ある層化法に就て

林　知己夫
丸山文行

§1 我々が Sampling Survey を行ふ時任意に population を層化して効果を上げる事を通常行つてゐる。此の効果と云ふ意味は我々の推定しようとしてゐる Universe の唯一種類のある数量化された性質の平均値の Variance を小さくすると云ふ事である。

従つて実態調査の様な立体的な実情を浮び上らせる様な調査に於ては（例へば国民所得の推移，構造の把握等 ——— 構造まで立入つて将来を予測しようとする有機的推定を，ミクロ的観点を通じてマクロ的関係を予測する計量社会学の問題は，益々重要となるであらう。——— ）唯一種類丈を鋭く推定する様な効果的 Sampling 計画は適当でない。

又益々重大になるであらう社会調査では Questionnaire をなる丈少くも調査を有効にする為には各問題はなる丈相関の小さいものにし（問題の重複をさけ多様性があり他から一方が推測しにくい又は不可能な様なもの）なければならないから唯一の標識に対して良い Sampling 他の標識に対しては全く悪くなつてしまふ場合が多いのである事を考へれば益々此の感を深くする。

又日本の現状では，費用の問題で同時に多数の項目を調べ

たい為に米国に行はれてゐると思はれる鋭い手先の利いた，sampling方法は用ひられない場合が多い。

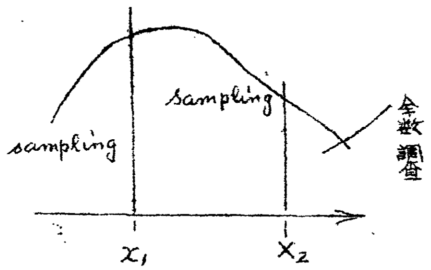
又別方面から考へれば社会統計をとるに当って直接欲する項目を調査する事が個人の権利を侵害する場合もあり又人の隠蔽したいと思ふ項目を調査する時虚偽の結果を得る場合もあり我々としては将来間接的に目的の特性を調査しなければならぬ様な事が増大して来ると考へられ，間接調査法を十分研究する事が必要になって来るものと思はれる。此の際先づ考へられるのは着目する特性を裏に包む所の多面的なDataの調査であらう。（此等Dataの綜合によって目的の特性（X）推定し得る。此の様な場合 $X=f(Y\ Z\ W\cdots)$ —— $Y, Z, W\cdots$ は調査し得る数量 —— なる時 f の形に応じて X のVarianceを小にする如きSampling方法を考へ得る場合もあり得ようがそうなし得ぬ場合もあらう）此の様にして同時に多くの物を調査する事が益々重大になって来る。この為我々としては或main pointに着目し他の推定しようとする項目を睨み合せ巧妙にUniverseを層化してなるべくSampleの数を少くする事を考へねばならない。従って細心に又ひろく考へて層化する事が必要でこの為Universeの実存的意義の把握，Universeに関する種々の基本資料の蒐集及其等資料の関係の明確化，而して其等の綜合を徹底的にする事が肝要になって来る。

此の様な立場は得られたSamplingの結果の統計的内的分析に対して有力な味方となるであらう。

§2.

今此の処では§1の様な一般論をはなれて極く特殊なmodel的Universeの通常のSampling法に対する層化法を考へて将来の高次のSamplingに対する一つの足場としたいと思ふ。

今迄，通常ある量を推定しようとする時其れの分布，或は，それと相関の高い量の分布が知られてゐるならば ── 予備知識としての分布；X年の分布を知ってゐる時 X＋1 年の平均値を推定しようとする問題にあらはれて来る。又全数調査の中の一部について特にある量をくわしく調査し全体を推測しよようとする時それと相関の高い量の分布を一部抽出により決定し，その分布からみて適当に全数から抽出して詳しい調査を行ふ様な問題にもあらはれて来る。──────



ある点 x_1，x_2 等で區切り x_2 以上の所は全数調査他は sampling 等のことを通常行ってゐるが x_1，x_2 等々を如何に定めるか又何個に定めるか（何個に Universe を區切るか）についての根據は未だ究明せられてゐない様であり sampling expert の勘に委ねられてゐる様に思はれる。

