白河市言語調査に於ける
サムプリング・調査計画

林　知己夫
石田　正次

此の研究は科学試験研究費によるものの一部である。国立国語研究所・民族学研究所との協同研究である。国研の中に，特別調査班（委員会）が設置せられ，国語研究所の所員，民族学研究所の研究員，統計数理研究所の所員によつて構成せられた。以下はこの委員会の中で主として統計数理研究所の研究員が行った成果の一部である。此の研究には外に丸山文行，西平重喜，村岡充子，田熊雅子が参加し，別の部分を担当した。

§1. 調査対象の決定と母集団の構成

白河市に居住するものの言語の実態を把握するのが調査の目的であるから調査対象は一先づ昭和２４年１０月に於ける白河市在住者と言ふことになる。

なお社会生活の調査実施の面を考慮し我々の場合は対象をさらに限定し白河市に居住する１５才から６９才迄の男女と定めた。　この様な人々の織りなす言語生活の Pattern を我々は調査することになるのである。　然しこの調査では Pattern のすべてを見極めようとするものではなく一定の基準によって作られた調査票（個人の経歴，社会環境

＊ 白河市とは旧白河市を指すものとする。

及び言語関係の調査項目 ―― 主として共通語化の程度をみるためのもの ―― ）に基きその項目にもられているものについて丈知識を得ようとするのである。 かう考えてくると我々の場合の母集団は次の様に定義するのが妥当であろうと思はれる。

「調査票にもられている調査事項に対する反応を標識とし且つ各個人の抽出確率が等しいと定められた、昭和２４年１０月白河市に居住する数え年１５才から６９才迄の男女」

我々の場合は調査は性格上当然面接によらねばならず且つ又微妙な内容をもつ関係上調査員は一定の素養を持ちさらにある程度の訓練を必要とするものであつて，為に適した調査員をさう多く得ることは出來ない。

その他調査にまつわる外的制約（予算，日時，等）を考え併せるとき２万に垂んとする母集団のすべてを抽出調査することはまづ不可能と言わざるを得ない。 ここに当然サムプル調査が行われることになるであろう。 我々の場合は云うまでもなく調査の精度（サムプルから母集団を推定する時の）の科学的保證のあるランダムサムプリングの方法によるのである。

§ ２　サムプル数の決定　その第一次的意味調査の実施は純理では行うことはできないのは云ふまでもないが，一定の制約の下で最適な結果を得る様理論的に考えてゆかねばならない。

此の調査では調査員は２２名程度（東京より出向くもの約１５人，現地で得られるもの約７人）と考えられ予算の関係で約４日位に最後の調査を完了しなければならない。 調査票の内容よりみて被調査者１名の調査時間は平均約１時間* 従つて調査員１名は１日に７人の調査を行えるものとみてよい。

※　調査時間は長きに亘るとき調査員にも被調査者にも苦痛を與え調査内容は不確かなものとなる虞れがある。したがって調査時間が長くとも１時間程度に終る様調査票はつくられてある。

こう考えると調査は４日間に約７００人の調査が行えることになる。しかし，此は全く順調に調査が進む場合の事であり，実際の場合こうはあり得ない。訪問する被調査者は常に在宅するとはかぎらず，しかもあくまで此を追究調査しなければならない。職場へ行き或は又夜間訪問をする場合もあるであろう。しかも調査員に無理を強いてはならない。（無理を強いるとき調査内容は不確かなものとなる）ことを思ふときサムプル数は高々５００程度が限度であろうと思われる。それでは５０.０のランダムサムプルで調査の精度は如何であろうか。

調査事項の反応が百分比率であらわされるものとする。サムプルの百分比率から母集団の百分比率を推定する時の精度は近似的に，

$$\text{信頼度 } 95\% \text{ で} \qquad \pm\ 1.96\sqrt{\frac{N-n}{N-1}\,\frac{p'(1-p')}{n}}$$

によってあたえられる。但し N は母集団の大さ我々の場合 17890，n はサムプルの大さ，我々の場合５００，p' はサムプルに於ける調査事項のある反応を示す比率である。

見透をよくするために信頼巾を示す．

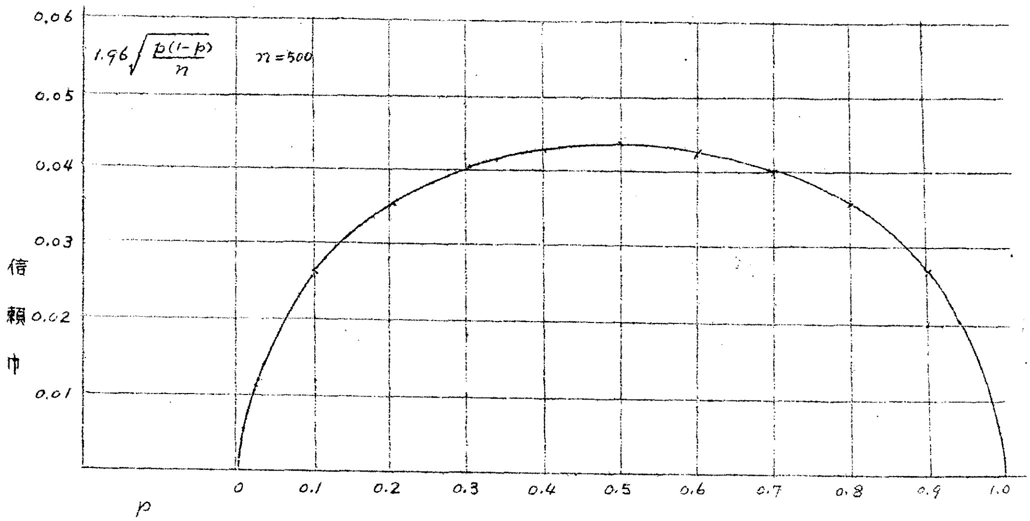
$$1.96\sqrt{\frac{N-n}{N-1}\,\frac{p(1-p)}{n}}$$

と p との関係を図示してみよう。こゝに p は母集団の比率を表すものとする。

我々の場合 N が n に比して十分大であるから

$$1.96\sqrt{\frac{p(1-p)}{n}}$$

を用いた。



此から直にわかる様に５００のサムプルがあればまず満足すべき結果が得られるであろう。

然し我々は更に精度をあげるために層化を行う事を考えよう。

§ 3. 層化とサンプルの第一次抽出

我々の場合調査事項の反応形体に強い影響を與えると考えられるものとしては性，年令、学厂，職業，居住厂，家庭環境等々である。

したがつて母集団を此等の標識を用いて層化するのがよい。然し層化の客観的資料として容易に用い得られるものは，性，年令，職業，現住所である。　従つて此等を用いて層化を行はうと思う。

