통계 89－01
(강의자료)

統 計 學(標本理論)

經濟企劃院 調査統計局

標本抽出에서 Bayes와 Minimax과정

(Bayes and minimax procedures in sampling)

한국외국어 대학교

申 敏 雄

目　　次

Ⅰ. 標本의 基礎 理論

1. 單純確率 抽出(simple random sampling)

① 標本調査(sample survey)… 母集團의 部分集合 위에서 수집한 資料

② 센서스(census)… 母集團 全體를 수집한 資料

③ 標本使用의 理由

…費用의 절감

…時間의 단축

…標本單位들로부터 더 詳細한 情報를 蒐集

…非標本誤差가 클 때에는 標本으로부터 더 正確한 結果를 얻을 수 있다.

④ 標本을 使用할 때의 問題點

…작은 地域 推定値들은 높은 精度(precision)를 提供하지 못한다.

…時間이 지나감에 따른 작은 變化를 찾기 위해 큰 標本이 要求될 수도 있다.

⑤ …必要한 標本은 크고, 地域이 작을 경우는 오히려 센서스가 費用이 적게 들 수도 있다.

(1) 定 義

① 母集團(population)

…調査에 關聯된 모든 元素들 U_1, U_2, …, U_N의 집합

② 特性値(characteristic)

…變數들에 대한 一般的 술어

③ 觀察値의 單位(unit)

…資料를 얻는 單位

④ 分析의 單位

…圖表 作成이나 分析하는 데 쓰이는 單位

⑤ 標本抽出 單位(sampling unit)

…母集團으로부터 標本을 抽出할 때의 單位

⑥ 統計値(statistics)

…標本 元素들 위에 정의된 함수

⑦ 母集團 母數(population parameter)

…母集團 元素들 위에 정의된 함수

⑧ 抽出臺帳(frame)

…母集團의 標本單位들 全部를 記錄한 장부(또는 틀이라고 부른다.)

⑨ 標本調査에 있어서 주요 단계

…目的의 明確한 言及

…母集團에 대한 定義

…蒐集될 資料에 대한 言及

…要望되는 程度에 대한 言及

…資料蒐集 方法에 대한 決定

…적절한 標本名簿에 대한 確認

…標本抽出

…豫備檢定

…Field作業의 組織

…資料의 要約, 分析 및 발간

…未來의 調査에 대한 情報獲得

⑩ 標本理論의 役割

…調査의 모든 段階에 응용

…最小의 費用으로 가장 精度높은 推定値

(2) 標本抽出로의 接近

① 非確率標本 (nonrandom sampling)

ⓐ 類型

…有意選定 (purposive selection),

判斷標本 (judgement sample)

ⓑ 有利한 點

…낮은 費用

ⓒ 不利한 點

…推定値에 대한 신뢰도를 알 수 없음.

② 確率標本 (random sampling)

ⓐ 알려진 確率로서 單位들의 選擇

ⓑ 有益性

…신뢰도를 推定할 수 있음.

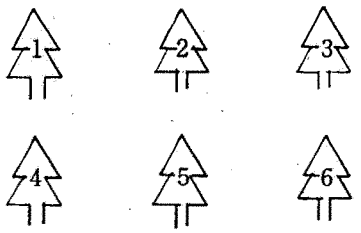
專門家의 判斷과 다른 利用 可能한 情報와의 結合.

例

母集團이 N = 6인 나무들의 集合에서 標本의 크기가 n = 2인 標本을 뽑아라.

…얼마나 많은 標本들이 抽出可能한가?

…다음과 같이 나무들이 배열되었다고 假定하자.

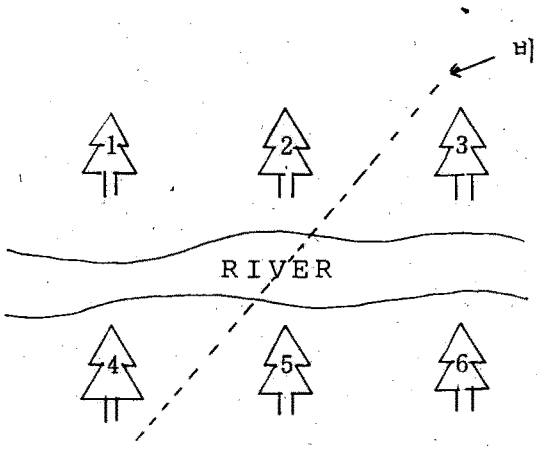


關係함 (irrigated)

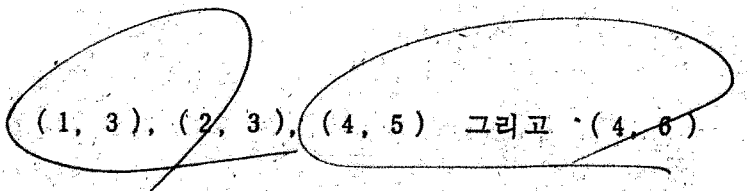
關係 안함

각 그룹에서 한나무씩 뽑을 수 있다. 이 方法을 層化 (stratification)라고 부른다.

…나무들이 다음과 같이 배열되었다고 假定하자.



강을 건너기가 費用이 많이 든다고 假定하자. 다음과 같은 標本들에 높은 確率을 割當하자.



나머지 標本들에는 낮은 確率을 할당한다.

이러한 技法을 조절된 選擇(controlled selection)이라고 한다.

(3) 定 義

① 標本誤差(sampling error)

…센서스에 의해서 얻어진 값과 標本에서 얻어진 推定値 사이에 差異

② 非標本 誤差(nonsampling error)

…標本이외의 원인으로서 일어나는 差異

③ 正確度(全體 誤差)

…標本에서 얻어진 값과 母集團의 값 사이의 差異

④ 精度(precision)

…非標本誤差는 무시하고, 標本誤差만 計算하여 그 값이 작은 程度

⑤ 單純確率標本(simple random sample, SRS)

…等確率로 抽出된 標本

⑥ 推定量(Estimator)

…標本값의 함수

⑦ 推定値(Estimate)

…推定量의 값

⑧ 推定量의 期待値

…모든 可能한 標本들의 推定値들의 平均値

⑨ 推定値의 偏倚

…推定量의 期待값과 참값(true value)사이의 差異

(4) 單純確率 抽出法(simple random sampling, SRS)

① 標本選擇의 方法

ⓐ 復元 抽出(with replacement)

…한번 뽑힌 單位가 다시 뽑힐 수 있다.

ⓑ 非復元 抽出

…한번 뽑힌 單位는 다시 뽑히지 않는다.

例

12가구로 構成된 母集團을 생각해 보자(N=12).

2가구의 單純確率標本을 抽出하라(N=2)

즉, N = 12 (SRS) n = 2

母 集 團

母集團 單位, U_i	收入, y_i	收入이 있는 사람, x_i
U_1	$ 1300	1
U_2	6300	2
U_3	3100	1
U_4	2000	1
U_5	3600	1
U_6	2200	1
U_7	1800	1
U_8	2700	1
U_9	1500	1
U_{10}	900	1
U_{11}	4800	2
U_{12}	1900	1

(5) 記 號

① N個의 母集團은 $y_1, y_2, \cdots, y_N$ 으로 表示한다.

② 標本은 $y_1, y_2, \cdots, y_n$ 으로 表示한다.

③ 一般的으로, 母集團은 大文字로 表示하고, 標本은 小文字로 表示한다.

	母 集 團	標 本
總 合	$Y=\sum_{i=1}^{N} y_i$	$y=\sum_{i=1}^{N} y_i$
平 均	$\bar{Y}=Y/N$	$\bar{y}=y/n$
比	$R=Y/X=\bar{Y}/\bar{X}$	$\hat{R}=y/x=\bar{y}/\bar{x}$
比 率	P	p

④ 母集團에 대한 SRS에 의한 推定

	推 定 量
$\bar{Y}$	$\hat{\bar{Y}}=\bar{y}$
Y	$\hat{Y}=N\bar{y}$
$R=\bar{Y}/\bar{X}=Y/X$	$\hat{R}=\bar{y}/\bar{x}=y/x$
P	p

(6) 推定値의 性質

① 任意의 推定値의 精度

…推定의 方法

…標本抽出 設計

② 一致性 (consistency)

…n = N일 때에 推定値는 母集團값이 된다면, 推定方法은 一致性이 있다.

…$\bar{y}$는 $\bar{Y}$의 一致推定量이다.

…$N\bar{y}$는 Y의 一致推定量이다.

…$\bar{y}/\bar{x}$는 R의 一致推定量이다.

③ 不偏推定 (unbiasedness)

…推定量의 期待값이 母集團의 값과 같은 경우

例

母集團은 6가구로 構成되었다. (N = 6)

標本으로 2가구를 選擇한다. (n = 2)

母集團 單位, U_i	가구의 크기, Y_i
U_1	4
U_2	3
U_3	3
U_4	5
U_5	2
U_6	1

$\bar{Y} = \frac{1}{N}\sum y_i = \frac{1}{6}(18) = 3$

$Y = N\bar{Y} = 6(3) = 18$

$E(\bar{y}) = \sum \bar{y}\left(\frac{6}{2}\right) = 45/15 = 3.0 = \bar{Y}$

$\hat{Y}$은 Y에 대하여 不偏推定値인가?

(7) 分散(variance)

① 기 호

θ -母集團 母數(Y, $\bar{Y}$, R, P)의 값

$\hat{\theta}$ -標本으로부터 θ의 推定値

$V(\hat{\theta})$-모든 可能한 $\binom{N}{n}$個의 標本들로부터 얻어진 $\hat{\theta}$값들의 分散

$v(\hat{\theta})$ - $V(\hat{\theta})$의 標本 推定値

② 여러가지 경우에 대한 分散

	分 散	
	定 義	代 數
y_i	$V(y_i) = E(y_i - \bar{Y})^2 = \sigma^2$	$V(y_i) = \frac{N-1}{N} S^2$
$\bar{y}$	$V(\bar{y}) = E(\bar{y} - \bar{Y})^2$	$V(\bar{y}) = (1-\frac{n}{N})\frac{S^2}{n} = (1-f)\frac{S^2}{n}$
$\hat{Y}$	$V(Y) = E(Y - Y)^2$	$V(\hat{Y}) = (1-\frac{n}{N})N^2\frac{S^2}{n}$
		$= (1-f)N^2\frac{S^2}{n}$

$S^{12} = \sum_{i=1}^{N} \frac{(y_i - \bar{Y})^2}{}$

$f = \frac{n}{N}$ (標本抽出率)

$(1-f)$ = 有限母集團修正因子 (finite population correction factor, FPC)

$\sqrt{V(\hat{\theta})} = \hat{\theta}$ 의 標準誤差

$\frac{n}{N} \leq 0.05$ 이면, 우리는 FPC를 무시한다.

(8) 標準誤差의 推定

① $\sqrt{V(\hat{Y})}$와 $\sqrt{V(\bar{y})}$의 使用

… SRS와 다른 標本抽出方法과의 精度比較

…標本크기의 推定

…精度의 推定

② 推 定

特性	母集團값	標本推定
	$S^{12}=\sum_{i=1}^{N}\frac{(y_i-\bar{Y})^2}{N-1}$	$s^2=\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{n-1}$
y_i	$V(y_i)=\frac{N-1}{N}S^2$	$V(y_i)=\frac{N-1}{N}s^2$
$\bar{y}$	$V(\bar{y})=(1-f)\frac{S^2}{n}$	$V(\bar{y})=(1-f)\frac{s^2}{n}$
$\hat{Y}$	$V(\hat{Y})=N^2(1-f)\frac{S^2}{n}$	$V(\hat{Y})=N^2(1-f)\frac{s^2}{n}$
	$=N^2V(\bar{y})$	$N^2V(\bar{y})$

※ 주의 : $\sqrt{V(\hat{\theta})} \equiv s\hat{\theta}$

ⓐ 信賴區間

i) θ에 대한 信賴區間 : $\hat{\theta} \pm t\sqrt{V(\hat{\theta})}$ 또는

$\hat{\theta} \pm t\, s\theta$

ii) $\bar{Y}$에 대하여 : $\bar{y} \pm ts_{\bar{y}}$

iii) Y에 대하여 : $y \pm ts_y$

實習 # 1 : 單純任意 標本抽出 (SRS)

부록Ⅳ에 3地域에 30個 마을들의 現在와 5年前 센서스에 의해 구해진 人口와 家口의 크기를 나열하였다.

즉, 다음과 같이 要約된다.

地域의 數 = 3

마을의 總數 = 30

總家口數 = 600

人口 = 3037

5年前 센서스의 인구 = 2815

전면적 = 270 ㎢

마을의 平均 家口數 = 20

마을의 平均 사람數 = 101.23

家口當 平均 人員數 = 5.06

1) 부록Ⅳ의 600家口에서 非復元 單純任意抽出로 20家口를 뽑고, 이 標本으로 全體 人口에 대해 推定하라.

(a) 全體 人口

(b) 家口當 平均 家口員 數

2) 1번 問題에서 얻어진 각 推定値에 대하여 分散과 標準誤差를 推定하라.

(a) 復元 單純 任意 抽出 (SRS-WR)

i) 非復元抽出이 復元抽出보다 더 使用된다 (有限 母集團의 경우)

ii) $V(\bar{y}_{SRS-WOR}) \leq V(\bar{y}_{SRS-WR})$, $1 < n < N$

(b) 部分 母集團 (subpopulation)의 平均의 推定

i) 母集團의 部分集合에 대한 總合, 平均, 比 等에 대한 推定

ii) 이 部分集合을 研究의 領域 (domain)이라고도 부른다.

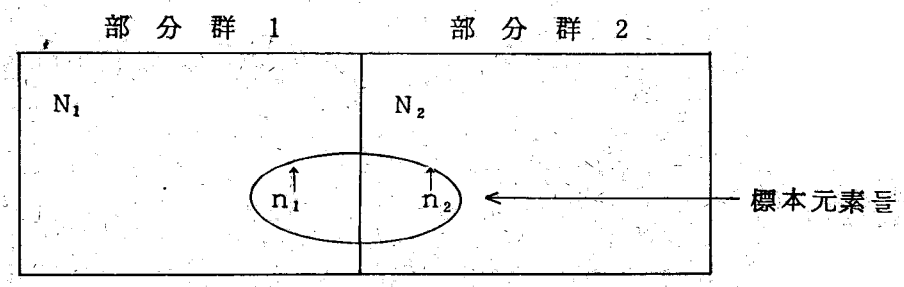
iii) 例

- 職業別 또는 收入別 家口에 대한 調査
- 社會調査에서 家族構造別 研究

iv) 母集團을 2個의 部分母集團으로 나눌 수 있다고 假定하자 (部分母集團은 各各 N_1과 N_2의 원소들로 構成된다.)

$N \xrightarrow{SRS-WOR} n$, 단, $n_1 + n_2 = n$

그림으로 表示하여,



N元素들의 母集團

$\bar{Y}_1$을 推定하고자 한다.

※ 주의 … $N \xrightarrow{SRS-WOR} n$ 이면,

$N_1 \xrightarrow{SRS-WOR} n_1$,

$N_2 \xrightarrow{SRS-WOR} n_2$이다.

$\bar{Y}_1$를 推定하여라. $\bar{y}_1 = \frac{1}{n_1}\sum_{k=1}^{n_1} y_{1k}$

$V(\bar{y}_1) = (1-\frac{n_1}{N_1})\frac{S_1^2}{n_1}$

$\hat{V}(\bar{y}_1) = (1-\frac{n_1}{N_1})\frac{s_1^2}{n_1}$

c 個의 部分母集團에 대하여

$N = N_1 + N_2 + \cdots\cdots + N_c$

$n = n_1 + n_2 + \cdots\cdots + n_c$

j 번째 部分母集團에 대한 推定

$\bar{y}_j = \frac{1}{n_j}\sum_{k=1}^{n_j} y_{jk}$

$V(\bar{y}_j) = (1-\frac{n_j}{N_j})\frac{S_j^2}{n_j}$

$\hat{V}(\bar{y}_j) =$

(c) 部分母集團들의 總合에 대한 推定

$N \xrightarrow{SRS-WOR} n$ 일 때에 Y_j를 推定하고자 한다.

2 가지 경우

i) N_j가 알려졌을 때

$\hat{Y}_j = N_j \bar{y}_j$

$V(\hat{Y}_j) = N_j^2 V(\bar{y}_j)$

$\hat{V}(\hat{Y}_j) = N_j^2 \hat{V}(\bar{y}_j)$

ii) N_j를 모를 때

$$\hat{Y}_j = \frac{N}{n}\sum_{k=1}^{n_j} y_{jk}$$

例

母集團은 N = 100 인 勤勞者들의 集合이다.

$N \xrightarrow{SRS-WOR} n = 5$

勤 勞 者	性 別, j	收入, y_{jk} ($1,000)
1	2	$20 = y_{21}$
2	1	$15 = y_{11}$
3	2	$35 = y_{22}$
4	1	$10 = y_{12}$
5	2	$25 = y_{23}$

※ 1 = 女子, 2 = 男子

男子의 收入, $\hat{Y}_2$를 推定하여라.

$V(\hat{Y}_2)$를 推定하라.

(c) 比 推定

i) 두 變數間의 比를 推定하고자 한다.

ii) $N \xrightarrow{SRS} n$ 아래서

$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$를 $\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$에 의해 推定한다.

iii) $E(\hat{R}) \neq R$, $\hat{R}$은 R의 偏倚推定量이다.

$n \to N$일 때에 偏倚(bias) $\to D$

iv) $MSE(\hat{\theta}) = VAR(\hat{\theta}) + BIAS^2$

v) n이 충분히 클 때는 $MSE(\hat{R}) = VAR(\hat{R})$

vi) $V(\hat{R}) =$

vii) $\bar{X}$를 모를 때 $\bar{X}$대신 $\bar{x}$로 대치하여라.

viii) $s\hat{R} = \sqrt{v(\hat{R})}$

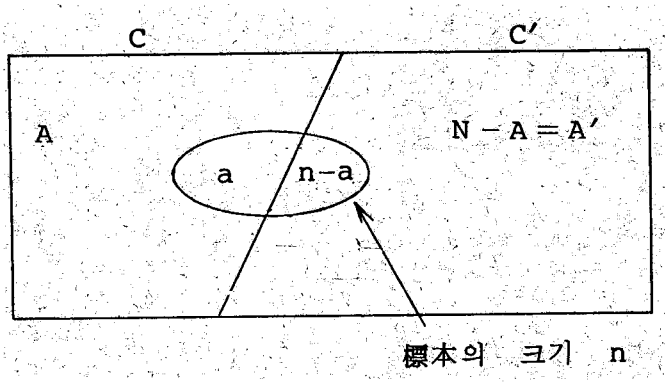
(e) 比率 推定

全體 중에서 어떤 特性을 갖는 部分에 대한 比率을 推定하고자 한다.

例…全 家口中 住宅 所有의 比率

기호…母集團의 모든 원소가 C나 C′中 하나에 속한다고 하자.

그리고 $N \xrightarrow{SRS-WOR} n$,



※ 주의 : A와 A′은 통상 모른다.

C에 속하는 元素의 比率은 $P = \frac{A}{N}$이다. 標本에서 C에 속하는 원소의 比率은 $p = \frac{a}{n}$이다.

P는 p로서 推定된다.

A는 $Np = N\left(\frac{a}{n}\right)$에 의하여 推定된다.

(f) 分散 推定

다음과 같이 變數를 定義하자.

$$y_i = \begin{cases} 1, & \text{元素 } i\text{가 } C\text{에 속할 때} \\ 0, & \text{元素 } i\text{가 } C\text{에 속하지 않을 때} \end{cases}$$

그러면, $Y = \sum_{i=1}^{N} y_i = A = NP$

$$\bar{Y} = \sum_{i=1}^{N} \frac{y_i}{N} = \frac{A}{N} = \frac{NP}{N} = P$$

그리고, $\bar{y} = \sum_{i=1}^{N} \frac{y_i}{n} = \frac{a}{n} = p$

SRS-WOR일 때에

$$E(p) = E(\bar{y}) = \bar{Y} = P,$$

$$V(p) = V(\bar{y}) = (1 - f)\frac{S^2}{n}$$

단, $S^2 = \sum_{i=1}^{N} (y_i - \bar{Y})^2 / (N-1)$

$$= \left(\sum_{i=1}^{N} y_i^2 - NY^2\right) / (N-1)$$

$$= \left(\sum_{i=1}^{N} y_i - N\bar{Y}^2\right) / (N-1)$$

$$= (A - NP^2) / (N-1)$$

$$= (NP - NP^2) / (N-1)$$

$$= NP(1-P) / (N-1)$$

$$= NPQ / (N-1)$$

즉, $V(P)=\frac{(1-f)}{n}\frac{NPQ}{N-1}$

$=\frac{N-n}{N}\frac{NPQ}{N-1}$

$=\frac{N-n}{N-1}\frac{PQ}{n}$

$\fallingdotseq(1-f)\frac{PQ}{n}$, $N-1\fallingdotseq N$ (N이 클 때)

$\hat{A}=Np$를 써서 $A=NP$를 推定

$E(\hat{A})=E(Np)=NE(p)=NP=A$

$V(\hat{A})=V(Np)=N^2V(p)=N^2\frac{N-n}{N-1}\frac{PQ}{n}$

$V(\bar{y})$의 不偏 推定量은 $v(\bar{y})=(1-f)\frac{S^2}{n}$

즉, $v(p)=v(\bar{y})=(1-f)\frac{S^2}{n}$

또한, $v(\hat{A})=v(Np)=N^2v(p)=N\frac{N-n}{n-1}pq$

例

3042명의 이름과 住所가 있는 名簿에서, SRS로 200名을 標本抽出한 結果 38名의 住所가 잘못 記錄되었다. 全體로 보아 몇 名이 잘못 記錄되었나를 推定하고, 그 推定値의 標準誤差를 구하라.

$\hat{A}=$

$s\hat{A}=$

(g) 比率의 推定

i) 部分母集團에 대한 比率과 總合

…다음과 같은 경우를 생각해 보라.

