

受入ID-1519990906B00047

収穫表及び材積表調製業務参考資料

経 営	28
測 定	8

W. G. コ ク ラ ン

標 本 調 査 法



02000-00044314-1

昭和30年9月

林 業 試 験 場 経 営 部

Sampling
Techniques

By

William G Cochran

Professor of Biostatistics

School of Hygiene and Public Health

The Johns Hopkins University

Copyright, 1953

New York. John Wiley & Sons, Inc

London. Chapman & Hall, Limited.

訳者まえがき

本編は、W. G. Cochran の "Sampling techniques" 1953 年版を訳したもので、着者が数年前 North Carolina State College の学生の講義用として用いたものである。本書の目的は実際の標本調査に役立つ様に抽出理論を分かり易く説明する事であつた。原著にはそのため各章の終りに練習問題及び参考文献が掲げられているが、本編に掲載出来なかつたのは残念である。本文は定理、証明、例の形式で書かれているので理解し易い。本編により標本調査法の理論と實際を理解し森林調査計画にあつて何等かの参考となれば幸である。

尚、§3.5, §3.9; §4.7; §5.8, §5.10, §5.12, §5.16, §5.17, §5.21; §6.8, §6.10; §7.4, §7.6, §7.7, §7.8, §7.9; §8.5, §8.6, §8.7, §8.9, §8.14; §9.5, §9.6, §9.12, §9.13; §10.4, §10.5, §10.8, §10.9; §11.9, §11.10, §11.11, §11.12, §11.13; §12.11, §13.4 は難解であれば最初はとはして進んでもよい。

昭和30年9月

測定研究室 西沢正久

目次

第1章 緒言	頁
§1.1 標本抽出法の利益	(1)
§1.2 標本調査に於ける主段階	(2)
§1.3 抽出理論の役割	(6)
§1.4 確率標本抽出	(10)
§1.5 偏倚とその影響	(11)
第2章 単純無作為抽出法	(16)
§2.1 単純無作為抽出	(16)
§2.2 定義及び記号	(17)
§2.3 推定量の特性	(18)
§2.4 推定量の分散	(19)
§2.5 有限母集団補正	(21)
§2.6 標本からの標準誤差の推定	(22)
§2.7 信頼限界	(24)
§2.8 正規性近似の妥当性	(25)
§2.9 推定分散への非正規性の影響	(32)
第3章 割合及び百分率の標本抽出法	(35)
§3.1 定性的性質	(35)
§3.2 標本推定量の分散	(35)
§3.3 標準誤差への P の影響	(39)
§3.4 二項分布	(41)
§3.5 p の一般分布	(42)
§3.6 信頼限界	(44)
§3.7 二つ以上の class への分類	(48)
§3.8 二つ以上の class である場合の信頼限界	(49)
§3.9 p の条件付分布	(51)
第4章 標本の大きさの推定	(54)
§4.1 假想的な例	(54)

§42 問題の分析	(35)
§43 正確度の分類	(56)
§44 割合に対する標本抽出に於ける n に対する公式	(58)
§45 連続資料での n に対する公式	(60)
§46 一つ以上の項目をもつ標本の大きさ	(62)
§47 二段抽出の <i>Stein</i> の方法	(64)
§48 標本の大きさの問題の一般解法の一つの試み	(66)
第5章 層化無作為抽出法	(69)
§51 緒言	(69)
§52 記号	(70)
§53 推定量の特性	(70)
§54 推定分散及び信頼限界	(76)
§55 最適割当て	(78)
§56 色々な費用での最適割当て	(79)
§57 層化無作為抽出と単純無作為抽出の相対的正確度	(81)
§58 最適からのかけはなれの影響	(83)
§59 以前の資料からの割当ての決定	(86)
§510 推定 S_y^2 に於ける誤差の影響	(88)
§511 数個の項目をもつ場合の割当て	(90)
§512 100%以上の抽出を必要とする割当て	(93)
§513 標本の大きさの推定	(94)
§514 割合に対する層化抽出	(97)
§515 層の作り方	(101)
§516 層化に対する基礎としての接近	(104)

§517 層化による利益の標本からの推定	(106)
§518 層の大きさに於ける誤差の影響	(111)
§519 層が前以て同一であると認め難い場合	(113)
§520 割堀抽出法	(114)
§521 層当り一つの単位での層化	(115)
§522 分析研究に於ける層化	(116)
第6章 比推定	(119)
§61 推定の方法	(119)
§62 比推定	(120)
§63 比推定の近似分散と偏倚	(122)
§64 推定分散	(127)
§65 標本の大きさ	(128)
§66 信頼限界	(129)
§67 単位当り平均値と比推定との比較	(131)
§68 比推定が最適である条件	(132)
§69 割合に対する標本抽出に於ける比推定	(135)
§610 比の分布の正規性への接近	(138)
§611 層化無作為抽出に於ける比推定	(140)
§612 組合せ比推定量	(142)
§613 比推定での最適割当て	(146)
第7章 回帰推定	(151)
§71 直線回帰推定	(151)
§72 大標本の理論	(152)
§73 基礎理論	(156)
§74 回帰推定の偏倚	(159)
§75 比推定と単位当り平均値との比較	(160)
§76 層化抽出に於ける回帰推定	(162)
§77 組合せ回帰推定	(165)
§78 推定分散	(169)
§79 回帰推定の二つの型の比較	(170)

第8章 組織的抽出法	(171)
§8.1 結 言	(171)
§8.2 他 の 見 方	(173)
§8.3 推定平均値の分散	(173)
§8.4 組織的抽出と層化無作為抽出との比較	(180)
§8.5 ランダムオーダーである母集団	(180)
§8.6 直線的傾向をもつ母集団	(183)
§8.7 端末補正	(185)
§8.8 周期的な変動をもつ母集団	(187)
§8.9 自己相関母集団	(187)
§8.10 自然母集団	(189)
§8.11 疑似周期の影響	(193)
§8.12 単一標本からの分散の推定	(194)
§8.13 層化組織抽出	(198)
§8.14 二次元に於ける組織抽出	(199)
§8.15 結 語	(201)
第9章 標本抽出単位の型	(203)
§9.1 最適単位	(203)
§9.2 簡単な例	(204)
§9.3 単位を比較する一般的方法	(206)
§9.4 調査資料から行う比較	(211)
§9.5 分散函数	(215)
§9.6 費用函数	(217)
§9.7 内部集落相関の項で表わした分散	(220)
§9.8 割合に対する集落抽出	(221)
§9.9 単位の大きさの尺度	(223)
§9.10 大きさに比例した確率で以ての標本抽出	(224)
§9.11 任意確率をもつ選択理論	(225)
§9.12 比推定との比較	(229)
§9.13 層化抽出への拡張	(232)

第10章 単位の大きさが等しい場合の副次抽出法	(233)
§10.1 結 言	(233)
§10.2 基礎理論	(235)
§10.3 他 の 副 次 抽 出 比 に 対 す る 分 散 の 予 測	(237)
§10.4 一般理論	(239)
§10.5 一般の場合に於ける分散の推定	(242)
§10.6 最適抽出比及び副次抽出比	(244)
§10.7 割合に対する副次抽出	(248)
§10.8 三段抽出法	(248)
§10.9 単位の層化抽出	(251)
第11章 単位の大きさが等しくない場合の副次抽出法	(253)
§11.1 結 言	(253)
§11.2 $n=1$ なるときの抽出方法	(255)
§11.3 推定大きさに比例した確率で以ての標本抽出	(259)
§11.4 $n=1$ なる場合に対する方法のとりまとめ	(259)
§11.5 $n>1$ なる場合の抽出法	(264)
§11.6 $n>1$ なる場合に対する方法のとりまとめ	(269)
§11.7 割合の推定	(270)
§11.8 暫定的註釈	(271)
§11.9 主な分散公式	(271)
§11.10 選択の最適確率	(276)
§11.11 偏倚対不偏推定量	(281)
§11.12 推定分散	(283)
§11.13 層化抽出への拡張	(287)
§11.14 ちすぶの註釈	(291)
第12章 二重抽出法	(293)
§12.1 方法の記述	(293)
§12.2 層化に対する二重抽出法	(294)
§12.3 最適割合	(297)
§12.4 層化に対する二重抽出に於ける推定分散	(299)

第 1 章 緒 言

Introduction

§ 1.1 標本抽出法の利益 *Advantages of the sampling method*
 我々知識、態度及び行動は大いに標本を基としている。これは日常生活及び科学的探究に於ても然りである。多くの事務を指導する立場にある人の意見は数年間彼がその立場で出会った一二の事によつてきまる事がよくある。外国で 10 日過ぎた旅行者がその国の住民が如何にして産業を復興し、政治形態を新たにし、予算を均衡し、ホテルの食事を改善したかを住民に知らせる本を面白く書いたとする。併し本当の意味では彼は経験が非常に少ない標本によつて結論してゐる点と彼の無知な点に気がついていないといふ点でその国で 20 年間生活と研究に身を捧げた政治学者とは違つてゐる。科学のあらゆる部門に於て我々が知識を押し進める現象の一断面より以上のものを研究する意欲に欠けてゐる。

最近迄如何にしたらよい標本を抽出する事が出来るかといふ事についての問題にはあまり注意が払われてゐなかつた。これは我々が標本抽出する資料がどの種類の標本でも殆んど同じ結果を与える様な一様なものであれば問題ではない。我々の健康状態がどうかといふ様な識別は研究室で血を数滴とつて検定される。この方法は体内を循環してゐる血は常によくまざつてをり、その一滴は他のものと同じ様子を知らせるものなどの仮定を基としてゐる。併し、資料が一様性からかけはなれてゐる場合は、此の様な事はごく普通の場合であるが、標本が得られる方法が批判され、信ずるに足る標本を得る方法についての研究が重要なものとなる。

之の本は良好な標本抽出法の背景を与える迄に今迄に組立てられてきた理論の大部分を含んでゐる。その応用に対して理論が組立てられた多くは、それについて情報が見られる集合が一都市の居住民とか工場の機械とか湖の魚とかいふ様に有限で限界があるものである。ある場合には完全調査又は集合体のセンサスを行つて

§ 12.5	回帰推定	(301)
§ 12.6	回帰での二重抽出対単一抽出	(304)
§ 12.7	回帰に対する二重抽出に於ける推定分散	(306)
§ 12.8	比推定	(307)
§ 12.9	同じ母集団のくりかえし抽出	(308)
§ 12.10	二つの事象での標本抽出	(310)
§ 12.11	二つ以上の継時事象の標本抽出	(313)
第 13 章	調査に於ける誤差の原因	(319)
§ 13.1	結 言	(319)
§ 13.2	無応答の影響	(320)
§ 13.3	無応答者向最適抽出	(325)
§ 13.4	無応答に対する他の方法	(330)
§ 13.5	測定誤差	(333)
§ 13.6	測定誤差に対する数学的模型	(335)
§ 13.7	一定偏倚の影響	(336)
§ 13.8	単位毎に独立である成分の影響	(338)
§ 13.9	違った単位上での誤差間の相関の影響	(341)
§ 13.10	相互費入副次標本	(342)
§ 13.11	結 び	(346)

正確な情報を得る事は容易である様に思われる。いつもセンサスばかりを取扱っていた人々は標本といふものには疑をもち、センサスの代りに標本を用いる事には気が進まない。此の態度は信用を失ふおそれはあるが、完全調査と比較して標本調査の主な利益を箇条書きする事にしよう。

1. 費用の節約

もしも資料が集合の小部分だけから得られるとすれば費用は完全調査が行われる場合よりも少なくて済む事は期得出来る。

2. スピードが大

同じ理由で資料は完全調査の場合よりも迅速に集める事が出来る。之は情報が急いで必要である時には大切な考え方である。

3. 視野が広い。

研究のある型に於ては資料を集める際に高度に熟練した人とか特別な装置とかの如く利用に対して限定されたものを用いなければならぬ。その事は完全調査では不可能である。

標本抽出か又は全くそうでないかによって情報を得る場合の選択がきまる。この様に標本抽出による調査は得る事が出来る情報の型に関して視野が広く又自由性に富んでゐる。一方もしも母集団の小部分又は一断片に対して情報が必要であればその際は完全調査が最もよい解答を与えてくれる事が分る。

4. 正確度が大

高度の特性をもつ人を調べる事が出来るし、又調査する人に完全な訓練をする事が出来るから標本は実際には取扱う事が容易である種類の完全調査よりも正確な結果を与える。

§ 1.2. 標本調査に於ける主段階 *The principal steps in a sample survey.* 理論が標本調査に如何なる役割を演ずるかを論議する始めとして簡単に調査の計画及び実行に含まれる段階を述べる事は便利であらう。調査はそれ等の複雑性によつて大いに違つてくる。5000のカードから標本を抽出するのにきれいに重ねて順序よく数えるのは容易な仕事である。

森林の中を通つてゐる川で交通している地方の住民の場合、地図がない場所、ノボの異つた方言を話す場所、住民が外人をいふ嫌う場所、物好きな他の人を非常に嫌う場所とかの標本抽出は自分から違つてくる。一つの問題で次の分岐問題は他の問題では些細な事か又は何ら問題にならない事であるかもしれない。

調査での主段階は幾分任意的であるが次の7つの主題にまとめられる。

1. 調査対象の叙述

明解な対象の叙述が最も助けとなるのである。これなくしては複雑な調査では計画の詳細についてあまり考えすぎて対象を忘れたり又対象と矛盾した決定を行う様な事がごく普通になる。

2. 標本抽出すべき母集団の決定

母集団なる言葉はそれから標本が選ばれる集合体を示すのに用いられる。母集団の定義が問題ではない場合がある。例えば電球の寿命の平均長さを推定するのに一箱の電球を抽出する場合の如きである。

一方農場の母集団を標本抽出する際には規則は農場に限つてしまい境界線如何という様な場合も起つてくる。これ等の規則は実際に使用出来るものでなければならぬ。即ち調査する人が島で母集団に属するかどうか疑わしい様な時に決定を躊躇する事なしに行い得るものでなければならぬ。

出来れば何時でも抽出すべき母集団は明らかにそれについて情報が望まれる母集団と一致すべきである。

時にはこの必要性はあまり困難であるため正しいが誤つてゐるか判断される。資料の蒐集にあつて複雑な測定問題をもちた様な新しい研究分野に於てはたとえこれはそれについての情報が真に望まれくゐるより広汎な母集団をなるべく抽出が完全で容易な母集団を選んで調査のこの点についての原因をつきとめるために決定される事がある。この様な場合は我々はより広汎な母集団に対する関係事項が調査出来るかどうかを示す助け

となる二つの母集団について比較出来る情報を集めなければならない。

III 集めらるべき資料の決定

すべての資料は調査目的に關係があり本質的な資料はオミットしてはならない事は明らかである。しばしばあまり多く資料をあつめすぎてその中のいくらかが後では決して検定をされない様な場合がある。

IV 測定方法

必要な種類の資料が決定した時には用いらるべき測定法に關しての選択となる。例えば人間の健康状態についての資料は本人が述べる事又は多少其医学的な検定から得られる。人間社会では問題を廣向する態度及び順序がその結果に大なる差を生ぜしめる。Payne (1951) を参照せられよ。

V 標本抽出単位を選択

標本を抽出するには先づ第一に母集団はある方法で標本抽出単位 (sampling unit) 又は単位 (unit) と呼ばれる部分に分割されねはならない。標本抽出単位はすべて母集団全体を構成しなければならず、母集団に於ける各要素は一ついわば一つの単位の中に居するという意味で重複してはならない。時には単位が一つの電球である場合には単位は電球の母集団を構成する様に明らかな場合がある。時には単位を選択は相当考慮して定めねはならぬ場合がある。町に住んでゐる人を抽出する際には、単位は個人、家族、同じ町のブロックに住んでゐる人でなければならぬ。農作物の標本抽出に於ては単位は形及び次元が我々の自由に定め得る土地の面積である。標本抽出単位の完全なリスト作成は標本抽出単位は時に枠 (frame) と呼ばれる。實際的には主たる問題の一つである。時には枠は湖の魚の如く作る事が不可能な場合がある。

VI 標本の選択

標本が選ばれる方法は非常に多い。又選択は標本の大きさにつ

いての決定の問題を包含してゐる。(抽出数) それには標本がそれをとるのに許容し得る費用の中で得られる様に調査費用の前以ての推定を必要とする。

VII 実際の仕事の組織

大抵の調査に於ては多くの事務管理の問題を含んでいる。これにたづさわる人は調査の目的及び用いらるべき測定法に熟練していなければならず又適當にそれ等の仕事を監督しなければならない。手紙を出して返答を求める際その返答を早くチェックする方法は無価値である。

無応答のものを取扱う計画をしなければならない。即ち無応答は標本に属する単位のあるものから情報を得る事に於て調査する人が失敗したものである。

VIII 資料の取纏めと分析

第一段階は記録の誤りを正す目的のために又は明らかに誤りのある資料を少くとも除く目的のために調査表を作成する事である。表示法につづく決定はある問題に対する回答がある応答者によつてオミットされるか又は作成過程に於て除かれねはならぬ場合に必要となる。その後推定量を導く表が作成される。推定の別の違った方法が同じ資料に対して利用される。

IX 将来の調査に対して利益となる情報

我々が最初に母集団について得た情報が多ければ多い程それだけ正確な推定量を与える標本を計画する事は容易となる。完全な標本は平均値又は標準偏差及び測定の変動の性質とか資料を得るために要した費用とかについて提供された資料を通じて将来の抽出計画を改善する本質的な手引となる。抽出計画の実行は順備さえよければその型の情報を集めたり記録したりする時に迅速にやれる。完全な標本が将来の抽出を容易にする点は又大切な事である。複雑な調査で計画された事は正確には行われぬ。すぐれた抽出計画者は自分の実行での誤りを認め、将来の調査にはその様な事が起らない様にする事を知つてゐる。

§ 1.3. 抽出理論の役割 *The role of sampling theory* :—

標本調査に於ける段階のこのリストは標本抽出が実際的な仕事である即ち技術の色々な異った型を要求するものであるという事を強調せんがために与えたものである。

此の段階のあるもの即ち母集団の決定とか、集めらるべき資料の決定とか測定方法とか又実際の仕事での組織という段階はこれ等に於て標本抽出理論は主たる役割を演ずるものである。これ等についてはこの本では最早論議しないが、その重要性については認識すべきである。

標本抽出は実行のあらゆる部面に於て細心の注意を要する。ある一局面で仕事が貪弱であれば他のあらゆるものが立派に行われていても調査全体をたいなしにするものである。

抽出理論の目的はより有効な標本抽出を行う事である。最少可能な費用で我々の目的に対して十分な正確度を与える標本の選択法推定法を進展させようとする。最小費用で特別な正確度というこの原則は、理論を示す際にくりかえし起ってくる。この原則を応用するためには我々が考えてゐる如何なる抽出方法に対しては期待されるべき正確度と費用を予測しなければならない。

正確度に関しては、如何に大きな誤差がある特別な立場での推定に於て表われるであらうかについては前以て正確に知る事は出来ない。何故ならばこれは母集団についての真の値の知識を必要とするからである。その代り標本抽出法の正確度はむしろその抽出法が同じ母集団に対してくりかえし適用されたならば推定量に対する一般的な分散分布を検討する事によつて判断される。

勿論、この事が正確度が統計理論によつて判断される標準の技術である。

一つの更に進んだ簡易化が導かれる。実際に於てごく普通の大きさの標本で以ての標本推定量は近似的に正規に分布すると考えられる。その結果推定値の抽出分散が換言すればその正確度の尺度を与えるために用いられる。

大部分の理論は色々な方法によつて得られた推定量の抽出分散の公式の計算を取扱つてゐる。無限母集団に対する抽出法の研究は古くから行われてをり充分基礎づけられてゐる。併し特に標本調査に適応するための理論の進展はごく最近の事である。この本の殆んどすべての関係事項は20年以内に研究されたものでありその大部分が10年以内のものである。標本調査理論の最初の利戟は情報を得る手段として標本調査が多く用いられる様になつてきたからであつた。標本調査理論における大部分は實際的に調査の指導に従事した人々によつて行われた。それに引きかえ理論の進展は抽出方法の範囲及び利用性を増した。そして調査に實際的に使用する事により更に進んだ成長をうながした。

標本調査理論と昔の標本抽出理論との間の一つの相違点は我々が実際の調査で取扱われはならぬ母集団は単位の有限の数を含んでゐるという事である。理論を証明するために用いた方法も違つてゐるし、その結果は標本抽出が無限母集団の代りに有限母集団であるためにかなり複雑となる。

有限及び無限母集団に対するこれ等の差は實際目的にはさして重要なものではない。応用面における大部分の場合に起る如く標本の大きさが母集団の大きさに比して相対的に小さい場合は無限母集団から導かれた結果で充分適當である。一般に有限母集団に対する結果が之の本では述べられる。ある困難な問題に対しては理論を簡単にするため無限母集団に対するものが用いられる事がある。

第 1 表

七

表

43.1

D. f	七の値がこれより大きくなる確率、符号は無視							D. f
	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	.986	1.061	1.386	1.888	2.920	4.303	6.965	9.925
3	.975	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.964	.941	1.190	1.533	2.132	2.776	3.747	4.604
5	.952	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.941	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.931	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.921	.887	1.108	1.397	1.860	2.306	2.896	3.355
9	.912	.879	1.100	1.383	1.833	2.262	2.821	3.250
10	.904	.871	1.093	1.372	1.812	2.228	2.764	3.169
11	.897	.864	1.088	1.363	1.796	2.201	2.718	3.106
12	.891	.858	1.083	1.356	1.782	2.179	2.681	3.055
13	.885	.852	1.079	1.350	1.771	2.160	2.650	3.012
14	.880	.847	1.076	1.345	1.761	2.145	2.624	2.977
15	.875	.842	1.074	1.341	1.753	2.131	2.602	2.947
16	.870	.837	1.071	1.337	1.746	2.120	2.583	2.921
17	.865	.832	1.069	1.333	1.740	2.110	2.567	2.898
18	.861	.828	1.067	1.330	1.734	2.101	2.552	2.878
19	.857	.824	1.066	1.328	1.729	2.093	2.539	2.861
20	.853	.820	1.064	1.325	1.725	2.086	2.528	2.845

21	.850	.817	1.063	1.323	1.721	2.080	2.518	2.831	21
22	.846	.813	1.061	1.321	1.717	2.074	2.508	2.819	22
23	.843	.810	1.060	1.319	1.714	2.069	2.500	2.807	23
24	.840	.807	1.059	1.318	1.711	2.064	2.492	2.797	24
25	.837	.804	1.058	1.316	1.708	2.060	2.485	2.787	25
26	.834	.801	1.058	1.315	1.706	2.056	2.479	2.779	26
27	.831	.798	1.057	1.314	1.703	2.052	2.473	2.771	27
28	.828	.795	1.056	1.313	1.701	2.048	2.467	2.763	28
29	.825	.792	1.055	1.311	1.699	2.045	2.462	2.756	29
30	.822	.789	1.055	1.310	1.697	2.042	2.457	2.750	30
35	.810	.777	1.055			2.030	2.434	2.724	35
40	.800	.767	1.055			2.021	2.404	2.704	40
45	.791	.758	1.055			2.014	2.690	2.690	45
50	.783	.750	1.055			2.008	2.678	2.678	50
60	.772	.739	1.055			2.000	2.660	2.660	60
70	.763	.730	1.055			1.994	2.648	2.648	70
80	.755	.722	1.055			1.990	2.638	2.638	80
90	.748	.715	1.055			1.987	2.632	2.632	90
100	.742	.709	1.055			1.984	2.626	2.626	100
125	.730	.697	1.055			1.979	2.616	2.616	125
150	.720	.687	1.055			1.976	2.609	2.609	150
200	.704	.671	1.055			1.972	2.601	2.601	200
300	.688	.655	1.055			1.968	2.592	2.592	300
400	.680	.647	1.055			1.966	2.588	2.588	400
500	.674	.641	1.055			1.965	2.586	2.586	500
1000	.667	.634	1.055			1.962	2.581	2.581	1000
∞	.667	.634	1.055	1.281	1.644	1.959	2.575	2.575	∞

(9)

§ 1.4 確率標本抽出 *Probability sampling*:-

それに対して理論が展開されてきたすべての抽出方法は、通常次の如き数学的特性をとつてゐる。

i 我々は次の明らかな標本の組を定義出来る。即ち $S_1, S_2, \dots, S_n$ 。これがもしも特別な母集団に適用されるとすれば、その方法は選択可能である。この事は我々がどの標本抽出単位が $S_1, S_2, \dots$ 等に属するかを正確に云ふ事が出来る事を意味する。例えば母集団がノからゴ迄の数の六つの数を含んでゐると假定しよう。大ききなる標本を送ぶ方法は次の三つのものと考えられる。

即ち $S_1 \sim (1, 4); S_2 \sim (2, 5); S_3 \sim (3, 6)$ 。大ききなるすべての可能な標本を示す必要はない事に注意せよ。

ii 各可能な標本 S_i は、それを既知の選択確率 π_i に割当てた。

iii そこで我々は、各 S_i が送ばれる使用確率 π_i をうける方法によつて S_i の一つを送ぶ。

例でいへばゴつの標本に等しい標本を割あてねばならぬ。従つて抽出そのものはノからゴ迄の乱数を送ぶ事により得られる。

もしもその数が g であれば S_g がとられる標本である。

iv 標本から推定量を計算する方法が述べられねばならず特別な標本に対して一様な推定量を導かねばならない。例えば推定量が標本に於ける個々の単位の測定値の平均値であるといふ事を述べるが如きである。これ等の特性を満足する如何なる抽出方法に対して、我々はもしも同じ母集団にくりかえし抽出を適用すれば一般化される推定量の度数分布を計算する立場となる。何故ならば我々はどの位特別な標本 S_i が送ばれるか又 S_i に於ける資料から推定量を如何にして計算するかを正確に知つてゐるからである。

従つて我々は理論の展開の詳細がたとえ複雑であつてもこの型の方法に対しての抽出理論を展開出来る事は明らかである。

確率標本抽出なる言葉はこの型の方法に関連するものである。

この言葉は未だ標準の定義はつけられてはいないし、ある人はそれをより抱束された意味で用いてゐる。この言葉の主目的はこの種の標本抽出を標本がある人によつて母集団の特に代表的なもの又は抽出に便利な様に単位を限る様な有意選択と明らかに区別する事である。有意選択は標本が小さい時は良好な結果を与えるが、無作為選択の要素が含まれていないから理論的な進展はあり得ない。

実際には我々は上に概略述べた様に S_i 及び π_i をすべて書く事によつて標本を抽出する事はめつたにない。これは標本抽出方法がノの億もの膨大な標本をとつてゐる大母集団では非常に面倒な事である。その抽出法は個々の単位に含まれてゐるものの確率を分類し、一つ一つ又は群として望まれる大きさの標本及び型が組立てられる迄抽出を行ふのが普通である。理論の目的に対してはもしも我々が望ましい時又無制限に時間があるときには S_i 及び π_i をすべて書く事が出来るであらうといふ事を知る事で充分である。

§ 1.5 偏倚とその影響 *Bias and its effects*:-

理論を示す際に簡単なため、巻目の *unit* での測定値 y_i はその *unit* に対して正しい値であると假定する。

測定誤差は無視する。この假定は勿論現実的ではない。そして第ノ3章で標準の結果の測定誤差を検定する。ある型の誤差に対しては、標準の結果はたゞ主な変化のみで妥当である。他の型の誤差ではより多くのひどい変化を必要とする。

併し、偏倚の影響をこの節で論じよう。何故なら偏倚推定量 (*biased estimate*) をわざと使用するときにはしばしば標本抽出調査で利益をもたらす事があるからである。

抽出方法はもしもそれが生ずる推定量の度数分布の平均値が正確に推定されてゐる母集団の特性に等しい時に不偏であるといわれる。前節の記号で π_i をして 選択確率 π_i をもつ標本 $S_i (i = 1, 2, \dots, n)$ によつて与えられた推定量とし θ_i をして推定

さるべき母集団値としよう。

もしも

$$\sum_{i=1}^n \pi_i \Sigma_i = 0$$

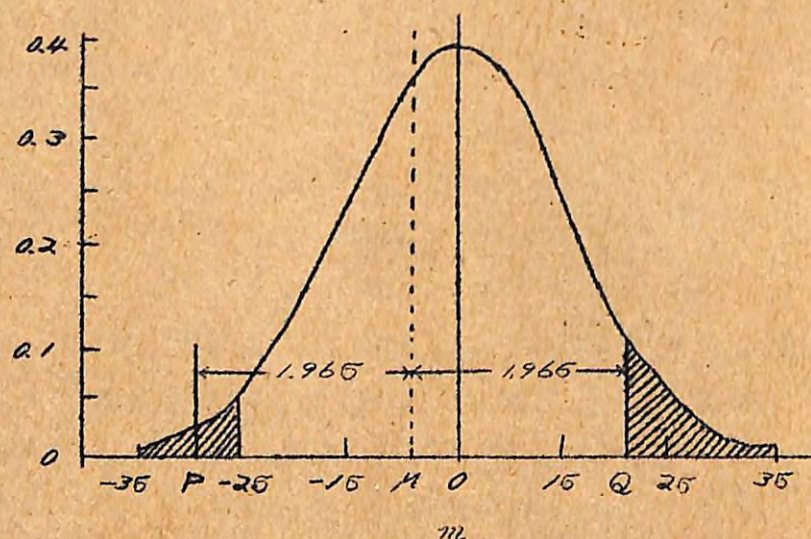
であれば抽出方法は不偏である。

もしも二つの量が等しくなければその差をその抽出方法での偏倚と呼ぶ 即ち

$$\text{偏倚} = \sum_{i=1}^n \pi_i \Sigma_i - 0$$

偏倚の影響を検定するため推定量区は 1/1 図に示す様に眞の母平均値 μ から B なる距離にある平均値 m のまわりに正規に分布してゐると假定しよう。偏倚の量は $B = m - \mu$ である。

1/1 図 推定の誤差への偏倚の影響



我々は偏倚が存在する事を知らないと假定しよう。我々は推定量の度数分布の標準偏差 σ を計算する——勿論これは眞値 μ まわりでなく度数分布の平均値 m まわりの標準偏差である。推定量の正確度についての論述として我々は推定量区が $\pm 1.96\sigma$ 以上の誤差を生ずる確率は 0.05 である事を知つてゐる。

我々は如何にして偏倚の存在がこの確率をゆがめるかを考えよう。このために推定量が $\pm 1.96\sigma$ 以上の誤差を生ずる確率を計算する。此処に誤差は眞値 μ から測定される。分布の二つの端 (tail) は別々に検定されねばならない。上 (右) の端に対して $+1.96\sigma$ 以上の誤差を生ずる確率は 1/1 図の Q 以上のハッチをつけた面積である。この面積は次の式で与えられる。

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{\mu+1.96\sigma}^{\infty} e^{-\frac{1}{2}(Z-m)^2/\sigma^2} dZ$$

$Z - m = 0$ とおく 0 に対する積分の範囲の下限は

$$\frac{\mu - m}{\sigma} + 1.96 = 1.96 - \frac{B}{\sigma}$$

$$\text{従つて面積は } \frac{1}{\sqrt{2\pi}} \int_{1.96 - (B/\sigma)}^{\infty} e^{-\frac{t^2}{2}} dt$$

同様に下 (左) の端即ち P 以上下のハッチした面積は

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1.96 - (B/\sigma)} e^{-\frac{t^2}{2}} dt$$

積分の形からデイスターバンスの量は偏倚と標準偏差の比のみによる事は明らかである。その結果は 1/1 表に示す。

1/1 表 $\pm 1.96\sigma$ より大きくなる誤差の確率への偏倚 B の影響

B/σ	誤差の確率		合 計
	$< -1.96\sigma$	$> 1.96\sigma$	
0.02	0.0238	0.0282	0.0500
0.04	0.0228	0.0274	0.0502
0.06	0.0217	0.0267	0.0504
0.08	0.0207	0.0261	0.0508
0.10	0.0197	0.0254	0.0511
0.20	0.0154	0.0232	0.0548
0.40	0.0091	0.0214	0.0685
0.60	0.0052	0.0201	0.0921
0.80	0.0029	0.0183	0.1259
1.00	0.0015	0.0165	0.1700
1.50	0.0003	0.0328	0.3231

1%以上の誤差の全確率に対しては偏倚は標準偏差の10分の1以下であれば大して影響はない。

この点では全確率は我々が考えている0.05の代りに0.05/10である。偏倚が更に増すとデイスターバンスは更にひどくなる。

$B=0$ ($B/\sigma=100$) 誤差の全確率は假定した値(0.05)の3倍以上即ち0.17である。

二つの端は別個に影響されてゐる。正の偏倚ではこの例では1%以上の誤差を低く見積る確率は假定した0.025から $B=0$ のときには無視出来る程急速に小さくなって来る。それに対応する過大推定の量は大きなものとなる。大抵の応用例では全体の誤差に最初に関心をもち、場合によつては一方の誤差に特に興味をもち、場合がある。

実行上の規則として推定量の正確度に対する偏倚の影響は、もしも偏倚が推定量の1/10以下であれば無視出来る。

もしも我々がそれに対して $B/\sigma < 0.1$ という事が分つた偏倚のある推定方法を行ふとすればその場合偏倚は大して不利益と与えていないといふ事が出来る。併し偏倚のある方法は注意して使用しなければならぬ。標本が母集団から一挙を通じて毎月抽出され月々の推定量が偏倚のある推定方法で行われと假定しよう。ノミの推定量の算術平均は月平均を得るため次々に計算される。もしも母集団が除々に変化しておればノミの推定量のうける偏倚は同じ符号をもち同じ大きさであるだらうと思われる。従つて月平均の偏倚は一月の数字での偏倚とほゞ同じである。もしも月々の標本が独立に抽出されるとすれば年平均推定量の標準誤差は月推定量の標準誤差の約 $1/\sqrt{12}$ 倍であらう。従つて年平均の標準誤差の偏倚との比は月の数字のそれの約 $\sqrt{12}$ 倍である。この大きくなつた比は無視出来ない。標本推定量が後の目的のために平均されるすべての方法を前以て知る事はいつも困難であるから、偏倚した推定量の使用は標準誤差に対する偏倚の比が非常に小さくなくてはさけるべきである。

偏倚を除くために非常に苦心してとられた時でも推定量には予期しない推定量が存在する。標本から得られた時推定量の標準偏差は偏倚の影響をうけていないからその *accuracy* よりもむしろ推定量の *precision* を測定する時にこの標準偏差を述べる事がより好ましい事である。 *accuracy* は通常真の平均値 μ からの偏差の大きさに関係し *precision* は標本抽出をくりかえし適用する事によつて得られた平均値 $\bar{x}$ からの偏差の大きさに関係する。

第 2 章 単純無作為抽出法

Simple random sampling

§2.1 単純無作為抽出 Simple random sampling

標本調査は有限の数 N なる *unit* を含む母集団から抽出された標本を取扱う。

N なる母集団より n なる標本を抽出する組合せは次の公式により
れる。

$$\binom{N}{n} = {}_N C_n = \frac{N!}{n!(N-n)!}$$

例えばもしも母集団が A, B, C, D, E なる 5 つの *unit* を含むとすれば大いさ 3 なる次の如き 10 の異なつた標本が出来る。

ABC, ABD, ABE, ACD, ACE

ADE, BCD, BCE, BDE, CDE

六つの標本 ABC, ACB, BAC, CAB, BCA 及び CB A は同じと考えられるから文字が標本に於て表われる順序には注意を払ふ必要はない。

Simple random sampling とは ${}_N C_n$ の中どの一つでもか送ばれる確率がすべて等しい様に N から n なる *unit* を選ぶ方法である。

単に random sampling と呼ばれる事もあるし、ある人は *unrestricted random sampling* (不抱束無作為抽出) なる言葉を推奨している。

実際には *simple random sampling* は *unit* 毎に抽出される。母集団に於ける *unit* の数は 1 から N 迄であり 1 から N 迄の間か乱数表又は書いたものをツボに入れてよくかきまぜる事により抽出される。抽出された数をどう *unit* が標本を構成する。抽出過程に於て前に抽出されなかつた n の数に対しては等しい機会を与える ${}_N C_n$ のすべての可能な標本は等しい機会をもつ事は容易に証明される。1 つの数カツボから抽出された時、これはもとにもとさない。従つて抽出は *without replacement*

ment ともいわれる。同様に乱数表を用いたときは前に引かれた数は二度と引かない。

単純無作為抽出法は抽出理論を導く場合に最も役に立つものである。

§2.2 定義及び記号 Definitions & Notations

標本調査に於て我々は *sample* に入つてくるあらゆる *unit* を測定し記録しようとするある特性を決定する。

unit のこれらの特性は *characteristics* 又は単に項目 *item* に関するものである。

母集団を構成する N *unit* に対する特性に対して得られる値は $y_1, y_2, \dots, y_N$ と書きそれに対応する標本 *unit* に対する値は $y_1, y_2, \dots, y_n$ とかく。

大文字は母集団に関するもので小文字は標本に関するものである。総計及び平均値に対して次の如く定義する。

母 集 団 値

標 本 値

$$\text{総 計 } Y = y_1 + y_2 + \dots + y_N = \sum_{i=1}^N y_i \quad y = y_1 + y_2 + \dots + y_n = \sum_{i=1}^n y_i \quad (2.1)$$

$$\text{平均値 } \bar{Y} = \frac{y_1 + y_2 + \dots + y_N}{N} = \frac{Y}{N} \quad \bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{y}{n} \quad (2.2)$$

標本抽出 は 多くの異つた目的に対して試みられるが主として次の三つの母集団の特性に集中される。

- (1) 母集団での全 *unit* 上のある特性に対する値の総計 Y (例えばある地方での小麦のエーカーでの総数)
- (2) *unit* 当りの平均値 $\bar{Y}$ (例えば畠当りの小麦のエーカーでの平均数)
- (3) あるきまつた級に入つてくる *unit* の割合 即ち百分率 (例えば小麦が生えていない畠の百分率)

次に母集団総計及び平均値の推定につき考える。記号 $\bar{y}$ は標本から得られた母集団特性の推定量を表わす。推定量の或る簡単な型のみを述べると

推定量

母集団平均値

$$\hat{\bar{y}} = \bar{y} = \text{標本平均}$$

母集団総計

$$\hat{Y} = N\bar{y} = \frac{N}{n}y$$

因子 $\frac{N}{n}$ は拡大係数 (raising factor) とよばれその逆数 $\frac{n}{N}$ は標本の大きさと母集団の大きさとの比であり抽出比 (sampling ratio) と呼ばれる。

§2.3 推定量の特性. Properties of the estimates:-

標本から得られた推定量の正確度は、それが標本資料から計算された方法及び抽出計画の両方による。此处では標本平均の正確度即ち単純無作為抽出の正確度についてのみ述べる。

公式を研究する前に抽出の方法及び公式が作られた推定方法を確実に把握しなければならない。推定方法は推定量が正確に $n = N$ 即ち標本が全母集団から成立つてゐる時正確に母集団の値と一致してゐるならば一致性 (Consistent) があるといわれる。

単純無作為抽出では $\bar{y}$ 及び $N\bar{y}$ は夫々母平均及び母集団総計の一致推定量 (Consistent estimate) である。

一致性は推定量の特性として望ましいものである。それに反して不一致推定量 (inconsistent estimate) は有用でないかといへばそうでもない。その理由は n が N に比して小さい場合は満足な正確度を与えるからその有用性は之の場合のみに限られる。総計理論では一致性に關しては主として無限母集団に対して論議されてきた。

定義としては

もしも $\epsilon > 0$ で

$$\lim_{n \rightarrow \infty} \Pr \{ |\bar{y} - \bar{Y}| > \epsilon \} = 0$$

であれば $\bar{y}$ は $\bar{Y}$ の一致推定量である。

この定義は有限母集団にも適用出来る方法がある。併し乍ら一致性の考え方は次の説明に対しては大して重要なものではない。

推定の方法はもしも与えられた大いさなるすべての可能な標本の上にとられた推定量の平均値が真の母集団値と正確に等しければ不偏 (unbiased) であるといわれる。 $\bar{y}$ が simple random sampling で不偏であるかどうかを検定するためにはすべての C_n なる標本に対して $\bar{y}$ の値を計算し推定量の平均値を求める。之が母平均値と等しければ不偏であるといわれる。 E なる記号はすべての可能な標本の上のこの平均を表わす。

定理 2.1

標本平均値 $\bar{y}$ は母平均値 $\bar{Y}$ の不偏推定量である。

(証明)

定義により

$$E(\bar{y}) = \frac{\sum \bar{y}}{nC_n} = \frac{\sum (y_1 + y_2 + \dots + y_n)}{n \frac{N!}{n!(N-n)!}} \quad (2.2)$$

分子の和の y_i が標本に於て表われる数は

$$(N-1)C_{(n-1)} = \frac{(N-1)!}{(n-1)!(N-n)!}$$

$$\text{従つて } \sum (y_1 + y_2 + \dots + y_n) = \frac{(N-1)!}{(n-1)!(N-n)!} (y_1 + y_2 + \dots + y_n)$$

従つて (2.3) から

$$\begin{aligned} E(\bar{y}) &= \frac{(N-1)!}{(n-1)!(N-n)!} \cdot \frac{n!(N-n)!}{nN!} (y_1 + y_2 + \dots + y_n) \\ &= \frac{y_1 + y_2 + \dots + y_n}{N} = \bar{Y} \end{aligned} \quad (2.4)$$

系

$\hat{Y} = N\bar{y}$ は母集団総計 Y の不偏推定量である。

定理 2.1 の証明より

$$E(\hat{Y}) = E(y_1 + y_2 + \dots + y_n) = E(N\bar{y}) = NE(\bar{y}) = Y$$

従つて

$$E(y_1 + y_2 + y_3 + \dots + y_n) \text{ は } (y_1 + y_2 + \dots + y_n) \text{ の } n \text{ 倍である} \quad (2.5)$$

2.4 推定量の分散 Variances of the estimates:-

有限母集団に於ける y_i の分散は次の様に定義される。

$$\sigma^2 = \frac{\sum (y_i - \bar{Y})^2}{N} \quad (2.6)$$

一致性並びに不偏であるためには上式は次の如くあらねばならない。

$$S^2 = \frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N-1} \quad (2.7)$$

分母が N でなく $(N-1)$ である事は注意を要する。

今我々は $\bar{y}$ の分散を考える。これはすべて nC_n なる標本でとられた $E(\bar{y} - \bar{y})^2$ を意味する。

定理 2.2

単純無作為標本からの平均値 $\bar{y}$ の分散は

$$V(\bar{y}) = E(\bar{y} - \bar{y})^2 = \frac{S^2 (N-n)}{n} \quad (2.8)$$

(証明)

$$n(\bar{y} - \bar{y}) = (y_1 - \bar{y}) + (y_2 - \bar{y}) + \cdots + (y_n - \bar{y}) \quad (2.9)$$

(2.5) を用いて

$$E\{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\} = \frac{n}{N} \{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\} \quad (2.10)$$

又

$$\begin{aligned} E\{(y_1 - \bar{y})(y_2 - \bar{y}) + (y_1 - \bar{y})(y_3 - \bar{y}) + \cdots + (y_{n-1} - \bar{y})(y_n - \bar{y})\} \\ = \frac{n(n-1)}{N(N-1)} \{(y_1 - \bar{y})(y_2 - \bar{y}) + (y_1 - \bar{y})(y_3 - \bar{y}) + \cdots \\ + (y_{n-1} - \bar{y})(y_n - \bar{y})\} \end{aligned} \quad (2.11)$$

(2.11) に於ける積和は 夫々標本及び母集団での unit のすべての対に關するものである。左辺の和は $n(n-1)/2$ 項を含む。右辺の和は $N(N-1)/2$ 項を含む。

(2.9) を平方してすべての単純無作為標本上で平均する。

(2.10) 及び (2.11) を用いて

$$\begin{aligned} n^2 E(\bar{y} - \bar{y})^2 &= \frac{n}{N} \{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2 + \frac{n(n-1)}{N-1} \\ &\quad \{(y_1 - \bar{y})(y_2 - \bar{y}) + \cdots + (y_{n-1} - \bar{y})(y_n - \bar{y})\}\} \\ &= \frac{n}{N} \left\{ \left(1 - \frac{n-1}{N-1}\right) \{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\} \right. \\ &\quad \left. + \frac{(n-1)}{(N-1)} \{(y_1 - \bar{y}) + \cdots + (y_n - \bar{y})\}^2 \right\} \end{aligned}$$

(2.1)

二番目の項は $\sum_{i=1}^N y_i = N\bar{y}$ であるから 0 である。従つて

$$\begin{aligned} V(\bar{y}) &= E(\bar{y} - \bar{y})^2 = \frac{(N-n)}{nN(N-1)} \sum_{i=1}^N (y_i - \bar{y})^2 \\ &= \frac{S^2}{N} \frac{N-n}{N-1} \end{aligned}$$

系 1 $\bar{y}$ の標準誤差 (Standard Error) は

$$\sigma_{\bar{y}} = \frac{S}{\sqrt{N}} \sqrt{\frac{N-n}{N-1}} \quad (2.12)$$

系 2 母集団総計の推定量としての $\bar{Y} = N\bar{y}$ の分散は

$$V(\bar{Y}) = E(\bar{Y} - Y)^2 = \frac{N^2 S^2}{n} \frac{(N-n)}{N-1} \quad (2.13)$$

系 3 $\bar{Y}$ の標準誤差は

$$\sigma_{\bar{Y}} = \frac{NS}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (2.14)$$

§ 2.5 有限母集団補正 The finite population correction:

無限母集団からの大きさ n なる無作為標本に対して平均値の分散は (2.8) 式で $N \rightarrow \infty$ とおき S^2/n となる。母集団が有限である時は因子 $(N-n)/N$ が導入されてくる。分散に対する $(N-n)/N$ 及び標準誤差に対する $\sqrt{(N-n)/N}$ は有限母集団補正 (f.p.c.) と呼ばれる。抽出比 n/N が小さいとこの因子は 1 に近づく標本平均の標準誤差には直接影響を与えない。例えば 200000 の母集団から 500 の標本は 100000 の母集団からの 500 の標本と殆んど同じ正確な母平均の推定量を与える。

實際上抽出比が 5% を越えないときは f.p.c. を無視してよい。又多くの目的に対しては 10% 位であつても無視してよい。補正を無視する事は $\bar{y}$ の標準誤差を過大に見積る事と安全側である。

有限母集団に対する抽出理論は Cornfield (1944), Gray (1950) 及び Wishart (1952) により進展した。次の理論は定理 2.1 の拡張であるがこの章では大して必要でない。

(22)

が後章の役に立つために証明をする。

定理 2.3.

もしも y_i, x_i が pop. での各 unit として定義された。

変量の対とし $\bar{y}, \bar{x}$ が大きい n なる simple random sample からのそれ等に対応する平均値とすればそれ等の共変量 (Covariance) は

$$E(\bar{y} - \bar{y})(\bar{x} - \bar{x}) = \frac{(N-n)}{nN} \frac{1}{(N-1)} \sum_{i=1}^N (y_i - \bar{y})(x_i - \bar{x}) \quad (2.15)$$

この定理はもしも変量 y_i 及び x_i がすべての unit で等しければ定理 2.2 と同じになる。

(証明)

変量 $u_i = x_i + y_i$ に定理 2.2 を適用する u_i の母平均は $\bar{u} = \bar{x} + \bar{y}$ であり定理 2.2 は次の如くなる。

$$E(\bar{u} - \bar{u})^2 = \frac{(N-n)}{nN} \frac{1}{(N-1)} \sum_{i=1}^N (u_i - \bar{u})^2$$

即ち

$$E\{(\bar{y} - \bar{y}) + (\bar{x} - \bar{x})\}^2 = \frac{(N-n)}{nN} \frac{1}{(N-1)} \sum_{i=1}^N \{(y_i - \bar{y}) + (x_i - \bar{x})\}^2$$

両辺の平方の項を展開すると 定理 2.2 より

$$E(\bar{y} - \bar{y})^2 = \frac{(N-n)}{nN} \frac{1}{(N-1)} \sum_{i=1}^N (y_i - \bar{y})^2$$

$$E(\bar{x} - \bar{x})^2 = \frac{(N-n)}{nN} \frac{1}{(N-1)} \sum_{i=1}^N (x_i - \bar{x})^2$$

であるから (2.15) でこの二つの項は両辺で消去されてなくなり積の項のみが残る。即ち

$$E(\bar{y} - \bar{y})(\bar{x} - \bar{x}) = \frac{(N-n)}{nN} \frac{1}{(N-1)} \sum_{i=1}^N (y_i - \bar{y})(x_i - \bar{x})$$

§ 2.4 標本からの標準誤差の推定 Estimation of the standard error from a sample:-

推定母平均及び推定母総計の標準誤差に対する公式は先づ第 1 に次の 3 つの目的に対して用いられる。

(23)

(i) 他の抽出方法と simple random sampling との正確度の比較

(ii) 計画されてゐる調査に必要な標本の大きさの推定

(iii) 実行された調査で実際に得られた正確度の推定

公式 (2.3) は母分散 S^2 を含んでゐるが実際にはこれが分からなくても標本資料から推定出来る。これに関連した定理を次の定理 2.4 で述べる。

定理 2.4

simple random sampling に対して $s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$ は $S^2 = \frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N-1}$ の不偏推定量である。

(証明)

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

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

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

$$E\{n(\bar{y} - \bar{y})^2\} = \frac{(N-n)}{N} S^2$$

従つて

$$E(s^2) = \frac{S^2}{(n-1)N} \{n(N-1) - (N-n)\} = S^2$$

系 $\bar{y}$ 及び $\bar{y} = \sqrt{N} \bar{y}$ の分散の不偏推定量は

$$v(\bar{y}) = s_{\bar{y}}^2 = \frac{s^2}{n} \left(\frac{N-n}{N} \right)$$

$$v(\bar{y}) = s_{\bar{y}}^2 = \frac{N^2 s^2}{n} \left(\frac{N-n}{N} \right)$$

標準誤差に対しては

$$s_{\bar{y}} = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N}} : s_{\bar{y}} = \frac{Ns}{\sqrt{n}} \sqrt{\frac{N-n}{N}}$$

こゝに注意を要するのは母分散及び推定分散(標本分散)に用いた記号である。即ち

$$\text{母分散} \quad V(\bar{y}) = \sigma^2$$

$$\text{推定分散} \quad v(\bar{y}) = s^2$$

分散に対しては V 及び v 、標準誤差に対しては σ 及び s を別個に用いた方が便利である。

§2.7 信頼限界 Confidence limits

推定量 $\bar{y}$ 及び $\hat{y}$ はそれに応ずる母集団値のまわりに正規分布をしてゐると假定する。この假定及びその限界に関しては §2.8 で考える。もしもこの假定が保持されると、母平均及び母総計に対する上下の信頼限界は次の如くなる。

平均値

$$\hat{y}_L = \bar{y} - \frac{t \cdot s}{\sqrt{n}} \sqrt{\frac{N-n}{N}} : \hat{y}_U = \bar{y} + \frac{t \cdot s}{\sqrt{n}} \sqrt{\frac{N-n}{N}} \quad (2.21)$$

総計

$$\hat{Y}_L = N\bar{y} - \frac{t \cdot N \cdot s}{\sqrt{n}} \sqrt{\frac{N-n}{N}} : \hat{Y}_U = N\bar{y} + \frac{t \cdot N \cdot s}{\sqrt{n}} \sqrt{\frac{N-n}{N}} \quad (2.22)$$

此処に t は欲する信頼確率に応ずる正規偏差の値である。このごく普通の値は次の如くである。

信頼確率(%)	50	80	90	95	99
t	0.67	1.28	1.64	1.96	2.58

もしも標本の大小 n が 30 以下であれば t の値は推定分散 s^2 の自由度である $(n-1)$ なる自由度に応ずる第一表の Student の t 表からとられる。 t 分布は観測値 y_i がそれ自身正規に分布し、 N が ∞ である場合のみ正確である。正規性から適当にかけ離れてゐるとしても大して影響はない。小標本に対しては非常に歪んだ分布をしてゐるときは特別な方法が必要である。

例

政願書の署名が 575 枚あつめられた。各紙片は一枚に 42 人の署名が出来るスペースがある。その多くは署名人の数は少なかつ

た。紙片あたり署名人の数は 2/1 表に示す如く 7% 標本である 50 枚の紙片の無作為標本で数えられた。2/1 表は 29 頁参照。署名人の総数を 95% 信頼限界で推定せよ。

解

標本抽出単位は一枚の紙片である。観測値 y_i は一枚あたり署名人の数である。約半数が最大の 42 人の署名がしつゝあるから度数分布として表はされた資料は最大の度数が上端にあるので正規分布からはかけはなれてゐる。併し 1/2 以下経験からやはり 50 枚の標本平均は近似的に正規分布すると假定してよい。

2/1 表から次の如く分かる。

$$n = \sum f_i = 50 : y = \sum f_i y_i = 1471 : \sum f_i y_i^2 = 54497$$

従つて署名人の推定総計は

$$\hat{Y} = N\bar{y} = (476) \frac{(1471)}{50} = 19888$$

標本分散

$$s^2 = \frac{1}{(n-1)} \left\{ \sum f_i (y_i - \bar{y})^2 \right\} = \frac{1}{(n-1)} \left\{ \sum f_i y_i^2 - \frac{(\sum f_i y_i)^2}{\sum f_i} \right\}$$

$$= \frac{1}{49} \left\{ 54497 - \frac{(1471)^2}{50} \right\} = 222.0$$

方程式 (2.22) から 95% 信頼限界は

$$19888 \pm \frac{t \cdot N \cdot s}{\sqrt{n}} \sqrt{\frac{N-n}{N}} = 19888 \pm \frac{(2.01)(476)(15.13)}{\sqrt{50}} \sqrt{1-0.0740}$$

$$= 19888 \pm 2796$$

之は 95% 信頼限界に対して 17092 ~ 22684 の間にある事を示す。

完全調査では 21045 であつた。

§2.8 正規性近似の妥当性

Validity of the normal approximation:-

正規性近似が多くの實際面に於て妥当であると信ぜられてゐるがそれには多くの原因がある。確率理論に於ては多くの研究が random sample の平均値の分布について行われてきた。有限の

標準偏差を有する母集団では標本平均の分布は n が増すにつれて正規に近づく事がFeller(1950)により証明された。この研究は無限母集団に関連するものである。Madow(1948)は有限母集団の大なるClassに対しては標本平均の分布はたとえ抽出比 n/N が無視出来なくとも又抽出がもとにもとすことなしに行われたとしても正規に近づく事を証明した。Madowは n 及び N 共に比 n/N がある数より小さいものである様な際には無限となる事を規定してゐる。彼の結果は例えば抽出比が75%であつても適用されるであらう。この知識の母体にも何か物足りないものが残つてゐる。次の食向に答える事は容易ではない。即ちその食向とは“この母集団に関して正規性近似が充分正確である様な n は如何なる大いさであるか?”