このためにはまづ母集団を表すリストがなくてはならない。　此のリストとして最も適当なものは物資配給台帳である。　然し此には今の場合対象外のものをも含むであるし，一世帯が同一頁に記載されてあり一人一人別になつて居らない。　従つて此を層化するためには対象たる１７８９０人のカードを新につくり此を層化する手続をとらねばならない。　此の事は事実上*不可能である。それは如何にすればよいか。

重ね抜きの方法を用いてみる。　　　　　　　　註次頁

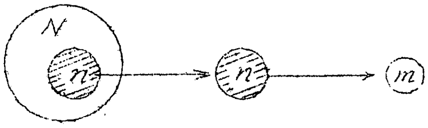
つまり第1次的に母集団からある人数大ランダムに抽出，しかる後慾する500人のサムプルをランダムに抽出する方法（此の時層化をほどこして）を採用することにする。この様な方法をとるとき欲するサムプルを得られることを證明しながら以下サムプルの抽出を述べてみよう。

（i）母集団の大さを N とする。その標識は $X_1, X_2, \cdots\cdots X_N$ であらわされ，又母集団分散は σ^2 によつて示されるものとする。

N 個からなる母集団からまず n 個のサムプルを抽出し、次にそのサムプルから m 個のサムプルを抽出する。

この時サムプル平均 $\bar{x} = \frac{1}{m} \Sigma x_i$ の精度を示す分散 S^2 はどうなるか、なおこの $\bar{x}$ は母集団平均の偏りのない推定量になつているのは云うまでもない。

此は副次抽出であるから



$$S^2 = \frac{n-m}{n-1} \frac{\sigma_w^2}{m} + \sigma_b^2$$ となる。

但し σ_w^2 は内分散 σ_b^2 は外分散であり次の関係をもつ

$$\sigma^2 = \sigma_w^2 + \sigma_b^2$$

$$\sigma_b^2 = \frac{N-n}{N-1} \frac{\sigma^2}{n}$$

したがつて

$$S^2 = \frac{n-m}{n-1} \frac{1}{m} (\sigma^2 - \sigma_b^2) + \sigma_b^2$$

註. *日時労力の上から不可能であり又莫大な犠牲をはらい此を実行することは調査の全体的構想よりみる時意味がない。

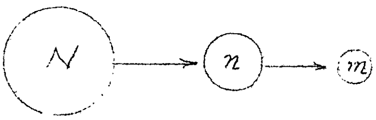
$$= \frac{n-m}{n-1}\frac{1}{m}\sigma^2 + \sigma_b^2\left(1-\frac{n-m}{n-1}\frac{1}{m}\right)$$

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

$$= \frac{N-m}{N-1}\frac{\sigma^2}{m}$$

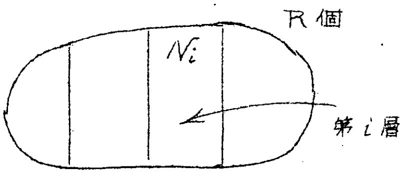
となる。 此は N 個の要素を持つ母集団から m 個のサムプルを直接ランダムに抽出した時の分散である。 即ち二度ぬきしても精度のかわらない事を示してゐる。 此処で大切なのは S^2 が第1次抽出の個数 n に全く依存しないと云う事である。

(ii)



前の場合と同様であるが，但し n が確率変数であるときどうなるか。（但し $n \geq m$ ）
先づ n を固定して S^2 をもとめてみればよいのであるが S^2 は全く n に関係しないので此の時の分散も亦母集団から直接 m 個のサムプルを抽出した時と同一となる。

(i)(ii) より次の様な考えが導かれる。 まづ母集団がR個の層に分けられたとする。



N_i は i 層の母集の大さ。今全体から n 個のサムプルをランダムに抽出すると i 層に属するサムプルを n_i だけ得る。

此の n_i は勿論確率変数であるが，n_i の待望値

$$E(n_i) = n \times \frac{N_i}{N}$$

である。 n_iから欲するサムプル数 m_i個を抽出（ $n_i \geq m_i$ ）此からサムプル平均をつくる時，その分散は n_i を無視し N_iから直接 m_iのサムプルを抽出した時の分散に等しい。

こういうわけであるから全体のサムプル数 m を定め，第 i 層の割当数 m_iをきめておき各層で上述の様なサムプリング方法を用いて（母集団全体を層化せず，全体よりランダムにとつた第1次サムプル n を層化しその中から定められた m_iのサムプルをランダムに抽出する）普通の場合の様に推定値をつくれば完全な層化抽出法の効果をあげうることになる。

云いかえれば層化抽出を行つて効果をあげるのに母集団を層化せず比較的少数の第1次サムプルを層化しその中より欲するサムプルを抽出しウエイトを用い偏りない推定値をつくればよいことになる。

層化の手続は著しく簡単になることであろう。

我々の場合，各層えのサムプルの割当は人口比例割当法によることにする。 蓋し此の割当法は複雑な分析を自由に簡単に行いうるという利点があるからである。 したがつて，

$$m_i = m p_i$$

但し $p_i = \frac{N_i}{N}$ とする。

m_iが決めきまるためには p_iがわかつておらねばならない。

白河市の場合残念ながら層化に用い様とする要因すべてについて p_iは判つておらない。

p_i は男女×年令についてのみわかつている。即ち、

年令	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	
男	9.52	6.84	5.31	4.28	4.06	4.28	3.89	3.18	2.39	1.97	1.28	
女	8.90	7.75	6.74	5.24	5.25	4.76	4.04	3.24	2.70	2.13	1.75	百分率