	領 域 1		領 域 2		…	領 域 k		總 計
	C	C′	C	C′	…	C	C′	
(母集團의 數)	A_1	A_1'	A_2	A_2'	…	A_k	A_k'	N
(標本의 數)	a_1	a_1'	a_2	a_2'	…	a_k	a_k'	n

ii) 比率의 推定

$$p_k = \frac{a_k}{n_k}$$

iii) 總計의 推定에 대한 2가지 경우

1) N_k를 알 때 ;

$$\hat{A}_k = N_k p_k = \left(\frac{N_k}{n_k}\right) a_k$$

$$s(\hat{A}_k) = N_k \sqrt{\left(1-\frac{n_k}{N_k}\right)\frac{p_k q_k}{n_k - 1}}$$

2) N_k를 모를 때 ;

$$\hat{A}_k = N\frac{a_k}{n}$$

$$s(\hat{A}_k) = N\sqrt{\left(1-\frac{n}{N}\right)\frac{pq}{n-1}}, \text{ 단, } p = \frac{a_k}{n}$$

그러면, N_k를 알 때와 모를 때, $s(p_k)$를 구하라.

例

어떤 市의 人口가 50,000名이다. SRS-WOR로 20%의 標本을 뽑았다. 標本에서 4,000名이 就業對象者이고, 그 중 200名이 未就業者

이다.

1) 이 市의 未就業率은 얼마인가?

2) 이 比率의 標準 誤差는 얼마인가?

3) 就業對象者중 未就業者의 總數를 推定하여라.

4) 이 推定値의 標準誤差를 구하라.

(h) 標本의 크기 推定

i) 적절한 標本 크기의 重要性

1) 너무 큰 標本은 낭비이다.

2) 너무 작은 標本은 그 結果의 有用性이 떨어진다.

3) 標本의 크기가 클수록 信賴度는 높아지나, 非標本 誤差때문에 正確性은 작아진다.

4) 標本의 크기 全體誤差를 최소화시키도록 뽑아져야 한다.

例

母集團이 N = 5000 사람이다. 標本 (SRS-WOR) 으로 n = 50 사람을 뽑았을 때 그 중 10 사람인 중국인이다.

중국인의 比率을 p라 할 때, p에 대한 95 % 信賴區間을 구하라.

$$p \pm ts_p = p \pm 2\sqrt{(1-f)\frac{pq}{n}}$$

$$= 0.2 \pm 2\sqrt{\left(1-\frac{50}{5000}\right)\frac{(0.2)(0.8)}{50}}$$

$$= (0.087,\ 0.312)$$

(i) 比率 推定을 위한 標本의 크기 決定

P가 $p \pm \alpha$ 밖에 있을 確率을 α라 하자.

n이 크고, P가 너무 작지는 않다고 하자.

P에 대한 信賴區間;

$$p \pm t\sqrt{\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}}$$

그러면,

$$\alpha = t\sqrt{\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}}$$

윗 식을 풀어서

$$n =$$

※주의 : 1) P가 Q를 모를 때는 이 公式을 쓸 수가 없다.

2) N이 충분히 클 때는

$$n_0 = \frac{t^2 PQ}{\alpha^2}$$

(j) 平均($\bar{Y}$)와 總合(Y)을 推定하기 위한 標本의 크기 推定

$\bar{Y}$를 推定하기 위해 標本의 크기를 決定하기 위한 公式은 앞절에서 σ^2 대신 PQ를 대치시키면 된다.

다음 公式을 회상하여라.

1) $P = \frac{\sum_{i=1}^{N} y_i}{N} = \bar{Y}$,

단, $y_i = \begin{cases} 1, & \text{單位 i가 C에 속할 때} \\ 0, & \text{그외에} \end{cases}$

2) $S^2 = \frac{NPQ}{N-1}$ 그리고 $\sigma^2 = \frac{N-1}{N} S^2$

즉,

$$\sigma^2 = \frac{N-1}{N} S^2 = \frac{N-1}{N} \left(\frac{NPQ}{N-1} \right) = PQ$$

n을 구하는 公式은

$$n = \frac{\frac{t^2\sigma^2}{d^2}}{1 + \frac{1}{N}\left(\frac{t^2\sigma^2}{d^2} - 1 \right)}$$

n이 클 때는

$$n_0 = \frac{t^2\sigma^2}{d^2}$$

$\frac{n_0}{N}$을 무시할 수 없으면,

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Y에 대한 $(1-\alpha)\,100\,\%$ 信賴區間은

$\hat{Y} \pm t\sigma_{\hat{Y}}$ 또는

$$N\bar{y} \pm tN\sqrt{(1-f)\frac{\sigma}{\sqrt{n}}}$$

精度의 程度는

$$d = tN\sqrt{(1-f)\frac{\sigma}{\sqrt{n}}}$$

n에 대해 풀면,

$$n = \frac{\frac{t^2(N^2\sigma^2)}{d^2}}{1 + t^2(\frac{N\sigma^2}{d^2})}$$

$n_0 = \frac{t^2N^2\sigma^2}{d^2}$ 으로 놓고, n_0/N을 무시할 수 없다면,

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

例 1

$N = 4000$

P_1 = 住宅所有의 퍼센트

P_2 = 自動車를 두대 所有한 家口의 퍼센트

우리가 원하는 精度는

$$d = \begin{cases} t\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{P_1Q_1}{n}} \leq 0.02 \\ t\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{P_2Q_2}{n}} \leq 0.01 \end{cases}$$

다음을 알고 있다.

$$\begin{cases} 0.45 < P_1 < 0.65 \\ 0.05 < P_2 < 0.10 \end{cases}$$

1) P_1에 대하여,

$$n_0 = \frac{t^2P_1Q_1}{t^2(\frac{N-n}{N-1})\frac{P_1Q_1}{n}} = \frac{P_1Q_1}{(\frac{N-n}{N-1})\frac{P_1Q_1}{n}}$$

※ 주의 : $P_1Q_1 = (0.5)(0.5) = 0.25$ 일 때 최대값을 갖는다.

이와같이 하여,

$$n_0 = \frac{(0.5)(0.5)}{(0.02)^2} = 625$$

$$\frac{n_0}{N} = \frac{625}{4000} = 0.156 > 0.05$$

$$n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{625}{1+\frac{625}{4000}} = 541$$

2) P_2 에 대하여,

$$n_0 = \frac{P_2Q_2}{(\frac{N-n}{N-1})\ \frac{P_2Q_2}{n}} = \frac{(0.10)\ (0.9)}{(0.01)^2} = 900$$

$$\frac{n_0}{N} = \frac{900}{4000} = 0.225 > 0.05$$

$$n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{900}{1+\frac{900}{4000}} = 735$$

例 2

625 家口에서 非復元 單純 任意 抽出로 50 家口를 뽑았다. 화장실에 所要되는 費用이 한달에 平均 $0.88 이고, 標準誤差가 $0.1 이었다. 이 情報를 利用하여 95 % 信賴水準으로 참값의 10 % 以內의 허용오차를 갖도록 標本의 크기를 決定하여라.

(k) 母集團 分散의 推定

標本의 크기를 決定하기 위해 母集團의 分散을 推定하기 위한 4 가지 方法

方法 1 : 二段 標本 抽出을 하여라.

任意로 n_1 元素를 抽出하여 s_1^2을 計算하여라. 이 값을 s^2과 P를 推定하는 데 使用하여라.

方法 2 : 豫備調査(pilot survey)의 結果를 利用하여라.

i) 方法 1

…이 方法은 가장 信賴度가 높다. 그러나 이 方法은 調査의 完了를 지연시키므로 자주 使用되지는 않는다.

…方法 1 이 使用된다면, 다음 結果들이 응용된다.

($n_1 \leq n$ 이라고 假定한다.)

1) $CV = \sqrt{c}$로 주어졌을 때 $\bar{Y}$를 推定

…假定 ; y_i ~正規分布

s_1^2은 첫번째 標本으로부터의 分散이다.

y_1는 첫번째 標本으로부터의 平均이다.

最終 標本의 크기를 定하기 위하여 부가적으로 원소들이 뽑힌다.

$$n = \frac{s_1^2}{c\bar{y}_1^2}\left(1 + 8c + \frac{s_1^2}{n_1\bar{y}_1^2} + \frac{2}{n_1}\right.$$

$\bar{Y}$를 推定하기 위하여 $\hat{\bar{Y}} = \bar{y}(1-2c)$를 使用하여라.

2) 주어진 分散 V로 $\bar{Y}$를 推定

最終 標本을 決定하기 위해 追加로 뽑는 元素들

$$n = \frac{s_1^2}{V}\left(1 + \frac{2}{n_1}\right)$$

3) 分散 V로 P를 推定

…p_1 =첫번째 標本으로부터의 P를 推定이라고 하자.

…最終 標本의 크기를 定하기 위하여 追加로 元素를 구하라.

하라.

$$n = \frac{p_1 q_1}{V} + \frac{3 - 8 p_1 q_1}{p_1 q_1} + \frac{1 - 3 p_1 q_1}{V n_1}$$

…P를 推定하기 위하여,

$$p = p + \frac{V(1-2p)}{pq}$$

4) $CV = \sqrt{c}$가 주어졌을 때 P를 推定

…追加로 元素를 뽑아

$$n = \frac{q_1}{cp_1} + \frac{3}{p_1 q_1} + \frac{1}{cp_1 n_1}$$

…P의 推定은

$$\hat{p} = p - \frac{cp}{q}$$

例

10 %의 CV로써 P를 推定하고자 한다. 첫번째 標本은 $n_1 = 396$으로서 $p_1 = 0.101$ 이다.

1) 우리가 원하는 精度로서 P를 推定하는 데 必要한 標本의 크기를 決定하여라.

2) 결합된 標本에서 $np = 88$ 이라고 假定하자. P의 最終 推定値를 구하라.

ii) 方法 4

…母集團에 관한 比較的 작은 情報로부터 S^2 에 관해 有用한 推定値를 구할 수 있다.

…데밍(1960)은 單純한 數學的 分布로부터 S^2 을 推定할 수 있음을 보였다. (領域은 h이다.)

分 布	σ 의 近似값
正規分布	h/6
이등변 삼각형	0.2h
직삼각형	0.24h
一樣分布	0.29h
二項分布	$h\sqrt{pq}$

例

미국의 어떤 大學에서 學生들을 4個의 學級으로 分類하였다. 各 學級의 標準偏差는 다음과 같았다.

	1	2	3	4
학 생 수	<1000	1000 ~ 3000	3000 ~ 10,000	10,000 이상
S_i	236	625	2008	10,023

만일 學級의 經濟(boundaries)는 알지만 S_i의 값을 모른다면 어떻게 S_i의 값을 추측할 수 있는가? 學級의 學生數는 200 以上

50,000 名 以下라는 假定하자.

(9) 調査項目이 한個以上일 경우의 標本의 크기

① 標本의 크기를 決定하는 한가지 方法

…重要한 項目에 대한 誤差의 限界를 구하여 各 項目別로 必要한 標本의 크기를 決定하여라.

…이러한 n들 중에서 豫算이 許用하는 範圍안에서 가장 큰 n을 구하라.

…n들이 變化가 크고, 가장 큰 n이 豫算을 초과한다면, 어떤 項目들에 대한 精度를 낮춘다. 또는 精度가 낮은 項目들은 調査에서 除外시킨다.

2. 層化抽出(stratified random sampling)

事前에 母集團에 대한 어떤 情報를 갖고 있다면, 이 情報를 層化하는 데 使用하여서 標本誤差를 減小시킬 수 있다.

(1) 層化 抽出法

…母集團의 元素들이 몇 個의 集團(group)으로 나누어(이 集團을 層이라고 부른다) 各 層에서 標本을 뽑는 方法이다.

…그 層은 國家의 各 地方을 意味하거나, 人口 密度別 地域 기타 宗教集團 等을 나타낸다.

…層化의 目的은 같은 性質을 갖는 元素들을 集團으로 묶어서 分析하는 데 있다.

(2) 層化의 基本的 法則

層안에 있는 元素들은 可能한한 같은 性質을 갖어야 하고, 서로 다른 層에 있는 元素들은 可能한 한 서로 달라야 한다.

대부분의 調査는 多目的으로 施行된다. 그러므로 같은 層안에 元素들은 어떤 重要한 特性에 대하여 같아야 한다. 다른 特性에 대해서는 다를 수도 있다.

(3) 層化를 하는 理由

① 母集團의 어떤 部分集團에 대해서는 精度높은 資料를 원한다.

② 行政的인 便宜性

③ 母集團의 서로 다른 部分에 대해서는 標本 抽出이 다를 수도 있다.

(4) 記 號

N_h = h번째 層안에 있는 元素들의 總數. $h = 1, 2 \cdots, L$.

n_h = h번째 層으로부터 抽出된 標本의 數

y_{hi} = h번째 層의 i번째 元素의 값

$$Y_h = \sum_{i=1}^{N_h} y_{hi}$$

$$y_h = \sum_{i=1}^{n_h} y_{hi}$$

$$\bar{Y}_h = Y_h / N_h$$

$$\bar{y}_h = y_h / n_h$$

$$N_h = \xrightarrow{SRS-WOR} n_h$$

(5) 推定과 分散

① 母集團의 平均, $\bar{Y}$의 推定

…만일 L개의 層이 있다면,

$N = N_1 + N_2 , + \cdots + N_L$

$$\bar{Y} = \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi} / N$$

…層化標本抽出에서, $\bar{Y}$의 不偏推定量은

$$\bar{y}_{st} = \sum_{h=1}^{L} \frac{N_h \bar{y}_h}{N}$$

$$= \sum_{n=1}^{L} W_h \bar{y}_h , \quad \bar{y}_h = \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h}$$

※ 주의 1 : $E(\bar{y}_{st}) = E(\sum_{h=1}^{L} N_h \bar{y}_h / N)$

$$= \bar{Y}$$

※ 주의 2 : $\bar{y}_{st} = \sum_{h=1}^{L} \frac{N_h \bar{y}_h}{N}$

$$= \sum_{h=1}^{L} \frac{\hat{Y}_h}{N}$$

$$= \frac{1}{N} \sum_{h=1}^{L} \hat{Y}_h$$

② $\bar{y}_{st}$의 分散

$$\cdots V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h} , \text{ 단}$$

$f_h = n_h / N_h$ 그리고

$$S_h^2 = \sum_{i=1}^{N_h} \frac{(y_{hi} - \bar{Y}_h)^2}{N_h - 1}$$

③ $V(\bar{y}_{st})$의 推定

… $V(\bar{y}_{st})$의 不偏推定量은

$$v(\bar{y}_{st}) = \sum_{h=1}^{L} w_h^2 (1 - f_h) \frac{s_h^2}{n_h}$$

… $\bar{y}_{st}$의 標本誤差는 $\sqrt{v(\bar{y}_{st})}$이다.

④ Y의 推定

$$\cdots Y = \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}$$

$$\bar{Y} = \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}/N = Y/N$$

즉, $Y = N\bar{Y}$

Y의 不偏 推定量은

$$\hat{Y}_{st} = N\bar{y}_{st}$$

$$E(\hat{Y}_{st}) = R(N\bar{y}_{st}) = NE(\bar{y}_{st}) = N\bar{Y} = Y.$$

⑤ $\hat{Y}_{st}$의 分散

$$\cdots V(\hat{Y}_{st}) = V(N\bar{y}_{st}) = N^2 V(\bar{y}_{st})$$

$$= N^2 \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{s_h^2}{n_h}$$

$$= \sum_{h=1}^{L} N_h^2 (1 - f_h) \frac{s_h^2}{n_h}$$

⑥ $V(\hat{Y}_{st})$의 推定

$$\cdots V(\hat{Y}_{st}) = V(N\bar{y}_{st}) = N^2 V(\bar{y}_{st})$$

$$= N^2 \sum_{h=1}^{L} W_h^2 (1-f_h) \frac{s_h^2}{n_h}$$

$$= \sum_{h=1}^{L} N_h^2 (1-f_h) \frac{s_h^2}{n_h}$$

… $V(\hat{Y}_{st})$의 不偏推定量은

$$v(Y_{st}) = \sum_{h=1}^{L} N_h^2 (1-f_h) \frac{s_h^2}{n_h}$$

… $\hat{Y}_{st}$의 標準誤差는 $\sqrt{v(\hat{Y}_{st})}$로 推定된다.

⑦ P의 推定

… P=母集團에서 어떤 推性을 갖는 比率 定義:

$$y_{hi} = \begin{cases} 1, & \text{層 h의 i번째 元素가 그 特性을 갖을 때.} \\ 0, & \text{그 以外의 경우.} \end{cases}$$

그런데,

$$P_{st} = \bar{y}_{st} = \sum_{h=1}^{L} \frac{N_h \bar{y}_h}{N} = \sum_{h=1}^{L} \frac{N_h P_h}{N}, \quad \text{단}$$

P_h = 層 h의 標本比率

⑧ P_{st}의 分散

… $V(P_{st}) = V(\bar{y}_{st})$

$$= \sum_{h=1}^{L} W_h^2 (1-f_h) \frac{s_h^2}{n_h}$$

$$= \sum_{h=1}^{L} W_h^2 (1-f_h) \frac{P_h Q_h}{n_h} \left(\frac{N_h}{N_h - 1} \right)$$

$$\fallingdotseq \sum_{h=1}^{L} W_h^2 (1-f_h) \frac{P_h Q_h}{n_h}$$

⑨ $V(P_{st})$의 推定

…$V(P_{st})$의 不偏 推定量은

$$v(P_{st})=\sum_{h=1}^{L} W_h^2(1-f_h)\frac{p_h q_h}{n_h-1}$$

…P의 標本誤差는 $\sqrt{v(P_{st})}$이다.

⑩ 母集團 比의 推定

$R=Y/X$의 推定値는

$$\hat{R}_{st}=\hat{Y}_{st}/X_{st}=\bar{y}_{st}/\bar{x}_{st}$$

$V(\hat{R})$의 推定値는

$$v(\hat{R}_{st})=\frac{1}{(\hat{X}_{st})^2}\left[v(\hat{Y}_{st})+\hat{R}_{st}^2 v(\hat{X}_{st})-2\hat{R}_{st}\,cov(\hat{Y}_{st},\hat{X}_{st})\right]$$

단,

$$cov(\hat{Y}_{st},\hat{X}_{st})=\sum_{h=1}^{L} N_h^2(1-f_h)\frac{s_{hxy}}{n_h}$$

그리고,

$$s_{hxy}=\sum_{i}^{n_h}\frac{(y_{hi}-\bar{y}_h)(x_{hi}-\bar{x}_h)}{n_h-1}$$

例

標本調査가 소의 總數를 推定하기 위하여 設計되었다. 母集團은 2072農家로서 土地 所有에 따라 5個의 層으로 나누었다. 各 層에서 非復元 單純 任意 抽出로 n_h農家를 뽑았다. 標本으로 뽑힌 農家의 소의 總數는 다음과 같다.

y_{hi} = 層 h의 i번째 農家의 소의 數

層別 農家의 소의 總數

層(acres)	層(N_h)의 農家의 數	標本(n_h)의 農家의 數	$\sum_{i=1}^{n_h} y_{hi}$	$\sum_{i=1}^{n_h} y_{hi}^2$
(1)	(2)	(3)	(4)	(5)
6～15	635	153	619	5579
16～30	570	138	1423	24253
31～50	475	115	1758	34082
51～75	303	73	1691	51419
76～100	89	21	603	18305
모든 層	2072(N)	500	6904	133658

1) 農家 當 平均 소의 數, $\bar{Y}_{st}$ 를 推定하여라.

2) 이 推定値의 標準誤差는 얼마인가?

(6) 實習－標本의 크기 推定

① 背 景 : 農家의 經濟에 대한 調査를 하고자 한다. 調査의 內容은 耕地面積, 山林面積, 소의 數, 그리고 젖소의 數 等이다.

調査地域인 洲은 두개의 地理的 地區로 나누어지고, 各 地區는 다시 小地區로 나눠진다. 農家의 목록(list)는 標本名簿로 使用된다(表1). 이 標本名簿는 그

洲의 모든 農家를 包含한다고 假定하자.

② 調査目的 : 調査目的은 다음 項目의 母數를 推定하고, 그 推定値의 信賴度를 測定하는 것이다.

곡물 경작 面積

옥수수 경작 面積

山林 面積

農場의 소의 總數

農場의 젖소의 總數

곡물 경작 農場 當 平均 面積

옥수수 경작 農場 當 平均 面積

山林에 있는 農場 當 平均 面積

農場 當 소의 平均 數

農場 當 젖소의 平均 數

山林을 1 에이커(acre) 以上 갖은 農場의 比率

25마리 以上의 젖소를 갖는 農場의 比率

③ 問 題 : 母集團에 대하여 아무것도 알려지지 않았으므로, 母集團 分散과 變異係數를 推定하기 위하여 豫備 標本 調査를 한다. 標本 名簿에서 單純 確率 標本을 10 個 뽑아라. 그 標本의 값을 <表 1 a>에 記錄하라. 또 <表 1 b>를 完成하여라.

推定値가 95 % 信賴水準에서 <表 1 c>에 必要한 精度를 갖도록 標本의 크기를 決定하여라.

<表 1 b>에서의 S^2과 V의 推定値들로서 <表 1 c>

에 必要한 精度를 갖는 標本의 크기를 推定하여라.

<表1 c>를 完成하라.

<表1 a> (實習 欄)

<表1 b> (實習 欄)

<表1 c> 單純確率 標本으로부터의 標本의 크기 推定

母 數	精 度	推定된 標本의 크기
곡물에 대한 全面積	±1000	
山林에 대한 全面積	±1500	
옥수수 재배農場에 대한 平均面積	±5	
소의 總數	±500	
25十젖소를 갖는 農場의 比率	±0.1	

調査(survey)를 위해서 얼마의 標本이 必要한가? 理由는?

④ 層에 標本 配定(allocation)

ⓐ 層化 標本의 定義는 各 層의 標本의 크기를 定해 주지는 않는다.

ⓑ 層別 標本의 크기를 定하는 데 두가지 重要한 基準이 있다.

i) 便宜性 (conveniency)

ii) 正確性

ⓒ 比例 配定은 두번째 基準을 만족시키며, 最小 標準 誤差를 提供한다.

⑤ 比例 配定

ⓐ 比例 配定일 때는 各 層에서 같은 比率로 標本이 抽出된다.

즉, $n_h = n\frac{N_h}{N}$ $\quad h = 1, \cdots\cdots, L$

ⓑ 推定 : 比例 配定에 있어서 n_h 대신 $n(N_h/N)$을 대입해도 좋다.

i) $\bar{y}_{st(prop)} = \sum_{h=1}^{L} \frac{N_h \bar{y}_h}{N}$

ii) $V(\bar{y}_{st(prop)}) = \sum_{h=1}^{L} W_h^2 (1-f_h)\frac{S_h^2}{n_h}$

iii) $\frac{N-n}{N} \approx 1$ 이면, 즉 $\frac{n}{N}$이 무시할 수 있으면,

$$v(\bar{y}_{st(prop)}) \fallingdotseq \sum_{h=1}^{L} W_h \frac{S_h^2}{n}$$

※ 주의 : 比例 配定은 決定値들을 製表(tabulate)하기 쉬운 매력이 있다. 서로 다른 層들을 各各 製表할 必要가 없다. 모든 標本資料를 合쳐서 使用할 수 있다.