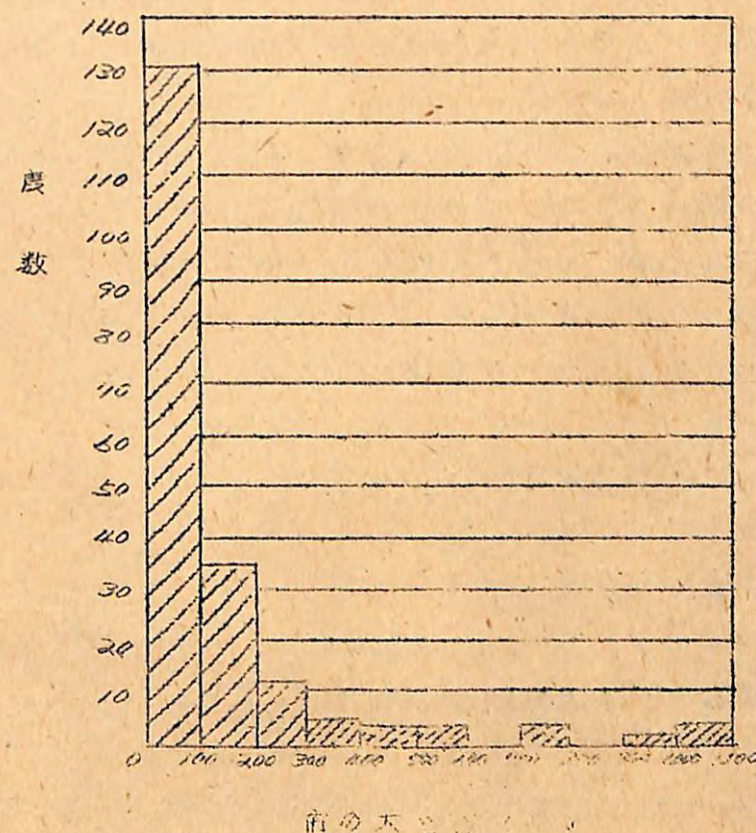
非正規性分布は正規分布とその性質及びへだたりの程度が非常に違う。標本抽出の実行面に於て遭遇する度数分布はすべて合理的に正規分布をしてゐると假定する事は出来ない。商店とか養鶏場とか町等の経済性をとつ企業の多くの型の分布は少数の大単位及び多数の小単位をもち正の歪を示す。同種類の歪は例えばブロック当りネズミの数とかハエの数の如き生物母集団にもあらわれる。正の歪分布の例として2/1図は1920年に調査されたアメリカの196の大都市の住民の数の度数分布を示す。2/2図はこの母集団から抽出された $n=49$ なる200のsimple random sampleの居住者の全数の度数分布を示す。

標本総計の分布及び平均値の分布は正規曲線と非常によく似てをりなお少しは正の歪を残してゐる。正規性近似の妥当性の論議に於て我々は正規性近似が“充分正確である。”といふ事は如何なる意味であるかを定義しなければならぬ。標本調査に於ては正規性近似は先づ第一に信頼限界の計算に用いられる。75%信頼限界が正規性近似によつて母平均 μ に対して計算されるとき我々は次の如き論述を。

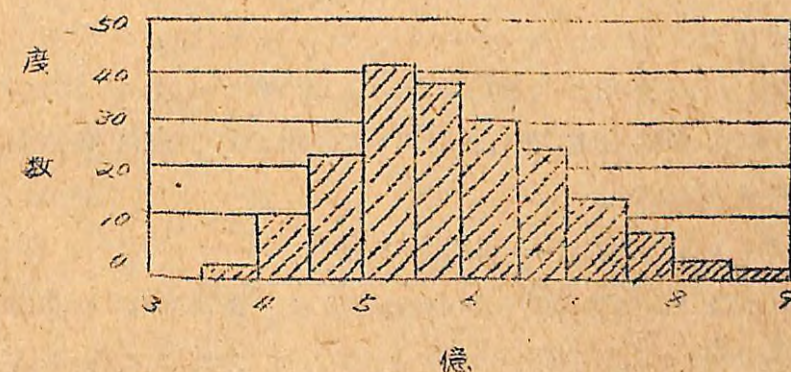
2-1

(2.23)

2/1図 1920年アメリカ合衆国196大都市の大きさの度数分布



2/2図 $n=49$ なる200の標本総計の総計の度数分布



repeated sampling ではこの種の論述にはその時の5%だけ誤りがある事を示してゐる。結論として正規性近似はむしろこの様な論述かその事象の4%と6%の間に実際には誤りがあるとするは充分正確であるといひ得る。

歪をもつ理論分布の研究及び現実の歪母集団からの抽出実験の結果からの論述は我々が正の歪母集団から抽出するとき信頼確率は通常どうなるかについて行われる。標本の大きさは又又図に示した様に $\bar{y}$ の分布が幾分正規に近づく程充分の大きさをもつてゐると假定される。そのときの論述は次の如くである。

(i) 主張 $\bar{y} - 1.96 \sqrt{\bar{y}} < \bar{y} < \bar{y} + 1.96 \sqrt{\bar{y}}$

が誤り度数は通常5%よりかなり大きい。

(ii) $\bar{y} > \bar{y} + 1.96 \sqrt{\bar{y}}$

なる度数は2.5%より大

(iii) $\bar{y} < \bar{y} - 1.96 \sqrt{\bar{y}}$

なる度数は2.5%より小さい。

例として Poisson 分布を考える。

変量 y_i は 0, 1, 2, ... の値をとり y_i が u なる値をとる確率は

$$Pr(y_i = u) = \frac{e^{-m} m^u}{u!}$$

此処に m は y_i の平均値である。

Poisson 分布に対して大いさ n なる simple random sample の総計 $\bar{y}$ の分布は母数 $m' = n m$ なる Poisson 分布とする事が分つてゐる。この分布表 (Molina 1949) から我々は n 及び m の色々な値に対して信頼限界に対して如何によく正規性近似をするかを覓出す事が出来る。

$m = 0.25$, $n = 400$ とする

$m = 0.25$ に対して $y_i = 0, 1, 2, 3$ 及び 4 である確率は 0.7728, 0.1947, 0.0243, 0.0020 及び 0.0001 である。最初の分布は明らかに歪をもつてゐる。

標本総計 y は 母数 $m' = (400)(0.25) = 100$ なる総計をもつ

Poisson 分布をする。

この分布に対して 次の理論的結果がある。

$$E(y) = m' \quad \sigma_y^2 = E(y - m')^2 = m'$$

従つて y は m' の不偏推定量である。 y の標準誤差として

$$\sqrt{y}$$

従つて m' に対する 95% 信頼限界は

$$y - 1.96 \sqrt{y} < m' < y + 1.96 \sqrt{y}$$

として正規性近似によつて作られる。

この論述が誤つてゐる確率を考える。

$$y = 82 \text{ のとき 上限 } y + 1.96 \sqrt{y} = 99.7$$

$$y = 83 \text{ " 上限 " " " } = 100.8$$

従つて $m' = 100$ なる Poisson 分布からの抽出に於て $y = 82$ 又は

それ以下であればいつでも上限は小さすぎる。Molina 表により y が 82 又はそれ以下である確率は 0.0369 である。

同様に下限は y が 122 又はそれ以上であれば大きすぎるそれに応ずる確率は 0.0181 である。

加え合せて

$$Pr(\text{低すぎる上限}) = 0.0369$$

$$+ Pr(\text{高すぎる下限}) = 0.0181$$

$$Pr(\text{信頼論述の誤り}) = 0.0550$$

誤まる全確率は 0.05 に近い値を示すが誤まつた論述をする場合の約 70% 於て誤りの m' は論述した上限より大きい (0.0550 × 0.70 = 0.0385)

もしも我々が推定量の誤差の絶対値だけが関心があれば $\bar{y}$ の分布での正の歪はかなりあつたにしても許容出来るが、もしも $\bar{y}$ が上の信頼限界を越える度数が 2.5% に近づくとき正規性近似は非常に小さい歪りか $\bar{y}$ の分布にない場合でない信用にならない。信頼限界を計算する際に正規性近似を用いる場合には如何なる大いさの n が必要かはならぬかに關しては完全な一般法則はな

(1) 正規性からの主たる偏差が著しい正の歪からなりたつている場合の母集団に対して次のような規則が有用である

$$n > 25 G^2$$

此処に G は Fisher の歪尺度 (Fisher's skewness) である。これは

$$G = \frac{E(y_i - \bar{y})^3}{\sigma^3} = \frac{1}{N\sigma^3} \sum_{i=1}^N (y_i - \bar{y})^3$$

この規則は 95% 信頼確率の論述が事象の 5% 以上は誤まらない様に計画されてある特別な母集団に対して G なる一つの推定量を計算する事によつて信頼限界を計算するために正規性近似の適用に必要な標本の大きさをラフに得る事が出来るがそれは何時でも抽出実験をしたあとでチェックしなければならない。

2.1 表 50 枚の歟願書の標本結果

署名人の数 y_i	度数 f_i	$f_i y_i$	$f_i y_i^2$
42	23	906	40572
41	4	164	6724
36	1	36	1296
32	1	32	1024
29	1	29	841
27	2	54	1458
23	1	23	529
19	1	19	361
16	2	32	512
15	2	30	450
14	1	14	196
11	1	11	121
10	1	10	100
9	1	9	81
7	1	7	49
6	3	18	108
5	2	10	50
4	1	4	16
3	1	3	9
計	$\sum f_i = 50$	$\sum f_i y_i = 1471$	$\sum f_i y_i^2 = 54497$

2.2 表 556 農場での歟物を作つてある畑の
ユーカーでの度数分布

級間隔 (エーカー)	級の中点 y_i	度数 f_i	$f_i y_i$	$f_i y_i^2$	$f_i y_i^3$
0 ~ 29	-2.9	47	-423	38.1	-343
30 ~ 63	0	143	0	0	0
64 ~ 97	1	154	154	154	154
98 ~ 131	2	82	164	328	656
132 ~ 165	3	62	186	558	1674
166 ~ 199	4	33	132	528	2112
200 ~ 233	5	13	65	325	1625
234 ~ 267	6	8	48	216	1296
268 ~ 301	7	4	28	196	1372
302 ~ 335	8	8	48	384	3072
336 ~ 369	9	2	18	162	1458
370 ~ 403	10	0	0	0	0
404 ~ 437	11	2	22	242	2662
438 ~ 471	12	0	0	0	0
472 ~ 505	13	2	26	338	4394
和		556	836.7	3469.1	20440.7

$$E(y_i) = \bar{y} = \frac{836.7}{556}$$

$$= 1.50486$$

$$E(y_i^2) = \frac{3469.1}{556}$$

$$= 6.23939$$

$$E(y_i^3) = \frac{20440.7}{556}$$

$$= 36.76385$$

$$\sigma^2 = E(y_i^2) - \bar{y}^2$$

$$= 3.97479$$

$$f_3 = E(y_i - \bar{y})^3$$

$$= E(y_i^3) - 3E(y_i^2)\bar{y}$$

$$+ 2\bar{y}^3$$

$$= 15.411$$

$$G = \frac{f_3}{\sigma^3} = \frac{15.411}{7.925}$$

$$= 1.9$$

例

2.2 表の資料は New York, Seneca 州での 556 農場の農家を
作る畑のエーカー数を表わす。

資料は West (1951) の研究の一部から引用したものである。
彼はこの母集団から太いさ 100 なる repeated sample
を抽出し、 $\bar{y}$, s 及び student の t を農場管理調査で関心をも
たれる数値の項目について度数分布を検討した。

G の計算が右に示されている。計算は假の尺度で行われている。

最初の級間隔の $\sim$ 27 は他のどのとかなり違っている事に注意

$G_1 = 1.9$ なる故 最小 n は次の如くなる。

$$n = (25)(1.9)^2 = 90$$

大いさ n の標本に対しては *West* はこの項目即ち穀物を作る畑のイーカー数では $\bar{y}$ の分布と *Student* の t 分布とそれに応ずる理論的分布と著しく違はない事を見出した。

抽出の実行面をうまくすると正規性近似を妥当なものとする傾向がある。正規性近似の失敗は標本平均を支配するある極端な値を母集団が含む場合に起る。この場合数が少なければ完全調査を行つてそれ等をうまく処理するために、さりはなすか別個な計画を立てるかする方が賢明である。母集団の主母体から極端な値を取り除く事は歪をなくし正規性近似をよくする。この技術は後章で述べる層化抽出の例である。

§ 2.7 推定分散 s^2 の非正規性の影響

Effect of non-normality on the estimated variance :-

非正規性の一つの影響は推定分散 s^2 が我々がもしも正規分布から標本抽出したと假定したとき期待する値よりは標本毎にかなり変動をするという事である。無限母集団に対して大いさ n なる *random sample* での s^2 の分散は (*Fisher* 1932)。

$$V(s^2) = \frac{2\sigma^4}{n-1} + \frac{k_4}{n} \quad (2.24)$$

第一項は s^2 の分散がその母集団分布が正規であるときに有する値である。第二項は非正規性の影響を表わす。量 k_4 は *Fisher* の四次の *cumulant* (*Fisher* 1932) で次の如くである。

$$k_4 = E(y_i - \bar{y})^4 - 3\sigma^4$$

G で測られたときはどの分布での歪は s^2 の安定性に影響はない事に注意せよ。重要な因子は母集団での四次モーメントである。

Cumulant k_4 は正規分布では 0 である。他の分布では正又は負の値をとるか抽出の実行面では k_4 は負よりも正の方が多いとあらわれ、ある母集団分布に対して高い値をもつ。今 (2.24) と次の如くかく

$$V(s^2) = \frac{2\sigma^4}{n-1} \left(1 + \frac{n-1}{2n} \frac{k_4}{\sigma^4} \right) = \frac{2\sigma^4}{n-1} \left(1 + \frac{n-1}{2n} G_2 \right)$$

此処に $G_2 = \frac{k_4}{\sigma^4}$ は *Fisher* の *kurtosis* の尺度である。

括弧内の量は s^2 の分散が非正規性により高められる因子を示す。この因子 $\frac{n-1}{2n} G_2$ が大であつても残るので殆んど n と無関係である事に注意せよ。

2.2 表の *West* の資料で G_2 は約 6 である事が分る。それで $V(s^2)$ はもしも正規分布として穀物を作る畑のイーカー数の最初の分布を考えたならば假定されたものの 4 倍に近くなる。

$$\text{訳註} \quad \left(1 + \frac{n-1}{2n} G_2 \approx 1 + \frac{6}{2} \approx 4 \right)$$

彼の抽出研究では 3 つの項目で標準偏差 s の分散でも同様に高められる事が分つた。 $V(s)$ と正規母集団からの s の理論分布との比は穀物を作る畑のイーカー数では 3.7、畑の全イーカー数では 2.1。生産してゐる人の仕事単位に対して 1.3.7 であつた。

抽出の実行面でこれに関連した事は一つの方法と他の方法の抽出の正確の比較のために s^2 の値を用いるか又は $\bar{y}$ の特別な正確度を得るに必要な標本の大きさを推定するのに s^2 を用いる事である (第 4 章参照) これ等の目的に対して s^2 が少数資料から計算されたとすると推定量 s^2 の正確度といふものに疑

向をとって考える幸になる。

前の結果を示す様に y^2 の分散を評価する正規公式の使用は y^2 の安定性について非常にあやまつた感銘を覚える幸になる。

第3章 割合及び百分率の標本抽出法

Sampling for proportions and Percentages

§ 3.1 定性的性質 *Qualitative characteristics*:-

我々がある特性又は属性をもつ或はある定まつた *class* に入つてある母集団での *unit* の全数又は割合又は百分率を推定したい場合がある。調査で記録された多くの結果は大抵この形のものである。無残の人の数とか住民の人口とかがその例である。その分類は簡単に“ハ、イ、エ”とかで答えられる様な直接に質問から導かれる様な類である。他の場合には最初の測定が多少とも連続でありその分類が結果の表示から導かれるものもある。或は回答者の最近の年令を記録し、その中から 60 才以上のものの割合を記録したりする場合である。

今母集団ですべての *unit* は二つの *class* *C* 及び *C'* の何れか/をとると仮定しよう。記号は次の如くである。

<i>C</i> なる <i>unit</i> の数		<i>C</i> なる <i>unit</i> の割合	
pop	sample	pop	sample
<i>A</i>	<i>a</i>	$p = A/N$	$p = a/n$

p の標本推定量は p であり *A* の標本推定量は Np 即ち Na/n である。

§ 3.2 標本推定量の分散 *Variances of the sample estimates*

簡単にこの立場にオズ章の理論を適用出来る。標本又は母集団のあらゆる *unit* に対して、もしも *unit* が *C* であれば y_i をとし *C'* であれば 0 であると定義しよう。 y_i なる値のこの母集団に対しては次の事が明らかである。

$$Y = \sum_{i=1}^N y_i = A \quad (3.1)$$

$$\bar{y} = \frac{\sum_{i=1}^N y_i}{N} = \frac{A}{N} = p \quad (3.2)$$

又標本に対しては

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{a}{n} = p \quad (3.3)$$

その結果 A 及び p を推定する問題は y_i が 1 か 0 かの母集団の総計及び平均値を推定する問題と考えられる。オ2章の理論を用いるため p 及び p の項で S^2 及び s^2 を最初に表はす

$$\sum_{i=1}^N y_i^2 = A = NP \quad ; \quad \sum_{i=1}^n y_i^2 = a = np \quad \text{であるから}$$

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

$$= \frac{1}{N-1} (NP - NP^2) = \frac{N}{(N-1)} p(1-p) = \frac{N}{(N-1)} pq \quad (3.4)$$

此処に
目標に

$$Q = 1-p \quad \text{である}$$

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{n}{n-1} pq \quad (3.5)$$

この母集団に定理 2.1, 2.2 及び 2.4 を適用すると分類される unit の simple random sampling は次の様な結果になる。

定理 3.1

標本の割合 $p = a/n$ は 母集団の割合 $P = A/N$ の不偏推定量である。

定理 3.2

p の分散は (3.4) 式を用いて

$$V(p) = E(p-P)^2 = \frac{S^2}{n} \left(\frac{N-n}{N} \right) = \frac{PQ}{n} \left(\frac{N-n}{N-1} \right) \quad (3.6)$$

系 1

もしも p 及び P が夫々 class C に入る sample 及び pop の百分率であれば公式 (3.6) は p の分散に対しても成立する。

系 2

$\hat{A} = NP$ の分散即ち class C に於ける unit の推定総数の分散は

$$V(\hat{A}) = \frac{N^2 PQ}{n} \left(\frac{N-n}{N-1} \right) \quad (3.7)$$

定理 3.3

標本から算かれた p の分散の不偏推定量は

$$v(p) = s_p^2 = \frac{(N-n)}{(n-1)N} pq \quad (3.8)$$

(証明)

定理 2.4 に於て連続変量 y_i に対して標本平均 $\bar{y}$ の分散の不偏推定量は

$$v(\bar{y}) = \frac{s^2}{n} \frac{(N-n)}{N} \quad (3.9)$$

である事が証明された

割合 p を $\bar{y}$ の代りに入れると (3.5) 式はその如くなる事を証明された

$$s^2 = \frac{n}{(n-1)} pq \quad (3.10)$$

依つて

$$v(p) = \frac{(N-n)}{(n-1)N} pq$$

もしも N が n に比して非常に大であれば $p \cdot C$ を無視して p の不偏推定量は

$$\frac{pq}{(n-1)} \quad \text{となる。}$$

多くの書物では $\frac{pq}{n}$ が普通用いられてゐるがこれは無限母集団に対してさえも不偏ではない。

系 $\hat{A} = NP$ の不偏推定量即ち class C での unit の推定総数の分散の不偏推定量は

$$v(\hat{A}) = s_{NP}^2 = \frac{N(N-n)}{n-1} pq \quad (3.11)$$

例 3042 名の名前及び宛先のリストから 200 名の名前の simple random sample をとり 38 名が誤った宛先であつた リストでの宛先の必要な補数の総数を推定し、この推定量の標準誤差

を計算せよ

$$\text{解 } N = 3042, \quad n = 200, \quad a = 38, \quad p = \frac{38}{200} = 0.19$$

誤った宛先の推定値は

$$\hat{A} = Np = (3042)(0.19) = 578$$

標準誤差は (3.11) より

$$S_{\hat{A}} = \sqrt{\frac{N(N-n)}{n-1} pq} = \sqrt{\frac{(3042)(3042-200)}{200-1} (0.19)(0.81)} = \sqrt{6685} = 81.8$$

誤った宛先の総数は

これから推論出来る事は 68% 信頼度で 578 ± 82 95% 578 ± 164

この例では抽出率は $n/N = 200/3042 = 0.07$ 即 7% であるから f.p.c. を無視して更に $n-1$ を n とおくと簡単な式を得る

$$S_{\hat{A}} = N \sqrt{\frac{pq}{n}} = (3042) \sqrt{\frac{(0.19)(0.81)}{200}} = 84.4$$

これは前の結果 81.8 と殆んど変わらない。

分散即ち p の推定分散に対する前の公式は、もしも unit が p が sample に於ける unit の総数に対する sample に於ける p なる unit の数との比である様にし、又は p が分けられた時にのみ成立する。それに対して unit が要素 (element) の群から作られてある事が普通の場合であり分類されるものは要素である。いくつかの例を述べると次の様である。

sampling unit	elements
家族	家族の員数
レストラン	従業員
卵の籠	個々の卵
梨の木	個々の梨の実

もしも unit の simple random sample が class C に属する pop での element の割合 p を推定するに抽出されたとする。前の公式は近似式とはなるけれども適用は出来ない。

各 sampling unit が M 個の element を含むとし $p_i = a_i/M$ は i unit での element の割合とし $p = \frac{\sum p_i}{n}$ を p の標本推定量としよう。正しい方法は量 p_i に対する章の公式も適用する事である。simple random sample では

$$V(p) = \frac{(N-n)}{Nn} \frac{1}{(n-1)} \sum_{i=1}^n (p_i - p)^2$$

でこの分数の不偏推定量は

$$v(p) = \frac{(N-n)}{Nn} \frac{1}{(n-1)} \sum_{i=1}^n (p_i - p)^2$$

である。

もしも M が unit 毎に異なれば問題により複雑となり、それに用いる方法は § 6.9 で述べる。

§ 3.3 標準誤差その p の影響 The effect of p on the standard error

方程式 (3.6) は一定の n 及び N に対して推定百分率の分散が p と共に如何に変わるかを示すものである。f.p.c. を無視すれば

$$V(p) = \frac{pq}{n}$$

函数 pq 及びその平方根を 3.1 表に示す。これ等の函数は夫々大きさ $1/(n-1)$ なる sample に対しての分散及び標準偏差と考えられる。

3.1 表 pq 及び $\sqrt{pq}$ の値 $p = \text{class C に於ける母百分率}$

	0	10	20	30	40	50	60	70	80	90	100
pq	0	900	1600	2100	2400	2500	2400	2100	1600	900	0
$\sqrt{pq}$	0	30	40	46	49	50	49	46	40	30	0

変数は母集団が等しい二つの class に分けられたとき最大の値をとりこの点を中心として対称である p の標準誤差は p が 30% と 70% の間にあるときは少ししか変化しない。 $\sqrt{pq}$ の最大の値は 50 でこの値では $n=100$ であれば $\sqrt{V(p)} = \frac{50}{\sqrt{100}} = 5\%$ である。換言すれば 5% の推定量の標準誤差を導くためには標本の大きさは 100 を必要とする。同様に 1% の標準誤差に対しては 2500 の標本の大きさが必要とする。

この考え方は我々が class C に属する pop での unit の総数に関心をもっている時には適用されない。これに対しては通常次の如き傾向が出来る。即ち、例えばこの推定量の 1% 以内で真の総計は推定されてゐるか？ この様な傾向に対しての答として一般に真の値 Np の出数又は百分率として標準誤差を考える様になる。

その出数

$$\frac{\sigma_{NP}}{NP} = \frac{N\sqrt{pq}}{Np} \sqrt{\frac{N-n}{N-1}} = \frac{1}{\sqrt{n}} \sqrt{\frac{q}{p}} \sqrt{\frac{N-n}{N-1}} \quad (3.12)$$

この量は推定量の変異係数 (Coefficient of variation) と呼ばれるものである。

f. p. c を無視すれば 変異係数は $\sqrt{\frac{q}{p}}$ となる。

f. p. c を無視して size 1 なる標本に対しての変異係数として考えられる比 $\sqrt{\frac{q}{p}}$ は 3.2 表に示されてゐる。

3.2 表 p の種々なる値に対する $\sqrt{\frac{q}{p}}$ の値

$p = \text{class } C \text{ に於ける母百分率}$

p	0	0.1	0.5	1	5	10	20	30	40	50	60	70	80	90
$\sqrt{\frac{q}{p}}$	∞	3.16	1.41	.99	.44	.30	.20	.15	.12	.10	.08	.07	.05	.03

一定の標本の大きさに対して class C に於ける推定総計の変異係数は C に於ける真の百分率が増すと共に減少する 変異係数は p が 5% 以下のときは大である。従つて母集団に於てごく稀な属性を有するものの総数の正確な推定量を得るには非常に大なる標本を要する。 $p=1\%$ に対して 20 即ち 10% の推定量の標準誤差を得るには

$$\frac{1}{\sqrt{n}} \sqrt{\frac{q}{p}} = 0.1 \quad \frac{1}{\sqrt{n}} 9.9 = 0.1 \quad \therefore \sqrt{n} = 99 \quad \therefore n = 9801$$

の大なる標本を要する。 simple random sample 又は一般目的に対して採用される抽出方法では稀な型の unit の総数を推定するには費用がかかる傾向がある事がこれによつても分る。例えばマグサ積みの中のハリの総数を推定するが如き場合である。

§ 3.4 二項分布 The binomial distribution:-

母集団が y_i が 0 か 1 である様な簡単な型である場合には単に平均値及び分散だけでなく推定量の度数分布を見出す事が出来る。母集団が class C に属する A unit 及び C' に属する $(N-A)$ unit を含むとする。比に $p = A/N$ である。もしも最初に抽出された unit が C であれば pop に残るものは C に属するものが $(A-1)$ unit C' に属するものは $(N-A)$ unit である。このとき C に属するものの割合は $(A-1)/(N-1)$ となる。

反対に最初に抽出されたものが C' であれば C を含む割合は A/N となる。

くりかえしを許さない (without replacement) 抽出では、割合は抽出を行うとこの様に変化する。この節ではこの変化は無視される即ち p は一定とする。これは A 及び $(N-A)$ は共に標本の大きさに比して相対的に大であると仮定する事である。

この假定の下に標本を抽く過程は n の試行の一連のものからなつてゐるとする。その各々に於て抽かれた unit が C である確率は p である。この立場は標本に於て C なる unit の数に対しても同様な二項分布を生ずる。標本が C なる属性を a unit 含む確率は

$$Pr(a) = \frac{n!}{a!(n-a)!} p^a q^{n-a} \quad (3.13)$$

この式から a 又は $p = a/n$ 又は推定総計 Np の度数分布が作られる。

最も分かり易い表は U. S. National Bureau of standards

(1950) に出ている。これは個々の項及び標本の大きさが49迄及び001の間隔で変る p の値の累積和を示してある。50と100の間の n の値に対しては Renig (1952) の表が参考になる。

§3.5 p の一般分析 The general distribution of p .

p の分布は pop が sample に比して非常に大であるといふ假定をしなくても見出す事が出来る。pop での二つの class C 及び C' に属する unit の数を夫々 A 及び A' とする。我々は標本でのそれ等に対応する数が a 及び a' に属する unit の数を夫々 A 及び A' とする。我々は標本でのそれ等に対応する数が a 及び a' である確率を計算する。此処に大々

$$a + a' = n \quad ; \quad A + A' = N$$

$$a \leq A \quad \text{及び} \quad a' \leq A' \quad \text{と假定する。何故ならもとにもと$$

すこのない標本抽出であるからこの不等式が成立しないと標本を抽出する事が不可能となるからである。

今最初に抽かれたすべての a unit が class C に属する場合の標本を考える。最初の抽出で C の一つが得られる確率は A/N である。1つの unit を抽いた後 pop には C なる unit は $(A-1)$ 残るから、二回目の抽出で C を得る確率は $(A-1)/(N-1)$ である。同様に三回目の抽出で C が得られる確率は $(A-2)/(N-2)$ である。従って抽出された最初の a unit がすべて C に属する確率はそれ等の積

$$\frac{A(A-1)(A-2) \cdots (A-a+1)}{N(N-1)(N-2) \cdots (N-a+1)} \quad (3.14)$$

今考えてある標本の型では今迄抽出された残りのすべての unit は C' に属さねばならない。

$(a+1)$ 回目の抽出では pop は $(N-a)$ unit の中 A' なる class C' に属する unit を含んでいる。

それから先の抽出で1つの C' を得る確率は夫々

$$\frac{A'}{N-a} \quad \frac{A'-1}{(N-a-1)} \quad \frac{A'-2}{(N-a-2)} \quad \text{etc}$$

従って残りのすべての unit が C' に属する確率はそれ等の積

$$\frac{A'(A'-1)(A'-2) \cdots (A'-a'+1)}{(N-a)(N-a-1) \cdots (N-a+1)} \quad (3.15)$$

(3.14) と (3.15) の積は最初の a unit がすべて C に属し残りの a' unit がすべて C' に属する標本を得る確率を与える。即ち

$$Pr = \frac{A(A-1)(A-2) \cdots (A-a+1) \cdot A'(A'-1) \cdots (A'-a'+1)}{N(N-1) \cdots (N-n+1)} \quad (3.16)$$

全確率を出すには假りに a が C に属し a' が C' に属するとする。一種類の a 及び他の a' が一つの線に沿って順序よく配列される数は nCa である。即ち $\frac{n!}{a!(a')!}$ である。

従って標本が C に属する a unit を含み C' に属する a' unit を含む全確率は

$$Pr(a, a'/A, A') = \frac{n!}{a!a'} \cdot \frac{A(A-1) \cdots (A-a+1) \cdot A'(A'-1) \cdots (A'-a'+1)}{N(N-1) \cdots (N-n+1)} \quad (3.17)$$

上式は次の如く簡単に表ゆせる

$$Pr(a, a'/A, A') = \frac{{}^nC_a \cdot {}^nC_{a'}}{{}^nC_n} \quad (3.18)$$

これが a 又は np の度数分布であり之から p の分布は直接に導かれる。この分布を超幾何分布 (hypergeometric distribution) といふ。

例

8人の家族が3人の男と5人の女からなっている。大きさ4なる simple random sample に於て男の数の度数分布を見出せ

解 $A=3, \quad A'=5, \quad N=8, \quad n=4$

公式 (3.18) より、男 a の度数分布は次の如くである。

a	確率
0	$\frac{4!}{0!4!} \cdot \frac{5 \cdot 4 \cdot 3 \cdot 2}{8 \cdot 7 \cdot 6 \cdot 5} = \frac{1}{14}$
1	$\frac{4!}{1!3!} \cdot \frac{3 \cdot 5 \cdot 4 \cdot 2}{8 \cdot 7 \cdot 6 \cdot 5} = \frac{6}{14}$

$$\begin{aligned} 2 & \frac{4!}{2!2!} \frac{3 \cdot 2 \cdot 5 \cdot 4}{8 \cdot 7 \cdot 6 \cdot 5} = \frac{6}{14} \\ 3 & \frac{4!}{3!1!} \frac{3 \cdot 2 \cdot 1 \cdot 5}{8 \cdot 7 \cdot 6 \cdot 5} = \frac{1}{14} \\ 4 & \text{不可能} = 0 \end{aligned}$$

これから簡単に男の平均数は $\frac{3}{2}$ で分散は $\frac{15}{28}$ である事が計算出来る。

この結果は §3.2 の公式の結果と一致する。即ち

$$E(np) = np = \frac{nA}{N} = \frac{(11)(3)}{(8)} = \frac{3}{2}$$

$$V(np) = npq \frac{N-n}{N-1} = 4 \cdot \frac{3}{8} \cdot \frac{5}{8} \cdot \frac{4}{7} = \frac{15}{28}$$

§3.6 信頼限界 Confidence limits :-

先づオーに定性的性質をもつ場合の信頼限界の意味を考える。標本に於ては N の中で a が class C に属する。関係事項は class C に属する pop での 1 の数について行われると假定しよう。 A に対する上の信頼限界に対しては、この値に対して a を得る確率即ち少くとも標本が C に属する確率は量 α 例えは 0.25 の如く小さい数である様に $\hat{A}_U$ なる値を計算する。公式的には $\hat{A}$ は次の公式を満足する

$$\sum_{j=0}^{\hat{A}} pr(j, n, \hat{A}_U, N - \hat{A}_U) = \alpha \quad (3.19)$$

此処に pr は (3.18) なる方程式で定義された *hypergeometric distribution* に対する確率の項である。前以て α が送はれると方程式 (3.19) はそれを満足するため $\hat{A}_U$ の非整数値を一般に必要とする。

併し乍らそれに反して概念的には $\hat{A}_U$ は整数値でなければならぬ。実際には我々は a が α より小さいか又は等しい様な A の最小の整数値として $\hat{A}_U$ を送ぶ。同様に下の信頼限界 $\hat{A}_L$ は次の如き最大整数値である。

$$\sum_{j=0}^{\hat{A}} pr(j, n, \hat{A}_U, N - \hat{A}_U) \leq \alpha_L \quad (3.20)$$

従つて p の信頼限界は $\hat{p}_U = \hat{A}_U/N$ $\hat{p}_L = \hat{A}_L/N$ によつて分る。正規性近似はこのときも役に立つ。定理 3.2 に於て p の標準誤差は

$$\sigma_p = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{pq}{n}}$$

p が

$$S_p = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{pq}{n}}$$

なる標準誤差をもち正規に分布してゐると假定すると信頼限界に対する正規性近似として次のものを得る

$$\hat{p}_L = p - t \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{pq}{n}} \quad ; \quad \hat{p}_U = p + t \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{pq}{n}}$$

此処に t は信頼確率に應ずる正規偏差である。正規性は連続的であり p の場合は不連続であるので連続に対する補正を求めてこれ等の公式を正さなければならぬ。何故に補正を行はねばならぬかの根本的理由は次の如くである。 30 の内 30 unit が class C に属すると観測されたと假定して $\hat{p}_U$ を近似的に求めたいとする。正確な分布を用いて $0, 1, 2, \dots, 30 \text{ unit}$ が C に属する確率の和が α である様な $\hat{p}_U$ を見出す。

もしも正確な分布を連続正規分布として近似させたいとすれば $29\frac{1}{2}$ と $30\frac{1}{2}$ との間の正規曲線の面積に應ずる 30 での最の横軸を考へる。 0 から 30 迄の項の和は $30\frac{1}{2}$ 点以下の正規曲線の面積に應ずるものである。補正としては $\hat{p}_U$ を $\hat{p}_U = p + \frac{1}{2} + t S_p$ として計算する。これは $\hat{p}_U$ が $\frac{1}{2}$ だけ増加する。同時に $\hat{p}_L$ は $\frac{1}{2}$ だけ減少する。従つて補正信頼限界は次の如くである。

$$p \pm \left\{ t \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{pq}{n}} + \frac{1}{2n} \right\} \quad (3.21)$$

連続に対する補正は近似に於ては少ししか長くはならない。併し補正しなければ信頼間隔は正規近似として計算すると通常狭すぎる筈であるが補正する事はこの欠点を補う助けとなるものである。

る。

正規性近似での誤差はすべて n , p , N 及び n/N なる量によって生ずる。その誤差が最も鋭敏にひびく量は np であり特に非常に数が少ないものを観測したりする場合に最もひどい。

3.3 表は正規性近似 (3.2) が用いられる時に用いられる規則を示す。

3.3 表 正規性近似の使用に対する np の最小値

p	np = これより小さい class で観測される数	n = 標本の大きさ
0.5	15	30
0.4	20	50
0.3	24	80
0.2	40	200
0.1	60	600
0.05	70	1400
10%	80	∞

* この意味は p が非常に小さいから np が Poisson 分布に従う。

3.3 表の規則は 95% 信頼限界でその限界 P を含まない真の度数が 5.5% 以下である様に作られたものである。更に上限が P 以下である確率は 2.5% と 3.5% の間であり下限が P 以上である確率が 2.5% 及び 1.5% の間である。誤差の一方の端での度数についてのこれらの制限は一般に二項分布が歪んでゐるから (3.2.8) である。確率に対して 5.5%, 3.5%, 1.5% を送るのは勿論自由である。

正規性近似で大きな誤差でもよければ n を小さくしてもよい。我々がとる立場が、正規性近似の妥当な範囲外にあるとき又は正確度が大でありたいときは Chung & DeLury (1950) により作られた超幾何函数の信頼限界の表を使用すればよい。これは母集団の大きさが 500, 2500 及び 10,000 に対する P に対する 90, 95, 99% 信頼限界を示してゐる。中間の母集団の大きさは挿入法で求める。正規性近似が適用されない他の場合は必要なる有限母集団補正を

ほとんどして二項分布に対する限界を用いる。 n が 50 以下の二項限界は *Tables of the binomial distribution* (U.S. National Bureau of Standards) が非常に速く見つける事が出来る。限界そのものに関しての便利な表は W. L. Stevens が作った Fisher & Yates の *Statistical tables* 3rd ed. VIII, 1 表がよい。

例 1.

大きさが 500 の母集団からの大きさ 100 なる simple random sample に於て class C に属する unit は 37 である。母集団に於ける割合及び総計の 95% 信頼限界を求めよ。

解

$$n=100, \quad N=500, \quad p=\frac{37}{n}=\frac{37}{100}=0.37$$

3.3 表によりこれは正規性近似が用いられる範囲内にある。 p の

推定標準誤差は

$$\sqrt{\frac{(N-n)pq}{(N-1)n}} = \sqrt{\frac{400(0.37)(0.63)}{499 \cdot 100}} = 0.0432$$

連続に対する補正 $1/2n = 0.005$ 従つて P に対する 95% 信頼限界は

$$p \pm \left(t \sqrt{\frac{(N-n)pq}{(N-1)n}} + \frac{1}{2n} \right) \\ = 0.37 \pm (1.96 \times 0.0432 + 0.005) = 0.37 \pm 0.090$$

$$\therefore \hat{p}_L = 0.280 \quad ; \quad \hat{p}_U = 0.460$$

Chung & DeLury の表での限界は夫々 0.285 及び 0.462 である母集団での総数に対する限界は夫々 N 倍して (500 倍)

$$N\hat{p}_L = 140, \quad N\hat{p}_U = 230$$

例 2

この例は二項限界が近似として用いられる場合を示す。前の標本で 100 の中 9 の unit だけが class C に属すると仮定する。

これは 3.3 表から正規性近似に対しての範囲外にある。

class C に属する期待数 np に対する 95% 二項限界は Fisher &

(48)

Yates の *Statistical Table VIII* 1 表から 4.20 と 16.40 である。それで割れば p に対する近似限界 0.042 及び 0.164 を得る。もしも抽出比が 5% 以下であればこの採りして見出される限界は大抵の目的に対してせまい限界である。この例では抽出比は 20% であるから $f.p.c$ が必要となる。

$$f.p.c \text{ は } \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{400}{499}} = 0.895$$

之を適用すると限界が狭くなる。漸正限界は次の採りである。

$$\hat{p}_L = 0.09 - (0.895)(0.09 - 0.042) = 0.047$$

$$\hat{p}_U = 0.09 + (0.895)(0.164 - 0.09) = 0.156$$

Chung & DeLury の表から得られる限界は夫々 0.045 及び 0.157 である。

§ 3.7. ニつ以上の class への分類 *Classification into more than two classes* 結果を表わす場合 unit が二つ以上の class に分類される場合がある。質問が簡単な“はい、又は”“いいえ”で答えられる採り時でも実際にはその結果は次の四つの class に分けられる。即ち“はい”“いいえ”“分らない”“答えられない”。この採り場合えの理論の拡張を此処では三つの class に分けられる場合を例として説明する。

i 番目の class に属する数が母集団では A_i 、標本では a_i とする。そのとき

$$N = \sum A_i \quad ; \quad n = \sum a_i \quad ; \quad p_i = \frac{A_i}{N} \quad ; \quad \pi_i = \frac{a_i}{n}$$

標本の大きさ n がすべての A_i に対して小さい場合には確率 p_i は標本の抽出を通じて一定と考えられる。観測される標本を抽出する確率は次の多項表示 (multinomial expression) で与えられる。

$$Pr(a_i) = \frac{n!}{a_1! a_2! a_3!} p_1^{a_1} p_2^{a_2} p_3^{a_3} \quad (3.22)$$

(49)

これは二項分布の拡張であり抽出比が小さいときは良い近似を与える。

観測される標本を抽出する確率に対する正しい表し方は

$$Pr(a_i/A_i) = \frac{A_1 C_{a_1} \cdot A_2 C_{a_2} \cdot A_3 C_{a_3}}{N C_n} \quad (3.23)$$

この表示は § 3.5 の超幾何分布の方程式 (3.18) の拡張である。此処では証明しない。

証明は超幾何分布に用いた方法で行えば (3.23) の結果を得る。

§ 3.8. ニつ以上の class である場合の信頼限界

Confidence limits when there are more than two classes.

二つの違った場合が明らかになり区別されねばならない。

Case I

$$p = \frac{\text{標本に於て一つの class に属する数}}{n} = \frac{a_i}{n}$$

又は

$$p = \frac{\text{class の group での割合}}{n} = \frac{a_i + a_j}{n}$$

例えば答が“はい”“いいえ”“分らない”“答えられない”の四つに分類されるとすれば p を“はい”と答えた標本に於ける割合とするか又は別に一定の答をあたえた標本に於ける割合即ち“はい”又は“いいえ”と答えたものの割合とする場合である。この場合何れの立場に於ても最初の分類は二つ以上の class を含んでいなければならない。そのものは n 個の unit を二つの class だけに分類して得られる。更に述べた理論がこの場合にも適用される。信頼限界は § 3.6 で述べたと同じ採りに計算される。

Case II

ある時には二つの部分に分けた残りの部分が失敗する^{こと}から p を計算するとある class がオミットされる場合がある。例えば“分らない”又は“答えられない”と答えた人はオミットして

「はい、+」い、と答えた人との比である。この形で作られた比は標本調査では偶然の比である。この採り比の分母は n ではなくてそれよりはるかにある数 n' である。

p の信頼分布は (3.24) よりも複雑である。何故ならば、 p の分子分母共にすべての標本は同じ総数 n であるけれども標本が変わるとに変わるからである。これは信頼限界の計算に障害となる。それ等の複雑性をなくは普通の統計理論即ち n 及び n' が与えられた p の条件付分布から信頼限界を計算する事を工夫すればできる事が出来る。この方法では割あてられた信頼標準で以て観測標本として同じ n 及び n' を持つすべての標本について真である信頼論述が出来る。

なぜこの方法が助けとなるかといえは p の条件付分布は次の節で示す様に通常の超幾何分布から得る事が出来るからである。もっと正確な結果を与えるため次の様に假定する

$$p = \frac{a_1}{a_1 + a_2} \quad ; \quad n' = a_1 + a_2 \quad ; \quad n = a_1 + a_2 + a_3$$

此処に a_3 はこの場合我々が関心をもちない class に属する標本での数である。そのとき a_1 及び a_2 の条件付分布は標本が大きさ n からなり pop の大きさが $N' = A_1 + A_2$ からなつてゐるとき超幾何分布をする時に p の条件は標準誤差は

$$\sigma_p = \sqrt{\frac{N' - n'}{N' - 1} \sqrt{\frac{pq}{n'}}$$

$p = A_1 / (A_1 + A_2)$ に対する信頼限界の正規近似は

$$p \pm \left(t \sqrt{\frac{N' - n'}{N' - 1} \sqrt{\frac{pq}{n'} + \frac{1}{2n'}} \right) \quad (3.24)$$

これにも一つの困難な事が残つてゐる。 n' は分つていても N' は標本からは分らないから方程式 (3.24) はそのまゝでは用いられない。 N' についての情報が手に入らなければ一つの方法は $n/n = N'/N$ と假定して N' を Nn/n として用いる等である。

ある。(3.24) にこの値を代入して近似限界は

$$\left. \begin{aligned} \hat{p}_1 &= p - t \sqrt{\frac{N - n}{N - 1} \sqrt{\frac{pq}{n'} - \frac{1}{2n'}}} \\ \hat{p}_2 &= p + t \sqrt{\frac{N - n}{N - 1} \sqrt{\frac{pq}{n'} - \frac{1}{2n'}}} \end{aligned} \right\} \quad (3.25)$$

が RC だけが N' に対して用いた近似に影響してゐる。

§3.9 p の条件付分布 The conditional distribution of p

この節では如何にして $p = A_1 / (A_1 + A_2)$ の条件付分布が得られるかを例を以て説明する。

三つの class では観測標本を抽出する確率は方程式 (3.23) で与えた様に

$$P_i(a_i/A_i) = \frac{A_i C_{a_i} \cdot A_2 C_{a_2} \cdot A_3 C_{a_3}}{N C_n} \quad (3.26)$$

条件付分布では $n' = a_1 + a_2$ は一定である。この n' の値を示さない大きさ n なるすべての標本は無視される。この値をもつ標本が方程式 (3.26) と相対的に同じ確率をもつてゐる。条件分布を見出すため (3.26) をすべての可能な標本に対する全確率で割る。此の全確率は n の中 n' が 1 または 2 の class に属する確率である。それは次の超幾何分布で与えられる

$$\frac{(A_1 + A_2) C_{(a_1 + a_2)} \cdot A_3 C_{a_3}}{N C_n} \quad (3.27)$$

(3.26) を (3.27) で割つて標本の条件付分布は次の如く得られる。

$$\begin{aligned} P_i(a_i/A_i, n') &= \frac{A_i C_{a_i} \cdot A_2 C_{a_2}}{(A_1 + A_2) C_{(a_1 + a_2)}} \\ &= \frac{A_i C_{a_i} \cdot A_2 C_{a_2}}{N' C_{n'}} \end{aligned} \quad (3.28)$$

此処に $N' = A_1 + A_2$, $n' = a_1 + a_2$

此れは大きさ N' なる母集団からの大きさ n' なる標本に対する超幾何分布である。

(52)

例として五つの unit A, B, C, D, E からなる母集団を考える。これは次の様に三つの class に属する。

class	A_i	class に属する unit
1	1	A
2	2	B, C
3	2	D, E

大いさ 3 なる simple random sample で我々は $P = A/(A+B)$ 即ちこの場合 P を推定したいとする。このとき $N=5$ $N'=3$ である。

この場合すべてが等しい初期確率をもつ大いさ 3 なる 10 の可能な標本が出る。これ等は次の様に n' の値によつてまとめられる。

$n'=1$					
標本	a_1	a_2	p	条件付確率	$(p-p)$
ADE	1	0	1	$1/3$	$\frac{2}{3}$
BDE or CDE	0	1	0	$2/3$	$-\frac{1}{3}$

もしも標本が a_1, a_2 の値で分類されるとすれば次の二つの型だけが得られる。即ち $a_1=1, a_2=0$; $a_1=0, a_2=1$ その条件付確率は夫々 $\frac{1}{3}$ 及び $\frac{2}{3}$ でこれは一般表示 (3.28) と一致する。

更に

$$E(p) = \frac{1}{3}$$

$$\sigma_p^2 = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right) + \left(\frac{2}{3}\right)\left(\frac{1}{3}\right) = \frac{2}{9}$$

推定 p は不偏でその分散は一般公式

$$\sigma_p^2 = \left(\frac{N \cdot n'}{N-1}\right) \frac{pq}{n} = \left(\frac{3-1}{3-1}\right) \left(\frac{1}{3}\right) \left(\frac{2}{3}\right) = \frac{2}{9} \quad \text{と一致する}$$

$n'=2$ に対しては a_1, a_2 の二つの組だけを与える六つの可能な標本がある。

(53)

 $n'=2$

標本	a_1	a_2	p	条件付確率	$(p-p)$
ABD, ABE, ACD or ACE	1	1	$1/2$	$2/3$	$\frac{1}{6}$
BCD or BCE	0	2	0	$1/3$	$-\frac{1}{3}$

このときも又推定量は不偏でその分散は一般公式から明かな様に

$$\sigma_p^2 = \left(\frac{2}{3}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{3}\right) = \frac{1}{6}$$

分散が $n'=1$ のときの $1/6$ にすぎない事は注意を要する。分散はそれが抽出された標本の配置とともに変化する。 $n'=3$ に対しては ABC なるたゞ一つの可能な標本があるだけである。これは正しい母集団の割合として p が計算する。 p の条件は分散は一般公式によつて示される如く 0 である。それは $n'=n'$ のときに 0 となる。

第4章 標本の大きさの推定

The Estimation of sample size

§4.1 仮想的な例 A hypothetical example

標本調査に於てその計画にあたって標本の大きさについての論議は重要なものである。あまり大に過ぎる標本は費用の浪費を意味し又あまり小に過ぎる標本は結果の有用性を減ずる。その論議は必ずしも満足なものは得られない。何故なら標本の大きさの選択がそのとき最良であるかといふ事の確信を得る情報は充分ないからである。その解決に到達する段階として仮想的な例を考える。ある一人の人類学者がある島の住民を調査研究しようとしてゐる。色々研究対象があるがその中で彼は血液型Oに属する住民の割合がどの位かを知りたい。simple random sample がとり易い構図であつた。どの位の大きさの標本をとればよいのか？

これに対する解答はまづ次の質問に対しての答が得られないと論議出来ない。即ち人類学者は血液型Oをもつ人の割合をどの位の正確度で知りたいか？ その答として百分率が±5%以内正しい。即ちもしも標本が43%の人が血液型Oであるという事を示せば全部の島では確かに38%から48%の間にある様に推定したいとする。

誤解を避けるため我々は一人一人を測定する以外には絶対に5%以内に正確度を抑える事は出来ないと人類学者に指摘する。許しを大にすると欲する25%より以上の誤差がある非常に悪い標本が入ってくる機会がある。人類学者はそのとき敢て悪い標本を得る機会を20回に1回位としたい。又彼が望んでゐるのはnの値であると答える。

さてこのときになつて我々はnのラフな推定量を得る立場に立たされる。簡単のためにf.p.c.を無視して標本百分率pが正規に分布してゐると仮定する。この仮定が合理的であるかどうかは最初のnが分つたときに明らかとなる。統計の言葉でpは20回に1

回の不平均が±5%の範囲内にある。pがpまゝの正規に分布してゐるから、20回に1回以外はp±2.5σp内にある。(厳密には、p±1.96σp) 更に

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

従つて

$$2\sqrt{\frac{pq}{n}} = 5 \quad \text{or} \quad n = \frac{4pq}{25}$$

この場合nは標本の大きさの推定の場合にごく普通起る困難な問題にぶつかる。nに対する公式は得られたがこのnは抽出さるべき母集団のある特性によつて定まる。この例ではその特性は我々が測定したい量pである。従つて我々は人類学者にpがどの位の値であらうかそれについて何かないだろうかと質問する。彼は他の人種の以前の資料から又この島の人類の歴史についての推測からpが30~60%以外にあつた様な事は恐らくないだろう。と答える。これでこの情報は有用な情報を与えるに充分である。pが30と60の間の値であればpqは2100とp=50でmax.の2500の間にある。それに加ふるnの値は336と400の間にある。安全側でnの推定量は400である。

この分析に於て行われた仮定は再び検定される。n=400でpが30と60の間であればpは正規に殆んど近い分布をしてゐる筈である。f.p.c.が必要かどうかは島の住民の数による。もし住民の人口が8000人以上であれば抽出率は5%以下でf.p.c.は必要でない。もしも必要であれば再検討をなどの方法は§4.4で述べる。

§4.2 問題の分析 Analysis of the problem

標本の大きさを定めるに當つての主なるとるべき段階は次の如くである。

- (1) 標本について何が期待されるかについての論議がなければならぬ。この論議は前の例の如くに誤差の欲する限界に因するものである。又標本の結果が分つたときとるべき行動又は

越さるべき論議に關するものである。論議を如何に作らねば責任は数字を表はしたいといふ必要性は多いけれども調査の結果を用いた人に歸する。

(ii) 望む標本の正確度とそれとの関係の方程式を見出さなければならぬ。その方程式は正確度についての論議の内容及び当面の標本抽出の種類により変ってくる。確率標本抽出の利益の一つはこの方程式を作る事が出来る事である。

(iii) この方程式はパラメーターとして母集団のある未知の特性を含んでゐる。これ等は特別な結果を与えるために推定しなければならない。

(iv) 資料がある母集団を分割してそれに対して欲する誤差の限界が各分割毎に作られる事がよくする。個々の計算は各分割でのそれに対して行われ全体のそれはすべてを加えて見出される。

(v) 一つ以上の項目又は特性が通常標本調査では測定される。時には項目の数が多き事がある。もしも欲する正確度が各項目に対して定められると各項目に対して夫々置つた n の値が出てくる。この n の値を一致させるためにある方法が考案されるべきである。

(vi) 最後に置かれた n の値は標本をとるのに利用する資源と矛盾がないかどうかを評価する必要がある。これは標本の欲する大きさをとるに要する費用、労力、時間及び材料の推定が必要となる。時には n を非常に小さくしなければならない様な事もある。この様なときには、その標本に小さい標本で押し進めていつてよいかどうか、正確度を、いてよいかどうか、資が出来る迄この仕事を放棄するかどうか等の困難な決定をしなければならない問題に当面する。

次節ではこれ等の問題のいくつかを詳細に述べてゆく事にする。

§ 4.3 正確度の分類 *The specification of precision*

欲する正確度についての論議は我々が標本推定量に於て許容したい誤差の量について行われる。この量は出来る限り標本の結果を

如何に利用するかによって決定される。特にその結果を色々利用する場合にはいかなる程度の誤差が許容されるべきかを定める事が困難な場合がある。

人類学者の例でいえば、例えば 4% 又は 6% よりも何故に 5% を正しく血液型のなもつ百分率がほしいのかを彼に質問したと假定しよう。彼はそれについて血液型のグループの資料は人種の種類にまづ用いられるものだと答える。彼はその島は約 35% の P をもつ人類型か又は約 50% の P をもつ人類型かに傾向をもつてゐる。5% の誤差限界はこれ等の一つの型の分類には充分小さいものである。併し彼は 4% 又は 6% の誤差限界でもさしてこの場合障害ではないのであらう。

この様に人類学者による 5% の誤差限界の選択はある程度任意的なものであつた。この奥でこの例は誤差限界が定められる方法の典型的なものである事実は人類学者は他の多くの科学者又は政治家よりも

彼が望むものについてはより確実ではあつた。最初に欲する正確度の問題が起るときにこの採入人々即ち科学者とか政治家はその問題に対して何も考へもたないとか答へるべき何等の思想をもたないと告白するであらう。併し経験からいつて少し論議をすると少くともラフに彼等に合理的に思える様な誤差限界の大きさを示す事が出来るであらう。

これより更に我々は實際的な立場には進む事は出来ない。困難な事は調査結果から行われる實際的な決定をする場合の落大方に影響する色々な大きさの誤差については充分でないといふ事である。この様な事は左に次々に左に単に受け入れるといふよりは研究に値する問題である。知識が増すにつれて欲する正確度の選択は容易となつてくる。併し、誤差の結果が分つた時でも、その結果が違つた人に用いられる多くの大切な調査があり、その目的のあるものは調査

計画された時に前以ては分らないものである。従つて推測する事が優秀である事がその場合の正確度の分類にとって大切である様に思はれる。

もしも標本が非常に特殊な目的例えば単にハイとかイ、エとかをきめる場合とか、ある事業で費される金額がいくらかを決めるとかする場合には必要な正確度は決定をする際の誤差の結果によつて明確に述べる事が出来る。この型の問題に対する一般的右解答は §48 で述べる。

§44 割合に対する標本抽出に於ける n に対する公式

The formula for n in sampling for proportions is:-
unit は二つの class C 及び C' に分けられる。class C に属する unit の推定割合 p に於ける誤差のある限界を d とすれば実際の誤差が d よりも大である様な事が小さい危険率 α しかない様にしたい。即ち

$$P\{ |p - P| \geq d \} = \alpha$$

simple random sample と假定し p は正規に分布してゐるとする。§3.2 定理 3.2 から

$$\sigma_p = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{pq}{n}}$$

従つて欲する正確度と n との関係の公式は

$$d = t \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{pq}{n}}$$

此処にも正規偏差である。 n について解くと、

$$n = \frac{\frac{t^2 pq}{d^2}}{1 + \frac{1}{N} \left(\frac{t^2 pq}{d^2} - 1 \right)} \quad (4.1)$$

実際的な使用に當つて p の前以ての推定量 p_0 がこの公式で代りに用いられる。 N が大であれば一次近似として

$$n_0 = \frac{t^2 p_0}{d^2} \quad (4.2)$$

α と t は比としてだけこの公式に入つてゐるのに注意 もしも

$$V = \frac{d^2}{t^2} = \text{標本割合の欲する分散}$$

とおくと

$$n_0 = \frac{p_0}{V}$$

実際的にはまづ n_0 を計算する。もしも n_0/N が無視出来れば n_0 は方程式 (4.1) の n の満足な近似である。無視出来なければ (4.1) と (4.2) を比較して次の様にして n が得られる。

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \approx \frac{n_0}{1 + \frac{n_0}{N}} \quad (4.3)$$

例) 假想的な血液型の例では

$$d = 0.05, \quad p = 0.5; \quad \alpha = 0.05, \quad t = 2$$

従つて

$$n_0 = \frac{(4)(0.5)(0.5)}{(0.0025)} = 400$$

寫には 3200 人を検査人であると假定しよう。 p, C が必要である。

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} = \frac{400}{1 + \frac{399}{3200}} = 356$$

n_0 に対する公式はもしも α, p 及び q が割合の代りに百分率で表わされたとしても成立する。積 pq は p が $1/2$ 即ち 50% になるにつれて増大するから n の安全側の値は p が存在とそうなる範囲に於て p を殆んど $1/2$ に近い値を送る事により得られる。例えば p が 5% と 9% の間にあれば n の推定量に対して 9% を假定する特別な施用での割合に対する標本の大きさについての良い論議は Cornfield (1951) が行つてゐる。

§4.5 連続資料での n に対する公式 The formula for n with continuous data :-

$\bar{y}$ を simple random sample からの観測値の平均値とすれば我々は

$$Pr \{ |\bar{y} - \bar{Y}| \geq d \} = \alpha$$

である杯にしたい。此処に d は誤差の許しを限界で α は小さい確率である。

$\bar{y}$ は正規に分布してゐると假定して定理 2.2 系ノから、その標準誤差は

$$\sigma_{\bar{y}} = \sqrt{\frac{N-n}{N-1}} \cdot \frac{S}{\sqrt{n}}$$

$$\text{従つて } d = t \sqrt{\frac{N-n}{N-1}} \cdot \frac{S}{\sqrt{n}} \quad (4.4)$$

$$\text{これから } n = \frac{(\frac{tS}{d})^2}{1 + \frac{1}{N} (\frac{tS}{d})^2}$$

前節の如く第一次近似として

$$n_0 = \left(\frac{tS}{d} \right)^2 = \frac{S^2}{V} \quad (4.5)$$

これは n_0/N が無視出来れば適当である。さうでないときは

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (4.6)$$

例 販売用の幼苗を生産する苗畑に於て晩冬又は早春に賢手に対しての順序をきめる方策をたてるため手元にとの位の健全苗木があるかを推定したい。苗木の全数を推定する抽出方法の研究が Johnson (1943) により行われた。1ft 巾で 430ft の長さの銀カエデの苗木の床から得られた資料は次の如きものであつた。sampling unit は床の長さの 1ft であるから $N=430$ である。苗木の完全調査によつて $\bar{y}=19$ $S^2=85.6$ である事が分つた。これ等は母集団の真の値である。simple random

sampling で 20 回に 1 回の確率で $\bar{y}$ を 10% 以内の誤差に推定するためにはどの位の unit をとつたらよいか?

方程式 (4.5) から

$$n_0 = \frac{t^2 S^2}{d^2} = \frac{(4)(85.6)}{(1.9)^2} = 95$$

n_0/N は無視出来ないから

$$n = \frac{95}{1 + \frac{95}{430}} = 78$$

全体の苗木のほん 20% を採る正確度を得るために数えればよい。

この例は母集団の値 S^2 が分つてゐる杯を例外なものである。実際には S^2 は同杯を又は関連した母集団の以前の標本抽出からか又はすぐれた推測によつて推定される。このために将来の標本調査の助けとして調査で得られた標準偏差を印刷するか少しも分り易い杯にして置く必要がある。

予備調査の杯に一寸したもののからでも S^2 の役に立つ推定量を得る事が出来る。例えば土壌中のハリガネ虫の母集団の推定に関する以前の研究に於ては表土の小さい標本例えば $9 \times 9 \times 5$ inch のものをする道具を使用した。標本の大きさの推定に対して抽出する人は標本が含むハリガネ虫の数の標準偏差を知る必要があつた。もしもハリガネ虫が表土に at random に分布してをれば小さい容積中に見出される数は Poisson 分布に従うであらう。従つて $S^2 = \bar{y}$ である。ハリガネ虫はあつまる傾向があるので安全係数として 1.2 をとり $S^2 = 1.2\bar{y}$ と假定した。そのものは分らないけれども経済的に重要性をもつ $\bar{y}$ の値が穀物の生長に損害をあたえるのに危険である。ハリガネ虫の密度の研究から導く事が出来るであらう。これ等の情報之二つが満足せよと証明された標本の大きさの決定を可能にした。

Deming (1950) は分布の range 巾及び型についてのある知識から S を推定する有用なヒントを与えている。

此処に与えられた \$n\$ に対する公式は標本平均が \$\bar{y}\$ の推定量として用いられた *simple random sampling* のみに適用され得る他の抽出及び推定方法に対してはそれに用いる公式はその場合々に説明する。

前述した際に \$n\$ は次のものを満足する大きさでなければならないと述べた。即ち

$$Pr\{|\bar{y} - \mu| \leq d\} = \alpha$$

此処に \$d\$ は誤差限界と呼ばれたものである。又標本の大きさは信頼確率 \$(1-\alpha)\$ で \$1/2d\$ の信頼間隔をあてえるに充分な大きさであるとして定義される。小標本以外ではこれら二つの方法は等価とされ \$n\$ に対して同じ推定量を導く。

§ 2.7 から信頼間隔の半分は

$$d = t \sqrt{\frac{N-n}{N}} \frac{s}{\sqrt{n}} \quad (4.7)$$

この方程式は (4.4) の \$S\$ の代りに \$s\$ とおいただけで同じである。たゞ \$n\$ が 30 以下であるとすれば \$d\$ に対しての (4.7) の \$t\$ の値は正規分布の代りに自由度 \$(n-1)\$ をもつ Student の \$t\$ 分布から得られる (表一表) 二つの方法は共に近似である。 (§ 4.7 参照)

§ 4.8 一つ以上の項目をもつ標本の大きさ *Sample size with more than one item*:-

大抵の調査では情報は一つ以上の項目について集められる。標本の大きさを決定する一つの方法はその調査で最も大切なと考えられる数項目に対して誤差限界をきめる事である。必要な標本の大きさの推定はまづこれ等の重要な項目の各々に対して別個に行われる。単一項目についての \$n\$ がきまつたときにその立場でその \$n\$ についてどうであるかを精査する。

必要とされる \$n\$ がすべて似かよつてゐる場合がある。もしも \$n\$ の中で最大なものがそれに要する費用の限界内にあればこの \$n\$ が選ばれる。普通の場合には予算面から又は最初に予想したよりも廣く大なる正確度が外のすべての標本で得られて無駄であるとい

ふ事案から最大の \$n\$ を選ぶ事が困難であるほど \$n\$ の間に変動が大きい。この場合には欲する正確度をより小さい \$n\$ を用いるために項目のあるものに対してすこしおとす事になる。

ある場合には、異なる項目に対して必要な \$n\$ が非常に異なつてゐるので利用出来る資源で以てはこれ等の項目に対して期待される正確度が全体的に不適當であるからある項目が調査の対象から除外されなければならない場合がある。困難はたゞに標本の大きさのみにはとゞまらない。ある項目は他の項目と違つた抽出方法をとりねばならない場合がある。くりかえしの抽出を行う母集団では、一般的に調査に経済的に組合してよい項目及び特別な方法が必要とする項目についての情報をあつめておく事が必要である。例えば地域の農業調査で経験によつて暗示された四つの型に項目を分類したものが 4/表である。この分類で一般調査とは例えば *simple random sample* の如く unit が地域上にかなり平等に分布してゐるものを意味する。

4/表 地域調査に於て項目の違つた型の例

型	項目の特性	必要な標本抽出の型
i)	地域のすべての部分に合理的な度数で以て表われる地域を通じて広くゆきわたつてゐるもの	抽出比の小なる一般調査
ii)	地域を通じて広くゆきわたつてゐるが低い度数をもつもの	抽出比の大なる一般調査
iii)	地域の大部分に合理的な度数を以て表われる他の部分にはなくある部分に多くあつまるのでまばらな分布をしてゐるもの	地域の異つた部分で違つた密度をもつ層化抽出 (第 5 章) 時にはある補助的な抽出で一般調査に含ませる。
iv)	非常にまばらな又地域の少	一般調査は適當でない。

部分にあらまつた分布をし

てあるもの

その分布型にあつた抽出
方法が必要である。

§ 47 = 二段抽出の Stein の方法 *Stein's method of two-stage sampling* ; -

もしも S 又は P が以前の資料又は推測から推定されるとすれば
 n に対して与えられた計算は誤差限界又は信頼限界が満足な大きさ
のものであるとの確信は与えられない。何故なら S 及び P の假り
の推定量は誤つてゐるかもしれない連続資料に対して *Stein* (1945)
は信頼間隔がある量より大きくはならない事を保証する方法を進
展させた。この方法では S についての情報は抽出される母集団か
ら得る。Stein の方法は母集団は正規であり実際の応用に當つて
は、この条件がほぼ満足される抽出方法に限ると假定してゐる。
併し、その方法は次の理由で大いに興味のある問題である。即ち
それは最初の推測の正確さには依存しない筈に正確度を与える
からである。標本が二つの部分でとられたいとする。大きさ n ,
なる最初の部分は S の推定量 s を通常の標本に計算して与えられる
又それは平均値 $\bar{y}$ の初期の推定量を与える。最初の部分がとられ
たとき s 又は n は特別な信頼間隔を得るために必要な附加的な観
測値の数を計算する方法を示してゐる。二つの部分共に調査され
てゐる母集団から抽出せしめなければならぬ。もしも母集団が時
と共に変化すれば第一巻目と二巻目の部分の時間の間隔はその変
化量が評価されない位充分小さなものでなければならぬ。
その方法は無限母集団に対して発展したから n が無限大なる場
合を先づ考えよう。第一の標本が得られたとき $\bar{y}$ に対しての信
頼間隔が計算出来る。この間隔の半分は

$$t_1 s / \sqrt{n_1}$$

此処に t_1 は自由度 $(n-1)$ 信頼確率 $(1-\alpha)$ なる t の値を示す。
もしもこの量が欲する信頼巾の半分 d に等しいか又はそれより

小さければ標本は己に充分な大きさのものである。もしもこの量
が d 以上になれば、全体の標本の大きさ n が次のものになる様に
附加的な観測値をとる。

$$t_1^2 s^2 / d^2$$

又 $\bar{y}$ をして全標本の平均値とすれば

$$Pr \{ |\bar{y} - \bar{Y}| \geq d \} \leq \alpha \quad (48)$$

証明の概略

証明には $y_1, y_2, \dots, y_n$ が $\bar{Y}$ のまわりに正規に分布してゐると
假定する。証明を通じて d, α 及び n は一定量である。全標本
の大きさ n は一定ではなくて、その値は最初の標本に於て見出さ
れた s の値によるから 任意変量 である。併し n を一定に
すれば n は一定となり、量

$$\sqrt{n} (\bar{y} - \bar{Y})$$

は平均値の次び分散 σ^2 をもつて正規に分布してゐる。従つてこ
の量は s が一定であらうがなからうが正規分布をしてゐる。更に
正規分布のよく知られた特性により、その分布は s の分布とは独立である。
従つて $\sqrt{n} (\bar{y} - \bar{Y}) / s$

は自由度 $(n-1)$ なる t 分布をする。この定義により

$$Pr \left\{ \left| \frac{\sqrt{n} (\bar{y} - \bar{Y})}{s} \right| \geq t_1 \right\} = \alpha \quad (49)$$

これが証明の鍵である。更に n が計算された方法によつて常に

$$\sqrt{n} \geq \frac{t_1 s}{d} \quad \text{or} \quad \frac{\sqrt{n}}{s} \geq \frac{t_1}{d} \quad (410)$$

従つて (49) から

$$Pr \left\{ \left| \frac{t_1 (\bar{y} - \bar{Y})}{d} \right| \geq t_1 \right\} \leq \alpha$$

$$\text{即ち } Pr \{ |\bar{y} - \bar{Y}| \geq d \} \leq \alpha$$

(例) $d=10, \alpha=0.05$ と假定する。以前の情報からこの推
測はかぎり誤差があるかもしれないが S は約 60 と推測されて

ある。Steinの方法の適用に当つてのオー投階は最初の標本の大きさ n_1 を送ぶ事である。このため、もしも假定した S の値が正しいとしたならばその位の最終の標本の大きさであらねばならぬかを知る必要がある。この値は

$$n = \left(\frac{tS}{d} \right)^2 = \left(\frac{2 \times 60}{10} \right)^2 = 144$$

此処にもは正規分布からとられた値である。 $n_1 = 50$ としたとしよう。これは S を推定するために必要な自由度の合理的な大きさを与えるもので S が我々が恐れているよりも小さくある場合があるから最初の標本はあまり大きくとつてはならない。オーの標本がとられて $s^2 = 1938$ であつたとする。

49 df に対する t は2.01であるから

$$\frac{t_1 S}{\sqrt{n_1}} = \frac{(2.01)(44.02)}{\sqrt{50}} = 12.51$$

従つて50の標本は信頼区間の半分として10ではなく12.51である。最後に n は次の様にしてきまる

$$n = \frac{t_1^2 s^2}{d^2} = \frac{(4.040)(1938)}{100} = 78.3$$

即ち全体の $n = 79$ となる様に附加的な観測値を29とすればよい

§ 48 標本の大きさの問題の一般解法の一つの試み *An attempt at a general solution of the sample size problem*:-

この節で論ずる問題は標本の結果からある実際的な決定が行われる様な場合に最も分り易い。この様な決定はもしも標本推定量が誤差が非常に大である場合よりも小なる誤差をもつてある場合の方が理解し易い。

