《抽樣方法》

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）

試題評析	本題是考抽樣誤差之基礎觀念，考古題中也曾出現，講義中也有清楚列示，獲得滿分不難。
考點命中	《高點・高上抽樣方法講義》第一回，趙治勳編撰，頁5。

答：

(一)抽樣調查中只是以一部份代表全母體，誤差是不可避免的，這類因為以樣本推論母體所產生之誤差稱為抽樣誤差。而抽樣誤差與非抽樣誤差最大之不同就是抽樣誤差是可以利用公式計算誤差大小並且可以控制的。

(二)只要利用機率抽樣法所產生之樣本是可以測量抽樣誤差的。

(三)降低抽樣誤差方法有二：

1.增加樣本數

2.選擇適當之抽樣方法，儘可能讓樣本結構接近母體，即「樣本與母體之相似度要高，以機率理論的角度來說即表示抽到明顯有偏差的樣本之機率要越小越好」。

二、(一)何謂簡單隨機樣本？（5分）

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

(三)請問下列敘述是否正確？「若任一母體元素 u_i，$i=1,\ldots,N$ 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）

試題評析	本題是考簡單隨機抽樣之基礎觀念，第(二)小題92年高考考題中也曾出現過，第(三)小題課堂中也有舉例說明，獲得滿分不難。
考點命中	1.《高點・高上抽樣方法講義》第一回，趙治勳編撰，頁10。 2.《抽樣方法申論題完全制霸》，高點文化出版，趙治勳編著，頁1-6範例7。

答：

(一)滿足每一組可能的樣本均有相等被抽出的機率，即為簡單隨機抽樣法，利用簡單隨機抽樣法所得之樣本稱為簡單隨機樣本。

(二)P(母體元素 u_i 被選中) $= n\times\dfrac{1}{N}=\dfrac{n}{N}$，

其中 $\dfrac{1}{N}$ 為母體N個元素中第i個元素 u_i 被選中之機率，

n 為第i個元素 u_i 可以為n個樣本中之任何一個。

(三)敘述並不正確。

舉反例： 母體(N=3)：u_1,u_2,u_3

以抽後放回且不考慮順序隨機抽取兩個樣本(n=2)

樣本	每一組可能樣本被抽中機率
u_1,u_1	$1/9$
u_2,u_2	$1/9$
u_3,u_3	$1/9$
u_1,u_2	$2/9$
u_1,u_3	$2/9$
u_2,u_3	$2/9$

由上表可得，每一組可能樣本被抽中機率並不相等。

三、(一)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)

(二)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)

(三)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柚樣時的母體平均估計量相同，試證明之。(10分)

試題評析	本題是考分層隨機抽樣之基礎觀念與公式證明，第(二)、(三)小題分別在88年高考與87年高考中曾出現過類似的考題，這些考題都有收錄在講義中，只要考生有認真練習，獲得滿分不難。
考點命中	1.《高點・高上抽樣方法講義》第一回，趙治勳編撰，頁7。 2.《高點・高上抽樣方法》補充講義(1)，趙治勳編撰，頁18補充19。 3.《抽樣方法申論題完全制霸》，高點文化出版，趙治勳編著，頁4-5範例4、頁4-7範例6。

答：

(一)分層隨機抽樣法是設計抽樣法之一種，採用設計抽樣法之原因是為了調整母體型態，方便資料蒐集及減少抽到明顯有偏差的樣本，可以有效降低抽樣誤差。

(二)假設總調查費用為 $C = c_0 + \sum_{i=1}^{L} c_i n_i$　　其中 c_0 表示固定成本

$$V(\bar{y}_{st}) = \sum_{i=1}^{L} W_i^2 (1-f_i)\frac{\sigma_i^2}{n_i} = \sum_{i=1}^{L}\frac{N_i^2}{N^2}\left(\frac{N_i-n_i}{N_i}\right)\frac{\sigma_i^2}{n_i} = \frac{1}{N^2}\sum_{i=1}^{L}\frac{(N_i\sigma_i)^2}{n_i} - \frac{1}{N^2}\sum_{i=1}^{L} N_i\sigma_i^2$$

