5通りの二段抽出法における標本平均値の偏りと標本抽出分散の評価のためのコンピューター・シミュレーション

——調査不能群をも想定して——

橋爪　浅治[1]・国枝　玲子[1]・高宮　義雄[2]
杉山　明子[2]・村瀬　久子[1]・中嶋　千春[1]

(1970年11月　受付)

Simulation Analysis for Sampling Bias and Variance of Sample Mean from Two-Stage Sampling

A. Hashizume[1], R. Kunieda[1], Y. Takamiya[2],
M. Sugiyama[2], H. Murase[1], T. Nakajima[1]

The two stage sampling method -- 1st stage: sample clusters are selected at random having no relation with cluster size, 2nd stage: a pre-determined number of samples are selected at random from the sample clusters -- is introduced in many text books of sampling survey, but rarely put into practice by field investigators, since there are problems in the point where sampling and calculation works are somewhat troublesome.

It is, however, difficult in general to seek exactly sampling bias and variance of the sample mean from two stage sampling methods utilized by field investigators.

By means of computer simulation, therefor, an attempt is made to evaluate these by five ways two stage sampling methods which are available from the points of theoretical and practical view, respectively.

Japan Womens University Computation Labolatory
NHK Public Opinion Research Institute

1. 目的および背景

標本調査において，二段抽出法が最も普遍的に用いられてきているが，これの利用頻度は別として，普通の text book に基礎理論として大てい書かれている方法は 各集落に第2次抽出として抜かれるべき標本数をあらかじめ定めておき，第1次抽出単位は，等確率で抽出し，この抽出された集落から，今述べたあらかじめ定められた数の標本を無作為に抽出する．——これの特別の場合を，3 で方法 D′ と名付けて説明されるであろう．——

方法 D′ による標本平均の不偏性と標本分散は，周知のように理論的な評価は容易である．しかしながら，もし総標本数を一定にしたいとか，抽出作業を楽にしたいため，必ずしも方法 D′ ではなく，費用またはた，抽出技術の観点からこれと類似した別の方法を用いることの方がむしろ普通である．このような場合，標本平均の偏り性と分散の理論的評価は，意外と困難であり，多くの標本調査者は，「方法 D′ の場合に理論的に評価される不偏性と分散の知識を基にして，これと類似の方法の場合のそれを経験のつみ重ねによる判断によって潜在的に評価する．あるいはもっと大ざっぱに単純無作為抽出法で得られるであろう結果から倍率をかけて類推する」ことによってこの問題を処理してきているのが実状であると思う．

---

* 本報告は日本女子大学計算研究所と NHK 放送世論調査所との共同研究による．

1) 日本女子大学計算研究所，Japan womens University Conputation Labolatory
2) NHK 放送世論調査所，NHK Public Opinion Pesearch Institute

さらにまた，母集団に調査不能群（標本入手困難な者等）がかなりの無視できない程度存在することは避けられないことである．こうなると問題は，ますます繁雑化して，前述の偏りと分散が，母調査不能率とどのような関数関係を持っているのか，その傾向を推察することができるとしても，数値的に掌握することは困難である．正確には，偏りと分散は，調査不能率ばかりでなく例えば，われわれの調査を．ある質問に Yes と反応する者と No と反応するものに分け，母 Yes 反応率を推定することに限定すれば，Yes 反応群，および No 反応群における調査不能率；調査可能群および不能群における Yes 反応率；母平均；全分散；級外分散；集落の大きさの分布等に影響されるであろう．

3 で説明される5通りの二段抽出法 A, B, D, B′, D′ を上述の母集団に用いた場合，標本平均値の偏りと標本分散の評価を考える．実際の標本調査においては，母集団も抽出方法も多種多様であるが，そのうちの比較的目に触れやすいと思われるもののごく一部，すなわち氷山の一角について理論的に評価しにくい偏りと分散をコンピューターシミュレーションによってわれわれは求めた．これによって，広汎なる各調査母集団のそれぞれの場合におけるアプローチの一つの手がかりとなることを念願とするものである．

2. 解析に用いる凝似母集団

調査母集団を特徴づける要素は，多岐多様であるが，そのなかで，調査実務家が最も遭遇しているものを見出し，それに基づく凝似母集団を構成しなければ有効でない．本報告で解析に用いられた凝似母集団の構成要素，特徴と用いた理由を述べてみる．

i) 母集団集落数 $R=50$ および 300

例えば，全国調査のような広汎な標本抽出の場合でも，母集団集数 R として，そう大きな数を用意する必要はなく，300 程度で十分であろう．何故ならば，このような場合，層別されるのが普通で，層別された一つ一つの層を母集団と考えれば，$R=300$ は，十分な大きさであろう．小規模な標本調査として $R=50$ の場合も考えるであろう．

ii) 集落の大きさの分布

図1の上段は，鳥取，岡山，広島，山口の各県における NHK が通常の調査で用いている集落の大きさ（住民数）の度数分布を示めす．これらは，平均: 1300～1800; SD: 360～500 の範囲の値をとっている．また図の中段は，東京都内における大学の学部を集落と見た場合の学生数の度数分布で，平均: 2080, SD: 1800 である．これら5つの度数曲線は山型となっており，われわれの凝似母集団の集落の大きさは，学生分布の平均，SD を参考にした正規分布に従っていると仮定しよう．もし．正規性以外の分布であったとしても，われわれの結論にそう大きな影響を及ぼさないであろう．これについては **5** で検討されるであろう．

実際の二段抽出調査において，人口の大きな集落は，—— 例えば，2000 を超えるものは，2000 以内になるよう——分割すること——例えば．3000 なら 1500+1500 と2つに，5400 なら 1800+1800+1800 と 3つの集落に——が抽出技術上行なわれる場合もある．図1．上段の4つの県の集落について人口 2000 を超えているものに，上述の修正を行なったものを図の下段左側に示す．

以上の点を考慮して，われわれの用いた凝似集落人口は，平均: 2100, SD: 1800 の正規乱数で，200 未満は，取除き，2000 を超えるものは，上述の方法で修正させたもの $S_1 S_2 \cdots S_R$ を，後で述べる MODEL-I MODEL-II に採用した．図1下段右は，修正を加えた 300 コの正規乱数の度数分布を示す．

iii) 母 Yes 反応率 R コからなる集落において，i-番集落における Yes 反応者総数，No 反応者総数，集落人口総数，Yes 反応率をそれぞれ Y_i, N_i, $S_i(=Y_i+N_i)$, $P_i(=Y_i/S_i)$ とし $(i=1,2\cdots,R)$，また全体としての母集団における対応するものを添数をつけないで記号づけると，

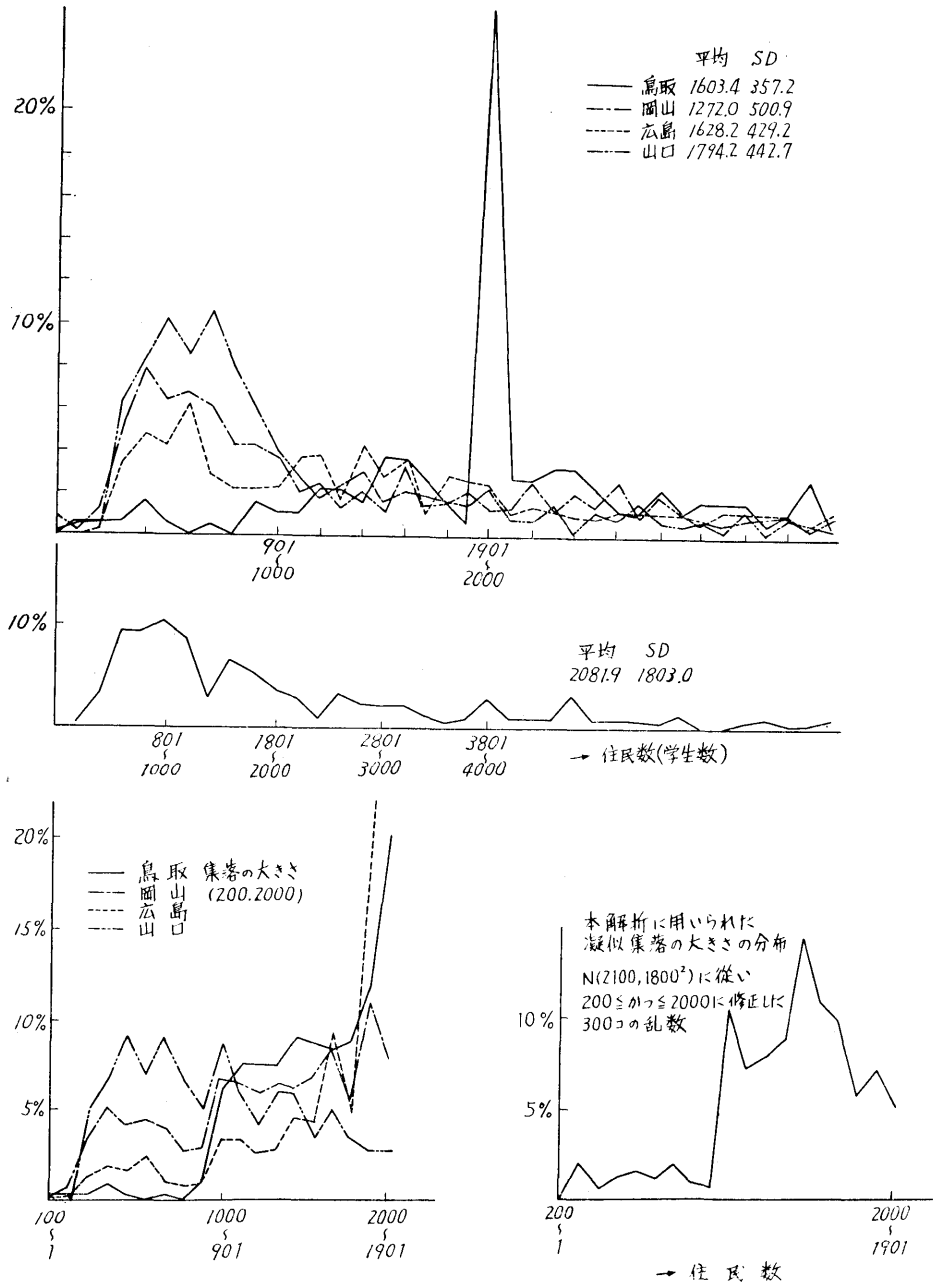


図 1.　4つの県住民の集落部内大学学部単位学生数および凝似集落の大きさについての分布

$$P = Y/S = \sum_{i=1}^{R} \frac{S_i}{S} \cdot P_i \qquad (1)$$

但し，

$$S = \sum_{i=1}^{R} S_i, \qquad Y = \sum_{i=1}^{R} Y_i$$

で，また全分散 S_t^2，級間分散 S_b^2，級内分散 S_w^2 に関して，周知のように

$$S_t^2 = S_b^2 + S_w^2 \qquad (2)$$

但し，

$$S_i^2 = P(1-P) \qquad (3)$$

$$S_b^2 = \sum_{i=1}^{R} \frac{S_i}{S} (P_i - P)^2 \qquad (4)$$

$$S_w^2 = \sum_{i=1}^{R} \frac{S_i}{S} P_i(1 - P_i) \qquad (5)$$

と分解される.

