《抽樣方法》

總評	今年高考抽樣方法，應屬最近10年來最有變化的一次，考試內容靈活，考題偏難，除計算外尚須搭配簡單說明及證明，須謹慎並把握時間，程度好的考生可拿70分以上，一般考生可達50分以上。

一、假定有一母體中共有$N=3$個單元，各單位編號i及其母體變數值（value of population variable of interest）y_i如下：

i	1	2	3
y_i	5	7	15

今以一取出放回之設計D，每次以下表之選擇機率（draw-by-draw selection probability）選擇一個單元觀察後放回，再以同樣的選擇機率選擇下一個單元，共選擇$n=2$個樣本單元，選擇機率如下：

i	1	2	3
$P_{(i)}$	1/6	1/3	1/2

(一) 試求設計D下第一個單元（$i=1$）之包含機率（inclusion probability），亦即以本設計選擇之一組樣本，其中包含第一個單元之機率。（5分）

(二) 若以觀察值樣本平均，記為$\bar{y}_n$，推估本母體平均μ，請問$\bar{y}_n$在設計D下是否為不偏估計？（5分）

(三) 請提出一個在設計D下，μ的不偏估計量，請說明其何以不偏，並計算您提出的不偏估計在樣本s = (1, 3)時μ之估計值。（10分）

試題評析	本題為Horvitz-Thomposon估計，在歷屆考題中亦曾出現過，要拿到分數並不困難。
考點命中	《高點抽樣方法講義》第四回上課補充，王俊彰編撰。

答：

(一) $\pi_1 = 1-\left(1-\frac{1}{6}\right)^2 = \frac{11}{36}$

(二) 令 $f(\bar{y}) = \begin{cases} \frac{1}{36}, \bar{y}=5 \\ \frac{1}{9}, \bar{y}=6 \\ \frac{1}{9}, \bar{y}=7 \\ \frac{1}{6}, \bar{y}=10 \\ \frac{1}{3}, \bar{y}=11 \\ \frac{1}{4}, \bar{y}=15 \end{cases} \rightarrow E(\bar{y}_n) = \frac{1}{36}\times 5+\frac{1}{9}\times 6+\cdots+\frac{1}{4}\times 15 = \frac{32}{3}$

$\mu = 5\times\frac{1}{6}+7\times\frac{1}{3}+15\times\frac{1}{2} = \frac{32}{3} \rightarrow E(\bar{y}_n) = \mu \rightarrow$ *所以$\bar{y}_n$為不偏估計量*

(三) 令 $\bar{y}_n = \frac{\hat{\theta}_{HT}}{n} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{\pi_i}$，*其中$\pi_i$表示第$i$個元素被抽取的機率*

再令 $I_i = \begin{cases} 1 & \text{觀測值被抽中} \\ 0 & \text{觀測值未被抽中} \end{cases} \rightarrow E(I_i) = P(I_i=1) = \pi_i$

$$E(\bar{y}_n) = \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{y_i}{\pi_i}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{E(I_i y_i)}{\pi_i} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} = \mu$$

$$\pi_3 = 1-\left(1-\frac{1}{2}\right)^2 = \frac{3}{4}$$

$$\bar{y}_n = \frac{\hat{\theta}_{HT}}{3} = \frac{1}{3}\sum_{i=1}^{n}\frac{y_i}{\pi_i} = \frac{1}{3}\left(\frac{5}{\frac{11}{36}} + \frac{15}{\frac{3}{4}}\right) = 12.1212$$

二、在某鎮所進行的年度家庭醫療支出調查中，調查戶為以簡單隨機抽樣取出不放回（simple random sampling without replacement）在全鎮10,000戶中所選取之1,000戶樣本戶。調查結束後所得之家戶醫療支出樣本平均為150仟元。今欲進一步探討家戶收入在50仟元／月以下之低收入戶醫療支出概況，因無財稅資料可供查考，故僅能由樣本資料判定受訪戶是否為符合前述定義之低收入戶，但無法得知本鎮符合定義之低收入戶總戶數。

檢視樣本資料後，符合此一定義之受訪戶共有100戶，而該100戶之年醫療支出總和為8,000仟元，另該100戶之年醫療支出平方和為2×10^{12}。（亦即若令y_i為第i戶之年醫療支出，s_d為樣本戶中之低收入戶集合，則$\sum_{i\in s_d} y_i = 8\times10^6$以及$\sum_{i\in s_d} y_i^2 = 2\times10^{12}$。）