⑥ 最適 配定(optimal allocation)

最適 配定에서 주어진 費用, c에 대하여 分散이 最小化되도록 n_h를 配定하거나, 주어진 分散에 대하여 費用이 最小化되도록 n_h를 配定한다.

費用 함수 :

c = 調査의 總費用

다음의 費用 함수를 假定하자.

$$C = c_0 + \sum_{h=1}^{L} c_h n_h, \text{ 단}$$

C_0 = 固定費用,

$\sum_{h=1}^{L} c_h n_h$ = 變數費用,

c_h = 層안에 한 元素 當 드는 費用

最適 配定에서

$$n_h = \frac{n N_h S_h / \sqrt{c_h}}{\sum N_h S_h / \sqrt{c_h}}$$

만일 모든 c_h가 같다면,

$$n_h = n \frac{N_h S_h}{\sum N_h S_h}$$

이것을 네이만 配定이라고 부른다.

더 큰 標本이 必要되는 層은

1. 그 層이 클 때

2. 그 層의 分散이 클 때

3. 그 層의 標本 費用이 저렴할 때

推 定:

1. $\bar{y}_{st(opt)} = \sum_{h=r}^{L} \frac{N_h \bar{y}_h}{N}$

2. 모든 層에 대하여 c_h가 같다면,

$$v(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 (1-f_h) \frac{s_h^2}{n_h}$$

※ 주의 : 最適 配定은 分散이 큰 資料에 가장 效果的이다. 즉, 個人의 收入, 추수량 등.

⑦ 分散이 計算

ⓐ 比例 配定할 때에 $V(P_{st})$?

$$V(P_{st}) = \frac{1-f}{n} \sum_h W_h P_h Q_h$$

ⓑ 最適 配定할 때에 $V(P_{st})$?

$$V(P_{st}) = \frac{\sum W_h \sqrt{P_h Q_h}}{n} - \frac{\sum W_h P_h Q_h}{N}$$

例 1

最近 센서스 資料에 의하면, 農場 當 소의 數는 아래와 같다. 農場들을 土地 所有(acre)에 따라 層化한다.(L = 5) 各 層(N_h)의 農場의 總數는 아래 表와 같다.

標本이 n＝500개의 農場에 대하여, 各 層의 標本의 크기를 計算하여라.

1. 比例 配定

2. 最適 配定

各 層의 總數, 標準偏差, 費用 等의 대략적 推定

層 h	N_h	S_h	C_h
Ⅰ	37800	28.5	3.50
Ⅱ	52600	18.6	2.75
Ⅲ	82000	27.6	2.25
Ⅳ	41600	21.2	3.00
Ⅴ	28800	16.8	2.50

⑧ 層의 定義

層을 定義하는 것은 標本 設計할 때에 첫 段階이다.

層을 決定하는 데 基礎가 되는 事項:

1. 事前知識
2. 個人的 직관(intuition)
3. 判斷(judgement)
4. 客觀的 통계적 情報

層을 決定하는 최선의 一般的인 過程은 없다.

가장 效果的인 層化 變數는 測定하고자하는 特性이다. 實際로 이것이 불가능하다. 이 경우에 測定하고자하는 變數에 가장 相關이 높은

變數의 層化는 分散을 감소시킨다. 費用을 考慮하면 層化를 하는 데는 더 單純해지나, 덜 效果的이다.

例

어떤 市의 家族 收入에 대한 標本分布

層化에 대한 여러가지의 可能性들;

1. 收入別로 差異가 나도록 地域의 層化

2. 家族들을 住居地別로 分類하여 層化

3. 住居 形態別로 層化 (外型的)

4. 住居 費用別로 層化

家族에 대한 平均 收入을 推定하고자 한다. 위의 方法中 어느 層化가 가장 效果的인가 ?

어느 方法이 가장 費用이 드는가 ?

⑨ 實習 : 層化 任意 抽出

부록Ⅳ의 3地帶 (zone) 에서 2마을씩을 非復元 單純 任意 抽出로 뽑는다. 家口數와 人口數를 地帶別로 그리고 3地帶를 합쳐서 推定하여라.

1) 總 家口數와 標準誤差

2) 總 人口數와 標準誤差

3) 平均 家口數와 標準誤差

연습 [1] 두가지 方法으로 層化 (크기와 保有期間別) 했을 때 比例層化 標本으로 300農場을 뽑았을 때 總 生産量에 대한 標準誤差를 計算하여라.

연습[2] 어느 層化 方法이 比例 標本에 대하여 效果的인가?

연습[3] 單純 任意 抽出로 300 農場을 抽出하여 總 生産量에 대한 標準誤差를 計算하여라.

연습[4] 두가지 層化 方法에 대하여, 300 農場의 標本에 대한 最適 配定을 使用하고 다음을 計算하여라.

a) 各 層안에 標本 農場의 數

b) 總 生産量의 推定値에 대한 標準誤差

연습[5] 이 分析에서 標本을 配定하는 5가지 方法中 어느 方法을 추천하겠는가?

解答

[1] $V(\hat{y}_{st(prop)}) = N^2 V(\bar{y}_{st(prop)})$

$$= N^2 \sum_{h=1}^{L} \frac{N_h}{N} \frac{s_h^2}{n}$$

$$= \frac{N}{n} \sum_{h=1}^{L} N_h S_h^2 \text{ (fpc 무시)}$$

農家의 크기

$$\cdots V(\hat{y}_{st(prop)}) = \frac{5900}{300}[\ 349,620,000,000\ ]$$

$$= 6.87586 \times 10^{12}$$

$$\sqrt{V(\hat{y}_{st(prop)})} = \$\ 2,622,186$$

保有期間에 의한 層化

$$\cdots V(\hat{y}_{st(prop)}) = \frac{5900}{300}[\ 289,200,000,000\ ]$$

$$= 5,6876 \times 10^{12}$$

$$\sqrt{V(\hat{y}_{st(prop)})} = \$\ 2,384,869$$

[2] 保有期間에 의한 層化가 가장 效率的이다. 왜냐하면 標準誤差가 더 작다.

[3] 單純 任意 抽出의 크기 $n = 300$ 에 대하여

$$\cdots \sqrt{V(\hat{y})} = \sqrt{N^2 (1-f) \frac{S^2}{n}}$$

$$= \sqrt{(5900)^2 \left(1 - \frac{300}{5900}\right) \frac{97,000,000}{300}}$$

$$= \$\ 3,268,476$$

※ 주의 : fpc는 무시된다.

[4] 크기에 의한 層化 (最適 配定)

a) $n_h = \frac{N_h S_h}{\sum N_h S_h}$ (c_h 들이 같다고 假定한다.)

$$n_1 = \frac{300(2502)}{38373} = 19.6 \approx 20$$

$$n_2 = \frac{300(6192)}{38373} = 48.4 \approx 48$$

.....

$$V(\hat{y}_{st(opt)}) = N^2 V(\bar{y}_{st(opt)})$$

$$= N^2 \left[\frac{1}{n} (\sum W_h S_h)^2 - \frac{\sum W_h S_h^2}{N}\right.$$

$$= N^2 \left[\frac{1}{n} \left(\frac{\sum N_h S_h}{N} \right)^2 - \frac{\sum \frac{N_h}{N} S_h^2}{N} \right]$$

$$= \frac{(\sum N_h S_h)^2}{n} - \sum N_h S_h^2$$

農場의 크기에 의한 層化를 위하여

$$V(\hat{y}_{st(opt)}) = \frac{(383,730,000)^2}{300} - (349,620,000,000)$$

$$= 4.5586704 \times 10^{12}$$

$$\sqrt{V(\hat{y}_{st(opt)})} = \$\ 2,135,104$$

保有期間(tenure)에 의한 層化에 대하여

$$n_1 = \frac{300(19536)}{39533} = 148.3 \approx 148$$

$$n_2 = \frac{300(6923)}{39533} = 52.5 \approx 53$$

.....

$$V(\hat{y}_{st(opt)}) = (\sum N_h S_h)^2 / n - (\sum N_h S_h^2)$$

$$= \frac{(39,533,000)^2}{300} - 289,200,000,000$$

$$= 4.920327 \times 10^{12}$$

$$\sqrt{V(\hat{y}_{st(opt)})} = \$\ 2,218,181$$

[5] 層化

農場의 크기	$\sqrt{V(\hat{y}_{st})}$	
比例	$ 2,622,186	
最適	$ 2,135,104	** 추천된다.
保有期間		
比例	$ 2,384,869	
最適	$ 2,218,181	
單純任意標本	$ 3,268,476	

3. 一段 集落 抽出法 : 같은 크기의 集落

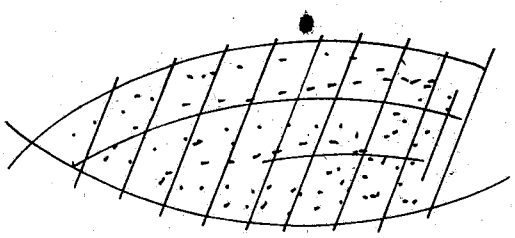
(1) 序 論

集落들을 標本 單位로 뽑아서 集落들 안에 모든 元素들을 調査하는 데는 몇가지 有益한 點이 있다. 이것을 一段 集落 標本(single stage cluster sampling)이라고 한다.

例

같은 크기의 集落들

왓슨 市



N = 5 個의 集落들이 있다.

各 集落은 M = 40 家口

不利：集落 標本은 單純 任意 標本보다 덜 效果的일 수 있다.

有利：1. 母集團을 作成하는 데 費用이 많이 들거나, 믿을 만한 母集團을 作成할 수 없을 수도 있다.

2. 資料 蒐集 等에 費用을 절감할 수 있다.

3. 非標本誤差를 줄일 수 있다.

(2) 標本計劃은 다음과 같다.

N集落들 $\xrightarrow{\text{SRS-WOR}}$ n集落들

M單位들 ——→ M單位들

※ 주의；같은 크기의 集落들에서 一段 集落 抽出을 하는 데, 各 集落들은 M單位들을 包含한다. 一般的으로, 하나의 集落은 M_i 單位들을 包含한다. $M_i = M$이라고 하자.

(또는 $M_i = \bar{M}$ 단, $\bar{M} = \sum_{i=1}^{N} M_i / N$)

Y에 대한 推定値는

$$Y = N\bar{Y} = N\sum_{i=1}^{n} \frac{y_i}{n} = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}$$

단, y_{ij} = i번째 集落에 j째 單位의 값

$$y_i = \sum_{j=1}^{M} y_{ij} = \text{集落 i에 대한 總計}$$

$$\bar{Y} = \sum_{i=1}^{n} \frac{y_i}{n} = \text{平均 集落値}$$

다음 分散을 計算할 수 있다.

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

$$= \frac{N(1-\frac{n}{N})}{n} \left(\frac{1}{N-1}\right) \sum_{i=1}^{N} (y_i - \bar{Y})^2$$

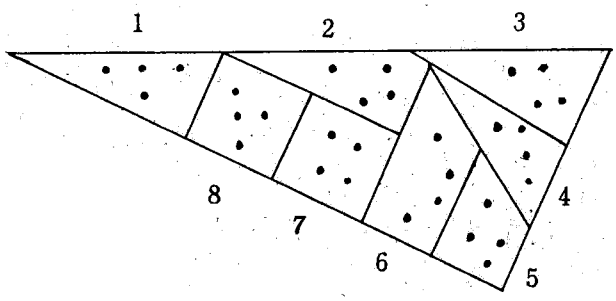
단, $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}$

그러면,

$$V(\hat{Y}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \left[\left(\frac{1}{n-1}\right) \sum_{i=1}^{n} (y_i - \bar{y})^2 \right]$$

단, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{M} y_{ij}$

例



• = 住居單位

N = 8 集落

$\bar{M}$ = 4 住居單位

위의 그림과 같이 32 住居 單位가 8 개의 集落으로 나뉘어졌다. 標本으로 n = 3 集落을 抽出하여, 이 3 個의 集落안에 모든 住居 單位를 調査한다. 標本 資料에서 y_{ij} = i 集落의 住居 單位 j안에 사람들의 數

$y_{21}=9$	$y_{41}=6$	$y_{81}=10$
$y_{22}=7$	$y_{42}=5$	$y_{82}=11$
$y_{23}=6$	$y_{43}=8$	$y_{83}=3$
$y_{24}=4$	$y_{44}=9$	$y_{84}=8$
$y_2=26$	$y_4=28$	$y_8=32$

$$\hat{Y}=\frac{N}{n}\sum y_i=\frac{8}{3}(26+28+32)=229.33\fallingdotseq 229\text{名}$$

$$V(\hat{Y})=\frac{N^2}{n}\left(1-\frac{n}{N}\right)\left[\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(y_i-\bar{y})^2\right]$$

$$=\frac{8^2}{3}\left(1-\frac{3}{8}\right)\left[\left(\frac{1}{3-1}\right)\{(26-28.67)^2+\cdots+(32-28.67)^2\}\right]$$

$$=124.4$$

$$\sqrt{V(\hat{Y})}=11.15\text{名}$$

95 % 信賴區間은

$$229\pm 2(11.15)$$

$$(206.7,\ 251.3)$$

(3) 母集團 平均 $\bar{\bar{Y}}$의 推定

$\bar{Y}=\sum_{i=1}^{N} y_i/N$ =集落의 平均

$\bar{\bar{Y}}=\sum_{i=1}^{N} y_i/NM=\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}/NM$=元素에 대한 平均

N개 集落의 母集團으로부터 n개 集落을 單純 確率 標本을 抽出

$\bar{\bar{Y}}$의 推定値는

(各 集落은 M個의 元素를 包含한다.)

$$\bar{\bar{y}} = \sum_{i=1}^{n} y_i / nM$$

$$= \bar{y}/M \quad (\bar{\bar{y}}\text{는 不偏이다.})$$

$$V(\bar{\bar{y}}) = V(\bar{y}/M) = \frac{1}{M^2} V(\bar{y}) = \frac{1}{M^2}\left[\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\right)\sum \frac{(y_i-\bar{y})^2}{N-1}\right]$$

$$v(\bar{\bar{y}}) = \frac{1}{M^2}\left[\left(1-\frac{n}{M}\right)\left(\frac{1}{n}\right)\sum_{i=1}^{n} \frac{(y_i-\bar{y})^2}{n-1}\right]$$

단, $\bar{y} = \sum_{i=1}^{n} y_i / n$

$\bar{\bar{Y}}$에 대한 $A(1-\alpha)\%$ 信賴區間은

$$\bar{\bar{y}} \pm t\sqrt{V(\bar{\bar{y}})}$$

例 ; 같은 크기의 集落

32 住居 單位를 8個의 集落으로 나누어, 이 중 다음과 같은 3個의 集落을 抽出하였다.

$y_{21} = 9$	$y_{41} = 6$	$y_{81} = 10$
$y_{22} = 7$	$y_{42} = 5$	$y_{82} = 11$
$y_{23} = 6$	$y_{43} = 8$	$y_{83} = 3$
$y_{24} = 4$	$y_{44} = 9$	$y_{84} = 8$
$y_2 = 26$	$y_4 = 28$	$y_8 = 32$

住居 單位 當 住居人의 平均數의 推定値는

$$\bar{\bar{y}} = \sum_{i=1}^{3} \frac{y_i}{nM} = \frac{(26+28+32)}{3.4} \fallingdotseq 7.17$$

$$V(\bar{\bar{y}}) = \frac{1}{M^2}\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\right)\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{n-1} \quad \text{단, } \bar{y} = \sum_{i=1}^{n}\frac{y_i}{n} = \frac{86}{3} = 28.67$$

$$\fallingdotseq \frac{1}{4^2}\left(1-\frac{3}{8}\right)\frac{1}{3}[(26-28.67)^2 + (28.67)^2 + (32-28.67)^2]/(3-1)$$

$$\fallingdotseq 0.12$$

$\bar{\bar{Y}}$에 대한 95% 信賴區間은

$$7.17 \pm 2\sqrt{0.12}$$

$$(6.48,\ 7.86)$$

(4) 母集團 比率, P의 推定

各 集落은 M個의 元素로 構成되어 있다. N個의 集落으로 構成된 母集團에서 어떤 集團에 속하는 元素들의 比率을 推定하고자 한다.

$$N \xrightarrow{\text{SRS-WOR}} n$$

$$M \longrightarrow M$$

$$y_{ij} = \begin{cases} 1, & i\text{째 集落에 } j\text{번째 元素가 集團에 속할 때} \\ 0, & \text{그 以外에 경우} \end{cases}$$

그러면,

y_i = 集落 i에 元素들 中에 그 集團에 속하는 元素의 數

$P_i = y_i/M$ = i번째 集落있는 元素들 중 그 集團에 속하는 比率

$$P = \sum_{i=1}^{N} y_i / NM = Y / NM = \sum_{i=1}^{N} P_i / N$$

母集團의 元素들이 1이나 0의 값을 갖는다면 P는 $\bar{\bar{Y}}$에 대응한다.

이와 같이, P의 推定値는

$$\hat{P} = \frac{1}{nM} \sum^{n} y_i = (\frac{1}{n}) (\sum_{i=1}^{n} y_i / M) = \frac{1}{n} \sum_{i=1}^{n} P_i$$

$$V(\hat{P}) = \frac{1}{n} (1 - \frac{n}{N}) \sum_{i=1}^{N} \frac{(P_i - P)^2}{N-1} \quad \text{그리고}$$

$$v(\hat{P}) = \frac{1}{n} (1 - \frac{n}{N}) \sum \frac{(P_i - P)^2}{N-1} \quad \text{단,} \quad P = \frac{1}{n} \sum_{i=1}^{n} P_i$$

P에 대한 信賴區間은

$$\hat{P} \pm t \sqrt{v(\hat{P})}$$

例 ; 같은 크기의 集落들 :

N = 8인 集落들에서 n = 3인 集落들을 抽出하는 例를 다시 생각해 보자. 水道가 있는 住居의 比率을 推定하고자 한다.

$$y_{ij} = \begin{cases} 1, & i\text{번째 集落안에 } j\text{번째 住居가 水道를 갖을 때} \\ 0, & \text{그 以外의 경우} \end{cases}$$

다음과 같은 資料를 얻었다 ;

$y_{21} = 1$	$y_{41} = 1$	$y_{81} = 1$
$y_{22} = 0$	$y_{42} = 0$	$y_{82} = 1$
$y_{23} = 0$	$y_{43} = 1$	$y_{83} = 1$
$y_{24} = 0$	$y_{44} = 0$	$y_{84} = 1$
1	2	4

그러면, $\hat{P} = \frac{1}{nM}\sum_{i=1}^{N} y_i$

$= \frac{1}{3.4}(1+2+4) = 7/12 \fallingdotseq 0.58$

$v(\hat{P}) = \frac{1}{n}(1-\frac{n}{N})[\frac{\sum_{i=1}^{3}(P_i-P)^2}{n-1}]$

$= \frac{1}{3}(1-\frac{3}{8})[\frac{(0.25-0.58)^2+(0.5-0.58)^2+(1-0.58)^2}{3-1}]$

$\fallingdotseq 0.0303$

P에 대한 95% 信賴區間은

$P \pm 2\sqrt{v(P)}$

$0.58 \pm 2\sqrt{0.0303}$

0.58 ± 0.349

(0.231, 0.929)

(5) 母集團 比의 推定

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}}{\frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} x_{ij}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}/n\cdot M}{\sum_{i=1}^{n}\sum_{j=1}^{M} x_{ij}/n\cdot M} = \bar{\bar{y}}/\bar{\bar{x}}$$

$$v(\hat{R}) = \frac{1}{\hat{x}^2}[v(\hat{Y}) + R^2 v(x) - 2\hat{R}\,cov(\hat{X}, \hat{Y})]$$

$$\mathrm{cov}(\hat{X},\ \hat{Y}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$$

$$\mathrm{cov}(\hat{X},\ \hat{Y}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{1}{n-1}\right)\left(\sum_{i=1}^{n}y_i x_i - n\bar{y}\bar{x}\right)$$

例

$N = 8 \xrightarrow{\text{SRS-WOR}} n = 3$ (集落)

$M \longrightarrow M = 4$ (住居單位)

y_{ij} = 集落 i 안에 家口 j 에 대한 家口收入

x_{ij} = 集落 i 안에 家口 j 에 대해 收入이 있는 사람의 數

$y_{21} = \$\ 30{,}000$	$x_{21} = 3$
$y_{22} = 25{,}000$	$x_{22} = 2$
$y_{23} = 25{,}000$	$x_{23} = 2$
$y_{24} = 28{,}000$	$x_{24} = 1$
$y_2 = \$108{,}000$	$x_2 = 8$

$y_{41} = 40{,}000$	$x_{41} = 2$
$y_{42} = 25{,}000$	$x_{42} = 1$
$y_{43} = 30{,}000$	$x_{42} = 3$
$y_{44} = 30{,}000$	$x_{43} = 4$
$y_4 = 125{,}000$	$x_4 = 10$

$y_{81} = 24{,}000$	$x_{81} = 2$
$y_{82} = 29{,}000$	$x_{82} = 3$
$y_{83} = 28{,}000$	$x_{83} = 1$
$y_{84} = 32{,}000$	$x_{84} = 2$
$y_8 = 113{,}000$	$x_8 = 8$

$$\hat{Y} = \frac{N}{n}\sum y_i = 922666.67$$

$$\hat{X} = \frac{N}{n}\sum x_i = 69.33$$

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\hat{Y}}{\hat{X}} = \frac{\frac{N}{n}\sum y_i}{\frac{N}{n}\sum x_i} = \frac{\frac{8}{3}\ (108,000 + 125,000 + 113,000)}{\frac{8}{3}\ (8 + 10 + 8)}$$

$$= \$13,308.33$$

$$v(\hat{Y}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{8^2}{3}\left(1 - \frac{3}{8}\right)(76333334)$$

$$= 1.01777 \times 10^9$$

$$v(\hat{X}) = 17.778$$

$$\mathrm{cov}(\hat{X}, \hat{Y}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\left(\sum x_i y_i - n\bar{x}\bar{y}\right)$$

$$= \frac{8^2}{3}\left(1 - \frac{3}{8}\right)\frac{1}{3-1}\left[3,018,000 - 3(11533.33)(8.666)\right]$$

$$= 128,888.9$$

$$v(\hat{R}) = \frac{1}{69.33}\left[1.01777 \times 10^9 + (13,308.33)^2\ (17.778)\right.$$

$$\left. - 2(13,308.33)(128888.9)\right]$$

$$= 10,614,000$$

$$\sqrt{v(\hat{R})} = \$\ 3257.91$$

(b) 같은 크기의 單位를 갖는 副次標本 (subsample)

① 序 論

다음과 같은 二段 抽出을 생각하자.

$$N \xrightarrow{SRS} n \text{ (一段)}$$

$$M \xrightarrow{SRS} n \text{ (二段)}$$

二段 抽出의 利點은 一段 抽出보다 융통성이 있다는 데 있다. 多段 抽出은 費用이 덜 들고, 操作上 더 용이하지만 二段 集落 抽出보다는 별로 낫지 못하다. 標本 變異度를 생각해 볼 때, 多段 抽出이 單純 任意 抽出보다 덜 效率的이지만 集落 抽出보다는 더 效率的이다 (이것은 全體 標本의 크기가 固定되었다는 假定일 때이다).