其中只有一項 $\sum_{i=1}^{L}\frac{(N_i\sigma_i)^2}{n_i}$ 與 n_i 有關，令 $V^* = \sum_{i=1}^{L}\frac{(N_i\sigma_i)^2}{n_i}$

令 $C^* = C - c_0 = \sum_{i=1}^{L} c_i n_i$ 及 $V = V(\bar{y}_{st})$

目標: $\min_{n_i} C\times V = \min_{n_i} C^*\times V^* = \min_{n_i}\left(\sum_{i=1}^{L} c_i n_i\right)\left(\sum_{i=1}^{L}\frac{(N_i\sigma_i)^2}{n_i}\right)$

由柯西不等式可得，

$$\left(\sum_{i=1}^{L} c_i n_i\right)\left(\sum_{i=1}^{L}\frac{(N_i\sigma_i)^2}{n_i}\right) = \left[\sum_{i=1}^{L}(\sqrt{c_i n_i})^2\right]\left[\sum_{i=1}^{L}\left(\frac{N_i\sigma_i}{\sqrt{n_i}}\right)^2\right] \ge \left(\sum_{i=1}^{L}\sqrt{c_i}N_i\sigma_i\right)^2$$

上式等號成立，當 $\dfrac{\sqrt{c_i n_i}}{\dfrac{N_i\sigma_i}{\sqrt{n_i}}}=\dfrac{n_i}{\dfrac{N_i\sigma_i}{\sqrt{c_i}}}=\dfrac{n}{\sum_{i=1}^{L}\dfrac{N_i\sigma_i}{\sqrt{c_i}}}, i=1,2,\cdots,L$ 時，這時 $C^*\times V^*$ 具有最小值，可得

$$n_i=n\left(\frac{\frac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{L}\frac{N_i\sigma_i}{\sqrt{c_i}}}\right)=n\left(\frac{\frac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{k=1}^{L}\frac{N_k\sigma_k}{\sqrt{c_k}}}\right)$$

(三)比例配置下兩者之母體平均數估計量會相同。

證明：比例配置法下，樣本配置權數 $w_i=\dfrac{n_i}{n}=\dfrac{N_i}{N}=W_i$

$$\bar{y}_{st}=\sum_{i=1}^{L}W_i\bar{y}_i=\sum_{i=1}^{L}w_i\bar{y}_i=\sum_{i=1}^{L}\frac{n_i}{n}\bar{y}_i=\frac{\sum_{i=1}^{L}n_i\bar{y}_i}{n}=\bar{y}$$

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，

$$MSB=\frac{n}{k-1}\sum_{i=1}^{k}(\bar{y}_i-\bar{\bar{y}})^2$$

$$MSW=\frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_j-\bar{y}_i)^2$$

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

（10分）

試題評析	本題是考系統抽樣法中系統樣本內任意兩個元素相關係數之證明題，這類證明不曾出現在考古題中，講義中也只有估計量變異數之證明，考生需要在考場中利用估計量變異數證明之討論方式(想成ANOVA)套用在相關係數之計算上，本考題要拿高分實屬不易。
考點命中	《高點・高上抽樣方法》補充講義(1)，趙治勳編撰，頁15補充17相關。

答：

由於系統抽樣之母體資料表格與一因子 k 水準變異數分析的一樣，故以變異數分析之原理討論（註：以下符號都用考題的）

第 i 組可能系統樣本	樣本資料	
1	$y_{11},y_{12},y_{13},\cdots,y_{1n}$ $(Y_{11},Y_{12},Y_{13},\cdots,Y_{1n})$	$\bar{y}_1$ $(\bar{Y}_1)$
2	$y_{21},y_{22},y_{23},\cdots,y_{2n}$ $(Y_{21},Y_{22},Y_{23},\cdots,Y_{2n})$	$\bar{y}_2$ $(\bar{Y}_2)$
⋮	⋮	
k	$y_{k1},y_{k2},y_{k3},\cdots,y_{kn}$ $(Y_{k1},Y_{k2},Y_{k3},\cdots,Y_{kn})$	$\bar{y}_k$ $(\bar{Y}_k)$
		$\bar{\bar{y}}$ $(\bar{Y})$