したがつて m_i を明確に定めるのは男女×年令について丈であるから第1次層化は男女×年令についてのみ行うことにした。

次に問題となるのは第1次サンプルたる n を幾何に定めたらよいかの問題である。 通常の層化サムプリングと同じ効果をもつサムプリングを行い得るためには常に

$$n_i \geqq m_i = m p_i$$

でなくてはならない。 此が成立する様 n を定めなければならない。

n_i は確率変数であり

$$E(n_i) = n p_i$$

である。

n_i の分散 $\sigma_{n_i}^2$ は

$$\sigma_{n_i}^2 = \frac{N-n}{N-1} \frac{p_i(1-p_i)}{n}$$

であたえられるから 99.7 % (99.9 %) の信頼を以て，

$n_i \geqq m p_i$ が成立するためには，

$$n\left(p_i - 3(4)\sqrt{\frac{N-n}{N-1}\frac{p_i(1-p_i)}{n}}\right) \geqq mp_i$$

なる様 n が定められねばならない。
我々の場合　$m = 500$　である。　一般に n と p との関係を図示してみると次の様になる。（別紙折込）

此による n としてまづ２０００程度抽出すればよい事が解るであろう。

さらに又，次の様なことも考えておこう。

$m_i = mp_i$　とするとき p_i の小さな層の m_i は小となりすぎてゞ層そのものの精度があまり落すぎ他との比較が意味をもたなくなることもあり得るので比例割当法でなく p の小な所に mp の２倍のサムプルを割当てることをするとしよう。

