集群(群集)抽樣

Cluster Sampling

集群抽樣概念

Obtain a specified amount of information about a population parameter

SRS及分層抽樣其抽樣技術主要是立基於研究者可確認出母群體中的每一個個體。假如母群體很小便很容易以上述的方法來抽樣，但是如果“**母群體很大**”，好比是一個城市 (**city**) 、州 (**state**) 、縣 (**province**) 或國家，則要確認每一個抽樣個體便非常困難且昂貴。在這種情況下，使用集群抽樣便較適當。

一般運用在廣大的地理區域，以減少抽樣與調查成本。抽樣前先將母體依特殊標準 (人為或自然因素) 分成若干集群，以集群為抽樣單位進行隨機抽樣, 自其中抽取一個或數個群集合成樣本進行調查。

全國環保政策民意調查 (依縣市區域劃分21個區域, 再隨機抽取6個區域進行全面性調查)

*注意區域的相似性

集群抽樣主要是立基於研究者將抽樣母群體分成若干個團體，此便成之為集群，然後使用SRS在每個集群中抽出所需數量的個體。集群的形成可以**地理位置的鄰近性**為基礎，或是以共同的特質，而這些特質與研究的主要變項是相關的 (就如同分層抽樣) 。依據劃分集群的層次，有時，可能要在不同的層次抽樣。 不同的層次構成劃分集群的不同階段 (**單階、雙階或多階**)。

想像假如你想調查大學生對澳洲高等教育問題的態度。高等教育機構在澳洲各省。另外，還有各種不同型態的高等教育機構，如大學、科技大學、學院、技術學院等等。在每一個教育機構中對大學部及研究所都提供各種不同的課程。每一個學術專業都要求四到五年的修業時間。可以想像此工程的浩大。在此情形下想要抽出一個隨機樣本，集群抽樣設計便非常有用(資料來源:胡龍騰等, 2001)。

集群抽樣的概念 (資料來源:胡龍騰等, 2001)

第一層的集群抽樣可能是省。各個集群可依據相似的特質來分。假如這不容易的話，則可考慮所有的省，而樣本則以機構作為層次的劃分。舉例來說，用SRS，在每一個省對各類高等教育機構都抽選一個機構，例如，每一省都抽選一所大學、一所科技大學、一所學院及一所技術學院。此是立基於一個假設，即同一個型態中的每一所高等教育機構是相似的。然後，在同一所教育機構中以隨機的方式抽選一個或多個系而在同一個系中可抽選特定年級的學生。而且也可依據年級的比例來抽選學生。此種抽樣過程的方式便稱之為**多階段集群抽樣** (資料來源:胡龍騰等, 2001)。

集群抽樣:

有時比起前述三種隨機抽樣方法，在相同成本下，給予更多的資訊。

總合來說，集群抽樣需在下列情形下，可以用最小的成本成為有效的設計來獲得特殊量的資訊：

1. 一個好的母群體架構列示要不是不適用或需耗資獲得，正當一個架構列示集群容易獲得時。

2. 取得觀察值的成本增加，當分離元素的距離增加。

集群抽樣是爲隨機抽樣，每個抽樣單位是爲元素的集合或集群。

（定義8-1）

與分層隨機抽樣之間的不同:

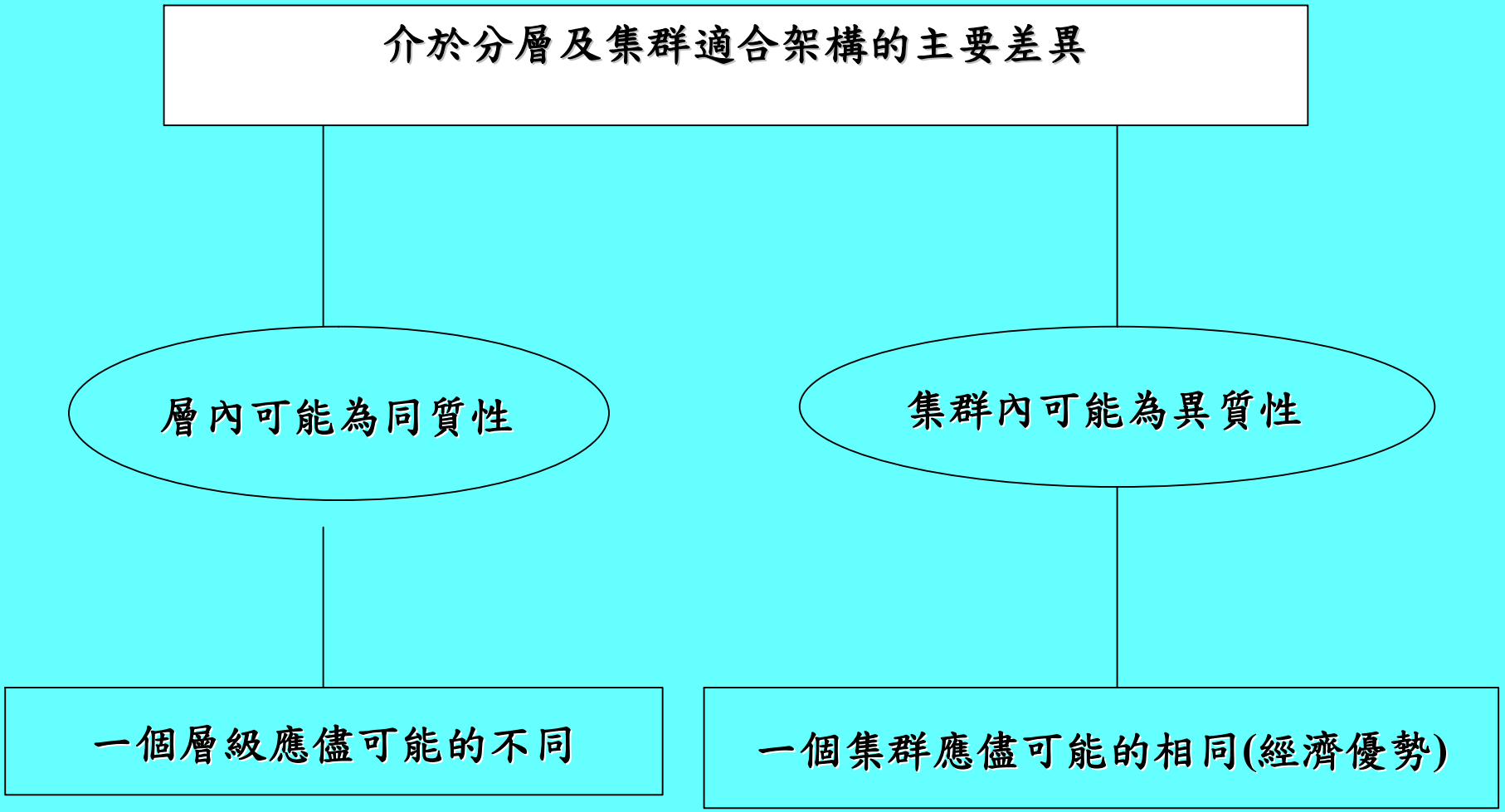
分層抽樣: 層間個體異質，層內個體同質

集群抽樣: 集群間個體同質，集群內個體異質

	優點	缺點	使用時機
隨機抽樣	1.各樣本均有被抽到的可能。亦即合於均等和獨立原則。 2.簡單易行。可用抽籤方式及使用隨機亂數表。	不適用於抽樣母群相當多的情形。即抽樣越多，需編碼也越多。	1.群體抽樣單位不太多時。 2.抽樣單位較同質時。
系統抽樣	由抽樣名單中有系統的每間距抽出若干單位。 1.簡便易行 2.只要有隨機排列名單即可行之。	易受樣本名單順序的影響，亦即若樣本名單中有週期性，若第一個樣本選定，隨後之樣本亦選定。	需先知道樣本名單不受規律性、週期性之安排。
分層隨機抽樣	1.可適用於樣本分配不均或異質時，使樣本也可以具有代表性。 2.分層的標準可同時用多種特徵。確切樣本易抽出。	分層標準不易釐清。	1.分層抽樣必須合於研究目的，即和要測量的變項有密切關係。 2.分層類別必須有確切的樣本、可靠的資料。 3.分層類別必須互斥。
集群抽樣	1.適用抽樣單位較大時。 2.適用抽樣單位分佈地區很散時。 3.可得到一個較完整之樣本團體。	因為抽樣單位是集合體，樣本點可能同質性高。	1.適合集群數較少或較多的情境。 2.可與分層隨機抽樣合併使用，成為多階段抽樣法。

資料來源: 中正大學85教研法考題

集群抽樣程序


介於分層及集群適合架構的主要差異
層內可能為同質性
集群內可能為異質性
一個層級應儘可能的不同
一個集群應儘可能的相同(經濟優勢)

母群體平均數及總數估計

集群抽樣是每個抽樣單位包含一些元素的簡單隨機抽樣，因此，母群體**平均數** μ 及**總數** τ 是類似於簡單隨機抽樣，特別，樣本平均數是為母群體平均數 μ 的良好估計值，在這個單元裏將討論 μ 的估計值及 τ 的兩個估計值。

以下是本章所用到的一些標示：

N = 母群體的集群數

n = 簡單隨機抽樣的集群數

m_i = 集群i中的元素數，i = 1, 2, …, N

$\overline{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ = 樣本的平均集群大小

$M = \sum_{i=1}^{N} m_i$ = 母群體中平均集群大小

$\overline{M} = \frac{M}{N}$ = 母群體中元素平均數

y_i = 在i個集群中總觀察數

母群體平均數 μ 的估計值是為樣本平均數 $\overline{y}$，如下所示：

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$$

母群體平均數 μ 的估計值：

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} \qquad (8.1)$$

$\overline{y}$ 的估計變異數：

$$\hat{V}(\overline{y}) = \left(\frac{N-n}{Nn\,\overline{M}^2}\right) s_r^2 \qquad (8.2)$$

當中

$$s_r^2 = \frac{\sum_{i=1}^{n} \left(y_i - \overline{y}\, m_i\right)^2}{n-1} \qquad (8.3)$$

*如果M未知，$\overline{M}$可以被 $\overline{m}$ 所估計

公式8-2的估計變異數是有偏差的且好的估計值只有在n很大的時候，諸如：$n \geq 20$；如果集群大小$m_1, m_2, \ldots, m_N$相等時，該偏差會消失。接下來舉例說明。

【例題8-1】：表8-1中有25個經過訪談集群的樣本，在該表中，有收入資料，使用表中的資料估計城市中的每資本收入（per-capital income）及估計誤差。

表8-1 每資本收入（美金）

集群i	居民數, m_i	每集群總收入, y_i	集群I	居民數, m_i	每集群總收入, y_i
1	8	$96,000	14	10	$49,000
2	12	121,000	15	9	53,000
3	4	42,000	16	3	50,000
4	5	65,000	17	6	32,000
5	6	52,000	18	5	22,000
6	6	40,000	19	5	45,000
7	7	75,000	20	4	37,000
8	5	65,000	21	6	51,000
9	8	45,000	22	8	30,000
10	3	50,000	23	7	39,000
11	2	85,000	24	3	47,000
12	6	43,000	25	8	41,000
2006/12/813	5	54,000	$\sum_{i=1}^{25} m_i = 151$	$\sum_{i=1}^{25} y_i = 1{,}329{,}000$	25

『解』：

當然，需先整理出相關的重要公式主成份，諸如：平均數、中位數及標準差（表8-2）。

表8-2每資本收入（美金）敘述統計

	N	平均數	中位數	標準差
居民	25	6.04	6.0	2.371
收入	25	53160	49,000	21,784
$y_i - \overline{ym}_i$	25	0	993	25,189

(a)母群體平均數:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} = \frac{\$\,1,329\ \ ,000}{151} = \frac{\$\,53\ \ ,160}{6\ .04} = \$\ 8801\ \ \ .3245$$

$$(b)\ \overline{m} = \frac{\sum_{i=1}^{n} m_i}{n} = \frac{151}{25} = 6\ .04$$

$$(c)\ \hat{V}\left(\overline{y}\right) = \left(\frac{N-n}{Nn\ \overline{M}^{2}}\right)\frac{\sum_{i=1}^{n}\left(y_i - \overline{y}m_i\right)^2}{n-1} = \left[\frac{415\ -25}{(415\ )(25\ )(6.04\ )^2}\right](25,189\ )^2 = 653\ ,785$$

$$(d) B = 2\sqrt{653{,}785} = 1{,}617$$

$$(e) \text{整體表示} : \bar{y} \pm 2\sqrt{\hat{V}\left(\bar{y}\right)} = 8{,}801 \pm 1{,}617$$