我々は金銭の項で推定量にある誤差量 Δ がある事によつて決定に際しての損失 $L(\Delta)$ を計算する事が出来る。 Δ の実際的な値は前以て予測出来ないけれども抽出理論は我々に特別な抽出方法に対しては標本の大きさ n によつて変る様な Δ の密度分布 $f(\Delta, n)$ を

見出す事を可能にしてくれる。

従つて与えられた標本の大きさに対する期待損失は

$$L(n) = \int L(\Delta) f(\Delta, n) d\Delta$$

標本をとる目的はこの損失を減少する事である。もしも $C(n)$ が標本の大きさ n をとるに要する費用とすれば、明らかに n は次のものを最小にする様にとられるべきである。

$$C(n) + L(n)$$

何故ならば、これは標本をとりそしてその結果から決定を下す際に含まれる全費用であるからである。 n の選択は標本の最適大きさ及び最も利益ある正確度の二つのものを決定する。

それとは別に同じ様な問題が標本の情報での誤差から生ずる損失の項よりもむしろ標本情報からもうかる金銭上の利益の項で提示する事が出来る。もしも金銭上の利益が用いられると標本の大きさ n からの期待利益 $G(n)$ が作られる。此処に $G(n)$ は標本がとられなければならないのである。我々は

$$G(n) - C(n)$$

を最大にする。この型に於ては原則は利益が最大となる様な古典的な経済に於ける規則と同じである。最も簡単な応用は損失函数 $L(\Delta)$ が入 Δ^2 である様な時に用いられる。此処に入は一定である。

従つて $L(n) = \lambda E(\Delta^2)$

例えば $\bar{y}$ が $\bar{Y}$ の標本推定量で $\Delta = \bar{y} - \bar{Y}$ とすれば simple

Random Sampling が用いられ fpc を無理すれば

$$L(n) = \lambda V(\bar{y}) = \frac{\lambda S^2}{n}$$

標本に対する費用函数の最も簡単な形

$$C(n) = C_0 + C_1 n$$

此処に C_0 は平均費用である。微分して、費用+損失を最小にする n の値は

$$n = \sqrt{\frac{\lambda S^2}{C_1}}$$

この結果のより一般的な型は *yates* (1949) によつて与えられる。同じ分析が推定量の分散が n に逆比例し費用が n の直線函数（一次函数）である場合の抽出及び推定の方法に適用される。

Blythe (1945) は販売の目的で棚積みの本枝数積の推定にこの原則を適用してゐる。

Nordin (1944) は製造者が入れたい市場に於ける可能な販売を推定するための標本の最適な大きさを論じてゐる。もしも販売が前以て正確に予測出来れば単位期間当の生産及び設備の量が製造者の期待利益を最大にする様に割あてゐる事が出来る。

この方法の応用は標本が特別な目的でとれる場合に限られる様であるけれども標本の大きさの問題に対しての此の方法は未だ実現されてない多くの応用面を残している様である。

それに関連した問題は一山が推定された品質を基として受け入れるかすてるかをきめるために大量生産過程での品物の山の標本抽出の問題がある。この場合標本の目的は Y の n の簡単な決定をする場合であるから決定に際しての誤差の結果を検定する事によつてきめる事が出来る。その方法の良い説明は *Tippett* (1950) 及び *Deming* (1950) が与えてゐる。

第5章 層化無作為抽出法

Stratified random sampling

§ 5.1 結 言 *Description*

層化抽出では、 N 個の *unit* をもつ母集団が夫々 $N_1, N_2, \dots, N_L$ なる *unit* をもつ副母集団にまづ分けられる。これ等の副母集団は重複することなく一様に母集団全体を構成する。即ち

$$N_1 + N_2 + \dots + N_L = N$$

副母集団は層 (*strata*) と呼ばれる。層化する事によつて充分な利益を得るためには、 N_k の値が分つていなければならない。層が決まると標本は各層から抽出され抽出は層毎に独立に行われる。層内の標本の大きさを夫々 $n_1, n_2, \dots, n_L$ で表わす。

もしも単純無作為抽出が各層で行われるとその方法は層化無作為抽出法と呼ばれる層化はごく普通に行われる方法である。これには多くの理由がある。主な理由をあげれば次の如くである。

- i. もしも既知の正確度をもつ資料が母集団のある部分でほしい場合には、各部分をその正当な要求として母集団として取扱う事がよい。
- ii. 層化を用いると行政上便利である。例えば調査を指導する機関はその各々が母集団の一つの部分に対して調査を監督出来る事務所をもつ事ができる。
- iii. 抽出問題は母集団の異なつた部分では著しく違つてくる。人間母集団では管造物（例えばホテル、病院、刑務所）に任んでいる人は普通の家に任んでいる人とは違つた層にいれる。何故ならば、抽出方法の違つた手引が二つの立場に対して用いられるからである。
- iv. 層化は、全母集団の特性の推定に際して正確度に利益をもたらす、基礎的な考え方は、不斉一な母集団を内部的に斉一な副集団に分ける事が可能であるという事である。これは層元の区分の意味で以て層という名によつて暗示される。もしも測定が

unit 毎に殆んど違わないという事に於て層が有るであれば層平均の正確な推定量はその層での小標本から得る事が出来る。

後で分る様にこれ等の推定量は、全母集団に対する正確な推定量に組合す事が出来る。

層化抽出の理論は層化標本からの推定量の特性及び最大正確度を得るための標本の大きさ n_h の最良選択の問題を取扱う。

この論議の展開にあたって勿論層は已に作られていると考える層を如何にして作るか及びどの位の数の層が必要かの前以ての問題は後の段階として後に述べる。(§ 5.15)

§ 5.2 記号 Notation

添字 h は層を表わし、 i は層内の unit を表わす。記号は前に用いたものの自然的な拡張である。次の記号はすべて層 h に関するものである。

N_h unit の総数
 n_h 標本での unit の数
 y_{hi} i 番目の unit で得られた値

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \quad \text{真の平均値}$$

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \quad \text{標本平均}$$

$$S_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1} \quad \text{真の分散}$$

分散に対しての分母は $(n_h - 1)$ である事に注意されたい。

§ 5.3 推定量の特性 Properties of the estimates :-

unit 当りの母平均に対して層化抽出に用いられる推定量の最も簡単な型は $\bar{y}_{st}$ (stratified に対して st をつけた) である。此処に

$$\bar{y}_{st} = \frac{\sum_{h=1}^L N_h \bar{y}_h}{N} \quad (5.1)$$

此処に $N = N_1 + N_2 + \dots + N_L$

推定量 $\bar{y}_{st}$ は一般に標本平均と同じではない。何故なら標本平均 $\bar{y}$ は次の様にかく事が出来るからである。

$$\bar{y} = \frac{\sum_{h=1}^L n_h \bar{y}_h}{n} \quad (5.2)$$

その相違は $\bar{y}_{st}$ では、個々の層からの推定量は補正の重み N_h/N をうけているという事である。 $\bar{y}$ は仮りに各層に於て

$$\frac{n_h}{n} = \frac{N_h}{N} : \text{又は} \frac{n_h}{N_h} = \frac{n}{N} = \text{一定}$$

とすれば、 $\bar{y}_{st}$ と一致する。

これは抽出比がすべての層で同じであることを意味する。この層化は n_h の比例割当 (proportional allocation) をした層化と呼ばれる。それは自己荷重をもつ標本を与える。もしも多くの推定量が必要であれば、自己荷重標本は時間の節約になる。

推定量 $\bar{y}_{st}$ の主な特性は次の定理で概略述べられる。最初の二つの定理は一般に層化抽出に用いられ、層化無作為抽出に限定されない。即ち如何なる層からの標本も単純無作為標本である必要はない。

定理 5.1

もしもすべての層に於て標本推定量 $\bar{y}_h$ が不偏であれば、 $\bar{y}_{st}$ は母平均 $\bar{Y}$ の不偏推定量である。

(証明) 推定量は個々の層で不偏であるから

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

併し、母平均 $\bar{Y}$ は次の様にかける

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

証明終り。

系 $\bar{y}_h$ は層内で単純無作為抽出に対して、 $\bar{Y}_h$ の不偏推定量であるから、 $\bar{y}_{st}$ は層化無作為抽出に対して $\bar{Y}$ の不偏推定

量である。

定理 5.2

層化抽出に対して母平均 $\bar{Y}$ の推定量 $\bar{y}_{st}$ の分散は

$$V(\bar{y}_{st}) = \frac{\sum_{h=1}^L N_h^2 V(\bar{y}_h)}{N^2} \quad (5.3)$$

此処に

$$V(\bar{y}_h) = E(\bar{y}_h - \bar{Y}_h)^2$$

之の定理には次の二つの制限がある：即ち (i) $\bar{y}_h$ は $\bar{Y}_h$ の不偏推定量でなければならない。(ii) 標本は層毎に独立に抽出されねばならない。

(証明)

$$\begin{aligned} \bar{y}_{st} - \bar{Y} &= \frac{\sum N_h \bar{y}_h}{N} - \frac{\sum N_h \bar{Y}_h}{N} \\ &= \frac{\sum N_h (\bar{y}_h - \bar{Y}_h)}{N} \end{aligned}$$

此処に和はすべての層で行われる。推定量に於ける $(\bar{y}_{st} - \bar{Y})$ なる誤差は個々の層内で行われた推定の誤差の荷重平均として表わされる事に注意せよ。従つて

$$(\bar{y}_{st} - \bar{Y})^2 = \frac{\sum N_h^2 (\bar{y}_h - \bar{Y}_h)^2}{N^2} + \frac{2 \sum N_h N_j (\bar{y}_h - \bar{Y}_h)(\bar{y}_j - \bar{Y}_j)}{N^2}$$

此処に右辺の項は層のすべての対に対して及ぼされる。

さてすべての可能な標本について平均する。積の項に対してはまず層 h での標本を固定して層 j でのすべての標本上で平均する。抽出は二つの層では独立であるからどんな標本が層 h に於て抽出されようと層 j に於ける可能な標本は同じで又同一の確率をもつ。併し、 $\bar{y}_j$ は不偏であると仮定されるから $(\bar{y}_j - \bar{Y}_j)$ の平均値は 0 である。従つて積の項はなくなる。

平方した項は

$$V(\bar{y}_{st}) = \frac{\sum N_h^2 E(\bar{y}_h - \bar{Y}_h)^2}{N^2} = \frac{\sum N_h^2 V(\bar{y}_h)}{N^2}$$

この結果について重要な点は、 $\bar{y}_{st}$ の分散はたゞ個々の層平均

$\bar{Y}_h$ の推定量の分散のみによるという事である。もしもすべての項目が層内で同じ値をもつ様に非常に変動の多い母集団を層に分割する事が可能であれば誤差なしに $\bar{Y}$ を推定する事が出来よう。証明の検定は、この結果を導く補正層荷重 N_h の使用であるという事を示す。

定理 5.3

層化無作為抽出に対して、推定量 $\bar{y}_{st}$ の分散は

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{S_h^2}{n_h} \quad (5.4)$$

(証明)

$\bar{y}_h$ は $\bar{Y}_h$ の不偏推定量であるから、定理 5.2 が適用出来る。更に定理 5.2 により個々の層に適用出来る。

$$V(\bar{y}_h) = \frac{S_h^2}{n_h} \frac{(N_h - n_h)}{N_h} \quad (5.5)$$

定理 5.2 の結果に代入すると

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L N_h^2 V(\bar{y}_h) = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

之の公式のある特別の場合を次の系に示す。

系 1

もしも抽出比 n_h/N_h がすべての層で無視出来れば

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum \frac{N_h^2 S_h^2}{n_h} = \sum \frac{W_h^2 S_h^2}{n_h} \quad (5.6)$$

此処に、 $W_h = N_h/N$ は層荷重である。これは有限母集団補正が無視し得るときに用いる公式である。層化抽出では因子は各層に個々に入ってくるから一般に単一の有限母集団補正ではない。

系 2

比例割当では、(5.4) 式に $n_h = \frac{n N_h}{N}$ を代入する。

分散は次の如くなる。

$$V(\bar{y}_{st}) = \sum \frac{N_h}{N} \frac{S_h^2}{n} \left(\frac{N - n}{N} \right) = \frac{(N - n)}{nN} \sum W_h S_h^2 \quad (5.7)$$

此の場合、単一の f.p.c. $(N-n)/N$ である。

系 3

もしも抽出が比例的ですべての層での分散が同じ値 S_w^2 であれば、次の如く簡単となる。

$$V(\bar{y}_{st}) = \frac{S_w^2}{n} \left(\frac{N-n}{N} \right) \quad (5.8)$$

定理 5.4

もしも $\hat{y}_{st} = N\bar{y}_{st}$ が母集団総計 Y の推定量であれば

$$V(\hat{y}_{st}) = \sum N_h(N_h - n_h) \frac{S_h^2}{n_h} \quad (5.9)$$

これは定理 5.3 から直ちに導かれる。

(例) 5.1 表はアメリカ合衆国の 64 の大都市の千単位での 1920 年及び 1930 年の居住者の数を示す。資料は 1920 年に居住者の総数でアメリカ合衆国で 5 番目から 68 番目迄の順位にある都市をとったものである。都市は二つの層に分けてある。その一つは 16 の最大の都市で、二番目の層は残り 48 都市である。

64 のすべての都市の居住者の 1930 年の総数は大きさ 24 の標本から推定される。次の抽出法に対して推定量の標準誤差を見出せ

(i) 単純無作為抽出法 (ii) 比例割当をした層化無作為抽出法 (iii) 各層から抽出した 12 の unit をもつ層化無作為抽出法

(解) この母集団はある unit 即ち大都市が総計に非常に影響するし残りのものよりも変動が非常に多いという事に於て多くの企業型体の母集団と類似している。

5.1 表 1920 年及び 1930 年の 64 都市の大きさ (1000 単位)

1920 年の大きさ (x_{hi})				1930 年の大きさ (y_{hi})			
層				層			
$h=1$	2			1	2		
777	314	172	121	900	364	209	113
773	298	172	120	822	317	183	115
748	296	163	119	781	328	163	123
734	258	162	118	805	302	253	154
588	256	161	118	670	288	232	140
577	243	159	116	1238	291	260	119
507	238	153	116	573	253	201	130
507	237	144	113	634	291	147	127
457	235	138	113	578	308	292	100
438	235	138	110	487	272	164	107
415	216	138	110	442	284	143	114
401	208	138	108	451	255	169	111
387	201	136	106	459	270	139	163
381	192	132	104	464	214	170	116
324	180	130	101	400	195	150	122
315	179	126	100	366	260	143	134

注意 都市は両年共同の順序で配列してある。

総計及び平方和

層	1920		1930	
	$\sum(x_{hi})$	$\sum(x_{hi}^2)$	$\sum(y_{hi})$	$\sum(y_{hi}^2)$
1	8,349	4,756,619	10,070	7,145,450
2	7,541	1,474,871	9,498	2,141,720

層総計及び平方和は 5.1 表の下に示してある。1930 年の完全母集団に対しては

(76)

$$Y = 19.568 : S^2 = 52.448$$

三つの Y の推定量を $\hat{Y}_{ran}$, $\hat{Y}_{prop}$ 及び $\hat{Y}_{equal}$ で表わす。

(i) 単純無作為抽出法 定理 2.2 系 2 から

$$V(\hat{Y}_{ran}) = \frac{N^2 S^2 (N-n)}{n} = \frac{(64)^2 (52.448)}{24} \left(\frac{40}{64} \right) = 5.594.453$$

$$\text{標準誤差は } \sigma(\hat{Y}_{ran}) = 2365$$

(ii) 個々の層に対して、分散は

$$S_1^2 = 53.843 : S_2^2 = 55.81$$

最大の都市を含む層は他の層の殆んど 10 倍の分散である事に注意されよ

$$\text{比例割当では } n_1 = 6 \quad n_2 = 18$$

公式 (5.7) から N^2 倍して

$$V(\hat{Y}_{prop}) = \frac{(N-n)}{n} \sum N_h S_h^2 = \frac{40}{24} \{ (16)(53.843) + (48)(55.81) \} = 1.882.293$$

$$\sigma(\hat{Y}_{prop}) = 1372$$

(iii) $n_1 = n_2 = 12$ に対して一般公式 (5.9) を用いる。

$$V(\hat{Y}_{equal}) = \sum N_h (N_h - n_h) \frac{S_h^2}{n_h} = \frac{(16)(4)(53.843)}{12} + \frac{(48)(36)(55.81)}{12} = 1.090.827$$

$$\sigma(\hat{Y}_{equal}) = 1044$$

この例では、二つの層で標本の大きさが等しい場合が比例割当よりも正確度がよくなっている。共に単純無作為抽出法よりもすぐれている。此処では利用しないが 1920 年の資料は後の例で用いられる。

§ 5.4 推定分散及び信頼限界 The estimated variance and confidence limits

単純無作為標本が各層内でとられると S_h^2 の不偏推定量は (定理 2.4 から) 次の如くである。

(77)

$$s_h^2 = \frac{1}{(n_h - 1)} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 \quad (5.10)$$

従つて次の定理を得る。

定理 5.5

層化無作為抽出では $\bar{y}_{st}$ の分散の不偏推定量は

$$v(\bar{y}_{st}) = s^2(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{s_h^2}{n_h} \quad (5.11)$$

計算の目的に対しては次の別な形を用いる。

$$s^2(\bar{y}_{st}) = \sum_{h=1}^L \frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^L \frac{W_h^2 s_h^2}{N} \quad (5.12)$$

右辺の二番目の項は f.p.c. による減少を表わす。

この推定量が計算出来るためには、各層から少くとも二つの単位が抽出されねばならない。たゞ一つの単位だけが選ばれる点で層化が実行される時の分散の推定は、§ 5.21 で論ずる。

系

ある応用の場合には S_h^2 はすべての層で同じ値であると仮定する事が合理的な事がある。標本の分散分析から、この普通の分散のまとめた推定量は

$$s_{st}^2 = \frac{\sum_{h=1}^L \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{(n - L)}$$

標本抽出は通常この場合には比例的であるから、 $\bar{y}_{st}$ の分散は次の簡単な形となる (定理 5.3、系 3 から)

$$v(\bar{y}_{st}) = \frac{s_{st}^2 (N-n)}{n}$$

此処に自由度は $(n - L)$ である。

信頼限界に対する公式は次の如くである。

$$\text{母平均} : \bar{y}_{st} \pm t s(\bar{y}_{st}) \quad (5.13)$$

$$\text{母集団総計} : N \bar{y}_{st} \pm t N s(\bar{y}_{st}) \quad (5.14)$$

これ等の公式は $\bar{y}_{st}$ は正規分布し、 $s(\bar{y}_{st})$ は良く決定されると仮定しているから乗数は正規分布の表から読む事が出来る。

もしも、各層自由度の数が少なければ、 $s(\bar{y}_{st})$ なる量につける抽出誤差を考慮に入れる通常の方法は正規分布の表の代りに、Student の t 表からの t の値を用いる事である。 $s(\bar{y}_{st})$ の分布は一般には非常に複雑であるから、この方法の厳密な応用は許されない。 $s(\bar{y}_{st})$ に有効な自由度の数を割当ててる近似方法は次の通りである。(Satterthwaite, 1946)

次の如くかく

$$s^2(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L f_h s_h^2$$

此処に $f_h = \frac{N_h(N_h - n_h)}{n_h}$

有効な自由度の数 n_e は

$$n_e = \frac{(\sum f_h s_h^2)^2}{\sum \frac{f_h^2 s_h^4}{n_h - 1}} \quad (5.15)$$

n_e の値は常に $(n_h - 1)$ なる値の最小値と、それ等の和との間にある。近似は s_h^2 は層毎に異なるという事実を考慮に入れているが、 y_{hi} なる値は正規に分布すると仮定しもしもこの条件が満足される時のみに値がある。

§ 5.5 最適割当て Optimum allocation :-

割当ての問題は夫々の層に於ける標本の大きさ n_h の選択に関するものである。定理 5.6 は Isaksson (1923) 及び Neyman (1934) により与えられた重要な結果である。

定理 5.6

層化無作為抽出では、推定平均 $\bar{y}_{st}$ の分散は標本の一定の総数に対してもしも標本が $N_h S_h$ に比例した n_h で割当てられると最小である。

証明

問題は $n_1 + n_2 + \dots + n_L = n$

なる拘束の下に

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

を最小にする事である。

我々は次を最小にする様に n_h 及び Lagrange 乗数 λ を選ぶ。

$$\begin{aligned} V(\bar{y}_{st}) + \lambda(n_1 + n_2 + \dots + n_L - n) \\ = \frac{1}{N^2} \sum \left(\frac{N_h^2}{n_h} - N_h \right) S_h^2 + \lambda(n_1 + n_2 + \dots + n_L - n) \end{aligned}$$

n_h に関して微分して次の方程式を得る。

$$-\frac{N_h^2}{N^2} \frac{S_h^2}{n_h^2} + \lambda = 0 \quad (5.16)$$

これは次の如くなる。

$$n_h = \frac{N_h S_h}{N \sqrt{\lambda}} \quad \text{又は} \quad n_h \propto N_h S_h \quad (5.17)$$

n_h の実際の値を見出すため、層の上で (5.17) を加える。すると

$$\sum n_h = n = \frac{\sum N_h S_h}{N \sqrt{\lambda}} \quad (5.18)$$

(5.18) に $\sqrt{\lambda}$ を代入して

$$n_h = n \frac{N_h S_h}{\sum N_h S_h} \quad (5.19)$$

この結果は層内での標本の大きさは層の大きさと層の標準偏差に比例する。換言すれば抽出比 n_h/N_h は標準偏差に比例しなければならぬ事を示す。他のものは等しいからより大きな標本が変動の多い層には必要である。

最小分散そのものに対する表示は $V(\bar{y}_{st})$ に対する一般公式に (5.19) により与えられた n_h の値を代入して得られる。これは次の如くである。

$$V_{min}(\bar{y}_{st}) = \frac{1}{N^2} \frac{(\sum N_h S_h)^2}{n} - \frac{1}{N^2} \sum N_h S_h^2 \quad (5.20)$$

§ 5.6 色々な費用での最適割当て Optimum allocation with

Varying Costs :-

より一般的な道路は特別な総費用に対しての最適割当てを考える事である。これに対して標本の大きさ n_h の頂で標本をとる費用を表わす費用函数を必要とする。

我々は次の簡単な函数だけを考える。

$$\text{費用} = C = a + \sum C_h n_h \quad (5.21)$$

層内では費用は直接標本の大きさによるが単位当りの費用 c_h は層毎に異なる。記号 a は一般費用を表わす。

定理 5.7

(5.21) なる形の費用函数では、推定平均 $\bar{y}_{st}$ の分散は n_h が $N a S_h / \sqrt{C_h}$ に比例するとき最小である。

証明

これは定理 5.6 の証明に同じである。次のものを最小にする。

$$V(\bar{y}_{st}) + \lambda (a + \sum C_h n_h) \\ = \frac{1}{N^2} \sum \left(\frac{N a^2}{n_h} - N a \right) S_h^2 + \lambda (a + \sum C_h n_h) \quad (5.22)$$

n_h に関して微分すると次の方程式を得る。

$$-\frac{N a^2 S_h^2}{N^2 n_h^2} + \lambda C_h = 0$$

$$\text{即ち} \quad n_h \sqrt{\lambda} = \frac{N a S_h}{N \sqrt{C_h}} \quad (5.23)$$

すべての層の上で加えると

$$n \sqrt{\lambda} = \sum \frac{N a S_h}{N \sqrt{C_h}} \quad (5.24)$$

最後に (5.23) と (5.24) の比をとれば

$$\frac{n_h}{n} = \frac{\frac{N a S_h}{\sqrt{C_h}}}{\sum \frac{N a S_h}{\sqrt{C_h}}} \quad (5.25)$$

この定理は手引になる次の規則をもたらす。あたえられた層で
もしも

- (i) 層が非常に大であれば
 - (ii) 層が非常に変動が大きければ
 - (iii) 抽出が層内で低廉であれば
- より大なる標本を要する。

更に進んだ一つの段階は割当てを計算する事が必要である事である。方程式 (5.25) は n_h の頂で n_h を与えるが我々は未だどの様な n_h の値であるか分らない。解は標本が特別な総費用 C に合うように又は $\bar{y}_{st}$ に対して特別な分散 V を与える様に選ばれるかというかによる。もし費用が一定であれば費用函数 (5.21) に於て、 n_h の最適の値を代入し n について解けばよい。これは

$$n = \frac{(C-a) \sum \frac{N a S_h}{\sqrt{C_h}}}{\sum N a S_h \sqrt{C_h}}$$

もしも V が一定であれば、我々は $V(\bar{y}_{st})$ に対する公式に最適 n_h を代入する。

即ち

$$n = \frac{(\sum W_h S_h \sqrt{C_h}) (\sum \frac{W_h S_h}{\sqrt{C_h}})}{V + \frac{1}{N} \sum W_h S_h^2}$$

此処に $W_h = N a / N$

§ 5.7 層化無作為抽出と単純無作為抽出の相対的正確度

Relative precision of stratified random and simple random sampling :-

うまく用いると層化抽出は殆んどいつもそれと比較出来る単純無作為抽出によつて与えられる推定平均又は推定総計に対してより小さい分散を与える。併しどの層化無作為標本も単純無作為標本よりは小さい分散を与えるとはいえない。

n_h の値が最適からかけはなれてくると、層化抽出は大きな分散を与える。事実、あてで分る様に一定の標本の総数に対して最適割当てをした層化抽出でさえも、この結果は実際に起りそうなものよりもむしろ學問的に奇異にさえ感ずるけれども大きな分散を与える。

この節では単純無作為抽出と比例割当て及び最適割当てをした層化無作為抽出との比較をする。この比較は如何なる方法で層化により利益が得られるかを示す助けとなるものである。先づ第一に f, p, c を無視する。こうすれば非常にすつきりした結果が得られるからである。より一般的な結果は後に示す。

推定平均の分散を夫々 V_{ran} , V_{prop} 及び V_{opt} で示す。
定理 5.8

n_h/N_h の項を無視すれば

$$V_{opt} \leq V_{prop} \leq V_{ran} \quad (5.26)$$

此処に最適割当ては一定 n 即ち $n_h \propto N_h S_h$ である
証明

f, p, c を無視すれば

$$V_{ran} = \frac{S^2}{n}$$

$$V_{prop} = \frac{\sum N_h S_h^2}{nN} \quad [\S 5.3 \text{ 方程式 (5.7) から}]$$

$$V_{opt} = \frac{(\sum N_h S_h)^2}{nN^2} \quad [\S 5.5 \text{ 方程式 (5.20) から}]$$

層化母集団の分散分析に対して標準の代数的等式より

$$\begin{aligned} (N-1)S^2 &= \sum_h \sum_i (y_{hi} - \bar{y})^2 \\ &= \sum_h \sum_i (y_{hi} - \bar{y}_h)^2 + \sum_h N_h (\bar{y}_h - \bar{y})^2 \\ &= \sum_h (N_h - 1) S_h^2 + \sum_h N_h (\bar{y}_h - \bar{y})^2 \quad (5.27) \end{aligned}$$

$1/N_h$ の項は無視出来るから此は次の如くかける。

$$NS^2 = \sum_h N_h S_h^2 + \sum_h N_h (\bar{y}_h - \bar{y})^2$$

依つて

$$\begin{aligned} V_{ran} &= \frac{S^2}{n} = \frac{\sum N_h S_h^2}{nN} + \frac{\sum N_h (\bar{y}_h - \bar{y})^2}{nN} \\ &= V_{prop} + \frac{\sum N_h (\bar{y}_h - \bar{y})^2}{nN} \quad (5.28) \end{aligned}$$

V_{opt} の定義により $V_{prop} \geq V_{opt}$ でなければならぬ。

この差は

$$\begin{aligned} V_{prop} - V_{opt} &= \frac{1}{nN} \left\{ \sum N_h S_h^2 - \frac{(\sum N_h S_h)^2}{N} \right\} \\ &= \frac{1}{nN} \sum N_h (S_h - \bar{S})^2 \quad (5.29) \end{aligned}$$

此処に $\bar{S} = \sum N_h S_h / N$ 方程式 (5.29) 及び (5.28) から

$$V_{ran} = V_{opt} + \frac{\sum N_h (S_h - \bar{S})^2}{nN} + \frac{\sum N_h (\bar{y}_h - \bar{y})^2}{nN} \quad (5.30)$$

この結果は単純無作為抽出から最適割当てへ変化する時の分散の減少に対して二つの成分がある事を示す。第一の成分(一番右端の項)層平均内の差の除去により生じ、第二の成分(右辺の真中の項)は層標準偏差内の差の影響の除去により生ずる。第二番目の成分は最適割当てと比例割当ての間の分散の差を示す。

より一般的な結果

f, p, c を無視しなければ、同じ分析の型は次の結果を導く

$$V_{ran} = V_{prop} + \frac{(N-n)}{nN(N-1)} \left\{ \sum N_h (\bar{y}_h - \bar{y})^2 - \frac{1}{N} \sum (N - N_h) S_h^2 \right\} \quad (5.31)$$

従つて単純無作為抽出は

$$\sum N_h (\bar{y}_h - \bar{y})^2 \leq \frac{1}{N} \sum (N - N_h) S_h^2$$

でなければ比例層化より大なる分散を与えるという事になる。

$\bar{y}_h$ がすべて等しい場合があり得るから、この様な場合はあり得るであろう。もしも $\bar{y}_h$ 内に差があれば、左辺は N_h オーダーであるのに右辺は 1 のオーダーであるから小さい層以外には不等性は満足されそうにない。

最適割当てに対して此処に述べた結果は最適割当てが逐行出来る場合のみに抽出の実行面に応用出来る。その様な試みは次の節で論ぜられる多くの問題をかもし出す。

§ 5.8 最適からのかけはなれの影響

Effects of deviations from the optimum :-

最適割当てに対する公式を実際に適用するのは未だ充分ではな

い。我々はどんなときに比例割当ての正確度の利益が信頼出来るか又それ等がどんなときに小さい様であるかを亦知る必要がある。更に実際面では S_n は分らないから最適に対しての近似だけを得る事が出来るだけで、我々はもしも割当てが最適からある程度かけはなれたらどの位正確度に損失があるかを知らねはならない。これ等の問題の解明は χ, P, C を無視したときに、二つの層を考える事により得られる。

この場合に於ける、 $\bar{y}_{st}$ の分散に対する一般公式は

$$V = \frac{1}{N^2} \left(\frac{N_1^2 S_1^2}{n_1} + \frac{N_2^2 S_2^2}{n_2} \right)$$

標本の大きさを一定にすると最適割当ては次の時に得られる。

$$n_{1,opt} = \frac{n N_1 S_1}{N_1 S_1 + N_2 S_2} \quad ; \quad n_{2,opt} = \frac{n N_2 S_2}{N_1 S_1 + N_2 S_2}$$

この値を一般公式に代入すると

$$V_{opt} = \frac{(N_1 S_1 + N_2 S_2)^2}{N^2 n}$$

最適割当てと比較したときの一般割当ての相対的正確度 (RP) は次により与えられる。

$$\begin{aligned} RP = \frac{V_{opt}}{V} &= \frac{(N_1 S_1 + N_2 S_2)^2}{n \left\{ \frac{N_1^2 S_1^2}{n_1} + \frac{N_2^2 S_2^2}{n_2} \right\}} \\ &= \frac{(N_1 U + N_2)^2}{n \left\{ \frac{N_1^2 U^2}{n_1} + \frac{N_2^2}{n_2} \right\}} \end{aligned} \quad (5.33)$$

此処に $U = S_1/S_2$

我々は最適の附近で、 n_1 と n_2 の値のシリーズに対してこの量を検定したい。最適な値からの n_1 及び n_2 のかけはなれは次の此の項で表わす事が出来る。

$$\varphi = \frac{n_1}{n_2} / \frac{n_{1,opt}}{n_{2,opt}} = \frac{n_1}{n_2} / \frac{N_1 S_1}{N_2 S_2} = \frac{n_1 N_2 S_2}{n_2 N_1 S_1} = \frac{n_1 N_2}{n_2 N_1 U}$$

従つて

$$\frac{n_1}{n_2} = \frac{\varphi N_1 U}{N_2} \quad (5.34)$$

$n = n_1 + n_2$ であるから次の如く n_1/n_2 の項で方程式 (5.33) を書き直す事が出来る。

$$RP = \frac{(N_1 U + N_2)^2}{\left(1 + \frac{n_1}{n_2}\right) \left(\frac{N_1^2 U^2 n_2}{n_1} + N_2^2\right)}$$

(5.34) からの n_1/n_2 を代入して簡単にすると

$$RP = \frac{\varphi (N_1 U + N_2)^2}{(\varphi N_1 U + N_2) (N_1 U + \varphi N_2)}$$

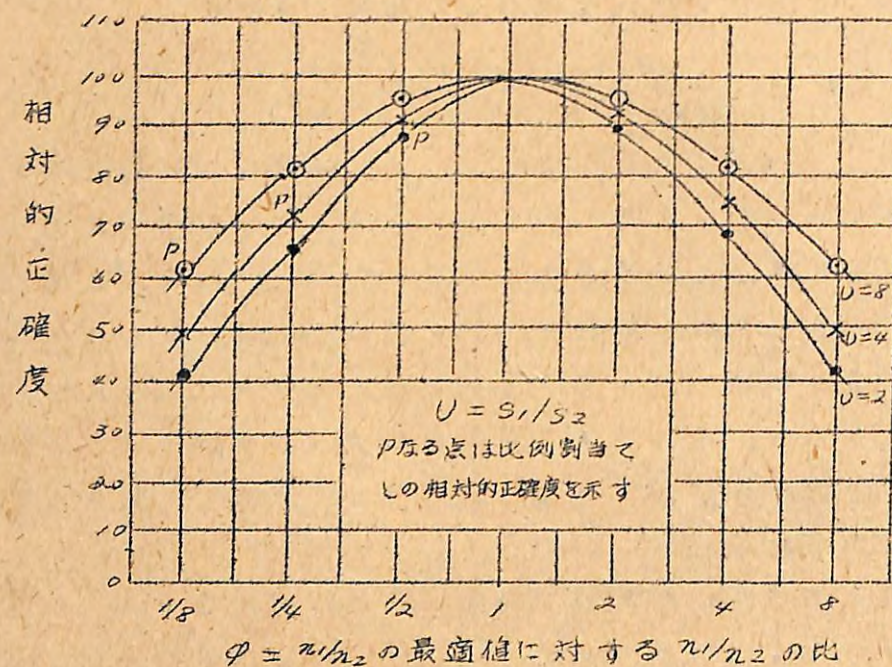
φ 及び U の与えられた値に対して、計算は相対的正確度が等しい場合からの層の大きさ N_1/N_2 の比に於ける適当なかけはなれに対して敏感でない事を示す。従つて $N_1 = N_2$ なる立場に対して結果が与えられるであろう。5.1 図は $S_1/S_2 = U = 2, 4, 8$ の三つの違つた値に対して φ に対してプロットされた RP (パーセント) を示す。 φ の尺度は対数である。各曲線上で比例割当てを示す φ の値が示されている。5.1 図を注意深く研究すると次の結論が暗示される。

- i. 最適比の 1/2 と 2 倍との間にある如何なる n_1/n_2 の値も、正確度の損失は三つの曲線に対して 10% 以下であるという意味で最適は安定している。
- ii. S_1/S_2 が 1 と 2 の間にあれば比例割当ての代りに最適割当てを行うのには異論はない。 $S_1/S_2 = 2$ では比例割当てによる正確度の損失は殆んど 10% で実際には近似的に最適であるというだけであるから小さいものである。(この結論は抽出比が大で、 χ, P, C がもはや無視出来なければ変える必要がある)
- iii. 比例割当ての相対的正確度は $S_1/S_2 = 4$ のとき 74%, $S_1/S_2 = 8$ のとき 62% となる。かくて S_1/S_2 が 2 以上になると最適から全くかけはなれた n_1 及び n_2 の値を得るしなは正確度に於ては比例割当て以上の価値ある増加を与える。

iv、一見して $U = \infty$ に対する曲線は最低で、 $U = 0$ に対しては最高であるという事には驚くであろう。この結果は選ばれた縦軸変量の結果である。

実際の分配規則として比例割当ては通常標本をとる前に推定したとき最適割当てからの正確度に於ける期待される利益が20%以上でなければ推奨出来る。

5.1図 最適割当てからのかけはなれによる正確度の損失



§ 5.7 以前の資料からの割当ての決定

Determination of the allocation from previous data

最適割当ては層の標準偏差 S_h の前以ての推定量を必要とする。これ等は以前の調査又はセンサスから得る事が出来る。

(例)

§ 5.3 の例は良好な以前の資料が手元にある場合である。5.1表 (p 75) に於て 64 の大都市は 1920 年の居住者の人数を基にして2つの層に分けられる。24 都市の層化無作為標本が 1930 年の総人口推定のためにとられる。一定の標本の太い

さに対して最適割当てが望ましい。1930 年の資料は 1930 年の標本が計画された時には利用出来ないから我々は 1920 年の S_h の値から割当てをする。計算は比較のため 1930 年の資料をも含めて 5.2 表に示してある。

5.2表 最適割当ての計算

層	N_h	1920 年の資料			1930 年の資料		
		S_h	$N_h S_h$	n_h	S_h	$N_h S_h$	n_h
1	16	163.30	2,612.80	11.56	232.04	3,712.64	12.21
2	48	58.55	2,810.40	12.44	74.71	3,586.08	11.79
合計	64		5,423.20	24.00		7,298.72	24.00

1920 年の資料は 1930 年の資料に対しては 12.21 なる真の最適に対してと同じ 11.56 なる n_h の値を示している。整数にまるめると二組の資料とも各層から 12 の標本の大きさをとる事に成り同じ割当てである。同時に 1930 年より 1920 年の S_h が小さい事に注意されたし。

1920 年の資料は割当ての計画に対してすぐれてはいるが、1930 年に得らるべき正確度の楽天的な感銘を与える。実際にはこの因子は常に考慮に入れておかねばならない。もしも都市の大きさが 1920 年と 1930 年の間に増大したとすれば、標本の大きさを選ぶに際して標本偏差に於けるそれに伴った増大に対して幾分斟酌されねばならぬ。

n_h は整数にまるめられるから常に理論的最適から正確度に於て幾分の損失がある。この損失は n_h が 6 位小さくても大して重要でない様である。標本に於けるこのまるめの影響は推定総計の分散の一般公式

$$V(\hat{y}_{st}) = \sum N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

に於て、 $n_1 = 12.21$ 、 $n_2 = 11.79$ を代入すると分る。これは

理論的最小分散 1090.157 を与える。12により実際に得られた分散即ち 12 の割当てが §5.3 の例で実行されたときの分散は 1090.827 であった。その差は僅少である。割当てが計画される時、比例割当てに対する相対的な正確度の利益を推定する事をおすすめする。§5.8 で我々は比例割当てはその自己荷重の特性の見地から最適割当てによる利益が少くとも 20% でなければ比例割当てがすぐれている事を暗示した。この資料では比例割当てとの比較は前の例 (§5.3) で行われた。相対的正確度は 1090.827/1090.157 即ち 100.07% であった。この種の計算は新しい調査での S_h は層毎に計算に用いた前の資料に同じ割合であると仮定するから比例割当てとの相対的な正確度の利益の過大視はされられないという事に注意されたい。

この例ではこの過大推定は無視出来る。何故ならば 1920 年の割当ては 1930 年のそれと同じであるからである。

§5.10 推定 S_h に於ける誤差の影響

Effects of errors in the estimated S_h

前の例では S_h の前以ての推定量が標本を割当てる際に完全にうまくゆく事が証明された。以前の資料がない場合は S_h の予測は割当ての基礎として用いるのに充分良好かどうかは疑わしい。

推定 S_h に於ける誤差の影響の検定は Evans (1951) により行われた。その分析は "推定最適" 割当てが比例割当てに比して利益があるかどうかを示す規則を導く。

割当てが S_h の推定量 $\hat{S}_h$ に基づいたと仮定しよう。即ち我々は

$$n_h = \frac{n N_h \hat{S}_h}{\sum N_h \hat{S}_h} \quad (5.35)$$

を用いる。

もしも記号 (CU) をしてすべての層で同じと仮定した S_h の変動係数を示すものとすれば、Evans は S_h の誤差による $V(\bar{y}_{st})$ に於ける平均の増大は近似的に

$$V(\bar{y}_{st}) - V_{opt} \doteq \frac{(CU)^2}{nN^2} \left\{ (\sum N_h S_h)^2 - \sum N_h^2 S_h^2 \right\}$$

になる事を示している。

併し方程式 (5.29) から

$$V_{prop} - V_{opt} = \frac{1}{nN} \left\{ \sum N_h S_h^2 - \frac{(\sum N_h S_h)^2}{N} \right\}$$

かくて平均して "最適" の層化はもしも

$$(CU)^2 < \frac{N \sum N_h S_h^2 - (\sum N_h S_h)^2}{(\sum N_h S_h)^2 - \sum (N_h S_h)^2}$$

であれば比例割当てよりも小さい分散を与える。

計算に対して、この結果は層の相対的な大きさ即ち $W_h = N_h/N$ の項で表わすにより便利である。

$$(CU)^2 < \frac{\sum W_h S_h^2 - (\sum W_h S_h)^2}{(\sum W_h S_h)^2 - \sum (W_h S_h)^2} \quad (5.36)$$

実際には我々はこの値よりも明らかに小さくすべき $(CU)^2$ を望まねばならない。何故ならば、この値に等しい $(CU)^2$ では仮定された "最適" がすぐれる確率は 50-50 のオーダーにすぎないからである。

(例) 1920 年の都市の大きさの資料 (§5.9, 5.1 表) に対してそれに関連した数字を 5.3 表に示す。

5.3 表 S_h に於ける許し得る誤差の推定

層	W_h	S_h	S_h^2	$W_h S_h$	$W_h S_h^2$
1	0.25	163.3	26,667	40.82	6.667
2	0.75	58.6	3,428	43.95	2.571

計算は次の如く行う。

$$\sum W_h S_h^2 = 9238 : (\sum W_h S_h)^2 = 7186 : \sum (W_h S_h)^2 = 3598$$

$$(CU)^2 < \frac{9238 + 7186}{7186 - 3598} = \frac{2052}{3588} = 0.57 \text{ 即ち } (CU) < 0.75$$

計算は 0.75 なる変動係数が S_h の推定には許され得るであろう事を示している。この値は恐らく用いられた近似の限界外にあるであろう。非常に貧弱な推定重が S_h の非常に違ってい

ると期待される時は、比例割当てよりも“最適”割当てをよくなるに充分である事を暗示している。

Sukhatme (1935) は選取が層化と単純無作為抽出との中間である時の SA^2 の前以ての推定量に対して必要な最初の標本の大きさについて論じた。彼は小さい最初の標本は“最適”割当てがより小さい分散をもつであろう高い確率を通常与える事を示している。Evans (1951) は平均して如何にして最適割当てが比例割当てよりもよくなるに必要な最初の標本の大きさを計算するかを示している。これ等の研究は正規分布に仮定している事に注意しなければならない。母集団が正規でない場合は最初の標本の大きさは SA^2 の分散が正規性からかけはなれる事に対して敏感であるから増さねばならぬであろう。(92, 93 参照)

5.5.1 数個の項目をもつ場合の割当て

Allocation with several items :-

一つの項目に対して最適である割当てが一般的に他の項目に対して最上のものではないからある妥協策が多くの項目をもつ場合の調査にとられねばならない。第一段階は最も重要と思われる相対的に小さな数に割当てに於て考えられる項目を減ずる事である。もしも以前の資料が利用出来れば各項目毎に最適割当ての計算をしてどの位一致していないかをみる事が出来る。特殊な型の調査では項目間の相関が高く割当てはお互にさして変らない場合がある。

(例)

Jessen (1942) により与えられた資料はこの種の農場調査の例である。Iowa 州がその主な農業企業によつて記しつけられた五つの地理的な地方に分けられた。これ等の地方は日々の農業で調査に層として用いられたと仮定しなさい。最も興味ある三つの項目は、一日当り乳をしぼる牛の数、一日当りミルクのガロン数及び日々の生産からの現金収入であった。1938 年に行われた調査から 5.4 表に示される層内推定標準偏差 SA を

得た。

5.4 表 層内標準偏差

層	$W_h = \frac{N_h}{N}$	SA 乳牛	SA ミルクのガロン	SA 日々の生産からの現金収入 (\$)
Northeast dairy	0.197	4.6	11.7	332
Cash grain	0.191	3.4	9.8	357
Western livestock	0.219	3.3	7.0	246
Southern pasture	0.184	2.8	6.5	173
Eastern livestock	0.208	3.7	9.8	279

我々は現在 SA は真の標準偏差であると仮定しなければならない。5.5 表には 1000 の農場の標本に於ての個々の項目に対する最適割当てが掲げてある。

5.5 表 層内の標本の大きさ

層	割 当 て				
	比 例	最 適			平 均 M_h
		牛	ガロン	収入	
Northeast dairy	197	258	258	236	250
Cash grain	191	182	209	246	212
Western livestock	219	203	171	194	189
Southern pasture	184	145	134	115	131
Eastern livestock	208	216	228	209	218

州には 200,000 以上の農場があるから F, P, C は無視される。かくて各項目に対しては

$$n_h = \frac{n W_h SA}{\sum W_h SA}$$

個々の最適割当てはお互に適当に違っているだけである。一つの例外以外はすべての三つは比例割当てから同じ方向にへだたっている。

かくて最初の層では比例割当てでは 197 農場を暗示し、個々の割当ては 236 と 258 の間の数である。三つの項目に対する最適の大きさの平均値が右辺の欄に与えてあり満足に中間的な割当てを与えている。

5.6 表は個々の最適、中間、比例割当てにより与えられたときの $\bar{y}_{st}$ の期待抽出分散を示している。

5.6 表 推定平均値の期待分散

割当ての型	牛	ガロン	収入
最適	0.0127	0.0800	76.9
中間	0.0128	0.0802	77.6
比例	0.0131	0.0837	80.9

公式は次の如くである。

$$V_{opt} = \frac{(\sum W_h S_h)^2}{n} : V_{comp} = \sum \frac{(W_h S_h)^2}{m_h} : V_{prop} = \frac{\sum W_h S_h^2}{n}$$

V_{comp} に対する公式は定理 5.3 系 1 に於ける $v(\bar{y}_{st})$ に対する一般公式から求められる。分母 m_h は中間割当てによる層で与えられる標本割当てである。中間割当ては各項目に対して個別の最適割当てを用い得たかの如く殆んど正確な結果を与えている。より注意すべき価値のあるものは、比例割当ては中間割当てよりも悪いけれども最適及び中間よりもひどくは正確度が劣っていないという事である。更に 5.6 表はこれ等の割当てが真の分散よりむしろ推定分散から行われたものであるから最適及び中間の正確度を過大視している。この結果はもしも割当てが最適と非常にかげはなれておれば推定量の分散は増加しないという点を指摘する重要な原則（前に論じた）を明らかにしたものである。5.6 表で比較すると比例割当ては推奨出来る選択法であろう。もしも S_h が層内で著しく違っておれば中間割当ては非常に満足なものであろう。

広範囲の項目についての調査では割当てはよりひどく違ってくる。最良中間はそのとき明らかでない。比例割当てがしばしばこの立場で用いられるとしてもそれよりすぐれた中間を見出す事は可能である。最初に我々は「最良」中間に対するある批判を必要とする。その一つの可能性は各々をその最適分散により分ける事によって項目に対する分散の和を最小にする事である。もしも f.p.c. を無視するとこれは次のものを最小にする n_h なる量を求める事である。

$$\sum' \frac{(\sum W_h S_h)^2}{n_h} : \frac{(\sum W_h S_h)^2}{n}$$

此処に $\sum'$ は項目についての和を表わす。併し広範囲の調査に於てそこに於て比例割当て以上の利益が後に於て「自己荷重」の形の利益を対照する事によつてひきたせるに充分である様な立場が普通あるかどうかをわからせる事が残っている。

§ 5.12 100% 以上の抽出を必要とする割当て

Allocation requiring more than 100 per cent sampling :-

割当ての計画に際して最適に対する公式はそれに応ずる N_h よりも大きな n_h をある層で生ずる事がある。§ 5.9 の都市の大きさの例を考えよう。2つの層の間に分布した 24 都市の標本は最初の層では 16 の中 12 都市、二番目の層では 4 から 12 都市を要した。標本の大きさが 48 であれば最初の層では割当ては、16 のや 24 都市を要するであろう。そこで行われ得る最良のものは公式により要求された 24 の代りに二番目の層に対して 32 都市をひき去つて第一の層のすべての都市をひく事である。この問題は全抽出比が非常に大きいときにのみ生じ、一つの層が他の層より変動が大である場合に生ずる。実際に二三の場合にこの様な事が生じた。

この際、この割当てに於ける期待分散を予測する正しい公式を用いる事又は他の割当てとの比較に際しての正しい公式を用いる

(94)

事に注意を払わねばならない。§5.3. 定理5.3での一般公式は、もしも修正最適割当てによつて与えられた n_h がその代りに用いられるとすれば使用出来る。一定の n に対する最小分散に対する公式 (5.20)

$$V_{\min}(\bar{y}_{st}) = \frac{1}{N^2} \frac{(\sum N_h S_h)^2}{n} - \frac{1}{N^2} \sum N_h S_h^2$$

はもはや正しくない。もしも層 l が過大抽出が示される唯一つの層であれば $V_{\min}$ に対する正しい公式は

$$V_{\min}(\bar{y}_{st}) = \frac{1}{N^2} \frac{(\sum' N_h S_h)^2}{n - N_l} - \frac{1}{N^2} \sum' N_h S_h^2 \quad (5.37)$$

此処に $\sum'$ は層 l 以外のすべての層についての和を表わす。

§5.13 標本の大きさの推定 Estimation of sample size :-

この問題はこの章では少し早く論じすぎたかもしれない。一方推定平均値及び総計の分散に対する公式は、 n_h 及び S_h を含むので標本の大きさは実際には S_h の前以ての推定量が利用されて割当てに対してある決定が行われる迄きめる事は出来ない。次の論議に於ては欲する正確度は誤差限界か又は信頼区間の半分即ち d の何れかで述べられると仮定しよう。

母平均の推定に対しては、層化に対する基礎公式は

$$d^2 = t^2 V(\bar{y}_{st}) \quad (5.38)$$

此処に t は欲する信頼確率に應ずる正規偏差である。 d 及び t は d^2/t^2 なる形のみで公式に入ってくるので、 $V = d^2/t^2$ とかける。此処に V は標本推定量での欲する分散である。 S_h をして S_h の推定量とし、 $n_h = W_h n$ としよう。此処に W_h は選ばれた場合である。

これ等の項で期待 $V(\bar{y}_{st})$ は (§5.3. 定理5.3 から)

$$\text{期待 } V(\bar{y}_{st}) = \frac{1}{N^2} \frac{\sum N_h^2 S_h^2}{n} - \frac{1}{N^2} \sum N_h S_h^2$$

即ち

$$V = \frac{1}{n} \sum \frac{W_h^2 S_h^2}{W_h} - \frac{1}{N} \sum W_h S_h^2 \quad (5.39)$$

(95)

此処に $W_h = N_h/N$ これは n に対する一般公式と同様

$$n = \frac{\sum \frac{W_h^2 S_h^2}{W_h}}{V + \frac{1}{N} \sum W_h S_h^2} \quad (5.40)$$

もしも f, p, c を無視すれば第一次近似として

$$n_0 = \frac{1}{V} \sum \frac{W_h^2 S_h^2}{W_h} \quad (5.41)$$

もしも n_0/N が無視出来ねば次の様にして n を計算する。

$$n = \frac{n_0}{1 + \frac{1}{N} \sum W_h S_h^2} \quad (5.42)$$

特別な場合には公式は計算に便利な色々な形をとる。次に二を
を示そう。

仮想最適割当て (Presumed optimum allocation) (n を一定)

: $W_h \propto W_h S_h$

$$n = \frac{(\sum W_h S_h)^2}{V + \frac{1}{N} \sum W_h S_h^2} \quad (5.43)$$

比例割当て (Proportional allocation) : $W_h = W_h = N_h/N$

$$n_0 = \frac{\sum W_h S_h^2}{V} : n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (5.44)$$

もしも V が推定母集団総計に対しての欲する分散であれば主な公式は次の如くなる。

$$\text{一般公式 : } n = \frac{\sum \frac{N_h^2 S_h^2}{W_h}}{V + \sum N_h S_h^2} \quad (5.45)$$

仮想最適 (一定 n) :

$$n = \frac{(\sum N_h S_h)^2}{V + \sum N_h S_h^2} \quad (5.46)$$

比 例 :

$$n_0 = \frac{N}{V} \sum N_h S_h^2 : n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (5.47)$$

(例)

この例は Cornell (1947) の論文からとったものである。1946-47 学年に対する入学者の推定のために U. S. Office of Education により 1946 年に抽出されたアメリカ合衆国の単科大学及び総合大学の標本である。此処に与えられる例は、196 の高等師範及び普通の学校の母集団に対してのものである。これ等は七つの層に配列され、その中の一つの小さな層は除かれた。最初の五つの層は学校の大きさにより作られ、六つ目のものは女のみの大学を含んでいる。Sh の推定値 sh は 1943-44 学年に対する結果から計算された。これ等の sh に基づいた最適層化が用いられた。

対象は推定総在籍者では 5% の変動係数であった。1943 年に於て単科大学のこのグループに対する総在籍者は 56,472 人であった。従つて欲する標準偏差は $(0.05)(56,472) = 2,824$ 従つて欲する分散は

$$V = (2,824)^2 = 7,974,776$$

在籍者は 1946 年には 1943 年よりも恐らく多いであろうし、この増加に対して考慮しなければならないであろうと考えられる。実際には計算は学校当り sh は 1943 年と 1946 年は同じだと仮定している。この仮定は不合理ではない。

5.7 表は n_h を決定する前に分っている。Nsh, sh 及び Nsh sh の値が示してある。

5.7 表 標本の大きさを推定する為の資料

層	Nsh	sh	Nsh sh	nsh
1	13	325	4,225	7
2	18	190	3,420	7
3	26	189	4,914	10
4	42	42	3,444	7
5	73	45	3,278	13
6	24	190	4,560	10
総計	196		26,841	56

n_h に対して用いた公式は、(5.46) である。これは総計推定に対しての最適割当てに用いる公式である。この母集団に於ける 196 unit だけでは、f.p.c. は無視出来るとは考えられない。併し 'R' 例であるから f.p.c. を無視した第一次近似を求めよう。これは

$$n_0 = \frac{(\sum N_h sh)^2}{V} = \frac{(26,841)^2}{7,974,776} = 90.34$$

補正は明らかに必要である。(5.46) に於ける補正 n_h に対しては

$$n_h = \frac{n_0}{1 + \frac{1}{V} \sum N_h sh^2} = \frac{90.34}{1 + \frac{4,642,387}{7,974,776}} = 57.1$$

56 の標本の大きさが選ばれた。* 個々の層に対する n_h は 5.7 表の右欄に示してある。

* 計算結果は Cornell (1947) のものとすこし違っている。
§ 5.14 割合に対する層化抽出 stratified sampling for proportions:

もしも我々があるきまつた class c に属する母集団での unit の割合を推定したければ、理想的な層化はもしも我々が第一の層に c に属するすべての unit をおき、二番目の層に属さない unit をおく事が出来れば達せられる。これに失敗すると我々は、class c に属する割合が層毎に出来るだけ変化する様に層を作る様に努める。

$$p_h = \frac{A_h}{N_h} : p_h = \frac{a_h}{n_h}$$

としよう。即ち前者は h 層に属する、c に属する unit の割合であるし、後者はその層からの標本での割合である。全母集団に対する割合に対して層化無作為抽出に用いる推定量は

$$p_{st} = \sum \frac{N_h p_h}{N} \quad (5.48)$$

定理 5.9

層化無作為抽出では、 p_{st} の分散は

$$V(P_{st}) = \frac{1}{N^2} \sum \frac{N_h^2(N_h - n_h)}{(N_h - 1)} \frac{p_h q_h}{n_h} \quad (5.49)$$

(証明)

これは推定平均値の分散の一般公式の特別な場合である。定理 5.3 から

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum N_h(N_h - n_h) \frac{s_h^2}{n_h}$$

y_{hi} をして、 C なる class に unit が属するとき 1 なる値をとり、それ以外のとき 0 なる値をとるとしよう。§ 5.2 に於て、この変量に対して

$$s_h^2 = \frac{N_h}{(N_h - 1)} p_h q_h$$

という事を示した。

これで証明終り。

注意

実際のなすすべての応用に当つては、 f, p, c が無視出来なくとも $1/N_h$ なる項は無視出来るであろう。そしてかなり簡単な公式

$$V(P_{st}) = \frac{1}{N^2} \sum N_h(N_h - n_h) \frac{p_h q_h}{n_h} \quad (5.50)$$

を用い得る。

系 1. f, p, c を無視出来る時は

$$V(P_{st}) = \frac{1}{N^2} \sum N_h^2 \frac{p_h q_h}{n_h} = \sum W_h^2 \frac{p_h q_h}{n_h} \quad (5.51)$$

系 2. 比例割当てでは

$$V(P_{st}) = \frac{(N - n)}{N} \frac{1}{nN} \sum \frac{N_h^2 p_h q_h}{(N_h - 1)} \quad (5.52)$$

$$\doteq \frac{(N - n)}{N} \frac{1}{n} \sum W_h p_h q_h \quad (5.53)$$

分散の標本推定量に対して、我々は上の公式のどれにでも未知の p_h, q_h の代りに p_h, q_h を用いる。これは不偏推定量を手

えないけれども信頼限界の計算には適当である。

$V(P_{st})$ を最小にするための n_h の最良選択は、§ 5.5 及び § 5.6 での一般理論から導かれる。

○ 一定の全標本の大きさに対する最小分散

$$n_h \propto N_h \sqrt{\frac{N_h}{N_h - 1}} \sqrt{p_h q_h} \doteq N_h \sqrt{p_h q_h}$$

従つて

$$n_h \doteq n \frac{N_h \sqrt{p_h q_h}}{\sum N_h \sqrt{p_h q_h}} \quad (5.54)$$

○ $Cost = a + \sum C_h n_h$ である一定費用に対する最小分散

$$n_h \doteq n \frac{N_h \sqrt{\frac{p_h q_h}{C_h}}}{\sum N_h \sqrt{\frac{p_h q_h}{C_h}}} \quad (5.55)$$

n の値は、§ 5.6 と同様に見出される。

○ 最適対比例割当て

もしも unit の抽出の費用がすべての層で同一で p_h が 0.1 と 0.9 の間にあれば、最適割当てが比例割当てより利益がある事はほんのすこしである。

f, p, c を無視して一定 n に対して即ち $n_h \propto N_h \sqrt{p_h q_h}$ で以て最適割当てが行われたとしよう。次の事が分る。

$$V_{opt} = \frac{(\sum W_h \sqrt{p_h q_h})^2}{n} : V_{prop} = \frac{\sum W_h p_h q_h}{n}$$

従つて最適割当てに対する比例割当ての相対的正確度は

$$\frac{V_{opt}}{V_{prop}} = \frac{(\sum W_h \sqrt{p_h q_h})^2}{\sum W_h p_h q_h}$$

もしも p_h が p_0 と $(1 - p_0)$ なる二つの値の間にあれば相対的正確度がとる最小の値に興味がある。簡単のため、等しい大きさ ($W_1 = W_2$) の二つの層を考えよう。最小相対的正確度は $p_1 = 1/2$ 及び $p_2 = p_0$ のときに得られる。

そのとき相対的正確度は次の如くなる。

$$\frac{V_{opt}}{V_{prop}} = \frac{(0.5 + \sqrt{p_0 q_0})^2}{2(0.25 + p_0 q_0)}$$

この函数のいくつかの値が 5.8 表に示してある。 p_0 が 0.1 に等しいか又は 0.9 位大きくても相対的正確度は 94% である。大抵の場合、比例割当との簡単さ及び自己荷重の特性は正確度のこのわづかな損失をおきながらあまりあるものである。

5.8 表 最適割当とに対する比例割当の相対的正確度

p_0	0.4 又は 0.6	0.3 又は 0.7	0.2 又は 0.8	0.1 又は 0.9	0.05 又は 0.95
RP (%)	100.0	99.8	98.8	94.1	86.6

この例の限界は注意しなければならない。違った層での抽出の違った費用を考慮に入れている。 p_0 が非常に小さいか例えば違った層で 0.001 から 0.05 の範囲にある様な調査がある。そこでは最適割当てからは非常に多くの利益があるであろう。

標本の大きさを決定する公式は 5.13 のより一般的な公式から導く事が出来る。 V をして全母集団に対する割合 p の推定に於ての欲する分散としよう。割当ての二つの主要型に対する公式は次の如くである。

比 例 :
$$n_0 = \frac{\sum W_h p_h q_h}{V} ; n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (5.56)$$

仮想最適 :
$$n_0 = \frac{(\sum W_h \sqrt{p_h q_h})^2}{V} ; n = \frac{n_0}{1 + \frac{n_0}{N} \sum W_h p_h q_h} \quad (5.57)$$

此処に n_0 は f, p, c を無視したときの第一次近似で n は f, p, c を考慮に入れて補正した値である。これ等の公式を展開する際に因子 $N_h / (N_h - 1)$ は 1 としてとられた。

この節でのすべての結果は割合の推定に適用される。もしも

百分率の項を考える方がよい場合はもしも p_h, q_h, V 等を百分率として表わすと同じ公式が適用される。class c に属する母集団の総数即ち NP の推定に際してはすべての分散を N^2 倍すればよい。

5.15 層の作り方 The Construction of strata : -

この表題は多くの問題を生ずる。層を作るための最良の特性は何か？、どの位の層であらねばならぬか？、層間の境界は如何にして決定さるべきか？、一般的な事項である答がそれに対して与えられたとしても多くの研究がこれ等の問題に行われないうちに残っている。

単一項目又は変量に対してもしもそれを得る事が可能であれば層化に対する最上の特性は変量そのものの度数分布である事は明かな様である。次に大切なものは例えば以前の調査でのその変量の値の如きその変量と高度の相関をもつ他の変量の度数分布である。層の数に関する限りは、その答は一見して愉快なものである。これは単純無作為抽出に対する層化無作為抽出の相対的正確度についての一般的な結果から導かれる。

5.57 では、もしも f, p, c が無視出来ると

$$V_{opt} \leq V_{prop} \leq V_{ran}$$

である事を示した。この結果は殆んどすべての場合にたとえ f, p, c が無視出来なくても妥当である事が暗示された。既に存在している層が新しい二つの層に分けられると仮定しよう。