$$SST=\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2 \text{ ，} SSR=\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_i-\bar{\bar{y}})^2=n\sum_{i=1}^{k}(\bar{y}_i-\bar{\bar{y}})^2$$

$$MSB=\frac{n}{k-1}\sum_{i=1}^{k}(\bar{y}_i-\bar{\bar{y}})^2=\frac{n\sum_{i=1}^{k}(\bar{y}_i-\bar{\bar{y}})^2}{k-1}=\frac{SSR}{k-1} \qquad \Rightarrow SSR=(k-1)MSB$$

題目所要求：$\rho=\dfrac{E[(y_{ij}-\bar{\bar{y}})(y_{ik}-\bar{\bar{y}})]}{\sqrt{E(y_{ij}-\bar{\bar{y}})^2}\sqrt{E(y_{ik}-\bar{\bar{y}})^2}}=\dfrac{E[(y_{ij}-\bar{\bar{y}})(y_{ik}-\bar{\bar{y}})]}{E(y_{ij}-\bar{\bar{y}})^2}$

$$=\frac{\dfrac{(k-1)nMSB-SST}{kn(n-1)}}{\dfrac{SST}{nk}}=\frac{(k-1)nMSB-SST}{(n-1)SST} \qquad \text{得證}$$

其中　分母部份：$E(y_{ij}-\bar{\bar{y}})^2=\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2}{nk}=\dfrac{SST}{nk}$

分子部份：

$$E[(y_{ij}-\bar{\bar{y}})(y_{ik}-\bar{\bar{y}})]=\frac{\sum_{i=1}^{k}\sum_{j<k}^{n}(y_{ij}-\bar{\bar{y}})(y_{ik}-\bar{\bar{y}})}{k\binom{n}{2}}=\frac{\sum_{i=1}^{k}\sum_{j<k}^{n}(y_{ij}-\bar{\bar{y}})(y_{ik}-\bar{\bar{y}})}{\dfrac{kn(n-1)}{2}}$$

$$=\frac{2\sum_{i=1}^{k}\sum_{j<k}^{n}(y_{ij}-\bar{\bar{y}})(y_{ik}-\bar{\bar{y}})}{kn(n-1)}\overset{\text{<註>}}{=}\frac{(k-1)nMSB-SST}{kn(n-1)}$$

<註> $SSR=\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_i-\bar{\bar{y}})^2=n\sum_{i=1}^{k}(\bar{y}_i-\bar{\bar{y}})^2=n\sum_{i=1}^{k}[\dfrac{\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})}{n}]^2=\dfrac{1}{n}\sum_{i=1}^{k}[\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})]^2$

$$=\frac{1}{n}\sum_{i=1}^{k}[\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2+2\sum_{j<k}^{n}(y_{ij}-\bar{\bar{y}})(y_{ik}-\bar{\bar{y}})]$$

$$=\frac{1}{n}[\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2+2\sum_{i=1}^{k}\sum_{j<k}^{n}(y_{ij}-\bar{\bar{y}})(y_{ik}-\bar{\bar{y}})]$$

$$=\frac{1}{n}[SST+2\sum_{i=1}^{k}\sum_{j<k}^{n}(y_{ij}-\bar{\bar{y}})(y_{ik}-\bar{\bar{y}})]$$

可得　$2\sum_{i=1}^{k}\sum_{j<k}^{n}(y_{ij}-\bar{\bar{y}})(y_{ik}-\bar{\bar{y}})=(k-1)nMSB-SST$

五、(一)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）

(二)考慮兩階段集群抽樣，證明母體平均估計量$\mu=\dfrac{1}{\bar{M}}\dfrac{\sum_{i=1}^{n}M_i\bar{y}_i}{n}$是母體平均$\mu$的不偏估計量。