何故に推定しようとする量の大であった所（又は推定しようとする量と正の相関の高い量の大である所）のものを全数調査するのであらうか．

今，若し我々がある一つの量の Universe に於ける全量を推定しようとする時 Neyman の best linear unbiased estimate に於ける optimum allocation の考に従ふものとすれば抽出すべき sample の總数 n_0 を分割された各層へ

$$\frac{M_i S_i}{\sum M_i S_i}；$$

M_i は i 層の Universe の總数

S_i^2 は i 層の Universe の variance；

に応じて　$n_0 \dfrac{M_i S_i}{\sum M_i S_i} = n_i$
比例の

式に振り分ける事になるのである。従つて Variance が大なる層に於ては Universe の總数より大なる Sample の数を振り当てねばならぬ場合が起り得る。此は Neyman の allocation を機械的に適用した為であつて我々としては唯全数を調査すればよいのである。Neyman の allocation には

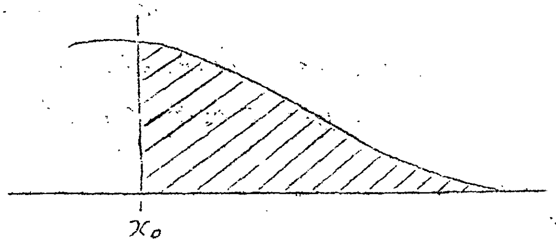
$$M_i \geqq n_i \geqq 0$$

（n_i は i 番目の層よりとる sample の数とする）

なる條件が自ら要求されてゐる事を忘れてはならない。

此の様な場合は通常の実際問題では先づおこらぬのであるが我々がある觀点に立つて Universe を層別しようとする時には必ず忘れてはならない條件である。

兎に角 Neyman 流に考へて行く時前述の様に推定しようとするのは言ひかへれば或る定められた量 x_0 以上のものを全数調査しようとするのは其処に於ける Variance の即ち斜線の部分の Variance が大であると云ふ事を暗に想定してゐる事に帰して了ふ事が出來る。



分布密度曲線

此が実際問題に対してよい層別であるか否かは十分個々の問題の内的性質に応じて検討せねばならぬ所であり唯無批判に大なる量を示すところの group は Variance 大であるとするのは考慮の余地があらう。

此の Neyman 流の立場に立って考へてゆくとき Universe の性質が次にのべる様な特殊なものである場合には前述の x_0 を稍々合理的に決定する事が出来よう。

Universe の総数は M 個である。

そして Universe の推定しようとする量に深い関係のある量の分布函数（前述の意味での —— X＋1 年の値を推定しようとする時その値の X 年の分布函数を用ひて層別する。又推定しようとする量と高い正の linear な相関をもつ量の分布函数を用ひて層別する等の時に用ひる分布函数の事をいふ ——）を $\Phi(x)$ とする。此は近似的に言って微分可能であり，即ち密度函数をもち此れも亦連続であるとしよう。x は推定しようとする量に関係の深い（上述の意味での）量とする。

即ち $y = k_1 x + k_2 + Z$ が推定しようとする量に近い（Z は確率変数，k_1，k_2 はある常数）と假定する。

此の時

$$M\int_{x_0}^{\infty} d\Phi(x) \geqq n_0 \frac{\sqrt{\int_{x_0}^{\infty} d\Phi(x)\cdot\int_{x_0}^{\infty}(x-b)^2 d\Phi(x)}}{\sqrt{\int_{x_0}^{\infty} d\Phi(x)\cdot\int_{-\infty}^{x_0}(x-a)^2 d\Phi(x)}+\sqrt{\int_{x_0}^{\infty} d\Phi(x)\int_{x_0}^{\infty}(x-b)^2 d\Phi(x)}}$$

但し

$$a = \int_{-\infty}^{x_0} x\, d\Phi(x) \Big/ \int_{-\infty}^{x_0} d\Phi(x)$$

$$b = \int_{x_0}^{\infty} x\, d\Phi(x) \Big/ \int_{x_0}^{\infty} d\Phi(x)$$

M は Universe の総数

n_0 は Sample の数を満足し右辺が最大になる様な x_0 を定め x_0 以上を全数調査すればよい事にならう。

［註］

$$y = a_n x^n + a_{n-1} x^{n-1} + \cdots\cdots + a_0$$

の様な時，$x^\ell y = C$ の様なときも適当に変数変換を行ひその変換された量について linear な関係となしその分布函数を求め同様に論を進めることは明かである。