この層に対して効果は単純無作為抽出を層化無作為抽出におきかえる事である。そしてそれは推定平均値に対して小さい分散を与え従つて全母集団の平均値に対して小さい分散を与える結果となる。かくて古い層の区分によつて新しい層を作る過程は推定量に於ける一連の減少分散をもたらす結果となる。

併し層を増すことはある不利益を伴う。割当てが比例的でなければ表に於て表われる荷重因子 W_h の数は推定さるべき層内分散が増すと割当て及び $\bar{y}_{st}$ の標準誤差を見出す両方に対して増加す

る。推定標準誤差に対して利用出来る自由度の数は標準誤差に対しての公式が層当り一つの unit で以て全く用いられなくなるまで減少する。もしも層の数に於ける増大が考慮に入れられるとすれば、我々は正確度の確実な利益がこれ等の複雑性に対して我々を償うに充分である事をたしかめたい。

各層が新しい層を作るために半分に分けられる過程を考えよう。層の数は次々に 2, 4, 8, 16 等となる。もしも我々が層を作るために y_i そのものの度数分布を用い得れば又 y_i の度数が連続的であれば過程に於ける各段階は推定量の分散に於て膨大な減少をもたらす。

層の数が大きくなる時各倍になったものは以前の値のほぼ近似的に $1/4$ の分散に減少する。これを示すために区分前に a 及び b の間の y_i のすべての値から一つの代表的な層がなりたつてしていると仮定しよう。もしも更に多くの層があれば、 a 及び b 間の y_i の分布は近似的に矩形となる。この矩形分布の分散は理論により、 $(b-a)^2/12$ である事が分っている。我々が層を半分にすると、その各々が $(b-a)/2$ の範囲で分散が $(b-a)^2/48$ なる二つの近似的に矩形の分布を生ずる。

比例割当てで f, p, c_i を無視して

$$V(\bar{y}_{st}) = \frac{\sum W_h S_h^2}{n}$$

従つて $V(\bar{y}_{st})$ へのこの層の寄与は区分する事により

$$\frac{W_h (b-a)^2}{12n}$$

から

$$\frac{W_h (b-a)^2}{2 \cdot 48n} + \frac{W_h (b-a)^2}{2 \cdot 48n} = \frac{W_h (b-a)^2}{48n}$$

に減少する。

例として $V(\bar{y}_{st})$ の値が度数函数 $f(y_i) = e^{-y_i} (y_i \geq 0)$ なる変量に対して 2, 4, 8 及び 16 の層に対して計算されたとする。

る。与えられた区分に於てすべての層はこれは最大正確度に対しての最良区分ではないけれども同じ大きさのものであつた。比例割当て及び最適割当ての両方（一定標本の大きさ）を含んでいた $V(\bar{y}_{st})$ の値及び上記の値との比が表 5.9 に掲げてある。

5.9 表 e^{-y_i} から作られた層に対する $nV(\bar{y}_{st})$ の値

層の 数	最 適		比	
	nV	上記の値に対する比	nV	上記の値に対する比
1	1		1	
2	0.27722	0.277	0.40747	0.407
4	0.06690	0.241	0.11822	0.290
8	0.01618	0.242	0.03077	0.260
16	0.00397	0.245	0.00778	0.253

比例割当てに対しては分散はまづ層の数を倍にすると $1/4$ に減する。何故ならば無限に広がる層は他のものよりは実質的には大きな分散をもつからである。併し比は両方の型に対して共にすみやかに $1/4$ に安定している。

実際には層化は y_i の度数分布に基いたものではない。 y_i と相関がある変量 z_i の度数分布に基いていると仮定しよう。 $y_i = z_i + e_i$ と仮定しよう。此処に変量 z_i 及び e_i は独立に分布している。 z_i の度数分布をきる事によつて作られた層は e_i の分散には影響しないであろう。多くの層がある時、半分に層を区分する事は

$$\frac{W_h (b-a)^2}{12n} + \frac{W_h \sigma_{e_h}^2}{n}$$

から

$$\frac{W_h (b-a)^2}{48n} + \frac{W_h \sigma_{e_h}^2}{n}$$

へのこの層の寄与を導く。

($k-2$)が σ^2_{k-2} の項が優位を示すほど小さくなるとすぐに層の数を更に増すと分散ではほんの少しの減少しか示さない。

要約すると遅かれ早かれ實際上層を区分しても正確度は最早増大しない。もしも σ^2_{k-2} が σ^2_{k-1} と適当に相関がある場合だけ少しの層で目的が達せられるであろう。

もしも層が度数分布をきる事によつて作られたとすると“区分の最良の点は何か?”という問題が残る。

規則は比例及び最適割当ての下の区分に対して *Dalenius* (1950) 及び *Gurney* (1951) により進展された。一つの結果を引用しよう。もしも区分がそれに基づいている変量 z_{ki} が調査で測定さるべき変量 y_{ki} に関して直線であれば、層 k 及び層 $(k+1)$ 間の区分点 z'_k は次の方程式を満足しなければならぬ。

$$2z'_k = \bar{z}_k + \bar{z}_{k+1}$$

此処に $\bar{z}_k$ 及び $\bar{z}_{k+1}$ は二つの層に於ける z_{ki} の平均値である。これは比例割当てに対する最も適した規則である。前以て層の数 k は選ばれていると仮定されるから $z'_1, z'_2, \dots, z'_{k-1}$ なる区分点は試算で見出される。

一つの変量に対して有効である層化は他のものに対してそうであるとは限らない。広範囲の項目にわたる調査では層を作るためのある妥協的な批判が見出されねばならない。経済的な項目に対する層化の基礎は *Stephan* (1941) 及び *Hagood & Bernert* (1945) により論せられ、農場の項目に対しては *King & McCarty* (1941) により論せられた。望まれるものはすべての主項目と調査に於て高度の相関をもつある変量である。

§5.16 層化に対する基礎としての接近

Proximity as a basis for stratification:

地域的な調査に於ては、隣りの単位は隔った単位よりは似通っている事が多い。お互に近くに住んでいる農民は同じ土壌型、同じ種類の農耕をやり、市場に同じ様な道路をもつ傾向がある。町の同じ場所に住んでいる人は同じ経済的水準をもち、彼等の生活

態度及び趣味も通常同じである様な傾向がある。これ等の点は類似性が強いといふ意味ではなくてたいそれが存在するというだけである。

この現象は完全な地理的な地方からなりたつ層を作る事によつて利用出来る。多くの代表的な農場の経済的項目に対してアメリカ合衆国に於ける州内の地理的層化の有効性についての資料が、5.10表に示してある。資料は *Jessen* (1942) 及び *Jessen & Houseman* (1944) により提供されたものである。

四つの大きさの層が示されている。即ち *township* (町)、*County* (郡)、“*type of farming*” *area* (農業の型の地域)、*state* (州)である。層の相対的大きさについてのある考え方が *Iowa* 州では約 1600 の *township*、100 の *County* 及び 5 の *area* があるという事実から得られる。

5.10表 地理的層化の色々な種類の相対的正確度 (%)

州	項目の数				
		<i>Township</i>	<i>County</i>	<i>Type of farming area</i>	<i>State</i>
<i>Iowa</i> 1938	18	115	100	16	91
<i>Iowa</i> 1939	19	121	100	97	91
<i>Florida</i> 1942					
<i>Citrus fruit area</i>	14	144	100		
<i>Truck farming area</i>	15	111	100		
<i>California</i> , 1942	17	113	100	97	

示された資料は表に示された項目の数について平均された平均値の推定率の平均相対的正確度である。*County* が各場合標準にとられている。

地理的層化の代表的なものと思われる様に層の数を増す事からの正確度の増大は大きいというよりはむしろ適當である。これは

上に示した類似性が弱いという事を示す。

§ 5.17 層化による利益の標本からの推定

Estimation from a sample of the gain due to stratification

層化無作為標本がとられた時には将来の調査の指導の手引として単純無作為抽出に対する相対的正確度に於ける利益を評価する事は興味ある事である。

標本から利用出来る資料は、 $N_h, n_h, \bar{y}_h$ 及び s_h^2 である。§ 5.4 から層化標本からの荷重平均の推定分散は

$$v(\bar{y}_{st}) = \sum \frac{W_h s_h^2}{n_h} - \sum \frac{W_h s_h^2}{N}$$

此処に $W_h = N_h/N$

問題は、この分散を単純無作為標本から得られる平均値の分散の推定量と比較する事である。しばしば用いられる方法は、標本平均からのよく知られている平均平方偏差

$$s^2 = \frac{\sum (y_{hi} - \bar{y})^2}{n-1}$$

を計算する事である。此処に層は無視されている。これは単純無作為標本に対して単位当りの分散の推定量としてとられる。この方法はもしも割当てが比例的であればうまくゆく、何故ならば、単純無作為標本が層内でそのものが近似的に比例的に分布しているからである。併し、もしも比例からかけはなれた割当てが採用されると、実際にとられた標本は単純無作為標本と似ても似つかぬものになり、この s^2 はつまらぬ推定量となる。一般的な方法を次にやってみよう。

単純無作為標本の真の分散は S^2 の代数的等号により

$$V_{run} = \frac{(N-n)}{nN} S^2 = \frac{(N-n)}{nN} \left[\frac{\sum (N_h-1) s_h^2 + \sum N_h (\bar{y}_h - \bar{y})^2}{(N-1)} \right]$$

(5.58)

括弧内の第一項の推定量を得る事はさして困難ではない。第二項は詳細な研究を要する。 $\sum N_h (\bar{y}_h - \bar{y})^2$ を推定するため $\sum N_h$

$(\bar{y}_h - \bar{y}_{st})^2$ を用いるのはごく自然である。この量は補正を要する過大推定量をもたらす。

我々は次の様にかく

$$\sum N_h (\bar{y}_h - \bar{y}_{st})^2 = \sum N_h \{ (\bar{y}_h - \bar{y}) + (\bar{y}_h - \bar{y}_{st}) - (\bar{y}_{st} - \bar{y}) \}^2$$

これを展開してすべての可能な標本上で平均する。 $(\bar{y}_h - \bar{y})$ を含む二つの積の各々の項は消去される事は明らかである。従つて

$$E \sum N_h (\bar{y}_h - \bar{y}_{st})^2 = \sum N_h (\bar{y}_h - \bar{y})^2 + E \sum N_h (\bar{y}_h - \bar{y}_{st})^2 + E \sum N_h (\bar{y}_{st} - \bar{y})^2 - 2 E \sum N_h (\bar{y}_h - \bar{y}_{st})(\bar{y}_{st} - \bar{y}) \quad (5.59)$$

併し、 $\bar{y}_{st}$ 及び $\bar{y}$ の定義により

$$\sum N_h (\bar{y}_h - \bar{y}_{st})(\bar{y}_{st} - \bar{y}) = N (\bar{y}_{st} - \bar{y})^2$$

従つて (5.59) の最後の二つの項を一緒にして、この表示は $\bar{y}_{st}$ の分散の N 倍であるから

$$- E N (\bar{y}_{st} - \bar{y})^2 = - \sum \frac{N_h (N_h - n_h)}{N} \frac{s_h^2}{n_h}$$

(5.59) の右辺の二番目の項に対しては

$$E \sum N_h (\bar{y}_h - \bar{y}_{st})^2 = \sum \frac{N_h (N_h - n_h)}{N_h} \frac{s_h^2}{n_h} = \sum (N_h - n_h) \frac{s_h^2}{n_h}$$

何故なら各層内での $\bar{y}_h$ は単純無作為標本の平均値であるからである。従つて

$$E \sum N_h (\bar{y}_h - \bar{y}_{st})^2 = \sum N_h (\bar{y}_h - \bar{y})^2 + \sum (N_h - n_h) \frac{s_h^2}{n_h} - \frac{\sum N_h (N_h - n_h)}{N} \frac{s_h^2}{n_h}$$

右辺の最後の二つの項の和は正であるから $\sum N_h (\bar{y}_h - \bar{y}_{st})^2$ は過大推定量を与える。

$\sum N_h (\bar{y}_h - \bar{y})^2$ の不偏推定量は次の如くである。

$$Q = \sum N_h (\bar{y}_h - \bar{y}_{st})^2 - \sum (N_h - n_h) \frac{s_h^2}{n_h} + \frac{\sum N_h (N_h - n_h)}{N} \frac{s_h^2}{n_h}$$

最後に方程式 (5.58) にこの推定量を代入して V_{run} の不偏推定量として次のものを得る

$$V_{run} = \frac{(N-n)}{nN} \left[\frac{\sum (N_h - 1) s_h^2 + Q}{N-1} \right] \quad (5.60)$$

計算の目的に対してはこれを相対的層の大きさ $W_h = N_h/N$ の項で表わすと便利である。(5.60)で Q の値が分るとすこし消去して次のものを得る。

$$U_{ran} = \frac{(N-n)}{nN} \left[W_h s_h^2 - \sum \frac{W_h s_h^2}{n_h} + \sum \frac{W_h s_h^2}{n_h} - \frac{\sum W_h s_h^2}{N} + \sum W_h \bar{y}_h^2 - (\sum W_h \bar{y}_h)^2 \right]$$

この表示は計算にはよくない面である。殆んど大抵の応用に於てある簡易化されたものが利用出来る。二つのものを示してみよう $N > 50$ これは殆んどすべての母集団に適用される。括弧内の四番目の項がなくなりこれは第一項を N で割ったものに等しい。従つて

$$U_{ran} = \frac{(N-n)}{nN} \left[\sum W_h s_h^2 - \sum \frac{W_h s_h^2}{n_h} + \sum \frac{W_h s_h^2}{n_h} + \sum W_h \bar{y}_h^2 - (\sum W_h \bar{y}_h)^2 \right] \quad (5.61)$$

すべての $n_h > 50$ 、括弧内の二番目と三番目の項がなくなり

$$U_{ran} = \frac{(N-n)}{nN} \left[\sum W_h s_h^2 + \sum W_h \bar{y}_h^2 - (\sum W_h \bar{y}_h)^2 \right] \quad (5.62)$$

(例)

計算は師範学校の標本(5.13)に於ける最初の三つの層からとつたものを例としよう。1946年の標本からの資料が5.11表に示してある。平均値は学校当り在籍者を十単位で表わしたものである。

5.11表 層化標本からの基礎資料

層	N_h	n_h	$\bar{y}_h$	s_h^2
1	13	9	2.200	1.615
2	18	7	1.638	0.063
3	26	10	0.992	0.077
計	57	26		

標本は非常に小さいから(5.61)なる方程式を用いる。

補助的な計算は5.12表に示してある。

5.12表 計算の配列

層	W_h	$W_h s_h^2$	$W_h s_h^2 / n_h$	$W_h s_h^2 / n_h$	$W_h \bar{y}_h^2$
1	0.228	0.36822	0.04091	0.00933	0.50160
2	0.316	0.01991	0.00284	0.00090	0.51761
3	0.456	0.03511	0.00351	0.00160	0.45235
計	1.000	0.42324	0.04726	0.01183	1.47156

公式は次の様にして計算される。

$$U_{st} = \sum \frac{W_h s_h^2}{n_h} - \sum \frac{W_h s_h^2}{N} = 0.01183 - 0.00743 = 0.0044$$

$$U_{ran} = \frac{31}{(57)(26)} [0.4232 - 0.0473 + 0.0118 + 2.4000 - 2.1653] = 0.0130$$

層化する事によつて単純無作為標本に対する値の約三分の一に分散が減っている。

○ 比例割当て

通常適当である推定量 U_{ran} はその平均値からの標本平均の偏差の平方和から得られる。即ち分散分析に於ける普通の等号により

$$s^2 = \frac{\sum (y_{hi} - \bar{y})^2}{n-1} = \frac{1}{(n-1)} \left[\sum (n_h - 1) s_h^2 + \sum n_h \bar{y}_h^2 - \frac{(\sum n_h \bar{y}_h)^2}{n} \right]$$

もしも $1/n_h$ が無視出来ればこれは次のものに等しい。

$$\sum W_h s_h^2 + \sum W_h \bar{y}_h^2 - (\sum W_h \bar{y}_h)^2$$

此処に $W_h = n_h/n$ これは方程式(5.62)の括弧内の量に等しい。従つて表示は

$$U_{ran} = \frac{(N-n)}{N} \frac{s^2}{n}$$

はもしも割当てが比例的であり $1/n_h$ なる項が無視出来れば満足である。

○ s_h^2 が一定なる場合の比例割当て

この場合は多くの応用に適用されると思われる。例えば農場

での標本抽出では層は高の区分である。 S_e^2 の一定の値即ち S_w^2 が層内のよせあつめた平均平方 S_w^2 により推定される。 $v(\bar{y}_{st}) = v_{st}$ に対して (5.6)なる方程式から

$$v_{st} = \left(\frac{N-n}{N} \right) \frac{S_w^2}{n}$$

v_{ran} に対して我々は $1/N$ なる項は無視出来ること仮定されるが $1/n_h$ は無視出来ない。何故ならば時には標本に於ては層あたり二、三の単位しかない場合があるからである。従つて我々は個々の S_e^2 の代りによせあつめた推定量 S_w^2 を求めるのに方程式 (5.61) から出発する。これは次の如くなる。

$$v_{ran} = \frac{(N-n)}{Nn} \left[\sum (W_h - \frac{1}{n} + \frac{W_h^2}{n}) S_w^2 + \sum W_h \bar{y}_h^2 - (\sum W_h \bar{y}_h)^2 \right]$$

$n_h = n W_h$ であるから S_w^2 の変動係数は次の如くかける。

$$\sum (W_h - \frac{1}{n} + \frac{W_h^2}{n}) = \frac{n-L+1}{n}$$

此処に L は層の数である。従つて

$$v_{ran} = \frac{(N-n)}{N} \frac{1}{n^2} \left\{ (n-L+1) S_w^2 + \sum n_h (\bar{y}_h - \bar{y}_{st})^2 \right\}$$

これ等の量は5.13表に示される様に二つの成分えの標本資料の分散分析から計算出来る。

5.13表 標本資料の分散分析

成分	d.f	m. s.
層間	$(L-1)$	$B = \sum n_h (\bar{y}_h - \bar{y}_{st})^2 / (L-1)$
層内	$(n-L)$	$W = S_w^2$

従つて次の如くなる

$$v_{ran} = \left(\frac{N-n}{N} \right) \frac{1}{n^2} \left[(L-1) B + (n-L+1) W \right]$$

(例)

プロット当りのハリガネ虫の数の推定に対する農事試験での標

本抽出に於て、25プロットが半分に分けられた。各半分から土壌の3つの無作為標本が小さい。ボーリング道具でとられた標本は深さ5インチで9平方インチの土壌の体積であつた。25プロットに対するハリガネ虫の数の組合せ分散分析は次の如くである。

	d.f	m. s.
層間 (半分のプロット)	25	$B = 90.76$
層内	100	$W = 38.44$

条件は上に提示した理論と少し違つてゐる。各プロットは二つの層に分けられた分離された母集団を表わす。従つて各プロットに対して $n=6$, $n_h=3$, $L=2$. 分散分析はこの型の25層化標本に対して組合せられた結果を与える。f, p, cは含まれる必要はない。

$$v_{st} = \frac{38.44}{6} = 6.41$$

$$v_{ran} = \frac{1}{36} [B + 5W] = \frac{1}{36} [90.76 + 5(38.44)] = 2.86$$

$$\text{相対的正確度} = \frac{2.86}{6.41} = 0.23$$

半分に層化する事によつてほぼ $1/4$ ほど正確度がました。

§ 5.18 層の大きさに於ける誤差の影響

Effects of errors in the stratum sizes: -

層化の欲する型に対して、層総計 N_h は日が経つてからのセンサスからか又は大標本から導かれるので正確には分らない。誤りのある荷重を層化の基礎におくことについての一定の論述はある特別な場合以外に行う事が出来ない。一般的な性質の二、三の結論が導かれる。f, p, cは無視する。

真の層の割合 W_h の代りに推定量 w_h を用いる。標本推定平均は $\sum w_h \bar{y}_h$ である。この推定量は偏倚がある。くりかえしの抽出でのその平均値は $\sum W_h \bar{y}_h$ である。偏倚は $\sum (w_h - W_h) \bar{y}_h$ である。その結果この推定量の誤差分散は次の二つの成分を含む

事になる。即ち推定量の平均値まわりの分散及び偏倚の平方である。もしも“最適”割当が（この推定量によつておきかえられた N_h で以て）用いられるとすれば第一の成分は $(\sum W_h S_h)^2/n$ である。全分散は

$$V(\bar{y}_{st}) = \frac{(\sum W_h S_h)^2}{n} + \left\{ \sum (W_h - \bar{W}_h) \bar{Y}_h \right\}^2 \quad (5.63)$$

この表示のより一般的な表示は Stephan (1941) により与えられた。彼が指摘している様に (5.63) の第一項は通常正しい荷重をもつ分散とは大して違わないであろう。——もしも分散がすべての層で同じであれば二つは正確に一致する。従つて正しくない荷重からの正確度の損失は個々の場合に小さいか又は大きい様な偏倚の大きさによる。誤りのある荷重の与えられた組に対して損失は標本の大きさと共に変化する。何故ならば、全分散の“偏倚”成分は標本の大きさと独立であるからである。標本の大きさを増すと“偏倚”の項が優勢になり、層化が無作為抽出よりも正確さが少くなる様な段階に到達する。

この論議は荷重に誤差がある事が分つている様な調査に於て層化するかどうかを考えるのには役に立たない。何故ならば偏倚の大きさは予測出来ないからである。場合によつては標準誤差はそれ等が推定された過程の知識が N_h の推定量に到達し得る。

もしも N_h の推定量が独立であり $\bar{y}_h$ について独立であれば、 $\bar{y}_{st}$ の全分散の偏倚成分の平均値はラフに (Cochran, 1939)

$$\frac{\sum (\bar{Y}_h - \bar{Y})^2 V(N_h)}{N^2}$$

此処に $V(N_h)$ は N_h の推定量の分散である。この量は N_h に於ける誤差による $\bar{y}_{st}$ の分散に於ける期待増大を測る尺度となる。

King, McCarty 及び McPeck (1942) はアメリカ合衆国の小麦地帯に於けるエーカー当り収獲量、蛋白質及び試験重量の推定に対する調査にこの公式を用いた。彼等は各州内の地区による層化の有用性を論じた。各地区に対する全面積 N_h はそれぞ

のものが標本調査で推定されたから $V(N_h)$ の知識が利用出来た。

この問題は二重抽出 (5.12.2) の場合にも起つてくる。もしも W_h が大きな無作為標本の大きさ n' ($n' \geq n$) で推定されどの位各層に属しているかが分ると $V(\bar{y}_{st})$ に於ける増大は近似的に

$$\frac{\sum W_h (\bar{Y}_h - \bar{Y})^2}{n'}$$

5.17 層が前以て同一であると認め難い場合

Case where the strata cannot be identified in advance

時には単位が属する層が単位から資料が得られる迄不明の場合がある。政治的な態度に関する調査に於ては個人に会つて分らない最終選択の個人の投票行動によつて層化が有用である。同様に態度は層化が収入、職業及び宗教関係の様な因子に基いている時にも起る。勿論この様な場合は、層の大きさ N_h は正確には分らない。併し我々はそれが分つていないと仮定する。一つの方法が単純無作為標本の大きさ n なるものをする事によつて行われる。標本資料があつめられたとき単位はそれ等について得られた情報によつて層に割当てられる。もしも $\bar{y}_h$ が n 層に属する単位の平均値とすれば、推定量として

$$\bar{y}_{ur} = \frac{\sum N_h \bar{y}_h}{N}$$

換言すれば我々は標本の重みづけられない平均値をとる代りに荷重平均を得るための荷重として真の層の大きさを用いている。

もしも標本が合理的な大きさであれば、この方法は殆んど比例層化抽出と正確度は同じである。 m_h をして n 層に属する標本の数としよう。此処に m_h は標本をとる毎に変わる。

m_h が一定である標本に対しては

$$V(\bar{y}_{ur}) = \frac{1}{N^2} \sum N_h^2 \frac{S_h^2}{m_h}$$

此処に S_h^2 は無視されている。

くりかえしの標本抽出に於てはこの量の平均値が今計算されなければならぬ。この方法はいささか注意を要する。何故ならば m_h の一つ又はそれ以上のものが 0 であり得るからである。もしもこの様な事が起れば推定量を作る前に二つ又はそれ以上の層を合併しなければならない。その結果は正確度の劣った層化及び $\bar{y}_w$ の大なる分散を得る事になる。これを増すと m_h が 0 である確率は小さいからこの原因からの分散への影響は無視出来る(この事は証明しなくても分る) m_h が 0 である場合を除けば *Stephan* (1945) は n^{-2} のオーダーの項に対して

$$E\left(\frac{1}{m_h}\right) = \frac{1}{nW_h} - \frac{1}{nW_h} + \frac{1}{n^2W_h^2}$$

である事を示した。此処に $W_h = N_h/N$ 従つて

$$E\{V(\bar{y}_w)\} = \frac{1}{n} \sum W_h S_h^2 + (n^{-2}) \text{ の項}$$

初項は比例層化抽出で得られる分散である。

§5.20 割堪抽出法 *Quota sampling* :-

この立場で用いられる他の方法は各層から望まれる n_h を前以てきめて必要な「割堪」が各層で得られる迄抽出を続けて測定する事である。もしも測定が最初に必要でないものを除いて無作為に単位を選ばば、この方法は層化無作為抽出に等しい。抽出の最終段階ではこれは調査する人にとってかなりの労力となる。何故なら隣り合つた単位の殆んどが割堪が既に終つた層に入るであろうからである。

この方法は実際に多くの機関によつて用いられるから測定者は単位を無作為に選ばない。その代り彼は割堪を出来るだけ早く満し得る様な情報を利用する(例えば裕福な人は決して貧民窟には住まない)。対象は単位を無作為に選ぶ事に於て生ずる高い費用をかけないで層化の利益を得る事である。範囲の量が変わる事は測定者に許される。抽出理論は確率抽出の要素を含まない。割堪

抽出法には通用出来ない。この様な方法の正確度についての情報はセンサスとの比較が可能になつた時又は信頼限界が計算出来る標本との比較が可能になつたときだけ得られる。

§5.21 層当り一つの単位での層化

Stratification with one unit per stratum :-

もしも母集団が非常に変動が多くて層化に対して多くの有効な批判が分つておれば、層化は標本が各層に一つの単位を含むところまで持つてゆかれる。この事象に於ては $V(\bar{y}_{st})$ の推定に際して前に用いた公式は用いる事は出来ない。推定量は層を対にまとめる事によつて得られる。対を形作る二つの層に於て我々は大きさ N_h は等しいと仮定する。等しい事からすこしはかけはなれてもこの方法は損われないであろう。対の二つの数に対する母平均 $\bar{y}_h$ はかなり違ふであろうが対への割当ては標本結果をみる前に行われねばならぬであろう。その理由は次に明らかになるであろう。層の数は少くとも 20 で推定分散に対して最小 10 以上でなければならない。

y_{j1} , y_{j2} を代表的な対に於ける観測値としよう。此処に j は 1 から $L/2$ まで変化する。そのとき、この対からのすべての標本を平均して

$$E(y_{j1} - y_{j2})^2 = (\bar{y}_{j1} - \bar{y}_{j2})^2 + \frac{(N_j - 1)}{N_j} (S_{j1}^2 + S_{j2}^2) \quad (5.64)$$

此処に $N_j = N_h$ は対に於ける各層の大きさである。次の推定量を考えよう。即ち

$$v(\bar{y}_{st}) = \frac{1}{N^2} \sum_{j=1}^{L/2} N_j^2 (y_{j1} - y_{j2})^2 \quad (5.65)$$

(5.64) によりこの量の期待値は

$$E v(\bar{y}_{st}) = \frac{1}{N^2} \left[\sum_{h=1}^L N_h(N_h - 1) S_h^2 + \sum_{j=1}^{L/2} N_j^2 (\bar{y}_{j1} - \bar{y}_{j2})^2 \right] \quad (5.66)$$

右辺の第一項は補正分散である ($n_h=1$ なる場合の定理 5.3 により)。第二の項は正の偏倚を表わす。偏倚の大きさは真の平均値から少ししか違わない層の対の情報に於て得られた成功如何による。推定 (5.65) の型は我々は標本の値を出来るだけ少ししか違わせない様にする事によつて対を作つてはならない事を警告している。何故ならこれはひどい過小評価を与えるからである。技術的には時には俗語で“くづれた層 (Collapsed strata)” と呼ばれているものである。

抽出の他に与るべき方策は層当り無作為に選ばれた $1/2$ 層をもち、二つの単位をもつ単一層として各層を用いる事である。この種の抽出に於ける $V(\bar{y}_{st})$ の不偏推定量は普通の公式から導かれる。諸君は次の事を証明出来るだろう。

$$V(\bar{y}_{st}) = \frac{1}{N^2} \left[\sum_{h=1}^L N_h(N_h-1) \frac{(2N_h-2)}{2N_h-1} S_h^2 + \sum_{j=1}^{L/2} N_j^2 \frac{(N_j-1)}{2N_j-1} (\bar{y}_{j1} - \bar{y}_{j2})^2 \right] \quad (5.67)$$

(5.66) と比較して推定 (5.65) は層当り 1 つの単位をもつ真の分散であるばかりでなく、もしも大きさ 2 倍の層が用いられたら適用される分散の過大評価である事が分る。この結果に照し合せてより小さい層がよりよいかどうかは問題となる。不幸にも、もしも対あたり 2 つの単位と比較したとき層当り 1 つの単位からの正確度に大いに利益があれば分散の大なる過大評価が又これにもある。

§ 5.22 分析研究に於ける層化

Stratification in analytical studies

抽出を含む多くの研究に於て、我々は例えば都市と田舎に差があるか、授業料を払う学校と払わぬ学校との間に差があるか、違った教育水準の人間に差があるかどうかを見出すために、母集団の一つの部分と他の部分とのある変量の平均値を比較したい場合がある。この様な研究は分析研究と呼ばれる。何故ならば母集団の特性についての単なる記述よりもむしろ母集団内部の関係の研究を含むからである。分析研究に於て我々は通常観測された差を

導いた下によてたわる原因の要因について更に、こゝで推測しようとする。コントロールされた実験でないに、この様な推測は面倒になる傾向があるが、研究の多くの分野では調査者は、彼の知識の多くに対して観測と分析の組合わせ以外にたよる手はない。

研究の領域なる言葉は我々が比較したい母集団のこれ等の部分に対する標本抽出に対して U.S. Subcommittee によつて与えられた。

分析調査に用いる計画及び分析の技術はこの本で今迄考えてきた詳細な調査に用いたものと異つてゐる。研究の各領域は別個な層を仮定する。これは分析調査を計画する最も簡単な方法の一つである。もしもこの様な二つの層が比較されると、 $\bar{y}_1 = \bar{y}_2$ であるかどうかを調べる科学的な興味は通常ない。何故ならばこれ等の平均値はごく稀な場合以外、同じ無限母集団から無作為に二つの層が抽出されたものであつても、稀にしか等しくはないからである。その代りに、我々は二つの層は同じ無限母集団から抽出されたものであるかどうかを考える事が出来るか否かを問題にする。その結果分析的比較に於て推定層平均の分散を計算するとき P, C を無視する。

層に於ける標本の大きさを割当てゐる規則は亦違つてくる。もしも二つの層しかない場合は標本平均間の差 ($\bar{y}_1 - \bar{y}_2$) の分散は

$$V = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \quad (5.68)$$

$$C = a + C_1 n_1 + C_2 n_2 \quad (5.69)$$

なる費用函数に対して V は

$$n_1 = \frac{\frac{n S_1}{\sqrt{C_1}}}{\frac{S_1}{\sqrt{C_1}} + \frac{S_2}{\sqrt{C_2}}} : n_2 = \frac{\frac{n S_2}{\sqrt{C_2}}}{\frac{S_1}{\sqrt{C_1}} + \frac{S_2}{\sqrt{C_2}}} \quad (5.70)$$

のとき最小となる

L 個の層 ($L \geq 2$) では最適割当ては周囲の状況による、一つの方法は層のすべての対間の差の平均分散を最小にする事である。

(5.68) からこの平均分散は

$$\bar{V} = \frac{2}{L} \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + \dots + \frac{S_L^2}{n_L} \right)$$

(5.70) と同じ規則によつて一定 C に対して、 $\bar{V}$ は最小になる。

即ち

$$n_h \propto \frac{S_h}{\sqrt{C_h}}$$

それに代る他の方法は、もしも C_h が層毎に少ししか違わないが S_h にひどい変化があるときは各平均値が同じ分散をもつ様に割当てる事が良い場合がある。これに対して次の様に n_h を選び

$$\frac{S_1^2}{n_1} = \frac{S_2^2}{n_2} = \dots = \frac{S_L^2}{n_L}$$

問題は多くの派生的なものを含んでいる。分散分析及び重回帰の標本の利益はもしも各領域が特別な層をもち単純無作為標本であれば用いる事が出来る、各領域が作られた層の部分覆う様な時は複雑な問題がおこる。

第6章 比推定

Ratio Estimates

§ 6.1 推定の方法 *methods of estimation:-*

過去三十年間の理論統計学の発展の一つの姿は如何にして資料からよい推定量を得るかを示す膨大な理論の出現である。特に標本調査に対する理論の発展に於てはこの知識の母体は相対的にほんの少ししか用いられて来なかった。これには私は二つの主な理由があると思う。一つはオーソドックスな調査ではこれは多くの項目を含むものであるが簡単な附加的なものより以上のものを少ししか必要としない推定方法に非常に利益があるが、一方最尤法の如き、統計理論に於けるすぐれた推定方法は推定量を見出す事が出来る以前に一連の次々の近似を必要とする。才二番目研究の二つの線に於ける態度に相違があった。理論統計学に於ける多くの推定方法は標本に於ける資料による度数分布の函数型は分つておるし、推定方法は注意深くこの分布型に合う様にする事は当然の事である。標本調査理論でやる事は、この度数分布についてはたゞ制約された假定、例えば非常に歪んでいるとか、対称であるとかを假定して論議中からその特種な分布函数を除く様なものである。この態度は分布型が一つの項目毎に変わり、如何にして各推定量を得るかを決定する前にこれ等のすべての分布を少しとゞまつて検定したくない様な多くの項目をもつ調査を取扱うのに合理的なものである。

その結果、標本調査の仕事の推定の技術は現在むしろ視野に制限をうけてゐる。二つの技術が考えられるであろう。その一つはこの章で取扱う比推定、今一つはオセ算で取扱う直線回帰法である。より複雑なる法を用いる事が増大する可能性はある。何故ならば推定のすぐれた方法からの正確度の利益は最終計算のみが影響するからしばしばかなり低減に導かれるからである。

§ 6.2 比推定 *The ratio estimate* :-

比推定に於ては y_i に関連した任意変量 x_i が標本に於ける各単位に対して得られる。 x_i の母総計 X は既知でなければならぬ。実際には x_i はしばしば完全調査が行われたときのある以前に於ける y_i の値である。この方法の目的は y_i と x_i の相関の利益を用いて正確度を増す事である。今我々は単純無作為抽出を假定しよう。

y_i の母総計 Y の比推定は

$$\hat{Y}_R = \frac{y}{x} X = \frac{\bar{y}}{\bar{x}} X \quad (6.1)$$

此処に y, x は夫々 y_i 及び x_i の標本総計である。もしも x_i がある以前に於ける y_i の値であれば比推定は以前に起つた Y/X なる相対的変化を推定するために標本を用いる。推定された相対的変化 y/x に現在の母総計の推定量を关えるため前の場合に於ける既知の母集団総計 X を乗ずる。もしも比 y/x_i がすべての標本抽出単位で殆んど同じであれば、 y/x の値は標本が變つても殆んど変わらず比推定は正確度が高い事になる。他の応用に於ては、 x_i は農場の全面積であり、 y_i はある穀物がまかれた面積である。比推定は、もしもすべての農夫がこの穀物の全面積に同じ割合でまければ、この場合成功するであろう。

もしも推定さるべき量が Y であれば (y_i の母平均) 比推定は

$$\hat{Y}_R = \frac{y}{x} X$$

しばしば我々は総計又は平均よりは比を知りたい場合がある。例えば小麦面積に対する穀物面積の比とか、全支出に対する労働支出の比とか、全資産に対する流動資産の比とかの場合である。標本推定量は $\hat{r} = y/x$ である。この場合 X は知る必要がない。

(例)

6.1 表は前に § 2.8 で論じた 196 の大都市の母集団から抽出した 49 都市の単純無作為標本の各々の居住者の数 (1000 単位) を

示す。向類は 1930 年に於ける。196 都市の総居住者数を推定する事である。眞の 1920 年の総計 X は分つてゐると假定する。その値は 22,919 である。

例は比推定に対して適当なものである。標本に属する都市の大部分は 20% のオーダーで 1920 年から 1930 年にかけて大きさが増大してゐる。標本資料から次のものを得る。

$$y = \sum y_i = 6262 \quad ; \quad x = \sum x_i = 5054$$

従つて、すべての 196 都市に対する 1930 年の総計の比推定量は

$$\hat{Y}_R = \frac{y}{x} X = \frac{6262}{5054} (22,919) = 28,397$$

それに応ずる都市当り標本平均に基いた推定量は

$$\hat{y} = N \bar{y} = \frac{(196)(6262)}{49} = 25,048$$

1930 年の正しい総計は 29,351 である。

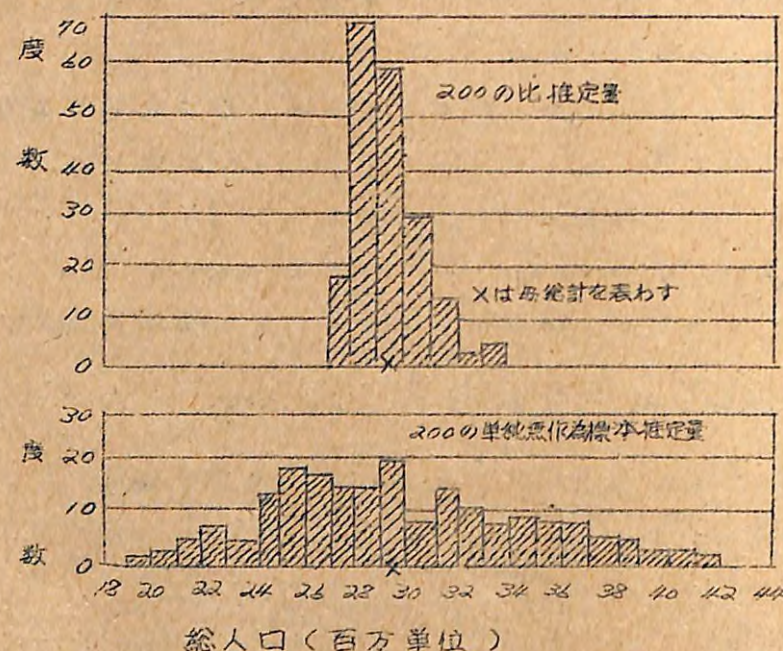
6.1 図は、この母集団から抽出された大きさ 49 の 200 個の単純無作為標本の各々に対する都市当り標本平均に基いた比推定量 $\hat{Y}_R$ と眞の比推定量 Y に対する比推定からの正確度に対する着しい改善は明らかである。

6.1 表 1920 年 (x_i) 及び 1930 年 (y_i) に於けるアメリカ

合衆国の 49 の大都市の大きさ (1000 単位)

x_i	y_i	x_i	y_i	x_i	y_i
76	80	2	50	243	291
138	143	507	634	87	105
67	67	128	260	30	111
29	50	121	113	71	79
381	464	50	80	256	288
23	48	44	58	43	61
37	63	77	89	35	57
120	115	64	63	94	85
61	69	64	77	13	50
387	459	58	142	298	317
23	104	40	60	36	46
112	183	40	64	161	232
18	106	38	52	74	93
66	86	136	137	45	53
60	57	116	130	36	54
46	65	46	53	50	58
				48	75

6.1 図 標本平均に基いた推定量と比推定量との実験的比較



8.6.3 比推定の近似分散と偏倚

Approximate variance and bias of the ratio estimate:-
 比推定量の分布は y 及び x は共に標本毎に変化するから面倒で処理しにくい事が証明された。今迄分った理論的結果は簡単にいえば我々が実際の適用に対して何か和らげたいかといふ事になる。主な結果を証明なしにまづ述べてみよう。

比推定量は一致推定量である。(これは明らかである。) 母集団のある特定の型以外では、たとえその偏倚が大標本で無視出来ようとも偏倚してゐる。 N が非常に大きくなると比推定の限界分布は、それから我々が抽出する母集団の型とすこし制限を設けると正規である。適当な大きさの標本に於ては分析は少くともその方法が非常にしばしば用いられる母集団の種類に対しては正の歪をもつ傾向がある。我々は偏倚及び推定量の抽出分散に対して正確な公式をもつてはいないが、大標本では妥当である一つの近似がある。

これ等の結果は次の如く標本が充分な大きさであれば困難ではない。

(i) 比が殆んど正規に分布してゐる、(ii) その分散に対する大標本の公式が妥当である。推論に於ける不備な点は (i) 次の問題に等すべき充分支持すべき規則がない事、即ちその問題は、何時標本の大きさが充分であるか? (iii) 小標本に対する信頼限界を推定するに役立つ方法、実行上の規則として大標本の結果はもしも標本の大きさが $\sqrt{N}$ の以上であれば用いられる。そして又 y 及び x の変動係数が共に 10% 以下である場合に用いられる。此の規則は本だむしる貧弱な証明ではある。

定理 6.1

大きさ N なる単純無作為標本では母集団総計 Y の比推定量 $\hat{Y}_R$ の分散は近似的に

$$V(\hat{Y}_R) = \frac{N(N-n)}{n(N-1)} \sum_{i=1}^n (y_i - R x_i)^2 \quad (6.2)$$

此処に $R = Y/X$ は母集団比である。近似は N が大であると仮定してゐる。

(証明の概略)

次の論議は分析に於て行われた近似が正当ではないといふ意味で厳密ではない。

推定母集団総計に於ける誤差は

$$\begin{aligned} \hat{Y}_R - Y &= \frac{\bar{y}}{\bar{x}} \bar{x} - Y \\ &= \bar{x} \left(\frac{\bar{y}}{\bar{x}} - \frac{Y}{X} \right) \\ &= \frac{N\bar{x}}{X} (\bar{y} - R\bar{x}) \end{aligned} \quad (6.3)$$

此処に $R = Y/X$

もしも標本が大であれば $\bar{y}$ は Y に近づき、近似は $\bar{x}/X$ を 1 とおく事にある。

$$\hat{Y}_R - Y \approx N(\bar{y} - R\bar{x}) \quad (6.4)$$

(124)

因子 N を除くと (6.4) に於ける推定量の近似誤差は変量 $u_i = y_i - Rx_i$ の標本値の平均である。我々は今単純無作為標本の平均値の分散に対する定理 2.2 を変量 u_i に適用する。これは次の如くなる。

$$V(\hat{y}_R) = N^2 V(\bar{u}) = \frac{N(N-n)}{n} \frac{\sum_{i=1}^N (u_i - \bar{u})^2}{(N-1)}$$

此処に $\bar{u}$ は変量 u_i の母平均である。併し R の定義により

$$\bar{u} = \bar{y} - R\bar{x} = 0$$

従つて

$$V(\hat{y}_R) = \frac{N(N-n)}{n(N-1)} \sum_{i=1}^N (y_i - Rx_i)^2$$

系 この結果には色々な変形がある。

$\bar{y} = R\bar{x}$ であるから次の様にかける。

$$\begin{aligned} V(\hat{y}_R) &= \frac{N(N-n)}{n(N-1)} \sum_{i=1}^N \{ (y_i - \bar{y}) - R(x_i - \bar{x}) \}^2 \\ &= \frac{N(N-n)}{n(N-1)} \left\{ \sum (y_i - \bar{y})^2 + R^2 \sum (x_i - \bar{x})^2 - 2R \sum (y_i - \bar{y})(x_i - \bar{x}) \right\} \end{aligned}$$

次の方程式により有限母集団に於ける y_i と x_i との相関係数 ρ を定義しよう

$$\rho = \frac{\sum_{i=1}^N (y_i - \bar{y})(x_i - \bar{x})}{(N-1) S_y S_x}$$

従つて次の如くなる

$$V(\hat{y}_R) = \frac{N(N-n)}{n} \{ S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x \} \quad (6.5)$$

これと同じ形で

$$V(\hat{y}_R) = \frac{(N-n)}{N} \frac{Y^2}{n} \left\{ \frac{S_y^2}{\bar{y}^2} + \frac{S_x^2}{\bar{x}^2} - \frac{2S_{yx}}{\bar{y}\bar{x}} \right\} \quad (6.6)$$

此処に $S_{yx} = \rho S_y S_x$ は y_i と x_i との共変量である。

この関係は又次の様にかける

$$V(\hat{y}_R) = \frac{(N-n)}{N} \frac{Y^2}{n} \{ C_{yy} + C_{xx} - 2C_{yx} \} \quad (6.7)$$

(125)

此処に C_{yy} , C_{xx} は y_i と x_i の変動係数 (C_v) の平方であり、 C_{yx} は同じものの相対的共分散である。

注意 1

母集団比 R の推定量としての比推定量は $\bar{y}$ と $\bar{x}$ との比を用いる。ある読者は個々の単位での比 $y_i = y_i/x_i$ の平均値 $\bar{y}$ が R のよりよい推定量ではないかといふ様な疑問を抱くであろう。詳細には述べるないが、これは母集団のある特殊な型以外では単純無作為抽出の場合とは思えない (§ 6.7 参照)。事実有限母集団では $\bar{y}$ はすべての抽出単位上でとられた $\bar{y}$ が R に等しくないから推定量 $\bar{y}$ は一致推定量ではない。更に $\bar{y}$ は y/x よりも計算がより面倒くさい。

注意 2

比推定の分散に対する近似公式は単位毎に変わる y_i の変動の量で表わす事が出来る。

方程式 (6.2) から

$$V(\hat{y}_R) = \frac{N(N-n)}{n(N-1)} \sum_{i=1}^N x_i^2 \left(\frac{y_i}{x_i} - R \right)^2 = \frac{N(N-n)}{n(N-1)} \sum x_i^2 (y_i - R)^2$$

和は母集団比 R からの y_i の偏差の平方の重みづけられたものである。もしもすべての y_i が等しければ、比推定は誤差なしであるから近似分散はなくなる。

注意 3

比 $\hat{R} = \hat{y}/\hat{x}$ の近似分散に対しては前の公式を X^2 で割る。その結果の二つの型は次の通りである。

$$\begin{aligned} V(\hat{R}) &= \frac{N(N-n)}{n(N-1)X^2} \sum_{i=1}^N (y_i - Rx_i)^2 \\ &= \frac{(N-n)}{nN} R^2 \{ C_{yy} + C_{xx} - 2C_{yx} \} \quad (6.8) \end{aligned}$$

注意 4. 備倚

定理 6.1 に於ける Y の比推定量の分散を見出す際、本質的な段階は次の近似を導く事であつた。

$$\hat{Y}_R - Y = \frac{N\bar{x}}{\bar{x}} (\bar{y} - R\bar{x}) \approx N(\bar{y} - R\bar{x}) \quad (6.9)$$

$E(\bar{y} - R\bar{x}) = 0$ であるから、分数公式に用いられる近似のオーダーに対しては比推定量は不偏である $\hat{Y}_R$ の偏倚に於ける初項を見出すため、我々は更に一つの段階の近似をとらねばならない。これは次の称にかく事によって得られる。

$$\frac{\bar{x}}{\bar{x}} = \frac{\bar{x}}{\bar{x}(1 + \frac{\bar{x} - \bar{x}}{\bar{x}})} = 1 - \frac{\bar{x} - \bar{x}}{\bar{x}}$$

これは Taylor 系列の第一項がとられてゐる。(6.9) に代入すると

$$\hat{Y}_R - Y = N(\bar{y} - R\bar{x}) \left(1 - \frac{\bar{x} - \bar{x}}{\bar{x}}\right)$$

さて

$$E\bar{x}(\bar{x} - \bar{x}) = E(\bar{x} - \bar{x})^2 = \frac{(N-n)}{N} \frac{S_x^2}{n}$$

同様に定理 2.3 (p.17) 及び ρ の定義により

$$E\bar{y}(\bar{y} - \bar{x}) = E(\bar{y} - \bar{x})(\bar{x} - \bar{x}) = \frac{(N-n)}{N} \frac{\rho S_y S_x}{n} \quad (6.10)$$

従って偏倚の初項は

$$E(\hat{Y}_R - Y) = \frac{(N-n)}{N\bar{x}} (R S_x^2 - \rho S_y S_x) \quad (6.11)$$

偏倚は正又は負の何れかである。標本の大きさを増すと、偏倚は $1/n$ をけ減じてくる。一方 $\hat{Y}_R$ の標準誤差は $1/\sqrt{n}$ をけ減ずる。特種な母集団に対しては偏倚が標準誤差に比して無視出来るほどの標本の大きさが存在する。

さて (6.5) から

$$\sigma(\hat{Y}_R) = \sqrt{\frac{N(N-n)}{n}} \sqrt{(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x)}$$

従って標準誤差に対する偏倚の比は近似的に

$$\frac{\text{偏倚}}{\sigma(\hat{Y}_R)} = \left\{ \sqrt{\frac{N-n}{N}} \frac{S_x}{\sqrt{n} \bar{x}} \right\} \left\{ \frac{(R S_x - \rho S_y)}{\sqrt{S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x}} \right\}$$

最初の括弧内の項は $\bar{x}$ の変動係数である。オニ第目の括弧内の項の絶対値は殆んど 1 に近い (これは平方すると明らかである) 従つて

$$\frac{|\text{偏倚}|}{\sigma(\hat{Y}_R)} \leq \bar{x} \text{ の変動係数} \quad (6.12)$$

もしも標本が $\bar{x}$ の CV が 0.1 以下である位の大きさであれば、偏倚は標準誤差に比して相対的に無視出来る (6.15)

6.4 推定分散 Estimated variance :-

方程式 (6.2) から

$$V(\hat{Y}_R) = \frac{N(N-n)}{n(N-1)} \sum_{i=1}^n (y_i - R x_i)^2$$

$$\frac{\sum_{i=1}^n (y_i - R x_i)^2}{N-1}$$

の標本推定量として

$$\frac{\sum_{i=1}^n (y_i - \hat{R} x_i)^2}{n-1}$$

を用いる。

この推定量は $1/n$ オーダーの偏倚がある事を示す事が出来る。即ち不偏推定量を得る事に対して利用出来る方法はない。

推定分散 $v(\hat{Y}_R)$ に対して

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

$$= \frac{N(N-n)}{n(N-1)} (\sum y_i^2 + \hat{R}^2 \sum x_i^2 - 2\hat{R} \sum y_i x_i) \quad (6.13)$$

これは計算が最もすみやかに出来る形である。更にこの表示を代数的に展開すると表示 (6.6) 及び (6.7) と同じ形となる。

比 $\hat{R}$ の推定分散に対しては (6.13) を X^2 で割る。即ち

$$v(\hat{R}) = \frac{(N-n)}{Nn(N-1)\bar{x}^2} (\sum y_i^2 + \hat{R}^2 \sum x_i^2 - 2\hat{R} \sum y_i x_i) \quad (6.14)$$

(6.13) 及び (6.14) に於ける平方和及び積和は標本平均に対する補正は含んでいない事に注意。もしも $\bar{x}$ が未知であれば標本推定量 $\bar{x}$ が代つて分母に用いられる。

(例)

これは母集団総計の比推定量の標準誤差の計算を示すものである。資料は前の8ノ表を用いる。

まず次のものを計算する。

$$\bar{y} = \sum y_i = 6262 : \quad \bar{x} = \sum x_i = 5054 : \quad \hat{R} = \bar{y}/\bar{x} = 1.239019$$

公式(6.13)を用いる。

$$v(\hat{Y}_R) = \frac{N(N-n)}{n(n-1)} (\sum y_i^2 + \hat{R}^2 \sum x_i^2 - 2\hat{R} \sum y_i x_i)$$

括弧内の項を計算するため平方和及び積和が次の如く乗数と並べてかけられる。

$$\sum y_i^2 = 1,527,882$$

乗 数

$$\sum x_i^2 = 1,044,504$$

$$1,53516 = \hat{R}^2$$

$$\sum y_i x_i = 1,251,630$$

$$2,478,038 = 2\hat{R}$$

従つて

$$v(\hat{Y}_R) = \frac{(196)(147)}{(49)(48)} (29,784) = 364.854$$

$$s(\hat{Y}_R) = 604$$

§ 6.5 標本の大きさ Sample size : —

特別な正確度を得るために必要な標本の大きさの推定は次の如くして行われる。最も便利な出発点は推定比 $\hat{R}$ の分散を与える方程式(6.8)である。この方程式を次の形にかきかえる。

$$C_{RR} = \frac{v(\hat{R})}{\hat{R}^2} = \frac{(N-n)}{Nn} (C_{yy} + C_{xx} - 2C_{yx}) \quad (6.8')$$

此処に C_{RR} は $\hat{R}$ の変動係数の平方である。もしも我々がこのCvの欲する値を特種化して即ち C_{RR} を特種化すれば方程式(6.8')は n に対して解ける。オ一段階はf.p.c $w = \frac{1}{N}$ を無視して n に対する近似式として

$$n_0 = \frac{C_{yy} + C_{xx} - 2C_{yx}}{C_{RR}}$$

もしもこの段階でf.p.cが必要である事が分れば(6.8')の正しい解は次の形におく事により得られる。

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

此の方法では $\hat{R}$ 及び推定母総計 $\hat{Y}_R$ 、及び単位当り推定母平均のCvはすべて等しい。従つて上の方程式はすべて三つの型の推定に適用される。この結果を用いるためには y_i 及び x_i のCv及び相関係数を前以て推定しなければならない。

§ 6.6 信頼限界 Confidence limits.