(一) 若以80仟元作為該鎮低收入戶之平均年醫療支出之估計量，請問該估計量是否為一不偏估計？並請說明理由或證明。（8分）

(二) 請估計(一)中估計量之95%信賴區間，請以仟元為單位，並請四捨五入至小數點下第二位。（8分）

(三) 請問該鎮所有低收入戶之醫療總支出之不偏估計量為何？請以仟元為單位，並說明或證明該估計量之不偏性。（10分）

(四) 請估計(三)中估計量之95%信賴區間，請以仟元為單位，並請四捨五入至整數位。（10分）

試題評析	本題是考母體估計，可搭配簡單隨機想法下筆，實屬冷門，但要拿分並不困難。
考點命中	《高點抽樣方法講義》第一回，簡單隨機抽樣觀念，王俊彰編撰。

答：

(一)母體平均數估計

$$\hat{\mu}_{低} = \frac{1}{n_{i\in s_d}}\sum_{i\in s_d} y_i = \frac{8\times10^6}{100} = 80000(元) = 80(仟元)$$

$$= \mu_{i\in s_d} \rightarrow 所以該估計量為不偏估計量$$

(二)$$\text{var}\left(\hat{\mu}_{低}\right) = \left(\frac{N_1 - n_{i\in s_d}}{N_1}\right)\times\frac{s^2_{i\in s_d}}{n_{i\in s_d}} = \left(\frac{1000-100}{1000}\right)\times\frac{\left(2\times10^{12} - \frac{(8\times10^6)^2}{100}\right)\Big/_{99}}{100} = 123636363.64$$

$\mu_{i\in s_d}$之95%信賴區間為

$$\left(\hat{\mu}_{低} \mp Z_{\alpha/2}\sqrt{\text{var}\left(\hat{\mu}_{低}\right)}\right) = \left(80000 \mp 1.96\sqrt{123636363.64}\right)$$

$$= (58206.3896，101793.6104)(元) = (58.21，101.79)(仟元)$$

(三)次母體總和數估計

$$\hat{\tau}_{低} = N_1\times\hat{\mu}_{低} = \frac{N_1}{n_{i\in s_d}}\sum_{i\in s_d} y_i = 1000\times\frac{8\times10^6}{100} = 80000000(元) = 80000(仟元)$$

$$\because \tau_{i\in s_d} = N_1\times\mu_{i\in s_d} \rightarrow$$

$$E\left(\hat{\tau}_{低}\right) = E\left(N_1\times\hat{\mu}_{低}\right) = N_1\times\mu_{i\in s_d} = \tau_{i\in s_d} 所以該估計量為不偏估計量$$

(四)$\text{var}\left(\hat{\tau}_{低}\right) = {N_1}^2\left(\frac{N_1 - n_{i\in s_d}}{N_1}\right) \times \frac{s^2_{i\in s_d}}{n_{i\in s_d}} =$

$$1000^2 \times \left(\frac{1000-100}{1000}\right) \times \frac{\left(2\times 10^{12} - \frac{(8\times 10^6)^2}{100}\right)/99}{100} = 123636363640000$$

$\tau_{i\in s_d}$*之*95%*信賴區間為*

$$\left(\hat{\mu}_{低} \mp Z\alpha_{/2}\sqrt{\text{var}\left(\hat{\mu}_{低}\right)}\right) = \left(80000000 \mp 1.96\sqrt{123636363640000}\right)$$

$$= \left(58206389.6\ ,\ 101793610.4\right)(元) = \left(58206\ ,\ 101794\right)(仟元)$$

三、在某鎮所進行的家庭月支出調查中其主抽樣設計如下：依家戶所得將全鎮家戶分為高收入戶（家戶月收入300仟元以上之100戶），一般收入戶（家戶月收入50仟元至299,999元之5,400戶）及低收入戶（家戶月收入49,999元以下共500戶），然後在各類收入家戶中選擇欲觀察之樣本。在各類收入戶中再視需要以不同之抽樣設計選擇該類收入戶樣本戶，在各類家戶中之次抽樣設計及調查方法如下：

1. 高收入戶採全查面訪：月支出平均值為200仟元，月支出標準差為30仟元。
2. 一般收入戶中，因其戶數眾多，故先在全鎮40個里中隨機選擇5個里，而在各選擇的里中，再以簡單隨機抽樣取出不放回選擇若干里內樣本戶。
3. 低收入戶中以簡單隨機抽樣取出不放回選擇200戶面訪調查；另因低收入戶之支出狀況為調查重點之一，為求慎重起見，再由該200樣本戶中以簡單隨機抽樣取出不放回選擇其中50戶，而該50戶另加以記帳調查蒐集其當月支出資料。

(一) 請問本調查中之主抽樣設計為何？同時請說明本調查採本設計之可能原因及優點。（8分）

(二) 在一般收入戶中調查資料如下：

里編號	里內總戶數	里內樣本戶數	里內月支出樣本平均（仟）
2	100	25	80
18	400	50	75
19	120	30	60
30	200	40	90
38	800	100	65

