})^2\\
&=\frac{1}{k}\sum_{i=1}^{k}\left[\frac{1}{n}\sum_{j=1}^{n}y_{ij}-\bar{\bar{y}}\right]^2\\
&=\frac{1}{kn^2}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})\right]^2\\
&=\frac{1}{kn^2}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2+\sum_{j\neq u}^{n}(y_{ij}-\bar{\bar{y}})(y_{iu}-\bar{\bar{y}})\right]
\end{aligned}
$$

$$\Rightarrow \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i-\bar{\bar{y}})^2=\frac{1}{kn^2}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2+\sum_{i=1}^{k}\sum_{j\neq u}^{n}(y_{ij}-\bar{\bar{y}})(y_{iu}-\bar{\bar{y}})\right]$$

$$\Rightarrow n^2\sum_{i=1}^{k}(\bar{y}_i-\bar{\bar{y}})^2=\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2+\sum_{i=1}^{k}\sum_{j\neq u}^{n}(y_{ij}-\bar{\bar{y}})(y_{iu}-\bar{\bar{y}})$$

$$\Rightarrow n(k-1)MSB=SST+\sum_{i=1}^{k}\sum_{j\neq u}^{n}(y_{ij}-\bar{\bar{y}})(y_{iu}-\bar{\bar{y}})$$

所以 $\sum_{i=1}^{k}\sum_{j\neq u}^{n}(y_{ij}-\bar{\bar{y}})(y_{iu}-\bar{\bar{y}})=n(k-1)MSB-SST$

因此 $\rho=\dfrac{\sum_{i=1}^{k}\sum_{j\neq u}^{n}(y_{ij}-\bar{\bar{y}})(y_{iu}-\bar{\bar{y}})}{(n-1)SST}=\dfrac{(k-1)nMSB-SST}{(n-1)SST}$

五、㈠考慮集群抽樣，若母體總數 M 已知，請問母體總和的估計量為何？若不知母體總數 M，但知道集群總數 N 時，請問母體總和的估計量為何？（5 分）

㈡考慮兩階段集群抽樣，證明母體平均估計量 $\hat{\mu}=\frac{1}{\bar{M}}\frac{\sum_{i=1}^{n}M_i\bar{y}_i}{n}$ 是母體平均 μ 的不偏估計量。其中 $\bar{M}=M/N$。提示：$E(\hat{\mu})=E_1E_{2|1}(\hat{\mu})$（10 分）

㈢考慮兩階段集群抽樣，由相等大小集群 M 抽取相等大小樣本 m，且當 N 很大時，證明在固定抽樣成本下，使 $V(\hat{\mu})=\frac{1}{n}(\sigma_b^2+\frac{\sigma_w^2}{m})$ 值最小的 m 為 $m=\sqrt{\frac{\sigma_w^2c_1}{\sigma_b^2c_2}}$，其中 c_1 是第一階段每一觀察值的成本，c_2 是第二階段每一觀察值的成本，σ_b^2 是集群平均間的變異數，σ_w^2 是集群內元素間的變異數。（10 分）

解題關鍵
1.《考題難易》：★★★★☆
2.《破題關鍵》：題(1)比較比率群集與簡單群集使用上的情境差異，這是近年常考的問題，如 108 年高考、108 年地特皆有相關命題，可參考王瑋 抽樣方法 P.12-25 與 P.12-32 頁相同例題演練。兩階段集群抽樣不偏性證明已強調有機會再次命題，過去在 101 年關務考過，可參考王瑋 抽樣方法 P.6-12 頁相同例題演練。考慮相等群集大小的兩階段集群抽樣與其樣本數的問題，在抽樣方法題庫班中有詳細介紹，但這樣的純證明題亦是需要考前多次演練過，一般考生不易取分。抽樣方法題庫班 P.34 頁完全相同證明演練。

【擬答】

㈠當母體總數 M 已知，可採用比率集群估計法，透過各個集群的大小作為輔助變數，來提升推估集群母體總和的準確度。母體總和的估計量為 $\hat{Y}_{rc}=M\bar{y}_{rc}=M\times\frac{\bar{y}_t}{\bar{M}}\xrightarrow{估計}Y$

當不知母體總數 M，但知道集群總數 N ，可採用簡單集群抽樣法來估計母體總和，母體總和的估計量為 $\hat{Y}_{cl} = N\bar{y}_t = N \times \frac{\sum Y_i}{n} \xrightarrow{\text{估計}} Y$

(二) $E(\hat{\mu}) = E_1\left(E_{2|1}\left(\frac{1}{\bar{M}}\frac{\sum M_i \bar{y}_i}{n}\right)\right) = E_1\left(\frac{1}{\bar{M}}E_{2|1}\left(\frac{\sum M_i \bar{y}_i}{n}\right)\right)$