もしも標本の大きさが正規性近似が用いられる程充分大であれば $\bar{y}$ 及び $\hat{R}$ に対する信頼限界は次の如くして得られる。

$$Y: \quad \hat{Y}_R \pm t \sqrt{v(\hat{Y}_R)} \quad (6.15)$$

$$R: \quad \hat{R} \pm t \sqrt{v(\hat{R})} \quad (6.16)$$

此処に t は送ばれた信頼確率に應ずる正規偏差である。

§ 6.3に於て、もしも標本の大きさが少くとも30で $\bar{y}$ 及び $\bar{x}$ のCvが共に0.1以下である場合には正規性近似は合理的に用いてよい事を暗示した。この条件が適用されない時は $v(\hat{R})$ に対する公式はあまり小にすぎる値を与える傾向があり、 $\hat{R}$ の分布に於ける歪の歪が著しくなってくる。信頼限界を計算する際のそれに代る一つの方法が生物的分析に用いられた(Fieller, 1932; Paulson, 1942)。

この研究は正規性近似より以上の二、三の假定を必要とするし $\hat{R}$ の分布の歪を幾分考慮に入れてゐる。

その方法は $\bar{y}$ 及び $\bar{x}$ は $(\bar{y} - R\bar{x})$ が正規に分布する様に副変量正規分布に従ふ事が必要である。

次の量

$$\frac{\bar{y} - R\bar{x}}{\sqrt{\frac{N-n}{Nn} (S_y^2 + R^2 S_x^2 - 2R S_{yx})}} \quad (6.17)$$

は近似的に平均値の、標準偏差ノなる正規分布に従う（我々はそれに施する母分散及び共分散に対して標本推定量 s_y^2 等を用いこれが無視出来る程度の誤差を生ずる位の標本の大きさを假定している。生物学的分析に於ては標本は全く小さく、上に述べた量は Student の t 分布に従うと考えられるであろう。）

R の値は未知であるがこの正規偏差を大きくすると予期される R の値は標本資料によって棄却したと考える。その結果 R に対する信頼限界は (6.17) を t とおいて R に対する二次方程式を解く事により得られる。信頼限界は近似的なものである。何故ならばもしも我々が既知の R をもつ一定母集団からくりかえし抽出する事によってそれ等をチェックしようとするれば $\bar{y}$ 及び $\bar{x}$ のある値は二つの平方根が虚根となる様な事が起る。この様な場合は、もしも $\bar{y}$ 及び $\bar{x}$ の C が 0.3 以下であれば稀となる。ある場合の後に二つの平方根は次の様に表わされる。

$$R = \hat{R} \pm \frac{(1 - t^2 C_{\bar{y}\bar{x}}) \pm t \sqrt{(C_{\bar{y}\bar{y}} + C_{\bar{x}\bar{x}} - 2C_{\bar{y}\bar{x}}) - t^2 (C_{\bar{y}\bar{y}} C_{\bar{x}\bar{x}} - C_{\bar{y}\bar{x}}^2)}}{1 - t^2 C_{\bar{x}\bar{x}}}$$

(6.18)

此処に

$$C_{\bar{y}\bar{y}} = \frac{(N-n)}{Nn} \frac{s_y^2}{\bar{y}^2}$$

は $C_{\bar{y}\bar{x}}$ 及び $C_{\bar{x}\bar{x}}$ も同様 $\bar{y}$ の推定 C の平方である。もしも $t^2 C_{\bar{y}\bar{y}}$, $t^2 C_{\bar{x}\bar{x}}$ 及び $t^2 C_{\bar{y}\bar{x}}$ が N に比して小さければ限界は次の如くなる

$$R = \hat{R} \pm t \hat{R} \sqrt{C_{\bar{y}\bar{y}} + C_{\bar{x}\bar{x}} - 2C_{\bar{y}\bar{x}}}$$

この表示は正規性近似 (6.16) と同じである。

Y に対する二次限界は方程式 (6.18) に於ける $\hat{R}$ を $\hat{y}_R$ とおきかえると見られる。

二次限界は常に少なくとも二つの假定によるから正規限界と同様に近似である。併しこれ等は計算が面倒で問題に対して完全解ではない。何故なら母集団から標本抽出に於て $\bar{y}$ 及び $\bar{x}$ の分布はそ

れ等自身歪んでいるからである。

6.7 単位当りの平均値と比推定量との比較

Comparison of the ratio estimate with the mean per unit

前章で研究した Y の定量の型は $1/\bar{y}$ である。此処に $\bar{y}$ は標本に対する単位当りの平均値（単純無作為抽出に於て）であるか又は単位当りの加重平均（層化無作為抽出に於て）である。

この種の推定量は単位当り平均値に基いた推定量又は単純拡張により得られた推定量と呼ばれる。

定理 6.2

大標本に於ては、単純無作為抽出では、比推定量 $\hat{y}_R$ はもしも

$$P > \frac{1}{2} \left(\frac{S_x}{\bar{x}} \right) / \left(\frac{S_y}{\bar{y}} \right) = \frac{r_c \text{ の変動係数}}{2 (y_i \text{ の変動係数})}$$

で単純拡張により得られた推定量 $\hat{y} = \dots \bar{y}$ よりは分散が小さい。

(証明)

$\hat{y}$ に対して

$$V(\hat{y}) = \frac{N(N-n)}{n} S_y^2$$

比推定に対して我々は (6.5) から

$$V(\hat{y}_R) = \frac{N(N-n)}{n} \{ S_y^2 + R^2 S_x^2 - 2RPS_y S_x \}$$

従つて、もしも

$$S_y^2 + R^2 S_x^2 - 2RPS_y S_x < S_y^2$$

即ち

$$P > \frac{RS_x}{2S_y} = \frac{1}{2} \left(\frac{S_x}{\bar{x}} \right) / \left(\frac{S_y}{\bar{y}} \right)$$

であれば比推定はより小さい分散をもつ。

この定理は、比推定が単純拡張よりは多少正確である事を示す。論議は y_i 及び x_i の相関係数及び二つの変量の C, V の大きさによる。任意変量 x_i の変動は重要な因子である。

即ち もしもその C が y_i の C の2倍であれば比推定は P が1を越える事が出来ないから常に正確度が劣る事になる。 x_i がある以前の時期に於ける y_i の値である時、二つの C は殆んど等しい。この場合には、比推定は P が 0.5 以上であればすぐれてゐる。

定理 6.2 は $V(\hat{y}_R)$ に対する公式が妥当である程標本の大きさが大である場合にみに適用される。より小さい標本では比推定は恐らく、近似公式は通常過小評価であるから、定理が暗示する程にうまくは比較されないだろう。

§ 6.8. 比推定が最適である条件

Conditions under which the ratio estimate is optimum

回帰理論でよく知られている結果は、それに対して比較推定が推定の無い級の中で最良である。母集団の型を示してゐる。定理は無限母集団のみに適用される。

定理 6.3

無限母集団からの単純無作為抽出では $\bar{y}$ の比推定量はもしも次の二つの条件が満足されるとすれば、'最良線型不偏推定量' である。

- (i) y_i と x_i の関係は原点を通る直線である。
- (ii) この線まわりの y_i の分散は x_i に比例する。

'最良線型不偏推定量' とは次の様に定義される。 y_i なる標

本値の直線函数即ち

$$l_1 y_1 + l_2 y_2 + \dots + l_n y_n$$

なる形のすべての推定量を考えよう。此処に l は x_i の函数であつても、 y_i にはよらないものである。

l の選択は $\bar{y}$ の不偏推定量を与えるものに限定される。最小分散をもつ推定量が '最良線型不偏推定量' である。

(証明)

数学的モデルは

$$y_i = Bx_i + e_i$$

此処に e_i は x_i と独立である。 x_i が一定である場合は e_i は平均値 0 で分散は λx_i である。

従つて

$$\bar{y} = B\bar{x}$$

$B\bar{x}$ の最良線型不偏推定量は $b\bar{x}$ である事を Gauss は示した。此処に b は B の最小自乗推定量である。(David and Neyman, 1938 参照)。最小自乗推定量は

$$b = \frac{\sum w_i y_i x_i}{\sum w_i x_i^2}$$

此処に

$$w_i = \frac{1}{\sigma_{e_i}^2} = \frac{1}{\lambda x_i}$$

これは

$$b = \frac{\sum y_i}{\sum x_i} = \frac{\bar{y}}{\bar{x}}$$

を与える。

従つて $\bar{y}$ の最良推定量は、比推定量 $(\bar{y}/\bar{x})$ である。

この結果を実際的なものに関連させるには、比推定量は単位当たり平均値よりすぐれているばかりでなく、推定のすべての級の最良のものである条件を暗示する事である。我々がどの種類の推定量を用いるかを試みる場合には y_i が x_i に対してプロットされたグラフが助けとなる。このグラフが直線が原点を通る事を示し、直線まわりの y_i の分散が x_i に比例して増してゐる様に見えるれば、比推定は最良のものである。時には y_i と x_i の関係が原点を通る直線であるが x_i を固定した列での y_i の分散が x_i に比例しない場合がある。ギリシヤの人口の標本で Jensen 等 (1947) は、分散はラフに x_i^2 に比例する事を見出した。これはそれに於て $w_i \propto 1/x_i^2$ の重みをつけられた回帰を暗示してゐる。最小自乗推定量 b に対してこれは

$$b = \frac{\sum w_i y_i x_i}{\sum w_i x_i^2} = \frac{1}{\bar{x}} \sum \left(\frac{y_i}{x_i} \right)$$

この場合での $\bar{y}$ の最良推定量は $\bar{y}/\bar{x}$ である。此処に b は個々の抽出単位上での比 y_i/x_i の平均値である。

定理 6.3 の条件の下では比推定は不偏である。この結果は比推定量が一般的に偏倚がある前の論述 (§ 6.3) とは矛盾しない。比推定量は如何なる単純無作為標本に対しても、もしも母集団が無限で y_i 、 x_i の関係が原点を通る直線であれば不偏である。証明は読者の例題として残して置く。

§ 6.7 割合に対する標本抽出に於ける比推定

The ratio estimate in sampling for proportions

此の方法は割合の推定に重要な役割を演ずる。単純無作為抽出では推定割合 p の分散に対する通常の公式は

$$V(p) = \frac{pq}{n}$$

此処に p は母集団の割合である。〔因子 $(N-n)/(N-1)$ は n, p が必要であればつけられる〕 § 5.2 で指摘した様にこの公式はもしも標本が単位の単純無作為標本であり、その単位の各々が割合が導き出される二つの級に分割されておればそのときのみ妥当である。例えばもしも小麦畑での病気にかつた植物の割合が標本抽出によって推定されるとすればこの公式はもしも個々の植物の単純無作為標本がとられたとすると病気か健全かの何れかに植物が分類される際に適用される。この標本抽出法が用いられるという事はありとしない。より代表的な単位は面積例えば $1/4$ 寸長さ3つの列の巾をもつ面積である。その各々の抽出された面積での植物が分類される場合である。

この様な立場に於て標本抽出単位は我々が要素 (element) と呼びより小さい単位のグループ又は集落 (cluster) からなつてゐる。各番目の単位は x_i の要素からなつてゐる。各要素はこの級 c 又は c' の一つに割当てられる。 y_i をして c に属する各番目の単位に於ける要素の数としよう。母集団に於ける c に属する単位の割合は

$$p = \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i} = \frac{Y}{X}$$

この割合の標本推定量は

$$p = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = \frac{y}{x}$$

(136)

構造的にはこれは典型的な比推定量である。従つて p の分散は比の方法に用いた公式から得られる。

近似分散に対する二つの等しい型は

$$V(p) = \frac{N(N-n)}{n(N-1) \times 2} \sum_{i=1}^N (y_i - p x_i)^2$$

$$= \frac{(N-n)}{nN(N-1) \bar{x}^2} \sum_{i=1}^N x_i^2 (p_i - p)^2$$

此処に $\bar{x} = X/N$ は集落単位での要素の平均数である。

推定分散に対しては

$$v(p) = \frac{(N-n)}{Nn(n-1)\bar{x}^2} \left[\sum y_i^2 + p^2 \sum x_i^2 - 2p \sum y_i x_i \right] \quad (6.19)$$

(例)

30 家族の単純無作為標本が Baltimore の Eastern Health District の 6 及び 7 病棟に於て 1947 年行われたセンサスから抽出された。母集団は約 15,000 家族を含んでいる。

6.2 表には各家族での人が (i) 彼等が過去 12 ヶ月に医者にかゝつたかどうか (ii) 性に関してはどうかに関して分類されている。

6.2 表 30 家族の単純無作為標本の資料

家族番号	人数 x_i	男の数 女の数		過去医者に かゝつたか	
		y_i	z_i	ハ イ	イ エ
1	5	1	4	5	0
2	6	3	3	0	6
3	3	1	2	2	1
4	3	1	2	3	0
5	2	1	1	0	2
6	3	1	2	0	3
7	3	1	2	0	3
8	3	1	2	0	3
9	4	2	2	0	4
10	4	3	1	0	4
11	3	2	1	0	3

(137)

12	2	1	1	0	2
13	7	3	4	0	7
14	4	3	1	4	0
15	3	2	1	1	2
16	5	3	2	2	3
17	4	3	1	0	4
18	4	3	1	0	4
19	3	2	1	1	2
20	3	1	2	3	0
21	4	1	3	3	2
22	3	2	1	0	3
23	3	2	1	0	3
24	1	0	1	0	1
25	2	1	1	2	0
26	4	3	1	2	2
27	3	1	2	0	3
28	4	2	2	2	2
29	2	1	1	0	2
30	4	2	2	1	3
計	104	53	52	30	74

我々の目的は比推定公式と用いられない二項公式とを対照する事である。まづ医者にかゝつた人の割合を考えよう。二項公式に対しては

$$n = 104 : p = \frac{30}{104} = 0.2885$$

従つて
$$v_{bin}(p) = \frac{pq}{n} = \frac{(0.2885)(0.7115)}{104} = 0.00194$$

比推定公式に対しては 30 グループがありとられた事に注意する

$$n = 30$$

x_i = i 番目の家族の総数

y_i = 医者にかゝつた i 番目の家族での数

$$p = 0.2885 \quad (\text{前の通り})$$

$$\sum y_i^2 = 86 : \sum x_i^2 = 404 : \sum y_i x_i = 113$$

ナ、P、Q を無視する。従つて (6.19) から

$$v(p) = \frac{(56) + (0.2885)^2(404) - 2(0.2885)(113)}{(30)(29)(3.4657)^2} = 0.00520$$

比推定により与えられた分散 0.00520 は二項公式により与えられた分散 0.00197 よりもすこく大である。これは色々な理由に対して家族のそのメンバーが医者にかかった度数が違う事によるものである。全体としての標本に対して医者にかかった割合は4人の中1人を一寸越すだけであるが、いくつかの家族はすべてのメンバーが医者にかかったのがある。同様な結果は同じ家族のメンバーは同様になる傾向があるという特性に対して得られたものである。

母集団に於ける男性の割合を推定する事に於てはその結果は違ってくる。上の計算と同じ型により

$$\text{二項公式} : v(p) = 0.00240$$

$$\text{比公式} : v(p) = 0.00114$$

此処では二項公式が分散は過大推定をしている。その理由は興味がある。大抵の家族は結婚の結果出来上つたもので従つて1人の男と1人の女を含んでいる。従つて家族あたり男の割合は二項公式から期待されるよりも1/2から小さく変化する。1つだけのメンバーを除いて30家族の中どれかが全く男を含んでいないし、たゞ1人の女を含んでいるだけである。もしも近似1/2の真の p で以て二項分布が用いられれば同じ性をもつすべてのメンバーをもつ家族は大きくなる家族の1/4及び大きさ4なる家族の1/8を形成するであろう。性比のこの特性は Hansen & Hurwicz (1942) により論ぜられた。

§ 6.10 比の分布の正規性への接近

The approach to normality of the distribution of the ratio - 大標本に於ける比 $\bar{y}/\bar{x}$ の限界分布は正規であるという結果は確率に於ける標準の理論から出てくる。我々は前提として Cramér (1946) により与えられた広い一般的結果を引用しよう。無限母集団を仮定するこの結果は二つの中央モーメント m_1 及び m_2

即ち y_i, x_i なる値の対の標本から計算された既知の函数の限界分布を与える。標本の中央モーメントによつて、我々は次の形の量を意味する。即ち

$$m = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^u (x_i - \bar{x})^w$$

此処に u 及び w は正の整数である。それに対応する全母集団のモーメントは母集団でのすべての単位上で平均して

$$M = E (y_i - \bar{y})^u (x_i - \bar{x})^w$$

前提 (Cramér)

もしも $m_1 = M_1, m_2 = M_2$ なる点の附近に於て函数 $H(m_1, m_2)$ が連続で m_1 及び m_2 に関して一次及び二次の導函数をもてば、 n が大きくなると無作為変量 $H(m_1, m_2)$ の限界分布は

$$\text{平均値} = H(M_1, M_2)$$

$$\text{分散} = \sigma_{m_1}^2 \left(\frac{\partial H}{\partial m_1} \right)^2 + 2 \text{Cov}(m_1, m_2) \left(\frac{\partial H}{\partial m_1} \right) \left(\frac{\partial H}{\partial m_2} \right) + \sigma_{m_2}^2 \left(\frac{\partial H}{\partial m_2} \right)^2$$

をもつて正規である。此処に部分導函数は $m_1 = M_1, m_2 = M_2$ なる点で計算される。

定理 3.4

もしも母分散 S_y^2, S_x^2 が有限で母平均 $\bar{y} \neq 0$ ならば無限母集団からの大きさ n なる無作為標本に於て $\bar{y}/\bar{x}$ の限界分布は

$$\text{平均値} = \bar{y}/\bar{x} = R$$

$$\text{分散} = \frac{1}{n} R^2 \left[\frac{S_y^2}{\bar{y}^2} - \frac{2\rho S_y S_x}{\bar{y} \bar{x}} + \frac{S_x^2}{\bar{x}^2} \right]$$

で以て正規である。

(証明)

前提を適用するため

$$m_1 = \bar{y}; \quad m_2 = \bar{x}; \quad H(m_1, m_2) = \frac{m_1}{m_2}$$

とおく、従つて

$$M_1 = \bar{y}, \quad M_2 = \bar{x}$$

函数 H は連続で $\bar{x}$ が 0 であれば、 $\bar{y} = \bar{Y}$ 、 $\bar{x} = \bar{X}$ の附近で一次及び二次の連続部分導函数をもつ。更にこの点で

$$\frac{\partial H}{\partial m_1} = \frac{1}{M_2} = \frac{1}{\bar{x}} ; \quad \frac{\partial H}{\partial m_2} = -\frac{M_1}{M_2^2} = -\frac{\bar{y}}{\bar{x}^2}$$

従つて前提から $\bar{y}/\bar{x}$ の限界分布は、

$$\text{平均値} = \frac{\bar{Y}}{\bar{X}}$$

$$\begin{aligned} \text{分散} &= \frac{1}{n} \left[\frac{S_y^2}{\bar{x}^2} - 2 \rho S_y S_x \frac{\bar{y}}{\bar{x}^3} + \frac{\bar{y}^2 S_x^2}{\bar{x}^4} \right] \\ &= \frac{1}{n} R^2 \left[\frac{S_y^2}{\bar{y}^2} - \frac{2 \rho S_y S_x}{\bar{y} \bar{x}} + \frac{S_x^2}{\bar{x}^2} \right] \end{aligned}$$

で以て正規である。

結果は ρ 、 C を除いた前述の方程式 (6.8) と同じである。実際の適用に関する限り、結果の一般性は y_i 及び x_i が有限分散をもち、 $\bar{x}$ が 0 という拘束の下のみで用いられる。

Madrow (1948) が与えた定理は母集団の性質に更に厳密でない拘束を設けて、有限母集団にその結果を拡張している。この展開は此処では論じない。

§ 6.11 層化無作為抽出に於ける比推定

Ratio estimates in stratified random sampling :-

母集団総計 Y の比推定量を作るには多くの方法がある。その一つは各層の総計の個々の比推定量を作りこれ等を合計する方法である。もしも y_h 、 x_h が、 h 層での標本総計とし、 X_h が x_h の層総計とすれば、この推定量 $\hat{Y}_{RS}$ (個々の意味で S をつけた *separate*) は

$$\hat{Y}_{RS} = \sum_h \frac{y_h}{x_h} X_h = \sum_h \frac{\bar{y}_h}{\bar{x}_h} X_h \quad (6.20)$$

真の比が層毎に一定であるとの仮定はいらない。併し推定量は個々の総計 X_h の知識を必要とする。

定理 6.5

もしも標本の大きさ n_h がすべての層で大きければ

$$V(\hat{Y}_{RS}) = \sum_h \frac{N_h(N_h - n_h)}{n_h} [S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2 R_h \rho_{y_h x_h} S_{y_h} S_{x_h}] \quad (6.21)$$

此処に $R_h = y_h/x_h$ は層 h の真の比であり、 $\rho_{y_h x_h}$ は各層で前に定義されたものである。

(証明)

$$\hat{Y}_{R_h} = \frac{y_h}{x_h} X_h$$

従つて

$$(\hat{Y}_{RS} - Y) = \sum_h (\hat{Y}_{R_h} - Y_h)$$

よつて

$$\begin{aligned} V(\hat{Y}_{RS}) &= E(\hat{Y}_{RS} - Y)^2 \\ &= \sum_h E(\hat{Y}_{R_h} - Y_h)^2 + 2 \sum_{h \neq j} E(\hat{Y}_{R_h} - Y_h)(\hat{Y}_{R_j} - Y_j) \end{aligned}$$

$\hat{Y}_{R_h}$ は層 h 内の単純無作為標本から作られた比推定量であるから、我々は $\hat{Y}_{R_h}$ の近似分散に対する公式 (6.5) を用いる。即ち

$$V(\hat{Y}_{R_h}) = \frac{N_h(N_h - n_h)}{n_h} [S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2 R_h \rho_{y_h x_h} S_{y_h} S_{x_h}]$$

積の項はなくなる。何故ならば抽出は層毎に独立であり、分散公式に用いた近似のオーガーに対して $\hat{Y}_{R_h}$ は Y_h の不偏推定量であるからである。(6.21) の結果となる。

この公式はもしも各層に於ける標本が各層で近似分散公式が適用される程大であれば妥当である。この限界は実際の適用にあつて注意を要する。我々は n_h が小さいとき、 $\hat{Y}_{RS}$ に対する真に信用のおける公式はもっていない。

更に n_h が小さくて L が大であれば、 $\hat{Y}_{RS}$ の偏倚は次の大さつはな論議が暗示する様にその標準誤差に比して相対的に無視出来ない。

単一層では、次の事が分つた (§ 6.3)。

$$\frac{|\hat{Y}_{Rk} \text{ での偏倚}|}{\sigma(\hat{Y}_{Rk})} \leq \bar{x}_k \text{ の変動係数}$$

もしもしばしばある様に偏倚がすべての層で同じ符号をもてば、 $\hat{Y}_{RS}$ での偏倚はラフに $\hat{Y}_{Rk}$ の L 倍であろう。併し $\hat{Y}_{RS}$ の標準誤差は、 $\hat{Y}_{Rk}$ の $\sqrt{L}$ 倍のオーダーにすぎない。

従って 比

$$\frac{|\hat{Y}_{RS} \text{ での偏倚}|}{\sigma(\hat{Y}_{RS})}$$

は $\sqrt{L}$ ($\bar{x}_k$ の変動係数) の大きさである。

例えば、50の層で $\bar{x}$ の CV が各層で 0.1 とすれば、 $\hat{Y}_{RS}$ の偏倚はその標準誤差の 0.7 倍の大きさである。

今の我々の知識に於ては $\hat{Y}_{RS}$ は $\sqrt{L}$ ($\bar{x}_k$ の CV) が 0.2 以下であればさけられる。この規則は恐らくひかえめすぎるであろう。何故ならば実際には偏倚は特にもしも各層で y_{ki} と x_{ki} との関係が近似的に原点を通る直線であれば上の限界よりは小さくなるであろうからである。

§ 6.12 組合せ比推定量

The Combined ratio estimate :-

その他の一つの推定量が単一組合せ比から導かれる (Hansen, Hurwicz & Gurney, 1946)、標本資料から、我々は次のものを計算する。

$$\hat{Y}_{st} = \sum_k N_k \bar{y}_k : \hat{X}_{st} = \sum_k N_k \bar{x}_k$$

これ等は層化標本から作られた夫々母総計 Y 及び X の標本の推定量である。組合せ比推定量 $\hat{Y}_{RC}$ (組合せに新しく C をつけた

Combined) は

$$\hat{Y}_{RC} = \frac{\hat{Y}_{st}}{\hat{X}_{st}} X = \frac{\bar{y}_{st}}{\bar{x}_{st}} X$$

此処に $\bar{y}_{st} = \hat{Y}_{st}/N$ 、 $\bar{x}_{st} = \hat{X}_{st}/N$ は層化標本からの推定母平均である。

推定量 $\hat{Y}_{RC}$ は X_k の知識の必要がなく、たゞ X が必要なだけである。

定理 6.6

もしも全標本の大きさが大であれば

$$V(\hat{Y}_{RC}) = \sum_k \frac{N_k(N_k - n_k)}{n_k} [S_{y_k}^2 + R^2 S_{x_k}^2 - 2R \rho_k S_{y_k} S_{x_k}] \quad (6.22)$$

(証明)

これは定理 6.1 と同じ論議に従う。現在の場合、鍵になる方程式 (6.4) は次の様になる。

$$(\hat{Y}_{RC} - Y) \doteq N(y_{st} - R\bar{x}_{st}) \quad (6.23)$$

さて変量 $u_{ki} = y_{ki} - R x_{ki}$ を考えよう。方程式 (6.23) の右辺は $N\bar{u}_{st}$ である。此処に $\bar{u}_{st}$ は層化標本に於ける変量 u_{ki} の荷重平均である。更に u_{ki} の母平均 $\bar{u}$ は $R = \bar{Y}/\bar{X}$ であるからである。

従って $\bar{u}_{st}$ に層化無作為標本からの推定平均値の分散に対する定理 5.3 を適用する。これは

$$V(\hat{Y}_{RC}) = N^2 V(\bar{u}_{st}) = \sum_k \frac{N_k(N_k - n_k)}{n_k} S_{u_k}^2$$

此処に

$$\begin{aligned} S_{u_k}^2 &= \frac{1}{(N_k - 1)} \sum_{i=1}^{N_k} (u_{ki} - \bar{u}_k)^2 \\ &= \frac{1}{(N_k - 1)} \sum_{i=1}^{N_k} \left\{ (y_{ki} - \bar{y}_k) - R(x_{ki} - \bar{x}_k) \right\}^2 \end{aligned}$$

平方を展開すると (6.22) の結果が得られる。

方程式 (6.21) 及び (6.22) から $\hat{Y}_{RS}$ と $\hat{Y}_{RC}$ との近似分散がたゞ違っているのは (6.21) で個々の層での母集団比 R_k が (6.22) では R になっただけで同じ一般型を仮定している事に注意する事は興味ある事である。

○ 二つの推定量の比較

次の如くかく

$$\begin{aligned}
 V(\hat{Y}_{RC}) - V(\hat{Y}_{RS}) \\
 &= \sum_h \frac{N_h(N_h - n_h)}{n_h} [(R^2 - R_h^2) S_{Yh}^2 - 2(R - R_h) \rho_h S_{Yh} S_{Xh}] \\
 &= \sum_h \frac{N_h(N_h - n_h)}{n_h} [(R - R_h)^2 S_{Xh}^2 + 2(R_h - R)(\rho_h S_{Yh} S_{Xh} - R_h S_{Xh}^2)]
 \end{aligned}$$

比推定が用いられる場合には右辺の最後の項は非常に小さくなる(もしも各層内で y_{hi} と x_{hi} の関係が原点を通る直線であれば)となる)。かくて R_h が層毎に一定でなければ各層内で別個な比推定を用いる事がより正確である様である。併しこの論議は各層内での標本が $V(\bar{Y}_{RS})$ に対する公式が妥当である位充分大であると仮定している。各層で標本が小さければ組合せ推定がそれを反対する充分な経験的証拠がなければ推奨出来るものである。

これ等の分散の標本推定量に対しては用いる場合に R_h 及び R の標本推定量をおけばよい。標本平均平方 S_{Yh}^2 及び S_{Xh}^2 はそれに応ずる分散に対して代りに用いられ $\rho_h S_{Yh} S_{Xh}$ に対しては標本共変量が用いられる。標本平均平方及び共分散は各層で別個に計算されねばならない。

(例)

資料は Iowa 州 Jefferson County に於けるすべての農場のセンサスからとつたものである。

この例では y_{hi} は穀物の面積、 x_{hi} は農場の全面積を表わす。母集団は二つの層に分けられる。第一の層は 160 エーカー迄の大きさの農場を含む。我々は 100 農場の標本を仮定する。

層化抽出が用いられる時、我々はこれは大ざつぱに最適割当てであるから、70 農場を層 1 から 30 農場を層 2 からとつたと考える。資料は 63 表に与えてある。

63 表 Iowa 州 Jefferson County からの資料

層	大きさ (農場I-0-1)	N_h	S_{Yh}^2	S_{YXh}	S_{Xh}^2	R_h
1	0-160	1580	312	494	2055	0.2350
2	160以上	430	922	858	7357	0.2109
全母集団に対して		2010	620	1453	7619	0.2242
層	$\bar{Y}_h$	$\bar{X}_h$	n_h	$Q_h = \frac{V_h^2}{N_h}$	V_h'	V_h''
1	12.40	82.56	70	0.008828	193	194
2	51.63	244.85	30	0.001525	887	907
全母集団に対して	26.30	117.28	100			

最後の三つの量 Q_h , V_h' , V_h'' は計算に用いられる任意の量であり最後の二つはあとで定義する。

農場当り穀物の母平均を推定する五つの方法を考えよう。ナ、P、C は無視する。

i. 単純無作為標本 : 農場当り平均値の推定

$$V_1 = \frac{S_y^2}{n} = \frac{620}{100} = 6.20$$

ii. 単純無作為標本 : 比推定

$$\begin{aligned}
 V_2 &= \frac{1}{n} [S_y^2 + R^2 S_x^2 - 2RS_{yx}] \\
 &= \frac{1}{100} [620 + (0.2242)^2 (7619) - 2(0.2242) \\
 &\quad (1453)] = 3.51
 \end{aligned}$$

iii. 層化無作為標本 : 農場当り平均値の推定

$$V_3 = \frac{1}{N^2} \sum_h \frac{N_h^2}{n_h} S_{Yh}^2 = \sum_h Q_h S_{Yh}^2 = 4.16$$

IV、層化無作為標本：各層に於て別個な比を用いる比推定

$$V_4 = \sum Q_h [S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{yhxh}] = \sum Q_h V_h' = 3.06$$

V、層化無作為標本：組合せ比を用いる比推定

$$V_5 = \sum Q_h [S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{yhxh}] = \sum Q_h V_h'' = 3.10$$

色々な方法の相対的正確度をまとめるに次の如くである。

抽出方法	推定方法	相対的正確度
(i) 単純無作為	農場当り平均値	100
(ii) 単純無作為	比	177
(iii) 層化無作為	農場当り平均値	149
(iv) 層化無作為	別個な比	203
(v) 層化無作為	組合せ比	200

結果は広汎な応用に対して興味ある論点をもたらす。農場の大きさによる層化は分母が農場の大きさである比推定と同じ一般的目的を満足している。二つの工夫は農場当り推定穀物平均面積の抽出誤差への農場の大きさの変動の影響をなくしている。例えば比推定からの正確度の利益は無作為抽出を用いたときの74%であり、層化抽出を用いたときの36%にすぎない(149に対して203)。標本の計画に於て層化にある因子を導入するか、推定方法にそれを利用するかどちらかの選択がある。最良の決定はその場合の周囲の状況による。関係した指摘点は次の通りである。

- (i) ある因子、例えば「地図上の位置」の如きものは推定方法よりも層化に容易に適入される。
- (ii) 論議は異なるものの関係の性質による。すべての簡単な相定方法は直線、曲線とき最も有効である。複雑か又は不連続な関係では層化はより有効である。何故ならば層の数が必要ならば層化は \$y_i\$ 及び \$x_i\$ の殆んどのある関係の影響をうちけすであろうからである。

§ 6.13 比推定での最適割当て

Optimum allocation with a ratio estimate:—

\$n_h\$ の最適割当ては、比推定が用いられる時と単位当り平均値が用いられる時とは違ってくる。

まづ変量 \$\bar{Y}_{RS}\$ を考えよう。定理 6.5 からその分散は

$$\begin{aligned} V(\bar{Y}_{RS}) &= \sum \frac{N_h(N_h - n_h)}{n_h} [S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{yhxh}] \\ &= \sum \frac{N_h(N_h - n_h)}{n_h} S_{dh}^2 \end{aligned}$$

此処に

$$S_{dh}^2 = \frac{1}{(N_h - 1)} \sum_{i=1}^{N_h} d_{hi}^2 \quad (6.24)$$

で、\$d_{hi} = y_{hi} - R_h x_{hi}\$ は \$R_h x_{hi}\$ からの \$y_{hi}\$ の偏差である。最適割当てを見出す第五章で与えた方法により (6.24) は

$$n_h \propto \frac{N_h S_{dh}}{\sqrt{C_h}}$$

なるとき \$\sum C_h n_h\$ なる形の総費用に対して最小になる。

単位当りの平均値では、最小分散に対して \$n_h\$ は \$N_h S_{yh} / \sqrt{C_h}\$ に比例して選ばれる事を思いおこしてみれば分る。

抽出計画に於て比推定での割当てはいささか面倒である。何故ならば \$S_{dh}\$ のありそうな値を見出す事が困難であるからである。次の二つの規則が助けとなる。比推定が最良線型不偏推定量であり、母集団では大抵 \$S_{dh}\$ は \$\sqrt{x_h}\$ に比例する(定理 6.3 により)。この場合は \$n_h\$ は \$N_h \sqrt{x_h} / \sqrt{C_h}\$ に比例しなければならない。時には \$d_{hi}\$ の分散は殆んど \$x_h^2\$ に比例する。これは、\$N_h x_h / \sqrt{C_h}\$ に比例する。即ち単位当り費目の平方根で割った \$x_{hi}\$ の層総計に比例する \$n_h\$ の割当てが導かれる。この種の例は小売店の売上げ推定に対して計画された標本に対して Hansen, Hurwitz 及び Gurney (1946) が論じている。もしも推定量 \$\hat{Y}_{RC}\$ が用いられるとしても、同じ一般的論法が適用される。

(例)

割当ての違った方法が 1946 年六月 North Carolina の 257 の商業用梨果樹園の完全調査で集められた資料で比較出来る。この調査の目的はこの地方の商用梨生産量を推定する最も有効な抽出方法を決定する事であつた。情報が梨の木の数及び各果樹園での推定全生産量について得られた。二つの変量間の高度の相関は比推定の使用を暗示した。非常に大きな一つの果樹園を除いた。

この例に対して面積は図的に三つの層に分けられる。果樹園の梨の木数は N_h で表われ梨のブッシェルでの推定生産量は y_h で表わす。

最初に比推定量 (各層で別個な比に基いた) のみが考えられる。何故ならば原則としては比推定の両方の型に対して同じであるからである。

次の割当ての違った方法に対して比較しよう。

- (i) n_h が N_h に比例する。
- (ii) n_h が $N_h S_{y_h}$ に比例する。
- (iii) n_h が $N_h \sqrt{x_h}$ に比例する。
- (iv) n_h が $N_h \bar{x}_h = x_h$ に比例する。

100 の標本の大きさを考える。此の比較に必要な資料は 6.4 表に示してある。

6.4 表 North Carolina の梨の調査からの資料

層	Sx_h^2	Sy_h^2	Sx_h	Sy_h	$\bar{x}_h$	$\bar{y}_h$	R_h	Sdx_h^2
1	5,186	6,462	72.01	23.27	53.80	69.48	1.29133	658
2	2,367	3,100	48.65	62.93	31.07	43.64	1.40475	573
3	4,877	4,817	69.83	85.51	56.97	66.39	1.16547	2,706
pop	3,898	4,434	62.43	80.06	44.45	56.47	1.27053	1,433

層	N_h	(i)	$N_h S_{y_h}$	(ii)	$\sqrt{x_h}$	$N_h \sqrt{x_h}$	(iii)	$N_h \bar{x}_h$	(iv)
1	47	18	4,384	22	7.33	344.5	20	2,529	22
2	118	46	8,016	40	5.57	657.3	39	3,666	32
3	91	36	2,781	38	2.55	232.1	41	5,184	46
pop	256	100	29,181	100	29.45	1,688.9	100	11,379	100

表の上の部分は基礎資料を表わす。四つの分散を計算するために用いた方法はまづ割当ての各型に対して n_h を見出す事である。これ等の値は表の下部分の (i) - (iv) なる表題の欄に示してある。かくて割当て (i) では $n_h = n N_h / N$ で第一の層では

$$n_1 = \frac{(100)(47)}{256} = 18$$

n_h が得られると、それに応ずる $V(\bar{y}_{RS})$ は次の公式に代入して得られる。

$$V(\bar{y}_{RS}) = \sum_h \frac{N_h(N_h - n_h)}{n_h} S_{dx_h}^2$$

此処に

$$S_{dx_h}^2 = S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h S_{y_h x_h}$$

量 $S_{dx_h}^2$ は四つの割当てに対して同じである。そしてこれは 6.4 表の上半分の一番右端に与えてある分散及び相対的正確度は 6.5 表に示してある

6.5 表 四つの割当て方法の比較

割当て方法： n_k は次のものに 比例する	分 散			相 対 的 正 確 度	
	層				
	1	2	3		
(i) N_k	49.824	105.833	376.215	531.872	100
(ii) $N_k S y_k$	35.144	131.847	343.446	510.437	104
(iii) $N_k \sqrt{x_k}$	41.750	136.964	300.312	479.026	111
(iv) $N_k \bar{x}_k$	35.144	181.710	240.888	457.742	116

期待していた様に n_k は四つの方法で大して違わないから違った割当てを選ぶ際に大して差はない。方法 (iv) はこれは層内での梨の木の数に比例する様に割当てるのであるが、他よりは少しはよい様である。

第 7 章 回 帰 推 定

Regression estimates

§ 7.1 直線回帰推定

The linear regression estimate

此推定と同様、直線回帰推定は y_c と相関がある。任意変量 x_c を用いて正確度を増すために計画される。 y_c と x_c との間の関係が検定されると、たとえその関係が近似的に直線であつても、直線が原点を通らない場合がある。此は二つの変量の比よりもむしろ x_c への y_c の直線回帰に基いた推定量を暗示してゐる。我々は、 y_c 及び x_c が標本に於けるすべての単位に対して各々が得られて x_c の母平均 $\bar{x}$ が既知であると假定しよう。 x_c への y_c の標本回帰が計算される。今我々は最小自乗回帰係数 β が用いられると假定する。此処に

$$\beta = \frac{\sum_{c=1}^n (y_c - \bar{y})(x_c - \bar{x})}{\sum_{c=1}^n (x_c - \bar{x})^2}$$

y_c の母平均 $\bar{y}$ の直線回帰推定量は、

$$\bar{y}_{cr} = \bar{y} + \beta(\bar{x} - \bar{x}) \quad (7.1)$$

此処に添字 cr は直線回帰 (linear regression) を示す。この推定量の基礎は、もしも $\bar{x}$ が平均値以下であれば、 x_c への y_c の回帰のため $\bar{y}$ も亦量 $\beta(\bar{x} - \bar{x})$ により平均以下となる事を期待する事にある。母総計 Y の推定量に対しては $\bar{y}_{cr} = Y/\bar{y}_{cr}$ である。Watson (ノラコク) は植物の葉の平均面積推定のため葉重量への葉面積の回帰を用いた方法は植物のすべての葉の重さを測る事であつた。標本平均葉面積は葉の重さへの回帰によって調整された。勿論その応用の論点は葉の重さは迅速に測定出来るが、その面積の決定は非常に時間がかかる事にある。この例は回帰推定が本能的に助けとなる一般的な立場を示してゐる。

る。我々はあらゆる単位に対してある特性をもつ迅速に得られる推定量 $\bar{y}_c$ をもつ事が出来亦いくらか費用がかかる方法で単位の単純無作為標本に対して特性をもつ正しい値 y_c を決定出来るかと仮定しよう。例えばネズミの調査家は都市の各ブロックでのネズミの数を迅速に目測出来せしめてワナをかける事によつてブロックの単純無作為標本の各々に於てネズミの実際の数を決定出来る。

Yates (ノタケタ)によつて述べられた他の応用に於ては、木材材積の目測が $\frac{1}{10}$ エーカープロットの各々で行われ眞の材積がプロットの標本で測定された例もある。

回帰推定量

$$\bar{y} + b(\bar{x} - \bar{x})$$

は迅速な推定への実際の測定の回帰によつて実際の測定の標本平均を補正する。迅速な推定量は偏倚を除く必要はない。もしも $\bar{x}_c - \bar{y}_c = D_c$ であり迅速な推定量が一定偏倚 D に対して以外は完全であれば $b = 1$ である事は明らかで回帰推定量は次の如くなる。

$$\bar{y} + (\bar{x} - \bar{x}) = \bar{x} + (\bar{y} - \bar{x})$$

= (迅速な推定量の母平均) + (偏倚に対する補正)

回帰推定の特性の我々の知識は殆んど比推定に対する我々の知識と同じ見方であるこれはつまらない意味であるが、標本が全母集団 $\bar{x} = \bar{y}$ で成立ち回帰推定が補正なしで行われる時には回帰推定は一致性があるといふ。これから先分る様に推定量は一般に偏倚があるか標準誤差に対する偏倚の比は、標本が大きいときには小さくなる。我々は推定量の分散に対し大標本の公式をもつてゐるが、小標本の推定量の分布及び大標本の結果を実際的に用いる際に必要な n の値についてより多くの情報が必要である。

§ 22 大標本の理論

Large-sample theory

直線回帰の理論は基礎統計学に於ける大抵のコースで非常に重要な役割を演ずる。この理論の標準の結果は標本調査には全く適当

でない。何故ならば $\bar{x}_c$ への y_c の母集団回帰が直線であり回帰直線まわりの y_c の残差分散は一定であるといふ抱索的假定を必要とするからである。もしもこれ等二つの假定が著しく違つてゐるとすれば直線回帰推定は恐らく正確でないであらうし、曲線回帰又は荷重直外回帰に基いた推定量がより良好である。

併し、我々はよりすぐれた方法からの正確度に対する利益が労力を払う価値があるのたろうかどうか疑はしい場合があるし又、それに於て回帰が直線であるといふ信すべ理由があつても、実際に直線であるといふ良好な証據をもたない他の場合がある。

従つて我々はまづ回帰が母集団で直線であると假定せず大標本のみに成立つ結果を与える理論を示そう。此の理論は此推定に対する大標本に対するものと同じである。

有限母集団直線回帰係数 B は次の関係により定義される。

$$B = \frac{\sum_{i=1}^N (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad (22)$$

残差変量 e_i は次の関係により定義される

$$y_i = \bar{y} + B(x_i - \bar{x}) + e_i \quad (23)$$

母集団のすべての単位上で (23) を加えると

$$\sum_{i=1}^N e_i = 0 \quad (24)$$

y_i と x_i の間には直線関係はないと假定してゐる事に注意を要する。母集団は (y_i, x_i) なる値の N 個の対から成立つてをり、方程式 (23) の明らかな直線回帰は B 及び e_i の定義により作られた。

定理 21

標本回帰係数 b の抽出誤差が無視出来る程 n が充分大である単純無作為抽出では

$$V(\bar{y}_{cr}) = \frac{(N-n)}{N} \frac{S_e^2}{n} \quad (25)$$

此処に

$$S_e^2 = \frac{\sum_{i=1}^V (e_i^2)}{V-1}$$

(証明)

定義により

$$\bar{y}_{cr} = \bar{y} + b(\bar{x} - \bar{x})$$

(21)

併し(23)から標本でのすべての単位上で平均して

$$\bar{y} = \bar{y} + B(\bar{x} - \bar{x}) + \bar{e}$$

(21)に代入すると

$$\bar{y}_{cr} = \bar{y} + (b-B)(\bar{x} - \bar{x}) + \bar{e}$$

(24)

従つて、もしも抽出誤差 $(b-B)$ が無視出来ると

$$\bar{y}_{cr} - \bar{y} = \bar{e}$$

故に $V(\bar{y}_{cr}) = E(\bar{y}_{cr} - \bar{y})^2 = E(\bar{e}^2)$

併し $E(\bar{e})$ は、方程式(24)により0であるから $E(\bar{e}^2)$ は単純無作為標本での量 e_i の平均値の分散である。従つて定理2.3により

$$V(\bar{y}_{cr}) = \frac{(V-n)}{V} \frac{S_e^2}{n}$$

系

有限母集団に於て y_i と x_i 間の相関係数 r が次の関係

$$r = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}} = B \sqrt{\frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2}} \quad (27)$$

によつて定義される

$$V(\bar{y}_{cr}) = \frac{(V-n)}{Vn} S_y^2 (1-r^2) \quad (28)$$

此処に和は母集団に於けるすべての単位について行われる。

(証明)

方程式(23)から、母集団全体に加えて

$$\sum e_i^2 = \sum \{(y_i - \bar{y}) - B(x_i - \bar{x})\}^2$$

$$= \sum (y_i - \bar{y})^2 - 2B \sum (y_i - \bar{y})(x_i - \bar{x}) + B^2 \sum (x_i - \bar{x})^2$$

$$= \sum (y_i - \bar{y})^2 - B^2 \sum (x_i - \bar{x})^2$$

(方程式(22)の B の定義により)

$$= \sum (y_i - \bar{y})^2 (1-r^2)$$

(方程式(27)の r の定義により)

$$\text{従つて } S_e^2 = S_y^2 (1-r^2)$$

定理2.2

標本回帰係数に於ける抽出誤差が無視出来る程 n が大である単純無作為抽出では標本からの $V(\bar{y}_{cr})$ の不偏推定量は

$$v(\bar{y}_{cr}) = \frac{(V-n)}{Vn(n-1)} \sum_{i=1}^n \{(y_i - \bar{y}) - b(x_i - \bar{x})\}^2$$

(証明)

定理2.4から標本からの S_e^2 の不偏推定量は

$$S_e^2 = \frac{1}{(n-1)} \sum_{i=1}^n (e_i - \bar{e})^2$$

さて 方程式(23)から

$$e_i - \bar{e} = y_i - \bar{y} - B(x_i - \bar{x})$$

$$= \{(y_i - \bar{y}) - b(x_i - \bar{x})\} + (b-B)(x_i - \bar{x})$$

もしも b に於ける抽出誤差が無視出来れば最後の項はなくなり

$$\sum_{i=1}^n (e_i - \bar{e})^2 = \sum_{i=1}^n \{(y_i - \bar{y}) - b(x_i - \bar{x})\}^2$$

従つて方程式(27)の如く定義された $v(\bar{y}_{cr})$ は方程式(25)により

$$E v(\bar{y}_{cr}) = \frac{(V-n)}{Vn} S_e^2 = V(\bar{y}_{cr})$$

定理2.1及び2.2は標本回帰係数 b を如何にして計算するかを明らかにしていない。もしも b が最小自乗回帰係数であれば

$v(\bar{y}_{cr})$ に於ける平方の和は次の形で迅速に計算される。

$$\sum (y_i - \bar{y})^2 = \frac{[\sum (y_i - \bar{y})(x_i - \bar{x})]^2}{\sum (x_i - \bar{x})^2}$$

もしも β が最小自乗回帰係数でなければ前の公式は成立しない。
平方和は次の様にして計算される。

$$\sum (y_i - \bar{y})^2 = 2\beta \sum (y_i - \bar{y})(x_i - \bar{x}) + \beta^2 \sum (x_i - \bar{x})^2$$

§ 23 基礎理論

Elementary theory

前の理論は次の問題に対して答えられないものである。即ち 標本はどの位の大きさでなければならぬか？ 如何なる母集団に対しても妥当である完全な解答は未だ分らないが標準的回帰理論の結果を検討する事によつてある情報が得られる。此れに於て母集団は無限と假定し、 y_i と x_i の関係は

$$y_{ij} = \bar{y} + B(x_i - \bar{x}) + e_{ij} \quad (23')$$

とする

形式的にはもう一つの添字 j が標準回帰理論に於ては x_i の各値に対して y_{ij} 及び e_{ij} の値の全度数分布又は列があるといふ注意として附加えられた以外は方程式(23')と同じである。

理論は単純無作為抽出と假定し、更に x_i を一定にした場合各列に於て $E(e_{ij}) = 0$: $E(e_{ij}^2) = S_e^2 = \text{一定}$ と假定する

このモデルから定理21と同じ分析により

$$\bar{y}_{er} - \bar{y} = \bar{e} + (\beta - B)(\bar{x} - \bar{x}) \quad (24)$$

さて、もしも β をして最小自乗回帰係数とすれば

$$\beta = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

此処にも j の添字はなくなつてゐる。方程式(23')から y_i 及び $\bar{y}$ に代入すると

$$\beta = B + \frac{\sum e_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \quad (25)$$

$$\beta - B = \frac{\sum e_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \quad (26)$$

x_i が標本が變つても一定であるくりかへしの抽出では $\bar{e}$ 及び $(\beta - B)$ の共分散は0であり

$$V(\beta) = E(\beta - B)^2 = \frac{S_e^2}{\sum (x_i - \bar{x})^2}$$

である事を証明するのは容易である。

従つて(210)から

$$\begin{aligned} V(\bar{y}_{er}) &= E(\bar{y}_{er} - \bar{y})^2 = \frac{S_e^2}{n} + (\bar{x} - \bar{x})^2 V(\beta) \\ &= S_e^2 \left\{ \frac{1}{n} + \frac{(\bar{x} - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\} \quad (213) \end{aligned}$$

これは基礎理論に於ける標準の結果である。前に述べた假定の下では如何なる標本の大きさに対しても正確である。

これ等の假定の下では、 $\bar{y}_{er}$ は $\bar{y}$ の不偏推定量である。此れは x_i の値が標本が變つても一定であるくりかへしの抽出を考える事によつて明らかとなる。假説により、この様な標本では $E(e_{ij}) = 0$ であるから $E(\bar{e}) = 0$ で、方程式(212)から $E(\beta - B) = 0$ となる。今や方程式(210)に於ては我々は $\bar{y}_{er}$ は $\bar{y}$ を一定にしたくりかへしの標本では不偏であり、従つてくりかへして単純無作為抽出で不偏である事が分る。

これ等の結果は定理21の更に一つ他の結果を導く。

定理 21a

この節の始めに述べた假定の下では、 $\bar{y}_{er}$ は方程式(213)に与えられた分散をもつ $\bar{y}$ の不偏推定量である方程式(213)で x_i を含む項は β の抽出誤差から生じたものである。大ききなる単純無作為標本に於けるこの項の平均値は x_i の度数分布の型による。もしもこの分布が正規であれば

$$E \left[\frac{(\bar{x} - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] = \frac{1}{n(n-3)}$$

である事が分る。

x_i が正規に分布していない時には、 x_i に於ける項の平均値は Cumulants の Fisher の方法(1923)により次の逆算のシ

リースに展開される。此のシリーズの初項は近似として $\sum (x_i - \bar{x})^2$ を $(n-1)S_x^2$ とおきかえる事によつて見出される。即ちこの近似のオーダーでは

$$E \left\{ \frac{(\bar{y} - \bar{y})^2}{(n-1)S_x^2} \right\} = \frac{S_y^2}{n(n-1)S_x^2} = \frac{1}{n(n-1)} = \frac{1}{n^2}$$

従つて $1/n^2$ オーダーの項に対して

$$E \{ V(\bar{y}_{er}) \} = \frac{S_y^2}{n} \left(1 + \frac{1}{n} \right) \quad (2.14)$$

この結果は、もしも n が 50 以上になれば最小自乗法に於ける抽出誤差の貢献は無視出来る事を示してゐる。

この結果は母集回帰が直線であるといふ假定をうけてゐる。この回帰が直線でないときは $V(\bar{y}_{er})$ への y の抽出誤差の貢献は n の逆数のシリーズで展開出来る。初項は方程式 (2.14) の如き $1/n^2$ オーダーであるが分子は ℓ_i 及び x_i の結合分布のあるモーメントの函数である。(Cochran, 1942)

基礎理論を完全にするため、如何なる標本の大きさに対しても等しい他の標準の結果は

$$s_{yx}^2 = \frac{1}{(n-2)} \sum_{i=1}^n \left\{ (y_i - \bar{y}) - b(x_i - \bar{x}) \right\}^2$$

は S_y^2 の不偏推定量であるといふ事である。これは定理 2.2 で与えた大標本の結果と分母が此処では定理 2.2 の $(n-1)$ に対して、 $(n-2)$ になつてゐるといふ事だけが違つてゐる又 F, R, C が含まれてゐない。

§ 2.4 回帰推定の偏倚

Bias of the regression estimate.

もしも y_i と x_i の関係が直線でなければ、 $\bar{y}_{er}$ は $1/n$ オーダーの偏倚をうける。我々は § 2.2 の有限母集団のモデルを仮定しよう。方程式 (2.6) から推定量の誤差は

$$\bar{y}_{er} - \bar{y} = \bar{e} + (b - B)(\bar{x} - \bar{x})$$

もしも b が B の最小自乗推定量であれば、方程式 (2.12) により

$$b - B = \frac{\sum \ell_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

従つて

$$\bar{y}_{er} - \bar{y} = \bar{e} + \frac{(\bar{x} - \bar{x}) \sum \ell_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

くりかえしの單純無作為標本では、 ℓ_i の母平均は 0 であるから、 $E(\bar{e}) = 0$ 。右辺の第二項の平均値はその逆数のシリーズで展開出来る。初項は次の最密ではないう減によつて得られる。我々は分母 $\sum (x_i - \bar{x})^2$ を近似 $(n-1)S_x^2$ とおきかえる。分子はその次にかける。

$$-(\bar{x} - \bar{x}) \left\{ \sum \ell_i (x_i - \bar{x}) - \sum \ell_i (\bar{x} - \bar{x}) \right\} \quad (2.16)$$

u_i をして変量 $\ell_i (x_i - \bar{x})$ としよう。そのとき有限母集団に於ける B の定義により

$$\begin{aligned} \sum_{i=1}^N u_i &= \sum_{i=1}^N \ell_i (x_i - \bar{x}) \\ &= \sum_{i=1}^N \left\{ (y_i - \bar{y}) - B(x_i - \bar{x}) \right\} (x_i - \bar{x}) = 0 \end{aligned}$$

従つて母平均 $\bar{U} = 0$ 。その結果 (2.16) の初項の平均値は次の減にかける。

$$-n E(\bar{x} - \bar{x})(\bar{U} - \bar{U}) = \frac{-n(N-n)}{N(N-1)} \sum_{i=1}^N (x_i - \bar{x})(u_i - \bar{U})$$

定理 2.3 (P17) により

$$= -\frac{(N-n)}{N(N-1)} \sum_{i=1}^N \ell_i (x_i - \bar{x})^2$$

(2.16) の第二項の平均値は $1/n$ オーダーとなり考えなくてよい。

従つて $(n-1)S_x^2$ で割つて $\bar{y}_{er}$ の偏倚の初項は $1/n$ オーダーの項

$$\text{に対して} \quad -\frac{(N-n)}{(n-1)NS_x^2} \left\{ \frac{\sum l_i (x_i - \bar{x})^2}{(N-1)} \right\} \quad (7.17)$$

括弧内の表示は l_i と $(x_i - \bar{x})^2$ の母共分散である；それは x_i への y_i の二次回帰からの貢献を表わす、そしてもしも y_i と x_i の関係が直線であればなくなる。 $\bar{y}_{er}$ の偏倚は $1/n$ オーダーであり、一方その標準誤差は、 $1/\sqrt{n}$ オーダーであるから、偏倚は大標本では無視出来る。

§ 7.5 比推定と単位当たり平均値との比較

Comparison with the ratio estimate and the mean per unit.

これ等の比較に対して標本の大きさ n は比及び回帰推定の分散に対する近似公式が妥当である程充分大でなければならぬ。推定母平均 $\bar{y}$ に対する三つの比較出来る分散は次の如くである。

$$V(\bar{y}_{er}) = \frac{(N-n)}{Nn} S_y^2 (1 - \rho^2) \quad (\text{回帰})$$

$$V(\bar{y}_R) = \frac{(N-n)}{Nn} (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x) \quad (\text{比})$$

$$V(\bar{y}) = \frac{(N-n)}{Nn} S_y^2 \quad (\text{単位当たり平均値})$$

回帰推定の分散は二つの分散が等しい場合の $\rho = 0$ でなければ単位当たり平均値の分散より小さい事は明らかである。

回帰推定の分散は、もしも

$$-\rho^2 S_y^2 < R^2 S_x^2 - 2R\rho S_y S_x \quad (7.18)$$

であれば、比推定の分散より小さい。

これは不等式 $(\rho S_y - R S_x)^2 > 0$ と同じである。

従つて回帰推定は $R = \bar{y}/\bar{x}$ であるから

$$\rho = \frac{R S_x}{S_y} = \frac{x_i \text{ の変動係数}}{y_i \text{ の変動係数}} \quad (7.19)$$

でなければ比推定よりも正確である。

方程式 (7.19) は y_i と x_i の関係が原点を通る直線であればい

も成立する。そしてこの様な場合には回帰と比推定は正確度が同じである。回帰推定か比推定が最良な偏推定量である時でさえも比推定と同じ正確度であるという事は興味ある事である。

実際に、比推定は直線回帰推定の特別な場合である。もしも回帰係数を $b' = \bar{y}/\bar{x}$ にとれば、考えられている値はもしも原点を通る直線であれば用いられ、

$$\bar{y}_{er} = \bar{y} + \frac{\bar{y}}{\bar{x}} (\bar{x} - \bar{x}) = \frac{\bar{y}}{\bar{x}} \bar{x} = \bar{y}_R$$

回帰推定は特に b を計算する労力のため比推定よりも計算が面倒である。大標本では、非有効推定量 b' はもしも比の方法が時間の節約になれば用いる事が出来る。§ 7.3 では最小自乗の抽出誤差からの $V(\bar{y}_{er})$ への貢献は分散の主成分の約 $1/n$ になる事を指摘した。従つて有効に資料の半分を用いる即ち 50% 効率での推定量 b' は $V(\bar{y}_{er})$ は

$$S_y^2 (1 - \rho^2) (1 + \frac{1}{n})$$

から $S_y^2 (1 - \rho^2) (1 + \frac{2}{n})$ に増大する

もしも n が大であれば、此の増大は些細なものとなる。従つて我々は資料の副次標本から R 及び $S_y^2 S_x$ を推定する。もしも真の回帰が直線であるとの良好な証拠があれば副次標本はいわば標本の b 分の 1 か又は 1 分の 1 位の単位からなる。もしも回帰が直線であるかどうか疑わしい場合は副次標本は無作為標本か又は本質的にこれに等しいものでなければならぬ。

時には以前の経験から b の値を推測するも可能である。 b の一定の値を例えば b^* に対してそれは標本の結果にはよらないものであるが、 $\bar{y}_{er}$ はくりかえしの抽出では

$$E \bar{y}_{er} = \bar{y} + b^* E(\bar{x} - \bar{x}) = \bar{y}$$

であるから不偏推定量である。

(例)

単位無作為標本からの回帰比及び単位当たり平均値の正確度は

§6.13 に述べた梨の果樹園の完全調査で集めた資料を用いて比較出来る。この例では y_i は果樹園で推定された梨の生産高で、 x_i は果樹園での梨の本数である。我々は 100 の果樹園の標本から推定されたものと 256 の果樹園の全生産高の推定量を比較する。標本は $\bar{y}$ 及び $\bar{x}$ の C_v が共に 10% より幾分上まわっているから分散公式を適用するのに充分の大きさであるか、どうかは疑わしいけれども、例は計算を示すのに役立つから用いてみよう。

基礎資料は次の如くである。

$$S_y^2 = 6409 \quad ; \quad S_{yx} = 4434 \quad ; \quad S_x^2 = 3898$$

$$R = 1.270 \quad ; \quad \rho = 0.887 \quad ; \quad n = 100 \quad ; \quad N = 256$$

$$V(\hat{y}_{er}) = \frac{N(N-n)}{n} S_y^2 (1-\rho)^2$$

$$= \frac{(256)(156)}{100} (6409)(1-0.887)^2 = 545,000$$

$$V(\hat{y}_R) = \frac{N(N-n)}{n} (S_y^2 + R^2 S_x^2 - 2RS_{yx})$$

$$= \frac{(256)(156)}{100} \{ 6409 + (1.613)(3898) - 2(1.270)(4434) \}$$

$$= 873,000$$

$$V(\hat{y}) = \frac{N(N-n)}{n} S_y^2 = 2,559,000$$

分散の性質から期待される様に、回帰推定と比較推定の向には殆んど差はない。共に単位当り平均値の場合より非常にすぐれている。

§7.6 層化抽出に於ける回帰推定

The regression estimate in stratified sampling

比推定の場合と同様、層化無作為抽出に適用出来る回帰推定には二つの型がある。第一の推定量 $\bar{y}_{ers}$ (S は個々のものの数) に対しては、我々は各層で個別の回帰係数を計算する。この推定量は次の形の数学的モデルに用いられる。

$$y_{hi} = \bar{y}_h + B_h (x_{hi} - \bar{x}_h) + e_{hi} \quad (7.20)$$

此処に通常の如く h は層を表わし、 i は層内の観測値を表わす。そして我々は B_h は層毎に変化すると信じている。

まず第一に各層平均の回帰推定量を計算する。即ち

$$\bar{y}_{h+k} = \bar{y}_h + B_h (\bar{x}_h - \bar{x}_h)$$

此処に B_h は B_h の最小自乗推定量である、従って

$$\bar{y}_{ers} = \frac{\sum_{h=1}^N N_h \bar{y}_{h+k}}{N} \quad (7.21)$$

単純無作為抽出で行われた二つの研究に於いて $\bar{y}_{ers}$ に対する抽出理論にも二つの研究がある。一方我々は各層に於ける母集団の大きさは無限で各層内で回帰は真に直線であると仮定すると標準の最小自乗理論の結果が適用される。これ等の仮定はある応用(例へば農事の抽出に於ては)ではあまり現実的なものではない。一方無限母集団又は直線回帰を仮定しない。(§7.2 の様式) 大標本理論がある。両方の結果は通常起るので $V(\bar{y}_{ers})$ に対する理論の二つの証明を与えよう。基礎理論を最初に示そう。

定理 7.3a

各層は無限と考え。

$$y_{hij} = \bar{y}_h + B_h (x_{hi} - \bar{x}_h) + e_{hij}$$

としよう。此処に一定 x_{hi} に対して

$$E(e_{hij}) = 0 \quad E(e_{hij}^2) = S_{eh}^2$$

そのとき層化無作為抽出では $\bar{y}_{ers}$ は

$$V(\bar{y}_{ers}) = \sum_{h=1}^N \left(\frac{N_h}{N} \right)^2 S_{yh}^2 (1-\rho_h^2) \left\{ \frac{1}{N_h} + \frac{(\bar{x}_h - \bar{x})^2}{\sum (x_{hi} - \bar{x}_h)^2} \right\} \quad (7.22)$$

なる分数をもつ $\bar{y}$ の不偏推定量である。

(証明) 層 h に定理 7.1c を適用すると $\bar{y}_{h+k}$ は

$$V(\bar{y}_{h+k}) = S_{yh}^2 (1-\rho_h^2) \left\{ \frac{1}{N_h} + \frac{(\bar{x}_h - \bar{x})^2}{\sum (x_{hi} - \bar{x}_h)^2} \right\}$$

の分散をもつ $\bar{y}_h$ の不偏推定量である事が分る

$$\bar{y}_{ers} = \sum_{h=1}^N \frac{N_h}{N} \bar{y}_{h+k}$$

であるから $\bar{y}_{ers}$ は $\bar{y}$ の不偏推定量であり、その分散は (7.22) に

与えたとおりである。

定理 7.3

もしも各層で標本が非常に大であれば

$$V(\bar{y}_{ers}) = \sum_{h=1}^L \frac{N_h(N_h - n_h)}{N^2 n_h} S_{y_h}^2 (1 - f_h^2)$$

(証明)

この証明に於ては、我々は直線回帰の存在を仮定しなり。§7.2と同様次のものを定義する。

$$B_h = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2}$$

同様に残差変量 e_{hi} は次の方程式により定義される。

$$y_{hi} = \bar{y}_h + B_h(x_{hi} - \bar{x}_h) + e_{hi}$$

§7.2の結果が今は $\bar{y}_{erh}$ に適用される。 $\bar{y}_{erh}$ の偏倚は $1/n_h$ オーダーで、その分散は近似的に

$$V(\bar{y}_{erh}) \doteq \frac{(N_h - n_h)}{N_h n_h} S_{y_h}^2 (1 - f_h^2)$$

従つて $\bar{y}_{ers}$ の定義によりその偏倚は殆んど $1/n_h$ オーダーである。此処に n_h は N_h の最小のものである。違つた層での抽出は独立であるから

$$V(\bar{y}_{ers}) = \sum_{h=1}^L \left(\frac{N_h}{N} \right)^2 V(\bar{y}_{erh}) \doteq \sum_{h=1}^L \frac{N_h(N_h - n_h)}{N^2 n_h} S_{y_h}^2 (1 - f_h^2)$$

系。

もしも各層に於て標本が大であれば実際に不偏である $V(\bar{y}_{ers})$ の

推定量は
$$v(\bar{y}_{ers}) = \sum_{h=1}^L \frac{N_h(N_h - n_h)}{N^2 n_h} s_{y_h}^2 \quad (7.23)$$

此処に
$$s_{y_h}^2 = \frac{1}{(n_h - 1)} \left\{ \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 - b_h^2 \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2 \right\}$$

この結果は亦 §7.2 の論議から導かれる。

推定量 $\bar{y}_{ers}$ はそれに対する比推定と同様困難な事がある。そ

れに於ける偏倚は各層で同じ符号である。もしも層の数が大であれば、その標準誤差に対する $\bar{y}_{ers}$ の偏倚の比はかなりのものになる。§7.4 に示した様に偏倚に於ける最初の項は x_{hi} への y_{hi} の曲線回帰から生ずるから、この危険性は変量間の関係が直線型よりもむしろ曲線型に近づくときにひどくなる。

§7.7 組合せ回帰推定

The Combined regression estimate

第二の推定量 $\bar{y}_{erc}$ (組合せに対して) は B_h がすべての層で同じであると仮定されるときに用いられる。そのときモデルは

$$y_{hi} = \bar{y} + B(x_{hi} - \bar{x}) + e_{hi}$$

$\bar{y}_{erc}$ を見出すため、先ず次のものを計算する。

$$\bar{y}_{st} = \frac{\sum N_h \bar{y}_h}{N} \quad ; \quad \bar{x}_{st} = \frac{\sum N_h \bar{x}_h}{N}$$

此等は簡化抽出に用いる通常の推定量である。そのとき

$$\bar{y}_{erc} = \bar{y}_{st} + B(\bar{x} - \bar{x}_{st}) \quad (7.24)$$

B に対しては習慣的な組合せ推定量

$$b = \frac{\sum_{h=1}^L \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_{h=1}^L \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2} \quad (7.25)$$

をとると、しばしば満足である。

此れは一般に B の最も正確な推定量ではない。層 h に於ける推定量 b_h の分散

$$V(b_h) = \frac{S_{y \cdot x_h}^2}{\sum_i (x_{hi} - \bar{x}_h)^2}$$

此処に $S_{y \cdot x_h}^2 = S_{y_h}^2 (1 - f_h^2)$ 理論的に B の最も正確な推定量はその分散の逆数を各 b_h に重みづける事により得られる。

此れは次のものになる事分かる。

$$b_{opt} = \frac{\sum \sum g_h (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum \sum g_h (x_{hi} - \bar{x}_h)^2}$$

此処に $f_h = 1/S_{y_{xh}}^2$ この推定量は、もしも残差分散がすべての層で同じであれば、そのときのみ (7.25) となる。実際には b_{opt} は用いられない、何故ならば我々はこれ等の推定量に於ける正確度の合成損失で以て $S_{y_{xh}}^2$ の標本推定量を求めなければならぬからである。これ等の誤差は層内の標本が大きい場合は小さいが、その場合に b の抽出誤差は $V(\bar{y}_{enc})$ に殆んど影響を及ぼさない。従って B の習慣的な組合せ推定量に対する改善は全標本の大きさから例えば 50 以下である様な場合及び、別な層での残差分散内に大きな差がある様な場合以外はいさくなるであろう。

$V(\bar{y}_{enc})$ の結果の基礎的形を表現する際に我々は標本回帰係数例えば b' は、その荷重が x_{hi} のみにする b_h のある荷重平均であると仮定しなければならぬ。この様な函数は特別な場合として共に習慣的な組合せ、 b 及び b_{opt} を含み、 $V(\bar{y}_{enc})$ をかなり一般的にのべる事を可能にする。

定理 7.4a

各層は無限と考え。

$$y_{hij} = \bar{y}_h + B(x_{hi} - \bar{x}_h) + e_{hij} \quad (7.26)$$

と仮定しよう、此処に一定 x_{hi} に対して

$$E(e_{hij}) = 0 \quad ; \quad E(e_{hij}^2) = S_{eh}$$

そのとき、層化無作為抽出では、推定量

$$\bar{y}_{enc} = \bar{y}_{st} + b'(\bar{x} + \bar{x}_{st})$$

は

$$V(\bar{y}_{enc}) = \sum_{h=1}^L \left(\frac{N_h}{N} \right)^2 \frac{1}{n_h} S_{yh}^2 (1 - P_h^2) + (\bar{x}_{st} - \bar{x})^2 V(b') \quad (7.27)$$

なる分散をもつ $\bar{y}$ の不偏推量である。

(証明)

(7.26) から

$$\bar{y}_{st} = \bar{y} + B(\bar{x}_{st} - \bar{x}) + \bar{e}_{st}$$

従つて推定量の誤差

$$\begin{aligned} \bar{y}_{enc} - \bar{y} &= \bar{y}_{st} + b'(\bar{x} - \bar{x}_{st}) - \bar{y} \\ &= \bar{e}_{st} + (b' - B)(\bar{x} - \bar{x}_{st}) \end{aligned}$$

上述の条件の下では通常の如く b' は B の不偏推定量であるから $\bar{y}_{enc}$ は不偏である。更に $\bar{e}_{st}$ と b' の共分散は 0 である事が示される。従つて

$$\begin{aligned} V(\bar{y}_{enc}) &= V(\bar{e}_{st}) + (\bar{x}_{st} - \bar{x})^2 V(b') \\ &= \sum_{h=1}^L \left(\frac{N_h}{N} \right)^2 \frac{1}{n_h} S_{yh}^2 (1 - P_h^2) + (\bar{x}_{st} - \bar{x})^2 V(b') \end{aligned}$$

系 1

$$\text{もしも } b' = b = \frac{\sum_h \sum_i (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_h \sum_i (x_{hi} - \bar{x}_h)^2}$$

とすれば

$$V(\bar{y}_{enc}) = \sum_{h=1}^L S_{yh}^2 (1 - P_h^2) \left\{ \left(\frac{1}{N} \right)^2 \frac{1}{n_h} + (\bar{x}_{st} - \bar{x})^2 \frac{\sum_h \sum_i (x_{hi} - \bar{x}_h)^2}{\left[\sum_h \sum_i (x_{hi} - \bar{x}_h)^2 \right]^2} \right\} \quad (7.28)$$

これを証明するため

$$b = \frac{\sum_h b_h \sum_i (x_{hi} - \bar{x}_h)^2}{\sum_h \sum_i (x_{hi} - \bar{x}_h)^2}$$

亦

$$V(b_h) = \frac{P_{yh}^2 (1 - P_h^2)}{\sum_i (x_{hi} - \bar{x}_h)^2}$$

直線回帰に於ける通常の公式を適用して結果を得る。

系 2.