２倍であるから全体の結果を出す場合さう複雑でないであろう。

此が出来うるために n はどうであろうか。

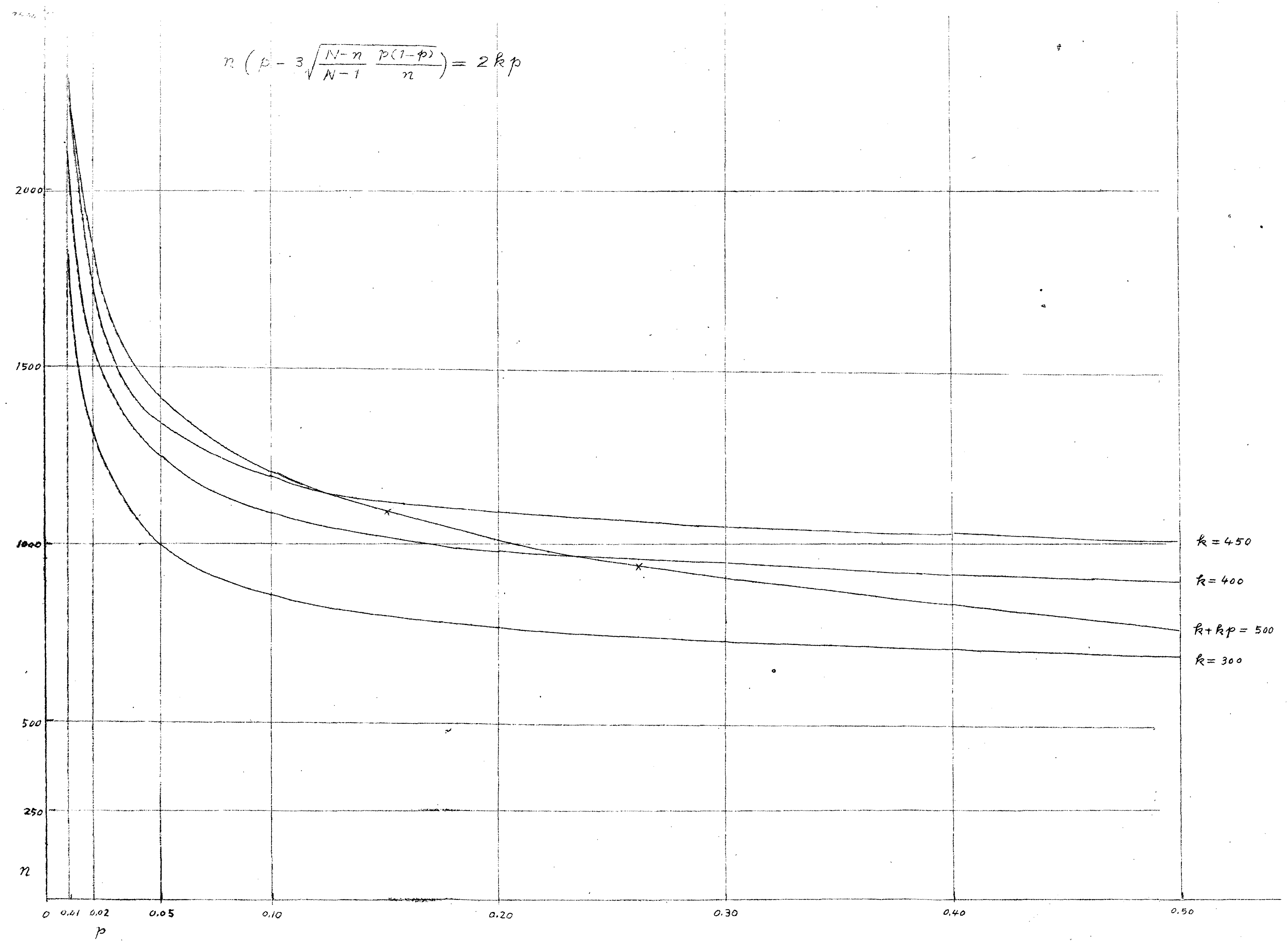
此の時 k 個のサンプルは前のとおり比例割当てするものとする。

次に，p の小なところは $2kp$ のサムプルを割当てるものとする

さうすると前と同様に

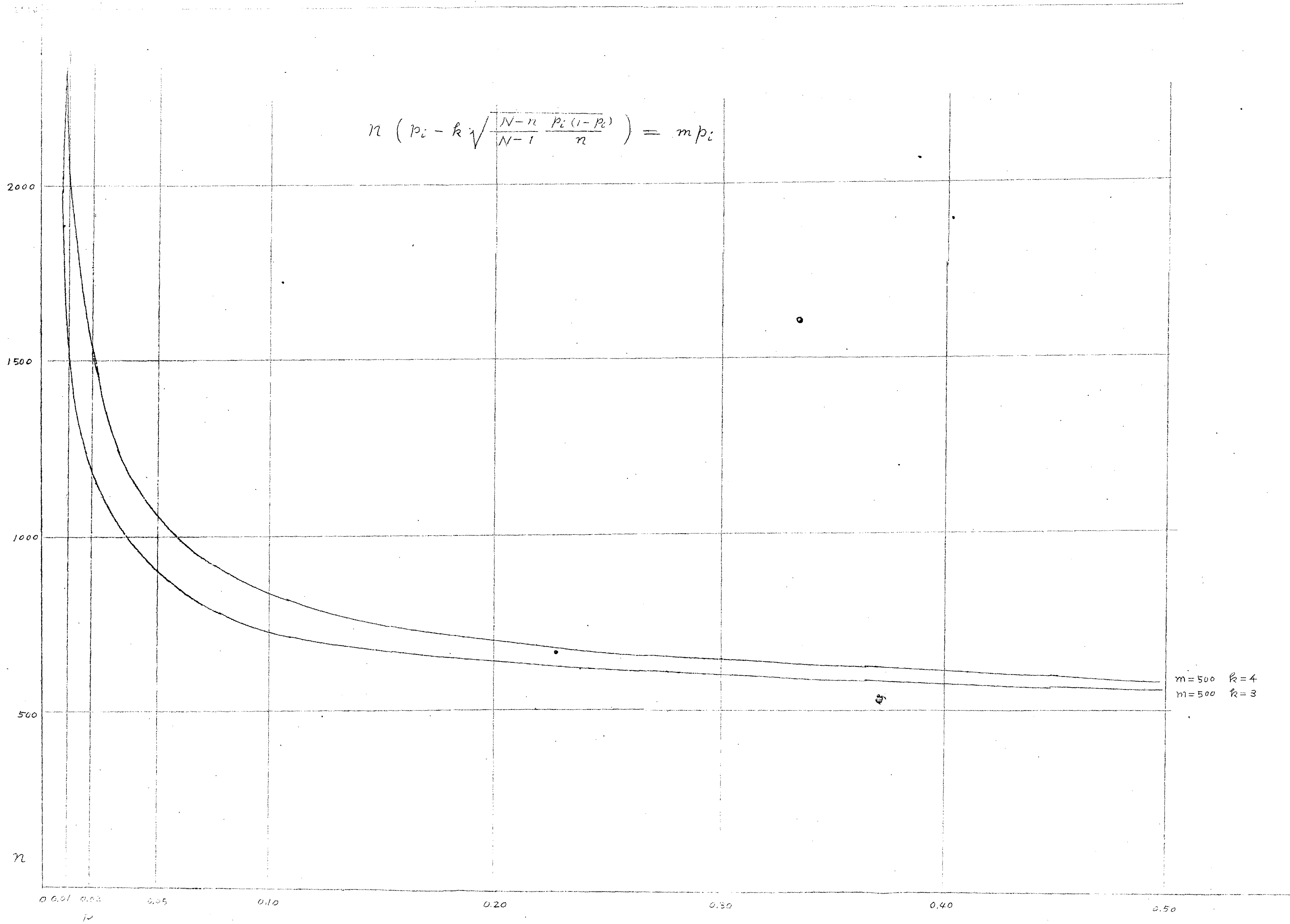
$$n\left(p - 3\sqrt{\frac{N-n}{N-1}\frac{p(1-p)}{n}}\right. \geqq 2kp$$

でなくてはならない。　k と p との関係を図示してみると次の様になる。　又　$k + kp = 500$　の関係があるとき即ち全サムプルを 500 と定めた時の関係も同時に示しておいた。



$n\left(p - 3\sqrt{\frac{N-n}{N-1}\,\frac{p(1-p)}{n}}\right) = 2kp$
2000
1500
1000
500
250
n
0
0.01
0.02
0.05
0.10
0.20
0.30
0.40
0.50
p
k = 450
k = 400
k + kp = 500
k = 300

折込、外



$n\left(p_i - k\sqrt{\frac{N-n}{N-1}\,\frac{p_i(1-p_i)}{n}}\right) = m p_i$
2000
1500
1000
500
n
0
0.01
0.02
0.05
0.10
0.20
0.30
0.40
0.50
p
m=500 k=4
m=500 k=3

以上綜合し略２０００程度の第１次サムプルを抽出すればまづ十分であろうかと思われる。

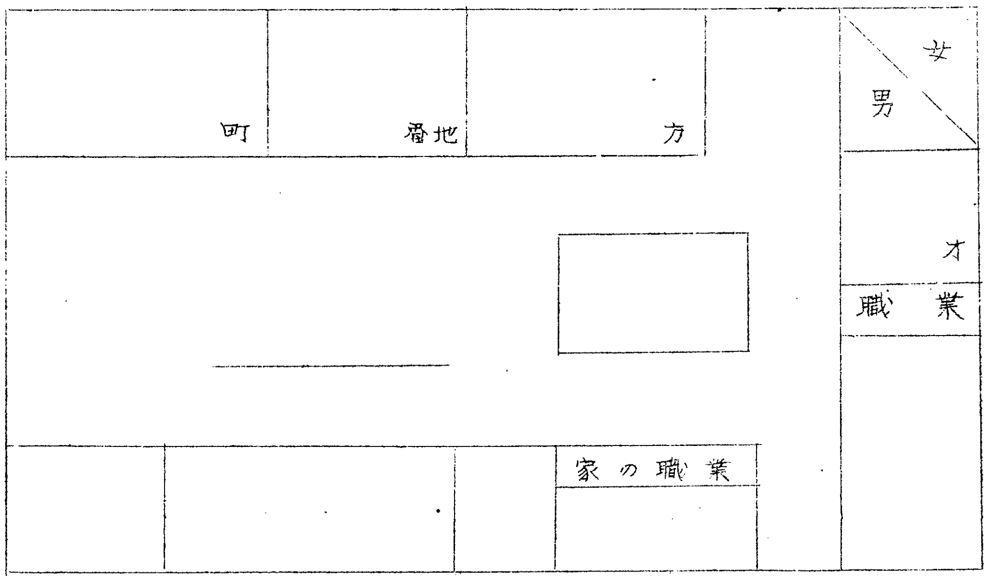
２０００のサムプルを抽出し此をカードにとり層化することは調査の実際的観点よりみて容易に行えるところである。

対象は１７８９０人である。 此より２０００人を抽出するのであるから抽出の便利及びシステマティックサムプリングの精度のよさ（よみ書き能力調査に於ける標本調査計画参照）を考えてランダム数を出発番号とし抽出間隔８を以て数えてゆく方法をとった。