請以比例估計（ration estimation）推估一般收入戶之平均月支出，並請說明在此採比例估計之可能優點及其原因。（8分）

(三) 在低收入戶中若面訪戶之樣本為s，其中記帳戶樣本為s'（s'為s之子集合），以下為樣本統計量（均以仟元為觀察單位）：

統計量	s	s'
面訪樣本平均	28	25
記帳樣本平均	NA	30
面訪樣本變異數	25	4
記帳樣本變異數	NA	9

另在s'中記帳與面訪所得之月支出相關係數為0.68，請問本類家戶中所使用的抽樣設計為何？並請以適當的方式推估低收入戶之家庭月支出平均。（8分）

(四) 請推估本鎮家戶月支出平均，並說明（不須計算）其估計量之變異數估計程序。（10分）

(五) 明年度的調查仍欲採類似方式進行，然而因經費所限，故高收入戶將改採簡單隨機抽樣取

出不放回之抽樣調查以取代全查，但仍欲將該類母體平均之最大推估誤差在95%之信心水準下，控制為10仟元以下，請問欲達此一精確度要求之所需最小樣本數。（10分）

試題評析	本題較偏向實務面操作，在主抽樣下又搭配其他次抽樣，考生不易下筆，整體觀念需搭配融會貫通，此題要拿高分有難度。
考點命中	《高點抽樣方法講義》第一、三、四回，王俊彰編撰。

答：

(一)

1.主抽樣設計為分層隨機抽樣，次抽樣分別採用全查及簡單隨機抽樣，其中在一般收入戶中，採比例兩階段估計亦即為兩階段群集抽樣，在低收入戶中，採兩階段簡單隨機。

2.採用本設計主要原因為將性質接近分在同一層，即層內的變異小，層間的變異大，主要目的為提高統計量的精確度。主要優點為抽樣較為有效，可靠性較高，亦可與其他抽樣技術結合使用，次要優點為可將大樣本按重要變項予以劃分，樣本太大可使用該技術等優點。

(二)$\hat{\mu} = \left(\frac{N}{M}\right) \times \frac{\sum_{i=1}^{5} M_i \bar{y}_i}{n} = \left(\frac{40}{5400}\right) \times \frac{100\times80+400\times75+\cdots+800\times65}{5} = 170.667(\text{仟元})$

一般收入戶中，採比例兩階段估計亦即為兩階段群集抽樣，採用本設計可能原因為了便利性且可與分層隨機抽樣合併使用，成為多階段抽樣法。。

主要優點：抽樣較節省成本且適用抽樣單位較大時，當抽樣單位分佈地區很散時，可得到一個較完整之樣本團體。

(三)$\hat{\mu}_{低} = \hat{R} \times \hat{\mu}_{s'} = \frac{28}{25} \times 30 = 33.6$

低收入戶抽樣採兩階段簡單隨機抽樣，其中除簡單隨機抽樣外，

另搭配比率估計推估低收入戶之家庭用支出平均，然其相關係數為推估變異數時使用。

(四)$\hat{\mu}_{st} = \sum_{h=1}^{3} W_h \widehat{\mu_h} = \frac{100}{6000} \times 200 + \frac{5400}{6000} \times 170.667 + \frac{500}{6000} \times 33.6 = 159.73(\text{仟元})$

$$\widehat{Var}(\hat{\mu}_{st}) = \sum_{h=1}^{3} W_h^2 (1 - f_h) \frac{s_h^2}{n_h}$$

1.令$X_1 \equiv$ 高收入戶，$X_2 \equiv$ 一般收入戶，$X_3 \equiv$ 低收入戶，其中$s_1^2 = 900$，

2.$s_2^2 = \frac{1}{135^2}\left(1 - \frac{4}{40}\right) \times \frac{s_b^2}{4} + \frac{1}{40\times5\times135^2} \sum_{i=1}^{5} M_i^2 (1 - f_i) \times \frac{s_i^2}{m_i}$

，s_b^2為一般收入戶中$M_i \bar{y}_i$的 variance，s_i^2為第i群的$variance$，$f_i = \frac{m_i}{M_i}$，

此題一般收入中缺少s_i^2故無法計算。

3.$s_3^2 = \frac{s_s^2 - s_r^2}{200} + \frac{s_r^2}{50} = \frac{25-6.1504}{200} + \frac{6.1504}{50} = 0.2173$，

其中$s_r^2 = 4 - 2 \times \frac{28}{25} \times 0.68 \times 3 \times 2 + \left(\frac{28}{25}\right)^2 \times 9 = 6.1504$

∴ 綜合以上可求得$\widehat{Var}(\hat{\mu}_{st}) = \sum_{h=1}^{3} W_h^2 (1 - f_h) \frac{s_h^2}{n_h}$

(五)$n_0 = \frac{Z_{\alpha/2}{}^2 \sigma^2}{B^2} = \frac{(1.96\times30)^2}{100} = 34.5744$，$n = \frac{n_0}{\frac{N-1}{N} + \frac{n_0}{N}} = \frac{34.5744}{\frac{99}{100} + \frac{34.5744}{100}} = 25.884 \rightarrow$ 取 $n = 26$ 即可