§3. Neyman の Optimum allocation は一つの量の推定によいが，他の量の推定に対してはあまりよい allocation ではないので Universe の総数に representative にとる。即ち i Strata Sample

$$n_i = n_0 \frac{M_i}{M}$$

にとるとする。

我々の場合 $\Phi(x)$ は前述の通りとし，此れを L 個の strata に分つときいかなる區分法が Universe のある一つの推定しようとする総量 X ($X = \sum M_i \bar{x}_i$　$\bar{x}_i$ は i strata の平均) の Sampling による Variance を minimum にするのであらうかを考えてみよう。

Representative にとるときの sampling の variance は，

$$\sigma^2 \fallingdotseq \text{Const.} \sum_i^L M_i S_i^2$$

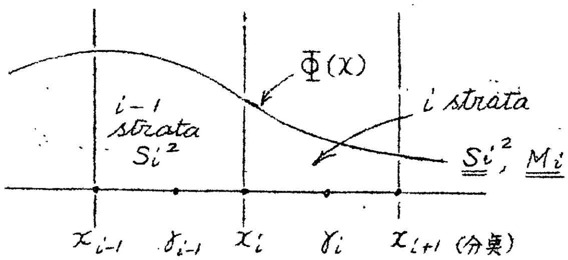
S_i^2 は，i strata の variance
M_i は，i strata の population の数とする。

今 i strata の mean を γ_i とし全体の variance, mean と夫々

$$\left.\begin{aligned} S^2 &= \left(\int_{-\infty}^{\infty} (x-\gamma)^2 d\,\Phi(x) \right) \\ \gamma &= \left(\int_{-\infty}^{\infty} x\, d\,\Phi(x) \right) \end{aligned}\right\}$$

とする時明らかに

$$MS^2 = \sum M_i S_i^2 + \sum M_i (\gamma - \gamma_i)^2 \quad となる。$$



したがって

$$\sigma^2 = Const\left(MS^2 - \sum_{i=0}^{L-1} M_i (\gamma - \gamma_i)^2 \right) \quad となる$$

我々はL個の strata に分つとき分点 x_1 ----- x_{L-1} を如何に定めたら σ^2 を最小にしうるか，即ち

$$\sum_{i=0}^{L-1} M_i (\gamma - \gamma_i)^2 \quad を$$

最大にしうるかを考へねばならない。

又

$$\sum_{i=0}^{L-1} M_i (\gamma - \gamma_i)^2 = M\gamma^2 - 2\gamma^2 M + \sum_{i=0}^{L-1} \gamma_i^2 M_i \quad で$$

あるから結局 $f = \sum_{i=0}^{L-1} \gamma_i^2 M_i$ を maximum にすればよい事になる。

さて、
$$\gamma_i = \frac{\int_{x_i}^{x_{i+1}} x\, d\Phi(x)}{\Phi(x_{i+1}) - \Phi(x_i)}$$
であるから

$$f(x_1 \cdots\cdots x_{L-1}) = \sum_{i=0}^{L-1} \frac{\left(\int_{x_i}^{x_{i+1}} x\, d\Phi(x)\right)^2}{\Phi(x_{i+1}) - \Phi(x_i)}$$

となる。 但し $x_0 = -\infty$ とする
$x_L = \infty$

註. x_0 は有限の P なる pt
x_L は有限の Q なる pt であって 6 同様である。

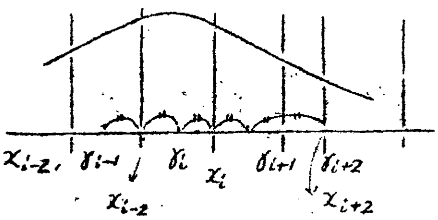
此を $x_1, \cdots\cdots, x_{L-1}$ について maximum にする事を考へればよい。 従って $\dfrac{\partial f}{\partial x_i} = 0 \quad (i = 1 \cdots\cdots L-1)$ を計算して書きおろせば

$$\frac{\partial f}{\partial x_i} = (\gamma_i - \gamma_{i+1})\left\{2x_i - (\gamma_i + \gamma_{i+1})\right\} = 0$$