蓋し此の方法は抽出速度がきわめて大きいからである。※

因みに，白河市の平均家族人数は４．９人であり抽出間隔８は此と同調していない。 此の結果サムプル数２１８４を得た。 第１次層別を示すと次の様になる。

なお被抽出者を記入するカードには次の様なものを用いた。



註．※ 対象外を除外して数える事は甚だ煩雑であり間違も起し易い

年令	男		女	
	Sample	Population	Sample	Population
15 — 19	228	208	176	194
20 — 24	132	149	205	169
25 — 29	137	116	140	147
30 — 34	86	94	107	115
35 — 39	77	89	105	115
40 — 44	85	93	99	104
45 — 49	99	85	79	88
50 — 54	79	69	81	71
55 — 59	65	63	60	59
60 — 64	37	43	45	47
65 — 69	29	28	33	38
計	1054	1037	1130	1147

こゝろみに母集団の構成と χ^2 検定を行うと

$$\chi^2 = 29.8 \qquad \text{自由度} \ 21$$

此より大なる χ^2 を得る確率は10%程度である。此の2184のサムプルは母集団からのランダムサムプルと考えられる程度のものであり強い偏倚を与えていないものと思われる。

§ 4. 第2次サムプルの抽出

§3で示した様な各層に属している第1次サムプルからいよいよ第2次サムプルを抽出することになる。さてこの様な男女×年令の層之のサムプルの割当は比例割当によるのであるから各層へのサンプル数は

$$m_i = m p_i \qquad m = 500$$

によってあたえられるのである。m_i は次の様になる。

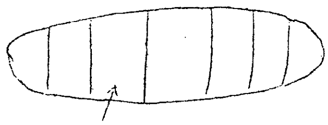
年令	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	計
男	48	34	27	21	20	21	19	16	15	10	6	237
女	45	39	34	26	26	24	19	16	14	11	9	263
計	93	73	61	47	46	45	38	32	29	21	15	500

此から見てもわかる様に明らかに $np_i \geq m_i$ は満足せられている。

各層から m_i 丈のサムプルを再びランダム数を出発番号としてシステイマテイックに抽出する。

此をなす時再び工夫をしてみよう。

各層の中を職業，現住所の性格（（町単位とし町をその性格により，商業地帯（中心部），商業地帯（非中心部），住宅地帯，農業地帯の4段階とした））によつて第2次的に層化し，此を一列に配列し此の中からシステイマテイックに所要数のサムプルを抽出することにした。 此の時は完全な層化抽出法ではない。 何となれば第2次層への割当数は母集団の比率を以てなされるのではなく第1次サムプルの比率を以てなされるからである。 かくするとき精度を示す分散はどうなるであろうか、順を追って示してみよう。



§ 5. 此のサムプリング法による精度の計算

(i) 層化抽出による分散（比例割当）

i 層の母集団の大さ N_i

i 層の 分散 σ_i

i 層の平均　$\overline{X}_i$

i 層のサムプルの大さ　m_i

母集団の分散　σ^2

〃　平均　$\overline{X}$

層の数　R

$$\sum_{i=1}^{R} N_i = N$$

$$\sum_{i=1}^{R} m_i = m$$

$$\frac{N_i}{N} = p_i$$

$$\text{又}\quad m_i = m p_i$$

とする。

此の時　$\overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$

の分散　S^2 は近似的に

$$S^2 = \frac{N-m}{N-1}\,\frac{\sigma^2}{m} - \frac{N-m}{N-1}\,\frac{1}{m}\sum p_i(\overline{X}_i - \overline{X})$$

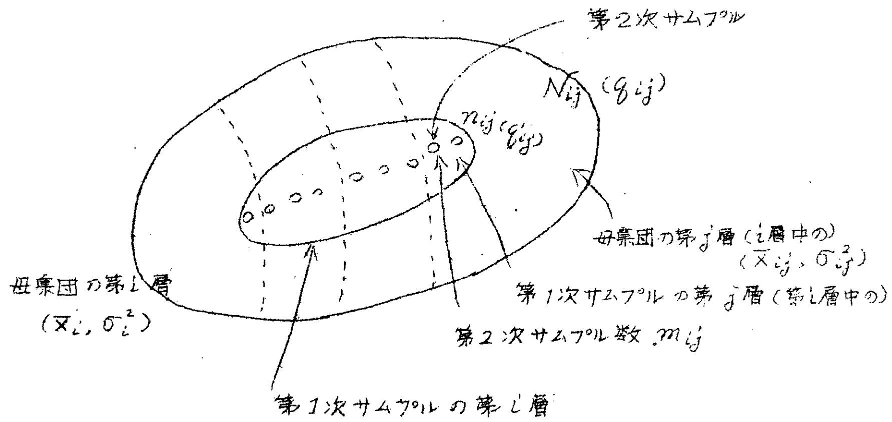
によってあたえられる。

第2項が層化の効果である。

我々の場合　男女×年令　の層化については此の様な精度を得る。但し $\overline{x}$ は推定比率（調査事項に対する反応形体を示す比率）である。重ね抜きサムプリングを行っても普通の層化抽出法と同一の分散を示す事が證明せられてあるからである。

(ii) 第1次層化による層化の効果

第2次サムプルを層化する。第1次サムプルを層化した時此の層の数を R_i とする。（今簡單のため第 i 層を固定して考える。）



i層中第j層の母集団の大さを N_{ij} とする。

$$\Sigma N_{ij} = N_i \qquad \frac{N_{ij}}{N_i} = q_{ij}$$

とする。

q_{ij} のサムプルの値（第1次サムプル）を q'_{ij} とする。

$$q'_{ij} = \frac{n_{ij}}{n_i}$$

n_i，n_{ij} は第1次のサムプルの第i層に属しているものの数，第i層中の第j層に属しているものゝ数である。

勿論 $n_i = \Sigma n_{ij}$ である。 q'_{ij} の分散は近似的に安全目に見積って

$$\frac{N_i - np_i}{N_i - 1} \frac{q_{ij}(1 - q_{ij})}{np_i} \doteqdot \frac{q_{ij}(1 - q_{ij})}{np_i}$$

である。

さて此から m 個のサムプルを抽出するのであるが層を一列に並べ此からシステイマテイツクサムプリングによって抽出するのでi層えのサムプルのサムプル割当は略 $m_i q'_{ij} = m_{ij}$ に等しくなる。

此の様なサムプルから推定値

$$\overline{x}_i = \frac{1}{m}\sum_{l=1}^{m} x_{il}$$

をつくると此の分散はどうなるであろうか。

j 層の母集団平均を $\overline{X}_{ij}$ 分散を σ_{ij}^2

全体の 〃 $\overline{X}_i$ 〃 σ_i^2

とする。

$$\overline{x}_i = \frac{1}{m}\sum_{l=1}^{m} x_{il} \fallingdotseq \sum q'_{ij}\,\overline{x}'_{ij}$$