今, P と S_b^2 の数値を予め与えておき, 平均 P, 分散 S_b^2 に等しい Beta 分布

$$f(x;\, p, q) = \frac{1}{B(p, q)} x^{p-1}(1 - x)^{q-1}$$

但し,

$$P = \frac{p}{p + q}, \qquad S_b^2 = \frac{pq}{(p + q)^2 (p + q + 1)}$$

からの R コの乱数で, $P_1, P_2 \cdots P_R$ を決めたとき, これらと, 前に述べた乱数 $\{S_i\}$ を (1) および (4) の右辺に代入して得られる P と S_b^2 は. 予め与えられた数値と多少の喰違いが生ずるであろう. しかし, (1) 式が正確に成り立つように乱数 $\{P_i\}$ に適当な補正を加えることはできる. 従って, 全分散を予め与えた数値に正確に等しく ((3) 式), 級間分散を予め与えた数値に接近している凝似母集団を作ることができる.

図2における4つの図形は, 平均 0.2 の Beta 分布において (級間) 分散を全分散の 1/∞, 約 1/10, 約 1/15, 1/1 倍した場合の比較を示す. 即ち, 左から a) $S_b^2 = 0$ の場合は, Yes 反応率は, 点 0.2 で固定され, b) 級間分散が小さいときは, 分布型は鋭い山型をしており, 級間分散が大きくなると山型はゆるみ, c) ついには谷型に変化し, d) 級間分散が全分散に一致すると, Yes 反応率は0か1かの両端に吸収されてしまう.

図2を参考にして $P=0.2$ および 0.4 の各々について, 級間分数は0と, 全分散に等しい場合を両端として, 中間に3コの値を持つ場合を考えることにする.

i), ii), iii) によって構成される凝収母集団を MODEL-I と名付けることにする. その種類は, 集落数×全分散×級間分散 $= 2 \times 2 \times 5 = 20$ 通りとなる.

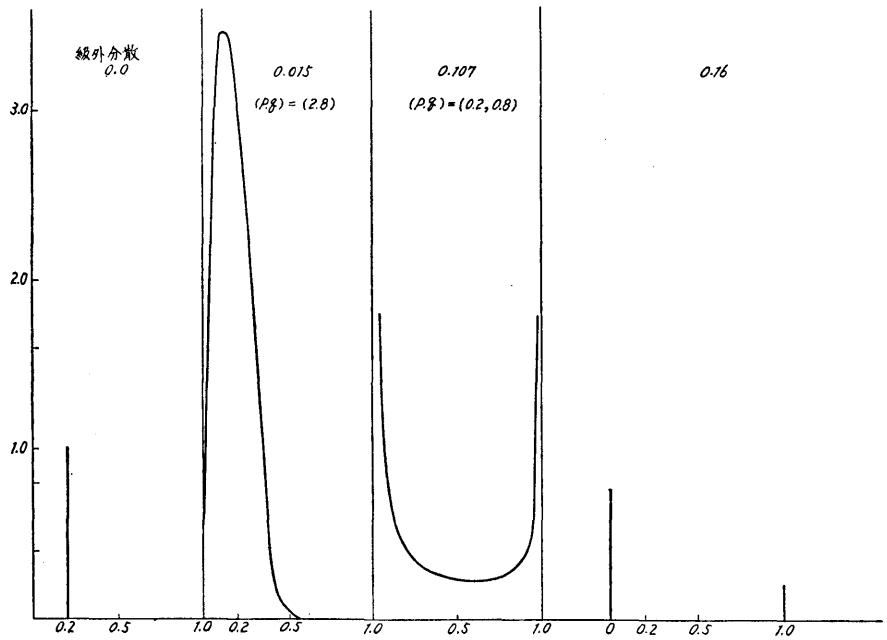


図 2. 平均 0.2 の Beta 分布の分散を変化させた比較

iv)　母調査不能率　NHK による最近数年間に行なわれる国民世論調査.（面接法による）の不能率は，20～30% である：

	不能率		不能率
第30回衆院選挙全国調査	26.9%	国民世論調査　43年6月調査	24.2%
第31回　〃	21.5%	〃　10月〃	25.1%
第32回　〃	31.9%	〃　10月〃	28.1%
第8回参院選　〃	24.2%	〃　44年3月〃	31.2%
第8回　〃	21.1%	〃　〃　10月〃	29.8%
		〃　11月〃	31.1%
		〃　12月〃	31.6%

第4次国民性調査（統計数理研究所）では，調査不能率は，23%で，各地方別調査不能率の分散は，0.0036 である．これは，全分散の $0.0036/(0.23\times0.77)\fallingdotseq 1/50$ に相当する．われわれは，各集落の調査不能率を平均 $u=10\%, 20\%, 30\%$，分散は，全分散 $u(1-u)$ の 1/20 の Beta 分布乱数で定めた．

擬似母集団に調査不能率を組込むのに，理論的観点および実用的観点から，それぞれ，次に述べる MODEL-II および III の構成方法が考えられる．

MODEL-II: MODEL-I で定義された各集落の Yes 反応群に調査不能率を iv) で説明した乱数で割り当てる．また，No 反応群における調査不能率は，今述べた乱数の一定倍 ——これを**伸縮率**と名付けおく——を用いることにする．伸縮率として 0.6, 0.8, 1.2, 1.4 を用いることにする．MODEL-II の種類は，MODEL-I×Yes 反応群の調査不能率×伸縮率＝20×3×4=240 通りとなる．

MODEL-III, i), ii) によって，集落数および集落の人口数を決めたら，iv) で説明した Beta 乱数によって，調査不能率と可能群の人口数を決める．次に，調査可能群における Yes 反応群を iii) の方法で割り当て，調査不能群における Yes 反応率は，一定倍したものを用いる．

MODEL-II は，MODEL-I の場合と同様，母 Yes 反応率を予め，与えられた数値に一致するよう構成することができるので，全分散が一定で，級間分散が変化している母集団における標本抽出による分散の変化を調べるのに便利であるが，MODEL-III では，この目的のためには不利である．しかしながら，MODEL-III は調査可能標本 Yes 反応率の母数に対する偏りが，調査不能率と，調査可能群における Yes 反応率——これらは，標本によって推定できる——とどのような関係にあるかを知ることができるので実用上の価値がある．

本報告では，MODEL-I および II のみについて述べ，MODEL-III は別[1] に論じられるであろう．MODEL 番号に続いて，集落の数が $R=50$ または，300 であることを表示するために，S または，L; 全体としての母集団における Yes 反応率 20%，または，40% の % 数字を記すことにする．例えば，MODEL-II-L-40 は，調査不能群を含む母集団 ——その構成内容は別記——で $R=300$, $P=40\%$ を意味している．

3. 標本抽出方法と推定式

母 Yes 反応率 P を推定するための二段標本抽出として，次の4通りの抽出方法，A, B, C, D を考えよう．大ざっぱに方法 A と C は第1段：確率比例，第2段：等確率で，方法 B と D は，第1段：等確率，第2段：確率比例的な抽出方法といえよう．これらの抽出法によって得られる全 Yes 標本数を全標本数で割ったものを，母数 P の推定式としている．

また，方法 B と D において，別の推定式（後述）を用いたものを，B′, D′ とすれば，抽出方法と推定方法により合計 A, B, C, D, B′, D′ の 6通りの組合せとなる．

方法 A（第1段：系統的抽出）集落を無作為な順序で並べたら，それを固定しておく．次に一直線上に，この順序で原点より長さ S_i の区間で分割点を作り，出発番号を $[1, 2\cdots, T]$（た

だし $T=S/r$) のなかから無作為に選び，抽出間隔 T で，系統的抽出法により，R コの集落から r コを第1次抽出単位として選ぶ．これから——その集落人口に無関係に——n/r の大きさの標本を，第2次抽出としてそれぞれ選ぶ．推定式：前述．これによって不偏推定量が得られる．

方法 B 第1段：R コの集落より r コを等確率で抜く．第2段：第2次抽出数は，第1次抽出集落の人口数に比例し，かつ，総数が一定 ($=n$) となるように決める．推定式：前述．これは明らかに不偏推定とならない．

方法 C 第1段：まず，集落1つを，その大きさに比例した確率で抜く．つぎに残りの集落から同じようにして抜く‥‥このようなことを r 回継続する．第2段：方法 A に同じ．推定式：前述

これの推定量は，$r=1$ の場合を除き，偏りを生ずる．また，$r=R$ の場合，第1段＝確率比例という意味が全く薄れてしまう．

方法 D 第1段：方法 B に同じ．第2段：第2次抽出数は，その第1次抽出集落の人口数に比例し，かつ，その比例定数は一定にしておく．推定式：前述

方法 B′ および D′ 抽出方法は，それぞれ方法 B および D でと同じあるが，推定式として

$$\frac{1}{S}\cdot\frac{R}{r}\sum_{i=1}^{r}S_i\times(i\text{番集落における標本 Yes 反応率}) \qquad \cdots\cdots(6)$$