② 記 號

y_{ij} = i번째 첫 單位 (primary unit) 안에 j번째 副次單位에 대한 값

$\bar{y}_i = \sum_{j=1}^{m} y_{ij}/m$ = i번째 첫 單位에 있는 2번째 單位에 대한 標本 平均

$$\bar{\bar{y}} = \sum_{i=1}^{n} \bar{y}_i / n = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}/mn$$

$$S_1^2 = \frac{\sum_{i=1}^{N} (\bar{Y}_i - \bar{\bar{Y}})^2}{N-1}$$

$$S_2^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{Y}_i)^2}{N(M-1)}$$

단, $Y_i = \sum_{j=1}^{M} x_{ij}$

③ 二段階 單位에 대한 母集團 平均, $\bar{\bar{Y}}$의 推定

ⓐ $\bar{\bar{Y}}$의 推定値는,

$$\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$$

※ 주의 : $E(\bar{\bar{y}}) = \bar{\bar{Y}}$

ⓑ $v(\bar{\bar{y}}) = (\frac{N-n}{N})\frac{S_1^2}{n} + (\frac{M-m}{M})\frac{S_2^2}{nm}$

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

※ 주의 : $E(S_1^2) = S_1^2 + \frac{(1-f_2)S_2^2}{m}$

단, $f_1 = \frac{n}{N}$ 그리고 $f_2 = \frac{m}{M}$

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{\bar{y}})^2$$

$$S_2^2 = \frac{1}{n(m-1)}\sum_i\sum_j(y_{ij} - \bar{y}_i)^2$$

例

한 地域이 N = 4 개의 集落으로 나누어졌다. 各 集落은 4 家口로 구성되었다.

y_{ij} = 集落 i 의 j번째 家口에 居住하는 사람의 數

集落 :

1	2	3	4
3	8	4	7
10	3	6	2
9	6	3	6
8	4	8	4
6	5	6	6

다음과 같은 標本計劃을 생각하자.

$$4 \xrightarrow{\text{SRS}} 2$$

$$5 \longrightarrow 3$$

標本 結果는,

$y_{41}, y_{42}, y_{45}, y_{32}, y_{34}, y_{35}$

$$\bar{\bar{y}} = \frac{1}{2}\sum_{i=1}^{2}\bar{y}_i = \frac{1}{2}\left(\frac{15}{3}+\frac{20}{3}\right) \fallingdotseq 5.83$$

$$S_1^2 = \frac{1}{2-1}\sum^{n}(\bar{y}_i - \bar{\bar{y}})^2 = (5-5.83)^2 + (6.67-5.83)^2 = 1.3945$$

$$S_2^2 = \frac{1}{2(3-1)}\sum\sum(y_{ij}-\bar{y}_i)^2$$

$$= \frac{1}{4}\left[(7-5)^2+(2-5)^2+(6-5)^2\right] + \left[(6-6.67)^2+(8-6.67)^2+\right.$$

$$\left.(6-6.67)^2\right] = 4.17$$

$$v(\bar{\bar{y}}) = \frac{(1-f_1)s_1^2}{n} + \frac{f_1(1-f_2)s_2^2}{mn}$$

$$= \frac{\left(1-\frac{2}{4}\right)(1.3945)}{2} + \frac{\left(\frac{2}{4}\right)\left(1-\frac{3}{5}\right)(4.17)}{3\cdot 2}$$

$$\fallingdotseq 0.35 + 0.14 = 0.49$$

$$\sqrt{v(\bar{\bar{y}})} = 0.7$$

$\bar{Y}$에 대한 95% 신뢰구간은 $5.83 \pm 2(0.7) \fallingdotseq (4.43, 7.23)$

④ Y의 推定

$$N \xrightarrow{SRS} n$$

$$M \xrightarrow{SRS} m$$

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n}\hat{Y}_i$$

$$= \frac{N}{n}\sum_{i=1}^{n}\frac{M}{m}\sum_{j=1}^{m}y_{ij} \qquad 단, \quad \frac{M}{m}\sum_{j=1}^{m}y_{ij} = \hat{Y}_i$$

$$= \frac{NM}{nm}\sum_{i}^{n}\sum_{j}^{m}y_{ij}$$

$$= NM\bar{\bar{y}}$$

$$v(\hat{Y}) = v(NM\bar{\bar{y}}) = N^2M^2\,v(\bar{\bar{y}})$$

$$= N^2M^2\left[(1-f_1)\frac{s_1^2}{n} + f_1(1-f_2)\frac{s_2^2}{nm}\right]$$

例

4 節의 標本에 대하여, 推定된 사람의 數와 그 標準 誤差는,

$$\hat{Y} = NM\bar{\bar{y}} = 4(5)(5.83) = 116.67$$

$$v(\hat{Y}) = (4)^2(5)^2(0.49) = 196$$

$$\sqrt{v(\hat{Y})} = 14$$

⑤ 比率의 推定

元素들을 두 集團으로 분류하고, 첫 集團에 속하는 元素의 比率을 推定한다.

y_{ij} = 1, ij번째 元素가 첫 集團에 속한다면,

0, 그 이외의 경우

$$P_i = \frac{a_i}{m} = \sum_{j=1}^{m} y_{ij}/m = \bar{y}_i$$

그러면, P의 推定値는,

$$\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} y_{ij}/m\right) = \frac{1}{n}\sum_{i=1}^{n} P_i = P$$

그리고,

$$v(\bar{P}) = \frac{(1-f_1)s_1^2}{n} + \frac{f_1(1-f_2)s_2^2}{mn}$$

단, $s_1^2 = \sum_{i=1}^{n}\frac{(p_i-p)^2}{n-1}$

$s_2^2 = \frac{m}{n(m-1)}\sum_{i=1}^{n} p_i q_i$

$p_i = \bar{y}_i$ 그리고 $p = \bar{\bar{y}}$

⑥ 比의 推定

$$\hat{R} = \hat{Y}/\hat{X} = \frac{NM\bar{\bar{y}}}{NM\bar{\bar{x}}} = \bar{\bar{y}}/\bar{\bar{x}}$$

$$v(\hat{R}) = \frac{1}{\hat{x}^2}\left[v(\hat{Y}) + R^2 v(\ ) - 2R\,\mathrm{cov}(X, Y)\right]$$

$$\mathrm{cov}(\hat{X}, \hat{Y}) = N^2M^2\left[(1-f_1)\frac{S_1xy}{n} + f_1(1-f_2)\frac{S_2xy}{nm}\right]$$

$$S_1xy = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{\bar{y}})(\bar{x}_i - \bar{\bar{x}})$$

$$S_2xy = \frac{1}{n(m-1)} \sum_i \sum_j (y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i)$$

⑦ 實 習

10,000家口의 調査表를 200個씩 묶어 50더미 (stack)로 나누었다. 이 중 5더미를 標本으로 뽑아 각 더미의 모든 調査表를 調査하고자 한다.

任意로 뽑은 5더미에서 다음과 같은 情報를 얻었다.

"前年度 일년간의 收入"에 대하여

y_1 = \$ 1,116,000 y_2 = \$ 1,246,000 y_3 = \$ 825,400

y_4 = \$ 771,800 y_5 = \$ 1,179,800

住宅 所有者의 數에 대하여

(a) y_1 = 40, y_2 = 53, y_3 = 28, y_4 = 37, y_5 = 32

(a) 10,000家口에 대한 平均 家口收入을 推定하여라.

(b) 10,000家口에 대한 住宅 所有 比率을 推定하여라.

추가로 다음는 情報를 수집하였다.

"收入이 있는 사람의 數"에 대하여

(c) y_1 = 500, y_2 = 460, y_3 = 420, y_4 = 400, y_5 = 520

(c) 10,000家口에 대하여 收入이 있는 사람에 대한 平均收入을 推定하여라.

(d) 위의 (a), (b) 그리고 (c)에 대한 分散을 推定하여라.

解

N = 50 ⟶ n = 5

M = 200 ⟶ $\mathcal{M}$ = 200

(a) $\bar{\bar{y}} = \frac{\sum\sum y_{ij}}{n M} = \frac{\sum y_i}{5(200)} = \frac{5,139,000}{1000} = \$\ 5,139$

(b) $\hat{P} = \frac{\sum P_i}{n} = \frac{\sum_i \sum_j y_{ij}}{nM} = \frac{\sum_i y_i}{n M} = \frac{190}{1000} = 0.19$

(c) $\hat{R} = \frac{\sum\sum y_{ij}}{\sum\sum x_{ij}} = \frac{5,139,000}{2300} = 2234.35$

(d) $var(\bar{\bar{y}}) = \frac{1}{M^2}\left[(1-\frac{n}{N})\frac{1}{n}\ \frac{\sum (y_i-\bar{y})^2}{n-1}\right]$

$= \frac{1}{(200)^2}\left[\ (1-\frac{5}{50})\frac{1}{5}\ \frac{1}{4}\ \{(1,116,000 - 1026800)^2 + \cdots\right.$

$\left. +\ \ (1,179,800 - 1026800)^2\}\right]$

$var(\hat{P}) = \frac{1}{n}\ (1-\frac{n}{N})\ \sum_{i=1}^{n} \frac{(P_i-P)^2}{n-1}$

$= \frac{1}{5}\ (1-\frac{5}{50})\frac{1}{4}\left[(\frac{40}{200} - 0.19)^2 + \cdots + (\frac{32}{200} - 0.19)^2\right]$

$var(\hat{R}) = \frac{1}{\hat{X}^2}\left[v(\hat{Y}) + R^2\, v(\hat{X}) - 2\, R\, cov(\hat{X},\ \hat{Y})\right]$

$\hat{X}^2 = (\frac{N}{n}\sum x_i)^2 = \left[\frac{50}{5}\ (2300)\right]^2 = 5.29 \times 10^8$

$\hat{R} = 4532.17$

$v(\hat{Y}) = N^2(1-\frac{n}{N})\frac{1}{n}\ \ \frac{1}{n-1}\ \ \sum_{i=1}^{n} (y_i - \bar{y}\ )^2$

$= (50)^2(1-\frac{5}{50})\frac{1}{5}\cdot\frac{1}{4}\{(1,116,000 - 1026800)^2 +\cdots$

$+\ \ (1,179,800 - 1026800)^2\}$

$v(\hat{X}) = N^2(1-\frac{n}{N})\frac{1}{n}\ \cdot\ \frac{1}{n-1}\ \ \sum_{i=1}^{n} (x_i - \bar{x})^2$

$$= (50)^2(1-5/50)\frac{1}{5}\frac{1}{4}\{(500-460)^2+\cdots+(520-460)^2\}$$

$$\mathrm{cov}(\hat{X},\hat{Y}) = N^2(1-\frac{n}{N})\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$$

$$= (50)^2(1-5/50)\frac{1}{5}\frac{1}{4}\{(1,116,000-1026800)(500-$$

$$400)+\cdots+(1,179,800-1026800)(520-460)\}$$

3. 一段 集落 標本 抽出 : 다른 크기의 集落들

① 序 論

대부분의 응용에 있어서, 集落 單位 (州, 市 등) 은 다른 數의 元素들을 포함한다 (家口, 사람 數), 다른 크기의 集落 單位에 대한 標本 抽出과 推定에 대하여 연구한다.

② 集落들의 單純 任意 標本 : 不偏推定値. 抽出과 推定 不偏推定値

$N \xrightarrow{SRS} n$

$M_i \longrightarrow M_i$ (全數調査)

$y_i = \sum_{i=1}^{M_i} y_{ij} = M_i\bar{y}_i$ = i 번째 集落 單位에 대한 總計

Y의 不偏 推定値는, 集落單位

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

$$v(\hat{Y}) = N^2\frac{(1-f)}{n}\left(\sum_{i=1}^{n}(y_i-\bar{Y})^2/(N-1)\right.$$

$$= N^2\frac{(1-f)}{n}\left(\sum_{i=1}^{n}(M_i\bar{y}_i-\bar{Y})^2/(N-1)\right)$$

단, $\bar{Y} = Y/N$

$$v(\hat{Y}) = N^2 \frac{(1-f)}{n} \left(\sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1) \right)$$

$$= N^2 \frac{(1-f)}{n} \left(\sum_{i=1}^{n} (M_i \bar{y}_i - \bar{y})^2/(n-1) \right)$$

※ 주의 : 推定値, $\hat{Y}$은 不偏 推定値이지만 精度가 낮다. $\bar{y}_i$가 서로 差異가 작지만, M_i는 單位와 單位사이에 클 때, $v(\hat{Y})$도 크다.

③ 集落의 單純 標本 抽出 : 크기 比推定値 (偏琦)

$M_0 = \sum_{i=1}^{N} M_i$ 이라고 하자.

M_i 가 모두 알고 있다면, M_0 도 알고, Y에 대한 比 推定値는

$$\hat{Y}_{ratio} = M_0 \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{N} M_i} = M_0 \bar{\bar{y}}$$

$$v(\hat{Y}_{ratio}) = M_0^2 v(\sum_{i=1}^{n} y_i / \sum^{n} M_i) = M_0^2 v(\sum_{i=1}^{n} y_i / \sum_{i=1}^{n} M_i) = M_0^2 u(\bar{\bar{y}})$$

$$\fallingdotseq \frac{(1-f)}{n} N^2 \sum_{i=1}^{N} (y_i - \bar{\bar{y}} M_i)^2 / (N-1)$$

$$= \frac{(1-f)}{n} N^2 \sum_{i=1}^{N} M_i^2 (\bar{y}_i - \bar{\bar{y}})^2 / (N-1)$$

$$v(\hat{Y}_{ratio}) \fallingdotseq N^2 \frac{(1-f)}{n} \sum^{n} (y_i - M_i \bar{\bar{y}})^2 / (n-1)$$

$$= N^2 \frac{(1-f)}{n} \frac{\sum_{i=1}^{n} y_i^2 - 2\bar{\bar{y}} \sum_{i=1}^{n} M_i y_i + \bar{\bar{y}}^2 \sum_{i=1}^{n} M_i^2}{(n-1)}$$

※ 주의 : $v(\hat{Y}_{ratio})$는 $\bar{y}_i$의 變異에 의존한다. y_i와 M_i는 相關이 높으므로, $v(\hat{Y}_{ratio})$는 $v(\hat{Y})$보다 보통 작다.

※ 주의 : $\hat{Y}_{ratio}$는 M_0에 대한 知識이 要求되지만, $\hat{Y}$은 그렇지 않다. 母集團의 平均을 推定할 경우는 그 반대이다.

즉, $\hat{\bar{\bar{Y}}} = \hat{Y}/M_0 = \frac{N}{nM_0} \sum^{n} y_i$

$$\hat{\bar{\bar{Y}}}_{ratio} = \frac{\hat{Y}_{ratio}}{M} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} M_i$$

例 ; $\hat{Y}$과 $\hat{Y}_{ratio}$의 比較

다음과 같은 크기가 다른 集落을 생각해 보자. 集落은 "家口의 크기"에 대하여 作成되었다.

集落		
1	2	3
3	1	3
6	4	6
4	8	3
7	4	6
	6	7
	5	4
		6
		3

$Y = 86, \quad \bar{Y} \fallingdotseq 28.67, \quad M_0 = 18$

$M_1 = 4$	$y_1 = 20$	$\bar{y}_1 = 5.00$	$M_1\bar{y}_1 = y_1 = 20$
$M_2 = 6$	$y_2 = 28$	$\bar{y}_2 = 4.67$	$M_2\bar{y}_2 = y_2 = 28$
$M_3 = 8$	$y_3 = 38$	$\bar{y}_3 = 4.75$	$M_3\bar{y}_3 = y_3 = 38$

※ 주의 ; $\bar{y}_i$는 조금 變하고, $M_i\bar{y}_i$는 더 많이 變한다.

다음과 같은 標本計劃을 생각해 보자.

$$N = 3 \xrightarrow{\text{SRS-WOR}} n = 2$$

$$M_i \longrightarrow M_i$$

Y과 $\hat{Y}_{ratio}$를 使用하여 Y의 모든 可能한 推定値를 생각해 보자.

$Y = \frac{N}{n}\sum^{n} y_i$에 대하여 ; 家門의 고기. 作木

	標 本	$\hat{Y}$	$\hat{Y}_{ratio}$
集落들	1,2	72	86.4
	1,3	87	87.0
	2,3	99	84.84

$$E(\hat{Y}) = \frac{72 + 87 + 99}{3} = 86 = Y$$

$\therefore$ $\hat{Y}$은 不偏推定値이다.

$$\hat{Y}_{ratio} = M_0 \left(\sum_{i=1}^{n} y_i / \sum^{n} M_i \right);$$

$$E(\hat{Y}_{ratio}) = 86.09$$

$\hat{Y}_{ratio}$ 는 약간 偏琦되었다.

어느 推定量이 더 큰 分散을 갖는가?

4. 確率比例抽出法

크기에 比例하는 確率로 標本을 抽出하는 方法을 PPS(Sampling with probability proportional to size)로 표시한다.

다음과 같은 6家口의 母集團과 各家口의 크기를 생각해 보자. 家口가 確率比例抽出法으로 뽑혀졌다면, 임의의 家口의 선택의 確率은 家口의 크기를 全體 母集團의 크기로 나눈 것이다. 즉,

선택의 확률 $= y_i / \sum y_i$

表. 確率比例抽出

家 口 #	家口의 크기	선택의 確率
i	y_i	$y_i / \sum y_i$
1	8	8/30
2	6	6/30
3	3	3/30
4	5	5/30
5	4	4/30
6	4	4/30
	30	1

注意 : $\sum_{i=1}^{N} \frac{y_i}{\sum y_i} = 1$

$N \xrightarrow{SRS} n = 1$ 에 대하여 抽出確率은 $1/N$이고, $Y = N\bar{Y} = N\sum_{i=1}^{N} \frac{y_i}{n}$

$= N y_i = y_i / (1/N)$

같은 方法으로

$$N \xrightarrow{PPS} n = 1$$

에 대하여 : 한 單位를 뽑을 確率은 $y_i / \sum_{i=1}^{N} y_i$. $Y_{pps} = y_i / \sum y_i$)

$$= \sum y_i = Y$$

例. 3家口가 PPS로 뽑혔다고 가정하면,

$$\hat{Y}_{pps} = 3/(3/30) = 30 = Y \text{ (母集團 총계)}$$

注意 : 모든 y_i는 旣知이므로, $Y_{pps} = Y$

補助變數에 比例하는 確率로 單位를 뽑는다고 가정하자. 補助變數의 크기 M_i는 單位값 y_i와 다음과 같은 관계가 있다.

$M_i = \beta y_i$ 단, $\beta > 0$ 그리고 β는 상수.

그러면 抽出確率은

$$PPS = M_i / \sum_{i=1}^{N} M_i = \frac{\beta y_i}{\beta \sum y_i} = \frac{y_i}{\sum y_i}$$

실제로 우리가 연구하는 變數의 참 크기를 알지 못한다. 그러나 보조변수를 통하여 硏究變數의 크기를 어림한다. 즉 SRS보다 더 큰 效率을 갖는 推定量을 얻기 위하여 PPZ으로 單位들을 뽑을 수 있다.

例. 硏究變數와 補助變數

硏究變數 (study variable)	補助變數 (auxiliary variable)
현재의 인구	前 센서스 人口
현재의 출생수	前 센서스 人口
현재의 총수입	前 센서스 人口
경작면적	총 지리적 면적
공장 생산량	노동자 총수

다음과 같은 推定 過程을 생각해 보자

$$N \xrightarrow{PPS-WR} n \geq 2 \text{를 생각하자}$$

i번째 單位의 1次抽出 確率은

$$M_i / \sum_{i=1}^{N} M_i = M_i / M_0$$

이다. 抽出된 各 標本單位들로 부터 총계의 不偏推定値를 얻는다. 즉,

$$y_i / (M_i / M_0) = M_0 y_i / M_i = M_0 \bar{y}_i \quad \text{단, } y_i = \sum_{i=1}^{Mi} y_{ij}$$

이들의 n개의 推定値들로 부터 不偏推定量은 :

$$\hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^{N} M_0 \bar{y}_i = \frac{M_0}{n} \sum_{i=1}^{N} \bar{y}_i$$

그러면, $V(\hat{Y}_{pps}) = \frac{M_0}{n} \sum_{i=1}^{N} M_i (\bar{y}_i - \bar{\bar{y}})^2$

그리고 $V(\hat{Y}_{pps}) = M_0^2 \sum_{i=1}^{N} (\bar{y}_i - \bar{\bar{y}})^2 / n(n-1)$

M_i가 다만 근사적으로 알려져 있다고 하자. 그러면, $Z_i = \frac{M_i'}{M_0}$라고 하자. 단. $M_0' = \sum_{i=1}^{N} M_i'$은 i번째 單位가 復元으로 抽出된 確率이다. 그러면,

$$\hat{Y}_{ppz} = \frac{1}{n} \sum y_i / Z_i \quad \text{단, } y_i = \sum_{i=1}^{Mi} y_{ij}$$

$$V(\hat{Y}_{ppz}) = \frac{1}{n} \sum z_i (y_i / z_i - Y)^2$$

$$V(\hat{Y}_{ppz}) = \sum (y_i / Z_i - Y_{ppz})^2 / n(n-1)$$

注意 : (1) $Z_i = M_i / M_0$ 이면, 前 結果와 같다.