$\overline{x}'_{ij}$ は j 層のサムプル平均である。

勿論

$$E(\overline{x}_i) = \sum q_{ij}\,\overline{X}_{ij} = \overline{X}_i \qquad \text{である。}$$

（不偏推定値）

$\overline{x}_i$ の分散を考えてみよう。こゝから煩雑をさけるため一応 i の指標をおとして考えてゆこう。

$$E(\overline{x} - \sum q_j\,\overline{X}_j)^2$$

$$= E(\sum q'_j\,\overline{x}'_j - \sum q_j\,\overline{X}_j)^2$$

$$= E(\sum q'_j\,\overline{x}'_j)^2 - (\sum q_j\,\overline{X}_j)^2$$

$$= E(\sum q'^2_j\,\overline{x}'^2_j) - (\sum q_j^2\,\overline{X}_j^2) + E(\sum_k \sum_{j\neq k} q'_k q'_j\,\overline{x}'_k\,\overline{x}'_j)$$

$$- \sum_k \sum_{j\neq k} q_k q_j\,\overline{X}_k\,\overline{X}_j$$

$$= \sum E(q'^2_j\,\overline{x}'^2_j - q_j^2\,\overline{X}_j^2)$$

$$+ E(\sum_k \sum_{j\neq k}(q'_j - q_k)(q'_j - q_j)\,\overline{x}'_k\,\overline{x}'_j)$$

$$+ \sum_k \sum_{j\neq k} q_k q_j\,\overline{X}_k\,\overline{X}_j - \sum_k \sum_{j\neq k} q_k q_j\,\overline{X}_k\,\overline{X}_j$$

$$= \sum_{j=1}^{R} E\,(q_j'\,\overline{x}_j' - q_j\,\overline{X}_j)^2$$

$$+ E\,(\sum_{k}\sum_{j\neq k}(q_k' - q_k)(q_j' - q_j)\,\overline{x}_k'\,\overline{x}_j')$$

第1項の Σ の中

$$E\,(q_j'\,\overline{x}_j' - q_j\,\overline{X}_j)^2$$

$$= E\,(q_j'\,\overline{x}_j' - q_j'\,\overline{X}_j + q_j'\,\overline{X}_j - q_j\,\overline{X}_j)^2$$

$$= E\,q_j'^2\,(\overline{x}_j' - \overline{X}_j)^2 + \overline{X}_j^2\,E\,(q_j' - q_j)^2$$

こゝで再び i の指標を入れ §3.の結果を用ふれば（此の場合と N_{ij} から抽出推定値をつくるときと分散が同一という事を用うる）

$\underbrace{\text{抽出推定値}}_{\text{直接}m_{ij}\text{を}}$

$$E\,q_{ij}'^2\,\frac{N_{ij}-m_{ij}}{N_{ij}-1}\,\frac{\sigma_{ij}^2}{m_{ij}} + \overline{X}_j^2\,\frac{N_i-np_i}{N_i-1}\,\frac{q_{ij}(1-q_{ij})}{np_i}$$

$$= E\,q_{ij}'\,\frac{N_{ij}-m_{ij}}{N_{ij}-1}\,\frac{\sigma_{ij}^2}{m_i} + \overline{X}_j^2\,\frac{N_i-np_i}{N_i-1}\,\frac{q_{ij}(1-q_{ij})}{np_i}$$

$\dfrac{N_{ij}-m_{ij}}{N_{ij}-1} \doteqdot 1$ と考え $\dfrac{N_i-m_i}{N_i-1}$ と近似すれば

$$\doteqdot \frac{N_i-m_i}{N_i-1}\,\frac{1}{m_i}\,q_{ij}\,\sigma_{ij}^2 + \overline{X}_{ij}^2\,\frac{N_i-np_i}{N_i-1}\,\frac{q_{ij}(1-q_{ij})}{np_i}$$

となる。 次に第2項について考える。 近似的に安全目にみて成立する

$$E\,(q_{ij}' - q_{ij})(q_{ik}' - q_{ik}) = -\,\frac{N_i-np_i}{N_i-1}\,\frac{q_{ij}\,q_{ik}}{np_i}$$

なる関係を用うれば，第2項は

$$-\frac{1}{np_i}\frac{N_i-np_i}{N_i-1}\sum_k\sum_{j\neq k}q_{ik}q_{ij}\overline{X}_{ik}\overline{X}_{ij}$$

$$=-\frac{1}{np_i}\frac{N_i-np_i}{N_i-1}\sum_k q_{ik}\overline{X}_{ik}\sum_{j\neq k}q_{ij}\overline{X}_{ij}$$

$$\sum_{j\neq 1}q_{ij}\overline{X}_{ij}=\overline{X}_i$$

。

したがって

$$\sum_{j\neq k}^{R}q_{ij}\overline{X}_{ij}\fallingdotseq\overline{X}_i-q_{ik}\overline{X}_{ik}$$ を用ふれば

$$=\frac{N_i-np_i}{N_i-1}\frac{1}{np_i}\left(-\overline{X}_i\sum_k q_{ik}\overline{X}_{ik}+\sum_k q_{ik}^2\overline{X}_{ik}^2\right)$$

$$=\frac{N_i-np_i}{N_i-1}\frac{1}{np_i}\left(-\overline{X}_i^2+\sum_k q_{ik}^2\overline{X}_{ik}^2\right)$$

$$=\frac{N_i-np_i}{N_i-1}\frac{1}{np_i}\left(-\overline{X}_i^2+\Sigma q_{ik}\overline{X}_{ik}^2-\Sigma q_{ik}\overline{X}_{ik}^2+\Sigma q_{ik}^2\overline{X}_{ik}^2\right)$$

$$=\frac{N_i-np_i}{N_i-1}\frac{1}{np_i}\left(\sigma_{bi}^2-\sum_k q_{ik}(1-q_{ik})\overline{X}_{ik}^2\right)$$