を用いる．

方法 B′, D′ は B, D と異なり，不偏推定量が得られる．

方法 B, B′ は，全標本数が一定で，抽出集落個々の標本数は，確率変数的であるのに対し，方法 D, D′ では，それが逆になっている違いがある．本報告では，方法および第一次抽出数にかかわらず，**標本数**はすべて 1000 である。以後このことは，ことわらない．

本報告では方法 C は取扱わない．次の章では，方法 A と B をよび．それらの比較について詳細に解析し，続く章では，MODEL-I を用いて，B, B′, D, D′ 間の比較検討を行なう．

4. 解析方法および結果

2. で述べた MODEL-I および II の凝似母集団から，**3** で述べた方法 A と B による (調査可能) 標本 Yes 反応率の期待値と抽出による SD を理論的に求めることが困難のため，標本 Yes 反応率を多数回，独立に繰返し求め，それらの平均値と SD で評価することにする．本報告では，繰返し数は，すべて 100 回としてあるが，これによる上述の期待値，抽出 SD の評価としての信頼性は，次章でふれることにする．

以降において，断らない限り，期待値 SD (MODEL-I)，SD_1, SD_2 (MODEL-II)，偏りを次の意味に限定して用いるであろう．

期待値：標本 Yes 反応率の繰返しに関する平均値

SD^2 (MODEL-I)，SD_1^2 (MODEL-II)：標本 Yes 反応率の期待値からの偏差二乗平均（繰返しに関する）

偏り：母 Yes 反応率の期待値からの偏差，即ち，これを標本 Yes 反応率に加えれば，不偏推定量が得られる．

SD_2^2 (MODEL-II)：標本 Yes 反応率の母平均からの偏差二乗平均

附録のフローチャートによって MODEL-I の構成と，抽出法 A による期待値と SD の計算過程が詳細に分かるであろう．

20 通りある MODEL-I について方法 A, B による，たゞし，第1次抽出数は，2, 10, 20, 30, 40, 60 (60 は $R=300$ の場合にのみ) とした場合についての解析結果は．図3にまとめられてある．即ち，図の左側1欄は，期待値，右側4つの欄は SD についての第1次抽出数 r

に関する傾向折線を示している．

期待値

図3から偏りの程度は，母 Yes 反応率 P，集落の数 R と殆んど無関係なことが観察される．偏りの絶対値は，r とともに減少し，また方法 A より B の方が大きい傾向を示しているが，それでも $20 \leq r$ で殆んど ±1% 以内となっている．しかし，$r \leq 20$ では ±3% 位のところがあり，厳密性が要求される実際調査においては警告を与えている．

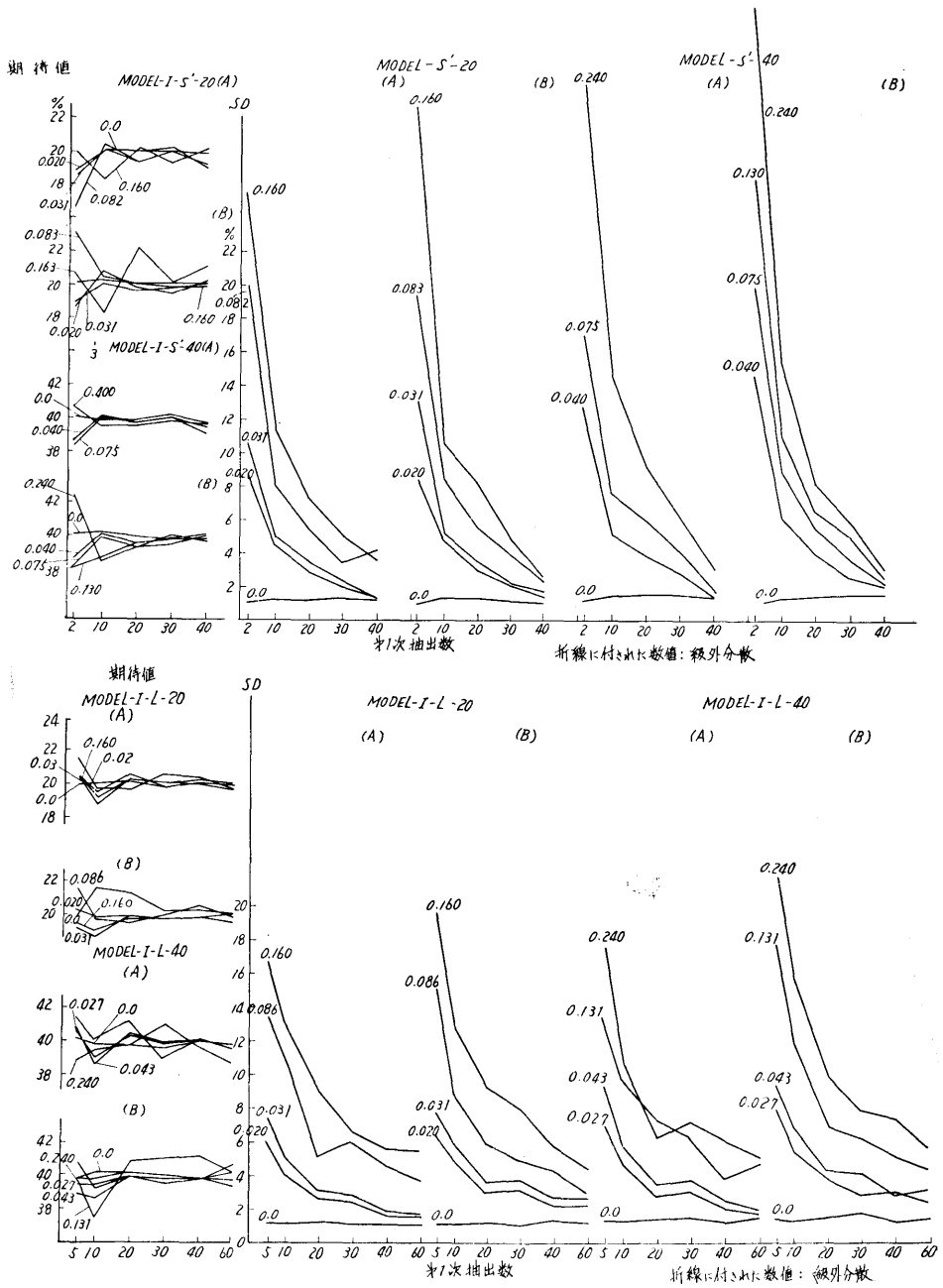


図 3.　全分散一定，級外分散をいろいろ変えた調査不能のない母集団から二段抽出法 A，B による標本平均のシミュレーション期待値および SD の第一次抽出数についての傾向折線の比較（全分散別）

SD

SD の r に関する傾向曲線は，級間分散 $S_b^2=0$ の場合は，$SD=1\sim2\%$ の範囲で，r に関して一様な水平線を示しているが，S_b^2 が増加した凝似母集団では，r に関する単調減少型の曲線となり，しかも SD の高さおよび，SD の減少率は S_b^2 と正の相関を持っていることが観察される．

20 通りの凝似母集団全部について，また，すべての r について，方法 A より B の方が SD が高くなっていることに関心を持つべきであろう．このことは，文献[2] の結果とも一致している．

傾向曲線の高さは，S_b^2 が一定の場合，P と正の相関を持っていることも図から見出されるが，P の違いによる比較は，全分散が違ってくるので意味がない．むしろ，上に述べたことは，S_b^2/S_t^2 を一定にした場合，傾向曲線は，P に関して一様であると解釈したい．

図3の上 ($R=50$) と下 ($R=300$) の傾向曲線を比較すると，$r(\leq30)$ の増加による SD の減少効果は，R が小さい程良いが——$S_b^2=0$ および $S_b^2=S_t^2$ の特殊な場合を除いて——$r=30\sim40$ 位が最も効果的で ($SD=2\sim5\%$)，それ以上 r を大きくしても相応する SD の減少効果は得られない．一般の標本抽出実務において，級間分散について，よほどの好条件の知識がない限り，第1次抽出数を 30 未満にすると危険であることを，この図は示唆している．極端な場合を図から読むと，MODEL-S-40 (B) $r=2$ の場合の SD は，実に 37% となっている．

240 種類の MODEL-II については，方法 A（のみ）による解析を行なった．ここで第1次抽出数の種類は，MODEL-I の場合と殆んど同じにしてある．偏りと SD_2 の第1次抽出数に関する傾向曲線を全分散，集落の数別に，5種類の S_b^2 群ごとに，調査不能率の組合せに対し，図4 (pp 99-103) の右半分と左半分に示してある．期待値と SD_1 の掲載は，本報告で割愛したが，もし必要なら

$$\text{期待値} = P - \text{偏り}$$

$$(SD_1)^2 = (SD_2)^2 - (\text{偏り})^2$$

なる関係があるから図4を用いて，直ちに算出できる．

標本 Yes 反応率に偏りが無い場合，および，有っても．それが評価できる場合には，不偏推定量を作ることができるから，これら二つの場合は，標本抽出の評価として，$SD(SD_1)$ は，意味を持つ．しかしながら，MODEL-II の場合，後述するように，かなりの偏りが生じ，この偏りと母集団の構成要素，例えば，Yes 反応群における調査不能率 u 等の間の関係が分っても，u の値そのものが調べられないので，偏りの評価ができない場合，SD_1 の実用上の利用価値は，薄いように思われる．偏り性と抽出精度の両方の性質を共有している SD_2 を MODEL-II で採用したのは，この理由による．