採用される割当ての型に応じてこの結果に色々特別な型がある
例えばもしも $S_{y_h}^2 (1 - \rho_h^2)$ がすべての層で一定で比例割当てを用
いれる。公式 (7.28) は

$$V(\bar{y}_{erc}) = S_{y_h}^2 (1 - \rho_h^2) \left\{ \frac{1}{n} + \frac{(\bar{x}_{st} - \bar{x})^2}{\sum_h \sum_i (x_{hi} - \bar{x}_h)^2} \right\} \quad (7.29)$$

単純無作為抽出では、 $V(\bar{y}_{erc})$ への b の抽出誤差の影響は全分散の
ほぼ $1/n$ 倍である事が分った。

不幸にもこの結果は必ずしも $V(\bar{y}_{erc})$ にはあてはまらない。方
程式 (7.29) に対しては結果は妥当であるが、より一般的な表示
(7.28) に於てはしはしは単一層例えば n 層から分散への主な貢
献は生ずる旅な事がある。今示される公式 (7.28) の検証はこの
立場で $V(b)$ からの貢献は全分散の $1/n_h$ 倍である事を示して
いる。適当な大きさの標本では従って $V(b)$ の貢献がそれをすこ
る前に無視出来るという事をチェックする事が望ましい。

直線回帰の仮定がゆるめられた場合の $\bar{y}_{erc}$ に対するより一般的
な理論は全く複雑となる。一般的な方法で何うなるかを示す事を
充分行わねばならない。層内回帰係数は次の様に定義される。

$$B_h = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_{i=1}^{N_h} (x_{hi} - \bar{x}_h)^2}$$

残差変量 e_{hi} は次の方程式により定義される。

$$y_{hi} = \bar{y}_h + B_h (x_{hi} - \bar{x}_h) + e_{hi}$$

従ってこの方程式から誘導して

$$\bar{y}_h = \bar{y}_h + B_h (\bar{x}_h - \bar{x}_h) + \bar{e}_h$$

$$\bar{y}_{st} = \bar{y} + \sum W_h B_h (\bar{x}_h - \bar{x}) + \bar{e}_{st} \quad (7.30)$$

此処に $W_h = N_h/N$

$$\text{さて、} \bar{y}_{erc} = \bar{y}_{st} + b'(\bar{x} - \bar{x}_{st})$$

(7.30) からの $\bar{y}_{st}$ に対して代入せよ。従って推定量の誤差は

$$\bar{y}_{erc} - \bar{y} = \sum W_h B_h (\bar{x}_h - \bar{x}) + b'(\bar{x} - \bar{x}_{st}) + \bar{e}_{st}$$

この点で、記号 $B' = E(b')$ を導入すると便利である。

前の表示は次の様にかける。

$$\begin{aligned} \bar{y}_{erc} - \bar{y} &= \sum W_h B_h (\bar{x}_h - \bar{x}) - B'(\bar{x}_{st} - \bar{x}) \\ &\quad - (b' - B')(\bar{x}_{st} - \bar{x}) + \bar{e}_{st} \\ &= \bar{e}_{st} + \sum W_h (\bar{x}_h - \bar{x})(B_h - B') - (b' - B')(\bar{x}_{st} - \bar{x}) \end{aligned}$$

之れには $\sum W_h \bar{x}_h = \bar{x}_{st}$ 及び $\sum W_h \bar{x}_h = \bar{x}$ を用いた。

この分析は推定量の誤差を三つの成分に分ける。第一の成分は
我々が良く知っている項 $\bar{e}_{st}$ である。第二の成分は真の層内回帰
係数 B_h に於ける層毎の変動から生じたものである。もしも B_h
がすべて B に等しければ、この項は b' がこの B の不偏推定量であ
ると仮定するとなくなる。 $E(\bar{x}_h) = \bar{x}_h$ であるから、この項は、
 $\bar{y}_{erc}$ に偏倚を等かなくて $\bar{y}_{erc}$ の分散に役立つ。第三の成分は

第三番目の項は b' の抽出誤差の貢献を表わす。単純無作為抽
出と同様に、この項の平均値は回帰が真に直線でない限りでは
ない。偏倚の初項は $2n$ への y への曲線回帰から生ずる。もしも変動性が
近似的にすべての層で同一であり比例割当てが用いられるとすれ
ば、偏倚は $1/n$ オーダーであるがもしも一つの層からの貢献が卓越
してくると大となる。同じ事は $V(b')$ からの貢献にも適用される。

もしも b' の項から生ずる偏倚及び変動が無視出来る時は $V(\bar{y}_{erc})$
の主要部分は

$$V(\bar{y}_{erc}) = \sum_h \frac{N_h(N_h - n_h)}{N^2 n_h} \{ S_{y_h}^2 (1 - \rho_h^2) + S_{e_h}^2 (B_h - B')^2 \}$$

§ 7.8 推 定 分 散

Estimated variances

$V(\bar{y}_{erc})$ に於て我々は $S_{y \cdot x}^2 = S_{y_h}^2 (1 - \rho^2)$ に対して標本推定量

$$\begin{aligned} s_{y \cdot x}^2 &= \frac{1}{(n_h - 2)} \sum_i \{ (y_{hi} - \bar{y}_h) - b_h (x_{hi} - \bar{x}) \}^2 \\ &= \frac{1}{(n_h - 2)} \left\{ \sum_i (y_{hi} - \bar{y}_h)^2 - \frac{\{ \sum_i (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}) \}^2}{\sum_i (x_{hi} - \bar{x})^2} \right\} \end{aligned}$$

を代入する。

組合せ推定量 $s_{y \cdot x_R}^2$ は次の様にとる。

$$s_{y \cdot x_R}^2 = \frac{1}{(n_R - 1)} \sum_i \{ (y_{Ri} - \bar{y}_R) - b(x_{Ri} - \bar{x}_R) \}^2$$

分母は $(n_R - 2)$ の代り $(n_R - 1)$ となる。何故ならば通常の b がすべての層で用いられるからである。

[直観的な近似として、分母は $(n_R - 1 - \frac{1}{L})$ がよりよい様に見える]

偏差の計算をさけるには $s_{y \cdot x_R}^2$ の計算は次の様になる。

$$\sum_i (y_{Ri} - \bar{y}_R)^2 - 2b \sum_i (y_{Ri} - \bar{y}_R)(x_{Ri} - \bar{x}_R) + b^2 \sum_i (x_{Ri} - \bar{x}_R)^2$$

$s_{y \cdot x_R}^2$ は層が変わっても変動しないという良い証拠がなければ一請に加えるよりはむしろ夫々の層で $s_{y \cdot x_R}^2$ の個々の値を求めた方がよい。

8.7.9 回帰推定の二つの型の比較

Comparison of the two types of regression estimate.

しっかりした規則は別個の推定量又は組合せ推定量の何れがある特殊な立場ですぐれているかをきめるのに作る事は出来ない。即ち判断を試みるのがその選択に必要となる別個の推定量の欠点は標本が個々の層内で小さいときには偏倚がかちであるとその分散は回帰係数に於ける抽出誤差から大きな貢献をうけるという事である。組合せ推定量の欠点はその分散はむしろ母回帰係数が層毎に変れば大きくなるという事である。もしも我々が回帰は直線で判断出来る限り B_R がすべての層で同じと思えるならば組合せ推定量がすぐれている。もしも習慣的な組合せ推定量 b が B に対して用いられると、 $V(\bar{y}_{HVR})$ の標本推定量は公式(7.24)に量 $s_{y \cdot x_R}^2$ を代入して得られる。

もしも回帰が直線である様で(従つて偏倚の危険性が小さい)あるが B_R が層毎に異なる様であれば別個な推定量がより、その分散の推定量は公式(7.22)に $s_{y \cdot x_R}^2$ の値を代入して得られる。もしも直線回帰推定が用いられる時回帰が曲線性をもつておれば、組合せ推定量は恐らく標本がすべての層で大きくなければより安全であろう。

第8章 組織的抽出法

Systematic sampling

8.1 緒言

Description

この抽出方法は一見してまづ単純無作為抽出とは全く違っている。母集団での N 個の単位があるオーダーで N から N 迄数えられると仮定しよう。 n 個の標本を選ぶには最初の n 個の単位から無作為に一つの単位をとり、次々に n 番目の単位をとる。例えば、もしも n が 15 で最初に抽かれた単位が 13 であれば、次々に抽出される単位は 28, 43, 58 等である。最初の単位の選択は全標本を決定する。この種の標本を n 番目毎の組織標本と呼ぶ。

単純無作為抽出よりもすぐれている利益は次の如くである。

1 標本の抽出が容易であり、しばしば誤りなしに実行する事がより容易である事。

これは抽出を野外で行ふ時に特に利益である。抽出が事務所内で行われる時でさえ大いに時間の節約となる。例えばもしも単位をすべて同じ大きさのカードに書き、それを積み重ねるとすれば、カードを積み重ねに沿つて尺で測定した各インチ毎に抽く事が出来る。この操作は迅速で、一方単純無作為抽出は遅い。勿論この方法は厳密な n 番目毎の抽出の規則とは違つたものであるが。

2 直観的に、組織抽出は単純無作為抽出よりも正確である様に見える。実際、それは最初が n 個の単位 n 番目毎 n 個の単位等から成立つてゐる n 個の層に母集団を層化してゐる。従つて組織標本は層当り一つの単位を n つそれに対応する層化無作為標本と殆んど同じ正確度が期待出来る。その相違は組織標本では層で相対的に同じ位置に標本がとられ、一方層化無作為標本では層での位置は各層内での無作為化により別々に決定される(よ/図参照)組織標本は母集団全体に一様にばらまかれている、そしてこの事實は時に組織抽出が層化無作為抽出よりもかなり正確度を大にするものである。

数個の公式が組織標本の平均値 $\bar{y}_{sg}$ の分散に対して発展してきた。下に与える最初の三つのものは集落がすべて n 個の要素を含み、標本が一つの集落からなつてゐるある種の集落抽出に適用される。

もしも $N = n$ であるならば、 $\bar{y}_{sg}$ は無作為におかれた組織標本に対する $\bar{Y}$ の不偏推定量である事は証明出来る。もしも $N \neq n$ ならば、この結果は偏倚は大きくて重要でない旅であるが成立しない。この偏倚はある標本に選ばれる確率を割当てる事により避ける事が出来る。

8.1 表の例を考えよう。もしも最初の三つの標本の各々に選ばれる確率を $5/23$ と与えて最後の二つの各々に $4/23$ の確率をあたえんとすれば標本平均は不偏である。

次の分析に於て、記号 y_{ij} は i 番目の組織標本の j 番目の数を表はす。そうすると、 $j = 1, 2, \dots, n$ $i = 1, 2, \dots, k$ である。

i 番目の標本の平均は $\bar{y}_{i\cdot}$ で表わす。

定理 8.1

組織標本の平均値の分散は

$$V(\bar{y}_{sg}) = \frac{(N-1)}{N} S^2 - \frac{k(n-1)}{N} S_{wsy}^2 \quad (8.1)$$

此処に
$$S_{wsy}^2 = \frac{1}{k(n-1)} \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{i\cdot})^2$$

は同じ組織標本内にある単位間分散である。この分散の分母 $k(n-1)$ は通常の分散分析の規則で作られる。即ち k 個の標本の各々は計算に於ける平方和に $(n-1)$ 個の自由度をもつてゐる。

(証明)

通常の分散分析の等号により

$$\begin{aligned} (N-1) S^2 &= \sum_i \sum_j (y_{ij} - \bar{Y})^2 \\ &= n \sum_i (\bar{y}_{i\cdot} - \bar{Y})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 \end{aligned}$$

併し $\bar{y}_{sg}$ の分散は定義により

$$V(\bar{y}_{sg}) = \frac{1}{k} \sum_{i=1}^k (\bar{y}_{i\cdot} - \bar{Y})^2$$

$$\text{依つて } (N-1) S^2 = n k V(\bar{y}_{sg}) + k(n-1) S_{wsy}^2$$

証明終り

系

組織標本の平均値は、もしも次の様であればそのときのみ単経無作為標本の平均値より正確である。即ち

$$S_{wsy}^2 > S^2 \quad (8.2)$$

(証明)

もしも $\bar{y}_{sg}$ が大さななる単経無作為標本の平均値であれば

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

方程式 (8.1) から、もしも

$$\frac{(N-1)}{N} S^2 - \frac{k(n-1)}{N} S_{wsy}^2 < \frac{(N-n)}{N} \frac{S^2}{n}$$

即ち

$$k(n-1) S_{wsy}^2 > \left\{ N-1 - \frac{N-n}{n} \right\} S^2 = k(n-1) S^2$$

であれば、そのときのみ $V(\bar{y}_{sg}) < V(\bar{y})$ である。

一般の集落抽出に適用するこの重要な結果は、もしも組織標本内分散が全体としての母分散より大であれば、組織抽出は単経無作為抽出よりも正確である事を示してゐる。組織抽出は同じ標本内の単位が不斉一であるとき正確であり、斉一であるときは不正確である。直観的にその結果は明らかである。もしも母集団の変動に比して組織標本内の変動が小さければ標本に於ける次々の単位は多少共同情報を含んでゐる。

他の分散の型を定理 8.2 で述べる

定理 8.2

$$V(\bar{y}_{sy}) = \frac{S^2}{n} \left[\frac{(N-1)}{N} + (n-1) \rho_w \right] \dots \dots \dots (8.3)$$

此処に

$$\rho_w = \frac{2}{h n (n-1) S^2} \sum_{i=1}^h \sum_{j < u} (y_{ij} - \bar{y})(y_{iu} - \bar{y})$$

この量は同じ組織標本にある単位間の相関係数である。分母の因子 $h n (n-1) / 2$ は積和に於ける明確な項の数である。

(証明)

$$\begin{aligned} n^2 h V(\bar{y}_{sy}) &= n^2 \sum_{i=1}^h (\bar{y}_{i.} - \bar{y})^2 \\ &= \sum_{i=1}^h \left[(y_{i1} - \bar{y}) + (y_{i2} - \bar{y}) + \dots + (y_{in} - \bar{y}) \right]^2 \end{aligned}$$

平方和は $\bar{y}$ からの偏差の全平方和即ち $(N-1) S^2$ となる。これは

$$\begin{aligned} n^2 h V(\bar{y}_{sy}) &= (N-1) S^2 + 2 \sum_{i < j} \sum_{j < u} (y_{ij} - \bar{y})(y_{iu} - \bar{y}) \\ &= (N-1) S^2 + h n (n-1) S^2 \rho_w \end{aligned}$$

従つて

$$V(\bar{y}_{sy}) = \frac{S^2}{n} \left[\frac{(N-1)}{N} + (n-1) \rho_w \right]$$

この結果は同じ標本に於ける単位間の正の相関があれば同じ標本平均の分散を高める事を示す。正の相関が小さくても乗数は $(n-1)$ であるから影響が大である。

今述べて二つの定理は S^2 の項で $V(\bar{y}_{sy})$ を表わしたものであるから、それは単純無作為標本に対する分散に関連してゐる。層が最初 h 単位次が h 単位等から成立つてゐる層化無作為標本に対しての分散の項で $V(\bar{y}_{sy})$ を表わす定理 8.2 の類似物である。我々の記号に於て y_{ij} の j を層を示すものとする。層平均は $\bar{y}_{.j}$ とかける。

定理 8.3

$$V(\bar{y}_{sy}) = \frac{S_{usy}^2}{n} \left[\frac{(N-n)}{N} + (n-1) \rho_{ust} \right] \quad (8.4)$$

$$\text{此処に } S_{usy}^2 = \frac{1}{n(h-1)} \sum_{j=1}^n \sum_{i=1}^h (y_{ij} - \bar{y}_{.j})^2$$

これは同じ層にある単位間の分散である。分母 $n(h-1)$ は h 個の層の各々が $(h-1)$ の自由度をもつてゐるから用いられる。更に

$$\rho_{ust} = \frac{2}{h n (n-1)} \sum_{i < j} \sum_{j < u} (y_{ij} - \bar{y}_{.j})(y_{iu} - \bar{y}_{.u}) / S_{usc}^2 \quad (8.5)$$

このむしろ複雑な量は同じ組織標本にある項の対の層平均からの偏差間の相関係数である。

証明は定理 8.2 と同様である。

系

組織標本は、もし $\rho_{ust} = 0$ であれば層当り $1/h$ の単位をもつ、それに応ずる層化無作為標本と同じ正確度をもつ。これは層化無作為標本のこの型に対する $V(\bar{y}_{st})$ は

$$V(\bar{y}_{st}) = \left(\frac{N-n}{N} \right) \frac{S_{ust}^2}{n}$$

であるからである。(定理 8.3, 系 2)

自己相関母集団 (autocorrelated population) に用いる。

$V(\bar{y}_{sy})$ に対する公式は、W.G. & L.H. madow (1944) により与えられた。彼等は組織標本抽出の正確度の最初の理論的研究をした。

(例)

おこ長の資料は h なり一定して増大する傾向を表わす小さい h 工母集団に対するものである。 $N=40$, $h=10$, $n=4$ である。各欄は組織標本を表わす。そして列は層を表わす。

例は層内相関係数が正である場合を示す。例えば α の標本では 4 つの数 0, 6, 18, 26 の各々はそれに属する層の平均以下である。

これは二三の例外を除いて最初の五つの組織標本ではすべて真で

ある。最後の五つの標本では、層平均からの偏差は大抵の場合正である。従つて f_{wst} に於ける積の項は大抵の場合正である。定種およびそれら残々は層あたり一つの単位をもつ層化無作為抽出よりも正確度が劣る事を組織抽出に期待出来る。

8.3表 $n=4$, $N=40$ なる10の組織標本に対する資料

層	組織標本番号										層平均
	1	2	3	4	5	6	7	8	9	10	
I	0	1	1	2	5	4	7	7	8	6	4.1
II	6	8	9	10	13	12	15	16	16	17	12.2
III	18	19	20	20	24	23	25	28	29	27	23.3
IV	26	30	31	31	33	32	35	37	38	38	33.1
合計	50	58	61	63	75	71	82	88	91	88	72.7

分散 $V(\bar{y}_{st})$ は組織標本合計から次の如く直接に分る

$$V(\bar{y}_{st}) = V_{st} = \frac{1}{n} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 = \frac{1}{nN} \sum_{i=1}^n (n\bar{y}_i - n\bar{y})^2$$

$$= \frac{1}{160} \left[(50)^2 + (58)^2 + \dots + (88)^2 - \frac{(727)^2}{10} \right] = 11.63$$

無作為及び層化無作為抽出に対しては“列間”及び“列内”に母集団分散の分析が必要である。これを8.4表に示す

8.4表 分散分析

	d.f.	SS	m.s.
列間(層)	3	4828.3	
層内	36	485.5	$13.49 = f_{wst}^2$
合計	39	5313.8	$136.25 = S^2$

従つて単純無作為及び層化無作為標本からの推定平均の分散は次の通りである。

$$V_{ran} = \left(\frac{N-n}{N} \right) \frac{S^2}{n} = \frac{9}{40} \cdot \frac{136.25}{4} = 30.66$$

$$V_{st} = \left(\frac{N-n}{N} \right) \frac{f_{wst}^2}{n} = \frac{9}{40} \cdot \frac{13.49}{4} = 3.04$$

層化無作為抽出及び組織抽出共に単純無作為抽出よりも有効であるが、予期した通り組織抽出は層化無作為抽出よりも正確度が劣る。

8.5表は同じ資料で二番目と四番目の層で観測値を逆にした順序を示す。これは f_{wst} は著しく異なる影響がある。何故ならば同じ組織標本内にある観測値の対に対して層平均からの偏差間の積の多くは負となるからである。

8.5表 層II及びIVの観測値を逆に並べた8.3表の資料

層	組織標本番号										層平均
	1	2	3	4	5	6	7	8	9	10	
I	0	1	1	2	5	4	7	7	8	6	4.1
II	17	16	16	15	12	13	10	9	8	6	12.2
III	18	19	20	20	24	23	25	28	29	27	23.3
IV	38	38	37	35	32	33	31	31	30	26	33.1
合計	73	74	74	72	73	73	73	75	75	65	72.7

例えば最初の組織標本では層平均からの偏差は -4.1 , $+4.8$, -5.3 , $+4.9$ である。偏差の対の六つの積の内四つは負である。大抵つばに同じ様な事が各組織標本に適用される。

この変化は V_{ran} 及び V_{st} には影響はない。組織標本では、8.5表での組織標本合計を8.3表のものと比較したとき分る様に、正確度が著しく増してくる。

さて、

$$V_{st} = \frac{1}{160} \left[(73)^2 + (74)^2 + \dots + (65)^2 - \frac{(727)^2}{10} \right] = 0.46$$

時には層内で質の相関を作る様に単位を数える事により、この結果を利用する事が可能である。母集団内の傾向の正確な知識が必要である。併し、後で分る様に、 $\bar{y}_{sy}$ の標準誤差の良好な推定量を標本から得る事は非常に困難である場合である。

§ 8.4 組織的抽出と層化無作為抽出との比較

Comparison with systematic with stratified-random sampling :-

層化又は単純無作為抽出に比して組織抽出の実行には大いに母集団の性質が関係してくる。組織抽出は最も正確度の高い母集団があり又単純無作為抽出よりも正確度の劣る母集団もある。

ある母集団及び y_i のある値に対して $V(\bar{y}_{sy})$ はより大きな標本からとられた時でも良好な場合からのおどろくべきかけはなれを増す事がある。かくて組織抽出が推奨出来る場合についての一般的な忠告を与える事は困難である。母集団構造の知識はその最も有効な使用に対して必要である。

この問題についての研究の二つの線は次の様に示されてきた。まづ第一は、 y_i がそのある簡単な函数である人工的母集団についての色々な抽出法を比較する事である。その二は自然母集団について比較する事である。その両方の研究の型は劣力がけり、又抽出計画をする人仕事をしたけつたと仮定して未だ彼に望んでる程多方面に亘るものはない。主な結果のいくらかを次々の節で述べてみよう。

§ 8.5 ランダムオーダーである母集団

population in "random" order

組織抽出は、その便利さのために、単位の数に有効的に無作為である母集団にしばしば用いられる。これはもしも測定すべき項目が何人の姓名に無関係であるものであれば、姓名のアルファベット順に並べられたカードから標本抽出する様な事に於て然りである。そのとき我々がその繰り込みによつて仕事を進めるとき y_i は一定の傾向又は層化というものはないであらうし、隣りの

値との相関もないであらう。

この立場に於て我々は本質的に単純無作為抽出と同じで又同じ分散をもつてであらう事に期待出来る。いかなる単一の有限母集団に対しても、 y_i 及び y_j の与えられた値で以て、この事は全く真ではない。何故なら y_i の自由度のみに基く V_{sy} はむしろ y_i が小さいときに誤差が多く、 V_{ran} よりも大きくなるか又は小さくなるのかの何れかである。大体に於て二つの平均値が等しい事を示す二つの結果がある。その二つの結果を、それ等が組織抽出の研究に色々な糸口を示すから、に報告してみよう。

定理 8.4

$y_1, y_2, \dots, y_N$ なる数の組の $N!$ の順序により作られてゐるすべての $N!$ 有限母集団を考えてみよう。そのとき平均してこれ等の有限母集団では

$$E(V_{sy}) = V_{ran}$$

V_{ran} はすべての順列に対して同じである事に注意

WG 及び L. H. madow (1944) によつて証明されたこの結果はもしも特別な有限母集団での項目の配列順序(オーダー)が $N!$ 順列から無作為に抽出されると考えるとすれば組織抽出は大体に於て単純無作為抽出と同じである事を示す。

第二の糸口は、ある特性をもつ無限超母集団から無作為に抽出された有限母集団を考えることである。

証明される結果は、単一有限母集団(例えば $y_1, y_2, \dots, y_N$ の値の特別な組)に対しては適用されなくて無限母集団から抽出出来るすべての有限母集団の平均に適用出来る。この糸口は一見して実際の応用には関係がない様に思われるが、この考え方は誤まりである。

如何なる抽出方法も有限母集団の全シリーズに実際に用いられる。与えられた抽出方法が有効である有限母集団のクラスを述べる一つの方法は、この様な有限母集団が無作為に抽出された無限超母集団を述べる事である。

(182)

記号 Σ をして、この超母集団から抽出出来るすべての有限母集団についての平均を表わす

定理 8.5

もしも、変量 y_i ($i=1, 2, \dots, N$) が次の性質をもつ超母集団から無作為に抽出されるとすれば

即ちその性質とは

$\Sigma y_i = N\mu$; $\Sigma (y_i - \mu)(y_j - \mu) = 0$ ($i \neq j$) ; $\Sigma (y_i - \mu)^2 = \Sigma b_i^2$
そのとき

$$\Sigma V_{sq} = \Sigma V_{ran}$$

厳密な条件はすべての y_i は同じ平均値 μ をもち即ち一定の傾向はなく二つの違った点での y_i と y_j の値の間の直線的相関はないといふ事である。分散の b_i^2 はそのシリーズの真価に変化する。

(証明)

特別な有限母集団に対して

$$V_{ran} = \frac{(N-n)}{Nn} \cdot \frac{\Sigma_{i=1}^N (y_i - \bar{y})^2}{(N-1)}$$

さて

$$\begin{aligned} \Sigma_{i=1}^N (y_i - \bar{y})^2 &= \Sigma_{i=1}^N \{ (y_i - \mu) - (\bar{y} - \mu) \}^2 \\ &= \Sigma_{i=1}^N (y_i - \mu)^2 - N(\bar{y} - \mu)^2 \end{aligned}$$

y_i と y_j は相関もないから ($i \neq j$)

$$\Sigma (\bar{y} - \mu)^2 = \frac{1}{N^2} \Sigma_{i=1}^N b_i^2$$

従つて

$$\Sigma V_{ran} = \frac{(N-n)}{Nn(N-1)} \left\{ \Sigma_{i=1}^N b_i^2 - N \frac{\Sigma b_i^2}{N^2} \right\}$$

即ち

$$\Sigma V_{ran} = \frac{(N-n)}{N^2 n} \Sigma_{i=1}^N b_i^2$$

(183)

V_{sq} を考えて、 $\bar{y}_u$ をして、 u 番目組織標本の平均値を表わす
としよう。特別な有限母集団に対して

$$\begin{aligned} V_{sq} &= \frac{1}{K} \Sigma_{u=1}^K (\bar{y}_u - \bar{y})^2 \\ &= \frac{1}{K} \left\{ \Sigma_{u=1}^K (\bar{y}_u - \mu)^2 - K(\bar{y} - \mu)^2 \right\} \end{aligned}$$

無限母集団からの無相関標本の平均値の分散に対して、定理から

$$\begin{aligned} \Sigma V_{sq} &= \frac{1}{K} \left\{ \frac{\Sigma_{i=1}^N b_i^2}{n^2} - \frac{K \Sigma_{i=1}^N b_i^2}{K^2} \right\} \\ &= \frac{(N-n)}{N^2 n} \Sigma_{i=1}^N b_i^2 = \Sigma V_{ran} \end{aligned}$$

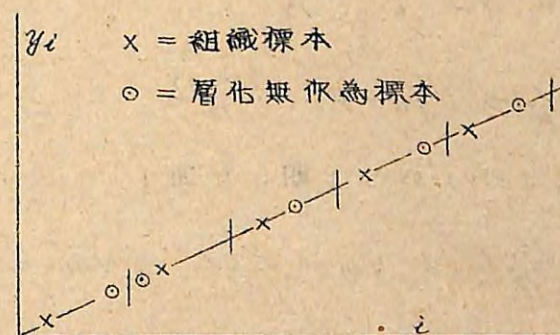
§ 8.6 直線的傾向をもつ母集団

populations with linear trend.

もしも、母集団が 8.2 図に示す様に単に直線的傾向のものからなつておれば、結果の性質を推測する事はかなり容易である。

8.2 図から、あたかも V_{sq} 及び V_{st} (層当り一つの単位) は V_{ran} よりも共に小さくなる様に見える。

8.2 図 直線的傾向をもつ母集団に於ける組織抽出



更に V_{sq} は V_{st} よりも大きい。何故ならば組織標本はもしも一つの層であまりに小さければすべての層であまりに小さくなりすぎ、一方層化無作為抽出は層内誤差を打消す様な機会があるからである。その効果を数学的に検定するため $y_i = i$ と仮定する。そのとき、

$$\sum_{i=1}^N i = \frac{N(N+1)}{2} \quad ; \quad \sum_{i=1}^N i^2 = \frac{N(N+1)(2N+1)}{6}$$

母分散 S^2 は次によつて与えられる

$$S^2 = \frac{1}{(N-1)} \{ \sum y_i^2 - N\bar{y}^2 \}$$

$$= \frac{1}{(N-1)} \left\{ \frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4} \right\} = \frac{N(N+1)}{12} \quad (8.6)$$

従つて単純無作為標本の平均値の分散は

$$V_{ran} = \frac{(N-n)}{N} \cdot \frac{S^2}{n} = \frac{n(k-1)}{n \cdot k} \cdot \frac{n \cdot k(N+1)}{12 \cdot n} = \frac{(k-1)(N+1)}{12} \quad (8.7)$$

層内分散 S^2_{iu} を見出すには (8.6) の N を k とおきかえるだけである。即ち

$$V_{st} = \frac{(N-n)}{N} \cdot \frac{S^2}{n} = \frac{n(k-1)}{n \cdot k} \cdot \frac{k(k+1)}{12 \cdot n} = \frac{(k^2-1)}{12 \cdot n} \quad (8.8)$$

組織標本に対して、 σ_1 の標本は σ_2 の標本よりその平均値は $1/n$ だけ大きいし σ_3 の標本の平均値は、 σ_1 の標本のそれより $1/n$ だけ大きい等である。従つて平均値 $\bar{y}_u$ は数 $1, 2, \dots, k$ でおきかえる事が出来る。従つて (8.6) を更に適用して

$$\sum_{i=1}^k (\bar{y}_u - \bar{y})^2 = \frac{k(k^2-1)}{12}$$

即ち

$$V_{sy} = \frac{1}{k} \sum (\bar{y}_u - \bar{y})^2 = \frac{(k^2-1)}{12} \quad (8.9)$$

公式 (8.7), (8.8), (8.9) から予期した通り

$$V_{st} = \frac{k^2-1}{12 \cdot n} \leq V_{sy} = \frac{k^2-1}{12} \leq V_{ran} = \frac{(k-1)(N+1)}{12}$$

$k=1$ のときのみすべてが正しい。従つて直線的傾向の影響を除くためには、それが疑わしくとも、疑わしくなくとも、組織標本は単純無作為標本よりも有効であるが、層化無作為標本よりも有効ではない。

§ 8.7 端末補正

End corrections

直線的傾向がある際に組織抽出をまづくやる事に対してこれを二三の方法で改善する事が出来る。

その一つは、中央におかれた標本を用いる事である。他の方法は重みのない平均値を標本のすべての中にある数は $1/n$ の荷重であるが、最初と最後の数は一般に $1/n$ と違った荷重をうける荷重平均に変換することである。この様な荷重を端末補正という前の様に、我々は 1 と k との間の無作為な数 i を選んで組織標本を作る。二つの端に割当てる荷重は選ばれた i の値による。荷重標本総計を計算する際に荷重平均を得るための n で割る前に荷重は次の通りである。

$$\text{最初の数} \quad 1 + \frac{n(2i-k-1)}{2(n-1)k}$$

$$\text{最後の数} \quad 1 - \frac{n(2i-k-1)}{2(n-1)k}$$

i の如何なる値に比しても、二つの荷重は常に 2 をつけ加える。

(例)

$n=4, k=3, N=12$ としよう。三つの可能な組織標本に対する荷重平均は次の通りである。

$$i=1 \quad \bar{y}_{wsy} = \frac{1}{4} \left(\frac{5}{9} y_1 + y_4 + y_7 + \frac{13}{9} y_{10} \right)$$

$$i=2 \quad \bar{y}_{wsy} = \frac{1}{4} (y_2 + y_5 + y_8 + y_{11})$$

$$i=3 \quad \bar{y}_{wsy} = \frac{1}{4} \left(\frac{13}{9} y_3 + y_6 + y_9 + \frac{5}{9} y_{12} \right)$$

この荷重のシステムが何故合理的であるかを大ざっぱにのべる。と $i=1$ であるときは標本での最初の数 y_1 は少ない荷重をうけている、何故ならば有限母集団の一端にそれがあるからである。一方 y_{10} は大きな荷重をうけている、何故ならば、この標本の他の数よりも、それにより近くある観測値 y_9, y_{11}, y_{12} をよく表わすからである。端末補正は数値積分に対する Euler -

MacLaurin の公式に於ける二つの端に割当てられる、係数 $\frac{1}{2}$ と同じである。

単に直線的傾向のみからなっている母集団に於て $\bar{y}_{wsy}$ は常に正しい母平均を与えるという事は容易に示される。従つて、 $V(\bar{y}_{wsy}) = 0$ である。直線母集団前の様に $y_i = j$ で以て表わし、 i で以て出発する組織標本を考えよう。

$$\begin{aligned} n \bar{y}_{wsy} &= i \left\{ 1 + \frac{n(2i-k-1)}{2(n-1)k} \right\} + \{ i+k \} + \dots \\ &+ \{ i + (n-2)k \} + \{ i + (n-1)k \} \left\{ 1 - \frac{n(2i-k-1)}{2(n-1)k} \right\} = ni \frac{k n(n-1)}{2} \\ \frac{n(2i-k-1)}{2} &= \frac{n(kn+1)}{2} \end{aligned}$$

従つて i の値如何に抱わらず $\bar{y}_{wsy} = (kn+1)/2 = \bar{y}$ である。これ等の端末補正は完全に母集団に於ける直線的傾向の影響を除く、実際の母集団ではシリーズの一つの数と他の数と全々無関係な“無作為”な変動と同様により複雑な傾向をもつ型が表われる。独立な変動に因する限り、端末補正はかなり正確度に損失をもたらず、何故なら、もしも y_i が同じ分散 s^2 を以て独立に変化するれば、

$$V(\bar{y}_{wsy}) = s^2 \sum_j w_j^2$$

此処に、 w_j は y_j につけるべき荷重である。荷重をうける平均 $\bar{y}_{wsy}$ では $\sum w_j^2 = 1/n$ である。 $\bar{y}_{wsy}$ では、 $\sum w_j^2$ は標本の出発の数による、 $i = 1, 2, \dots, k$ なる範囲でとられた $\sum w_j^2$

$$\frac{1}{n} \left\{ 1 + \frac{n(k^2-1)}{6(n-1)^2 k^2} \right\}$$

分散の増大は n が小なとき以外は無視出来る、 $n=10$ では増大因子は殆んど約 2% である。

端末補正は Yates (1948) により最初提案された、彼は、最初

と最後の数にすこし違つた荷重

$$1 \pm \frac{(2i-k-1)}{2k}$$

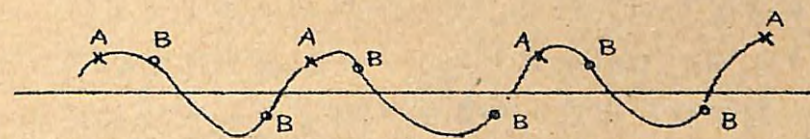
を割当てた、これらは前に与えた荷重と $(n-1)/n$ だけ違う、五つの自然母集団での彼の端末補正の有効性の検定に於て (86 表)、Yates は五つの場合の内四つに正確度に相当の増大がある事を見出した。

§ 8.8 周期的な変動をもつ母集団

populations with periodic variation:—

もしも母集団が周期的な傾向、例えばサインカーブからなっているとすれば、組織標本の有効性は k の値による、これは 8.3 図をみれば分る。

8.3 図 周期的変動



この表示に於て、曲線の高さが y_i なる観測値である、標本 A は組織標本についてあまり好きしくない場合を示す。この場合は k がサインカーブの周期に等しいか又は周期の整数倍の時に何時でも成立する組織標本内のすべての観測値は正確に同じであるから標本が母集団から無作為にとられた単一観測値ほど正確ではない。

最も好ましい場合は (標本 B) k が周期の半分の偶数倍のときである。各組織標本は真の母平均に正確に等しい平均値をもつ。何故ならば、中央線の上下の偏差はお互には打消すからである。従つて平均値の抽出分散は 0 である。これ等二つの場合の中間では標本は k と波長との間の関係により色々な有効性を示すことになる。

§ 8.9 自己相関母集団

Auto correlated populations:—

多くの自然母集団では、二つの観測値 y_i, y_j が i 及び j がそのシリーズであるに因んでいるときよりもくつついている時がより似通っているのであろう事が期待される。これは我々がそのシリーズに沿って進むとき自然の力がゆつくりとした変化を導いている時はいつても起る。この影響に対する数学的モデルに於て、我々は y_i と y_j との間には正の相関があり、その相関はたゞ、その距離 $|i-j|$ のみの函数であり、距離が増すと小さくすると仮定する。このモデルはあまり簡単化しすぎているけれども、多くの自然母集団の著しい特性の一つを支えている。

このモデルを一つの母集団に適用するかどうかを研究するために、我々はこれらの距離だけ因んでいる項目に対して相関 ρ_{ij} の組を計算してこの相関を i に対してプロット出来る。この曲線又それを表わす函数をコレログラムと呼ぶ。このモデルは妥当であつても、コレログラムは有限母集団ではスムーズな函数はない、何故なら母集団の有限の性質により不規則性を生ずるからである。このモデルに対して層化無作為抽出と組織抽出を比較する様に、これ等の不規則性は単一の有限母集団に對比の結果を導くのを困難にさせる。比較はモデルを適用する無限超母集団から無作為に抽出された有限母集団の全シリーズの平均上で行われる。この方法は既に定理 8.5 に適用した。

今我々は、観測値 y_i ($i = 1, 2, \dots, N$) は

$$E(y_i) = \mu, \quad E(y_i - \mu)^2 = \sigma^2, \quad E(y_i - \mu)$$

$(y_i + u - \mu) = \rho_{iu} \sigma^2$ (8.10) なる超母集団から抽出されたものとする。

距離に u になるときは何時でも

$$\rho_{uu} \geq \rho_{uu} \geq 0$$

この超母集団から y_i の一組を抽くと大きさ N なる一つの有限母集団が出来る。

組織標本に対する平均分散は次で示される。

$$E V_{oy} = E E (\bar{y}_{oy} - \bar{y})^2$$

この母集団のクラスに対して層化無作為抽出が単純無作為抽出よりもすぐれている事を示す事は容易であるが組織抽出について一般的結果は得られない。クラス内で組織抽出が層化無作為抽出よりもすぐれている超母集団はあるけれども組織抽出があるものの値に対し層化無作為抽出より劣っている超母集団も亦ない。

もしも更にコレログラムが上に凹であれば一般的定理を得る事が出来る。

定理 8.6

もしも (8.10) の条件に加えて

$$\delta_i^2 = \rho_{ii} + 1 - \rho_{i-1} + 2\rho_{i-1} \geq 0 \quad (i = 2, 3, \dots, (N-2))$$

であれば、如何なる標本の大きさに対しても

$$E V_{oy} \leq E V_{st} \leq E V_{ran}$$

更に $\delta_i^2 = 0$, $i = 2, 3, \dots, (N-2)$ でなければ

$$E V_{oy} < E V_{st}$$

証明は Cochran (1946) が与えたし此処では再びくりかへさない Quehenille (1949) は定理 8.6 の不等式は条件の二つが

$$E(y_i) = \mu_i; \quad E(y_i - \mu_i)^2 = \sigma_i^2$$

である様にゆめられると妥当である事を示した。

この場合三つの平均分散の各々は同じ量だけ増大する。

実際的な応用に因する限り、上に凹なるコレログラムは特殊な自然母集団に対してのモデルとして数人の研究者によつて提案されてきた。函数 $\rho_{iu} = \tanh(u - \frac{1}{2})$ は u だけ離れた二つの観測所での二つの雨量間の相関である事を Fisher 及び MacKenzie (1922) が暗示した又 $\rho_{iu} = e^{-\lambda u}$ は林業及び土地使用調査に対して

Urbanc (1942) が又 $\rho_{iu} = (1-u)/e$ は経済的時系列のある型に対して Wold (1938) により暗示された。

§ 8.10 自然母集団

Natural populations

色々な自然母集団についての調査が行われた。その資料を 8.6 表に示す

8.6 表 組織抽出の研究に用いられた自然母集団

関係事項	N	資料の型
<i>Gates</i> (1948) 13表	232	兵器調査地図から 0.1 マイル間隔で読まれた高度
<i>Osborne</i> (1942)	米	引いた平行線上の (i) 耕地 (ii) 灌木 (iii) 草地 (iv) cover - type map 上で林地の面積の百分率
<i>Osborne</i> (1942)	米	cover-type map 上で引いた平行線上の <i>Douglas fir</i> の面積百分率
<i>Gates</i> (1948)	192	192 の連続した日の土壌温度 (草下 12 インチ)
<i>Gates</i> (1948)	192	192 の連続した日の土壌 (裸地下 4 インチ)
<i>Gates</i> (1948)	192	192 の空気温度
<i>Gates</i> (1948)	96	馬鈴薯の 96 列の収穫
<i>Finney</i> (1948)	160	3 フィート巾で色々な長さの帯当り定規し得る木材材積 (<i>Mt Stuart forest</i>)
<i>Finney</i> (1948)	288	254 インチ中 80 フィート 長さの帯当り原始林の材積 (<i>Black's Mountain forest</i>)
<i>Finney</i> (1950)	292	2 フィート巾で色々な長さの帯当りの立木材積 (<i>Pikru Dun forest</i>)
<i>Johnson</i> (1943)	400+	広葉樹苗床の 4 つの床の 1 フート 床幅当りの苗木本数
<i>Johnson</i> (1943)	400	松の苗床の 3 つの床の 1 フート 床幅当りの苗木本数

Johnson (1943) 400+ 松床苗床苗の 6 つの床の 1 フート 床幅当りの苗木本数

注意 * 理論的にももしも線が無限にうすいと考える事が出来れば N は無限である。

+ 近似的である。数は床毎に変化している。

最初の三つの研究は地図で行われた。最初の研究に於ては、有限母集団は起伏のある地方での 0.1 マイルの連続した間隔での 282 の高度からなっている。次の二つの研究では資料はある地表 (cover) の型 (例えば草地) に属する cover-type map から抽出された線の長さの一部分である。これ等の例は数字的意味に於て連続変動に最も近いと考えられる。

次の三つは 192 の連続した日に対する次の温度に基いたものである。(i) 土壌下 12 インチ (ii) 土壌下 4 インチ (iii) 空気。この三組は遅い季節的な影響と比較したときに誤りの多い天候の日の変化のより大きな影響の方向の順序づけを示す。

残りの研究は一つの線に沿つてある継続時の植物又は木の収穫を取り扱っている。グループの代表的なものである馬鈴薯の研究に於ては有限母集団は島の 96 列の総収穫からなっている。文献を徹底的に調べなかったから、まだ更に進んだ資料が利用できるかもしれない。

ある研究では V_{sy} が層の大きさが 2 層で層当り 2 単位をもつ層化無作為標本に対する分散 V_{st2} と比較される。この比較は V_{st2} が不偏推定量が標本資料から得る事ができるから興味ある問題である。

これは V_{st} (大きさ 2 なる層と層当り 1 単位) 及び V_{sy} に対しては行う事はできない。他の研究者は V_{st} 、及び V_{st2} 共に V_{sy} と比較して報告している者もある。多くの資料はすみやかに利用できる形で V_{ran} と比較を示すものではなくて一般に V_{st2} か V_{ran} よりも正確度に利益があるように思われる。

Yates 及び *Finney* による論文では、その比較は各有限母集団

内の n 及び n の値の範囲に対して与えてある。これ等の場合では
8.7 表の資料は n の個々の値に対する分散比の幾何平均である。

8.7 表 組織抽出と層化無作為抽出の相対的正確度

資料	n の範囲	組織対層化の相対的正確度	
		V_{st}/V_{sy} 2.87	V_{st}/V_{sy} 5.68
高度	2-20		
面積百分率 (4 Cover type)		--	4.40
" (Douglas fir)		--	1.83
土壌湿度 (12 インチ)	2-24	2.42	4.23
" (4 インチ)	4-24	1.45	2.07
空気湿度	4-24	1.26	1.65
馬鈴薯	3-16	1.37	1.90
立木枚数 (Mt. Stuart)	2-32	1.07	1.35
" (Block's mt)	2-24	1.19	1.44
" (Dehra Dun)	2-32	1.39	1.89
広葉樹苗床	14	--	1.89
松苗床	14-24	--	2.22
松床苗苗	12-22	--	0.93

他の研究者は母集団当り n の1つの値だけについて計算しているが、同じ自然母集団の数個のものに対して又違った数値の事項についての資料を与えている。此処でも再び分散比の幾何平均が与られている。

資料は内容に於て限定されているが、その結果は役に立つ。 V_{st} との比較をした研究では組織抽出はどう目立たなくても価値ある正確度の一致した利益を示している。 V_{st} との比較での利益も大したものである。

結果の内部的傾向はこれは少数の研究の見方からしかこれについて行われなかつたけれども予想した通りである。利益は変化が連続して最も近いと推測される資料型で最大である。土壌から空気湿度への V_{st}/V_{sy} の低下は亦この見地から予想されるであろう。

最後の三つの項目(林木苗木の資料)では年数を少しとった又は他よりも一である松の床苗床のみが利益を示している。

8.8.11 疑似同期の影響

Quasi-periodic effects:—

これ等の研究の多くに於ては、分散比 V_{st}/V_{sy} は検定された n の違った値に対して合理的に母定していた。例外は Finney により研究された Dehra Dun 森林と L. H. Madaw (1946) により研究された広葉樹苗木本数の床の一つである。(ある研究ではたゞ一つの n の値しか含んでいないから多くの例外がある) 例外に於ては V_{sy} はかくされた同期性を幾分か見出す米口がある事を暗示する様子で標本の大きさと共に変化した。8.8 表は L. H. Madaw により提供された資料を示す。

8.8 表 疑似同期の影響を暗示する資料

k	n	V_{sy}	V_{st}	V_{ran}	V_{st}/V_{sy}
42	10	4.21	7.21	10.29	1.71
20	21	3.06	3.00	4.77	0.98
15	28	2.42	2.09	3.52	0.86
14	30	0.69	1.90	3.26	2.25
10	42	1.74	1.29	2.26	0.74
7	60	0.26	0.82	1.51	3.15
5	84	1.22	0.50	1.00	0.41

床は 420 ft の長さで単位は床幅の 1 ft であった。

n が q の倍数である時には組織標本の正確度は層化抽出のそれと比較すると高い。中間の値に対しては正確度はほぼ同じか低いのである標本の大きさを増すと V_{sy} が誤りが多くなるのはこの現象の他の反映である。Finney の資料 (1950) は同様の影響を表わしている。

どの位自然母集団での影響があるか分らない。短いシリーズよりも長いシリーズ(例えば N が大)が少ない様に思われる。8.8 表の5から22迄の n の値の範囲以上で組織抽出が行われると

大体系化無作為抽出と同じである事が観測された。従つて批判は組織抽出を行う事は実行不可能ではなく、一杯に悪い事ではないという事である。

§ 8.12 単一標本からの分散の推定

Estimation of the variance from a single sample

$n > 1$ なる単純無作為標本の結果から、我々は母集団の型はどうであろうと推定量は不偏であるから標本平均の分散の不偏推定量を計算できる。組織標本は $n = 1$ なる単純無作為標本として考える事ができるから、この有用な特性は組織標本に対して成立する。例として、サインカーブの例を考えよう。

$$y_i = m + a \sin(\pi i / 2)$$

としよう。此処に $h = 4$ で、 $i = 1, 2, \dots, 4n$

母集団における次々の観測値は

$$(m+a), m, (m-a), m, (m+a), m, (m-a), m$$

-----である。

もし、 $i = 1$ が最初の数として選ばれれば、組織標本のすべての数は $(m+a)$ なる値である i の他の三つの可能な選択に対しては矢々すべての数は $m, (m-a)$ 又は m である。従つて単一標本から我々は決して a の値を見出すか又は推定する手段をもたない。併し、組織標本の平均値の真の抽出分散は $a^2/2$ である。

例えばもしも同期的変動が存在すれば不偏である推定分散を作る事が不可能である事を示している。

これらの結果は何も行う事ができないという意味ではない。同期的変動を含んでいる場合、我々は適当に存在している変動型を表わす数値的モデルを依り上げるために母集団の構造について充分よく知っていなければならぬ。

それから我々はたとえそれが他のモデルに対してひどく偏倚があつても、このモデルに対して近似的に不偏である推定分散

に対する公式を作る事ができる。これらのモデルの一つを用いる事を決定するには抽出する人の判断によるものである。不幸にも我々は論議から明らかな様に母集団の構造についての資料をもたないし与えられたモデルが満足であるという確信はない。

それ等に応ずるある簡単なモデルを次に示そう。証明は与えない。最も簡単なモデルは

y_i が一つの傾向に加えるに無作為な成分からなっている母集団に適用する。

即ち

$$y_i = \mu_i + e_i$$

此処に μ_i は i のある函数である。無作為な成分に対して我々は

$$E(e_i) = 0 : E(e_i^2) = b_i^2 : E(e_i e_j) = 0 \quad (i \neq j)$$

なる超母集団があると仮定する。

推定分散に対する要求される公式 s_{sy}^2 は s^2 と

$$E E(s_{sy}^2) = E V_{sy}$$

であれば即ち超母集団から抽出来る。すべての有限母集団上で不偏であれば不偏であるといわれる。

I ランダムオーダーに於ける母集団

$$\mu_i = \text{一定} \quad (i = 1, 2, \dots, N)$$

$$s_{sy}^2 = \frac{(N-n)}{Nn} \frac{\sum (y_i - \bar{y}_{sy})^2}{(n-1)}$$

この場合は我々がオーダーが測定さるべき項目に関して本質的に無作為であると確信するときに適用する。

分散公式は単純無作為標本に対してのものであり、もしもモデルが正しければ不偏である。

II 層化の影響のみ

$$\mu_i = \text{一定} \quad (r h + 1 \leq i \leq (r+1)h)$$

$$s_{sy}^2 = \frac{(N-n)}{Nn} \frac{\sum (y_i - y_{ieR})^2}{2(n-1)}$$

この場合には平均値は各個の単位をもつ各層内で一定である。平均平方の次の差に基づいた推定量 s_{sy}^2 は不偏ではない。隣りの層にある u 間の差からの望ましくない貢献を含んでおり、最初と最後の層は分散の無作為成分を推定する際に荷重があまりに小さすぎる。合理的に大きな標本では、この推定量はモデルが正しいと仮定しても一般的に高すぎるであろう。

Ⅲ 直線的傾向

$$\mu_i = \mu + \beta i$$

$$s_{sy}^2 = \frac{(N-n)}{N} \frac{n'}{n^2} \frac{\sum (y_i - 2y_{i-1} + y_{i+2})^2}{6(n-2)}$$

推定量は y_i の継時における次の自乗の項に基づいている。

平方和は $(n-2)$ 項である。

直線的傾向では、我々はその傾向は端末補正で消去できる事が分った (§8.7) 項 n/n^2 は $\bar{y}_{sy}$ における荷重の平方和である。 n が小さくなければ n/n^2 は通常の因子 $1/n$ とおきかえる事ができる。端の方の層は少ししか荷重をうけないから、推定者は ϵ_i^2 が一定でなければ、不偏でないが、もしも n が大きくモデルが正しければ満足である。

より複雑な型の連続的変動が存在すれば、前の公式は貧弱な結果を与える。8.9表に於て

二番目と三番目の公式が六つの森林苗床に適用された。

(Johnson 1943)

8.9表 苗木の標本平均数の分散 (Johnsonの資料)

	床	実際の V_{sy}	s_{sy}^2	s_{sy}^2
銀カエデ	1	0.91	2.8	2.5
	2	0.74	3.6	2.9
アカカエデ	1	4.8	28.4	12.6
	2	15.5	22.6	18.6
白トウヒ	1	5.5	72.2	11.2
	2	2.0	11.6	6.4

白 松 / 2.2 21.0 21.9

二次公式は次の差に基づいたものよりもかなりよいが共にかなり過大推定をしている。

色々な他の公式が工夫できる。高次の適合した多項式からの残差はもしも μ_i が連続的に変化してあまり急速に変化しなければ有効である。この方法に対する表を DeLury (1950) が与えた。コログラムの性質についての簡単な仮定から発表した公式が、Osborne (1942), Cochran (1946), 及び Matern (1947) により論ぜられた Yates (1949) は

$(y_u + y_{u+2} + y_{u+4} + \dots) - (y_{u+1} + y_{u+3} + \dots)$ なる型の量に基づいた推定量を研究した。

標本における次の項は交互に + 及び - の符号を与える。もしもこの表示が全標本上でとられると一つの $d.f$ しか利用できない。より多くの自由度を与えるためには、標本資料は Yates が暗示したが各々 9 つの観測値を含む部分に分ける事ができる。もしも我々が y'_1, y'_2 等によって組織標本における次の観測値を表わし最初と終りの項に $1/2$ なる荷重を与えると次の様にかける。

$$d_1 = (\frac{1}{2}y'_1 + y'_3 + y'_5 + y'_7 + \frac{1}{2}y'_9) - (y'_2 + y'_4 + y'_6 + y'_8)$$

次の差 d_2 は y'_7 から出発する。次も亦同じ。従って $\bar{y}_{sy}$ の推定分散に対して

$$s_{sy}^2 = \left(\frac{N-n}{Nn} \right) \sum_{u=1}^g \frac{du^2}{2.5g}$$

因子 2.5 は du における係数の平方和であり、 g は標本が与える差の数である (g は近似的に n/g)。Yates が推定した自然母集団では、この型の公式は次の差に基づいた公式 s_{sy}^2 よりはすぐれているがまた $\bar{y}_{sy}$ の真の分散の過大推定量である。

結論として推定分散に対する公式には不足はないが、すべて応用に当っては限られた範囲がある様である。

§ 8.13 層化組織抽出

Stratified systematic sampling:—

我々は已に組織抽出そのものは一種の層化抽出である事が分った。層化が望ましいと思えるある応用に当つては組織標本の使用を導く事ができる。大都市におけるブロックの標本を考えてみよう。通常この標本は居住面積の違つた型をすべて表わした良好な地図であるという事を確保する事が望ましい。もしもブロックが連続的に数えられると即ち螺旋状にある側から他の側へ数えられると組織標本はしばしば適当な地形層化を与えるであろう。

その立場は一つの層が他の層と合同しない様に明らかである場合とか別個な推定量が各層に対して必要である場合は違つてくる。此処に於ては我々は各層に対して新しい出発点で他の違つた値で似て別個な組織標本をとる。この方法はもしも層内の組織標本が層内の単純無作為標本よりも正確であれば層化無作為標本よりも正確であろう。

もしも $\bar{y}_{syh}$ をして各層における組織標本の平均値とすれば母平均 $\bar{y}$ の推定量は通常の如く

$$\bar{y}_{stgy} = \frac{\sum N_h \bar{y}_{syh}}{N}$$

定理 5.2 から推定量は各層で不偏と仮定して

$$V(\bar{y}_{stgy}) = \frac{\sum N_h^2 V(\bar{y}_{syh})}{N^2}$$

層の数が少ない場合はこの量の標本推定量を見出す問題は前節で論じた方法の一つによつて各層で $V(\bar{y}_{syh})$ の満足な標本推定量を見出す問題となる。層の数が20以上のときは層の対向の差に基づいた推定量がより好ましい。層の大きさ及び標本の大きさは近似的に対の数と同じでなければならぬ。もしも前節の二つの層が対を作れば、

$$E(\bar{y}_{sy1} - \bar{y}_{sy2}) = V(\bar{y}_{sy1}) + V(\bar{y}_{sy2}) + (\bar{y}_1 - \bar{y}_2)^2$$

その結果、推定量は

$$V(\bar{y}_{stgy}) = \frac{1}{N^2} \sum N_h^2 (\bar{y}_{syh} - \bar{y}_{syj})^2$$

此処に和は層の対上に反ぼすこの推定量はたとえ周期的な影響が層内に存在しても大体過大推定量となる。過大推定の量は $(\bar{y}_1 - \bar{y}_2)^2$ の項による。予測できる限り同じ対における層は従つてほぼ同じ母平均でなければならぬ。この工夫は § 5.21 ですでに論じた「不成功な層」(collapsed strata)の方法の応用である。

§ 8.14 二次元に於ける組織抽出

Systematic sampling in two dimensions:—

二次元である場合の抽出問題は一次元組織標本をとる事ができる様な単位を数える事によつて取扱われる。前節で述べた市のブロックの標本はその一つの例である。他の例は通常森林調査に用いられる方法である。

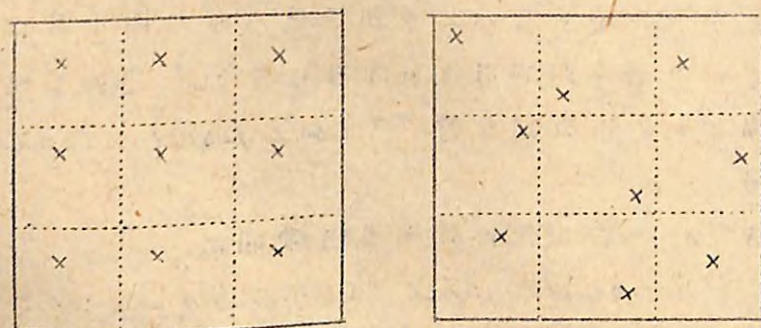
等距離の平行な帯のシリーズが森林の全域にわたつて地図がとられる。各帯に於ける立木材積が時には帯の全林母木が時には木の標本を測定する事により推定される。もしも y_i をして i 番目の帯の全立木材積とすれば、一次元理論が適用される森林の幅が変化するとき、自然の模型は任意変量として i 番目の帯の面積 a_i を考えて比又は回帰推定を適用する。

正方形の面積に対する二つの組織標本の型を a 図に示す。正方形格子に似ている左側の標本は上の左側の単位の座標を定めるため乱数の対を選ばると完全に決定される右側の標本は水平及び垂直の次々の層における単位間の距離が常に等しいからこれも亦組織的である。

しかし § 8.4 a とちがつて単位は同じ線上にはない。この種の一回線上にない標本を選ぶためには我々は乱数の対によつて上の左側の単位の座標を決定する。

二つの附加的な座標が層の最初の欄における残りの単位の水平座標を決定する。他の二つは層の最初の列における残りの単位の垂直座標を決定するために必要である。

8.4図 二次元組織標本の二つの型



(a) 一線上にある又は方形格子標本 (b) 一線上にない標本

Quenauille (1949) 及び Das (1950) はある簡単な二次元コレログラムに対しての理論的研究において層化抽出の色々な型をもつ二次元組織標本を比較した。その結果は一線上にない組織標本はしばしば層化無作為抽出よりもすぐれているという事を示している。

Quenauille の分析は方形格子は一線上にない標本ほど正確ではない事を暗示している。この暗示は Haynes (1948) の自然母集団についての研究によって支持された。十四の場合の一樣な試行に於て彼は格子は二次元の単純無作為抽出とほぼ同じ正確度にすぎない事を見出した。

方形格子を相対的にまづく行方という事は、我々が直線的傾きの影響を考えると予期されない事はない。もしも面微の水平側に平行に着しい直線的な傾きがあれば、8.4図における方形格子はたゞ三つの英だけでこの傾きを抽出するだけである。それは直線的に一線上にない標本の如く九つの違った英で傾きを抽出する方法がよりすぐれている様に見える。

一線上にない標本の卓越性に対する更にも一つの証拠はラテン方格 (Latin square) が矩形の畝で処理を配列するために正確な方法である事がわかった場合の実験計画における経験から得られる。

8.5図の 5x5 ラテン方格はその各々が文字で表わした五つの組織標本に畝を区分するものとして考えられる "knight move" ラテン方格と呼ばれる。この特別な方形は恐らく一線上にある事が引及び行も同様に対角線に存在しないであろうから無作為に選ばれた 5x5 ラテン方格よりかなり正確であるという証拠がある。

ラテン方格の原理は燕麦の矩形畝を抽出する事に於て Nomeyer and Black (1946) によって用いられた。

各畝は 21 プロットを含んでいる。三つの可能な組織標本は 8.5図に夫々 A, B, C なる文字で示されている。各畝と無作為に選ばれた文字の一つで以ての配列は層として列での層化無作為抽出よりの約 25% 以上の正確度の増大を与える。配列は全くラテン方格の特性を満足しない。何故ならば、各文字は一つの欄に 3 回、他の欄に 2 回あらわれているからである。しかし、この特性に出来るだけ接近はしている。

8.5図 ラテン方格に基いた二つの組織標本

A B C D E	A B C
D E A B C	B C A
B C D E A	C A B
E A B C D	A B C
C D E A B	B C A
	C A B
	A B C

(a) "knight move" ラテン方格 (b) 3x7 矩形畝に対する組織標本

組織標本は抽出並びに実行上便利である。この章で報告した多くの研究に於て即ち人工的にしても自然母集団にしても共にかなりよく層化無作為抽出との正確度を比較したその不利益な事は予期しない周期性が存在するときは正確度が悪くなり、標本から $V(\bar{y}_{sy})$ を推定する信頼するに足る方法が分らないという事である。

これ等の結果に照してみれば抽出研究をした人達は組織標本抽出を推奨する事に対して意見が一致していない。併し乍ら、組織抽出は次の立場では安心しておすすめる様に思われる。

i 母集団の配列が本質的に無作為であるか又は殆んどわずかでも層化を含んでいる場合。この場合組織抽出は便利に使用され、正確度もすこしの利益が期待される。合理的に不偏な誤差の標本推定量が利用できる。(§ 8, 12)

ii 多くの層での層化が用いられ、各層から独立な組織標本が抽出される場合。この立場ではかくれた周期性の影響を打消す傾向があり、過大推定量であると分っている誤差の推定量が得られる。

(§ 8, 13) その他我々は層の数を半分にして用い、独立なランダムスタートで以て二つの組織標本を抽出できる。この方法は誤差の不偏推定量を与える。

iii 単位の副次抽出に対して用いられる(第10章) この場合抽出誤差の不偏推定量が大抵の実際の立場で得る事ができる事が指摘される。

iv 仮りに抽出誤差の推定量が正しく必要でないとするれば、連続型の変動をもつ母集団の抽出に用いられる。もしもこの種の型の調査のシリーズが行われるとするれば、抽出誤差をその場合々に4エックする事が有効である。Gates (1948) は如何にして補助的な観測値をとつてこれを行うかを論じた。