但し σ_{bi}^2 は i 層中の各層間の外分散（ between の Variance ）である。

以上綜合すれば， $\overline{x}_i$ の分散 S_i^2 は近似的に

$$S_i^2\fallingdotseq\sum_j\frac{N_i-m_i}{N_i-1}\frac{1}{m_i}q_{ij}\sigma_{ij}^2+\sum_j\overline{X}_{ij}^2\frac{q_{ij}(1-q_{ij})}{np_i}\frac{N_i-np_i}{N_i-1}$$

$$+\frac{N_i-np_i}{N_i-1}\frac{1}{np_i}\sigma_{bi}^2-\frac{N_i-np_i}{N_i-1}\sum_k\overline{X}_{ik}^2\frac{q_{ik}(1-q_{ik})}{np_i}$$

$$\fallingdotseq\left\{\frac{\sigma_i^2}{m_i}-\sigma_{bi}^2\left(\frac{1}{m_i}-\frac{1}{np_i}\right)\right\}$$

となる。

常に $\frac{1}{m_i} > \frac{1}{np_i}$ であるから此の層化によつて第2項の分散

分散が小となり得をすることになる。

註,

$$S_i^2 \doteqdot \frac{\sigma_i^2}{m_i} - \sigma_{bi}^2 \left(\frac{1}{m_i} - \frac{1}{np_i} \right)$$

を考えてみる。

もし $\sigma_{bi}^2 = 0$ ならば結果は0となる。

又 $m_i = np_i$ に於ても同様である。母集団の比率 q_{ij} がわかっており $q'_{ij} = q_{ij}$ となしうるならば

$$S_i^2 \doteqdot \frac{\sigma_i^2}{m_i} - \frac{\sigma_{bi}^2}{m_i}$$

となりもつとも効果がある場合となる。 $-\frac{1}{np_i}$ の項は q'_{ij} が確率変数であるために起る損失である。

さてそれで全体の推定値

$$\overline{x} = \Sigma p_i \overline{x}_i$$

はどうなるか、

此は勿論 $\overline{X}$ の偏りのない推定値である。

この時の分散 S^2 は，近似的に

$$S^2 = \sum_{i=1}^{R} p_i^2 S_i^2$$

$$\doteqdot \sum_{i=1}^{R} \frac{N_i - m_i}{N_i - 1} \frac{1}{m_i} p_i^2 \sigma_i^2 - \frac{N-n}{N-1} \Sigma \left(1 - \frac{m_i}{np_i}\right) \frac{p_i^2 \sigma_{bi}^2}{m_i}$$

$$\doteqdot \frac{\sigma^2}{m} - \frac{1}{m} \Sigma p_i (\overline{X} - \overline{X}_i)^2 - \Sigma \left(1 - \frac{m_i}{np_i}\right) \frac{\sigma_{bi}^2}{m} p_i$$

$$\doteqdot \frac{\sigma^2}{m} - \frac{1}{m} \left(\Sigma p_i (\overline{X} - \overline{X}_i)^2 + \Sigma \left(1 - \frac{m_i}{np_i}\right) \sigma_{bi}^2 p_i \right)$$

によつてあたえられる。　層化の効果は第2項によつてあたえられている。　即ち完全な層化抽出法でなくとも一般に効果のあることが保證された。

我々の場合　$\dfrac{m_i}{np_i} \fallingdotseq \dfrac{1}{4}$　であるから

$$S^2 \fallingdotseq \frac{\sigma^2}{m} - \frac{1}{m}\left(\Sigma p_i (\bar{X}-\bar{X}_i)^2 + \frac{3}{4}\Sigma \sigma_{bi}^2 p\right)$$

とあらはせる。

以上の論議の保證によつて我々は各層の中を層化し欲する丈のサムプルをランダムに抽出することにした。

此によつて相当分散を小になし得，500のサンプルであつても層化しない時のより大きなサムプルに相当する精度をもたせることが出來るであろう。　云いかえてみるならば §2 でのべたよりもさらによい精度を得ることになるであろう。

§5. その他

確実にサムプル500を得るために止むを得ざる調査不能者を見込んで各層から補助サムプルを抽出，万一の場合にそなえた。

此の数は116である。

然し此はサンプルの追及調査を忽にしてよいと云ふ事ではない。

上述の様に細かく組むだサムプルであるから，調査はますます嚴密にし，あくまでも追及調査し，移転，死亡，重病，長期の旅行等，真に止むを得ず調査できぬもの丈サムプルの差し替えを行い，予定のサムプル数を確保したいが爲である。

§6. 此の様な計画は実際に可能であろうか、一番困難と思われるのは面接に行つた時不在であるものを追及することである。　このためあらかじめ在宅者率と言ふものをしらべておく事は大切なことである。