図4からの所見は，次の通りである．

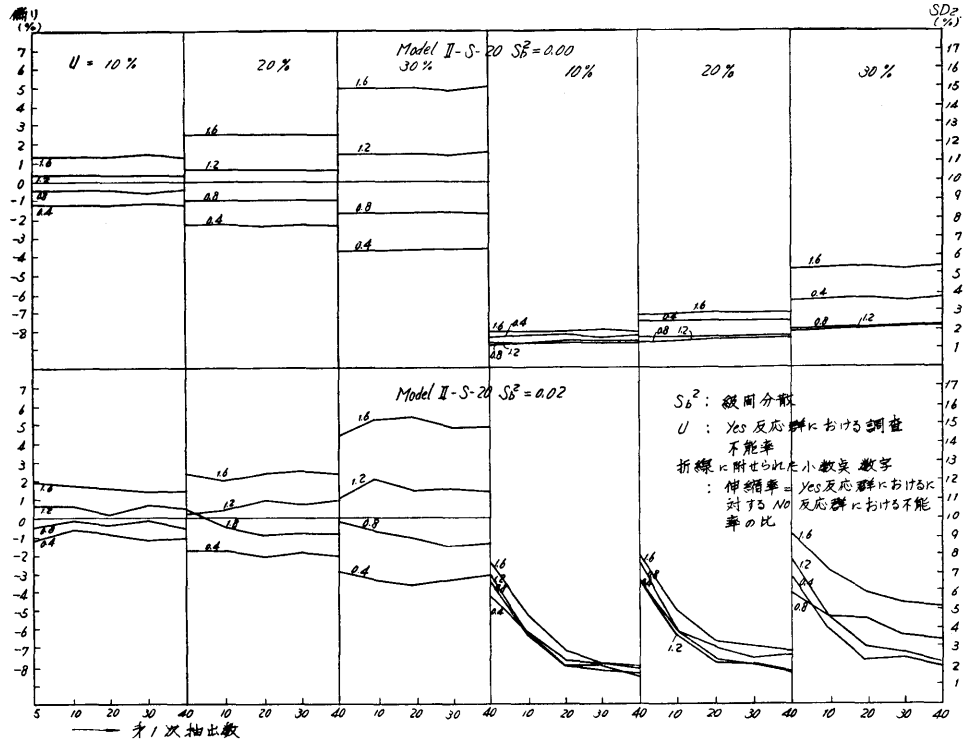
偏り

MODEL-I の場合とやゝ異なり，第1次抽出数に関し，何等の傾向も見出されていない．

伸縮率が1より小さい（大きい）程，偏りは負（正）の方向に向い，また伸縮率が同じ場合は，P の値が大きい程，偏りは，絶対値において大きい値を示している．その程度を図より拾うと下表が得られる．

P		20			40		
u		10	20	30	10	20	30
伸縮率	0.4	−2	−2	−4	−2	−4	−7
	〜	〜	〜	〜	〜	〜	〜
	1.6	2	3	5	2	5	8

数字は % 単位



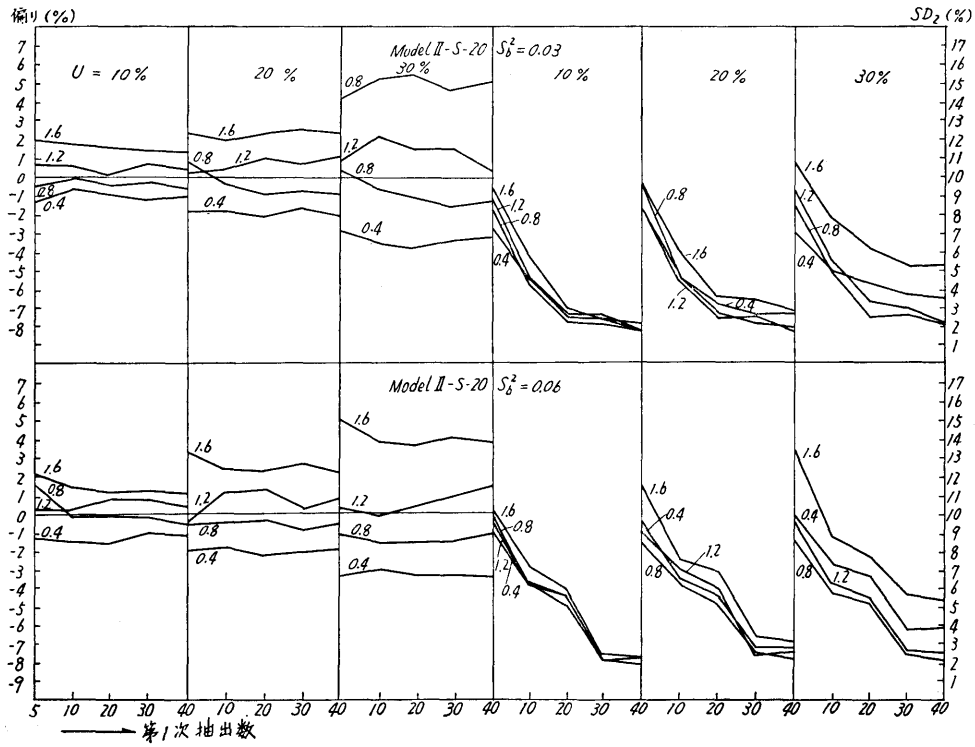
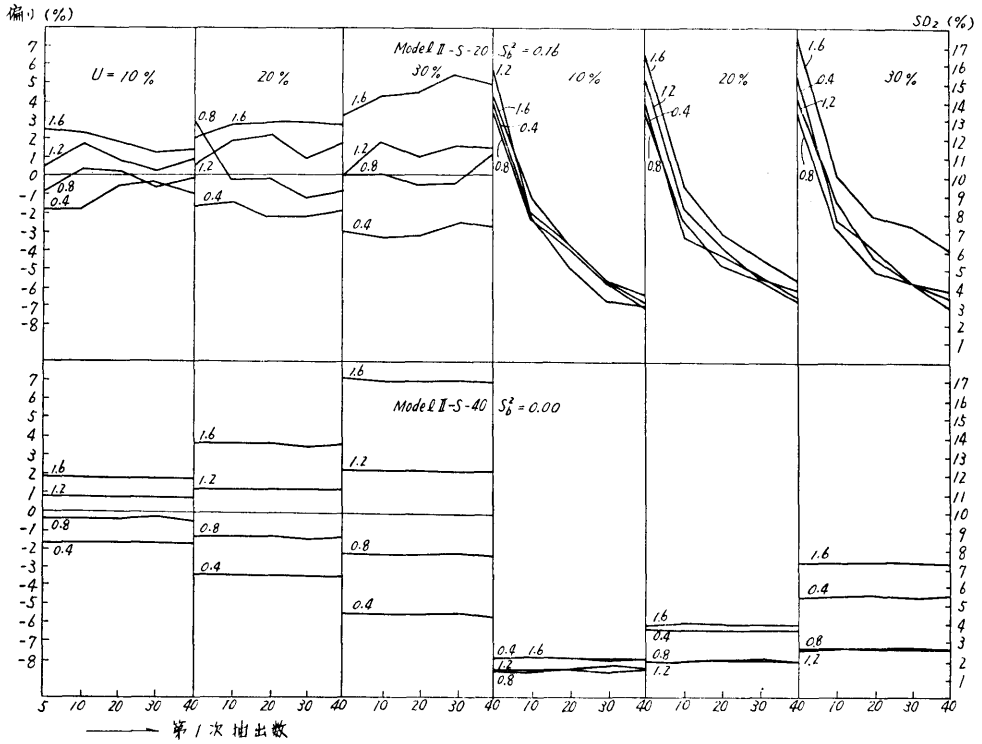


図 4.　全分散一定，5通りの級間分散，12通りの調査不能率の組合せの疑似母集団から二段抽出法Aによる標本 yes 反応率の偏りと SD_2 の第1次抽出数についての傾向折線（全分散別）

図4（続き）


偏り (%)
SD₂ (%)
U = 10 %
20 %
30 %
Model II-S-20 S_b^2 = 0.16
10 %
20 %
30 %
Model II-S-40 S_b^2 = 0.00
第1次抽出数

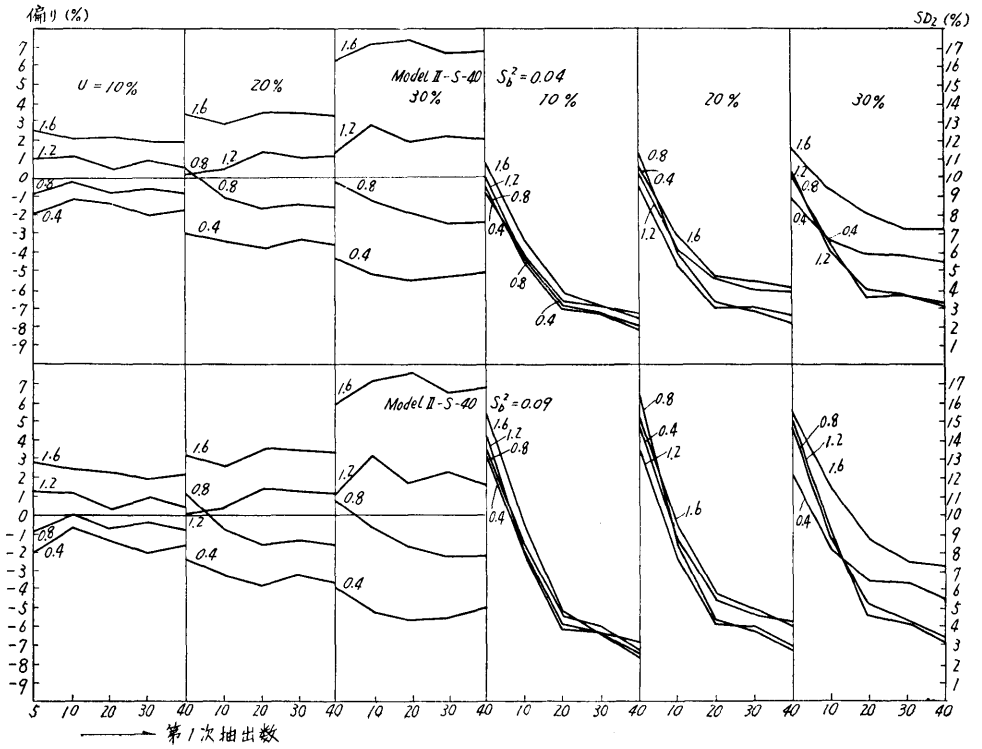
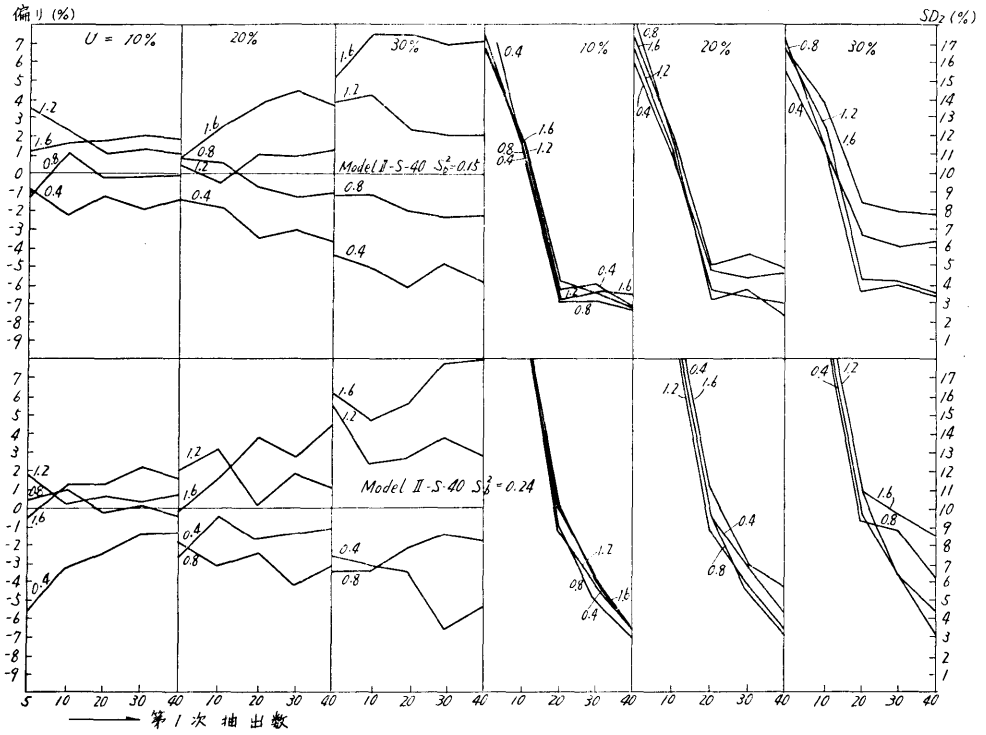