(f) 平均每資本收入（美金計算）是$8,801，估計誤差在95%信賴水準下應該少於$1,617，這樣的估計誤差相當大，改進方式可採用抽取更多集群或增加樣本大小

既然公式8-1中的是一個比例估計值，所以可以使用第6章裏的替換計算公式，表8-2的進一步敘述統計表如下：

	N	平均數	標準差
y	25	53.16	21.78
m	25	6.04	2.317
y及m的相關係數 = .303			

在此，計算是以千元（少3個0）為單位，現今，可以被計算為：

$$\hat{V}\left(\overline{y}\right)=\left(\frac{N-n}{Nn}\right)\left(\frac{1}{\overline{m}^{2}}\right)\left[s_{y}^{2}+r^{2}s_{m}^{2}-2r\hat{\rho}s_{y}s_{m}\right]$$

其中，

$$s_{m}^{2}=\frac{1}{m-1}\sum_{i=1}^{m}\left(m_{i}-\overline{m}\right)^{2}$$

在這裏的r是公式8-1的，這樣計算產生：

$$\hat{V}(\bar{y}) = \left[\frac{390}{415(25)}\right]\left(\frac{1}{6.04^2}\right)\left[(21.78)^2 + (8.801)^2(2.371)^2 - 2(8.801)(.303)(21.78)(2.371)\right] = .6542$$

且

$$2\sqrt{\hat{V}(\bar{y})} = 1.616$$

此相等於$1,616，與上面計算的估計誤差相近。

母群體總數 τ 現今變成M μ 因為M表示在母群體元素的總數，相繼的，如簡單隨機抽樣，M提供 τ 的估計值。

$$M\ \overline{y} = M\ \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} \qquad (8.4)$$

$M\ \overline{y}$ 的估計變異數：

$$\hat{V}\left(M\ \overline{y}\right) = M^2\ \hat{V}\left(\overline{y}\right) = N^2 \left(\frac{N-n}{Nn}\right) s_r^2 \qquad (8.5)$$

注意 $M\overline{y}$ 估計值是為有用只有母群體元素數M，為已知。

【例題8-2】：利用表8-1的資料來估計城市所有居民的總收入並估計誤差，總共有2,500個居民。

『解』：

$$M\ \overline{y} = 2500\ (8801\ ) = \$22\ ,002\ ,500$$

$$M\ \overline{y} \pm 2\sqrt{\hat{V}\left(M\ \overline{y}\right)} = M\ \overline{y} \pm 2\sqrt{M^{2}\hat{V}\left(\overline{y}\right)}$$

$$\downarrow$$

$$22\ ,002\ ,500\ \pm 2\sqrt{\hat{V}\left(M\ \overline{y}\right)} = M\ \overline{y} \pm 2\sqrt{M^{2}\hat{V}\left(\overline{y}\right)}$$

$$\downarrow$$

$$22\ ,002\ ,500\ \pm 2\sqrt{(2500\ )^{2}(653\ ,785\ )}$$

$$\downarrow$$

$$22\ ,002\ ,500\ \pm 4\ ,042\ ,848$$

同樣，估計誤差很大，可以藉由增加樣本大小來減少。

母群體總數 τ 的估計值，不依靠M：

$$N\ \overline{y}_t = \frac{N}{n}\sum_{i=1}^{n} y_i \qquad \textbf{(8.7)}$$

$N\ \overline{y}$ 的估計變異數：

$$\hat{V}\left(N\ \overline{y}_t\right) = N^2\ \hat{V}\left(\overline{y}_t\right) = N^2\left(\frac{N-n}{Nn}\right)s_t^2 \qquad \textbf{(8.8)}$$

其中

$$s_t^2 = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}_t\right)^2}{n-1} \qquad \textbf{(8.9)}$$

如果集群大小有很大的變異量，且如果集群大小與集群總數是高度相關，$N\bar{y}_t$ 的變異數通常比 $M\bar{y}$（公式8-5）來得大，$N\bar{y}_t$ 的估計值並不是使用集群大小$m_1, m_2, \ldots, m_n$所提供的資訊，因此可能較不精確。

【例題8-3】：利用表8-1的資料來估計城市所有居民的總收入，如果M是未知，並估計誤差。

『解』：

$$N\bar{y}_t = \frac{N}{n}\sum_{i=1}^{n} y_i = \frac{415}{25}(1{,}329{,}000) = \$22{,}061{,}400$$

此數字很接近在例題8-2中的估計值

$$s_t^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}_t\right)^2}{n-1} = (21{,}784)^2$$

$$N\bar{y}_t \pm 2\sqrt{N^2\left(\frac{N-n}{Nn}\right)S_t^2} = 22{,}061{,}400 \pm 3{,}505{,}532$$

相同集群大小；與簡單隨機抽樣的比較

為了更詳盡的研究介於集群抽樣及簡單隨機抽樣之間的關係，將限制所有的m_i為共同值 – 例如，m。假設整個集群母群體是為真，就如每箱罐頭食品抽取一箱裏確實有24罐的情形一樣，在此案例中，M=Nm，以及總樣本大小是nm元素（每個m元素的n集群）。

μ及τ的估計值當所有集群大小相等時（亦即，$m_1 = m_2 = \ldots = m_N$）擁有特殊的特性，首先，公式8-1中的是為母群體平均數μ的不偏估計值，第二，公式8-2的是變異數的不偏估計值，最後，$M\bar{y}$及$N\bar{y}_t$為母群體總數τ是為相等。

公式8-1中每元素的母群體平均數會被標示為相同集群大小 $\bar{\bar{y}}_c$，並且計算如下：

$$\bar{\bar{y}}_c = \frac{1}{m}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}$$