(2) $\hat{Y}_{ppz}$은 총계의 不偏推定量이다.

例.　　集　落

1	2	3
3	1	3
6	4	6
4	8	3
7	4	6
	6	7
	5	4
		6
		3

$Y = 86$　　$M_0 = 18$

$\bar{y}_1 = 5.00$

$\bar{y}_2 = 4.60$

$\bar{y}_3 = 4.75$

$Z_i = \frac{M_i}{M_0}$ 라 하자.

모든 가능한 標本에 대하여,

$$\hat{Y}_{pps} = \frac{M_0}{n} \sum_{i=1}^{n} \bar{y}_i$$

가능한 標本	$\hat{Y}_{pps}$	前例에서부터 $Y=N/n\sum Y_i$	$\hat{Y}_{RATIO} = M_0 \sum y_i / \sum M_i$
1,2/2,1	87.00	72	86.4
1,3/3,1	87.75	87	87.0
2,3/3,2	84.75	99	84.86
1,1	90.00	不偏	偏倚
2,2	84.00		
3,3	85.50		

$$E(\hat{Y}_{pps}) = 2(4/18)(6/18)(87.00) + 2(4/18)(8/18)(87.75)$$
$$+ 2(6/18)(8/18)(84.75) + (4/18)(90.00)$$
$$+ (6/18)(6/18)(84.00) + (8/18)(8/18)(85.50) = 86$$

$\hat{Y}_{pps}$ 는 不偏推定値이다.

抽出過程

方法 1 : M_i 를 累積하여라

N = 7 개의 單位들이 아래와 같이 주어졌다고 하자.

單位 PPS를 抽出하는 것은 1과 30사이에 亂數를 뽑는 것이다.

單位 i	크 기 M_i'	M_i	할당된 영역
1	3	3	1 − 3
2	1	4	4
3	11	15	5 − 15
4	6	21	16 − 21
5	4	25	22 − 25
6	2	27	26 − 27
7	3	30	28 − 30

例. (1) 亂數表에서 01과 30사이에 數를 뽑아라 06이 뽑혔다면, 單位 3이 뽑힌 것이다.

(2) 다음, 數 12가 뽑혔다면, 單位 3이 다시 뽑힌 것이다.

(3) 다음 亂數 (RN) 06이 뽑힌다면, 이미 뽑힌 수이므로 제외시킨다.

(4) 다음 亂數 30이 뽑히면 單位 7이 標本으로 들어간다.

PPZ 系統抽出法을 사용하면, n개 單位들의 標本을 抽出하기가 쉽다. 많은 경우에 系統 標本抽出이 잘 수행되도록 單位들을 순서있게 정돈할 수 있는 事前 情報가 있다. 그러나 $V(\hat{Y}_{pps-sys})$의 不

偏推定値는 없다.

例 : $N \xrightarrow{PPZ-SYS} n=3$

$n=3$ 單位를 얻기 위하여, $I=\frac{M_0'}{n}=\frac{30}{3}=10$ 이라고 하자. 01-10 사이에 數 07를 抽出하여라. 그러면, 07에 대응하는 單位들 $07+10=17$, $07+2(10)=27$은 標本에 포함된다. 즉, 單位들 3, 4 그리고 6은 標本에 포함된다.

다음과 같은 영역들을 생각할 수 있음을 注意하라.

1 - 10

11 - 20

21 - 30

y_i = 1次 抽出單位 i에 대한 총계라 하자. 그러면 層 i에 대하여, 特性 y에 대한 총계의 推定値는 :

$$\hat{Y}_i = W_i y_i = \frac{1}{nZ_i}\, y_i = \frac{1}{n}\,\frac{y_i}{z_i}$$

3개의 層에 대하여 총계에 대한 推情値는 :

$$\hat{Y}_{ppz-sys} = \sum \hat{Y}_i = \sum \frac{1}{n}\,\frac{y_i}{Z_i} = \frac{1}{n}\sum \frac{y_i}{Z_i}$$

$V(\hat{Y}_{ppz-sys})$의 不偏推定値는 존재하지 않는다. 그러나 여러가지의 分散推定 過程이 여러 조건아래서 수행된다.

例. 위에서 抽出한 標本單位들에 대한 PSU 총계를 다음과 같다고 가정하자.

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{Z_i} = \sum_{i=1}^{n}\frac{y_i}{nz_i}$$

$$= 55\cdot\frac{1}{11/10} + 28\cdot\frac{1}{6/10} + 12\cdot\frac{1}{2/10}$$

$$= 156.67$$

크기가 다른 單位들에 대한 副次標本 다음과 같은 標本抽出을 생각하자.

$$N \longrightarrow n \qquad M_i \longrightarrow m_i$$

等確率로 抽出된 單位들 (不偏推定量)

$$N \xrightarrow{SRS-WOR} n$$

$$M_i \xrightarrow{SRS-WOR} m_i$$

母集團 총계, Y를 推定하기 위하여

$$\hat{Y}_v = \frac{N}{n}\sum Y_i = \frac{N}{n}\sum M_i \bar{y}_i = \frac{N}{n}\sum M_i \sum y_{ij}/m_i$$

$\hat{Y}_i$ 은 PSU i 에 대한 推定된 총계이다.

注意 : $f_{2i} = \frac{m_i}{M_i}$가 상수이면, 즉 $\frac{m_i}{M_i} = \frac{m}{M}$이면.

$$\hat{Y}_u = \frac{N}{n}\frac{M}{m}\sum\sum y_{ij}$$

즉, 設計 結果는 自體加重 推定量이다.

$$V(\hat{Y}_u) = \frac{N^2}{n}(1-f_1)\sum\frac{(y_i-\bar{Y})^2}{N-1} + \frac{N}{n}\sum f_{2i}(1-f_{2i})\frac{S_{2i}^2}{m_i}$$

단, $S_{2i}^2 = \frac{1}{M_i-1}\sum(y_{ij}-\bar{Y}_i)^2$

$$V(\hat{Y}_u) = \frac{N^2}{n}(1-f_1)\sum\frac{(\hat{Y}_i-\hat{\bar{Y}}_u)^2}{n-1} + \frac{N}{n}\sum M_i^2(1-f_{2i})\frac{s_{2i}^2}{m_i}$$

단, $s_{2i}^2 = \frac{1}{m_{j-1}}\sum(y_{ij}-\bar{y}_i)^2$ 그리고 $\bar{Y}_v = Y_v(\frac{1}{N})$

Ⅱ. 유한 모집단과 무한 모집단으로 부터의 표본추출에서 Bayes와 미니맥스(minimax)과정

요 약 : 표본조사에서 표본추출 방법으로 손실(loss)과 위험(risk) 함수를 생각할 수 있다. 손실함수는 추정치의 오차의 자승합과 표본추출 비용(cost)으로 이루어지는 함수이다.

1. 서 론

통계학자는 표본조사에서 주어진 조사비용에 대하여 最大精度(precision)의 추정치를 얻고자 한다. 또는 최소의 비용으로 주어진 精度의 추정치를 얻고자 한다.

그러나 추정치에서의 오차(error)와 표본추출 비용으로 부터의 손실을 함께 생각하여, 즉 전체 기대 손실을 최소화시키는 추정 과정을 생각할 수 있다. 따라서 손실함수를 두 성분의 합으로 생각할 수 있는데, 한 성분은 추정치의 오차의 자승에 비례하고, 다른 성분은 표본을 얻는 비용에 비례한다. 이 문제는 의사결정론(decision theory)의 손실함수를 써서 해결되는 바, 표본조사(samplimg survey)에 의사결정론 개념이 도입된 것이다.

2. 베이즈와 미니맥스 추정치

X를 표본공간이라고 하자. 표본공간 X위에서 확률분포의 공간 $P\Omega = \{P_\omega : \omega \in \Omega\}$, Ω는 모수공간, 그리고 Ω위에서 정의된 $g(\omega)$를 생각

하자. 비확률화 결정함수(nonrandomized decision function)는 X위에서 정의되는 함수 δ로 각 x에 대하여 $a \in A$를 지정한다. 즉 x가 관찰되었을때 a로 $g(\omega)$를 추정한다. 행동공간(action space) A는 원소 a들의 모임으로 A는 실직선(real line)이다. 손실함수 R은

$$R(\omega, \delta) = E_\omega L(\omega, \delta)$$

로 정의된다. 첨자 ω는 고정된(fixed)수로 간주한다.

ω는 확률분포 λ을 갖는 확률변수이다. 事前分布(prior distribution) λ에 대한 베이즈 추정치는 평균 위험 $\int R(\omega, \delta) d_\lambda(\omega)$를 최소화시키는 추정치 δ로 정의된다. 통계학자가 λ를 안다면, 이 추정치가 최선의 행동이다. λ에 대한 지식이 없다면 미니맥스 전략(strategy)를 쓴다. 미니맥스 추정치는 $\sup_{\omega \in \Omega} R(\omega, \delta)$를 최소화시키는 추정치 δ로 정의한다. 최소유리분포(least favorable distribution)는 $\inf_\delta \int R(\omega, \delta) d_\lambda(\omega)$를 최대화시키는 분포 λ로 정의된다.

정리 2.1. 만일 베이즈 추정치 δ_λ가 고정위험 $R(\omega, \delta_\lambda) = \gamma$을 갖는다면, δ_λ는 미니맥스이고, λ는 최소유리분포이다.

정리 2.2. $\{\lambda_n\}$이 事前確率分布의 수열이고, $\{\lambda_n\}$이 그에 대응하는 베이즈 위험이라고 하자. $n \to \infty$일 때 $\gamma_n \to \gamma$이고, 모든 ω에 대하여, $R(\omega, \delta) \leq \gamma$인 추정치 δ가 존재한다면, δ는 미니맥스 추정치이다.

정리 2.3. δ는 미니맥스 과정, γ는 미니맥스 위험 그리고 관찰치

X는 확률분포 $\omega \in \Omega^*$를 따른다고 하자. 만일 Ω Ω^*이고, δ에 대응하는 위험이 γ보다 크지 않다면 δ는 Ω에서 미니맥스 과정이고, γ은 미니맥스 위험이다.

증명. 가설에 의하여

$$\gamma = \sup_{\omega \in \Omega^*} R(\omega, \delta) \leqq \sup_{\omega \in \Omega} R(\omega, \delta) \leqq \gamma$$

3. 손실함수

손실을 오차 자승으로 간주하여 유한 모집단의 평균을 추정하고자 한다. 분산의 크기가 제한되어 있지 않으면, 위험이 임의로 커질 것이다. 따라서 유한모집단의 분산의 기대치가 유계되었다고 가정한다.

명백히, 결정문제는 다음과 같이 요약된다.

(a) 표본공간은 (X, Ω, P)이다. X는 N-차원 유크리드공간 R^n, Ω는 모든 분포 ω의 집합으로 $x_1 + x_2 + \cdots\cdots + x_n = N\mu_\omega$이다. 그리고 $E_\omega \Sigma (x_i - \mu_\omega)^2 \leqq (N-1)\sigma^2$

(b) 행동공간 A는 실직선 R^1이다.

(c) 손실함수 L은 $(\Omega \times A)$위에서 정의되며

$$L(\omega, a) = (a - \mu_\omega)^2$$

으로 주어진다.

(d) 만일 $\delta = (i_1, \cdots, i_n : f)$이면, $\delta(x) = f(x_{i1}, \cdots, x_{in})$이다.

유한모집단 $(x_1, \cdots, x_n)$의 평균 μ_ω의 미니맥스 추정치를 구하는 문제는 다음과 같다. μ_ω는 평균이 0이고 분산이 θ^2인 정규분

포를 한다. 주어진 μ_ω에 대하여 x_i의 평균은 μ_ω이고, 분산은 $\sigma^2(N-1)/N$이고, 두 성분 x_i와 x_j에 대한 공분산은 $-\sigma^2/N$이다.

베이즈 추정치를 事前分布의 수열 $\{\lambda_\theta\}$에 대하여 구할수 있으며, 그에 대응하는 베이즈 위험 $\{\gamma_\theta\}$가 존재하면, $\theta\to\infty$일 때에 $\{\gamma_\theta\}$의 극한은 γ이다. 그러면 만일 어떤 추정치 δ에 대하여 $R(\omega,\delta)\leqq\gamma$이면 δ는 미니맥스 추정치이다. 여기서 표본추출을 비복원 단순확률추출법으로 한다.

4. 유한모집단의 평균추정에 대한 베이즈와 미니맥스 과정(비복원 추출)

사전분포 ξ와 결정함수 δ에 대응하는 δ에 대응하는 평균 위험은

$$R(\xi,\delta)=\underset{\Omega}{E}\,\underset{x}{E}\left[(\delta(x)-g(\omega))^2\,|\,\omega\right]$$

이다. 단, x는 $(x_1,x_2,\cdots,x_n)$을 나타내고, δ는 $g(\omega)$의 추정치이다. 위식의 위험에서 기대치의 순서를 바꾸므로서

$$R(\xi,\delta)=\underset{x}{E}\,\underset{\Omega}{E}\left[(\delta(x)-g(\omega))^2\,|\,x\right]$$

이 식은 각 x에 대하여 $E[(\delta(x)-g(\omega))^2\,|\,x]$를 최소화시키는 $\delta(x)$를 선택하므로서 최소화 된다. 즉

$$\delta_\xi(x)=E[g(\omega)\,|\,x]$$

이것이 $E[(\delta(x)-g(\omega)^2\,|\,x]$의 최소값으로 $\sigma_g^2(\omega)|x$로 표시하는데, 주어진 x에 대한 $g(\omega)$의 조건부 분포의 분산이다. 그러면, 베이즈 위험 γ_ξ는

$$\gamma_\xi=\underset{x}{E}\,\sigma_g^2(\omega)|x$$

로 주어진다.

이 결과는 일반적으로 $g(\omega)$의 추정치에 대한 손실함수가 $L(\omega, a) = [a - g(\omega)]^2$ 일때에 성립한다.

표본 $(x_1, \cdots x_n)$의 분포는 각 x_i에 대하여 평균 $\mu\omega$ 그리고 분산 $\sigma^2(N-1)/N$인 n-변량 정규분포를 하고, 각 쌍(i,j), $i \neq j$에 대한 공분산은 $-\sigma^2/N$이다. 표본평균 $\bar{x}$는 $\mu\omega$에 대한 충족통계량이므로, 베이즈 추정치는 $\delta_\theta(x) = E(\mu\omega|x) = E(\mu\omega|\bar{x})$이고, 베이즈 위험은 $\gamma_\theta = E\,\sigma^2_{\mu\omega|x} = E\sigma^2_{\mu\omega|x}$이다. 이제 $\mu\omega$는 $N(o, \theta^2)$이고, 주어진 μ_ω에 대해서 $\bar{x}$는 $N(\mu_\omega, v)$이다. 단, $v = (n^{-1} - N^{-1})\sigma^2$, 주어진 $\bar{x}$에 대하여 μ_ω의 조건부 분포는 평균 $\theta^2\bar{x}(\theta^2 + v)^{-1}$와 분산 $\theta^2 v(\theta^2 + v)^{-1}$를 갖는다. 이 분산은 x에 독립이므로 이들이 각각 베이즈 추정치 $\delta_\theta(x)$와 베이즈 위험 γ_θ이다.

$$\gamma = \lim_{\theta \to \infty} \gamma_\theta = v = \frac{N-n}{Nn}\sigma^2$$

어떤 추정치 δ에 대해, 그 위험이 γ보다 크지 않으면, δ는 미니맥스 추정치이다. $\delta(x) = \bar{x}$라 하면, δ에 대응하는 위험은

$$R(\omega, \delta) = E_\omega(\bar{x} - \mu_\omega)^2 \leqq \frac{N-n}{Nn}\sigma^2 = \gamma$$

그러므로 $\bar{x}$는 미니맥스 추정치이고 그 위험은 $(N-n)\delta^2/nN$을 초과하지 않는다.

5. 층화표본추출에서의 베이즈와 미니맥스 과정

모집단을 층화하여서 각 층에서 독립적으로 표본추출을 한다. i번

관위원소에 대한 표본추출 비용, 그리고 δ는 표본 $\{X_{ij}: i=1, k: j=1, \cdots, n_i\}$의 함수이다. 단, X_{ij}는 i번째 층으로부터 j번째 관찰치이다.

더 실질적으로 손실함수를

$$L(U,\delta)=\alpha(\delta-U)^2+\Sigma c_i n_i$$

로 나타낼 수 있으며, α는 주어진 상황에서의 비용이다. 그러나 c_i 대신 c_i/α를 대입하므로

$$L(U,\delta)=(\delta-U)^2+\Sigma c_i n_i$$

식을 얻을 수 있다. 주어진 ω에 대하여, 각 층은 분산이 σ^2인 정규분포를 한다고 하자. 또 μ_i가 평균이 0이고, 분산이 θ^2인 정규분포를 하고, 각각 독립인 분포를 한다고 가정하자. 그리고 μ_i의 분포로부터 U의 사전분포의 수열 $\{\lambda_\theta\}$에 대응하는 베이즈 해를 구한다. δ_θ를 그에 대응하는 U의 베이즈 추정치라 하자. δ_θ에 대응하는 베이즈 위험 r_θ의 극한값이 r이라 할 때, 어떤 추정치의 위험이 r보다 작거나 같으면, 그것이 미니맥스 추정치가 된다.

우리는 위험

$$R(U,\delta^*)=E(\delta^*-U)^2+\Sigma c_i n_i$$

가 최소화되도록 n_i를 선택하고자 한다. δ^*는 주어진 n_i에 대한 미니맥스 추정치이다.

μ_i들이 독립적으로 그리고 평균이 0이고 분산이 θ^2인 정규분포를 한다. 주어진 μ_i에 대하여 $\bar{X}_i$들은 평균이 μ_i 그리고 분산이 σ_i^2/n_i인 정규분포를 한다. 주어진 $\bar{X}_1, \cdots\cdots, \bar{X}_k$에 대하여 μ_i의 조건부 분포는 평균이

$$y_i = \frac{n_i \theta^2 \bar{X}_i}{\sigma_i^2 + n_i \theta}$$

그리고 분산이

$$v_i = \frac{\theta^2 \sigma_i^2}{\sigma_i^2 + n_i \theta^2}$$

이다. 주어진 $\bar{X}_1, \cdots, \bar{X}_k$에 대하여 $\mu_1, \cdots, \mu_k$는 서로 독립이다. 그래서, 주어진 $\bar{X}_1, \cdots, \bar{X}_k$에 대하여, U의 분포는 평균이 $\Sigma a_i y_i$ 그리고 분산이 $\Sigma a_i^2 v_i$인 정규분포를 한다. 따라서 베이즈 추정치는 $\delta_\theta(x) = \delta_\theta(\bar{X}_1, \cdots, \bar{X}_k) = \Sigma a_i y_i$, 그리고 U의 조건부 분포의 분산이 $\bar{X}_1, \cdots, \bar{X}_k$에 독립이므로 베이즈 위험은 $r_\theta = \Sigma a_i^2 v_i$ 이다.

주어진 n_i에 대한 미니맥스 추정치. $\theta \to \infty$일 때, $r_\theta \to r$ 임을 우리는 알았다. 단, $r = \Sigma a_i^2 \sigma_i^2 / n_i$, 그래서 어떤 추정치 δ^*가 r보다 같거나 작은 위험을 갖으면, δ^*는 미니맥스 추정치이다. 베이즈 추정치의 극한은

$$\lim_{\theta \to \infty} \delta_\theta(x) = \Sigma a_i \bar{X}_i = \delta^*(x)$$

이다. $\bar{X}_i$들은 평균이 μ_i 그리고 분산이 σ_i^2/n_i인 정규분포를 하며, 독립이므로, $\sum a_i \bar{x}_i$는 평균이 $\sum a_i \mu_i = U$이고, 분산어 $\sum a_i^2 \sigma_i^2 / n_i$ 인 정규분포들이다. 그러므로 추정치 $\delta^*(x) = \sum a_i \bar{X}_i$에 대응하는 위험은 r과 같고, 따라서 $\sum a_i \bar{X}_i$는 주어진 n_i에 대하여 U의 미니맥스 추정치이다.

정규분포라는 가정의 제거 : X_{ij}의 분포가 정규라는 가정을 제거하자, X_{ij}의 결합분포가 무엇이든지 간에, i번째 층의 분포는 미지의 평균 μ_i를 갖고, 표본 평균 $\bar{X}_i$는 σ_i^2/n_i보다 크지 않은 분산을

갖는다. 이것은 i번째 층에서 X_{ij}가 평균이 μ_i 그리고 분산이 σ^2 이며 독립이라는 것보다 더 일반적이다. 먼저 미니맥스 추정치 $S^*(x)=\Sigma a_i\bar{X}_i$에 대응하는 위험 R를 계산해 보자. 주어진 n_i에 대하여

$$R=E(\Sigma a_i\bar{X}_i-U)^2$$

$$=E[\Sigma a_i(\bar{X}-\mu_i)]^2\leqq a_i^2\sigma_i^2/n_i=r$$

n_i를 선택하는 미니맥스 전략 : 서로 다른 층들에 대해 모집단의 분산이 알려져 있다면, 위험을 n_i의 함수로 보아 최소화시키므로서 최적의 n_i를 선택할 수 있다. 미니맥스 전략을 최대로 허용된 분산 σ_i^2에 대하여 최적의 n_i를 선택하는 것이다.

정리 6.1. 전략들의 공간이 $D=U_cD_c$라고 하자. δ_c가 D_c에서 미니맥스이고, 각 C에 대하여 $R(\omega,\delta_c)$가 상수값을 갖는다고 하자. 즉 $R(\omega,\delta_c)=r_c$, 그러면 δ_c는 r_c를 최소화 한다.

증명. $\delta\in D$인 임의의 전략 δ를 생각해 보자. 그러면 $\delta\in D_c$,

$$\max_{\omega}R(\omega,\delta)\geqq\max_{\omega}R(\omega,\delta_c)=r_c$$

δ_c^*가 r_c를 최소화시키는 δ_c라 하고, δ_c^*에 대응하는 위험을 r_c^*로 표시하자. 그러면

$$r_c\geqq r_c^*=\max_{\omega}R(\omega,\delta_c^*)$$

따라서 모든 $\delta\in D$에 대하여,

$$\max_{\omega}R(\omega,\delta_c^*)\leqq\max_{\omega}R(\omega,\delta)$$

이므로, δ_c^*는 D에서 미니맥스이다.