偏り (%)
SD₂ (%)
U = 10 %
20 %
30 %
Model II-S-40 S_b^2 = 0.04
10 %
20 %
30 %
Model II-S-40 S_b^2 = 0.09
第1次抽出数

図4（続き）


偏り (%)
SD2 (%)
U = 10%
20%
30%
10%
20%
30%
Model II-S-40 S0²=0.15
Model II-S-40 S0²=0.24
第1次抽出数

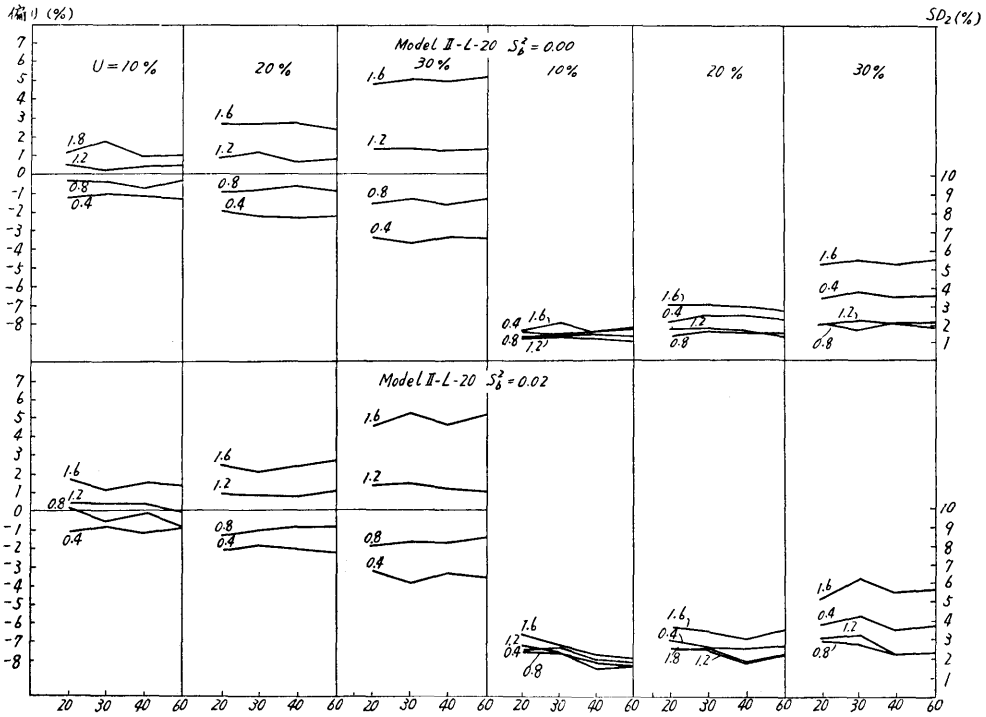
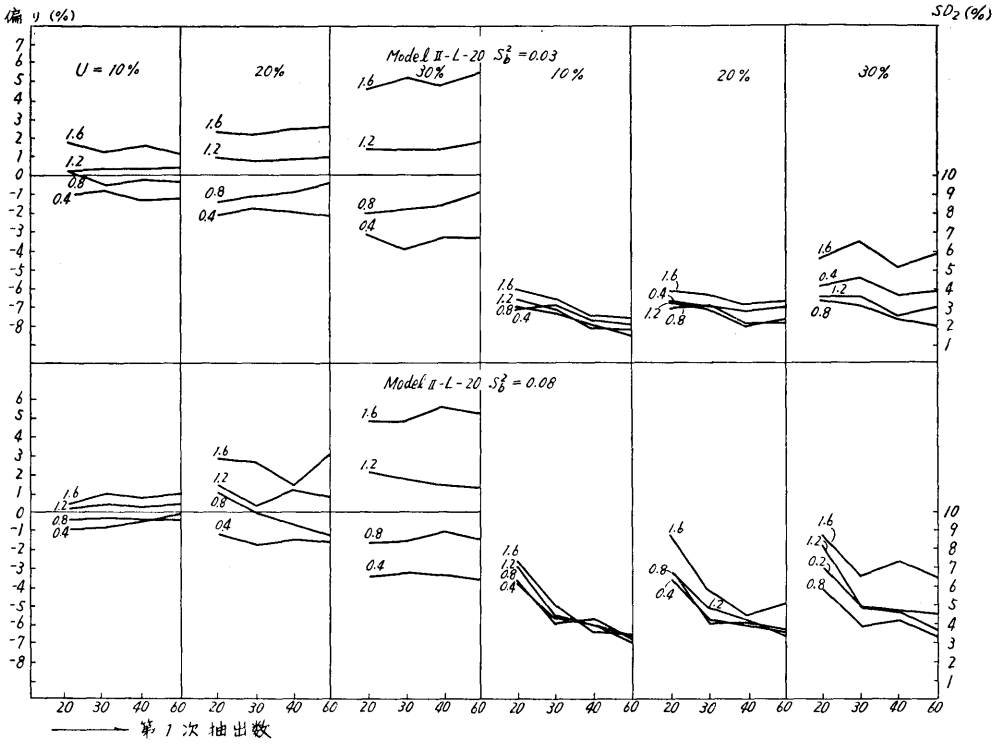

偏り (%)
SD2 (%)
U=10%
20%
30%
10%
20%
30%
Model II-L-20 S0²=0.00
Model II-L-20 S0²=0.02

図4（続き）


偏り (%)
SD₂ (%)
U = 10%
20%
30%
Model II-L-20 S_b^2 = 0.03
Model II-L-20 S_b^2 = 0.08
第1次抽出数