となる。

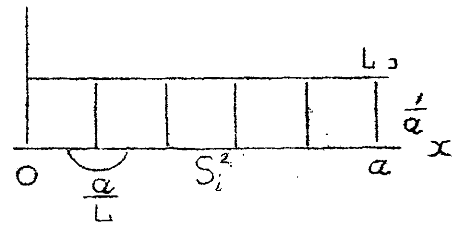
此から $x_i = \frac{1}{2}(\gamma_i + \gamma_{i+1}) \qquad i = 1, \cdots\cdots, L-1$

を得る。 此等のものが唯一の解を與へたらば ($\Phi(x)$ は、近似的に言つて連続な密度函数をもつ故に) 此の様な x_i の組が、f を maximum にすることは明かである。即ち、分点が其を相隣れる層の平均値の中点になる如く分割すると言ふ事になるのである。



§4. 以上は strata の個数Lを一定と考へたのであるが，今度は個数Lを Variable と考へ（§3にのべた分割法をとるとき）此の個数によつて variance が如何に変化するかを考へよう。

今総数 N sample の数を n strata の数をLとし population は一様分布と考へてみよう。



sampling によるxなる量の総平均の variance は

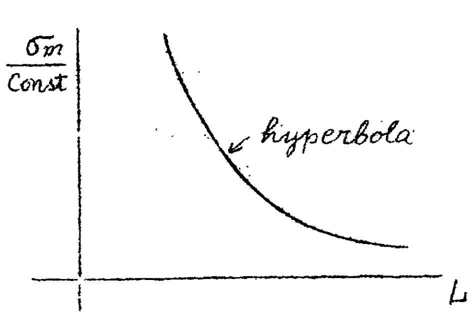
$$\sigma_m^2 = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{L}\sum S_i^2$$ となる

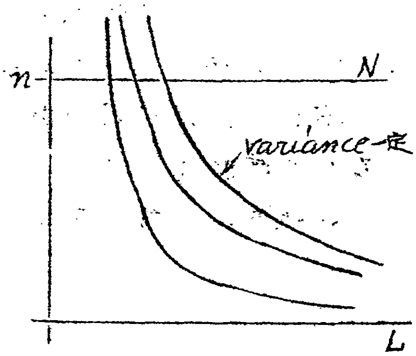
S_i^2 は i strata の variance である。

此から

$$\sigma_m^2 = Const.\left(\frac{1}{n} - \frac{1}{N}\right)\cdot\frac{a^2}{L^2}$$ を得る

即ち variance 一定と考へれば strata の数と sample 数とは略ゝ一乗と二乗との関係にある。





strata を2倍にすれば n 一定で variance は $\frac{1}{4}$ となるこ

とが言へる。

次に

$$\frac{1}{\sqrt{2\pi}\,\sigma^2} e^{-\frac{x^2}{2\sigma^2}} \text{ で specify される normal}$$

の場合を考へる。此の時の §3 でのべた区分法を得るには *successive* にやらねばならなかった。その結果は第二表の通りである。

l は *strata* の数

x_i は 分点

φ_i は 分点に於ける $\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{x_i^2}{2\sigma^2}}$ の値

Φ_i は $-\infty$ から分点までの確率をあらはす

即ち

$$\Phi_i(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{x_i} e^{-\frac{y^2}{2\sigma^2}} dy$$