其中的yij代表i集群中的第j個觀察值，注意 $\bar{\bar{y}}_c$ 可以被看作nm測量樣本的整體平均，或抽樣及群總數除以m的平均數，從後者的觀點，可以很容易的看到：

$$\hat{V}\left(\bar{\bar{y}}_c\right) = \left(\frac{N-n}{N}\right)\left(\frac{1}{nm^2}\right)\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(y_i - \bar{y}_t\right)^2$$

其中

$$\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_i = m\,\bar{\bar{y}}_c$$

如果讓集群i的樣本平均數標示為 $\bar{y}_i$，這樣就變成 $\bar{y}_i = \frac{y_i}{m}$，或 $y_i = m\bar{y}_i$，公式就可以重寫為：

$$\frac{1}{m^2 n(n-1)}\sum_{i=1}^{n}\left(y_i - \bar{y}_t\right)^2 = \frac{1}{m^2 n(n-1)}\sum_{i=1}^{n}\left(m\bar{y}_i - m\bar{\bar{y}}_c\right)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\bar{y}_i - \bar{\bar{y}}_c\right)^2$$

為了簡化變異數計算與探索集群抽樣及簡單隨機抽樣之間的關係，可以利用平方和相似於古典ANOVA的鑑定，如下表示：

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}-\bar{\bar{y}}_c\right)^2=\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}-\bar{y}_i\right)^2+\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\bar{y}_i-\bar{\bar{y}}_c\right)^2=\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}-\bar{y}_i\right)^2+m\sum_{i=1}^{n}\left(\bar{y}_i-\bar{\bar{y}}_c\right)^2$$

以上三個詞，從左邊開始，分別為**平方和總數（SST）**，**集群內平方總合（SSW）**及**集群間平方總合（SSB）**，上述恆等式為：

SST = SSW + SSB

隨著適當的除數，這些平方和變成ANOVA的均方根，因此，MSB計算如下：

$$MSB = \frac{SSB}{n-1} = \frac{m}{n-1}\sum_{i=1}^{n}\left(\bar{y}_i-\bar{\bar{y}}_c\right)^2$$

MSW計算如下：

$$MSW = \frac{SSW}{n(m-1)} = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(y_{ij} - \bar{y}_i \right)^2$$

接續的是：

$$\hat{V}\left(\bar{\bar{y}}_c \right) = \left(\frac{N-n}{N} \right) \frac{1}{nm} MSB$$

【例題8-4】：報社發行經理想要估計在給予社區中每個家庭購買報紙的平均數，家庭與家庭間的旅費支出是顯而易見的，因此，在社區中的4,000個家庭，列示於400個地理集群及包含10個家庭，且簡單隨機抽取4個集群，採用訪談的方式，結果如所附表格，請估計社區每個家庭平均報紙數及估計誤差。

集群	報紙數										合計
1	1	2	1	3	3	2	1	4	1	1	19
2	1	3	2	2	3	1	4	1	1	2	20
3	2	1	1	1	1	3	2	1	3	1	16
4	1	1	3	2	1	5	1	2	3	1	20

『解』：

從公式8-1知道 $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$，當$m_1 = m_2 = \ldots = m$，此方程式計算如下：

(a) $\bar{\bar{y}}_c = \dfrac{\sum_{i=1}^{n} y_i}{nm} = \dfrac{19 + 20 + 16 + 20}{4(10)} = 1.875$

標準ANOVA計算利用統計軟體計算如下:

ANOVA			
Source	DF	SS	MS
Factor	3	1.07	.36
Error	36	43.30	1.20
Total	39	44.38	

在本案例中，”**Factor**”代表集群間的計算，以及”**Error**”代表集群內計算，因此，**MSB = .36及MST=1.20**，所以

（b） $\hat{V}\left(\bar{\bar{y}}_c\right)=\left(\frac{N-n}{N}\right)\frac{1}{nm}MSB=\left(\frac{396}{400}\right)\frac{1}{4(10)}(.36)=.0089$且$2\sqrt{\hat{V}\left(\bar{\bar{y}}_c\right)}=.19$

（c）結論

每個家庭報紙數最佳估計值是 $1.88\pm.19$

選擇樣本大小來估計母群體平均數及總數

集群樣本的資訊量被兩個因素所影響，**集群數及相對集群大小**，第二個因素到目前還未探討，在關於估計每州家庭是否有不合適火險時，集群可能是郡、選區、學區、社區，或任何便利的家庭組別。如前所述，誤差的大小很重要的決定於集群總數的變異性，因此，嘗試達到最小的誤差，在這些總數中儘可能選擇最小變異集群，假設集群大小（抽樣單位）已被選擇且將會考慮只有選擇集群數n的問題。

從公式8-2知道的估計變異數是：

$$\hat{V}\left(\bar{y}\right)=\frac{N-n}{Nn\,\overline{M}^{2}}\left(s_r^{2}\right)$$

其中：

$$s_r^{2}=\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}m_i\right)^{2}}{n-1} \qquad \textbf{(8.10)}$$

$\bar{y}$ 的真實變異數近似於：

$$V\left(\bar{y}\right)=\frac{N-n}{Nn\overline{M}^{2}}\left(\sigma_r^{2}\right) \qquad \textbf{(8.11)}$$

其中 σ_r^{2} 被 s_r^{2} 所估計。

因為不知道σ_r^2或平均集群大小$\overline{M}$，樣本大小，即需要購買特別量關於母體參數及群數的資訊是困難的，這裏的問題可以運用比例估計的概念來克服，亦即，從先前研究所估計獲得的σ_r^2或 $\overline{M}$ ，或選擇含有n'元素的初步樣本。σ_r^2及 $\overline{M}$ 可從初步資料計算獲得，並且用來獲得近似樣本大小的總數，因此，有關樣本大小的選擇，2個標準差的估計值為估計誤差B，這方面的界定在於研究者他們可以忍受最大誤差的界限，亦即：

$$2\sqrt{V\left(\overline{y}\right)} = B$$

利用公式8-11，可以解n。當使用去估計母群體總數時，可獲得相類似的結果，因為 $V\left(M\bar{y}\right)=M^2V\left(\bar{y}\right)$ 。