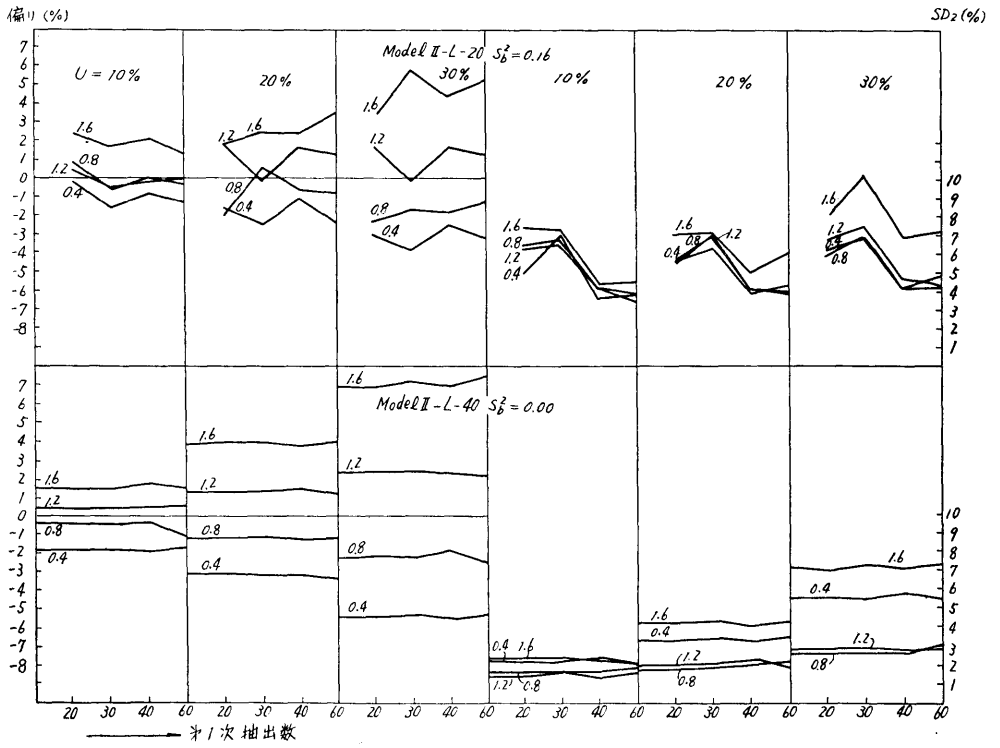
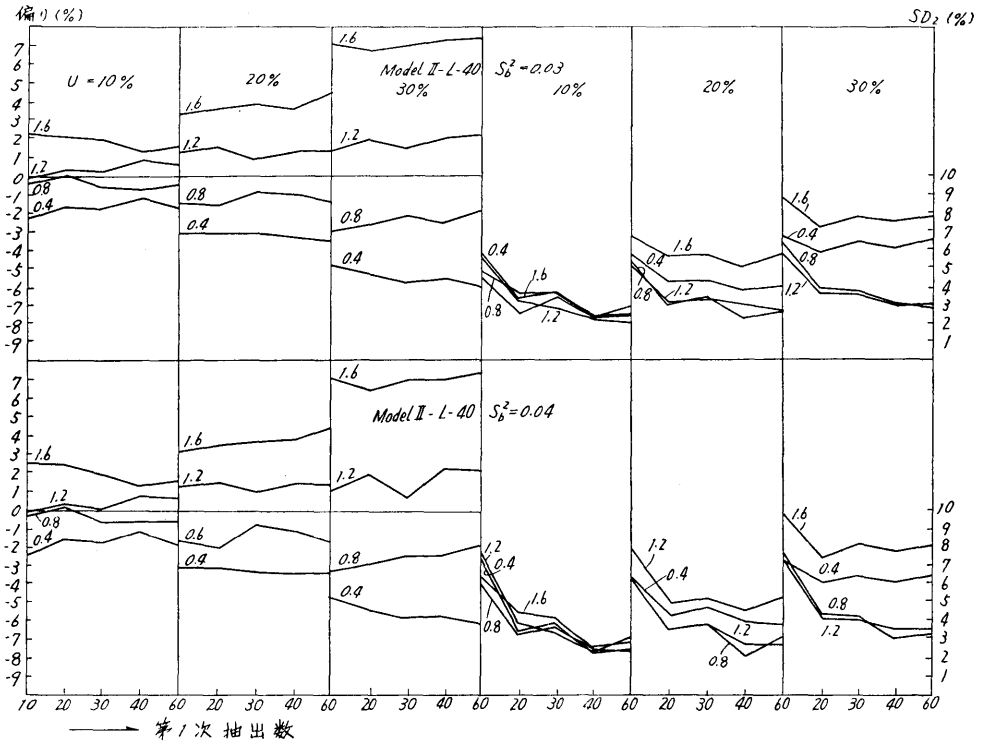
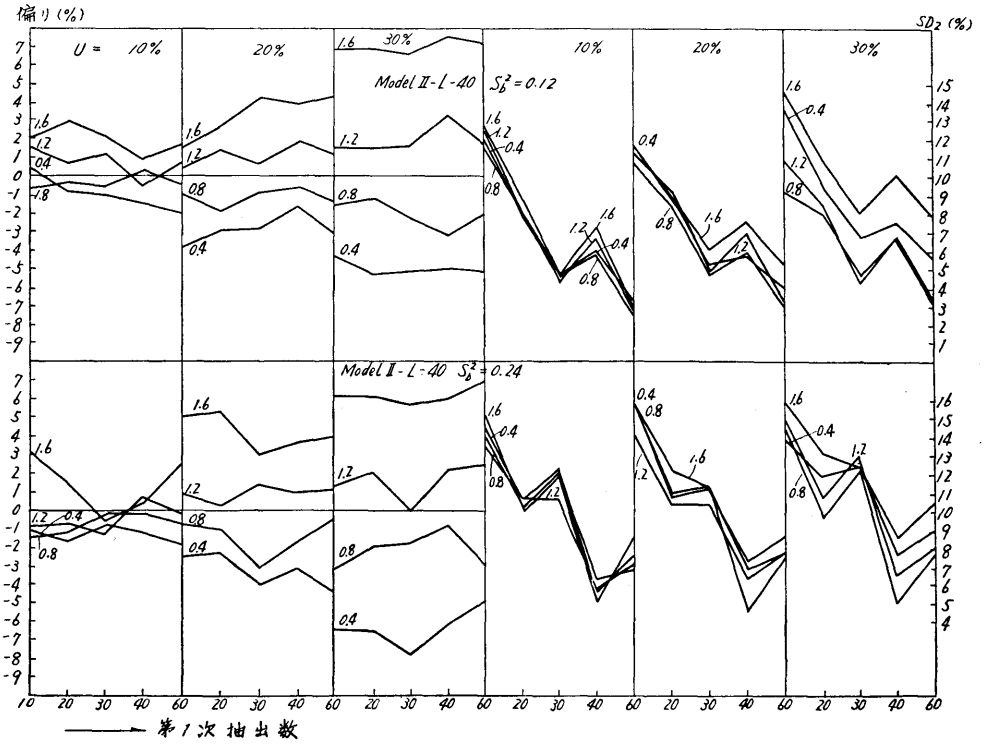

偏り (%)
SD₂ (%)
U = 10%
20%
30%
Model II-L-20 S_b^2 = 0.16
Model II-L-40 S_b^2 = 0.00
第1次抽出数

図4（続き）


偏り（%）
SD₂（%）
U＝10%
20%
30%
Model Ⅱ-L-40
S_b^2=0.03
Model Ⅱ-L-40
S_b^2=0.04
第1次抽出数


偏り（%）
SD₂（%）
U＝10%
20%
30%
Model Ⅱ-L-40
S_b^2=0.12
Model Ⅱ-L-40
S_b^2=0.24
第1次抽出数

上の表は，20～30% 程度の調査不能率は避けられない通常の標本調査において，伸縮率が1に近いという保証がなければ，無視し得ない偏りが生じ得るとゆうことをわれわれに教えている．

伸縮率 0.4 と 1.6; 0.8 と 1.2 の対は互いに相反する方向に同程度に偏りを起こさせていることが図からも観測される．即ち，偏り=0 の水平線に関し，これら対同志の折線は，対称的に書かれている．

SD_2

MODEL-I において，級間分散が SD の r に関する傾向曲線に及ぼす効果は．調査不能群が加わっている MODEL-II においても保存されていることが図から分かる．即ち，全分散が一定の場合，級間分散の大きいもの程，SD_2 は高い．そして，第1次抽出数が小さいもの程これが著しい．SD_2 を SD_1 におきかえても同じことがいえるが，図は省略する．

調査不能率の組合せが，SD_2 に及ぼす影響を調べるために，図4の右側の同一 S_b^2，かつ同一伸縮率の r–SD_2 曲線を左から右へたどってゆくと，すなわち，u を増加させると，曲線は，だんだん上に上ってきている．SD_2 が Yes 反応群における調査不能率 u と正の相関を持つのは，言うまでもなく，2つの要因がある．1つは，u の増加によって調査可能標本が減少し，従って抽出分散が多くなることと，他は．SD_2^2 は，SD_1^2 に偏りの二乗が加わっており，また偏りの二乗は u と正の相関をしている．(図4の左側を見よ.)

u が同一の場合 SD_2 は，必ずしも伸縮率と正の相関を持たず，大部分は．r–SD_2 曲線は，伸縮率 0.8, 1.2, 0.4, 1.6 の順に上にあがって書かれてある．この理由は，伸縮率が大きいと調査可能標本が減少し，抽出分散が大きくなるが，偏りのところで述べたように，伸縮率 0.8 と 1.2，および 0.4 と 1.6 は．それぞれ，偏りの二乗に同程度に関与しているからであろう．

$S_b^2=0$ および，$S_b^2=S_t^2$ は実用上特種の場合なので．これらを除いた各 r–SD_2 曲線を眺めると，大部分は，r=30～40 のところで SD_2 の減少はゆるやかになる．今，r=30 における SD_2 の値を図から取り上げてみると下表のようになる．

$r=30$ のときの SD_2(%)

P		20%			40%		
u		10%	20%	30%	10%	20%	30%
R	50	2～3	2～3	3～6	2～4	3～5	4～8
	300	3～5	3～6	3～7	3～5	3～6	4～8

表における小さい（大きい）値は，大体，伸縮率 0.8 ないし，1.2 (0.4 ないし 1.6) における SD_2 を示している．仮りに，SD_2=5% を標本抽出における偏りとばらつきを許容する限界だと決めると，伸縮率が 0.8～1.2 程度のわれわれの凝似母集団に対し，抽出法 A (r=30～40) は合格である．しかし，伸縮率 <0.4 または >1.6 のとき，P=20% の一部および P=40% であるものの大部分の凝似母集団における大きさ 1000 の標本抽出法 A は危険である．

5. 凝似母集団および解析方法に対する検討

ここでは，**2** で構成された凝似母集団とこれらとかなりかけ離れていると思われる別のものとの間に，方法 A によって得られる SD にかなりの喰違いが生ずるかどうか; **4** で得られた結果の信頼性はどうか; **3** で触れた別の抽出方法—推定式と前章で取扱ったものとの比較について一部論じてみたい．

集落の数 が，300 (-R) である MODEL-I-L-20 と，これを 1000 (-R) に増やしたもの

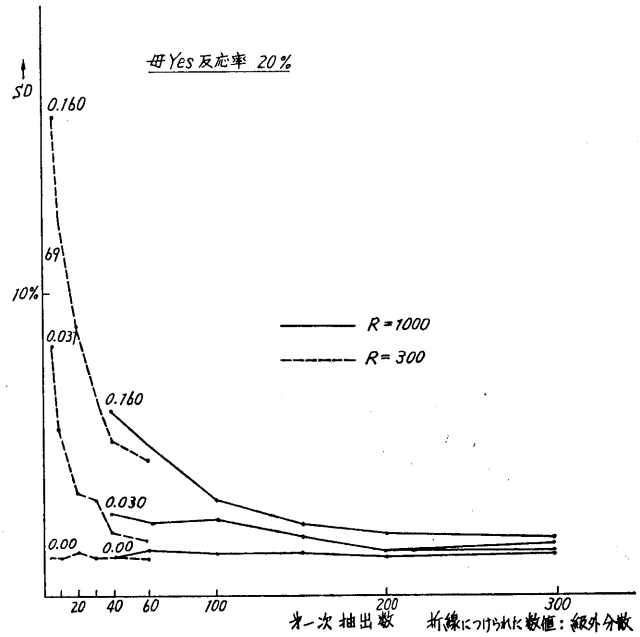


図 5. 集落数 300 および 1000 の種々の級外分散を与えた調査不能のない母集団からの二段抽出法 A による標本平均の SD の第 1 次抽出数に関する傾向折線の比較（標本数：1000）