をあらはす。

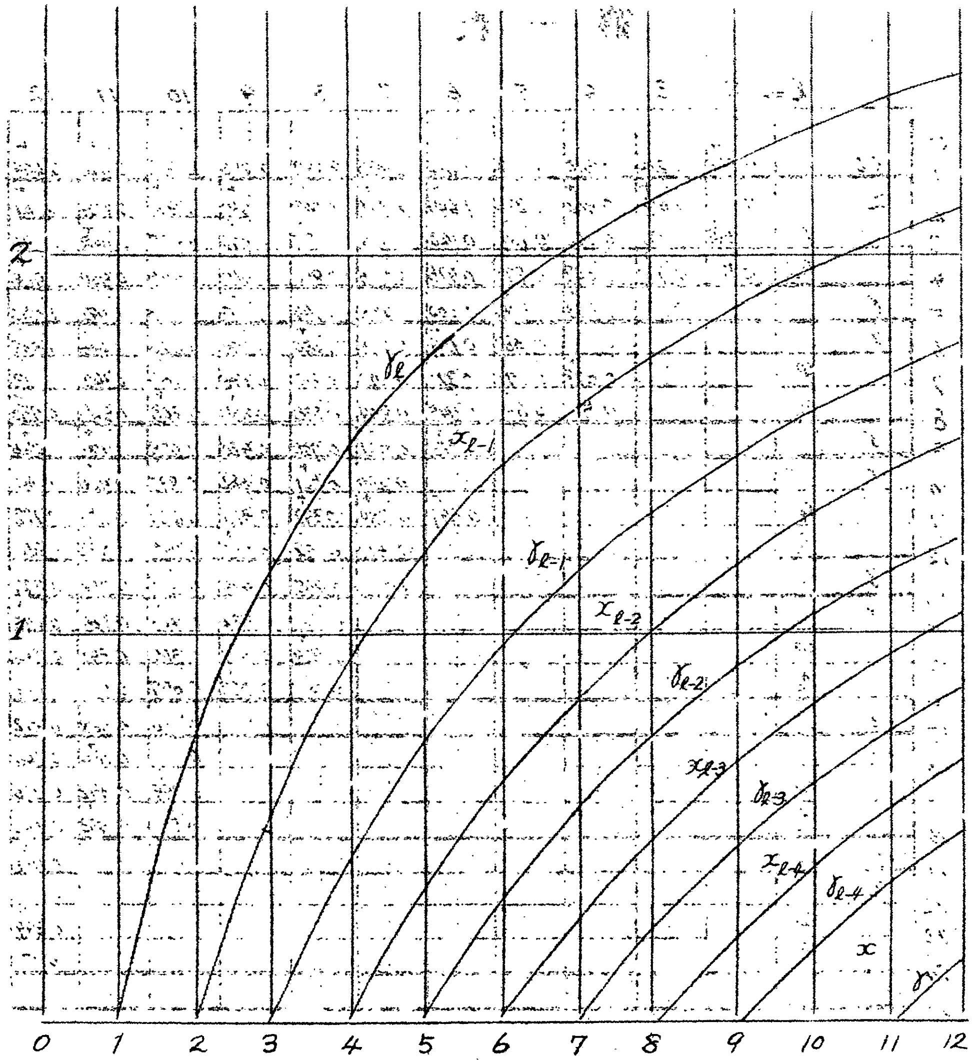
第二表は如上の関係を図示したものであり横軸 $l=i$ の所を縦にみてゆけば分点及び各々の *strata* 内の平均値を得る事が出来る。かくして n 一定として *strata* の数により、*sampling* の *variance* $\sigma^2 = \text{Const.} \sum M_i C_i^2$ が如何に減じてゆくかは、第三表に示した。

比較のため $\frac{\sigma^2}{\text{const.}} \cdot l = \text{Const.}$ なる *hyperbolic* 的のものを併せ図示した。

以上の様に考へうるならば *sample* すべき総数 n が一定であつても l を大にすればよい精度が得られるし又精度を一定とすれば l のつくり方によつて *sample* すべき数を小となしうる一応の目安が出来るものと言へよう。