우리는 각 층의 분산이 σ_i^2 일 때에, 최적의 n_i 를 선택한다. 주어진 n_i 에 대하여, δ^* 에 대응하는 위험은

$$R(\omega, \delta^*) = \Sigma\left[\frac{a_i^2\sigma_i^2}{n_i} + c_i n_i\right]$$

이다. 이제 우리는 이 위험이 최소가 되는 n_i 를 선택하고자 한다.

$f(n_i) = a_i^2\sigma_i^2/n_i + c_i n_i$ 로 놓자. 그러면

$$f(n_i + 1) - f(n_i) = c_i - a_i^2\sigma_i^2/n_i(n_i + 1)$$

위 식의 차이가 양이 되는 최소의 양정수를 구하면 된다. 그러면

$n_i = \sqrt{\frac{a_i^2\sigma_i^2}{c_i} + \frac{1}{4}}$ 에 가장 가까운 정수.

B. 유한모집단

x_{ij} ($i = 1, \cdots, k$: $j = 1, 2, \cdots, N_i$)가 i번째 층에 j번째 원소라 하자. N_i 는 기지이고, 각층들의 평균 μ_i 는 미지라 하자. 또 각 층들의 모집단의 분산은 유계 (bound)라 하자. 즉

$$\sigma_i^2 \geqq 1/(N_i - 1)\, E\Sigma\,(x_{ij} - v_i)^2$$

우리는 선형함수

$$T = \Sigma a_i v_i$$

를 추정하고자 한다. 손실함수 L은

$$L(T, \delta) = (\delta - T)^2 + \Sigma c_i n_i$$

로 주어지며, δ 는 T의 추정치이다.

주어진 $\bar{X}_i$ 에 대하여 v_i 의 조건부 분포는 평균이 $y_i = \theta^2\bar{X}_i(\theta^2 + v_i)^{-1}$ 그리고 분산이 $\theta^2 v_i(\theta^2 + v_i)^{-1}$ 인 정규분포이다. 단, $v_i = (n_i^{-1} - N_i^{-1})\sigma_i^2$, 그래서 주어진 $\bar{X}_1, \cdots \bar{X}_k$ 에 대하여, $T = \Sigma a_i v_i$ 의 조

건부 분포는 평균이 $\Sigma a_i^2\theta^2 v_i(\theta^2+v_i)^{-1}$ 이다. 따라서 T에 대한 베이즈 추정치는

$$\delta_\theta(x)=\Sigma a_i y_i=\Sigma a_i\theta^2\bar{X}_i(\theta^2+v_i)^{-1}$$

이다.

T의 조건부 분포의 분산은 x와 독립이므로, 베이즈 위험은

$$r_\theta=\Sigma a_i^2\theta^2 v_i(\theta^2+v_i)^{-1}+\Sigma c_i n_i$$

이다.

T에 대한 미니맥스 추정치를 구하기 위하여, 수열 $\{r_\theta\}$ 이 $\theta\to\infty$ 일 때 극한을 갖는가를 조사해 본다.

$$r=\lim_{\theta\to\infty} r_\theta=\Sigma a_i^2 v_i+\Sigma c_i n_i$$

$$=\Sigma a_i^2\frac{N_i-n_i}{N_i n_i}\sigma_i^2+\Sigma c_i n_i$$

어떤 추정치의 δ^*의 위험이 r보다 크지 않으면 그것은 주어진 n_i에 대하여 미니맥스는 추정치이다.

$\delta^*(x)=\Sigma a_i\bar{X}_i$ 의 위험은

$$R(\omega,\delta^*)=E\omega(\delta^*-T)^2+\Sigma c_i n_i$$

$$=E\omega[\Sigma a_i(\bar{X}_i-v_i)]^2+\Sigma c_i n_i$$

각 층들이 독립이므로

$$R(\omega,\delta^*)\le\Sigma a_i^2\frac{N_i-n_i}{N_i n_i}\sigma_i^2+\Sigma c_i n_i=r$$

그러므로 주어진 n_i에 대하여 $\Sigma a_i\bar{x}_i$는 미니맥스 추정치이다.

n_i를 선택하는 미니맥스 전략 : 우리는 이제 미니맥스 위험이 최소가 되도록 하는 n_i를 선택하고자 한다. 이 위험은

$$\gamma = \Sigma \left[a_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) \sigma_i^2 + c_i n_i \right]$$

이다.

특수한 경우로, 모집단의 평균 $\mu = N^{-1} \sum_{i=1}^{k} \sum_{j=1}^{N=1} x_{ij}$ 를 추정하는 문제를 생각해 보자. $a_i = N_i/N$, $T = \Sigma a_i v_i = \mu$ 로 놓으므로서, 평균 μ 를 추정하는 미니맥스 과정을 n_i 를 다음과 같이 구하는 것이다.

$n_i = \sqrt{\frac{N_i^2 \sigma_i^2}{N^2 c_i} + \frac{1}{\lambda}}$ 에 가까운 정수, 그리고 $\leqq N_i$.

Ⅲ. 이단표본추출(two-stage sampling)에서 베이즈와 미니맥스 과정

1. 서 론

표본조사에서 이단표본추출은 모집단의 원소들이 집락(Cluster)으로 나누어져서, 표본추출이 이단으로 수행된다. 처음에는 집락의 표본이 추출되어 집락이 첫 단계 표본추출단위가 된다. 이를 기본표본 추출단위(primary sampling unit)라 부르고 P.S.U 로 간단히 나타낸다. 그러면 표본은 각각의 추출된 P.S.U들로 부터 뽑혀진다. 최후로 뽑혀진 모집단의 단위는 이단표본추출 단위이다. 층화추출과 집락추출법은 이단표본추출의 특수한 경우이다. 즉, 첫번째와 두번째 단계에서 각각 표본추출율이 100%인 경우이다. 각 단계에서 표본추출율이 100%보다 작을 때가 이단표본추출이다.

2. 무한모집단, 두번째 단계에서 다른 크기의 표본추출

기지의 상계 σ_0^2 을 갖는 변수 Y의 평균 μ를 추정한다고 가정한다. 먼저 미리 결성된 크기 m의 표본을 추출한다. 각 집락들의 평균 $\mu_1, \cdots, \mu_m$은 미지이다. $\mu_2, \cdots, \mu_m$ 에 대응하는 확률변수 X_i 들의 조건부 분포는

$$E(X_i \mid \mu_i) = \mu_i, \quad E[(X_i - \mu_i)^2 \mid \mu_i] \leq \sigma_i^2, \quad i = 1, \cdots, m$$

이다.

우리는 $\sigma_0^2, \sigma_1^2, \cdots, \sigma_m^2$ 은 알고 있다고 가정하여 μ를 추정한다.

손실함수 L은 다음과 같이 주어진다.

$$L(\mu,\delta) = \alpha(\delta-\mu)^2 + C_b m + \sum C_i N_i$$

여기서 δ는 관찰된 표본의 함수로 μ를 추정하는데 사용되고, C_b는 첫 단계에서 표본$(\mu_1,\cdots,\mu_m)$을 선택하는데 필요한 표본추출 비용, 그리고 C_i는 각 i에 대하여 X_i의 표본추출단위에 대한 비용이다. 일반성을 잃지 않고 $\alpha=1$로 놓을 수 있다. 우리가 결정해야 할 문제는 μ를 추정하기 위해서 어떤 함수 δ를 선택하여야 할것인지, 또 m과 n_i는 어떤 값을 취해야 할 것인지를 정해야 한다. 손실함수를 $\alpha=1$로 놓아

$$L(\mu,\delta) = (\delta-\mu)^2 + C_b m + \sum C_i n_i$$

라 하자

3. 베이즈와 미니맥스 전략

정리 3.1 전략들의 공간을 $D=U_\lambda D_\lambda$라 가정하자 λ가 임의로 선택된다고 가정하여, λ가 선택되면 D_λ로 부터 하나의 전략을 사용해야 한다. δ_λ가 D_λ에서 미니맥스라고 하자 λ가 선택될 때에는 δ_λ를 사용하는 전략을 δ라 하자. 그러면 전략 δ는 다음의 조건이 만족하면 미니맥스이다.

(i) 위험 $R(W,\delta_\lambda)=r_\lambda$가 각 λ에 대하여 성립한다.

(ii) r_λ는 λ의 유계함수이다.

(iii) 모든 λ에 대하여 베이즈 위험은 r_λ에 수렴한다.

증명. λ가 선택되면 δ_λ를 사용하므로

$$R(W,\delta) = E_{(\lambda)}\, R(W,\delta_\lambda)$$

단, 확률변수 λ에 대하여 기대치가 취해진다.

$$\max_w R(W,\delta) = \max E_{(\lambda)} R(W,\delta_\lambda)$$
$$\leqq E_{(\lambda)} \max_w R(W,S_\lambda)$$
$$= E_{(\lambda)} r_\lambda$$

δ'는 λ가 선택되면, δ'_λ D_λ를 선택하는 전략이다. 그러면

$$\max_w R(W,\delta') = \max E_{(\lambda)} R(W,\delta'_\lambda)$$
$$\geqq \mathrm{Lim}_\theta E_{(\lambda)} R(\theta,\delta'_\lambda)$$
$$= E_{(\lambda)} \mathrm{Lim}_\theta R(\theta,\delta'_\lambda)$$
$$\geqq E_{(\lambda)} \mathrm{Lim}_\theta R(\theta,\delta_\lambda)$$
$$= E_{(\lambda)} r_\lambda$$

따라서

$$\max R(W,\delta) \leqq E_{(\lambda)} r_\lambda \leqq \max_w R(W,\delta')$$

위 식이 모든 δ'에 대하여 성립하므로 δ는 미니맥스이다.

이제 집락들의 고정된 선택에 대응하는 베이즈와 미니맥스 추정치를 구하는 문제를 생각해 보자. 우리는 μ의 사전분포들의 수열$\{\lambda_\theta\}$에 대응하는 베이즈 추정치들의 수열을 구하여야 하는데, λ_θ는 평균이 0이고, 분산이 θ^2인 정규분포를 한다. 즉 $N(0,\theta^2)$이다.

n_i가 고정되었다고 간주하여, 손실함수에서 $C_b m+\sum c_i n_i$를 생략할 수 있으므로, 손실함수는 단순히 평균자승오차의 함수이다. 베이즈 추정치 δ_θ는 주어진 표본에 대한 μ의 조건부 분포의 평균이고, 베이즈 위험 r_θ는 이 조건부 분포의 분산의 기대치와 같다.

표본을 $X=\{X_{ij}\ ;\ i=1,\cdots,\ m\ ;\ j=1,\ \cdots,\ n_i\}$로 나타내자. X_{ij}는 주어진 μ_i에 대한 X_i의 조건부 분포에서의 j번째 관찰치이다.

표본추출물은 각 X_i에 대하여 독립이다. Y는 $N(\mu, \sigma_i^2)$이다. 각각의 $\bar{X}_i = n_i^{-1} \sum X_{ij}$는 μ_i에 대하여 충족통계량이고, μ_i의 분포는 μ에 의존하므로, $\bar{X}^* = (\bar{X}_1, \cdots, \bar{X}_m)$은 μ에 대한 충족통계량이다. 표본 X를 집합 $\bar{X}^*$로 대치하면

$$\delta_\theta = E(\mu | \bar{X}^*)$$

$$\gamma_\theta = E\,\sigma_\mu^2 | \bar{X}^*$$

이다.

주어진 μ_i에 대하여 $\bar{X}_i$는 $N(\mu_i, \sigma_i^2/n_i)$로 분포되고, 주어진 μ에 대하여 μ_i는 $N(\mu, \sigma_b^2)$이다. 그 분포들이 독립이므로 주어진 μ에 대하여 $\bar{X}_i$는 $N(\mu, \sigma_b^2 + \sigma_i^2/n_i)$이다. μ의 사전분포는 $N(0, \theta^2)$이므로 $\bar{X}_i\,(i = 1, \cdots, m)$와 μ의 결합분포는 $(m+1)$변량 정규분포이므로 주어진 $\bar{X}^*$에 대하여 μ의 조건부 분포를 구할 수 있다. 여기서

$$\alpha = \sum W_i \bar{X}_i / (\sum W_i + \theta^{-2})$$

$$\beta = [\sum W_i + \theta^{-2}]^{-1}$$

그리고

$$W_i = n_i / (n_i \sigma_b^2 + \sigma_i^2)$$

따라서 베이즈 추정치는 $\delta_\theta = \alpha$이고, 베이즈 위험은 $\gamma_\theta = \beta$이다.

4. 마니맥스 추정치

$\theta \to \infty$일 때에, $\gamma_\theta \to \gamma$이다. 여기서

$$\gamma = (\sum W_i)^{-1} = [\sum n_i / (n_i \sigma_b^2 + \sigma_i^2)]^{-1}$$

이다. 우리는 그 위험이 γ 보다 크지 않은 추정치 δ^*를 찾아서 한다. 그러한 추정치는

$$\text{Line}\ \delta_{\theta(\lambda)} = \sum W_i \bar{X}_i / \sum W_i = \delta^*(x)$$

$\bar{X}_i$들은 정규분포를 하고 독립이며, 평균은 μ, 분산은 $1/W_i$이다. 따라서 $\delta^*(x)$는 $\bar{X}_i$의 선형함수로서 평균은 μ이고 분산은 $(\sum W_i)^{-1}$이다. 그러므로 추정치 δ^*에 대응하는 위험은 γ이고, δ^*는 μ의 미니맥스 추정치이다.

5. 정규성의 가정을 제거

이제 Y의 분포에 대한 가정들과 X_i의 조건부 분포에 대한 가정들을 제거한다. Y는 평균이 μ, 분산이 σ_0^2 보다 작은 확률변수이고, 각 i에 대하여, 주어진 μ_i에 대하여, X_i의 조건부 분포는 평균이 μ_i이고, 분산은 σ_i^2보다 작거나 같다. 더 일반적인 조건들 아래서 μ의 추정치 δ^*에 대응하는 위험은 $R(\mu, \delta^*) \leqq \gamma$ 이므로, δ^*는 미니맥스 추정치이다.

6. 표본의 크기의 미니맥스 선택

주어진 m집합들에 대한 미니맥스 표본추출은 위험이 최소가 되도록 n_i들을 선택하는 것이다. 즉, 주어진 m에 대하여 n_i의 선택은 위험

$$R(\mu, \delta^*) = [\sum n_i / (n_i \sigma_0^2 + \sigma_i^2)]^{-1} + \sum C_i n_i$$

를 최소화함으로서 구해진다.

7. 이 단계에서 같은 크기의 표본추출

각 집락에서 같은 크기의 표본들에 대하여, $n_i = n$, $C_i = C_w$ 그리고 $\sigma_i^2 = \sigma_w^2$ 이라고 모든 i에 대하여 가정한다. 손실함수는

$$E(X_i \mid \mu_i) = \mu_i \quad E[(X_i - \mu_i)^2 \mid \mu_i] \leqq \sigma_w^2$$

이다. 그리고

$$L(\mu, \delta) = (\delta - \mu)^2 + C_b m + C_w mn$$

이다.

앞 절에서 언급한 과정에 의하여, n_i를 n, σ_i^2을 σ_w^2 그리고 c_i를 c_w로 대치하여 다음의 결과를 얻는다.

(i) 베이즈 추정치 $\delta_\theta = n\theta^2 \sum \bar{X}_i / (mn\theta^2 + n\sigma_b^2 + \sigma_w^2)$

(ii) 베이즈 위험 $r_\theta = [mn/(n\sigma_b^2 + \sigma_w^2) + \theta^{-2}]^{-1} + C_b m + C_w mn$

(iii) 미니맥스 추정치 $\delta = \bar{X}$

(iv) 미니맥스 위험 $= (\sigma_b^2/m) + (\sigma_w^2/mn) + C_b m + C_w mn$

8. m과 n을 선택하는 미니맥스 전략

위험 $(\sigma_b^2/m) + (\sigma_w^2/mn) + C_b m + C_w mn$을 최소화시키는 m과 n을 구하는 방법은 계산상 어렵다. 그러나 어렵지 않게 근사해를 구할 수 있다. 즉,

$$m \cong \sigma_b / C_b^{\frac{1}{2}}$$

$$n \cong (C_b/C_w)^{\frac{1}{2}} (\sigma_w/\sigma_b)$$

9. 표본의 크기의 미니맥스 선택
(근사적으로 구하는 방법)

모든 i에 대하여 $C_i\sigma_i^2 = C$라 가정하면, 주어진 m집락들에 대하여 미니맥스 위험은

$$R(\mu, \delta^*) = \{\sum 1/[\sigma_b^2 + (C/C_i n_i)]\}^{-1} + C_b m + \sum C_i n_i$$

이다. 이 식은 주어진 $\sum C_i n_i$에 대하여 최소화되고, $\sum (C_b^2 + (C/C_i n_i))^{-1}$일 때에 최대화된다. 그런데 모든 i에 대하여 $C_i n_i$가 같아야 될 것이 요구된다.

또한 $\sigma_i^2/n_i = C_i\sigma_i^2/C_i n_i = k$ 이다. 즉 W_i는 모든 i에 대하여 같고, 추정치 δ^*는 (집락)표본평균들의 평균이 된다. 따라서

$$R(\mu, \delta^*) = (\sigma_b^2/m) + (k/m) + C_b m + (Cm/k)$$

이다. 이 위험은 $k = mc^{\frac{1}{2}}$에 대하여 최소이고, 주어진 m에 대하여 n_i 값이 결정된다.

$n_1, n_2, \cdots, n_m$에 대한

$$R(\mu, \delta^*) = [\sum n_i/(n_i\sigma_b^2 + \sigma_i^2)]^{-1} + C_b m + \sum C_i n_i$$

의 최소값은

$$R(\mu, m, c_i, \cdots, C_m, \sigma_1^2, \cdots, \sigma_m^2) = (\sigma_b^2/m) + C_b m + 2C^{\frac{1}{2}}$$

로 주어진다.

C_i와 σ_i^2의 결합분포의 기대치를 취하여,

$$ER(\mu, m, c_1, \cdots, c_m, \sigma_1^2, \cdots, \sigma_m^2) = (\sigma_b^2/m) + c_b m + 2c^{\frac{1}{2}}$$

이다. 그러면 이 식을 최소로 하는 m의 값은

$m = [(\sigma_b^2/c_b) + \frac{1}{4}]^{\frac{1}{2}}$에 가까운 정수

n_i의 미니맥스 선택은 w_i를 n_i에 비례하게 만들므로서 얻어지는데,

$$n_i = \bar{n} + (\sigma_w^2 - \sigma_i^2) / \sigma_b^2$$

이 된다. 여기서, $\bar{n}$과 σ_w^2은 각각 n_i와 σ_i^2의 평균이다.

10. 유한모집단 크기가 같지 않은 집락들

유한모집단이 M개의 부분-모집단으로 이루어졌고, i번째 부분-모집단으로 이루어졌고, i번째 부분-모집단($i=1,\cdots,M$)은 N_i개의 단위들로 이루어 졌다고 가정하자. X_{ij}는 i번째 부분-모집단에서 j번째 단위를 나타내고, 평균은 $\mu_i = N_i^{-1} \sum X_{ij}$이다. 이단표본추출로서 모집단 평균 $\mu = (\sum N_i)^{-1} \sum N_i \mu_i$를 추정하고자 한다. 첫 단계에서 단순확률추출로서 M집락에서 m집락을 뽑아낸다. 두번째 단계에서는 첫 단계에서 뽑힌 집락의 크기가 N_i라면 크기 n_i의 단순확률표본을 추출한다. 이렇게 해서 총 $\sum n_i$개의 표본이 추출된다. 만일 C_b가 첫 단계에서 하나의 집락을 추출하는 비용, C_i는 i번째 집락에서 한 단위의 부차-표본추출하는 비용이라면 손실함수 L은 다음과 같이 주어진다고 가정한다.

$$L(\mu, \delta) = (\delta - \mu)^2 + c_b m + \sum C_i n_i$$

여기서, δ는 μ에 대한 추정치이다.

베이즈 추정치 δ_ξ와 베이즈 위험 r_ξ는 다음과 같이 주어진다.

$$\delta_\xi(x) = E(\mu/x)$$

$$r_\xi(x) = E[E(\mu - \delta_\xi(x))^2 | x]$$

11. 베이즈 추정치들

μ는 $N(0,\theta^2)$인 정규분포를 한다. 그리고 주어진 μ에 대하여 ω_b의 분포는 평균이 μ, 분산이 $\sigma_b^2(M-1)/M$인 M-변량 정규분포를 한다. 그리고 ω_i의 분포는 평균이 μ_i, 분산이 $\sigma_i^2(N_i-1)/N_i$인 정규분포를 한다. 그리고 x_{ij}, $i=1,\cdots,m$; $j=1,\cdots,n_i$는 $\sum(n_i+1)$-변량 정규분포를 한다. μ_i들의 분포는 μ에 의존하고, 집합 $(\bar{x}_1,\cdots,\bar{x}_m)$는 집합 $(\mu_1,\cdots,\mu_m)$에 충족통계량이며, 또한 μ에 대해서도 충족통계량이다.

주어진 μ_i에 대하여 $\bar{x}_i$의 분포는 $N(\mu_i,v_i)$이다. 단, $v_i=(n_i^{-1}-N_i^{-1})\sigma_i^2$이다. 표본 추출이 각 집락에서 독립이므로 $\bar{x}_i$의 결합분포는 m개의 정규분포들의 곱으로 표시된다. 표본추출된 μ_i들은 평균이 μ, 분산이 $\sigma_b^2(M-1)/M$이고, 공분산이 $-\sigma_b^2/M$인 m-변량 정규분포를 한다. $\bar{x}^*$, μ^* 그리고 e는 성분들이 각각 $(\bar{x}_1,\cdots,\bar{x}_m)$, $(\mu_1,\cdots,\mu_m)$ 그리고 $(1,\cdots\cdots 1)$인 $m\times 1$ 열벡터를 나타낸다. 그리고 I는 $m\times m$ 단위행렬이다. 만일 $N(u,\sum)$가 평균벡타 v 그리고 공분산 행렬 $\sum$인 k-변량 정규분포를 표시한다면, 주어진 v에 대하여, μ^*는 $N(\mu_e,A)$인 분포를 한다. 단, $A=\sigma_b^2(I-M^{-1}ee')$이다. 그리고 주어진 μ^*에 대하여, $\bar{x}^*$의 분포는 $N_m(\mu^*,B)$이다. 단, B는 대각원소들이 v_i인 대각행렬이다. 주어진 μ에 대하여 $\bar{x}^*$는 $N(\mu_e,W)$로 분포된다. 단, $W=A+B$, μ는 사전분포 $N(0,\theta^2)$를 하고, μ와 $\bar{x}^*$의 결합분포는 $(m+1)$-변량 정규분포를 하므로, 주어진 $\bar{x}^*$에 대한 μ의 조건부 분포는 $N(\alpha,\beta)$이다. 단, $\alpha=e'W^{-1}\bar{x}^*/(e'W^{-1}e+\theta^{-2})$ 그리고 $\beta=(e'W^{-1}e+\theta^{-2})^{-1}$

이와같이 하여, 베이츠 추정량은 $\delta_\theta = \alpha$, 그리고 β의 분산은 x^*에 독립이므로 베이즈 위험은 $r_\theta = \beta$이다.