準備調査のとき此を行つてみた。（実際には同時に媾和問題につき，世論調査を行つて政治的関心をもあわせて調査しておいた。）

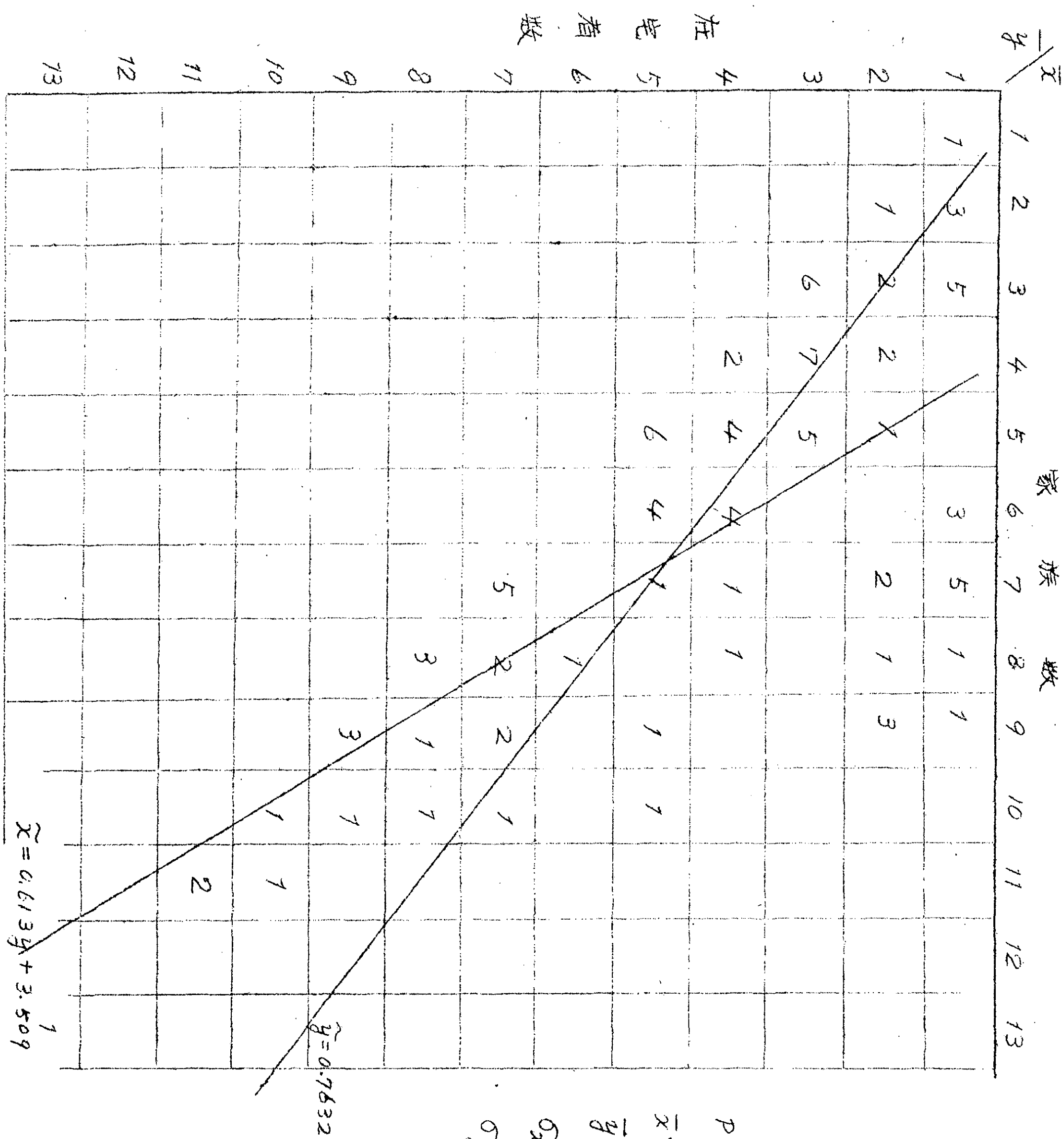
サムプリングの方法は各世帯を調査単位とした。即ち等しい確率を以て世帯を抽出した。

各町を層としてこゝに含まれている世帯数に比例してサムプルを割あてた。この方法によつて今の在宅者と家族人数を面接によって調査した。こうしてから在宅者率を Regression Estimate を用いて推定することにした。此の方法によるとき9時――4時迄（此の時間が実際調査した所の時間）の間の平均在宅者率（時間平均*）が求められるであろう。

さて，

家族人数と調査時に於ける在宅者数との相関関係を求めてみると次の様になる。（次頁）

註。*各人の訪問される時刻は等しい確率で決定されるとの仮定をしていることになる。



$P = 0.684$

$\bar{x} = 6.091$

$\bar{y} = 4.212$

$\sigma_x = 2.499$

$\sigma_y = 2.790$

y の x に対する回帰直線は

$$\tilde{y} = 0.763x - 0.435 \quad \text{となる。}$$

$\bar{x} = 4.9$ と云ふ事（白河市の一世帯当り平均家族人数）は既知であるから平均在宅者（一世帯当り）が回帰線からよみとれる比は 3.3 である。こゝで $\bar{y}$ の分散を計算してみると，

$$V(\bar{y}) = \frac{\sigma_y^2(1-\rho^2)}{n} = \frac{7.78(1-0.684^2)}{99} = 0.042$$

となる。したがつてその在宅率は

$$A = \frac{3.3 \pm 0.4}{4.9} = 0.67 \pm 0.08$$

（信頼巾　95 %）

と推定される，なお此の様な推定が正規分布にしたがふ事が石田によつて説明（講究録、火災推定参照）されてあるのでそれにしたがふ信頼巾がつけられてある。

これから本調査の場合 30 % の不在者を思はしめるが農閑期に入るからこれよりはいくぶん在宅率は上るであろう。

以上の様な目安によつて期日的にみて此の調査計画はまづ無理はないものとみてよい。

§ 7. 附　録

調査結果からみたサムプリングの精度

調査の主な目的は共通語化の程度と言ふ事であった。 その共通語化の程度は音韻の反応によって測定された。 しかもその程度は一応共通語的反応を示したところの項目の数によってあらわされることになった。

したがって此の標識をもとにしてサムプリングの精度を計算しておこう。 勿論推定は母集団平均の推定についてである、即ちサムプル平均*についてゞある。

*の信頼巾

まづ，層別の効果について考えてゆこう。

此を出すためには各層の分散を計算してゆかねばならない。 しかし重ねぬきをした時考えた層のそのについてはサムプル数が過少であったため分散を推定することは意味が少いと考えたので男女×年令の層別についてのみ話をすゝめてゆくことにする。 重ねぬきの方の時用いた所の層に就ても層の平均の間に標識の上で差のあることは結果の分析から判明してゐるので以上の様な検討は相当控え目なものであることが言われるであろう。

さて層別した時の分散を計算すると

$$\sigma_{st}^2 \fallingdotseq \frac{N-n}{N-1}\,\frac{1}{n}\sum_{i=1}^{R} p_i \sigma_i^2$$

$N = 17890$

$n = 476$

p_i は既述

σ_i は i 層の分散

$\sigma_{st}^2 = 0.03236$　　$\sigma_{st} = 0.1799$

層別しないとした時の分散　σ_r^2 は

$\sigma_r^2 = 0.03854$　　$\sigma_r = 0.1963$

となる。　層別の効果は，標準偏差の意味で

$$\frac{\sigma_{r} - \sigma_{st}}{\sigma_{r}} = 0.08$$

となる。　即ち約 8％ 効果があった事になる。

層別の効果が比較的少なかったのは意外であった。

さて母集団平均の推定値 $\bar{x}$ は

$$\bar{x} = 19.6$$

であるから，サムプル平均の変異係数をもとめると

$$\frac{\sigma_{st}}{\bar{x}} = 0.01$$

となり，サムプリングの精度は相当高いと言われるであろう。

何とならば，信頼巾をしめす相対精度

$$k \frac{\sigma_{st}}{\bar{x}}$$

（ k は此の場合，信頼度 95％ で 1.96 と言へてよい）
は，2％ 以下となるからである。

サムプリングの結果はまづ満足すべきものと思う。