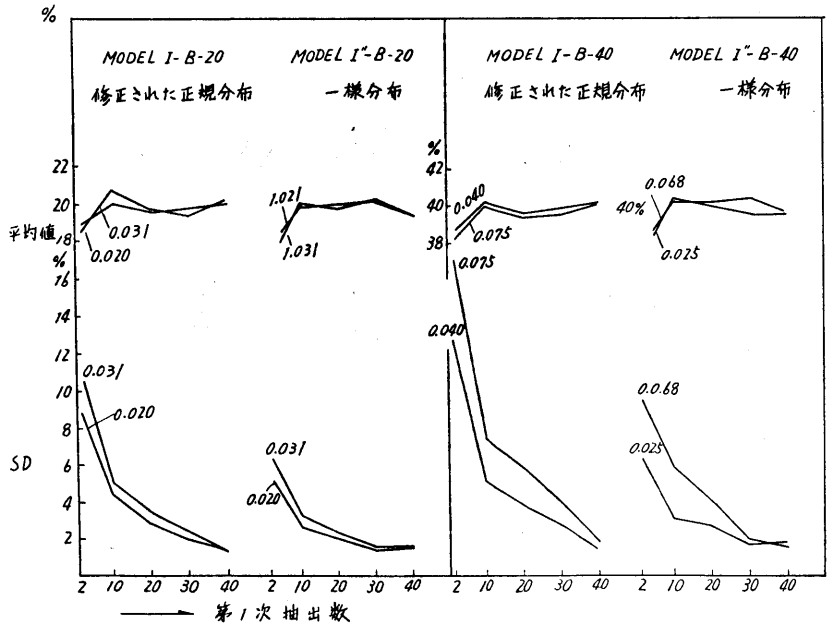


図 6. 集落人口を修正附正現分布および一様分布乱船で定めた母集団における標本 yes 反応率の期待値，SD の比較（方法A）

——これを MODEL-I'-L-20 と名付けることにする——との間の r-SD 曲線の比較を，S_b^2 が接近しているもの同志で，図 5 で行なった．図における点線は，MODEL-I-L-20，実線は，MODEL-I'-L-20 の r-SD の曲線を示す．Sb^2 が 0.16 と 0.16; 0.31 と 0.30 ; 0.00 と 0.00 の対同志について，共通している r の範囲で，傾向曲線は，比較的一致している．**2** で

集落数は，あまり大きな数を用意する必要はないと述べたが，もしこれの大きな母集団に当面したとしても，図5から前章で取扱った母集団で十分間に合うと思われる．

集落人口数の分布型 MODEL-I における各集落人口は，[200, 2000] で定義された修正附切断正規分布乱数であった．MODEL-I-S-20 および I-S-40 において，集落人口のみを [200, 2000] で定義される一様分布乱数に変更したもの——それを MODEL-I″-S-20 および I″-S-40 と名付ける——についての比較を $r-SD$ 曲線について行ったものを図6に示す．（図6において MODELI-B-20(40) I″-B-20(40)，の B は S に訂正されたし）ここで方法 A が用いられている．両者における推定値の不偏性は，あまり変らないが，後者は，前者より SD は，低く現われている。(しかし $r \geqq 20$ では±1%以内の違いである。) この理由は，人口分布の分散が，前者より，後者の方が小さいことにも原因があるように推測される．人口数 2000 を超えるものを 2 で説明したような修正を加えると，かなり特異な型をした分布型となる．（図1）

繰返し数に対する検討

文献 2) では，本報告の予備解析として同種の計算を繰返し数 $l=300$ で行なわれている，またそこでは，繰返し数を，100, 200, 300 とした場合の SD の変化に触れているが，0.1% 程度の違いしか見出せなかった．この経験をもとにして，本報告では，$l=100$ として，本解析を4で行なった．更に詳しい検討を図7で行なおう．

MODEL-I-L-40 の $S_b^2=0.042$ と $S_b^2=0.24$（＝全分散）の二つの凝似母集団を——前者は普遍的なもの，後者は，特種なものの例として——とり上げ，これらについて，それぞれ，繰返し数＝100 で SD を計算することを1試行とし2試行……50 試行まで行なったときの

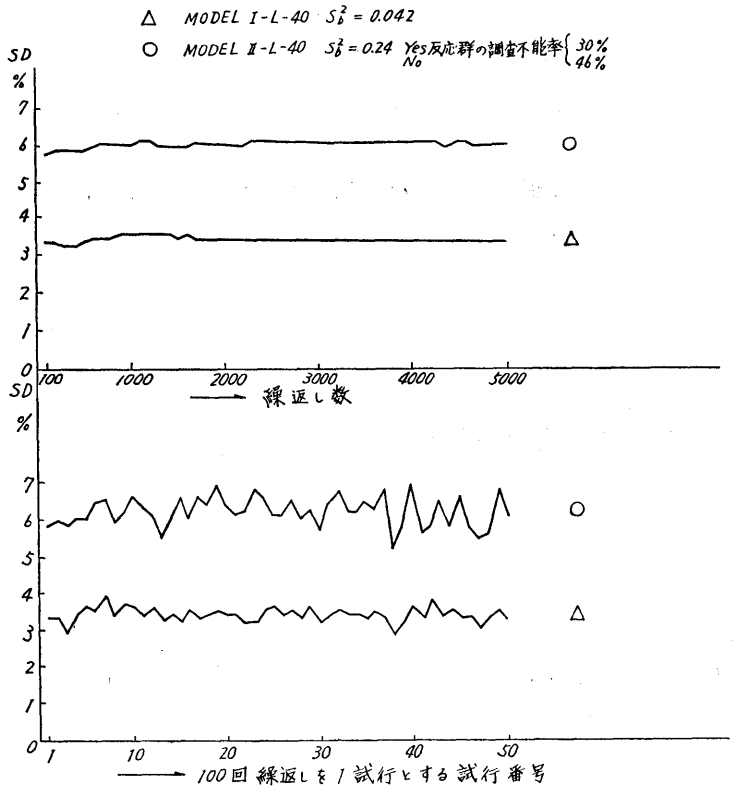


図 7. MODEL I-L-40 および II-L-40 に対する抽出法 A による 100 回繰返しによるシミュレーション SD の独立な 50 この評価とこれらの累積繰返しによる評価

各試行における SD を 図7 の下段に示す．ここで方法は A を用いた．SD は，約 3.3% ($S_b^2=0.04$ の場合) および，6% ($S_b^2=0.24$) の値を中心にして，各試行ごとに ±1%以内の範囲で変化している．この二つの場合について．それぞれの 50 コの SD についての標準偏差を計算すると，

	SD の SD*
$S_b^2=0.042$ の場合	0.0018
$S_b^2=0.240$ の場合	0.0040

である．この数値から判断する限り，4 で得られた結果は，きわめて信頼度が高いと言うことができるのであろう．

しかしながら．われわれが今迄得た r-SD 曲線をふり返ってみると，必ずしも，滑らかに書かれてはおらず，また或る構成要素——例えば伸縮率——の二つの接近した値を持つ擬似母集団から，それぞれ得られる r-SD 曲線は途中で交叉することもあった．全体からみると，これらは，数少なく起っているが．SD の SD が，今述べたように小さい値であるにしては何故なのであろうか？

憶測として考えられることは，擬似母集団を構成するときに用いた乱数と，擬似標本を抽出するときの乱数が同調したのであろうか？もしそうゆうことが起っていたとしても恐らく，4 で得られた諸結果は，1 で述べた目的には，十分な信頼度があると確信している．幸なことに，方法 D′ による SD の理論的評価は，可能なので，これとシミュレーション SD との比較を後で行ってみよう．

図7上段の二つの横ばい線は，試行回数1回ごとに，前に得られたものを累積して繰返し数とした場合の SD の値を示す．すなわち，$S_b^2=0.042$ および 0.240 両方とも，きわめて安定し，例えば，繰返し数を 100 としたものと，5000 にもふやしたものとでは，その差は，たった 0.1%，0.4% 程度である．

S_b^2	100	5,000	差
0.042	3.3	3.4	0.1
0.240	5.8	6.2	0.4

抽出方法についての検討

こゝでは引用される擬似母集団はすべて MODEL-I-L-40 (S_b^2 別) とする．

方法 A と B については，すでに 3 で検討されたがこゝでは5通りの方法の比較を検討してみよう．

方法 D では i 番集落に割り当てられる第2次標本数 s_i は $s_i=gS_i$ ($i=1\cdots\cdots, R$) としてあるので，総標本数は，前に述べたように確率変数となる．これの期待値が 1,000 となるよう．すはわち

$$1,000=E\{s_{i_1}+\cdots\cdots+s_{i_r}\}=gE\{S_{i_1}+\cdots\cdots+S_{i_r}\}$$

$$=g\cdot\frac{r}{R}\sum_{i=1}^{R}S_i=g\cdot\frac{r}{R}\cdot S$$

によって g を決めてから，方法 D による期得値，SD の r についての傾向曲線を図8下一左に書き，また上一右には方法 B によるものを示した（方法 B および D においては，推定式は集落には無関係に単純平均したものであるが，これを，3 の (6) 式で計算しなおすと，それぞれ方法，B′, D′ となる．これらの傾向曲線は図の下一中および下一右に掲げられてある．参

* 普通の意味の SD

考までに方法 A による傾向曲線は上一左に示しておいた．図8では第1次抽出数は，5から最高 250 の範囲の値が用意されている．

方法 B と D (B′ と D′) との比較を，偏り性・*SD* について行なうのは，厳密には無理があるが，抽出比 g について，前に述べた操作を加えた後の比較は，そう不合理ではないと思う，図8についての検討は次の通りである．

期待値に関する傾向は，5通りの方法とも，きわめて安定した水平線で，偏りの絶対値は幾分 r とともに減少している ($r \geqq 30$ において，すこしの例外を除けば1%以内)．5通りの方法中，強いて云えば方法 A が一番不偏性が強いことが観察される．