結論として、更に進んだ研究は推定量 $V\bar{y}_{sy}$ を主旨とする公式の妥当性及び応用範囲の我々の知識を広めてくれるだろうし、改善された公式を導いてくれるであろう。

第9章 標本抽出単位の型

Type of Sampling Unit

§ 9.1 最適単位

The Optimum Unit

時には母集団は色々な方法で単位に分ける事が出来る。市は多くのブロックからなり、多くの家族からなり又は多くの人間からなっている。土壌の標本抽出に於ては、標本をとる道具は色々な大きさ及び形に作る事が出来、その各々が畝の色々な区分を抽出単位に分ける。単位の形の变化は通常標本をとる費用及びそれから得られる正確度の両方による。従って単位の最適の型の決定は抽出の経済性に於て重要である。

最適単位は最小費用で標本に対して欲する正確度を与える単位又は一定費用に対して最大の正確度を与える単位である。与えられた標本の大きさに対して、大きな単位は小さな単位よりも殆んど常に費用はかゝらないが、正確度は劣る。単位の選択は相対的正確度と相対的費用との間のバランスをとる問題を含む。殆んどの実際的な決定に於ける様に、次の如き評価されない様な因子がある。即ち単位の一つの型はある特別な便利さがあるか又は費用の計算に於て含まるべき困難である不利益さをもっている。穀物の標本抽出に於ては、ある経験で単位の正確な境界については不正確であるため、小さい単位は偏倚があることを暗示している。例えば Homeyer & Black (1946) は 2×2 尺の単位は 3×3 尺の単位よりもほぼ8%高い燕麦の収穫を与えた事を見出した。これは恐らく抽出者が疑わしい時に単位の内部に境界の穀物を入れる傾向があったためであろう。Sukhatme (1947) は小麦及び米に対して同様な結果を例証している。

§ 9.2 簡単な例

A simple example: -

White pineの苗木床に対するJohnsonの資料は(1941)色々な単位を比較するための方法の簡単な例を与える。床は4例で各々の長さは434 ftであった。床を抽出単位に分けるためには多くの方法がある。単位の四つの型に対する資料を9.1表に示す。床は完全調査されたから、資料は正しい母集団の値を示す。

単位は次の如きであった。即ち

単一例の1フート
単一例の2フート
床幅の1フート
床幅の2フート

9.1表 抽出単位の四つの型に対する資料

予備資料	単位の型			
	1フート 列	2フート 列	1フート 床	2フート 床
単位の相対的大きさ	1	2	6	12
N_u = 母集団での単位の数	2604	1302	434	217
S_u^2 = 単位当り母分散	2.537	6.746	23.094	68.558
15分間で数え得る列の フート数	44	62	73	103

最初の二つの単位では、抽出は列毎の層化と仮定されるから、 S_u^2 は列内分散を表わす。単純無作為抽出は最後の二つの単位に対して仮定された。

主な費用は単位を配置して数える費用であるから、費用は時間研究によって推定された。(9.1表の最後の列)より大なる単位では標本の大きなかたまりを15分間で数える事が出来る。というのは一つの単位から他の単位に移動する時間が少くてすむからである。

推定さるべき項目は、苗木の母集団総数である。この型の研究に於ては、母集団総計は母平均よりも論議の廉便利である。何故ならば2フート床の単位に対して単位当り平均値は1フート列単位に対する単位当り平均値と全く違った量であるからで母集団総計はすべての単位に対して同じ量であるからである。

もしも $f \cdot P \cdot C$ を無視すると、推定母集団総計の分散は

$$\frac{N_u^2 S_u^2}{n_u}$$

此処に $u = 1, 2, 3, 4$ は単位の型を表わす。この分散はすべての単位に対して同じであるべきである。もしも最小単位が標準単位として送られると、最小単位と同じ正確度を与える他の n_u の値は次の方程式から得られる。

$$\frac{N_1^2 S_1^2}{n_1} = \frac{N_u^2 S_u^2}{n_u} \quad \text{即ち } n_u = n_1 \left(\frac{N_u}{N_1} \right)^2 \frac{S_u^2}{S_1^2}$$

例えば n_1 と比較さるべき n_2 の値は

$$n_2 = n_1 \left(\frac{1}{4} \right) \frac{6.746}{2.537} = 0.665 n_1$$

この資料は9.2表の最初の欄に与えてある。

9.2表 比較された標本の大きさと費用

計算の逐次段階	単位の型			
	1フート 列	2フート 列	1フート 床	2フート 床
n_u の比較値	n_1	$0.665 n_1$	$0.253 n_1$	$0.188 n_1$
比較された標本の 大きさ (1フート列単位で)	n_1	$1.330 n_1$	$1.518 n_1$	$2.256 n_1$
比較された費用	C_1	$0.944 C_1$	$0.856 C_1$	$0.919 C_1$
相対的正確度	100	106	117	104

次の段階は列の単一フートの項で比較さるべき標本の大きさを見出す事である。何故ならば費用はこれ等の項で表わされるからである。 n_2 に対して単位は列の2フートを含んでいるから(9.2

表の2番目)上の欄を2倍する。単位が大きさが増すにつれて同じ正確度を得るに要する標本の大きさは又増大する。事実2フィート床の単位では標本はノコート列の単位の2/4倍でなければならぬ。

最小費用の n_1 をとるに要する費用は次の如くして表わされる。

$$C_1 = \frac{n_1}{44}$$

何故ならこれは15分間隔で必要とされる時間である。同様に2番目の単位での標本の費用は

$$\frac{1.330 n_1}{62} = 1.330 \left(\frac{44}{62} \right) C_1 = 0.944 C_1$$

これは9.2表に示してある。すべてより大きい単位はそう大して差はないが、最小単位よりも費用がかゝらない。ノフト床が最良の様に思われる。9.2表の最後の欄は単小単位を100とした場合の費用の逆数である。表に於てこれ等の数字は相対的正確度と呼ばれた。何故ならこれはもしも同じ比較が分散の代りに費用を一定にして行われるとこれ等の数字は推定母集団総計の相対的分散に逆比例し従って相対的正確度を測る尺度となるという事が分るからである。

§ 9.3 単位を比較する一般的方法

General Procedure for Comparing Units:-

この例での分析は次の如き一般的形式で表わす事ができる。

定理 9.1

単位の u 番目の型に対して次の如くかく

単位の相対的大きさ $= Mu$

単位での項目総計間分散 $= Su^2$

単位当り相対的費用 $= Cu$

f. P. C を無視した単純標本作為抽出で母集団総計は簡単に拡大によって推定されると仮定する。そのとき同じ正確度に対する相対費用は次のものに比例する。

$$\frac{Cu Su^2}{Mu^2} \quad (9.1)$$

(証明)

これは前の数値例に用いられた論議に従う。

u 番目の単位に対して

母集団に於ける単位数 $\propto 1/Mu$

推定母集団総計の分散 $\propto Su^2/n_u Mu^2$

等しい正確度に対する標本の大きさ(n_u) $\propto Su^2/Mu^2$

等しい正確度に対する相対的費用 $\propto Cu Su^2/Mu^2$

Su^2 及び Cu の定義は注意しなければならない。何故ならば資料の編輯に於てこれ等の量を最初はある他の形で表わすのが便利であるからである。例えば数値例では費用の資料は与えられた時間に数える事ができる標本の大きさの項で与えられた。

系 1

定理の条件の下では一定費用に対する推定母集団の分散は又次のものに比例する。

$$\frac{Cu Su^2}{Mu^2} \quad (9.2)$$

これは同じ論議により証明される。

系 2

分散分析に於て、違った大きさをもち単位に対する分散はしばしば所謂通常の基礎のもとに計算される。大きな Mu なる単位の総計間分散 Su^2 は Mu で割られる。例えば列のノフト当りの分散の項で9.1表の分散をすべて表わしたいと仮定しよう。

2フート列単位は2つの単一フートの総計であるから、分散6.746を2で割って、3.373を得る。同様に三番目及び四番目の単位に対する分散は天々も及びノで割る。その結果は次の通りである。

列の単一フート当りの比較分散

単位の型			
1フート 列	2フート 列	1フート 床	2フート 床
2.537	3.373	3.849	5.713

通常の基礎の下でもなお、分散は単位が大きくなると一定に増大する。

$S_u^2 = S_u^2/M_u$ としよう。そうすれば量 S_u^2 は通常の基礎の下にある。亦 C_u' をして標本のあたえられたかたまりをとる費用としよう。そのとき $C_u' \propto C_u/M_u$ である。従って定理9.1は次の様に述べられる。

等しい正確度に対する相対的純費用 $\propto C_u' S_u^2$

相対的純正確度 $\propto 1/C_u' S_u^2$

この結果はもしも我々が標本をとる費用を一定にすれば（即ち C_u' を一定にすれば）相対的純正確度は S_u^2 に逆比例する事を示している。換言すれば、標本の同じ総量に対して違う単位を比較するためには、それに関連した量は通常の基礎に導かれた単位間分散である。

定理9.1及びその結果はもしも層が同じ大きさであり又 S_u^2 , S_u^2 が層内の平均分散を表わせば、比例割当てをした層化抽出に対しても妥当である。これは上に述べた条件の下では f.p.c. を無視した推定母総計の分散は $N^2 S_u^2/n$ であり、従って単純無作為抽出の形と同じものであると仮定しているからである。定理9.1はより複雑な抽出形式には成立しない。

前の結果は一般的な方法の例として単に述べただけである。「単位間の比較は常に実際に用いられるべき抽出の種類に対して又はも

しも、これがきめられなかった場合は考えられている種類に対して行わなければならない」抽出又は推定の方法を変える事は、違った単位の相対的純正確度に変化をもたらす。一定の抽出及び推定法に対してさえも、相対的純正確度はもしも費用が大きさの直線函数でないか又はその大きさが f.p.c. を考慮に入れなければならぬほど大きければ、標本の大きさと共に変化するであろう。通常一つ以上の項目が考えられる場合がある。一つの米口は費用を一定にして単位の各々の型及び各項目に対して相対的純正確度を見出す事である。単位の一つの型が一様にすぐれていなければ最も重要な項目に主として重みをつけてある受動的な決定を行う。結果に影響する多くの要因の見地から、広汎な調査に於ける単位の最適の大きさの研究は大きな仕事である。農場の抽出に対する良い例が Jessen (1942) により述べられた。

彼の結果からの引用は9.3表に与えてある。これは単位の四つの大きさを比較している。即ちその単位は $1/4$ section, $1/2$ section section 及び2つの隣りあった section からなる block である。section は平均4つ以下の農場を含む1マイル平方の面積である。この比較に於て全現地調査費用（\$1000）、質問の長さ（完全に対して60分）及び旅費（マイル当り5セント）はすべて定められている。何故ならば相対的純正確度はこれ等の変量が変わると変わってくるからである。費用は1939年の水準である。

9.3表 単純無作為抽出での単位の四つの大きさに対する推定標準誤差（%）

項 目	S/4	S/2	S	2S	最良単位
白鳥の数	5.0	4.9	5.3	6.2	S/2
馬の数	3.4	3.3	3.6	4.2	S/2
羊の数	17.4	15.7	14.9	14.3	2S
鶏の数	3.0	3.0	3.3	3.8	S/4 S/2
作田生んだ卵の数	5.7	5.2	4.9	4.7	2S

牛の数	4.7	4.6	4.8	5.5	S/2
乳牛の数	3.7	3.6	3.8	4.4	S/2
ミルクのガロン数	4.4	4.2	4.4	4.9	S/2
日々の生産物収入	5.5	5.2	5.4	6.0	S/2
農場の工-カー数	2.9	2.8	3.0	3.5	S/2

穀物の工-カー数	3.7	3.5	3.8	4.4	S/2
燕麦の工-カー表	4.6	4.8	5.6	7.0	S/4
穀物収量	1.6	1.7	2.0	2.5	S/4
燕麦収量	1.6	1.5	1.6	1.8	S/2
商用飼料の経費	12.6	13.6	16.7	21.8	S/2

全経費経営者	7.8	8.1	9.6	12.0	S/4
全収入経営者	6.2	6.5	7.7	9.8	S/4
純現金収入経営者	6.8	6.9	7.8	9.5	S/4

表での資料は、18項目に対する農場当り推定平均の相対的標準誤差(%)である。すべて項目に対して最良の単位はない。併し S/2 及び S/4 はすべての内2項目を除いてより大きな単位よりすぐれていてその二つの間にはすこしか差がない。S/2の方が恐らくよりよいであろう。何故ならば境界を正確に示す問題はより容易であるからである。

この種類の比較をするためには、我々はそれに含まれている単位の各型に対して単位間の変動を知らねばならない。如何にしてこの様な資料を得るか？ 苗木床の例に於ける如くその一つの手段は単位の各型に対して別個な標本を抽く事である。この様な方法は母集団が完全で資料を得るにあまり費用がかゝらない場合に用いられる。併し広汎な母集団では単位の違った型を比較する目的で単に調査をする事は決して容易ではない。単位の最適型についての情報は主目的が推定を行う様な調査の巧妙な副産物として通

常得られる。これを行うある技術を次節で概説しよう。

§ 9. 4 調査資料から行う比較

Comparisons made from survey data: -

調査に於て単位が M 個のより小さい単位に分割出来ると仮定しよう。標本に於ける各大単位に対しての総計のみを記録する代りに、M 個の小単位の各々に対して別個に記録する。そのとき大単位と小単位の正確度の比較が出来る。大きさ n なる単純無作為標本を最初に仮定しよう。9.4 表の分散分析が標本から計算出来る。

9.4 表 標本資料の分散分析 (小単位を基礎として)

	d. f	m. S
大単位間	(M - 1)	δb^2
大単位間小単位間	$n(M - 1)$	δw^2
標本に於ける小単位間	(nM - 1)	$\delta^2 = \frac{(n-1)\delta b^2 + n(M-1)\delta w^2}{n(M-1)}$

大単位 (小単位を基礎とした) の推定分散は δb^2 である。小単位の分散の使用推定量は標本に於て すべての小単位間平均平方と考えられる。即ち

$$\delta^2 = \frac{(n-1)\delta b^2 + n(M-1)\delta w^2}{(nM-1)} \quad (9.3)$$

この推定量は大抵の場合は満足であるが偏倚がある。何故ならば標本は M 個の単位の隣りあった群で抽出されたから小単位の単純無作為標本ではないからである。

不偏推定量は N 個の大単位及び NM 個の小単位を含む全母集団に対して 9.5 表の様な分散分析と作って標本から得られる。

9.5表 全母集団に対する分散分析 (小単位を基礎として)

	d. f	m. s
大単位間	(N-1)	S_b^2
大単位間小単位間	$N(M-1)$	S_w^2
母集団に於ける小単位間	$S^2 = \frac{(N-1)S_b^2 + N(M-1)S_w^2}{(NM-1)}$	

定義により小単位間母分散は表の最後の欄で与えられる。即ち

$$S^2 = \frac{(N-1)S_b^2 + N(M-1)S_w^2}{(NM-1)}$$

単純無作為標本では9.4表の S_b^2 は S_b^2 の不偏推定量であることは容易に示される。従って母集団におけるすべての小単位間分散 S^2 の不偏推定量は

$$S^2 = \frac{(N-1)S_b^2 + N(M-1)S_w^2}{(NM-1)} \quad (9.4)$$

n が 50 以上であれば、共に近似的に

$$S^2 \div S^2 \div \frac{S_b^2 + (M-1)S_w^2}{M} \quad (9.5)$$

となるから、実際的に(9.3)と(9.4)は同じである。

二つの推定量 S_b^2 (最初の単位に対して) 及び S^2 (小単位に対して) は従って推定母総計に対する分散公式に導入される。

もしも標本が大きければ小単位は大単位の無作為副次標本に対して測定される。(例えば 600 から 100) その他各大単位から無作為に抽出された二つの小単位が測定される場合がある。一つの大きさより以上の小単位は仮りに各小単位に対して S_w^2 の不偏推定量を与える資料をとるとすれば同時に調査される。

層化抽出では大単位と小単位に対する分散は各層で別個の方法により推定出来る。従って層化標本からの推定量の分散に対する使用公式に代入出来る。

(例)

農場労働者を推定するため、1942年に North Carolina でと

られて農場標本から資料をとる。(Finkner, Morgan & Monroe 1943)。標本抽出の方法は地図に点を無作為におき、各点に最も近い三つの農場を抽出単位として選ぶ。この方法は推奨出来ない。何故ならば大農場は小農場よりも標本に含まれる機会が多く、孤立した農場は一処にかたまつた農場地帯の農場より多くとられる機会をもつからである。この偏倚の影響はこの例では無視しよう。標本は層化された。その層は農場母集団の密度及び穀物地帯と農場地帯との比に等しい町のグループであった。五月にとられた標本に対する資料は9.6表に示されている。

9.6表 母集団と標本の大きさ

	母集団	標本
層の数	587	572
標本抽出単位の数	72,849	1,397
農場の数	217,976	4,165

小数の層は抽出されなかったし、当位当りの農場が3つ以下のものがある事は注意を要する。このへだたりを無視しよう。抽出比は 1.9% であった。

標本資料から我々は抽出単位として個々の農場と三つの農場グループとを比較出来る。我々は厳密な分析よりはかなり簡単な分析を行ってみよう。f.p.c. は無視出来る。抽出は層化されているから、推定母総計の分散は

$$V(\hat{Y}_{st}) = \sum_h \frac{N_h^2 S_h^2}{n_h}$$

各層で標準の方法で単位の二つの型に対して S_h^2 の推定量が計算出来る。この公式に入れられる。層は一般に 300 と 450 の間の農場を含んでいた。そして二つ又は三つの三農場単位の何れかが近似

的に比例抽出される様に各層でとられた。比例的即ち $n_h/N_h = n/N$ と仮定して

$$V(\hat{Y}_{st}) = \frac{N}{n} \sum N_h S_h^2 \doteq \frac{N^2}{n} \bar{S}_h^2$$

更にもしも S_h^2 が層間であまり大して違わないと仮定すればその平均 $\bar{S}_h^2$ で置きかえる事が出来る。 $\bar{S}_h^2$ の推定量は単一農場を基礎として 9.7 表の分散分析から得られる。

9.7 表 標本分散分析 (日雇労働者の数)
(単一農場を基礎として)

	d.f	m.s
層内単位間	825	6.218
単位内農場間	2768	2.918
層内農場間	2593	3.676

3農場のグループに対しての平均平方 $\bar{S}_h^2 = 6.218$ は $\bar{S}_h^2$ の推定量として役立つ。個々の農場に対する層内分散の推定量を得るため、全母集団に対して分散分析をする。(9.8 表) 自由度は 9.6 表から分り最初の二つの平均平方は 9.7 表から分る。

9.8 表 全母集団に対して作られた分散分析

	d.f	m.s
層内単位間	72,262	6.218
単位内農場間	145,127	2.918
層内農場間	217,389	

従って層内農場間分散は次の様に計算される。

$$\hat{S}^2 = \frac{(72,262)(6.218) + (145,127)(2.918)}{217,389} = 4.015$$

推定分散 3農場グループに対して 6.218 及び個々の農場に対して 4.015 は通常の基礎の下のものであるから、これ等の二つの数字は標本の一定総大きさに対する二つの単位の相対的正確度を表わす。農場のグループは単一農場の正確度の約 2/3 にすぎない。費用を考慮に入れるとおそらく 3農場グループがよいように思われる。

§ 9.5 分散函数

Variance function: -

前の例で我々は 3農場単位と、2又は6又は10農場からなる単位とを比較したいと仮定しよう。我々は単位の大きさ M の函数として母集団に於ける単位間分散 S_b^2 を予測するある方法が必要である。分散分析により S_b^2 はもしも (i) 母集団に於けるすべての要素間分散 S^2 及び (ii) 同じ単位に属する要素間分散 S_w^2 が分れば見出す事が出来る。ここに示す方法ではその条口は S_w^2 及び S^2 を予測し、分散分析により S_b^2 を見出すことである。

標本資料は実際に用いた単位の大きさに対して S^2 及び S_w^2 の推定量をもたらす。 S^2 は要素間分散であるから単位の大きさには影響されない。併し、 S_w^2 は影響される。それは大単位の大きさが増すにつれて増大する事が期待される。もしも検定さるべき大単位が実際に用いた単位と大きさに於てあまり差がなければ第一次近似は標本資料によって与えられた推定量を用いて S_w^2 を一定と考える事である。Mc Vay (1947) による研究は、この近似はしばしば満足であることを暗示している。

よりよい近似としての試みが如何に S_w^2 が単位の大きさと共に変化するかを予測する一般的な法則を発展させるために行われた。

(Jessen, 1942; Mahalanobis, 1944; Hendricks, 1944)

二三の農業調査においては S_w^2 は次の実験方式により M と関係

している様に思われる

$$S_w^2 = AM^g \quad (g > 0) \quad (9.6)$$

ここに A 及び g は M によらない常数である。この公式に於て S_w^2 は M が増すと増大する。通常 g は小さい。この型の曲線は共に接近している要素に同様な影響を及ぼす力がある時に期待される。天候、土壌型、地形及び市販路は隣接農場に似た特性をもたせる傾向がある。

理論的に公式は S_w^2 は M が増すと、とが事なしに増大するから異論をうけ易い。もしも我々が合理的と思える様に、あまりかけはなれた要素間には相関がないと仮定すると S_w^2 は大きな M で上の限界に近づく公式がより有用であろう。併しもしも調査研究下の M の範囲で充分適合しておればどの公式も満足なものである。若しこの公式が適合してをれば、 $\log S_w^2$ は $\log M$ に対して直線としてプロットされる。

M の最小二つの値に対して S_w^2 の値が、常数 $\log A$ 及び g を推定するために必要である。少くとも三つの M の値は、適合の遠徴性の評価には必要である。

9.5表の分散分析から

$$\begin{aligned} S_b^2 &= \frac{(NM-1)S^2 - N(M-1)S_w^2}{N-1} \\ &= \frac{(NM-1)S^2 - N(N-1)AM^g}{N-1} \quad (9.7) \end{aligned}$$

もしも大単位の数 N が大であれば、これは次の如く簡単な型となる

$$S_b^2 = MS^2 - (M-1)AM^g \quad (9.8)$$

* 引用文献では、 N は通常母集団の要素の数として定義されているから、公式(9.7)及び(9.8)は違った形となる。

Hendrick (1944) は完全母集団は NM 要素を含む単一の大標本抽出単位と考える事を指適した。もしも公式(9.6)が成立すれば、 $S = A(NM)^g$ 。この考え方の利益は A 及び g が M の一つの値しか用いられない様な調査に対しての資料から推定出来るという事である。推定量を導く二つの公式は

$$\begin{aligned} \log S_w^2 &= \log A + g \log M \\ \log S^2 &= \log A + g \log (NM) \end{aligned}$$

(9.7) から S_b^2 に対する公式は

$$S_b^2 = \frac{AM^g \{ (NM-1)N^g - N(M-1) \}}{N-1}$$

この方法は公式(9.6)の正確さについてのチェックを提供しない。公式は M の小さい値に対して十分成立するが、 NM の如く大きい値に対しては失敗である。この場合はより一般的な公式(9.7)及び(9.8)が用いられなければならない。

公式(9.6)は一般法則としてよりはむしろ方法論の例として提供される。同様な問題に直面する読者は、公式の如何なる型がその資料に最もよく用いられるかどうかを作って検定しなければならない。ある場合には $\log S_b^2$ は M の簡単な函数であるかもしれない。

§ 9.6 費用函数

a Cost function.

広汎な調査では調査費用の性質は最適単位決定の際に重要な役割を演ずる。費用の因子の役割の例として我々は、大単位が隣接農場の集落である農場調査に対して Jessen (1942) により述べられた費用函数を述べてみよう。

調査費用の二つの成分が区別されている。 C, Mn なる成分は要素(農場)の総数と共に直接に変化する費用からなる；従って $C,$

は訪問及び集落内を農場から農場へ旅行する費用を含む。

第二の成分 $C_2\sqrt{n}$ は 集落内を旅行する費用を測る。地図上での検定で、母集団を一定とした場合この費用は近似的に集落の数の平方根と共に変化する事が分る。

従って総費用は

$$C = C_1 M n + C_2 \sqrt{n} \quad (9.9)$$

単純無作為抽出と仮定し、f.p.c. を無視すると、要素 $\bar{y}$ 当り、平均値の分散は S^2/nM である。(9.8) からこれは次のものに等しい。

$$V(\bar{y}) = \frac{S^2 - (M-1)AM^{g-1}}{n} \quad (9.10)$$

単位の最適の大きさを決定するため、 M を見出すと同時に一定の C に対して V を最小にする様な n を見出す。一般解はその数値問題への適用はさして困難ではないが複雑である。

ある操作によって我々は最適 M を与える方程式を得る事が出来る。先づ最初に、方程式 (9.9) を $\sqrt{n}$ についての二次式として解くこれは

$$\frac{2C_1 M \sqrt{n}}{C^2} = \left(1 + \frac{4C_0 M}{C_2^2}\right)^{1/2} - 1 \quad (9.11)$$

最小にさるべき方程式は

$$C + \lambda V = C_1 M n + C_2 \sqrt{n} + \lambda V$$

微分して $2V/2n = -V/n$ なる事に注意すると次の方程式を得る。

$$n: \quad C_1 M + \frac{1}{2} C_2 n^{-1/2} = \frac{-\lambda 2V}{\partial n} = \frac{\lambda V}{n} \quad (9.12)$$

$$M: \quad C_1 n = - \frac{\lambda 2V}{\partial M}$$

(9.13) を λ を消去するため (9.12) で割る。すると

$$\frac{n}{V} \frac{\partial V}{\partial M} = \frac{-C_1 n}{C_1 M + \frac{1}{2} C_2 n^{-1/2}} \quad \text{即ち}$$

$$\frac{M}{V} \frac{\partial V}{\partial M} = - \frac{1}{1 + \frac{C_2}{2C_1 M \sqrt{n}}} \quad (9.14)$$

もしも (9.11) からの $\sqrt{n}$ を代入すると、簡単にして次のものを得る

$$\frac{M}{V} \frac{\partial V}{\partial M} = \left(1 + \frac{4C_0 M}{C_2^2}\right)^{-1/2} \quad (9.15)$$

この方程式の左辺を全部書き出し、両辺の符号を変えると

$$\frac{AM^{g-1}\{gM - (g-1)\}}{S^2 - (M-1)AM^{g-1}} = 1 - \left(1 + \frac{4C_0 M}{C_2^2}\right)^{-1/2}$$

この方程式は最適 M を与える。左辺は分散函数のみによっているから費用の因子を含んでいない。答えられた母集団に対して、これが一定の分散函数をもっているとする、最適 M は

$$\text{量} \quad C_0 M / C_2^2$$

が一定である様に費用因子が変化する様に反応する。

さて、もしも訪問の長さが増せば C_1 は増加するが一方旅費がやすくなるか又は与えられた面積に於て農場がより密度が高くなれば C_2 は減少する。これ等の事実は単位の最適の大きさが次のときにより小さくなるという結論を導く：

訪問の長さが増す。

旅費がよりやすくなる。

要素(農場)がより密集している。

所要全費用(C)が増す。

結論は費用函数の型の一つの結果であり、違った方程式では再検討を要すであろう。最適単位は母集団の一定の特性でなくて調査の型及び資金の水準によるという事を示している。

§ 9.7 内部集落相関の項で表わした分散

Variance in terms of intra-cluster correlation:-

抽出単位がM個の要素の集落である時には、分散公式は時に同じ集落に於ける要素間相関係数Pの項で表わされる事がある。この条口への例はすでに組織抽出で与えた(§8.3)

y_{ij} をしてi番目の単位j番目の要素に対する観測値とし、 $\bar{y}_i$ をして単位の統計としよう。集落抽出では我々は、次の二種類の平均値を明らかに区別する必要がある。即ち単位当り平均値 $\bar{Y} = \sum y_{ij}/N$ 及び要素当り平均値 $\bar{y} = \sum y_{ij}/NM = \bar{Y}/M$ 。要素間分散は

$$S^2 = \frac{\sum (y_{ij} - \bar{Y})^2}{NM - 1}$$

内部集落相関係数Pは次の様に定義される。

$$P = \frac{\sum_i \sum_k (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{NM(M-1)S^2/2}$$

分母の数値項は分子の和に於ける積 $NM(M-1)/2$ の数である。要素当り推定平均値 $\bar{y}$ の分散を考えよう。n個の完全集落の単純無作為標本では $\bar{y} = \bar{Y}/M$ であるから

$$V(\bar{y}) = \frac{1}{M^2} V(\bar{Y}) = \frac{(N-n)}{N^2 M^2} \frac{\sum (y_{ij} - \bar{Y})^2}{(N-1)} \quad (9.16)$$

併し

$$(y_{ij} - \bar{Y}) = (y_{i1} - \bar{Y}) + (y_{i2} - \bar{Y}) + \dots + (y_{iM} - \bar{Y})$$

すべてのN集落上で、平方して加えると

$$\sum (y_{ij} - \bar{Y})^2 = \sum_i \sum_j (y_{ij} - \bar{Y})^2 + 2 \sum_i \sum_{j < k} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})$$

$$= (NM - 1) S^2 + NM(M-1) P S^2$$

$V(\bar{y})$ に対する公式(9.16)に代入して

$$V(\bar{y}) = \frac{(N-n)}{N} \frac{1}{NM} S^2 \left\{ \frac{(NM-1)}{(N-1)M} + \frac{N(M-1)P}{(N-1)} \right\}$$

Nが大であれば、これは次の如くなる。

$$V(\bar{y}) = \frac{(N-n)}{N} \frac{1}{NM} S^2 \{ 1 + (M-1)P \} \quad (9.17)$$

もしもnM要素の単純無作為標本がとられるとすれば $V(\bar{y})$ に対する公式は括弧内の項に対して以外は(9.17)と同じである。

因子 $1 + (M-1)P$

は抽出単位として要素の代りに集落を用いる事によってどの位分散が変化するかを示す。もしも $P > 0$ であれば集落はあたえられた標本の大きさでは正確度が少ない。もしも $P < 0$ であればこの様な事はしばしば起ることであるが、集落はより正確である。

§ 9.8 割合に対する集落抽出

Cluster sampling for proportions:-

同じ方法は割合に対する集落抽出にも適用される。集落に於てM要素が二つの組に分類出来るとし $p_i = a_i/M$ はi番目の集落でのcなる組に属する割合としよう。n個の集落の単純無作為標本がとられ、標本に於ては観測値 p_i の平均値 $\bar{p}$ が母集団の割合Pの推定量として用いられる。

我々は $V(P)$ を見出すためには二項理論を用いる事は出来なく、 p_i に対する連続変量に対する公式を適用しなければならぬ事を思い出さなくてはならない(§3.2)これは

$$V(P) = \frac{(N-n)}{Nn} \frac{\sum_{i=1}^N (p_i - P)^2}{N-1} = \frac{(N-n)}{N^2} \sum (p_i - P)^2$$

更にもしも我々が nM 要素の単純無作為標本をとればPの分散はもしもNが大であるとすれば次の如き二項理論によって得ら

れる。

$$V_{bin}(P) = \frac{(NM - nM)}{NM - 1} \frac{PQ}{nM} \div \frac{(N - n)}{N} \frac{PQ}{nN}$$

従って 因子

$$\frac{V(P)}{V_{bin}(P)} \div \frac{M \sum (p_i - P)^2}{N P Q} \quad (N \text{が大}) \quad (9.18)$$

は集落の使用による分散に於ける相対的变化を示す。この因子の数値は集落抽出で標本の大きさの初めの推定量を作る助けとなるものである。先づ第一に必要な標本の大きさは二項公式によって推定されそれから、集落抽出で必要とされる大きさを示す因子を乗ずる。その例は Corn-field (1951) を参照されよ。

もしも集落の大きさ M_i が変化するとすれば、推定量 $P = \sum a_i / \sum M_i$ は比推定量である。その分散は次の公式により近似的に求まる。(9.19)

$$V(P) \div \frac{N - n}{N n M} \frac{\sum M_i^2 (p_i - P)^2}{N - 1}$$

此処に $\bar{M} = \sum M_i / N$ は集落の平均の大きさである。

もしも標本が nM 要素の単純無作為標本と比較されるとすれば (9.18) の一般化したものとして次を得る。

$$\frac{V(P)}{V_{bin}(P)} \div \frac{\sum M_i^2 (p_i - P)^2}{N \bar{M} P Q} \quad (9.19)$$

連続変数の場合と同様、集落の大きさと集落内分散との関係は、 $\bar{M}$ の函数として (9.18) 及び (9.19) 式に於ける因子を表わすか又は集落内分散と $\bar{M}$ との間の関係を求める事によって求める事が出来る。もしも我々が C なる組に属する単位を 1 として他のものを 0 として割当てると、一定 M に対する分散公式の基礎的な分析は

$$N M P (1 - P) = M \sum (p_i - P)^2 + M \sum p_i (1 - p_i)$$

$$\text{全 } S.S. = \text{集落間 } S.S. + \text{集落内 } S.S.$$

この関係から集落内の平均平方が計算出来 M の函数としてプロット出来る。Mc Vay (1947) は如何にしてこの分析が最適集落の大きさを求めるのに用いる事が出来るかを述べている。

§ 9.9 単位の大きさの尺度

Measures of the size of a Unit :-

多くの調査に於ては単位はその大きさが変化する。適当に定義された家又は住居はしばしば人間母集団の調査に於ては便利な単位であるが、何処でもひととふ人以上の間の人を含んでゐる。この種の例では我々は、それが含む要素の数 M_i として集落の大きさを定義する事が出来る。

如何にしてその大きさが測定さるべきかは判然としないが、明らかに単位同の大きさに差が存在する他の母集団がある。農場、銀行及び料理店はその例である。大きい農場もあれば小さい農場もある。併し農場の大きさの尺度として我々は、全面積か又は穀物に利用出来る全面積か農場生産物の全部の値か又なを他の量を計画する

どの種類の大きさの尺度が抽出者にとって有用であるか？ ある項目 y_i が農場の単純無作為標本の各々で測定されると假定しよう。抽出者は y_i は分散が大きいのではないかと恐れている。

何故ならば、年々 y_i の大きい値を答える。農場がある一方にえず小さな値を答える他の農場があるからである。必要なるものは調査前に得られて、 y_i の値が大きい小さいかを予測する任意変数 x_i である。従って大きさの尺度を見出す問題は、この項目のある意味で y_i と高度の相関をもつ任意変数を見出す問題となる。全面積、耕地面積及び全部の値間の選択は平均して三つの値の中、調査に含まれる項目と最も相関の高いものを検定する事によって行われる。今我々は我々の関心は特別な定義にあるよりはむしろ、大きさの良質な尺度の一般的概念にあるので、この

平均的な相関が如何にして計算されるかは論じない。

同じ調査に対して大きさの最良の尺度は項目如何による。もしも項目が最近の調査で調べられたとすれば、最もよい任意変量と、以前のセンサスでこの項目の値である事が分る事がしばしばある。この様な場合大きさの一般的尺度は各項目に対して別個な尺度よりは劣っている。併し利用出来る以前のセンサスの資料は、新しい調査であるべき項目を含まなくて数個の一般的な大きさの尺度を与える。大きさの一般的尺度が与えられると、前章で論じた二つの方法の中一つを利用出来る。

我々は大きさによって層化する y_i の分散は通常 x_i と共に増加するから、抽出比は通常層毎に違ふ。最大の単位で層を完全調査する事をおすすめする。オニの方法は任意変量としての x_i との比推定又は回帰推定を利用する事である。これはある因子をコントロールするのに用いるべき層化を与える。二つの方法の組合せは時には効果が多い。

§ 9.10 大きさに比例した確率で以ての標本抽出

Sampling with probability proportional to size

Hansen 及び Hurwitz (1943) により暗示された第三の方法は、標本が抽出されるとき大きな単位に対して高い確率を割当てる事である。この方法は副次抽出(オニ11章)を用いる調査に主として利用される事を見出したが又、この問題にも応用出来る。大きさに比例した確率抽出は7つの単位をもつ小母集団に関する次の例に示される。

単位	大きさの尺度	尺度の和	割当範囲
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

大きさの尺度の累積和が作られる。単位を送るには、1と30の間の乱数を抽く。これが19であると假定しよう。和に於て19は16から21を含む4なる単位に属する。この抽出方法では単位が選ばれ、確率は単位に割当てられた大きさの尺度に比例してゐる。もしオニ番目の単位が選ばれるとすれば、その方法は1と30との間の新しい乱数で以ておきかえる。併し、前の実行と違ふのはオニ番目に4なる単位の選択を許すことである。おきかえを1以上で大きさに比例した選択確率を保つためにはおきかえを許す抽出が必要である。これは極限の場合 $n=7$ によって分る。

もしも選択がおきかえなしに行われるとすれば、すべての単位はたとえ我々が大きさに比例した確率での選択方法を行ったとしても自動的に送られるであろう。これが1と7の間の値に対しておきかえなしの選択は等しい確率と大きさに比例した確率の中間の確率になる。

規則としておきかえを許す抽出はおきかえを許さない抽出より正確度は劣る。併し n/N が小さいとき同じ単位が標本に二度表われる確率は小さいからおきかえを許す抽出は実際にはおきかえを許さない抽出と同じである。

§ 9.11 任意確率をもつ選択理論

Theory for selection with arbitrary probability:-

我々は先づ確率のシステムでの選択の下に二三の一般公式を作ろう。 x_i をして i 番目の単位の選択確率としよう。此処に x_i は1に加える正の数の組である。大きさ n なる標本が前節で述べたおきかえを許して送られる。

t_i をして i 番目の単位が特別な標本の大きさ n に於て表われる回数としよう。此処に t_i は $0, 1, 2, \dots, n$ なる値をとる。母集団に於けるすべての N 単位に対して t_i の結合度数分布を考えよう。

標本を抽出する方法は n 個のボールが N 個の箱の中に投げこまれる即ち各投入に於て i 番目の箱に入るボールは z_i である確率である。標準の確率問題と同じである。従って t_i の結合分布は次の多項式で表わされる。

$$\frac{n!}{t_1! t_2! \dots t_n!} z_1^{t_1} z_2^{t_2} \dots z_n^{t_n}$$

多項式に対して t_i の度数の次の特性はよく知られている。

$$E(t_i) = n z_i : V(t_i) = n z_i (1 - z_i) : \text{Cov}(t_i, t_j) = -n z_i z_j$$

この選択システムの下で標本平均を $\bar{y}_z$ で表わす。平均値は標本に於ける n 個の単位に対する項目の値を加えて n で割る通常の方法で計算される。これはもしも一つの単位が二度抽出されたとすると、次の項目の値は $\bar{y}_z$ の計算には重みを 2 づけることを意味する。

定理 9.2

もしも $\bar{y}_z$ が z_i に比例した確率で引て抽出された標本平均とすれば

$$E(\bar{y}_z) = \sum_{i=1}^N z_i y_i = \bar{y}_z \quad (9.20)$$

$$V(\bar{y}_z) = E(\bar{y}_z - \bar{y}_z)^2 = \frac{1}{n} \sum_{i=1}^N (y_i - \bar{y}_z)^2 \quad (9.21)$$

(証明)

標本に於ける如何なる単位も、それが抽出された回数 t_i の重みをうけるから次の様にかける。

$$\bar{y}_z = \frac{1}{n} (t_1 y_1 + t_2 y_2 + \dots + t_N y_N) = \frac{1}{n} \sum_{i=1}^N t_i y_i$$

母集団に於けるすべての y_i がこの表示で表われてゐるのに注意。くりかえしの抽出では t_i は無作為変数であるが y_i は一定数の組である。従つて

$$E(\bar{y}_z) = \frac{1}{n} \sum_{i=1}^N y_i E(t_i) = \sum_{i=1}^N z_i y_i = \bar{y}_z$$

更に

$$V(\bar{y}_z) = \frac{1}{n^2} \left\{ \sum_{i=1}^N y_i^2 V(t_i) + 2 \sum_{i < j} y_i y_j \text{Cov}(t_i, t_j) \right\}$$

$$= \frac{1}{n} \left\{ \sum_{i=1}^N y_i^2 z_i (1 - z_i) - \sum_{i < j} y_i y_j z_i z_j \right\}$$

$$= \frac{1}{n} \left\{ \sum z_i y_i^2 - \left(\sum z_i y_i \right)^2 \right\}$$

$$= \frac{1}{n} \sum z_i (y_i - \bar{y}_z)^2$$

定理 9.3

標本からの $V(\bar{y}_z)$ の不偏推定量は

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

(証明)

通常の数式的等式により次の様にかける。

$$\sum_{i=1}^n (y_i - \bar{y}_z)^2 = \sum_{i=1}^n (y_i - \bar{y}_z)^2 - n(\bar{y}_z - \bar{y}_z)^2$$

依つて

$$E \sum_{i=1}^n (y_i - \bar{y}_z)^2 = E \sum_{i=1}^n (y_i - \bar{y}_z)^2 - n V(\bar{y}_z)$$

之は $V(\bar{y}_z)$ の定義により右辺の二番目の項の平均値は $-n V(\bar{y}_z)$ であるからである。

変量 t_i を導入すると

$$E \sum_{i=1}^n (y_i - \bar{y}_z)^2 = E \sum_{i=1}^N t_i (y_i - \bar{y}_z)^2 - n V(\bar{y}_z)$$

今 y_i を一定数、 z_i を無作為変量と考えよう。 $E(z_i) = nZ_i$ であるから定理 9.2 により

$$\begin{aligned} E \sum_{i=1}^n (y_i - \bar{y}_2)^2 &= n \sum_{i=1}^n Z_i (y_i - \bar{y}_2)^2 - nV(\bar{y}_2) \\ &= n^2 V(\bar{y}_2) - nV(\bar{y}_2) = n(n-1)V(\bar{y}_2) \end{aligned}$$

これで完全に証明された。

標本からの分散の推定に於て我々は z_i によって y_i を重みづけない。何故ならば、この重みづけは標本の選択に於て既に等入されているからである。これ等の結果は今や単位が等しくない大きさである場合の母集団統計の推定に用いられる。

定理 9.4

大きさ n なる標本が、大きさの尺度 $z_i = M_i / \sum M_i$ に比例した確率で抽出される。標本に於ける単位に対する項目総計は $y_1, y_2, \dots, y_n$ である。比 y_i/z_i に抽出はおきかえを以て行われるから、同じ単位は一度以上表われない。母統計 Y の推定量として我々は $\hat{Y}_{PPS}$ とかく (*Probability Proportional to size*) 推定に

$$\hat{Y}_{PPS} = \frac{1}{n} \left(\frac{y_1}{z_1} + \frac{y_2}{z_2} + \dots + \frac{y_n}{z_n} \right) \quad (9.23)$$

そのとき $\hat{Y}_{PPS}$ は分散

$$V(\hat{Y}_{PPS}) = E(\hat{Y}_{PPS} - Y)^2 = \frac{1}{n} \sum_{i=1}^n Z_i \left(\frac{y_i}{z_i} - Y \right)^2 \quad (9.23)$$

をもつ Y の不偏推定量である。

(証明)

変量 y_i/z_i は定理 9.2 を適用せよ。そのとき

$$E(\hat{Y}_{PPS}) = \sum_{i=1}^n Z_i \left(\frac{y_i}{z_i} \right) = \sum_{i=1}^n y_i = Y$$

$$V(\hat{Y}_{PPS}) = \frac{1}{n} \sum_{i=1}^n Z_i \left(\frac{y_i}{z_i} - Y \right)^2$$

結果は如何なる標本の大きさに対しても正確である。抽出はおきかえを以て行われるから分散は f.p.c. を含まない。推定量 $\hat{Y}_{PPS}$ は迅速な計算には不適当である。何故ならば標本に於ける各単位上での比、 y_i/z_i を見出さねばならぬからでなくは用いられない様である。

この推定量では大きさの最適尺度はこれに対して z_i が

$$V(\hat{Y}_{PPS}) \frac{1}{n} \sum Z_i \left(\frac{y_i}{z_i} - Y \right)^2 = \frac{1}{n} \left\{ \sum \frac{y_i^2}{z_i} - Y^2 \right\}$$

を最小にする数 M_i の組である。

もしも y_i がすべて正であれば $M_i \propto y_i$ のとき $V(\hat{Y})$ が 0 となる事を見出す事は容易である。従つて大きさの最良の尺度は単位での項目総計 y_i である。この結果はもしも y_i が前以て分つてをれば標本が不必要であるから実際的には興味のない事である。

併しこの結果はもしも項目が相対的に期間を通して安定してをれば y_i の最近の利用出来る以前の値が採用すべき最良の大きさの尺度である事を暗示してゐる。

§ 9.12 比推定との比較

Comparison with the ratio estimate :-

$\hat{Y}_{PPS}$ と等確率での抽出からの比推定量 $\hat{Y}_R$ との間の比較は、実際的に興味のある問題である。 $Y(\hat{Y}_R)$ は大標本の場合のみ分つてから比較はこの場だけに限られる。

比推定に対して

$$\hat{Y}_R = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i} \left(\sum_{i=1}^N M_i \right)$$

定理 6.1 により比較推定の近似分散は

$$V(\hat{Y}_R) = \frac{N(N-n)}{N(N-1)} \sum_{i=1}^N (y_i - R x_i)^2$$

$x_i = M_i : R = Y / (\sum M_i)$ であるからこれは次の様にかける。

$$\begin{aligned} V(\hat{Y}_R) &= \frac{N(N-n)}{N(N-1)} \sum_{i=1}^N \left(y_i - \frac{M_i Y}{\sum M_i} \right)^2 \\ &= \frac{N(N-n)}{N(N-1)} \sum_{i=1}^N (y_i - Y Z_i)^2 \quad (9.24) \end{aligned}$$

P.P.S 抽出では不偏推定量に対して (9.23) から

$$V(\hat{Y}_{PPS}) = \frac{1}{n} \sum_{i=1}^N Z_i \left(\frac{y_i}{Z_i} - Y \right)^2 = \frac{1}{n} \sum_{i=1}^N \frac{1}{Z_i} (y_i - Y Z_i)^2 \quad (9.25)$$

n/N を小さいと仮定すると、二つの比較出来る量は

$$n V(\hat{Y}_R) = N \sum (y_i - Y Z_i)^2 \quad (9.26)$$

$$n V(\hat{Y}_{PPS}) = \sum \frac{1}{Z_i} (y_i - Y Z_i)^2 \quad (9.27)$$

ある母集団に対しては比推定はすぐれているし、他の母集団に対しては P.P.S 推定量がすぐれている。

私にはよりよい推定量を予測する簡単な法則は分らない。公式 (9.26) 及び (9.27) は母集団資料が利用出来るときに比較をするのに用い得る。

一つの結果を示しておこう。

$$y_i = Y Z_i + e_i$$

と仮定しよう。此処に e_i は確率の意味で Z_i と独立である。 Z_i が一定である組では

$$E(e_i) = 0 ; \quad E(e_i)^2 = \sigma Z_i^g \quad (g > 0)$$

と仮定する。

もしも $g = 1$ であればこのモデルは比推定が最良線型不偏推定量である条件を満足する (8.6.8)

(9.28) から

$$n V(\hat{Y}_R) = N \sum_{i=1}^N e_i^2 = N^2 E(e_i^2) = \alpha N^2 E(Z_i^g)$$

此処に平均は Z_i のすべての値上でとられる。

同様に (9.27) から

$$\begin{aligned} n V(\hat{Y}_{PPS}) &= \sum_{i=1}^N \frac{1}{Z_i} (y_i - Y Z_i)^2 = N E \left(\frac{e_i^2}{Z_i} \right) \\ &= \alpha N E(Z_i^{g-1}) \end{aligned}$$

依ってもしも $E(Z_i^g) - \frac{1}{N} E(Z_i^{g-1}) > 0$ であれば

$$V(\hat{Y}_R) > V(\hat{Y}_{PPS})$$

Z_i は 1 に加えるから

$E(Z_i) = 1/N$ であるから不等式は次の様にかける。

$$E(Z_i Z_i^{g-1}) - E(Z_i) E(Z_i^{g-1}) > 0$$

この表示は変量 Z_i 及び Z_i^{g-1} の共分散である。もしも $g = 1$ であれば共分散はなくなる。もしも $g > 1$ であるは変量 Z_i は 0 と 1 の間にあるから共分散は正である。もしも $0 < g < 1$ であれば共分散は負である。

要約すると我々は y_i と Z_i の間の関係は原点を通る直線であると仮定した。もしもこの直線まわりの y_i の分散が Z_i よりも早く増大すると P.P.S 抽出はより正確である。 $V(e_i)$ と Z_i

の関係を表わす資料を報告した研究から見ると、分散は αZ_1 と αZ_1^2 との間の何処かである程度増大した。

この比較は二つの方法に於て等しい大きさの標本で行われた。もしも小単位からよりも大単位から資料を得るのに費用が多くかゝれば、比較はより大きな単位に集中する傾向がある P. P. S 抽出は利益となる様に偏ってくる。更に比推定は計算力より簡単である。

§ 9.13 層化抽出の拡張

Extension to stratified sampling: -

大きさに比例した確率抽出は層化が大きさ以外のある変量によつて行われた場合有用である様である。もしも各層内の標本が小さく、総標本があまり小さくなければ、比較度に対する利用出来る分散公式は幾分疑わしく比推定の公式はいちじるしく偏倚がある事が分った(§ 6.11)

P. P. S 抽出では推定総計は個々の層からの推定量の和である。即ち

$$\hat{Y}_{PPS} = \sum_h \frac{1}{n_h} \sum_{i=1}^{n_h} \left(\frac{y_{hi}}{Z_{hi}} \right)$$

前の定理から次を得る。

$$V(\hat{Y}_{PPS}) = \sum_h \frac{1}{n_h} \sum_{i=1}^{n_h} Z_{hi} \left(\frac{y_{hi}}{Z_{hi}} - Y_h \right)^2 \quad (9.28)$$

$$V(\hat{Y}_{PPS}) = \sum_h \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} \left\{ \frac{y_{hi}}{Z_{hi}} - \left(\frac{y_{hi}}{Z_{hi}} \right) \right\}^2 \quad (9.29)$$

第10章 単位大きさが等しい場合の副次抽出法

Subsampling with units of equal size

§ 10.1 緒言

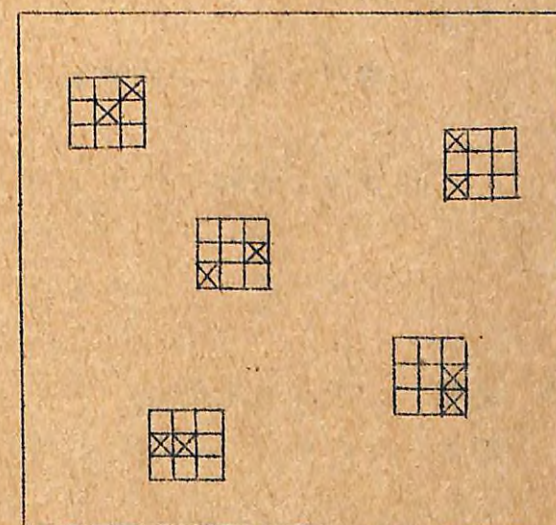
Introduction

母集団に於ける各単位が多くより小さい単位即ち要素に分ける事が出来ると假定しよう。九個の単位をもつ標本が選ばれた。もしも選ばれた単位内の要素が同様な結果を具えれば、それ等をすべて測定する事は不経済である様に思われる。通常の実行は選ばれた単位に於ける要素の標本を選んで測定する事である。この方法は、単位は完全には測定されなくて、それそのものが抽出されるから副次抽出法と呼ばれる。Mahalanobis によれば標本が二段階でとられるから他の言葉で二段抽出法 (two-stage sampling) と呼ばれる。まず第一に一次単位 (primary unit) と呼ばれる単位の標本が選ばれ、次の段階で各選ばれた一次単位から要素の標本が選ばれる。

副次抽出は標本調査の直接の力の及ばぬかぎり広い多くの応用面をもっている。その過程が少数の資料で遂行出来る化学、物理又は生物学的検定を含むときは何時でも、これはそのものが標本であるより大きな量からの副次標本として抽出される様である。

この章では、我々は各単位は要素の同じ数 M 個を含みその中の m 個の単位が副

10.1 図 二段抽出の図的表示
($N=81, n=5, M=9, m=2$)



⊗ は標本に於ける要素を表はす

次抽出されたとき選ばれる最も簡単な場合を考えよう。

$M=9$ 及び $m=2$ なる二段抽出の図的表示が 10.1 図に示されている。

二段抽出の主な利益は一段抽出よりも融通性があるという事である。 $m=M$ のとき一段抽出となるが、これは m に対して最良の選択でなければ我々はより有効と思えるあるより小さい値をとる機会をもつ。通常の様には論議は統計的正確度と費用との間のバランスの問題となる。同じ単位内の要素が非常に近似している時には、正確度の考え方は m の小さい値を暗示する。一方それを副次抽出する際に単位全体を測定すると殆んど同じ位低廉な事がある。例えば単位が家族とかたば一つの解答とかは家族全員についての正確な資料を与える事が出来る。

注意

多段抽出では記号は面倒となる。何故なら数個の平均値即ち一次単位当り平均値、副次単位当り平均値等を区別する事が必要であるからである。

基礎的な組織は二段抽出に対して次の如くである。記号 y_{ij} は i 番目の単位で j 番目の要素（副次単位）に対して得られた観測値を表わす。前の様に y 及び Y は標本及び母集団の天々の総計を表わす。

$$i \text{ 番目の単位での要素当りの標本平均} = \bar{y}_i = \sum_{j=1}^m y_{ij}/m$$

要素当り全体の標本平均

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

一次単位当りの標本平均

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

同様な定義が母平均 $\bar{Y}_i$, $\bar{Y}$ 及び $\bar{Y}$ に対しても成立する。記号の本質は一つの棒は一段での平均を表わし、二つの棒は二段での平均を表わす。（もしもあれば）添字は何が一定であることを表わす、即ち i が一定である。 y_{ij} の平均は $\bar{y}_i$ であり、単位上での $\bar{y}_i$

の平均は $\bar{y}$ である。

亦 N 及び n は一次単位の数に対して、 M 及び m は一次単位当りの副次単位の数に対して用いられる事に注意されよ。ある学者は、この役割をかえて用いているから、参考文献を引用するときには記号に関して充分注意する事が肝要である。

§ 10.2 基礎理論

Elementary theory:-

副次抽出の理論は農事試験に於けるプロットの抽出と関連して発達した。これ等の応用に於て抽出比 n/N 及び副次抽出比 m/M 共に通常小さく f , p , c は無視出来る。その結果の理論は卓越しており多くの応用に適當であるからまづそれを述べよう。実際には基礎理論は n/N が無視出来る。例えば 0.05 以下であるという事だけ必要である。この事は正確な理論が示された時に分るであろう。

i 番目の単位の j 番目の要素に於ける観測値 y_{ij} は次の形をとると假定する

$$y_{ij} = \bar{Y} + u_i + w_{ij} \quad (10.1)$$

此処に項 u_i は単位に関連した成分を表わし単位に於けるすべての要素に対して一定である。項 w_{ij} は単位内の要素毎に変わる成分を表わす。変量 u_i 及び w_{ij} は確率の意味で独立でありその平均値は 0 である。変量 u_i は分散 S_u^2 (unit に対して u) をもち、 w_{ij} は分散 S_w^2 (within に対して w) をもつ。

N 及び M の値は無限と假定する。単位は母集団から無作為に選ばれ、要素は単位から無作為に選ばれる。モデルの結果として要素当り標本平均値 $\bar{y}$ は $\bar{Y}$ の不偏推定量である事を示すのは容易である。

定理 10.1

このモデルでは要素当り標本平均 $\bar{y}$ の分散は

$$V(\bar{y}) = \frac{S_u^2}{n} + \frac{S_w^2}{nm} \quad (10.2)$$

(証明)

(10.1) から

$$\bar{y} = \bar{Y} + \frac{u_1 + u_2 + \dots + u_n}{n} + \frac{w_1 + w_2 + \dots + w_{nm}}{nm}$$

従って無限母集団からの平均値の分散に対する公式により

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2 = \frac{S_u^2}{n} + \frac{S_w^2}{nm}$$

n が増すと単位内分散 S_w^2 からの貢献は少なくなり、 n が増すと分散の二つの成分からの貢献が少なくなる事に注意されよ。

定理 10.2

$V(\bar{y})$ の不偏推定量は次の如く標本から得られる

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

右辺の始めの項は計算に最も便利な形である。

(証明)

代数的等式により

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

方程式 (10.1) から標本に於ける n 番目の単位に於ける m 個の要素上に平均して

$$\bar{y}_i - \bar{Y} = u_i + \frac{(w_{i1} + w_{i2} + \dots + w_{im})}{m}$$

$$\text{依って } E(\bar{y}_i - \bar{Y})^2 = S_u^2 + \frac{S_w^2}{m}$$

$$E \sum_{i=1}^n (\bar{y}_i - \bar{Y})^2 = n S_u^2 + \frac{n S_w^2}{m}$$

定理 10.1 により

$$n E(\bar{y} - \bar{Y})^2 = n V(\bar{y}) = S_u^2 + \frac{S_w^2}{m}$$

従って差引く事により

$$E \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 = (n-1) \left(S_u^2 + \frac{S_w^2}{m} \right) = n(n-1) V(\bar{y})$$

証明終り

推定分散は副次抽出計画でない場合と同様単位間の変動のみから計算される事に注意されよ、たとえ推定分散に於て明らかであったにしても単位内分散は定理 10.1 に於ける $V(\bar{y})$ での S_w^2/nm 項から明らかな様に考慮に入れられている。

§ 10.3 他の副次抽出比に対する分散の予測

Prediction of the variance for other sampling ratios: -

数式的モデルから我々はそれに対して我々が資料を得る調査に於て用いたものと違った抽出比及び副次抽出比に対して $\bar{y}$ の分散を予測出来る。この情報はいずれも同じ母集団の型についての将来の標本の計画に役立つものである。

最初の標本が n 個の単位及び単位当り m 個の要素をもつと假定しよう。もしもこれ等の数が夫々 n' 及び m' に変化すれば、定理 10.1 によって標本平均の分散は次の如くなる。

$$V(\bar{y}') = \frac{S_u^2}{n'} + \frac{S_w^2}{n'm'} \quad (10.3)$$

この公式を用いるためには S_u^2 及び S_w^2 の標本推定量が必要である。これ等は 10.1 表に示されている標本の分散分析から得られる。

10.1 表 標本分散分析 (要素を基礎としての)

	d.f	m.S	次のものの推定量
単位間	$(n-1)$	$S_b^2 = \frac{m \sum (\bar{y}_i - \bar{y})^2}{n-1}$	$S_u^2 + m S_w^2$
単位内要素間	$n(n-1)$	$S_w^2 = \frac{\sum \sum (y_{ij} - \bar{y}_i)^2}{n(n-1)}$	S_w^2

右の欄は平均平方 S_b^2 及び S_w^2 が不偏である量を示す、 S_b^2 に対する結果は $S_b^2 = nm V(\bar{y})$ であるから定理 10.2 から直ちに導かれる、 S_w^2 に対する結果は容易に証明される。従って

S_u^2 の不偏推定量は

$$S_u^2 = \frac{S_b^2 - S_w^2}{m}$$

従って $V(\bar{y})$ の不偏推定量は

$$v(\bar{y}) = \frac{S_b^2}{n'} + \frac{S_w^2}{n'm'} = \frac{1}{n'} \left[\frac{S_b^2}{m} + S_w^2 \left(\frac{1}{m'} - \frac{1}{m} \right) \right], (10.4)$$

(例)

King と Jebe (1940) は North Dakota の小麦畝の標本抽出 (1938) に於ける次の如き分散分析を報告している。

二つの標本が各畝からとられた、そして畝は地区によって層化されている。

10.2 表 小麦収量の分散分析 (エーカー・ブッシュ)

	m.S
地区内畝間	$S_b^2 = 180$
畝内副次単位間 (要素)	$S_w^2 = 38$

畝は無作為に選ばれなかったけれども、次の手段によって面積の良好な事項をそえる様に計画されたから、畝間平均平方は、畝の無作為標本から得られる分散の 過大推定量である。このディスターバンス及び畝の大きさに於ける変動の影響は無視しよう。

我々は次のものによって如何に標本平均の分散が影響されるかを検定しよう。

- (i) 畝当り二つの副次標本で以て畝の数を倍にする
- (ii) 畝の数は変えないが畑当り四つの副次標本をとる。
- (iii) 畝の数は変えないが、畑を完全収穫する。

n を最初の標本の畑の数としよう。公式 (10.4) から次の標本平均の推定分散が得られる。

($m=2$ なる事に注意)

最初の標本: ($n'=n, m'=2$) $v = \left(\frac{1}{n}\right) \left(\frac{180}{2}\right) = \frac{90}{n}$

(i) の場合: ($n'=2n, m'=2$) $v_i = \left(\frac{1}{2n}\right) \left(\frac{180}{2}\right) = \frac{45}{n}$

(ii) の場合: ($n'=n, m'=\infty$) $v_{ii} = \left(\frac{1}{n}\right) \left(\frac{180}{2} - \frac{38}{2}\right) = \frac{85}{n}$

(iii) の場合: ($n'=n, m'=\infty$) $v_{iii} = \left(\frac{1}{n}\right) \left(\frac{180}{2} - \frac{38}{2}\right) = \frac{85}{n}$

(iii) の場合の完全収穫は標本でのすべての畑からすべての可能な標本をとる事と同じであると假定される。副次標本の大きさは畝の大きさに比して非常に小さいから、これは $m'=\infty$ なる事を意味する。

(ii) の場合及び (iii) の場合は畝の数は一定のまま抽出比を増加すると分散に於て最も適当な減少を生ずる事を示している。もしも正確度を著しく増加したければ、畝の数を増さねばならない。

§ 10.4 一般理論

general theory:—

さて我々は $\frac{1}{N}$ 及び $\frac{1}{M}$ が小さく数率的モデルが成立するという假定を除いてみよう。なお量 S_b^2 及び S_w^2 の項で分散を表わす事は便利であるが、10.3 表に示す様に完全母集団に対する分散分析からこれ等を作らなければならない。従って S_b^2 及び S_w^2 は分散分析の二欄で述べられた二つの方程式が妥当である様に定義される。これ等の定義は基礎理論の自然の拡張として表わすための一般理論を可能にするから選ばれたのである。

10.3 表 完全母集団の分散分析
(要素を基礎とした)

	d.f	m.S	次のものとして定義される
単位間	$(N-1)$	$M \frac{\sum_{i=1}^N (\bar{y}_i - \bar{y})^2}{N-1}$	$= S_w^2 + MS_b^2$
単位内要素間	$N(M-1)$	$\frac{\sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{y}_i)^2}{N(M-1)}$	$= S_w^2$

ある母集団では量 S_b^2 は負となる; これは同じ単位内の要素が負の相関がある場合に起る。不安な感じは S_w^2 及び S_b^2 の代りに S_{io}^2 及び ρ (単位内相関係数) の項で結果を表わす事によって避