估計 μ 具有估計誤差B的近似樣本大小：

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2} \qquad (8.12)$$

其中 σ_r^2 是由 s_r^2 所估計且 $D = \frac{\left(B^2 \overline{M}^2\right)}{4}$

【例題8-5】：假設表8-1代表城市中收入的初步樣本，在未來的調查中，應抽取多少樣本來估計平均每資本收入μ具有$500美金的誤差？

『解』：

使用公式8-12，必須估計σ_r^2，最好的估計值是s_r^2，此資料已在表8-2中算出，所以	
s_r^2	$s_r^2 = \dfrac{\sum_{i=1}^{n}\left(y_i - \overline{y}m_i\right)^2}{n-1} = (25,189)^2$
$\overline{M}$ 可以用表8-2的 $\overline{m} = 6.04$ 來估計，然後D是近似於	
（a）D	$\dfrac{B^2\overline{m}^2}{4} = \dfrac{(500)^2(6.04)^2}{4} = (62,500)(6.04)^2$
（b）n	$n = \dfrac{N\sigma_r^2}{ND+\sigma_r^2} = \dfrac{415(25,189)^2}{415(6.04)^2(62,500)+(25,189)^2} = 166.58$
（c）結論	應抽取167個集群

使用 $M\bar{y}$ 估計 τ 具有估計誤差B的近似樣本大小：

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2} \qquad (8.13)$$

其中 σ_r^2 是由 s_r^2 所估計且 $D = \frac{B^2}{4N^2}$

【例題8-6】：在未來的調查中，應抽取多少樣本（表8-1中）來估計所有居民總收入 τ 具有$1,000,000美金的誤差？城市裏有2,500個居民（M = 2,500）

『解』：

利用公式8.13及用 $s_r^2 = (25{,}189)^2$ 來估計 σ_r^2	
（a）D	$= \frac{B^2}{4N^2} = \frac{(1{,}000\ {,}000\ )^2}{4(415\ )^2}$
（b）ND	$= \frac{(1{,}000{,}000)^2}{4(415)} = 602{,}409{,}000$
（c）n	$n = \frac{N\sigma_r^2}{ND + \sigma_r^2} = \frac{415(25{,}189)^2}{602{,}409{,}000 + (25{,}189)^2} = 212.88$
（d）結論	應抽取213個集群（在總收入誤差為$1,000,000）

公式8-7中的 $N\bar{y}_t$，當M未知是用來估計 τ，公式8-8的估計變異數公式如下：

$$\hat{V}\left(N\bar{y}_t\right) = N^2\hat{V}\left(\bar{y}_t\right) = N^2\left(\frac{N-n}{Nn}\right)s_t^2$$

其中：

$$s_t^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}_t\right)^2}{n-1}$$

因此，$N\bar{y}_t$ 的母群體變異數為：

$$\hat{V}\left(N\bar{y}_t\right) = N^2V\left(\bar{y}_t\right) = N^2\left(\frac{N-n}{Nn}\right)\sigma_t^2 \qquad (8.14)$$

其中母群體 σ_t^2 是被 s_t^2 所估計。

τ 的估計值具有估計誤差B產生下列的方程式：

$$2\sqrt{V\left(N\bar{y}_t\right)} = B$$

應用公式8-14可以解n。

使用 $N\bar{y}_t$ 估計 τ 具有估計誤差B的近似樣本大小：

$$n = \frac{N\sigma_t^2}{ND + \sigma_t^2} \qquad (8.15)$$

其中母群體 σ_t^2 是被 s_t^2 所估計且 $D=\frac{B^2}{4N^2}$ 。

【例題8-7】：假設表8-1代表城市中收入的初步樣本，在未來的調查中，應抽取多少樣本來估計所有居民總收入τ具有$1,000,000美金的誤差？（M未知）

『解』：

σ_t^2 必須由 s_t^2 來估計，從表8-2可以找到結果，	
（a） S_t^2	$= \frac{\sum_{i=1}^{n} (y_i - \overline{y}_t)^2}{n-1} = (21,784)^2$
（b）D	$= \frac{B^2}{4N^2} = \frac{(1{,}000{,}000)^2}{4(415)^2}$
（c）n	$= \frac{N\sigma_t^2}{ND + \sigma_t^2} = \frac{415(21{,}784)^2}{\frac{415(1{,}000{,}000)^2}{4(415)^2} + (21{,}784)^2} = 182.88$
（d）結論	因此，在$1,000,000估計誤差下，需要183個集群樣本

母群體比例估計

假設一個實驗者想要估計母群體比例或分數，諸如家庭沒有寬頻網路的比例，或是公司總裁具有研究所學歷的比例，母群體比例**p**最好的估計值是為樣本比例 $\hat{p}$ ，令a_i為感興趣特性**i**集群中元素的總數，然後，在**n**集群擁有特性樣殊元素比例公式如下：

$$\hat{p} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i}$$

其中，m_i是為i集群中的元素數，i = 1, 2, …, n；注意，與有相同的型態【如公式8-1】，除了y_i是被a_i所取代，的估計變異數相等於 $\bar{y}$ 。

母群體比例p的估計值：

$$\hat{p} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i} \qquad (8.16)$$

$\hat{p}$ 的估計變異數：

$$\hat{V}\left(\hat{p}\right) = \left(\frac{N - n}{Nn\,\overline{M}^{2}}\right) s_{p}^{2}$$

(8.17)

其中：

$$s_{p}^{2} = \frac{\sum_{i=1}^{n}\left(a_i - \hat{p}\,m_i\right)^2}{n-1}$$

(8.18)

變異數公式8-17，是好的估計值當樣本大小n為大的時候 - 如 $n \geq 20$。如果 $m_1 = m_2 = ... = m_N$ ，然後 $\hat{p}$ 是p的不偏估計值，且公式8-17中的 $\hat{V}\left(\hat{p}\right)$，是為任何樣本大小實際變異數的不偏估計值。

【例題8-8】：除了問到之前例題的收入外，例題8-1中的居民還被問到他們的房子是自己的或是租的，請使用下列表8-3來估計租房子居民的比例及估計誤差。

『解』：

(a) $\hat{p}$	$= \dfrac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i} = \dfrac{72}{151} = \dfrac{2.88}{6.04} = .48$
(b) s_p^2	$= (0.726)^2$