第一表

$l=$		2	3	4	5	6	7	8	9	10	11	12
γ_i		0.998	1.224	1.510	1.724	1.894	2.033	2.152	2.256	2.346	2.426	2.499
x_i		0	0.612	0.982	1.244	1.447	1.611	1.748	1.867	1.970	2.059	2.141
	φ_i	0.399	0.331	0.246	0.184	0.140	0.109	0.087	0.070	0.057	0.048	0.040
	Φ_i	0.5	0.730	0.837	0.893	0.926	0.948	0.960	0.969	0.976	0.980	0.984
γ			0	0.453	0.765	1.000	1.188	1.344	1.478	1.593	1.692	1.783
x				0	0.382	0.659	0.874	1.050	1.199	1.326	1.436	1.535
	φ			0.399	0.371	0.321	0.272	0.230	0.194	0.166	0.142	0.123
	Φ			0.5	0.649	0.745	0.809	0.853	0.885	0.908	0.924	0.938
γ					0	0.318	0.561	0.756	0.920	1.059	1.178	1.286
x						0	0.280	0.501	0.682	0.835	0.966	1.081
	φ					0.399	0.384	0.352	0.316	0.281	0.250	0.222
	Φ					0.5	0.610	0.692	0.753	0.798	0.833	0.860
γ							0	0.245	0.444	0.611	0.752	0.877
x								0	0.222	0.406	0.560	0.695
	φ							0.399	0.389	0.367	0.341	0.313
	Φ							0.5	0.588	0.657	0.71[illegible]	0.756
γ									0	0.200	0.368	0.512
x										0	0.184	0.340
	φ									0.399	0.392	0.376
	Φ									0.5	0.573	0.633
γ											0	0.169
x												0
	φ											0.399
	Φ											0.5


2
1
γ_ℓ
$x_{\ell-1}$
$\gamma_{\ell-1}$
$x_{\ell-2}$
$\gamma_{\ell-2}$
$x_{\ell-3}$
$\gamma_{\ell-3}$
$x_{\ell-4}$
$\gamma_{\ell-4}$
x
γ
0 1 2 3 4 5 6 7 8 9 10 11 12

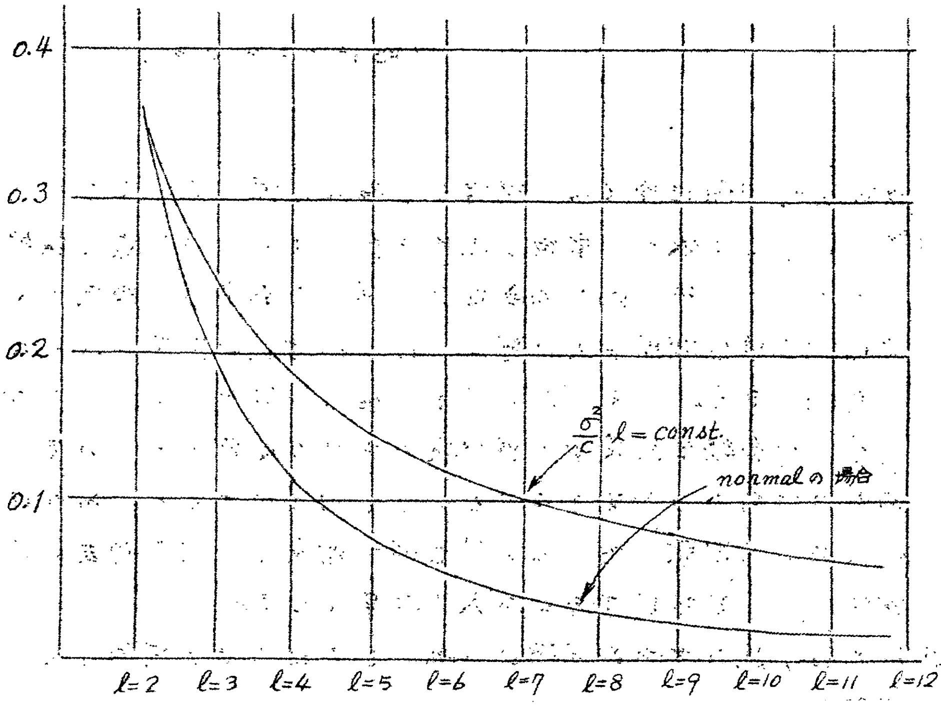
第 二 表

Effect of Stratification
in Normal Population

$\frac{\sigma^2}{c}$: Variance × const.

l : Number of Strata

$$\frac{\sigma^2}{c} = \Sigma \left\{ (x_i \varphi_i - x_{i+1} \varphi_{i+1}) + (1-\gamma_i^2)\left[\Phi(x_{i+1}) - \Phi(x_i)\right] \right\}$$



第 三 表