12. 미니맥스 추정량

$\theta \to \infty$일 때에, $r_\theta \to r = (e'W^{-1}e)^{-1}$이다. 이제 우리는 추정량 δ^*가 존재하여, 그 추정량의 위험이 r보다 크지 않은지 찾아 보아야 한다. 그러면

$$\delta^*(x) = \lim_{\theta \to \infty} \delta_\theta(x) = e'W^{-1}\bar{x}^*/e'W^{-1}e$$

추정량 δ^*에 대응하는 위험은 표본추출 비용과 관계없이 다음과 같이 주어진다.

$$E_w(\delta^*-\mu)^2 = [1/(e'W^{-1}e)^2]E[e'W^{-1}(\bar{x}^*-\mu e)]^2$$
$$= e'W^{-1}W_1W^{-1}e/(e'W^{-1}e)^2$$

단, $W_1 = A_1 + B_1$, $\quad A_1 = \eta_b^2(I - M^{-1}ee')$

그리고 B_1은 대각원소가 $v_i = (n_i^{-1} - N_i^{-1})\eta_i^2$

인 대각행렬이다. 단,

$$\eta_b^2 = (M-1)^{-1}E\{\sum(\mu_i-\mu)^2/\mu\} \leqq \sigma_b^2$$

그리고

$$\eta_i^2 = (N_i-1)^{-1}E\{\sum(X_{ij}-\mu_i)^2 | \mu_i)^2 | \mu\} \leqq \sigma_b^2, \quad i = 1, \cdots, M$$

대수적 계산을 하여, $e'W^{-1}W_1W^{-1}e \leqq e'W^{-1}e$, 따라서 $E(\delta^*-\mu)^2 \leqq (e'W^{-1}e)^{-1} = r$. 즉 $\delta^*(x)$는 μ의 미니맥스 추정량이다. 이 추정량은 $\bar{x}_1, \cdots, \bar{x}_m$의 가중평균으로

$$\delta^*(x) = \sum \omega_i \bar{x}_i / \sum \omega_i$$

로 쓸 수 있다. 단, W_i는 W^{-1}의 i번째 원소로

$$w_i = [d_i(1-(\sigma_b^2/M)\sum d_j^{-1})]^{-1}$$

로 주어진다. 단, $d_i = (\eta_i^{-1} - N_i^{-1})\sigma_i^2 + \sigma_b^2$

표본추출 비용과 관계없이, 최대로 가능한 위험은

$$r = (e'w^{-1}e)^{-1} = (\sum W_i)^{-1} = (\sum d_i^{-1})^{-1} - M^{-1}\sigma_b^2$$

으로 주어진다.

13. 표본크기의 미니맥스 선택

표본으로 추출된 m개의 집락들에 대하여 n_i를 "최적"이 되도록 정하고자 한다. 즉, 주어진 m개의 집락에 대하여 다음의 위험이 최소가 되도록 n_i를 정한다.

$$R(\mu, \delta^*) = (\sum d_i^{-1})^{-1} - M^{-1}\sigma_b^2 + c_b^2 + c_b m + \sum c_i n_i$$

위의 위험을 n_i가 양수가 되도록 최소화시켜야 한다. 이 문제는 수치해석의 방법으로 풀 수 있다.

14. 유한모집단, 같은 크기의 집락들에서 같은 크기의 표본추출

특별한 경우로서 각 i에 대하여, $N_i = N$ 그리고 $n_i = n$이라고 가정하자. 우리는 각 i에 대하여 $c_i = c_w$ 그리고 $\sigma_i^2 = \sigma_w^2$ 이라고 가정하자. 각각의 추출된 집락으로부터 관찰치들의 수는 같다.

m과 n이 이미 결정되었다고 가정하여, 이 고정된 m과 n에 관하여 μ에 대한 베이즈와 미니맥스 추정량을 구한다. 우리는 11절과 같은 과정을 통하여

베이즈 추정량 $\delta_\theta = \theta^2\bar{x} \cdots /(\theta^2 + V)$

베이즈 위험 $\gamma_\theta = (\theta^{-2} + V^{-1})^{-1} + c_b m + c_w mn$

단, $\bar{x}\cdot\cdot$와 V는 다음과 같이 정의된다.

$$\bar{x}.. = m^{-1} \sum \bar{x}_i$$

그리고

$$V = (m^{-1} - M^{-1})\,\sigma_b^2 + m^{-1}(n^{-1} - N^{-1})\,\sigma_w^2$$

주어진 m과 n에 대하여 전체에 대한 평균 $\bar{x}..$는 미니맥스 위험은

$$R = (m^{-1} - M^{-1})\,\sigma_b^2 + m^{-1}(n^{-1} - N^{-1})\,\sigma_w^2 + c_b m + c_w mn$$

미니맥스 전략은 최대로 허용된 분산에 대해서 m과 n의 "최적"의 값을 구하는 것이다. 근사값으로

$$m \cong [(\sigma_b^2 - \sigma_w^2/N)/c_b]^2$$

$$m \cong [(c_b/c_w)(\sigma_w^2/(\sigma_b^2 - \sigma_w^2/N))]^{\frac{1}{2}}$$

Ⅳ. 베이지안 층화 二相(two-phase) 표본추출
(Bayesian stratified two-phase sampling)

요약. 이 논문에서는 첫번째 相(first-phase)에서 얻어진 정보를 사용해서, 이상표본추출(two-phase)과정의 두번째 相(phase)에서 표본의 최적 할당을 한다. 여기서 두 가지 다른 방법을 시도하였다.; 베이지안 事後(posterior) 分析과 베이지안 사전 사후분석 (preposterior). 모든 장애모수(nuisance parameter)들을 모른다는 조건하에 두가지 다른 할당방법을 유도하고 예를 들어서 설명했다. 모든 장애모수들을 안다고 했을 때의 할당방법은 1965년 Ericson에 의해 밝혀진 바 있다.

1. 서 론

k개 층의 관찰치들을 x_{ij} 또는 y_{ij} ($i=1, \cdots, k$)라 하자. 문자 x는 첫번째 相의 관찰치를, y는 두번째 相의 관찰치를 의미한다. x_{ij} (또는 y_{ij})가 평균 μ_i와 분산 σ_i^2을 갖는 정규분포를 한다고 가정하고, 총 모집단에 대한 i번째 층의 비율을 π_i라 하자 단, $\pi_1 + \cdots + \pi_k = 1$, c_i는 i번째 층에서 얻어지는 관찰치에 대한 비용(cost)이고, 그 비용을 안다고 하자 C를 이상(two-phase) 표본추출할 때 총비용이라 하자. 비용 αc를 첫번째 상의 관찰치들을 추출할 때 드는 비용이라 하자. ($0 < \alpha < 1$, α는 총비용에 대한 비율) 여기서, i번째 층의 표본의 크기는 n_i이다. 그러면 두번

째 상에서의 비용은 $(1-\alpha)C$이다. N_i는 두번째 상의 i번째 층에서 얻어진 관찰치들의 갯수이다. 따라서

$$\sum_{i=1}^{k} c_i n_i = \alpha c, \qquad \sum_{i=1}^{k} c_i n_i = (1-\alpha)c$$

표본추출의 주요 목적은 가능한한 精度가 높게 전체평균 $\mu = \pi_1 \mu_1 + \cdots + \pi_k \mu_k$를 추정하는 것이라고 가정하자. 두번째 상에서 N_i의 크기를 어떻게 잡을 것인가? 이 질문의 대답은 모수에 관하여 어떤 가정을 하느냐와 방법론에 달려있다.

2. 베이지안 접근; π_i를 알 때

우리는 π_i를 알고, α, c_i, c, k 그리고 n_i는 이미 고정되어 있다고 가정하자. 첫번째 할당(allocation)을 하기 전에 사전 정보를 안다고 가정한다. 즉 μ_i와 σ_i^2에 대한 사전분포가 독립이고, 국소 일양분포를 한다고 하자. 따라서

$$P(\mu_i)d\mu_i \propto d\mu_i, \quad p(\sigma_i^2)d\sigma_i^2 \propto d\sigma_i^2/\sigma_i^2$$

만약 관찰치들 x_{ij}가 이상추출 과정중 첫번째 상(first phase)에서 얻어졌다면, 우도함수는 아래와 같다.

$$\prod_{i=1}^{k} (\sigma_i^2)^{-\frac{1}{2}n_i} (2\pi)^{-\frac{1}{2}n_i} \exp\left\{-\frac{n_i(\bar{x}_i-\mu_i)^2+(n_i-1)S_i^2}{2\sigma_i^2}\right.$$

단, $n\bar{x}_i = x_{i1} + x_{i2} + \cdots\cdots + x_{in_i}$ 그리고

$$(n_i-1)\,S_i^2 = \sum_{j=1}^{n_i} (x_{ij}-\bar{x}_i)^2$$

첫번째 할당후에 μ_i와 σ_i^2 $(i=1, 2, \cdots, k)$의 結合事後分布는 사전분포와 우도함수의 곱으로

$$\prod_{i=1}^{k} p(\mu_i, \sigma_i^2 \mid X_i) \propto \pi(\sigma_i^2)^{-\frac{1}{2}(n_i+2)}$$

$$\exp\left\{-\frac{n_i(\bar{x}_i-\mu_i)^2+(n_i-1)s_i^2}{2\sigma_i^2}\right\}$$

이 분포는 두번째 상의 표본추출 전에 사전분포로 사용될 수 있다. y_{ij}가 두번째 상에서 얻어진다면, 우리는 n_i, $\bar{x}_i$ 그리고 s_i^2 대신에

N_i; $\bar{y}_i = \sum y_{ij}/N_i$ 그리고 $W_i^2 = \sum (y_{ij}-\bar{y}_i)^2/(N_i-1)$ 를 앞에 우도함수에 대치하므로 새 우도함수를 얻는다. 그러면, 새로운 μ_i와 σ_i^2에 대한 事後分布를 얻을 수 있다. 즉,

$$\prod_{i=1}^{k} p(\mu_i, \sigma_i^2 | x_i, y_i) d\mu_i \, d\sigma_i^2 \propto \prod_{i=1}^{k} (\sigma_i^2)^{-\frac{1}{2}(N_i+n_i+2)} \exp(-\tfrac{1}{2} Q_i/\sigma_i^2)$$

단, $Q_i = n_i(\bar{x}_i-\mu_i)^2 + N_i(y_i-\mu_i)^2 + (n_i-1)s_i^2 + (N_i-1)w_i^2$

$= (N_i+n_i)(\mu_i-\bar{Y}_i)^2 + SS_i$

그리고 $\bar{Y}_i = (n_i\bar{x}_i + N_i\bar{y}_i)/(n_i+N_i)$,

$SS_i = n_iN_i(\bar{x}_i-\bar{y}_i)^2/(n_i+N_i) + (n_i-1)s_i^2 + (N_i-1)w_i^2$

σ_i^2 $(i=1,2,\cdots,k)$에 대하여 적분하므로서 μ_i의 주변분포를 얻는다.

註. $\int_0^\infty x^{-\frac{1}{2}(N+1)} \exp(-A/x)dx = A^{-\frac{1}{2}(N-1)} \Gamma(\frac{1}{2}N-\frac{1}{2})$

따라서

$$\prod_{i=1}^{k} p(\mu_i | x_i, y_i) \propto \prod_{i=1}^{k} Q_i^{-\frac{1}{2}(N_i+n_i)} \propto \prod_{i=1}^{k} \{1+T_i^2/(N_i+n_i-1)\}^{-\frac{1}{2}(N_i+n_i)}$$

단, $T_i^2 = (N_i+n_i)(N_i+n_i-1)(\mu_i-\bar{Y}_i)^2/(SS_i)$

명백히 T_i는 자유도가 N_i+n_i-1인 t분포를 한다. 그리고 T_j, $j \neq i$와 독립이다. 그러므로 $E(T_i) = 0$이다. 따라서

$E(\mu_i) = \bar{Y}_i$ 그리고

$$V(T_i) = E(T_i^2) = (N_i+n_i-1)/(N_i+n_i-3)$$

또한

$$V(\mu_i) = E(\mu_i - \bar{Y}_i)^2 = \{SS_i/(N_i+n_i-3)\}/(N_i+n_i)$$

N_i에 대한 할당을 하기 위하여, 이상 관찰 전에 결정을 해야 한다. 우리는 $V(\mu_i)$의 기대치 대신에 $E(SS_i)$를 사용한다. 그런데 $E(SS_i)$는 주어진 x_{ij}에 대하여 관찰치 y_{ij}의 분포위에서 취한 기대치이다. $E(SS_i)$의 첫번째와 세번째의 항은

$$\frac{n_i N_i}{N_i+n_i}\left(1-\frac{1}{n_i}\right)\left(1+\frac{n_i}{N_i}\right)\left(\frac{s_i^2}{n_i-3}\right)+(n_i-1)S_i^2 = (n_i-1)\frac{n_i-2}{n_i-3}S_i^2$$

이다.

$\sigma_i^2 = (n_i-1)S_i^2 X_{ni-1}^{-2}$ 이므로 세번째 항

$E\{(N_i-1)w_i^2\}$은 $(N_i-1)(n_i-1)s_i^2/(n_i-3)$ 이다.

따라서

$$v_i^2 = \sum(x_{ij}-\bar{x}_i)^2/(n_i-3) = (n_i-1)S_i^2/(n_i-3)$$

이라고 정의하면

$$E(SS_i) = (N_i+n_i-3)(n_i-1)S_i^2/(ni-3) = (N_i+n_i-3)v_i^2$$

그러므로

$$E_y\{V(\mu)\} = V(\sum \pi_i \mu_i) = \sum_{i=1}^{k} \pi_i^2 v_i^2/(N_i+n_i)$$

다음의 제한 조건아래서

$$N_i \geqq 0,\ \sum c_i N_i = (1-\alpha)c \quad (\text{제한 조건})$$

$E_y\{V(\mu)\}$를 최소화하기 위한 Ni을 선택한다.

Lagrange 승수법을 써서

$$N_i = \frac{c}{c_i}q_i - n_i$$

단, $q_i = \pi_i v_i c_i^{\frac{1}{2}} / \sum_{i=1}^{k} \pi_j u_j c_j^{\frac{1}{2}}$

이차 도함수의 행렬이 양정치이므로 ~~극소값이 된다.~~

어떤 N_i가 음수일 때 재할당

q_i가 작고, n_i가 클 때는 N_i가 음수가 될 수 있다. 이것은 i번째 층이 초과해서 추출되었음을 뜻한다. N_i가 음수일 때는 $N_i(-)$, 영일 때는 $N_i(0)$, 양수일 때는 $N_i(+)$로 나타내자. 첫번째 재할당을 N_i이라 하자. 만일 $N_i = N(-)$ 또는 $N_i(0)$이면, $N_i' = 0$으로 놓는다. 즉 그 그룹에서 더 이상 뽑지 않는다. 이것은 나머지 관찰치들을 나머지 그룹들에 할당하는 것을 의미한다. 새로운 조건

$$\sum_{+} c_i N_i' = (1-\alpha)C$$

에 대해 $E(V(\mu))$를 최소화하여, $N_i(+)$ 대신에

$$N_i' = \frac{\{(1-\alpha)C + \sum_{+} n_j c_j\}}{c_i} \times \frac{\pi_i v_i c_i^{1/2}}{\sum_{+} \pi_j v_j c_j^{1/2}} - n_i$$

를 얻는다. 만일 어떤 N_i'이 음수이거나 영이면, $N_i'' = 0$으로 놓고 재할당을 한다.

3. 사전사후 (preposterior) 분석 ; π_i를 알때

우리는 2절과 같은 작업과 기호를 사용하지만 다른 형태의 분석을 하여 다른 해를 얻는다. 이 사전사후 분석에서 첫번째 위상(phase) 표본추출이후에 사후분포를 사용한다. 이것은 모수들을 적분하므로서 이상 (two-phase)에서 표본평균들 $\bar{y}_i$의 미래 분포를 제공하기 위해서 이다. 그러면 $V(\sum \pi_i \bar{y}_i)$를 최소화하기 위한 N_i을 선택한다.

첫 위상이후에, i번째 층에서 n_i개의 관찰치 $x_{ij}(j=1, \cdots, n_i)$

가 취해졌다면, μ_i 와 σ_i^2 ($i=1,\cdots,k$)의 결합사후분포는 $\prod_{i=1}^{k} p(\mu_i, \sigma_i^2 | X_i)$ 이다. 새로운 표본 $y_{i1},\cdots,y_{iN_i}$ 가 i번째 층으로 부터 뽑혀진다면, 주어진 μ_i, σ_i^2 에 대한 평균들 $\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_k$ 의 결합확률분포는

$$p(\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_k | \mu_i, \sigma_i^2) = \pi \frac{1}{\sigma_i}\sqrt{\left(\frac{N_i}{2\pi}\right)} \exp\left\{-\frac{N_i(\bar{y}_i-\mu_i)^2}{2\sigma_i^2}\right\}$$

이다.

$\bar{y}_i, \mu_i, \sigma_i^2$ ($i=1,\cdots,k$)의 결합분포는

$$\pi P(\bar{y}_i, \mu_i, \sigma_i^2) d\bar{y}_i\, d\mu_i\, d\sigma_i^2$$

$$= \pi P(\bar{y}_i, \cdots, \bar{y}_k | \mu_i, \sigma_i^2) \times \pi P(\mu_i, \sigma_i^2 | X_i)$$

$$\alpha\ \pi \frac{N_i^{\frac{1}{2}}}{(\sigma_i^2)^{\frac{1}{2}(n_i+3)}} \exp\left[-\frac{1}{2\sigma_i^2}\left\{(\bar{y}_i-\mu_i)^2 N_i + (\bar{x}_i-\mu_i)^2 n_i + (n_i-1)S_i^2\right\}\right] d\bar{y}_i\, d\mu_i\, d\sigma_i^2$$

위 식에서 괄호 안을 다시 정리하여

$$Q_i' = (n_i+N_i)\left\{\mu_i - \frac{n_i\bar{x}_i+N_i\bar{y}_i}{n_i+N_i}\right\}^2 + (n_i-1)S_i^2 + \frac{n_iN_i(\bar{x}_i-\bar{y}_i)^2}{n_i+N_i}$$

로 쓸 수 있다.

μ_i 와 σ_i^2 에 대하여 적분하므로서 $\bar{y}_i$ 의 사전사후 분포는

$$\pi P(\bar{y}_i) \ \alpha\ \pi\left(1+\frac{t_i^2}{v_i}\right)^{-\frac{1}{2}(\nu_i+1)}$$

단, $t_i^2 = n_iN_i(\bar{y}_i-\bar{x}_i)^2/\{S_i^2(N_i+n_i)\}$

그리고 $v_i = n_i-1$

$\bar{y}_i$들은 독립임을 주목하라. 따라서

$E(\bar{y}_i) = \bar{x}_i$ 그리고 $V(t_i) = (n_i-1)/(n_i-3)$

이므로,

$$V(\bar{y}_i) = (n_i-1)(N_i+n_i)S_i^2/\{(n_i-3)n_iN_i\}$$

$$= (1-\frac{1}{n_i})(1+\frac{n_i}{N_i})(\frac{s_i^2}{n_i-3})$$

첫 위상이후에 $\sum\pi_i\mu_i$는 사후분포로 기대치 $\sum\pi_i\bar{x}_i$를 갖고, $\sum\pi_i\bar{y}_i$는 사전사후로 기대치 $\sum\pi_i\bar{x}_i$를 갖으므로 $\sum\pi_i\bar{y}_i$는 $\mu=\sum\pi_i\mu_i$의 事前事後 추정치이다. 이제 이 추정치의 분산은

$$f = V(\sum\pi_i\bar{y}_i) = \sum A_i(1+\frac{n_i}{N_i})$$

단, $A_i = \pi_i^2(1-\frac{1}{n_i})(\frac{S_i^2}{n_i-3}) = \pi_i^2 v_i^2/n_i$

제한조건 $\sum c_iN_i = (1-\alpha)C$아래서 f를 최소화시키면 할당

$$N_i^* = (1-\alpha)Cq_i/c_i$$

을 얻는다.

4절에서는 공식의 사용법을 실제 데이타 분석으로 설명한다.

4. 데이타의 분석

$k=6$개의 그룹들, $c_i=1$ 그리고 $\sum(n_i+N_i)=120$인 경우에 대한 데이타 분석 결과를 다음 표에 나타낸다. α, n_i 그리고 π_iv_i의 값이 알려진 경우 그 결과는 다음과 같다. 우리는 n_i의 그룹들을 A,B,C로 표 1에 그리고 π_iv_i의 그룹들을 a,b,c,d로 표 2에 나타낸다.

표 1. n_i의 그룹들

α	n_1	n_2	n_3	n_4	n_5	n_6	표시
0.5	10	10	10	10	10	10	A
0.5	5	5	5	15	15	15	B
0.25	3	3	3	3	9	9	C

표 2. $\pi_i v_i$ 의 그룹들

$\pi_1 v_1$	$\pi_2 v_2$	$\pi_3 v_3$	$\pi_4 v_4$	$\pi_5 v_5$	$\pi_6 v_6$	표시
1	2	3	4	5	6	a
1	1	1	1	1	1	b
1	1	1	5	5	5	c
3	2	1	1	2	3	d

공식 $N_i = \frac{C}{c_i} q_i - n_i$, N_i', $N_i^* = (1-\alpha) C q_i / c_i$ 를 써서 계산된 표본의 크기들은 다음과 같다.