表8-3 租用者數

集群	居民，m	租用者，a	$a_i - \hat{p}\, m_i$
1	8	4	0.16
2	12	7	1.24
3	4	1	-0.92
4	5	3	0.60
5	6	3	0.12
6	6	4	1.12
7	7	4	0.64
8	5	2	-0.40
9	8	3	-0.84
10	3	2	0.56
11	2	1	0.04
12	6	3	0.12
13	5	2	-0.40
14	10	5	0.20
15	9	4	-0.32
16	3	1	-0.44
17	6	4	1.12
18	5	2	-0.40
19	5	3	0.60
20	4	1	-0.92
21	6	3	0.12
22	8	3	-0.84
23	7	4	0.64
24	3	0	-1.44
25	8	3	-0.84

	N	Mean	Median	SD
居民	25	6.04	6.00	2.371
租用者	25	2.88	3.00	1.509
$a_i - \hat{p}\, m_i$	25	-0.019	0.12	0.726

(c) $\overline{m}$	$= \dfrac{\sum_{i=1}^{n} m_i}{n} = \dfrac{151}{25} = 6.04$
(d) $\hat{V}\left(\hat{p}\right)$	$= \left(\dfrac{N-n}{Nn\overline{M}^2}\right) s_p^2 = \dfrac{(415-25)}{415(25)(6.04)^2}(.726)^2 = 0.00054$
(e) 整體表示	$\hat{p} \pm 2\sqrt{\hat{V}\left(\hat{p}\right)} = .48 \pm 2\sqrt{.00054} = .48 \pm .05$

因此，租房子居民比例最好的估計值是為.48，在95%信賴水準下，小於0.05的估計誤差，事實上若是用簡單隨機抽樣來估計比例，就可以免去計算之苦，因此可以考慮 $\hat{V}\left(\hat{p}\right)$ 其他版本的計算，表8-3可以進一步整理如下：

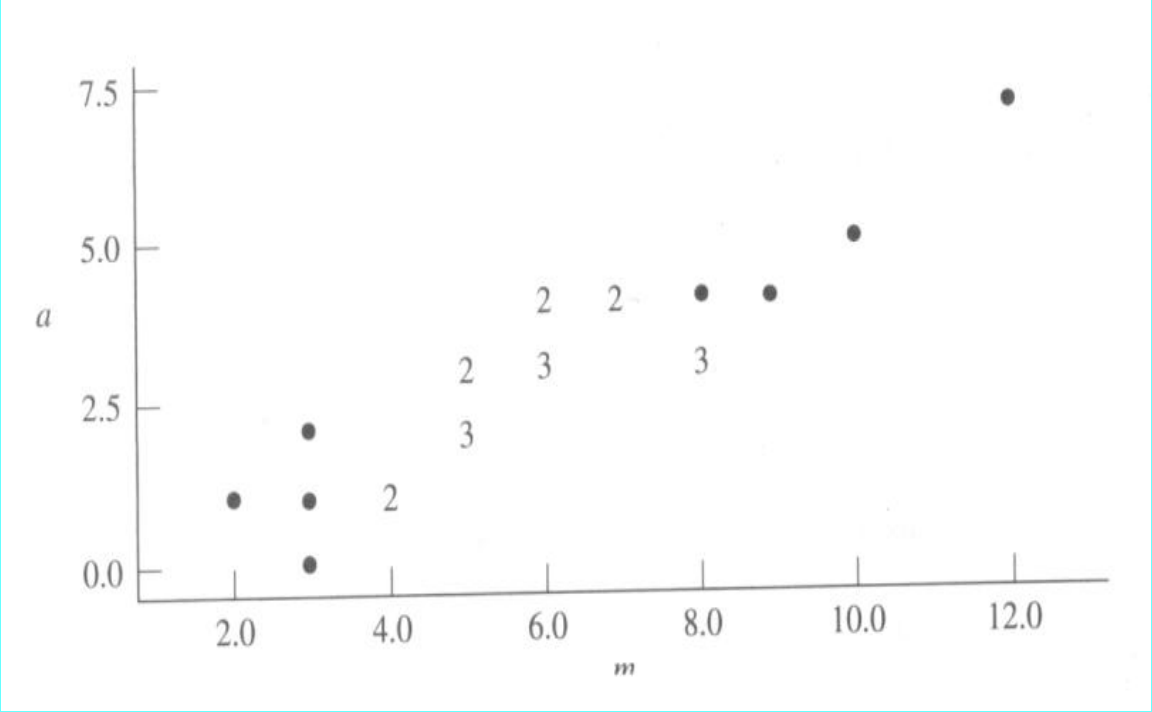
	N	平均數	標準差
a	25	2.88	1.509
m	25	6.04	2.371
a及m的相關係數 = .886			

從上述的整合，可以很容易的計算：

$$\hat{V}\left(\hat{\bar{p}}\right)=\left(\frac{N-n}{Nn}\right)\left(\frac{1}{\bar{m}}\right)^2\left[s_a^2+\hat{\bar{p}}^{\,2}\,s_m^2-2\,\hat{\bar{p}}\,\hat{\rho}\,s_a s_m\right]$$

$$=\left(\frac{390}{415(25)}\right)\left(\frac{1}{6.04}\right)^2\left[(1.509)^2+(.477)^2(2.371)^2-2(.477)(.886)(1.509)(2.371)\right]=.00055$$

以上的結果與先前的計算一致，下列有關上述表格中a及m的資料圖形可以顯示強烈的線性關係（圖8-1）。



選擇樣本大小來配置估計比例所需樣本

母群體比例p具有估計誤差B單位代表實驗者想要找出下列答案：

$$2\sqrt{V\left(\hat{p}\right)} = B$$

這個方程式可以解n，而且解答與公式8-12類似，亦即：

$$n = \frac{N\sigma_p^2}{ND + \sigma_p^2}$$

其中，$D = \frac{B^2\overline{M}^2}{4}$，而且$\sigma_p^2$被估計為：

$$s_p^2 = \frac{\sum_{i=1}^{n}\left(a_i - \hat{p}m_i\right)^2}{n-1}$$

【例題8-9】：表8-3的資料過時了，在同樣的城市裏有新的研究出爐，主要目的在於估計租用房子居民的比例，在估計誤差為.04下，需要多大的樣本來估計p？

『解』：

(a) S_p^2	$= \dfrac{\sum_{i=1}^{n}\left(a_i - \hat{p}\, m_i\right)^2}{n-1} = (.726)^2 = .527$
(b) D	$= \dfrac{B^2 \overline{M}^2}{4} = \dfrac{(.04)^2 (6.04)^2}{4} = .0146$
(c) n	$= \dfrac{N\sigma_p^2}{ND + \sigma_p^2} = \dfrac{(415)(.527)}{(415)(.0146) + .527} = 33.20$
(d) 結論	在估計誤差爲.04下，應抽取34個集群