SD 曲線　方法 B と D とでは *SD* 曲線に，ほとんど差が見られないが，級間分散の小さ

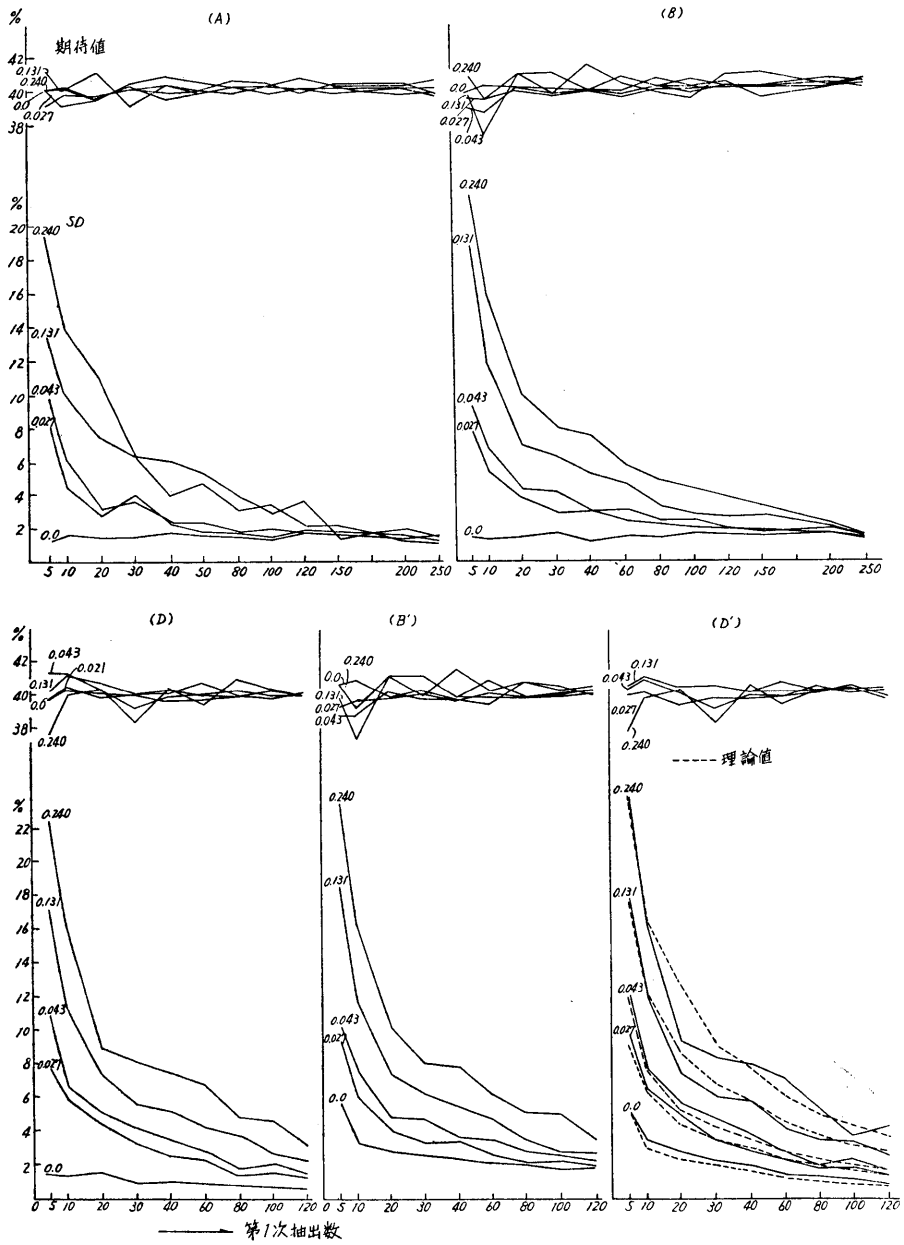


図 8. MODEL-I-L-40 (級頂分散別) に対する5通りの方法によって得られる期待値と SD の比較および方法 D′ における SD の理論的計算値とシュミレーション評価値の比較

い母集団では，方法 D の方が，$r \geqq 30$ でわずかばかり（±0.5% 以内）低下している．

方法 B′ と D′ についても，同様のことが見出されている．

方法 B と B′ では，S_b^2 が小さい程，推定式が単純集計である方法 B の方が，わずかばかりではあるが SD が低下している．

方法 D と D′ についても，同様に単純集計方式の D の方が SD が低い．

図8上一左と残りの4つを比較すると，これら5つのうち方法 A が一番好ましいように思われる．しかしながら好ましいと云っても他との差は僅少である．われわれの擬似母集団において，各集落の人口数，Yes 反応率は，ランダムに配列されてあった。もしそうでない特殊な配列をしている母集団に，現実に出くわした場合，方法Aは好ましくなくなるかもしれない．この点を考慮すると：

5通りの二段抽出法 A, B, D, B′, D′ による推定値の不偏性と標本抽出 SD は，ほゞ同程度であるから，抽出技術の難易等によって方法を決めればよく，特に各集落人口数および Yes 反応率がランダムに並んでいると思われるときは，方法 A が好ましい．また偏りを避けるための推定式 (6) は単純集計による推定式に較べ特に利点は見出せなかった．

SD の理論値とシミュレーション評価値

方法 D′ では，SD の理論値を求めることができる．図8下右の実線はシミュレーションによる値，点線は，理論値を示す．

理論式は，次式による．

$$SD = \sqrt{\frac{1}{S^2} V(X)} = \frac{R^2(R-r)}{R-1} \frac{\sigma_i^2}{r} + \frac{R}{r} \sum_j^R S_j \frac{S_j - s_j}{S_j - 1} \frac{\sigma_j^2}{p_j}$$

ただし

$$\sigma_i^2 = \frac{1}{R} \sum (P_i S_i)^2 - \left(\frac{\sum P_i S_i}{R} \right)^2$$

$$\sigma_j = P_j(1 - P_j)$$

理論値とシミュレーションによる SD との差は $S_b^2 = S_i^2$ の場合を除き，±1%以内である．もし $r \geqq 30$ に限定すれば，±0.5% 以内である．これによって，このオーダーのシミュレーション解析による SD の評価は十分に信頼にたるものと思われる．

謝　辞

本研究は，統計数理研究所，林　知己夫先生の示唆により始められたものであるが，その途上において，筆者らの最も苦慮した点の一つは，複雑多岐な，擬似二段抽出母集団のうちより，いかにしたら標本調査実務家が良く遭遇しうるようなものを作り出すかということである．林先生は，この点についても，われわれに有効な助言を与えて下さった．ここに感謝の意を表する．

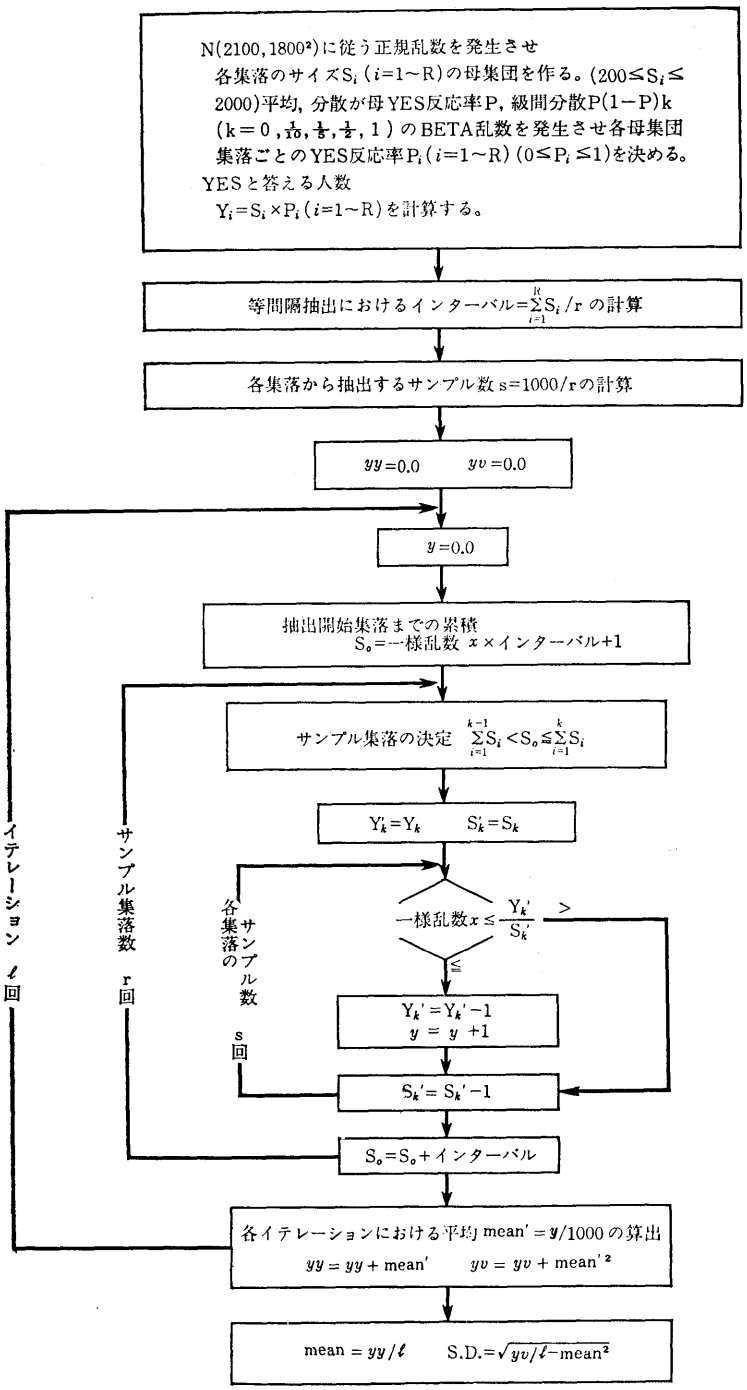
日本女子大学計算研究所
NHK 放送世論調査所

文　献

1) サンプリング研究会：サンプリングをめぐる諸問題 (5)，文研月報，1970，10.
2) サンプリング研究会：サンプリングをめぐる諸問題 (4)，文研月報，1970，9.

付録 サンプリング A のプログラム・フロー・チャート

（1段 proportional, 2段 equal）



R; 母集団集落数　r; サンプル集落数　ℓ; イテレーション回数　s; 集落ごとのサンプル数　S_i; 母集団集落 i の人口数　Y_i; 母集団集落 i の YES 反応数