其中$\bar{M}=M/N$。提示：$E(\mu)=E_1E_{2|1}(\mu)$（10分）

(三)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 $V(\mu)=\frac{1}{n}\left(\sigma_b^2+\frac{\sigma_w^2}{m}\right)$ 值最小的m為 $m=\sqrt{\frac{\sigma_w^2 c_1}{\sigma_b^2 c_2}}$，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）

試題評析	本題是考簡單估計法與比率估計法下群集抽樣法及兩階段抽樣法之證明題，第(一)小題需要猜想出題老師之想法才能夠寫出正確答案。第(二)小題兩階段抽樣法下估計量之不偏性證明，於101年關務有考過一次，課堂上也有說明過。第(三)小題利用固定總成本及最小化變異數下得到樣本數之證明題，這是第一次命題，但考生可以參照分層抽樣法下樣本數之證明過程回答本題，有充分準備證明之考生於本題獲得高分不難。
考點命中	1.《高點・高上抽樣方法講義》第一回，趙治勳編撰，頁51、頁77。 2.《高點・高上抽樣方法》補充講義(1)，趙治勳編撰，頁52補充53。

答：

(一)在母體總和值之估計問題上，簡單估計法是用不到M，反之比率估計法會用到M，所以猜想出題老師希望考生比較兩種估計法下估計量之差異。

若母體總數M已知，母體總和值之估計量為 $\hat{Y}_{rcl}=M\overline{y}_{rcl}=M\frac{\overline{y}_t}{\overline{m}}$

若母體總數M未知，母體總和值之估計量為 $\hat{Y}_{cl}=M\overline{y}_{cl}=N\overline{y}_t$

其中 $\overline{y}_t=\frac{1}{n}\sum_{i=1}^{n}\overline{y}_i$, $\overline{m}=\frac{1}{n}\sum_{i=1}^{n}M_i$

(二) $E(\hat{\mu})=E(\frac{1}{\overline{M}}\frac{\sum_{i=1}^{n}M_i\overline{y}_i}{n})=E_1[E_2(\frac{1}{\overline{M}}\frac{\sum_{i=1}^{n}M_i\overline{y}_i}{n})]=E_1[\frac{1}{\overline{M}n}\sum_{i=1}^{n}M_iE_2(\overline{y}_i)]$

$=E_1(\frac{1}{\overline{M}n}\sum_{i=1}^{n}M_i\frac{Y_i}{M_i})=E_1(\frac{1}{\overline{M}}\frac{\sum_{i=1}^{n}Y_i}{n})=\frac{1}{\overline{M}}E_1(\overline{y}_{1A})=\frac{1}{\overline{M}}N\overline{Y}=\frac{Y}{\overline{M}}=\mu$

(三)假設總調查費用為 $C=c_0+c_1n+c_2nm$，其中 c_0 表示固定成本，

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

令 $C^*=C-c_0=c_1n+c_2nm$ 及 $V=V(\hat{\mu})$

目標: $\min_m C\times V=\min_m C^*\times V=\min_m(c_1n+c_2nm)(\frac{1}{n}(\sigma_b^2+\frac{\sigma_w^2}{m}))=\min_m(c_1+c_2m)(\sigma_b^2+\frac{\sigma_w^2}{m})$

由柯西不等式可得，

$(c_1+c_2m)(\sigma_b^2+\frac{\sigma_w^2}{m})=[(\sqrt{c_1})^2+(\sqrt{c_2m})^2][(\sigma_b)^2+(\frac{\sigma_w}{\sqrt{m}})^2]\geq(\sqrt{c_1}\sigma_b+\sqrt{c_2}\sigma_w)^2$

上式等號成立，當 $\frac{\sqrt{c_1}}{\sigma_b}=\frac{\sqrt{c_2m}}{\frac{\sigma_w}{\sqrt{m}}}$ 時，這時 $C^*\times V$ 具有最小值，可得 $m=\sqrt{\frac{\sigma_w^2c_1}{\sigma_b^2c_2}}$。