與分層混合的集群抽樣

誠如其他抽樣方法，集群抽樣可與其他抽樣方法混合，觀念上母群體可以分為L層，在每一個層級中可抽取一個集群樣本。

回顧公式8-1中，有比例估計值且被認定為平均集群總數對到平均集群大小的比例估計值，然後，思考比例估計值，在層級內有兩種方式來形成母群體平均數估計值：分開估計值及混合估計值。如果使用分開估計值的方式，在每一層元素的總數應為已知，為了給予適當『層重』（stratum weights）；既然因為這些量通常為未知，所以只能用集群抽樣混合的比例估計值，以下舉一具體實例作說明：

【例題8-10】：例題1中，令表8-1中的資料變成層級1樣本，$N_1 = 415$及$n_1 = 25$，一個相鄰的城市作為層級2，在層級2中，n2是從N2=168抽取出的10個集群，請估計兩城市混合每資本收入及估計誤差，相關資料如附表。

集群i	居民數目m_i	每個集群總收入y_i
1	2	$18,000
2	5	52,000
3	7	68,000
4	4	36,000
5	3	45,000
6	8	96,000
7	6	64,000
8	10	115,000
9	3	41,000
10	1	12,000

『解』：

(a) $\overline{y}_{t1}$	(集群1平均總數) $= 53{,}160$
(b) $\overline{y}_{t2}$	(集群2平均總數) $= 54{,}700$
(c) $\overline{m}_1$	(集群1平均大小) $= 6.04$
(d) $\overline{m}_2$	(集群2平均大小) $= 4.90$
(e) 母群體平均集群總數	$\frac{1}{N}\left(N_1 \overline{y}_{t1} + N_2 \overline{y}_{t2}\right)$
(f) 平均集群大小	$\frac{1}{N}\left(N_1 \overline{m}_1 + N_2 \overline{m}_2\right)$
(g) 元素母群體平均數估計值	$\overline{y}_c = \frac{N_1 \overline{y}_{t1} + N_2 \overline{y}_{t2}}{N_1 \overline{m}_1 + N_2 \overline{m}_2}$

以上方程式並沒有混合比例估計的型態，如同單元6-6的觀念，$\overline{y}_c$ 的變異數可以估計如下：

$$\overline{V}\left(\overline{y}_c\right)=\frac{1}{M^2}\left\{\frac{N_1\left(N_1-n_1\right)}{n_1}s_{c1}^2+\frac{N_2\left(N_2-n_2\right)}{n_2}s_{c2}^2\right\}$$

當M為母群體元素總數且未知時，可以被 $N_1\overline{m}_1+N_2\overline{m}_2$ 來估計，第一個變異數 s_{c1}^2 是為層級1中 $\left(y_i-\overline{y}_c m_i\right)$ 的變異數；第二個變異數 s_{c2}^2 是為層級2 $\left(y_i-\overline{y}_c m_i\right)$ 中的變異數。

Cluster Sampling with Probabilities Proportional to Size
(比例機率集群抽樣大小)

上述的探討中，可說是把每一集群所含基本單位數 N_i 假設為大致相等或近似相等，因此大都是以均等機率 **(equal probability)**，即用SRS技術進行抽樣。

但在每一集群所含基本單位數 N_i 不相等的情況下，最好不用均等機率抽樣技術，而用比例機率 **(proportional probability sampling)** 抽樣技術抽樣。

統計文獻上咸認為比例機率抽樣以用於集群單位或多段組織的第一段抽樣單位上最為適宜。即比例機率抽樣，以用於大單位較合適，因單位大，各單位的輔助訊息可能相差懸殊，如用均等機率抽樣，殊難達到高效率統計效果。

比例機率又稱不等機率 (pps sampling)

若是M元素在母群體中，母群體平均數估計值 $\hat{\mu}_{pps}$，簡單地可以表示如下：

$$\hat{\mu}_{pps} = \frac{1}{M}\hat{\tau}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i$$

$\hat{\mu}_{pps}$ 的估計變異數同樣地也很容易計算，所有相關公式歸納如下：

母群體平均數 μ 的估計值：

$$\hat{\mu}_{pps} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i \qquad \textbf{(8.19)}$$

其中 $\bar{y}_i$ 是i個集群中的平均數。

$\hat{\mu}_{pps}$ 的估計變異數：

$$\hat{V}\left(\hat{\mu}_{pps}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\bar{y}_i - \hat{\mu}_{pps}\right)^2 \qquad \textbf{(8.20)}$$

母群體總數 τ 的估計值：

$$\hat{\tau}_{pps} = \frac{M}{n}\sum_{i=1}^{n}\bar{y}_i \qquad (8.21)$$

$\hat{\tau}_{pps}$ 的估計變異數：

$$\hat{V}\left(\hat{\tau}_{pps}\right) = \frac{M^2}{n(n-1)}\sum_{i=1}^{n}\left(\bar{y}_i - \hat{\mu}_{pps}\right)^2 \qquad (8.22)$$

【例題8-11】：一位審計人員想要抽取一家大型公司的病假紀錄，為了估計上一季每個員工平均病假天數，此公司有8個部門，每個部門有不同的員工，既然每個部門病假天數應與員工數高度相關，所以該審計人員決定抽取n=3個機率比例員工數的部門，請顯示員工數分別為1,200, 450, 2,100, 860, 2,840, 1,910, 290, 3,200的樣本選取方法。