$= E_1\left(\frac{1}{\bar{M}}\frac{\sum M_i \bar{Y}_i}{n}\right)$

$= E_1\left(\frac{N}{M}\hat{Y}\right) = \frac{N}{M}E_1(\hat{Y})$

$= \frac{N}{M}\bar{Y} = \frac{Y}{M} = \mu$ 為不偏估計式

(三)首先導出相等大小集群的兩階段母體平均數估計

假設 $M_1 = M_2 = \ldots = M_N = M_0$ 且 $m_1 = m_2 = \ldots = m_n = m$

二階段平均數估計式為 $\bar{y}_{cl} = \frac{N}{Mn}\sum M_i \bar{y}_i \xrightarrow{\text{估計}} \bar{\bar{Y}}$

其中 $\bar{y}_{cl} = \frac{N}{Mn}\sum M_i \bar{y}_i = \frac{N}{NM_0 n}\sum M_0 \bar{y}_i = \frac{1}{n}\sum \bar{y}_i$

兩階段母體平均數之變異數為

$$Var(\bar{y}_{cl}) = \frac{1}{M^2}\left(N^2(1-f_1)\frac{S_{1b}^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i^2(1-f_{2i})\frac{S_{2i}^2}{m_i}\right)$$

其中 $f_1 = \frac{n}{N}$ ， $f_{2i} = \frac{m_i}{M_i}$ ， $S_{1b}^2 = \frac{\sum (Y_i - \bar{Y})^2}{N-1}$ ， $S_{2i}^2 = \frac{\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)^2}{M_i - 1}$

所以當 $M_1 = M_2 = \ldots = M_N = M_0$ 且 $m_1 = m_2 = \ldots = m_n = m$ 時，

$f_{2i} = \frac{m}{M_0} = f_2$ ， $\bar{Y} = M_0\bar{\bar{Y}}$ ，

$$S_{1b}^2 = \frac{\sum (Y_i - \bar{Y})^2}{N-1} = \frac{\sum (M_0\bar{Y}_i - M_0\bar{\bar{Y}})^2}{N-1} = M_0^2\frac{\sum(\bar{Y}_i - \bar{\bar{Y}})^2}{N-1} = M_0^2 S_b^2$$

又 $S_{2i}^2 = \frac{\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)^2}{M_i - 1}$ ， $\sigma_w^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_0}(y_{ij} - \bar{Y}_i)^2}{N(M_0 - 1)} = \frac{\sum_{i=1}^{N} S_{2i}^2}{N}$

故 $\frac{N}{n}\sum_{i=1}^{N} M_i^2(1-f_{2i})\frac{S_{2i}^2}{m_i} = \frac{N}{n}\frac{M_0^2}{m}(1-f_2)\sum_{i=1}^{n} S_{2i}^2 = N^2 M_0^2(1-f_2)\frac{\sigma_w^2}{nm}$

$$Var(\hat{\mu}) = \frac{1}{N^2 M_0^2}\left(N^2(1-f_1)\frac{M_0^2 S_1^2}{n} + N^2 M_0^2(1-f_2)\frac{S_2^2}{nm}\right)$$

$$= (1-f_1)\frac{S_1^2}{n} + (1-f_2)\frac{S_2^2}{nm}$$

當 N 很大， $f_1 = \frac{n}{N} \approx 0$ 且 $f_2 = \frac{m_0}{M_0} \approx 0$ ，

$$Var(\hat{\mu}) \approx \frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{nm} = \frac{1}{n}(\sigma_b^2 + \frac{\sigma_w^2}{m})$$

考慮總成本函數為 $C = c_0 + c_1 n + c_2 nm$ ，為固定成本 c_0 加上第一階段成本 c_1 與第二階段成本 c_2 的總和。故僅考慮影響樣本數的變動成本

$C^* = c_1 n + c_2 nm = n(c_1 + c_2 m)$

而兩階段相等大小集群母體平均數之變異數在此只考慮與樣本數有關的部分，

$$V^* = \frac{1}{n}(\sigma_b^2 + \frac{\sigma_w^2}{m})$$

故需求解 $\min C^* \cdot V^*$

$$C^* \cdot V^* = n(c_1 + c_2 m) \cdot \frac{1}{n}(\sigma_b^2 + \frac{\sigma_w^2}{m}) = (c_1 + c_2 m) \cdot (\sigma_b^2 + \frac{\sigma_w^2}{m})$$

利用柯西不等式可知，最小值等號成立於

$$\frac{\sqrt{c_1}}{\sigma_b} = \frac{\sqrt{c_2 m}}{\frac{\sigma_w}{\sqrt{m}}} \Rightarrow m = \frac{\sigma_w \sqrt{c_1}}{\sigma_b \sqrt{c_2}} = \frac{\sqrt{\sigma_w^2 c_1}}{\sqrt{\sigma_b^2 c_2}}$$