표 3. 베이지안 접근 (N,N′) 과 사전사후 분석을 사용해 계산된 표본의 크기들

Aa	N	-4.29	1.43	7.14	12.86	18.57	24.29
	N′	0	1.0	6.5	12.0	17.5	23.0
	N*	2.86	5.71	8.57	11.43	14.29	17.14
Ab	N=N*	10	10	10	10	10	10
Ac	N	-3.33	-3.33	-3.33	23.33	23.33	23.33
	N′	0	0	0	20	20	20
	N*	3.33	3.33	3.33	16.67	16.67	16.67
Ad	N	20	10	0	0	10	20
	N*	15	10	5	5	10	15
Ba	N	0.71	6.43	12.14	7.86	13.57	19.29
	N*	2.86	5.71	8.57	11.43	14.29	17.14

Bb	N	15	15	15	5	5	5
	N*	10	10	10	10	10	10
Bc	N	1.67	1.67	1.67	18.33	18.33	18.33
	N*	3.33	3.33	3.33	16.67	16.67	16.67
Bd	N	25	15	5	-5	5	15
	N'	23.64	14.09	4.55	0	4.09	13.64
	N*	15	10	5	5	10	15
Ca	N	2.71	8.43	14.14	19.86	19.57	25.29
	N*	4.29	8.57	12.86	17.14	21.43	25.71
Cb	N	17	17	17	17	11	11
	N*	15	15	15	15	15	15
Cc	N	3.67	3.67	3.67	30.33	24.33	24.33
	N*	5	5	5	25	25	25
Cd	N	27	17	7	7	11	21
	N*	22.5	15	7.5	7.5	15	22.5

5. 결과의 음미

실제 데이타의 예에서 $\sum(N_i+n_i) = 120$ 인 경우만을 생각해 보았다. 작은 예산으로 베이지안 접근은 음수의 N_i가 나올수도 있으나 재할당시에는 어떤 층에서는 더이상의 관찰치가 필요하지 않다. 이상 표본추출시에 더 많은 관찰치들이 이용되면은 음수의 Ni는 나타나지 않고, 모든 층에서 표본추출이 된다. 같은 결론이 Ericson

(1965)에 의해 σ^2이 기지일 때에 밝혀졌다. 사전사후 접근에서는 오직 이 단계(second stage)에서의 관찰치들이 할당되므로 양수의 할당이 이루어진다.

6. 베이지안 접근 ; π_i를 모를 때

π_i가 더이상 알려져 있지 않을 때도 2절과 같은 과정을 밟는다. 우리는 π_i의 사전분포가 디리슈레 분포

$$P(\pi_1, \cdots, \pi_k) = \frac{P(v_1+\cdots+v_k)}{\pi P(v_j)} \pi_1^{v_1-1} \cdots \pi_k^{v_k-1}$$

이다. 단, $\pi_1+\cdots+\pi_k=1$ 디리슈레 분포는 여러가지의 이유로 사전분포의 중요한 것중에 하나이다. 특히 모든 $v_i=1$이면, 국소 균일 사전분포이다. μ_i와 σ_i^2에 대한 사전분포는 2절과 같다고 하자.

π_i가 알려져 있지 않으므로, 첫 위상에서의 표본추출은 다음과 같다. 우리는 첫 위상에서 예산의 βC를 소비한다고 가정하자 모두 $[\beta C/(\max c_i)]$개의 표본을 뽑아서 얼마나 많은 n_i가 i층에 속하는지를 계산해 보자 $\sum n_i c_i = \alpha C$라 하자. 우도함수를 얻기 위하여, 우리는 x_{ij}에 의존하는 우도에 n_i에 의존하는 다항 우도를 곱하므로서 얻어진다. 즉,

$$\pi \frac{1}{\sigma_i^{n_i}} \exp\{-\frac{n_i(\bar{x}_i-\mu_i)^2+(n_i-1)S_i^2}{2\sigma_i^2}\} \frac{(n_1+\cdots+n_k)!}{n_1!\cdots n_k!}$$

$$\pi_1^{n_1} \cdots \pi_k^{n_k}$$

이 식에 $P(\mu_i)$, $P(\sigma_i^2)$ 그리고 $P(\pi_1, \cdots, \pi_k)$의 사전분포를 곱하므로서 첫 위상이후의 μ_i, σ_i^2 그리고 π_i에 대한 사후분포를 얻는다. 이것은 이상(second phase)에서 사전분포로 쓰이고 우리는 i번째 층에 얼마나 더 추가적인 관찰치들 N_i를 취할 것인지를 정하고자 한다. 이상(second phase)에서 N_i가 지정되었을 때에는 우도함수는

$$\prod_{i=1}^{k} \frac{1}{\sigma_i^{N_i}} \exp\{-\frac{N_i(\bar{y}_i-\mu_i)^2+(N_i-1)w_i^2}{2\sigma_i^2}\}$$

에 비례한다. 이것과 사전분포를 결합하여 사후분포를 얻는다. 즉,

$$\pi P(\mu_i, \sigma_i^2, \pi_i | \bar{x}_i, n_i, \bar{y}_i, N_i) d\mu_i d\sigma_i^2 d\pi_i \propto \pi \sigma_i^{-(N_i+n_i)-2}$$

$$\times \exp\{-\frac{n_i(\bar{x}_i-\mu_i)^2+(n_i-1)S_i^2+N_i(\bar{y}_i-\mu_i)^2+(N_i-1)w_i^2}{2\sigma_i^2}\}$$

$$\pi_i^{n_i+\nu_i-1}$$

만일 $V(\mu)$가 계산되어, y의 미래분포 위에서 그 기대치를 취하면

$\sum \frac{D_i^2 v_i^2}{N_i+n_i}$ + N_i를 포함하지 않는 항

으로 된다. 단, $D_i = (n_i+u_i)/\{\sum(n_i+v_i)\} = E(\pi_i)$

그리고 $v_i^2 = (n_i-1)S_i^2/(n_i-3)$

이와 같이 하여서 우리는 정확히 $V(\sum \pi_i \mu_i)$를 얻지만 π_i가 미지일 경우에는 π_i 대신에 $E(\pi_i)$를 사용한다. 2절에서와 같이 풀면,

$$N_i = \frac{C}{C_i} q_i - n_i$$

단, $q_i = \nu_i C_i^{\frac{1}{2}} / \sum \pi_j \nu_j C_j^{\frac{1}{2}}$

7. 사전사후 분석 ; π_i 를 모를 때

이 절은 3절과 같으나 π_i 가 미지이므로 적절한 수정을 한다. 첫 위상의 표본추출 이후에 μ_i, σ_i^2, π_i 에 대한 사후분포를 우도함수와 사전분포로 구할 수 있다. π_i 에 대해 적분하므로 μ_i, σ_i^2 의 결합주변분포를 얻는다. $\bar{y}_1, \cdots, \bar{y}_k$ 의 결합분포함수는

$$P(\bar{y}_1, \cdots, \bar{y}_k \mid \mu_i, \sigma_i^2) = \pi \frac{1}{\sigma_N} \sqrt{(\frac{N_i}{2\pi})} \exp\{ -\frac{N_i(\bar{y}_i - \mu_i)^2}{2\sigma_i^2} \}$$

이다. μ의 추정량으로, 우리는 $\sum E(\pi_i)\bar{y}_i$ 를 사용하고,

$$\sum c_i N_i = (1-\alpha)C, \quad N_i \geqq 0$$

라는 조건에서 그것의 분산을 최소화시키는 N_i 를 선택한다. 분산을 최소화시키는 것은 3절에서 같은 계산을 D_i 를 π_i 로 대치하면 된다. 이와같이 π_i 를 D_i 로 대치하여 최선의 할당을 하면

$$N_i^* = \frac{\{(1-\alpha)C\} D_i v_i C_i^{\frac{1}{2}}}{c_i \sum D_j v_j c_j^{\frac{1}{2}}}$$

이다. π_i 가 미지일 때에, '$\hat{\pi}_i$ 가 기지인' 경우에 유사한 대답을 π_i 의 추정치 D_i 를 넣으므로서 얻어진다.

Ⅴ. 事前分布를 使用하는 최적층화 標本抽出
(optimum stratified sampling using prior information)

1. 序 論

事前 情報가 이용가능 할 경우에 k層에 최적 할당을 하고자 한다. 事前 情報로 正規分布를 함을 알 때에 그 情報를 알므로서의 효과를 찾는다.

母集團이 k개의 層으로 분할된다고 가정하자. μ_i를 i번째 층에서의 未知의 母集團 平均이라 하자. π_i는 i번째 층의 母集團에 대한 旣知의 비율이고, π와 μ는 π_i와 μ_i의 행벡터라 가정하자.

母平均 μ는 $\mu = \mu\pi^t$

이다. 첨자 "t"는 전치를 나타낸다.

층화표본은 $n = (n_1, \cdots, n_k)$로 정의되는데, $n_i \geq 0$는 i번째 층으로 부터 독립적으로 추출된 관찰치의 갯수이다.

$\bar{X} = (\bar{x}_1, \cdots, \bar{x}_k)$는 표본평균의 벡터이고, σ_i^2 ($i = 1, \cdots, k$)은 i번째 층의 旣知의 分散을 나타낸다고 하자. 주어진 μ에 대하여, $\bar{x}$는 k-변량 정규분포를 한다. 평균은 μ 그리고 분산-공분산 행렬은 V로 그 대각 원소는 $u_{ii} = \sigma_i^2/n_i$, $i = 1, \cdots, k$이다. 비용벡터는 $C = (C_1, \cdots, c_k)$로 정의되며, c_i는 i번째 층의 단위 관찰치당 비용이다.

μ_i에 대한 事前 情報나 지식이 존재한다고 가정한다. 즉, μ는 k-변량 정규분포로 평균 벡터가 m' 그리고 분산-공분산 행열

$V' = [v_{ij}]$ 임을 事前에 안다고 하자, 그러면 μ의 事後分布는 事前分布와 標本 情報로 부터 얻어진다.

μ에 관한 事後分布는 주어진 층화표본 $n > 0$와 관찰치 $\bar{x}$에 대하여, 평균 벡타가

$$m'' = [xN + m'N'][N'']^{-1}$$

그리고 분산-공분산 행렬

$$V'' = [N'']^{-1} = [N + N']^{-1}$$

이다. 단, $N = V^{-1}$ 그리고 $N' = (V')^{-1}$

모평균 μ는 正規 事前分布로 평균 $m' = m'\pi^t$ 그리고 분산 $v' = \pi V'\pi^t$, 正規 事後分布는 평균 $m'' = m''\pi^t$ 그리고 분산 $v'' = \pi (N + N')^{-1} \pi^t$ 이다.

우리는 주어진 조건 $cn^t = c$, (단, C는 이용가능한 전비용) 아래서 μ의 事後分散을 최소화시키는 층화표본 $n > 0$을 구하는 것이다.

2. 데이타의 分析

분산-공분산 행렬, V'이 대각행렬인 경우에는, 제한 조건이

$$\sum_{i=1}^{k} c_i n_i = c$$

일 때

$$N(V'') = \sum_{i=1}^{k} \frac{\pi_i^2}{(n_{ii} + n_i/\sigma_i^2)}$$

을 최소화시키는

$$n_i > 0, \quad i = 1, \cdots, k$$

를 찾는 문제가 된다. 단, $n'_i = 1/v'_i$는 대각행렬 N'의 원소이다.

$B_0 = \infty$, 그리고 $B_i = (\sigma_i\sqrt{c_i})/(\pi_i v'_i)$라 하고,

$$B_0 > B_1 \geq B_2 \geq \cdots \geq B_k$$

라 하자. 구간 $C > 0$가 다음과 같이 분할 될 수 있다. $r = 0, 1, \cdots, k-1$에 대하여

$$I_r = \{C \mid C_r < C \leq C_r - 1\}$$

라 하자. 단, $C - 1 = \infty$

그리고

$$C_r = B_{r+1} \sum \pi_i \sigma_i \sqrt{c_i} - \sum c_i \sigma_i^2 n'_i$$

최적의 표본크기는 $C \in I_r$에 대하여

$$n_i^0 = 0, \quad i \leq r$$

그리고

$$n_i^0 = \frac{\pi_i \sigma_i}{\sqrt{c_i}} \left[\frac{C + \sum \frac{c_j \sigma_j^2}{v'_i}}{\sum \pi_j \sigma_j \sqrt{c_j}} \right] - \frac{\sigma_i^2}{v'_i}, \quad i > r$$

事前 情報가 없을 때에는, 즉 事前 分散, u'_i가 매우 클 때에는, $v'_i \to \infty$로 놓으므로서 최적할당의 근사값을 얻는다.

그러면

$$n_i^0 = \frac{\pi_i \sigma_i}{\sqrt{c_i}} \frac{C}{\sum \pi_j \sqrt{c_j} \sigma_j}, \quad i = 1, \cdots, k$$

이 식은 잘 알려진 층화표본추출 할당 공식이다.

특히 $k = 2$일 때에는 다음과 같은 공식을 얻을 수 있다. 즉, 다음과 같은 조건

$$n''_i \geq n'_i, \quad i = 1, 2$$

그리고

$$d_1^2 = n''_{11} + d_2^2 n''_{22} = D$$

아래서

$$v'' = \frac{\pi_1^2 n''_{22} + \pi_2^2 n''_{11} - 2\pi_1 \pi_2 n'_{12}}{n''_{11} n''_{22} - (n'_{12})^2}$$

를 최소화시키는 n''_{11}과 n''_{22}을 선택하는 문제이다. 단, $n''_{ii} = n'_{ii} + (n_i / n\sigma_i^2), d_i = \sqrt{c_i}\ \sigma_i$ 그리고

$$D = C + d_1^2 n'_{11} + d_2^2 n'_{22}, \quad i = 1, 2$$

만일 "음이 아닌"이라는 조건 $n''_{ii} \geq n_{ii}$을 무시한다면, 이 문제는 라그랑지 승수법의 문제로 돌아간다. 그러면

$$L = v'' + \lambda (d_1^2 n''_{11} + d_2^2 n''_{22} - D),$$

단, λ는 승수이다. D가 변하므로 직선들

$$n''_{11} = e_{11} n''_{22} + e_{12}$$

그리고

$$n''_{11} = e_{21} n''_{22} + e_{22}$$

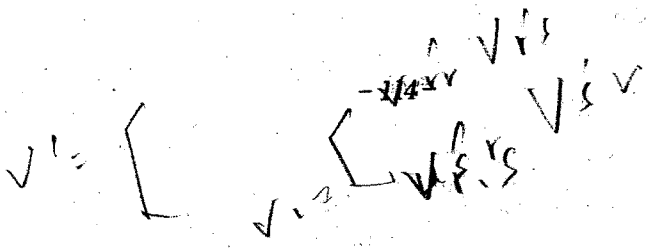
에 따라서 L의 임계점들이 속하게 된다.

k개의 층들을 r개의 層과 나머지 $S = k - r$ 層으로 분할하자. 未知의 모평균 μ를 $\mu = (\mu_r, \mu_s)$로 분할하고, 事前, 事後의 공분산 행열과 그 역행렬들 다음과 같이 분할하자. 즉

$$V' = \begin{pmatrix} V'_{rr} & V'_{rs} \\ V'_{sr} & V'_{ss} \end{pmatrix} \qquad N' = (V')^{-1} = \begin{pmatrix} N'_{rr} & N'_{rs} \\ N'_{sr} & N'_{ss} \end{pmatrix}$$

또한

$$V_{rr}^{(r|s)} = (N'_{rr})^{-1} = [v_{ij}^{(r|s)}] \quad i, j = 1, 2, \cdots, r$$



$$B_{rs}^{(r|s)} = (N'_{rr})^{-1} N'_{rs} = [b_{ij}^{(r|s)}],\ i = 1, \cdots, r;\ j = r+1, \cdots, k$$

그리고

$$N_{ss}^{(s)} = N'_{ss} - N'_{sr}(N'_{rr})^{-1} N'_{rs} = [n_{ij}^{(s)}],\ i, j = r+1, \cdots, k$$

만일 $r = 0$ 이면,

$$B_{rs}^{(r|s)} = 0 \quad \text{그리고} \quad N_{ss}^{(s)} = N_{ss} = N'$$

따라서,

$$d_j = \sigma_j \sqrt{c_j}, \qquad j = 1, \cdots, k$$

라 하고, 또

$$\varphi_j^{(r)} = \pi_j - \sum \pi_i b_{ij}^{(r|s)}, \quad j = r+1, \cdots, k$$

라 하자. $s_j, j = r+1, \cdots, k$는 양 또는 음인 1인 수이다.

整理. 영과 $k-1$ 사이에 임의의 정수 r에 대하여, 1 또는 -1인 $s_{r+1}, \cdots, s_k$와 모든 $C > 0$에 대하여 다음 식이 만족한다고 하자.

1) $r^{(r)} = \sum_{j=r+1}^{k} s_j d_j \varphi_j^{(r)} \neq 0$

2) $\lambda_1^{(r)} \leq 0,\ 1 \leq r$

단, $\phi^{(r)} = r^{(r)} / (C + \sum\sum s_i d_i s_j d_j n_{ij}^{(s)})$

그리고

$$v_l^{(r)} = s_l [\varphi_l^{(r)} / \phi^{(r)} - \sum s_i d_i n_{il}^{(s)}] \geq 0,\ l > r$$

그러면 모든 그러한 C'에 대하여 最適標本의 크기는

$$n_l^0 = 0, \quad l \leq r,$$

그리고

$$n_l^0 = \frac{s_l \sigma \varphi_l^{(r)}}{\sqrt{c_l}} \left[\frac{C + \sum\sum s_i d_i s_j d_j n_{ij}^{(s)}}{r^{(r)}} \right]$$

$$-\frac{S_\ell \sigma_\ell}{\sqrt{c_\ell}} \sum s_i d_i n_{i\ell}, \quad \ell > r$$

실제 資料를 가지고, 最適標本의 크기를 구해 보자.

$k = 4$ 그리고 양-정치 사전 정보행렬을

$$N' = (V')^{-1} = \begin{pmatrix} 0.03 & 0.02 & -0.01 & 0.05 \\ 0.02 & 0.04 & 0.03 & 0.06 \\ -0.01 & 0.03 & 0.10 & 0.01 \\ 0.05 & 0.06 & 0.01 & 0.20 \end{pmatrix}$$

이라고 하자.

또

$$\pi = (0.02, 0.08, 0.20, 0.70$$

$$c = (1, 25, 4, 9)$$

그리고

$$\sigma_1 = 200, \quad \sigma_2 = 40, \quad \sigma_3 = 50, \quad \text{그리고} \quad \sigma_4 = 100$$

따라서

$$d_1 = 200, \quad d_2 = 200, \quad d_3 = 100 \quad \text{그리고} \quad d_4 = 300$$

C값이 변함에 따라 n_1^0와 n_2^0의 값이 어떻게 움직이나 살펴보자.

표 1의 계산과정을 설명하기 위하여, 다음을 주목하라

$$\sum\sum d_i d_j n_{ij} = 38,000, \quad \sum \pi_i d_i = 250, \quad \max_\ell \frac{1}{\pi_\ell} \sum d_i n_{i\ell} = 1200$$

그리고 $C_0 = 1200(250) - 38000 = 262000$. 어떤 구간 $c_1 \leq C \leq 262000$에 대하여, $r = 1$, $s_2 = s_3 = s_4 = +1$ 인 경우에 최적 n_i가 주어진다. c_1을 구하기 위하여 n_i^0, $v_{11}^{(1/3)}$, $B_{1,3}^{(113)}$ 그리고 $N_{2,3}^{(3)}$ 의 값이 필요하다. 공식에 의하여

$$V_{11} = v_{11} = \frac{1}{0.03}$$

$$B_{1,3} = [b_{1,2},\ b_{1,3},\ b_{1,4}] = \frac{1}{0.03}[0.02\ -0.01\ 0.05]$$

$$= [\frac{2}{3} \quad -\frac{1}{3} \quad \frac{5}{3}]$$

그리고

$$N_{3,3} = \begin{bmatrix} 0.04 & 0.03 & 0.06 \\ 0.03 & 0.10 & 0.01 \\ 0.06 & 0.01 & 0.20 \end{bmatrix} - \frac{1}{3}\begin{bmatrix} 0.04 & -0.02 & 0.10 \\ -0.02 & 0.01 & 0.05 \\ 0.10 & -0.05 & 0.25 \end{bmatrix}$$

$$= \frac{1}{3}\begin{bmatrix} 0.08 & 0.11 & 0.08 \\ 0.11 & 0.29 & 0.08 \\ 0.08 & 0.08 & 0.35 \end{bmatrix}$$

따라서, $r^{(1)} = 234$ 그리고

$$\phi^{(1)} = \frac{234}{C + 18000}$$

다음을 쉽게 증명할 수 있다.

$121600 \leq C \leq 262000$ 이면, $\lambda_1 \leq 0$

$C \geq 40870$ 이면, $\nu_2 \geq 0$

$C \geq 950645$ 이면, $\nu_3 \geq 0$

그리고

$C \geq -3707$ 이면, $\nu_4 \geq 0$

이와 같이 $121600 \leq c \leq 262000$에 대하여 앞의 정리의 모든 조건이 적합하다. 그러나 $\ell = 1$에 대하여 $\lambda_\ell \leq 0$인 조건이 성립하지 않는다.

표. C > 0 에 대한 최적의 표본크기

C	n_1^0	n_2^0	n_3^0	n_4^0
$0<C<343.71$	0	C / 25	0	0
$343.71<C<745.46$	.434472 C - 149.33	.022621C+5.97	0	0
$121,600\leq C\leq 262,000$	0	.002279C-93.15	0.022080C-209	0.094967C+352.04

References

(1) O.P.Aggarwal, Bayes and minimax procedures in sampling from finite and infinite population I, Ann, Math Statist.30(1959)

(2) _____, Bayes and minimax procedures for estimating the arithmetic mean of a population with two-stage sampling,Ann,Math,Statist,37 (1966)

(3) J.O. Berger, statistical decision theory and Bayesian analyeis 2nd edition, Springer-Verlag,(1985)

(4) N.R.Draper and I.Guttman,Some Bayesian stratified two-phase sampling results,Biometrika (1968)

(5) W.A. Ericson, Optimum stratified sampling using prior information,J.Am. statist.Ass.60 (1965)

(6) V.M.Josh;, Note on a minimax design for cluster sampling, Amm.Math. Statist.39(1968)

(7) Raiffa,H.and Scblaifer,R.,Applied Statistical Decision Tueory, Boston,Division of Researcb, Harvard Business School,1961