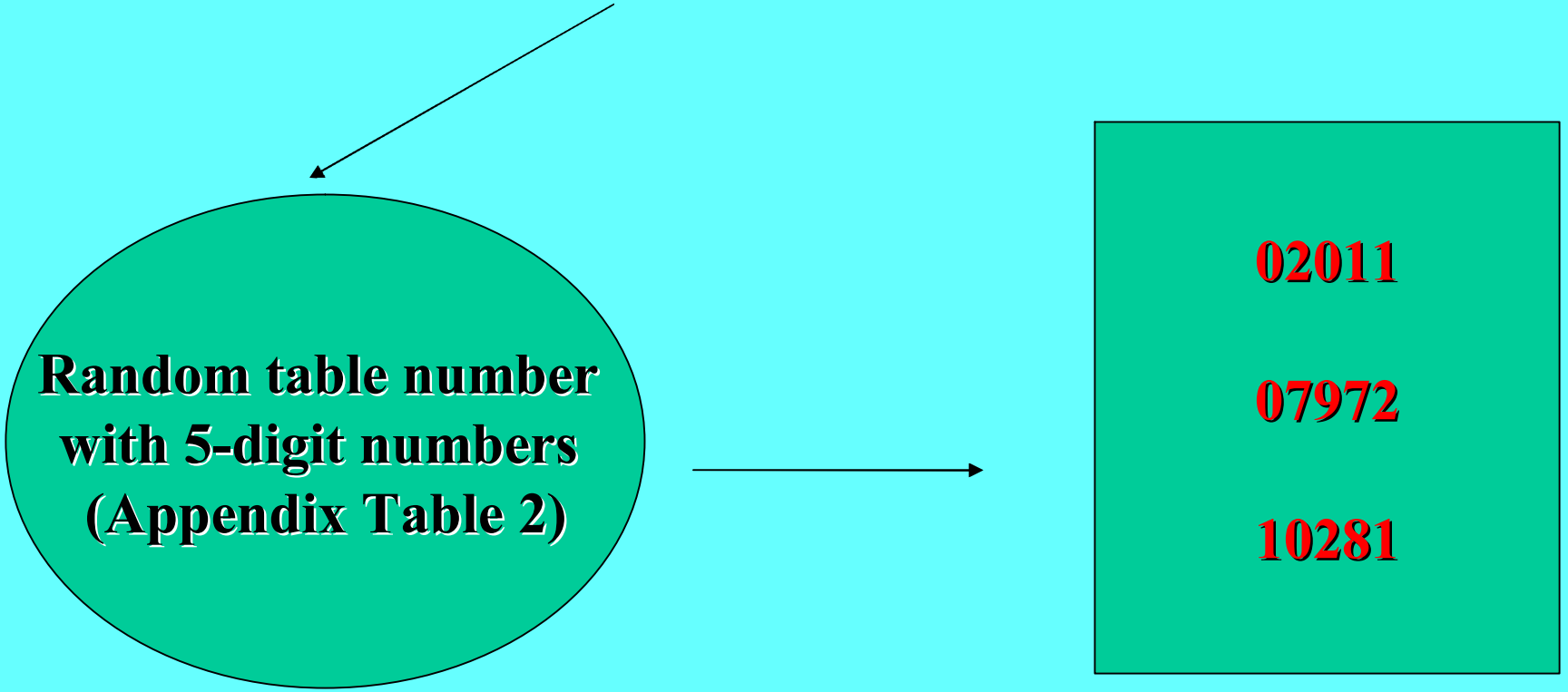
『解』：

第一個列出員工數及每個部門累計範圍如下：

部門	員工數	累積範圍
1	1,200	1-1,200
2	450	1,201-1,650
3	2,100	1,651-3,750
4	860	3,751-4,610
5	2,840	4,611-7,450
6	1,910	7,451-9,360
7	390	9,361-9,750
8	3,200	9,751-12,950
	12950	

(a) 判斷

Since n = 3 divisions are to be sampled, we must select three random numbers between 00001 and 12,950.



(b) 結論

We first list t… …oyees and the cumulative range for eac… …vs:

02011

07972

10281

Division	…er of …yees	Cumulative range
1	1,200	1 – 1200
2	450	1201 – 1650
3	2,100	1651 – 3750
4	860	3751 – 4610
5	2,840	4611 – 7450
6	1,910	7451 – 9360
7	390	9361 – 9750
8	3,200	9751 – 12,950
	12,950	

【例題8-12】：假設過去一季在3個抽取部門總病假數分別為：

$y_1 = 4320$ $y_2 = 4160$ $y_3 = 5790$

估計整個公司每個人病假的平均數及估計誤差。

『解』：

先算出抽出來集群的集群平均數	
(a) $\overline{y}_1$	$= \frac{4,320}{2,100} = 2.06$
(b) $\overline{y}_2$	$= \frac{4,160}{1,910} = 2.18$
(c) $\overline{y}_3$	$= \frac{5,790}{3,200} = 1.81$
(d) $\hat{\mu}_{pps}$	$= \frac{1}{n}\sum_{i=1}^{n} \overline{y}_i = \frac{1}{3}(2.06 + 2.18 + 1.81) = 2.02$
(e) $\hat{V}\left(\hat{\mu}_{pps}\right)$	$= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\overline{y}_i - \hat{\mu}_{pps}\right)^2 = \frac{1}{3(2)}\left[(2.06-2.02)^2 + (2.18-2.02)^2 + (1.81-2.02)^2\right]$ $= .0119$
(f) B	$2\sqrt{.0119} = .22$
(g) 結論	公司每個員工上一季平均病假天數及估計誤差為 $2.02 \pm .22$

在集群抽樣中有3種母群體總數的估計值：（a）比例估計值（ratio estimator）【公式8-4】、（b）不偏估計值（the unbiased estimator）【公式8-7】及所謂的（c）pps估計值。該如何判定那一種是最好的？以下有幾種參考指引：

1. 如果y_i與m_i不相關，然後不偏估計值會比其他估計值來得好；
2. 如果y_i與m_i相關，然後比例及pps會比不偏估計值來的精準；
3. pps估計值比比例估計值來得好，如果集群內變異不會隨著m_i改變而改變；
4. 比例估計值比pps估計值好，如果集群內變異會隨著m_i改變而改變。

在例題8-11及8-12中，病假數應跟著員工數增加而增加，因此，不偏估計值在此就是很差勁的選擇，然而，在每個部門病假天數的變異，橫跨部門間可以維持相當的穩定，在這樣的情形下，pps估計值是為最好的選擇。