を得る。然しすべての公式は S_u^2 が負である時は正しいまいである。

二段抽出では期待値は抽出されるすべての可能な n 単位の標本で見出されるばかりでなく単位の選ばれた組から抽出されるすべての可能な副次標本上でも見出されなければならない、次々の段階で平均してゆく事はしばしば助けとなる。この目的のため次の二つの記号を導入しよう。

E_n = n 番目の単位からのすべての副次標本上での平均
 E_n = n 個の単位の一組からのすべての副次標本上での平均

もしも選ばれた単位に於て n 個の要素が単純無作為抽出で選ばれとすれば

$$E_n \bar{y}_i = \bar{Y}_i$$

従って

$$E_n \bar{y} = \bar{Y}_n$$

此処に $\bar{Y}_n$ はもしも標本に於て n 個の単位が完全調査されるとすれば得られるであろう平均値を表わす。

更にもしも n 個の単位が亦単純無作為抽出によって選ばれたとすれば

$$E \bar{Y}_n = \bar{Y}$$

これは $\bar{y}$ が $\bar{Y}$ の不偏確定量である事を示す。

定理 10.3

もしも n 個の単位及び各選ばれた単位から n 個の要素が単純無作為抽出で選ばれたとすれば

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2 = \frac{(N-n)}{N} \frac{S_u^2}{n} + \frac{(MN-mn)}{MN} \frac{S_w^2}{mn} \quad (10.5)$$

(証明)

$$\bar{y} - \bar{Y} = (\bar{y} - \bar{Y}_n) + (\bar{Y}_n - \bar{Y}) \quad \text{とかく}$$

両辺を平方して n 個の単位の一組上で平均すれば

$$E_n(\bar{y}) = \bar{Y}_n$$

であるから右辺の積の項はなくなる。

右辺の第一項を考えよう。 n 個の単位の各々は M 個の要素からなる層として考えられる。これ等の単位からの標本は n 個の要素が各層からとれるから、比例層化標本である。従って定理 5.3 の公式即ち層化無作為標本の平均値の分散公式が適用される。即ち

$$E_n(\bar{y} - \bar{Y}_n)^2 = \frac{1}{(nm)^2} \sum_{i=1}^n \frac{M(M-m)}{m} S_{wi}^2$$

此処に S_{wi}^2 は i 番目の単位内分散である。これは又次の如く書ける

$$E_n(\bar{y} - \bar{Y}_n)^2 = \frac{(M-m)}{M} \cdot \frac{1}{mn} \bar{S}_{wn}^2 \quad (10.6)$$

此処に $\bar{S}_{wn}^2$ はこれ等 n 個の単位内平均分散である。

更に n のすべての可能な組上で平均すれば 10.3 表から定義された様に $\bar{S}_{wn}^2$ の平均値は S_w^2 である事は明らかである。従って

$$E(\bar{y} - \bar{Y}_n)^2 = \frac{(M-m)}{M} \cdot \frac{1}{mn} S_w^2 \quad (10.7)$$

方程 (10.5) の右辺の第二番目の項からの貢献は $\bar{Y}_n$ は n 個の単位をもつ単純無作為標本に対して $\bar{Y}_i$ なる値の平均値であるから容易に示す事が出来る。従って 定理 2.2 及び 10.3 表の S_u^2 の定義により

$$E(\bar{Y}_n - \bar{Y})^2 = \frac{(N-n)}{Nn} \frac{\sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2}{(N-1)} = \frac{(N-n)}{Nn} \left(S_u^2 + \frac{S_w^2}{M} \right) \quad (10.8)$$

(10.7) 及び (10.8) から最後に次のものを得る

$$\begin{aligned} V(\bar{y}) = E(\bar{y} - \bar{Y})^2 &= \frac{(M-m)}{M} \frac{S_w^2}{mn} + \frac{(N-n)}{Nn} \left(S_u^2 + \frac{S_w^2}{M} \right) \\ &= \frac{(N-n)}{N} \frac{S_u^2}{n} + \frac{(MN-mn)}{MN} \frac{S_w^2}{mn} \end{aligned}$$

もしも n/N が無視出来れば $V(\bar{y})$ は基礎理論の場合と同じく

$$\frac{S_u^2}{n} + \frac{S_w^2}{mn}$$

となる。

§ 10.5 一般の場合に於ける分散の推定

Estimation of the variance in the general case-

第一段階は S_u^2 及び S_w^2 の標本推定量を見出す事である。もしも標本分散分析が 10.1 表の様に行われると平均平方 $\mathcal{S}_b^2$ 及び $\mathcal{S}_w^2$ はなお表で与えられた期待値を有している事になる。即ち

$$E(\mathcal{S}_b^2) = S_w^2 + m S_u^2 \quad (10.9)$$

$$E(\mathcal{S}_w^2) = S_w^2 \quad (10.10)$$

$\mathcal{S}_w^2$ に対する結果は容易に証明されるが、 $\mathcal{S}_b^2$ に対しては明らかでない。それに次の様な直接の方法で証明される。

$$\mathcal{S}_b^2 = \frac{m \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}})^2}{n-1}$$

$$\text{さて} \quad \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}})^2 = \sum_{i=1}^n \bar{y}_i^2 - n \bar{\bar{y}}^2 \quad (10.11)$$

$$\text{次の様にかく} \quad \bar{y}_i = \bar{Y}_i + (\bar{y}_i - \bar{Y}_i)$$

$\bar{y}_i$ は大きさ m なる無作為副次標本の平均値であるから

$$E(\bar{y}_i^2) = \bar{Y}_i^2 + \frac{(M-m)}{M} \cdot \frac{S_{wi}^2}{m}$$

此処に平均値はこの単位からとられたすべての副次標本上でとられる。従って n 単位の一様選択に対して

$$\begin{aligned} E \sum_{i=1}^n \bar{y}_i^2 &= \sum_{i=1}^n \bar{Y}_i^2 + \frac{(M-m)}{Mn} \sum_{i=1}^n S_{wi}^2 \\ &= \sum_{i=1}^n \bar{Y}_i^2 + \frac{(M-m)}{Mm} n S_{wn}^2 \end{aligned} \quad (10.12)$$

更に次の様にかく

$$\bar{\bar{y}} = \bar{Y}_n + (\bar{\bar{y}} - \bar{Y}_n)$$

$$\text{従って} \quad E \bar{\bar{y}}^2 = \bar{Y}_n^2 + E(\bar{\bar{y}} - \bar{Y}_n)^2$$

此処に $\bar{Y}_n$ は標本のすべての n 個の単位の真の平均値である。方程式 (10.6) に於て

$$E(\bar{\bar{y}} - \bar{Y}_n)^2 = \frac{(M-m)}{M} \cdot \frac{1}{mn} S_{wn}^2$$

である事が証明された。

従って

$$E n \bar{\bar{y}}^2 = n \bar{Y}_n^2 + \frac{(M-m)}{Mn} S_{wn}^2 \quad (10.13)$$

(10.12) 及び (10.13) から

$$E \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}})^2 = \sum_{i=1}^n (\bar{Y}_i - \bar{Y}_n)^2 + \frac{(n-1)(M-m)}{Mn} S_{wn}^2$$

即ち

$$\frac{n}{(n-1)} E \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}})^2 = m \frac{\sum_{i=1}^n (\bar{Y}_i - \bar{Y}_n)^2}{n-1} + \frac{(M-m)}{M} S_{wn}^2$$

さて n 個の単位のすべての可能な選択上で平均しよう。定理 2.4 により右辺の第一項は $\bar{Y}$ の母分散の m 倍の不偏推定量である。これは次のものをさえる

$$\begin{aligned} E(\mathcal{S}_b^2) &= m \sum_{i=1}^N \frac{(\bar{Y}_i - \bar{Y})^2}{N-1} + \frac{(M-m)}{M} S_w^2 \\ &= \frac{m}{M} (S_w^2 + M S_u^2) + (1 - \frac{m}{M}) S_w^2 = S_w^2 + m S_u^2 \end{aligned}$$

最後の式は S_u^2 を定義する 10.3 表の方程式を利用している。これで証明は終った。

この結果は $V(\bar{\bar{y}})$ の不偏推定量を導く。

定理 10.4

標本からの $V(\bar{\bar{y}})$ の不偏推定量は

$$v(\bar{\bar{y}}) = \frac{1}{mn} \left\{ \frac{(N-n)}{N} \mathcal{S}_b^2 + \frac{(M-m)}{M} \frac{n}{N} \mathcal{S}_w^2 \right\} \quad (10.14)$$

(証明)

$v(\bar{\bar{y}})$ に於ける $\mathcal{S}_b^2$, $\mathcal{S}_w^2$ の期待値を代入する。

$$\begin{aligned} E v(\bar{\bar{y}}) &= \frac{1}{mn} \left\{ \frac{(N-n)}{N} (S_w^2 + m S_u^2) + \frac{(M-m)}{M} \frac{n}{N} S_w^2 \right\} \\ &= \frac{(N-n)}{N} \frac{S_u^2}{n} + \frac{(MN-mn)}{MN} \frac{S_w^2}{mn} \quad (S_w^2 \text{ の項をまとめて}) \end{aligned}$$

もしも $m=M$ であれば 公式 (10.14) は単位の単純無作為抽出に用いたものとなる。

もしも $n=N$ であれば 公式は単位は層と考えられそのすべ

てが抽出されるから比例層化抽出に於ける公式と同じである。

(同様に二階抽出は層として単位をもつ不完全層化の一種である)

n/N が無視出来る普通の場合には

$$v(\bar{y}) = \frac{s_b^2}{mn} = \frac{\sum_{i=1}^n (\bar{y}_i - \bar{y})^2}{n(n-1)} \quad (10.15)$$

これは定理 10.2 の基礎理論からの結果と一致する。

$m=1$ なるときは標本は s_w^2 の推定を与えない。これは s_w^2 が $v(\bar{y})$ に入っていないという事実で n/N が無視出来るという事を假定した問題ではない。この結果の一つの応用は副次抽出が組織的であるときに起る。組織副次標本はより複雑な要素の型及び $m=1$ なる単純無作為副次標本に等しいから、公式 (10.15) は n/N が大きくなければ組織副次標本抽出にも妥当である。然しむしろ一般抽出が組織的であれば公式はむしろ単位の組織標本が単純無作為標本に等しい場合はその場合のみ成立する。

§ 10.6 最適抽出比及び副次抽出比

Optimum sampling and subsampling fractions:-

これ等は費用函数の型による。有用であると分っている一つの型は

$$C = C_u n + C_e n m \quad (10.16)$$

第一項 $C_u n$ は標本に於ける単位の数に比例する。第二項 $C_e n m$ は要素の総数に比例する

我々は次のものを最小にする様に n 及び m を選ぶ

$$V(\bar{y}) + \lambda(C - C_u n - C_e n m) \quad (10.17)$$

n は組合せに於てのみ V 及び C に入ってくるから $k = nm$ とかさ (10.17) を次の様に書く。

$$\left(\frac{1}{n} - \frac{1}{N}\right) S_u^2 + \left(\frac{1}{k} - \frac{1}{MN}\right) S_w^2 + \lambda(C - C_u n - C_e k)$$

n と k について微分すると夫々

$$\frac{S_u^2}{n^2} = -\lambda C_u$$

$$\frac{S_w^2}{k^2} = -\lambda C_e$$

従って

$$\frac{k}{n} = m_{opt} = \frac{S_w}{S_u} \sqrt{\frac{C_u}{C_e}} \quad (10.18)$$

この公式の構造は期待した通りである。単位間の変動に比して単位内変動が大となればなる程最適 m は大きくなる。同様に単位内の要素から資料を得る費用 C_e に比して単位を得る費用 C_u が大となればなる程最適 m は大きくなる。

n の値は費用又は分散の何れかが一定であるかにより費用方程式又は分散方程式の何れかを解く事によって得られる。

実際の使用に対しては、 S_w 及び S_u の推定量は n 及び m の特別な値をもつ標本の分散分析 (10.1 表) から通常得られる。 m_{opt} の標本推定量は

$$\hat{m}_{opt} = \frac{S_w \sqrt{m}}{\sqrt{s_b^2 - s_w^2}} \sqrt{\frac{C_u}{C_e}} \quad (10.19)$$

見出される値は整数ではないであろう。Carmerson (1951) はこれは全く最上の規則でない事を指摘したけれどもそれに最も近い整数をとれば充分である。もし $\hat{m}_{opt}$ が二つの整数 m 及び $(m+1)$ の間にあればもし $m_{opt}^2 > m(m+1)$ であれば $(m+1)$ を選ばねばならぬ。さもなければ m を選ぶ。従ってもし $\hat{m}_{opt}$ が $1.414 = \sqrt{2}$ と 2 との間にあれば 2 に切上げてしまふ。もし $\hat{m}_{opt}$ が M よりも大きい又は s_b^2 が s_w^2 より小さければ $m=M$ として一段抽出を用いる。

推定量 $\hat{m}_{opt}$ はそれそのものはその大きさが数個の因子によるけれども、それから s_b^2 が計算される標本に於ける単位の数 n に主としてよる抽出誤差をうける。 $\hat{m}_{opt}$ の信頼限界は n が 10 以下になれば広くなる傾向がある。併し最適の範囲は広く、 m に於ける

少数単位の誤差は正確度の小さい損失だけを生ずる。次の例はこの小論を研究するための方法を示す。

(例) $C_u = 10 C_e$; $S_w = 1.3 S_u$ とおくと

$$m_{opt} = 1.3 \sqrt{10} = 4.1$$

全費用を一定と考え $\bar{y}$ の分散が m と共に如何に変化するかをみよう。 N を M 共に大きいと仮定する。

$$V(\bar{y}) = \frac{S_u^2}{n} + \frac{S_w^2}{nm}$$

$$= (S_u^2 + \frac{S_w^2}{m}) (\frac{C_u + m C_e}{C_e})$$

上式の最後は、費用方程式によって n を消去したものである。この例では V は次の様にかかる。

$$V(\bar{y}) = \frac{S_u^2 C_e}{C_e} (1 + \frac{S_w^2}{m S_u^2}) (\frac{C_u}{C_e} + m) = \frac{S_u^2 C_e}{C_e} (1 + \frac{1.69}{m}) (10 + m)$$

一定因子を除くと相対分散が m の違った値に対して計算出来る。10.4 表はこれ等の分散及び相対的正確度 (標準として $m=4$ に対するものを最大正確度とした) を示す。

10.4 表 m の種々の値に対する相対的分散及び正確度

$m =$	1	2	3	4	5	6	7	8	9	10
相対的分散	29.57	22.14	20.32	19.92	20.07	20.51	21.10	21.80	22.56	23.38
相対的正確度	0.67	0.90	0.98	1.00	0.99	0.97	0.94	0.91	0.88	0.85

2 と 9 との間の m の値に対しては最適に対する相対的正確度の損失は 1.2% 以下である。英国の病気の月々の調査に対するこの種の分析の興味ある応用が Gray & Corlett (1950) によって与えられている。

さて我々は $m=4$ 及び n の色々な値をもつ最初の標本から如何によく m_{opt} が推定出来るかを考えよう。(10.19) から

$$\hat{m}_{opt} = \frac{\sqrt{m}}{\sqrt{(S_w^2/S_u^2) - 1}} \sqrt{\frac{C_u}{C_e}} = \frac{6.324}{\sqrt{(S_w^2/S_u^2) - 1}}$$

もしも初期の変量 y_{ij} が正規に分布しているとすれば

S_w^2/S_u^2 は

$$F(1 + m \frac{S_u^2}{S_w^2}) = 3.367 F$$

なる分布をしている事が分散分析の理論から分る。但し F は自由度 $(n-1)$ 及び $n(m-1)$ なる F 分布に従う (Eisenhart, 1947) これは次を与える

$$\hat{m}_{opt} = \frac{6.324}{\sqrt{3.367 F - 1}}$$

この結果は m_{opt} に対する信頼限界を与える。 $n=10$ として 80% 限界を考えよう。自由度は 9 及 30 である。

F の 20% 有意水準から (Morrington & Thompson, 1943) 次が分る

$$F_{20}^{**}(9, 30) = 1.8490$$

$$F_{80}^{***}(9, 30) = 1/F_{10}(30, 9) = 1/2.2547 = 0.4435$$

F の之等の値を代入すると

$$\text{下限} : m_{opt} = 2.8$$

$$\text{上限} : m_{opt} = 9.0$$

10.4 表から分る様に、この範囲内の如何なる m の値もかなり最適に近い正確度の程度を与える。従って $n=10$ では正確度の損失が小さい確率は 10 回中 8 回である。

$n=5, 10, 20$ に対する 80% 及び 90% 信頼限界は 10.5 表に与えてある。三つの場合にある $m=\infty$ の上限は一段抽出を意味する。

10.5 表 m_{opt} に対する信頼限界

n	80%	95%
5	2.5, ∞	1.8, ∞
10	2.8, 9.0	2.3, ∞
20	3.1, 6.4	2.7, 4.1

註 * 原著では 10%, ** 原著では F_{10} , *** 原著では F_{90} になっている

要約すると、 $n=20$ では実際に得られた正確度は最適に殆んど近い位充分な $Mopt$ を推定する事は殆んど確実である。これは $n=5$ では真ではない。結果は費用及び分散の他の値では勿論違ってくる。

§ 10.7 割合に対する副次抽出

Subsampling for proportions:—

もしも要素が二つのクラスに分類され第一のクラスに入る割合を推定したいとすれば、 y_{ij} をこのクラスにそれに応ずる要素が入れば1としさもなければ0と定義する通常の工夫によって前の公式を適用出来る。

$p_i = a_i/n$ をして i 番目の単位からの副次標本に於ける最初のクラスに入る割合としよう。二つの推定分散 S_b^2 , S_w^2 は次の如くして計算される。

$$S_b^2 = \frac{n \sum_{i=1}^n (p_i - \bar{p})^2}{n-1}$$

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

此処に $\bar{p} = \sum p_i/n$ 。従って二段抽出に於ける推定分散に対する公式は(定理 10.4 により)

$$V(\bar{p}) = \frac{(N-n)}{N} \frac{1}{n(n-1)} \sum_{i=1}^n (p_i - \bar{p})^2 + \frac{(M-n)}{M} \frac{1}{Nn(m-1)} \sum_{i=1}^n p_i q_i$$

§ 10.8 三段抽出法

Three-stage sampling:—

副次抽出の方法は時には要素を完全調査する代りに副次単位(要素)を抽出する事によって三段に実行される場合がある。例えばインドの穀物生産高の推定調査に於て(Sukhatme, 1947)村は便利な抽出単位である。村の中で問題になっている穀物が生育し

ている畠のいくらかだけが選ばれるから畠は副次単位である。畠が選ばれるとその中のある部分だけがエーカー当り收穫を決定するため刈りとられる。この様に副次単位そのものが抽出される。もしも穀物の物理化学的分析が含まれるとすれば、これ等の決定は畠から刈られた標本の一部分だけについてのみ行われるから附加的な副次抽出が用いられる。

基礎理論に対する結果は簡単に与えられる。母集団は N 個の単位を含み、その各々が M 個の副次単位を含む、又その各々が k 個の副副次単位(sub-subunit)を含む、それに対応する標本の数は夫々 n, m 及び k である。モデルは

$$y_{ijz} = \bar{Y} + u_i + w_{ij} + e_{ijz}$$

此処に u_i, w_{ij}, e_{ijz} なる成分は夫々平均値 0、分散 S_u^2, S_w^2 , 及び S_{ew}^2 なる独立な分布をする、副副次単位当り標本分散は次の如くなる。

$$V(\bar{y}) = \frac{S_u^2}{n} + \frac{S_w^2}{nm} + \frac{S_{ew}^2}{nmk} \quad (10.20)$$

副副次単位を基とした標本の分散分析が 10.6 表に与えてある。

10.6 表 三段抽出に対する分散分析

	df	m, S	次のものの推定量
単位間	$(n-1)$	$S_b^2 = nk \sum_i (\bar{y}_i - \bar{y})^2 / (n-1)$	$S_w^2 + k S_u^2 + mk S_{ew}^2$
単位内副次単位間	$n(m-1)$	$S_w^2 = k \sum_i \sum_j (\bar{y}_{ij} - \bar{y}_i)^2 / (n(m-1))$	$S_{ew}^2 + k S_w^2$
副次単位内副副次単位間	$nm(k-1)$	$S_{ew}^2 = \sum_i \sum_j \sum_z (y_{ijz} - \bar{y}_{ij})^2 / nm(k-1)$	S_{ew}^2

右欄の期待値はモデルから容易に証明される。方程式 (10.20) から $V(\bar{y})$ の不偏推定量は S_b^2/nmk である。

二段抽出に於けると同様、不偏推定量は亦用いられたものとは違ひ n, m 及び k の値に対して $V(\bar{y})$ に対して得る事が出来る。

一般理論に於ては、第一段階は二段抽出に於ける 10.3 表と同様

に完全母集団に対する分散分析により分散成分 S_u^2 , S_w^2 及び S_{uw}^2 を定義する事である。(10.7表)

10.7表 母集団に対する分散分析

	df	m.s	次のものに等しいと表される
単位間	$(N-1)$	$Mk \sum (\bar{y}_k - \bar{\bar{y}})^2 / (N-1)$	$S_{uw}^2 + kS_w^2 + MkS_u^2$
単位内副次単位間	$N(M-1)$	$k \sum \sum (\bar{y}_{ij} - \bar{\bar{y}}_i)^2 / N(M-1)$	$S_{uw}^2 + kS_w^2$
副次単位内副次単位間	$NM(k-1)$	$\sum \sum \sum (y_{ijk} - \bar{\bar{y}}_{ij})^2 / NM(k-1)$	S_{uw}^2

定理 10.5

単純無作為抽出がすべての三段抽出に用いられるとすれば

$$V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_u^2 + \left(\frac{1}{nm} - \frac{1}{NM}\right) S_w^2 + \left(\frac{1}{nmk} - \frac{1}{NMK}\right) S_{uw}^2$$

(証明)

証明は二段抽出の場合の自然の拡張であるから主段階のみを示す。次の如くかく

$$\bar{y} - \bar{\bar{y}} = (\bar{y} - \bar{\bar{y}}_{nm}) + (\bar{\bar{y}}_{nm} - \bar{\bar{y}}_n) + (\bar{\bar{y}}_n - \bar{\bar{y}})$$

此処に $\bar{\bar{y}}_{nm}$ は選ばれた nm 個の副次単位の母平均、 $\bar{\bar{y}}_n$ は選ばれた n 単位の母平均である。

平方して平均すれば幾の項はなくなる。平方した項に役立つものは次のものである。

$$E(\bar{y} - \bar{\bar{y}}_{nm})^2 = \left(\frac{k-k}{K}\right) \frac{1}{nmk} S_{uw}^2$$

$$E(\bar{\bar{y}}_{nm} - \bar{\bar{y}}_n)^2 = \left(\frac{M-m}{M}\right) \frac{1}{nm} \left(\frac{1}{k} S_{uw}^2 + S_w^2\right)$$

$$E(\bar{\bar{y}}_n - \bar{\bar{y}})^2 = \left(\frac{N-n}{N}\right) \frac{1}{n} \left(\frac{1}{MK} S_{uw}^2 + \frac{1}{M} S_w^2 + S_u^2\right)$$

三つの項を加えると定理を得る。

定理 10.6

標本からの $V(\bar{y})$ の不偏推定量は

$$v(\bar{y}) = \frac{1}{nmk} \left[\frac{(N-n)}{N} S_u^2 + \frac{(M-m)}{M} \frac{n}{N} S_w^2 + \frac{(k-k)}{K} \frac{n}{N} \cdot \frac{m}{M} S_{uw}^2 \right]$$

証明は S_u^2 , S_w^2 及び S_{uw}^2 の期待値に対する 10.6 表で述べた結果が妥当である事を示す事になる。

詳細にやると面倒であるが結果は明らかである。

更に上の段階への抽出にこの結果を拡張する事は公式の構造から明らかである。二段及び三段での最適抽出が二段抽出の場合と同時に研究出来る。

§ 10.9 単位の層化抽出

Stratified sampling of the unit:—

副次抽出は如何なる単位の抽出型にも組合される。副次抽出そのものは層化抽出又は組織抽出に適用される。これ等の型に対する分散公式はより簡単な方法で公式から作る事が出来る。結果は単位の層化抽出と副次抽出との組合せに対して得られる。我々は単位の大きさは与えられた層では一定であるが層毎に違ふと假定する添字 h は層に属するものである。母分散 S_{uh}^2 及び S_{wh}^2 は 10.3 表と同様な分散分析により各層で別個に一般に定義される。それに応ずる標本の数は n_h 及び m_h である。要素当り推定母平均は

$$\bar{y}_{st} = \frac{\sum M_h N_h \bar{y}_h}{\sum M_h N_h}$$

此処に

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{m_h} y_{hij}}{n_h m_h}$$

各層内で定理 10.3 を適用して

$$V(\bar{y}_{st}) = \frac{\sum (M_h N_h)^2 \left\{ \frac{(N_h - n_h)}{N_h} \frac{S_{uh}^2}{n_h} + \frac{(M_h N_h - m_h m_h)}{M_h N_h} \frac{S_{wh}^2}{m_h m_h} \right\}}{(\sum M_h N_h)^2}$$

定理 10.4 から不偏標本推定量は

$$V(\bar{y}_{st}) = \frac{\sum_h \frac{(M_h N_h)^2}{M_h N_h} \left\{ \frac{(M_h - n_h)^2}{N_h} S_{1h}^2 + \frac{(M_h - m_h)}{M_h} \frac{n_h}{N_h} S_{0h}^2 \right\}}{(\sum M_h N_h)^2}$$

もしもすべて単位に対する抽出比 n_h/N_h が小さければこの公式は次の如く簡単となる。

$$V(\bar{y}_{st}) = \frac{\sum_h (M_h N_h)^2 \left(\frac{S_{1h}^2}{n_h} + \frac{S_{0h}^2}{m_h N_h} \right)}{(\sum M_h N_h)^2}$$

$$V(\bar{y}_{st}) = \frac{\sum_h \frac{(M_h N_h)^2}{M_h N_h} S_{0h}^2}{(\sum M_h N_h)^2}$$

第 11 章 単位の大きさが等しくない場合の副次抽出法

Subsampling with units of unequal sizes

§ 11.1 緒言 Introduction: -

もしも一次単位の大きさに変化があるときに二級抽出が用いられるとすれば、一つの方法は層内の単位の大きさが等しいか又はほぼ等しくなるように単位の大きさによって層化する事である。§ 10.9 の公式がそのときは適当な近似となる。併し時には単位の大きさが非常に変化しているから層のあるものの中には非常に大きな差が残る事になり、時には大きさよりも変量についての層化を基とした方がよい場合がある。一次単位として大抵は区で以て民族の大きさの標本とする英国社会調査の見方では、Gray と Corlett (1950) は大きさは先ず第一に層化に対する変量の一つとして含ませる事を指摘したが、母集団の特性がよりよく分ったときに他の変量がより望ましい事が分った。

単位の大きさが変化しているときに多級抽出の良好な実地的知識を得るためには、ある集中的な努力を必要とする。というのはその技術は非常に融通性があるからである。単位は等確率が大きさに比例した確率が又は大きさのある推定量に比例した確率の何れかで選ばれる。副次抽出比を定めるために色々な規則が工夫出来又色々な推定方法が利用出来る。違った方法の利益は母集団の性質、調査費用及び我々の手元にある補助的な資料によって定まる。

この章の最初の部分は使用される主な方法を概略述べるのに費される。我々はまず一つの層からなっている母集団から始めなければならない。層化抽出への拡張は、前章の様に、層について近似分散公式を加え合せる事によって行ふ事が出来る。簡単なため、我々は先ず第一にたゞ一つの一次単位が選ばれる。即ち $n = 1$ の場合を仮定する。この場合は一見して思われる程非実地的なもので

はない、何故ならば非常に多くの層があるときには我々は $n=1$ のときでさえも推定に於て満足な正確度を得る事が可能であるからである。雇用の数を推定するためアメリカ合衆国調査局によつて行われた月々の調査に於て、郡 (county) が一次単位である。これは大きな単位であるが費用を節約できる大きな利益をもっている。郡はその特性は一様性からかけはなれているから、郡の層化がこれらの調査に於ては一つの郡だけが各層から送られる事まで拡張される。その結果此処に論ぜられるべき理論はこの種の抽出計画に於て単一層に適用出来る。前章に於ける様に推定さるべき量は母集団総計 Y 、母平均 (通常、要素別平均値 $\bar{y}$) 又は二つの変量の比である。現在の論議は要素当りの母平均値に限定する。注意

i 番目の単位内の j 番目の要素に対する観測値を y_{ij} で表わす。その記号は i 番目の単位に関するものである。

	母 集 団	標 本
要素の数	M_i	m_i
要素当り平均値	$\bar{y}_i$	$\bar{y}_i$
総 計	$Y_i = M_i \bar{y}_i$	$y_i = m_i \bar{y}_i$

又の記号は全母集団又は標本に関するものである。

	母 集 団	標 本
要素の数	$M = \sum M_i$	$m = \sum m_i$
総 計	$Y = \sum Y_i$	$y = \sum y_i$
要素当り平均値	$\bar{y} = Y/M$	$\bar{y} = y/m$
一次単位当り平均値	$\bar{y} = Y/N$	$\bar{y} = y/n$

記号は一つの点で10章の記号と違っている。我々は母集団及び標本に於ける要素の総数として夫々 M 及び m を定義している。一方10章ではこれ等の記号は一次単位に於けるそれ等に対する総計である (すべての単位は等しいから)。記号を一致させるためには記号 N 及び n が10章で用いられたものでなければならぬ。

§ 11.2 $n=1$ なるときの抽出方法

Sampling methods when $n=1$

i 番目の単位が選ばれそれは M_i 個の単位を含みその中 m_i 個が無作為に抽出されたと仮定しよう。我々は $\bar{y}$ を推定する次の三つの方法を考へる。単位が等しい確率で送ばれる場合

I 推定量 $= \bar{y}_E = \bar{y}_c$ 単位が等しい確率で送ばれる場合

推定量は要素当りの平均標本である。それは偏倚がある。何故ならば同じ単位からのくりかえし抽出では、 $\bar{y}_c$ の平均は $\bar{y}_c$ であり、各単位は同じ確率で送ばれるから $\bar{y}_c$ の平均は

$$\frac{1}{N} \sum_{i=1}^N \bar{y}_c = \bar{y}_a \quad (\text{とおく})$$

併し 母平均は

$$\bar{y} = \frac{\sum_{i=1}^N M_i \bar{y}_c}{M} \quad \text{此処に } M = \sum_{i=1}^N M_i$$

従つて偏倚は $(\bar{y}_a - \bar{y})$ に等しい。

$V(\bar{y}_E)$ を見出すには次の様にかく

$$\bar{y}_c - \bar{y} = (\bar{y}_c - \bar{y}_i) + (\bar{y}_i - \bar{y}_a) + (\bar{y}_a - \bar{y})$$

平方してすべての可能な標本上で期待値をとる。積の項はすべてなくなる。平方項の期待値は10章で与えた方法によつて容易に導かれる。即ち、

$$V(\bar{y}_c - \bar{y}) = \frac{1}{N} \sum_{i=1}^N \frac{(M_i - m_i)}{M_i} \frac{s_c^2}{m_i} + \frac{1}{N} \sum_{i=1}^N (\bar{y}_c - \bar{y}_a)^2 + (\bar{y}_a - \bar{y})^2 (M_i)$$

単位内 単位間 偏倚

此処に

$$s_c^2 = \frac{1}{(M_i - 1)} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_c)^2$$

は i 番目の単位に於ける要素間分散である。

$\bar{y}_E$ の分散は次の三つの成分を含んでいる。即ち単位内の変動

から生ずる成分、単位の真の平均値間の変動から生ずる成分及び偏倚から生ずるものである。

m_c の値は特殊化されなかった。最も普通の選択はすべての m_c を等しくとるか又は M_c に比例した m_c をとるかである。即ちどんな単位が選ばれてもその一定割合の副次抽出をする事である。 m_c の選択は分散の三つの成分の最初のもの即ち単位内変動から生ずるもののみに影響される。

II 単位が等しい確率で選ばれる場合

$$\text{推定量} = \bar{y}_{II} = NM_c \bar{y}_c / M$$

この推定量は不偏である。 $\bar{y}_c$ は $\bar{y}_c$ の不偏推定量であるから積 $M_c \bar{y}_c$ は単位総計 Y_c の不偏推定量である。従って $N M_c \bar{y}_c$ は母集団総計 Y の不偏推定量である。 M で割れば即ち母集団に於ける要素の総数で割れば我々は $\bar{y}$ の不偏推定量を得る。 $V(\bar{y}_{II})$ を見出すには

$$\begin{aligned} \bar{y}_{II} - \bar{y} &= \frac{N M_c \bar{y}_c}{M} - \bar{y} \\ &= \frac{N M_c}{M} (\bar{y}_c - Y_c) + \left(\frac{N M_c}{M} \bar{y}_c - \bar{y} \right) \end{aligned}$$

さて $M_c \bar{y}_c = Y_c$ 即ち単位に対する総計であり、 $\bar{y}$ が単位当りの母平均である時には $\bar{y} = N \bar{Y} / M$ である。

これは

$$\bar{y}_{II} - \bar{y} = \frac{N M_c}{M} (\bar{y}_c - \bar{Y}_c) + \frac{N}{M} (Y_c - \bar{Y})$$

となる。

従って

$$V(\bar{y}_{II}) = \frac{N}{M^2} \sum_{c=1}^M M_c (M_c - m_c) \frac{S_c^2}{m_c} + \frac{N}{M^2} \sum_{c=1}^M (Y_c - \bar{Y})^2 \quad (11.2)$$

この分散の“単位間”の成分(右辺の第2項)は単位総計 Y_c 間の変動を表わす。この成分は単位毎の M_c の変動及び要素当り平均値 $\bar{y}_c$ の変動の両方によって影響を受ける。もしも単位

が大きさによってかなり変動があれば、この成分はたとえ要素当り平均値 $\bar{y}_c$ が実際には単位毎に一定であっても大きい。しばしばこの成分は非常に大きいので $\bar{y}_{II}$ は偏倚推定量 $\bar{y}_I$ よりも大きな分散をもつ事がある。従って方法 I も方法 II もまだ充分満足なものではない。

III 単位が大きさ按比例した確率で選ばれた場合。

$$\text{推定量} = \bar{y}_{III} = \bar{y}_c = \text{標本平均}$$

この方法は Hansen 及び Hurwitz (1943) によったものである。それは不偏な標本平均を与え、方法 II に於ける様な分散の増大はない。

くりかえし抽出に於ては、 c 番目の単位は相対度数 M_c/M で以て表われる。従って

$$E(\bar{y}_{III}) = \sum_{c=1}^N \frac{M_c}{M} \bar{y}_c = \bar{y}$$

更に

$$\bar{y}_{III} - \bar{y} = (\bar{y}_{III} - \bar{y}_c) + (\bar{y}_c - \bar{y})$$

まずそれに於て c 番目の単位が選ばれた標本上で平均する。

$$E_c(\bar{y}_{III} - \bar{y})^2 = \frac{(M_c - m_c)}{M_c} \frac{S_c^2}{m_c} + (\bar{y}_c - \bar{y})^2$$

さてすべての可能な単位の変動されたものの上で平均する。 c 番目の単位は相対度数 M_c/M で以て選ばれるから

$$V(\bar{y}_{III}) = \frac{1}{M} \left\{ \sum_{c=1}^N (M_c - m_c) \frac{S_c^2}{m_c} + \sum_{c=1}^N M_c (\bar{y}_c - \bar{y})^2 \right\} \quad (11.3)$$

方法 I に於ける様に“単位間”成分は次々の単位に於ける要素当り平均値 $\bar{y}_c$ 間の差から生ずる事に注意を要する。もしもこれ等の要素当り平均値が殆んどすべて等しければ、この成分は小さい。

(例)

これ等の結果を人工的に作られた小母集団に適用してみよう。

資料を 11.1 表に示す。夫々 2, 4 及び 6 なる要素をもつ三つの単位がある。 Y_u , S_u^2 , 及び $\bar{Y}_u$ で与えられる数字はもう知っているであろう。母平均 $\bar{Y}$ は $33/12$ 即ち 2.75 である。 $\bar{Y}_u$ の重みづけない平均値は $2.167 = \bar{Y}_a$ であるから、方法 I に於ける偏倚は -0.583 である。その平方即ち分散に役立つものは 0.340 である。

11.1 表 大きさが等しくない単位をもつ人工母集団

単位	g_{ij}	M_u	Y_u	S_u^2	$\bar{Y}_u$	$\bar{Y}_u - \bar{Y}$
1	0, 1	2	1	0.500	0.5	-2.25
2	1, 2, 2, 3	4	8	0.667	2.0	-0.75
3	3, 3, 4, 4, 5, 5	6	24	0.800	4.0	+1.25
計		12	33			

一つの単位が選ばれそれから二つの要素が抽出される。我々は次の四つの方法を考えよう、その中の二つは方法 I の変形である。

方法 Ia

選択：等確率で以ての単位， $m_u = 2$

推定量： $\bar{y}_u$ (偏倚)

方法 Ib

選択：等確率で以ての単位， $m_u = \frac{1}{2} M_u$

推定量： $\bar{y}_u$ (偏倚)

方法 II

選択：等確率で以ての単位， $m_u = 2$

推定量： $N M_u \bar{y}_u / M$ (不偏)

方法 III

選択： M_u/M なる確率で以ての単位， $m_u = 2$

推定量： $\bar{y}_u$ (不偏)

方法 Ia (比例副次抽出) は 2 なる大きさをもつ標本とは一致しなくて (それは 1, 2 または 3 である) 平均の標本の大きさは 2 である。

抽出誤差公式 (11.1), (11.2) 及び (11.3) を適用する事によって 11.2 表の結果を得る。

11.2 表 $\bar{Y}$ の標本推定量の分散

方法	次のものからの分散に役立つ			全分散
	単位内	単位間	偏倚	
Ia	0.145	2.056	0.340	2.541
Ib	0.183	2.056	0.340	2.579
II	0.256	5.792	0.000	6.048
III	0.189	1.813	0.000	2.002

例は人工的であるけれども、その結果は多くの母集団で行われた比較に表われるものの代表的なものである。方法 II は単位間分散が最小であるから最小分散を与えている。方法 II は不偏ではあるが、非常に劣っている。方法 Ia は (副次標本の等しい大きさ) 方法 Ib (比例副次抽出) よりも少しはよい。

これ等の方法のいくつかの比較が本実験の母集団について行われた。六つの項目 (全労働者、全農業労働者、全非農業労働者、これを性別即ち男性及び女性に対して別個に推定された) に対して Hansen 及び Hurwitz (1943) は方法 II は不偏な方法 II と比較したとき単位間変動からの貢献に大いに減少があり、方法 I と比較したとき平均 30% の減少がある事を見出した。(彼等は単位内変動からの貢献は無視されると仮定した) North Carolina 州に対する典型的な農場項目を推定する際に Jebe (1952) は、I の型の方法と比較したとき 15% のオーガーで全分散に減少がある事を報告した。この研究に於て共に一次単位は郡であった。

§ 11.3 推定大きさに比例した確率で以ての標本抽出

Sampling with probability proportional to estimated size

単位の大きさ M_i が大ざっぱにしか分っていない様な場合がある。単位がブロックで要素が家族である様な町の標本抽出に於てはブロック当り家族の数は町の地図か又は以前の調査資料から通常得られる。そして共にそれぞれは多少共時期遅れのものである。

pps 標本抽出の利益のいくらかは大きさの最良推定量に比例した確率で以ての抽出により得る事が出来る。 Z_i をして i 番目の単位に割当てた確率としよう。此処に Z_i は 1 以上の正数の組である。我々はなお $n = 1$ と仮定しよう。

方法 IV

$\bar{y}$ の不偏推定量は

$$\bar{y}_{IV} = \frac{\sum_{i=1}^N M_i \bar{y}_i}{\sum_{i=1}^N M_i} \quad (11.4)$$

これはくりかえし抽出では i 番目の単位は相対度数 Z_i で以て表われるからである。従って

$$E(\bar{y}_{IV}) = \sum_{i=1}^N Z_i \left(\frac{M_i \bar{y}_i}{\sum_{i=1}^N M_i} \right) = \sum_{i=1}^N \frac{M_i \bar{y}_i}{M} = \bar{y}$$

この方法では

$$m_i = \frac{k M_i}{Z_i} \quad (11.5)$$

なる如く m_i を選ぶのが普通である。此処に k は常数である。そのとき 推定量は

$$\bar{y}_{IV} = \frac{\sum_{i=1}^N m_i \bar{y}_i}{\sum_{i=1}^N m_i} = \frac{\sum_{i=1}^N \bar{y}_i}{\sum_{i=1}^N 1} = \bar{y} \quad (11.6)$$

此処に $\bar{y}_i$ は標本総計である。

量 k は期待されるすべての抽出比として表わされる。というのは

$$E(m_i) = \sum_{i=1}^N Z_i m_i = k \sum_{i=1}^N M_i = k M$$

従って

$$k = \frac{\text{標本に於ける要素の期待数}}{\text{母集団に於ける要素の数}}$$

$m_i = k M_i / Z_i$ を選ぶ事に於ける利益は今の簡単な場合には表われてこない。 $n > 1$ 及び層化抽出ではこの選択が推定量を自己荷重にするという事が後に分かるであろう。

(例)

これは m_i を如何にして定めるかを示す。層は 11.3 表に示される様な家族の推定数をもつ 6 つのブロックを含んでいる。我々はすべて 5% の抽出比をもつようにしたいと仮定しよう。

11.3 表 m_i の計算例

ブロック	家族推定数	累積和	割当て範囲
1	10	10	1 - 10
2	30	40	11 - 40
3	17	57	41 - 57
4	25	82	58 - 82
5	23	105	83 - 105
6	16	121	106 - 121

$k = \frac{1}{20}$ とする。通常の pps 標本抽出と同様に、単位はまづ 1 と 121 の間の乱数を抽く事によって選ばれる。選ばれたブロックは累積和において 83 から 105 迄の間にある。20.5 である。

訪問者がこのブロックを訪れそこですべての家族を前以てリストにのせるか又は数えるかする。調査して 31 家族あったとしよう。方程式 (11.5) を適用して

$$m_i = \frac{k M_i}{Z_i} = \left(\frac{1}{20} \right) (31) \frac{121}{23} = 8 \quad (\text{一番近い整数})$$

欲する副次抽出比

$$\frac{m_i}{M_i} = \frac{k}{Z_i} = \frac{121}{(20)(23)} \approx \frac{1}{4}$$

はブロックがリストされる前に分る。従って訪問者は前以てこのブロックから k 家族の中 1 家族をとる事を知る事が出来る。この規則は副次標本が実際に示はしはる様な組織的である時

に便利である。

$\bar{y}_{1v}$ の分散は通常の方法で得られる。

$$\begin{aligned}\bar{y}_{1v} - \bar{y} &= \frac{M_c \bar{y}_c}{\sum z_c} - \bar{y} && \text{((11.4) により)} \\ &= \frac{1}{M} \left\{ \frac{M_c}{z_c} (\bar{y}_c - \bar{y}_c) + \left(\frac{M_c}{z_c} \bar{y}_c - M \bar{y} \right) \right\}\end{aligned}$$

分散に於て、各平方は荷重 z_c をうける。従つて

$$V(\bar{y}_{1v}) = \frac{1}{M^2} \left\{ \sum_{c=1}^N \frac{M_c(M_c - m_c)}{z_c} \frac{S_c^2}{m_c} + \sum_{c=1}^N z_c \left(\frac{M_c \bar{y}_c}{z_c} - M \bar{y} \right)^2 \right\} \quad (11.7)$$

もしも $h_c = m_c z_c / M_c$ であれば、これは次の如くかなり簡単な形にかゝれる。

$$V(\bar{y}_{1v}) = \frac{1}{M^2} \left\{ \sum_{c=1}^N \frac{(M_c - m_c) S_c^2}{h_c} + \sum_{c=1}^N z_c \left(\frac{y_c}{z_c} - \bar{y} \right)^2 \right\} \quad (11.7')$$

もしも $z_c = M_c / M$ であれば公式 (11.7) は $V(\bar{y}_{II})$ に対しての公式 (11.3) となる。もしも $z_c = 1/N$ (初期確率が等しい) であれば公式 (11.7) は 確率が等しい場合の不偏推定量の分散に対しての公式 (11.2) となる。

$z_c = M_c / M$ であれば (11.7) に於ける "単位間" 成分は要素間平均値 $\bar{y}_c$ に於ける変動によつてと同様大まき M_c に於ける変動によつてある程度影響される。

(例)

11.4 表は 11.1 表の人工母集団に於て $V(\bar{y}_{1v})$ を求める基礎的な計算を示す。 $M = 12$ であり欲する標本の大きさは 2 であるから $h_c = 1/6$ とおく。 z_c は 0.2, 0.4 及び 0.4 としてとらえた。

11.4 表 $V(\bar{y}_{1v})$ の計算

単位	M_c	M_c/M	z_c	$m_c = M_c/6z_c$	$b(M - m_c)$	S_c^2	y_c	y_c/z_c	$y_c/z_c - \bar{y}$
1	2	0.17	.2	10/6	2	0.500	1	5	-2.8
2	8	0.33	.4	10/6	14	0.667	8	20	-1.3
3	6	0.50	.4	15/6	21	0.800	24	60	+2.7

実際には m_c は整数にまるめられる。これはこの例では行わなかった。公式 (11.7') から分散は次の様になる

$$\text{"単位内" からの貢献} = b \sum (M_c - m_c) S_c^2 / M^2 = 0.188$$

$$\begin{aligned}\text{"単位間" からの貢献} &= \sum z_c \left(\frac{y_c}{z_c} - \bar{y} \right)^2 / M^2 = 3.583 \\ &= 3.771\end{aligned}$$

総計

11.2 表と比較すると方法 IV は単位が等確率で選ばれた不偏の方法 II よりも小さい分散であるが方法 IV は方法 I 又は III よりも劣っている事が分る。もしも大きさが知らなかったならば、方法 II (PPS) は用いる事が出来ないであろう。併し等確率での抽出からの方法 I で得られた偏倚推定量を用いる事が出来る。この例では明らかに方法 IV は不偏推定量を得るために高価な費用を費している。方法 IV の失望させる様々やり方は必ずしも代表的なものではない。大きさのより接近した推定量では方法はより利益がある事を示す。併し標本平均 (方法 I の様な) が方法 IV で採用された推定量よりもよりよいものであるかどうかをきえる事はあたりまえの事である。

方法 V 単位を推定大まきに比例した確率で選ぶ場合

$$\text{推定量} = \bar{y}_v = \bar{y}_c = \text{標本平均}$$

この推定量は

$$E(\bar{y}_v) = \sum z_c \bar{y}_c = \bar{y}_z \quad (\text{とおく})$$

であるから偏倚している。

もしも z_c が良好な推定量であれば、 $\bar{y}_z$ は正しい平均値 $\bar{y} = \sum M_c \bar{y}_c / M$ に近いし偏倚は小さくなる。

もしも $\bar{y}_v = (\bar{y}_c - \bar{y}_c) + (\bar{y}_c - \bar{y}_z) + (\bar{y}_z - \bar{y})$ とかけば

分散の三つの成分は次の様に算かれる。

$$V(\bar{y}_v) = \sum_{c=1}^N \frac{z_c(M_c - m_c)}{M_c} \frac{S_c^2}{m_c} + \sum_{c=1}^N z_c (\bar{y}_c - \bar{y}_z)^2 + (\bar{y}_z - \bar{y})^2$$

(例)

もしも z_c と m_c の値が 11.4 表と同じ様に選ばれるならば、

読者は $\bar{y}_i$ の分散の成分は 11.5 表に示す様になる事を証明出来るだろう。

11.5 表 方法 V に於ける分散の成分

単位内	単位間	偏倚	全分散
0.178	1.800	0.062	2.040

これは方法 III (pp 5) 以外のすべての方法よりもまさり方法 III と殆んど同じ位良好である。

§ 11.4 $n=1$ なる場合に対する方法のとりまとめ

Summary of methods for $n=1$

要素当り平均値 $\bar{y}$ を推定する五つの方法と数値例に於ける分散を 11.6 表にまとめてある。

11.6 表 二級抽出法 ($n=1$)

方法	単位を返る確率	$\bar{y}$ の推定量	偏倚の採子	例に於ける分散
I	等確率	$\bar{y}_i$	偏倚	Ia: 2.541 Id: 2.579
II	等確率	$\frac{N M_i \bar{y}_i}{M}$	不偏	6.048
III	$\frac{M_i}{M}$ の大きさ	$\bar{y}_i$	不偏	2.002
IV	$\bar{y}_i$ の推定大きさ	$\frac{m_i \bar{y}_i}{k M}$	不偏	3.771
V	$\bar{y}_i$ の推定大きさ	$\bar{y}_i$	偏倚	2.040

§ 11.5 $n > 1$ なる場合の抽出法

Sampling methods when $n > 1$

この節では主な抽出方法及び推定量を一つ以上の単位が送られる通常の立場に対して述べる。論議はなお単一の層に限定する。

先づキーに等確率での単位の抽出を考えよう。例として「アメリカの科学者」の一卷から 20 頁を無作為に選ばれたとする。頁当り伝記の数 M_i は一般に約 14 から 21 に変化している。各選ばれた頁

で二つの伝記が無作為に選ばれ、科学者の年齢が記録された。資料を 11.7 表に示す。目的は全冊に於ける伝記をかいた人の平均年齢を推定する事である。

11.7 表 「アメリカの科学者」なる本に於ける 40 人の科学者の年齢 ($n=20, m=2$)

単位番号	M_i	年 令		合計 k_i	$M_i \bar{y}_i$
		y_{i1}	y_{i2}		
1	15	47	30	77	572.5
2	19	38	51	89	845.5
3	19	43	45	88	836.0
4	16	55	41	96	968.0
5	16	59	45	104	1322.0
6	19	39	38	77	731.5
7	18	43	43	86	774.0
8	18	49	51	100	900.0
9	18	45	35	80	720.0
10	18	46	59	105	945.0
11	20	71	64	135	1350.0
12	18	35	46	81	729.0
13	19	61	54	115	1092.5
14	19	45	87	132	1254.0
15	18	31	38	69	621.0
16	16	64	39	103	824.0
17	16	63	47	110	880.0
18	19	36	33	69	655.5
19	19	61	39	100	950.0
20	19	54	34	88	832.0
合計	359			11904	17121.5

標本平均が推定量を与える方法 I は二つの類似性をもっている。
まづ第一に我々は通常の標本平均をとる。即ち

$$\bar{y}_I = \frac{\sum y_i}{\sum m_i} = \frac{1908}{40} = 47.6 \text{ 支}$$

この推定量は偏倚がある。これは m を一定にしたとき

$$E(\bar{y}_I) = \frac{1}{N} \sum_{i=1}^N \bar{y}_i$$

であるのに要素当り標本平均は $\bar{y} = \sum M_i \bar{y}_i / M$ であるので容易に分る。偏倚推定量はより小さな単位にあまりに大きな重みを与えすぎる。もしも M_i と $\bar{y}_i$ の間に相関がなければ歪んだ重みは大した問題ではなく偏倚は大標本ではさして重要ではなくなる。併し、もしも M_i と $\bar{y}_i$ の間に相関があれば偏倚は大標本でも無視出来なくなる。この例では M_i と $\bar{y}_i$ の間には少しの負の相関がある事が期待される。何故ならば頁あたりの数をさめる伝記が長ければ長い程科学者の年令は年をとっている傾向があるからである。

この論議は方法 I に於ける別の推定量として荷重平均

$$\bar{y}_{I'} = \frac{\sum_{i=1}^N M_i \bar{y}_i}{\sum_{i=1}^N M_i} = \frac{17,121.5}{359} = 47.7 \text{ 支}$$

を暗示する。

これは分子及び分母共に標本が変ると変化するから典型的な比推定である。比推定の特徴である様に偏倚は大標本では無視出来る。もしも副次抽出が比例的であれば、この推定量は通常の標本平均となり、方法 I のものと一致する。

m がすべて等しい場合は、重みづけられない標本平均はしばしば荷重平均より小さい分散をもつ。併し偏倚を生ずる傾向が多いことから重みづけられない標本平均はより危険である。

不偏推定量(方法 II)は次によって与えられる。

$$\bar{y}_{II} = \frac{N}{nM} \sum_{i=1}^n M_i \bar{y}_i$$

この推定量に於ては標本に於ける n 個の単位の総計の不偏推定量である。量 $\sum M_i \bar{y}_i$ は抽出比の逆数 N/n 倍されそれを要素の総数 M で割る。この例では本の総頁数は 2823 で M (本に於ける伝記の総数) は "約 50,000" と序文に与えてある。例にこの数字をあてはめると

$$\bar{y}_{II} = \frac{2823}{(20)(50,000)} (17,121.5) = 48.3 \text{ 支}$$

$n=1$ に於ける場合と同様この推定量はもしも M_i 間に変動が多ければ正確度が悪くなる。

大きさ又は推定大きさに比例した確率で以ての抽出は、この例では採用されそうもない。何故ならば 2823 頁での記入事項の数を数えたり推定したりする仕事が含まれるからである。これ等の方法に対する推定量は代数的な形で与えられるだろう。

§ 9.10 で指摘した様に我々は大きさ又は推定大きさに比例した確率をもつたためにおきかえて以て標本抽出せねばならぬ。もしも同じ単位が二度抽かれるとすれば副次標本は又おきかえて以てとられる。推定量の偏倚を検討するに際して我々は § 9.11 と同じ数学的な工夫を用いねばならない。無作為変量 x_i は i 番目の単位が特別な標本にあらわれる回数を表わす。くりかえしの抽出に於ては $E(x_i) = n\pi_i$ である。此処に π_i は選択確率である。

方法 III (pps 抽出で標本に於けるすべての単位に対して m は等しい) に於ては、標本平均 $\bar{y}_{III}$ は不偏である。何故ならば

$$\begin{aligned} E(\bar{y}_{III}) &= \frac{1}{n} E(x_1 \bar{y}_1 + x_2 \bar{y}_2 + \dots + x_N \bar{y}_N) \\ &= \frac{1}{n} \sum_{i=1}^N \frac{n M_i \bar{y}_i}{M} = \bar{y} \end{aligned}$$

であるからである。

方法IV (Z_i に比例した確率で以ての抽出) に於ては、 m_i に割当てた値如何に拘らず常に不偏である推定量は

$$\bar{y}_{IV} = \frac{1}{nM} \sum_{i=1}^n \frac{M_i}{Z_i} \bar{y}_i$$

である。

何故ならば

$$E(\bar{y}_{IV}) = \frac{1}{nM} E\left(\sum_{i=1}^n \frac{M_i}{Z_i} \bar{y}_i\right) = \frac{1}{nM} \sum_{i=1}^n m_i \bar{y}_i = \bar{y}$$

であるからである。

この推定量は

$$\bar{y}_{IV} = \frac{1}{n k M} \sum_{i=1}^n m_i \bar{y}_i = \frac{\bar{y}}{n k M}$$

となるから、もしも

$$m_i = \frac{k M_i}{Z_i} \quad (11.7)$$

とすれば自己荷重となる。

$n=1$ なる場合には、 k はすべての期待抽出比であった。

$n>1$ に対しては (11.5) から新しい分数のシリーズの分子分母を加え合わせる事によって

$$k = \frac{m_i Z_i}{M_i} = \frac{\sum_{i=1}^N m_i Z_i}{M} \quad (11.8)$$

であるからすべての期待抽出比は k である。併し標本に於ける要素の平均数は

$$E\left(\sum_{i=1}^n m_i\right) = E\left(\sum_{i=1}^N x_i m_i\right) = n \sum_{i=1}^N Z_i m_i$$

従って (11.8) から

$$n k = (\text{標本に於ける要素の期待数}) / M$$

すべての抽出比 (標本に於ける要素の数と母集団に於ける要素の数の比) は一次単位抽出比 n/N 及び副次抽出比 m_i/M_i と明らかに区別しなければならぬ。

$n=1$ なる場合と同様この推定量は不偏推定量IIと同じ種類の分散の増大をある程度うける。 $n>1$ なる時方法IVの数値の可能な拡張がある。少くともかなりの偏倚をうけかねない推定量は荷重標本平均

$$\bar{y}_V = \frac{\sum_{i=1}^n \frac{M_i \bar{y}_i}{Z_i}}{\sum_{i=1}^n \frac{M_i}{Z_i}} \quad (11.9)$$

である。

分子は n の不偏推定量であり、分母は nM の不偏推定量である。もしも方法IVの採に $m_i = k M_i / Z_i$ であればこの推定量は重みづけられない標本平均 $\bar{y}$ となる。

§ 11.6 $n>1$ なる場合に対する方法のとりまとめ。

Summary of methods for $n>1$

方法は 11.8 表に示す。自己荷重標本の利益の見地から推定量に自己荷重させる m_i の値も又与えてある。

11.8 表 二段抽出法 ($n>1$)

方法	単位を選ぶ確率	m_i	$\bar{y}$ の推定量		偏倚の様子
			一般	自己荷重	
I	等確率	$\bar{m}$	$\bar{y}$	$\bar{y}$	偏倚
I'	等確率	$k N M_i$	$\frac{\sum M_i \bar{y}_i}{\sum M_i}$	$\bar{y}$	偏倚
II	等確率	$k N M_i$	$\frac{N}{nM} \sum M_i \bar{y}_i$	$\frac{\bar{y}}{n k M}$	不偏
III	M_i/M	$\bar{m}$	$\bar{y}$	$\bar{y}$	不偏
IV	Z_i	$k M_i / Z_i$	$\frac{1}{nM} \sum \frac{M_i}{Z_i} \bar{y}_i$	$\bar{y} / n k M$	不偏
V	Z_i	$k M_i / Z_i$	$\frac{\sum \frac{M_i}{Z_i} \bar{y}_i}{\sum \frac{M_i}{Z_i}}$	$\bar{y}$	偏倚

方法IV及びVに於ては、これ等の値は $m_i = k M_i / Z_i$ である。

方法 I' 及び II に於ては、比例的副次抽出は自己荷重推定量を生ずる。この比例の枠子は $m_c = kNM_c$ によって示される。抽出が等確率である時 $Z_c = 1/N$ であるから、記号 k は方法 I', II, IV 及び V に対しての 11.8 表に於て同じ意味である。

一つの推定量に対する二つの代数的表示が示される処では (方法 I', II, IV 及び V の如き) 第一のものは m_c の如何なる選択に対しても適用する推定量の一般型であり、二番目のものは m_c が前の欄で選ばれたときのみに成立する自己荷重の型である。

§ 11.7 割合の推定

The estimation of proportions :-

11.8 表に於ける推定量は、その目的があるきまったクラス C に属する母集団に於ける要素の割合 (例えば スノエ 以上の人の割合) を推定する事である場合にも適用出来る。我々は要素が C に属するならば 1、さもなければ 0 であるとして y_{ci} を定義する通常の工夫を採用するだけが必要である。

もしも割合の分母が母集団に於けるすべての要素を含まなければ、その立場は直ってくる。両性に関する一般調査に於ては、この種の割合の例は

14 文以上の雇用男性の数

14 文以上の男性の総数

もしも要素が 14 文以上の雇用男性であれば $y_{ci} = 1$ 、さもなければ 0 とし、需要が 14 文以上の男性であれば $x_{ci} = 1$ でさもなければ 0 であるとき、母集団の割合は比 y/x となるから比推定となる。11.8 表の方法 V の一般比である比推定を考えよう

$$\hat{\theta} = \frac{\sum_{c=1}^n \frac{M_c \bar{y}_c}{Z_c}}{\sum_{c=1}^n \frac{M_c \bar{x}_c}{Z_c}} \quad (11.10)$$

分子及び分母は夫々 nY 及び nX の不偏推定量である (11.8 表

方法 IV により)、もしも $m_c = kM_c/Z_c$ であれば、この推定量は $\hat{\theta} = y/x$ という簡単な型となる。

明らかな様にこの型の推定量は割合及び連続変量の両方の理論を提示するのに非常に有用である。その分散は § 11.9 で与えられる。

§ 11.8 暫定的註釈

Interim Comments

色々な抽出方法及び推定方法を述べた際に我々は既に二つの主な結果を示した。先ず第一は要素の与えられた数 Z 、 m_c 及び標本に於て与えられた一次単位の数 n に対して $p/p/s$ での抽出はしばしば最も正確な推定量を与えるということである。第二は一次単位が等確率又は等しくない確率で選ばれようが要素当り荷重又は重みづけのない標本平均に基いた推定の偏倚の方法は同じ資料から作り得る不偏推定量よりもしばしば正確であるという事である。これ等の結果の何れも数学的には確かではないが、ある理論的調査と同様に経験はそれ等が多くの母集団及び多くの項目の型に実際に成立する事を暗示している。

この章の残りの内容は次の通りである。

§ 11.9 では、此処に述べたすべての方法に用いる一般的な分散公式を展開させる。

§ 11.10 は Hansen 及び Hurvitz (1949) による調査費用を考慮に入れたときに一次単位の選択の最適確率を決定する方法を示す。

§ 11.11 は標本平均に依する不偏推定量よりも標本平均に基いた推定方法がより正確であるための一次単位総計 Y 及び一次単位の数 M との間で成立しなければならない関係の型を示す。

§ 11.12 は分散の推定量を見出すための一般公式が得られる。

§ 11.13 は如何なる方法が層化抽出に拡張されるかを示す。

§ 11.9 主な分散公式

The principal variance formula

各方法に対して何々の公式を導くのを避けるため我々は方法 V はもしも $Z_c = M_c/M$ であれば方法 II となり、 $Z_c = 1/N$ であれば

方法 I' となる事に注意する。(等確率では我々はおきかえで以て抽出する事を除いて)、同様に方法 IV は $\Sigma x_i = 1/n$ のときに方法 II となる。従つて方法 V 及び IV に対する分散公式が大抵の必要情報を与える。方法 I はしばしば用いられるけれどもなごい偏倚があるためこれを除いておこう。

方法は公式 (11.10) で表わされた比推定を考える事によって更に簡単に出来る。単位はある大きさの推定量 Σx_i に比例した確率で選ばれる。

推定量

$$\hat{R} = \frac{\sum_{i=1}^n \frac{M_i \bar{y}_i}{\Sigma x_i}}{\sum_{i=1}^n \frac{M_i \bar{x}_i}{\Sigma x_i}} \quad (11.10)$$

もしも x_{ij} が母集団に於ける各要素に対して 1 なる値をもつ "みせかけ" の変量即ち $x_{ij} = 1$ であるとするとき $\hat{R}$ は 11.8 表で与えられた様な $\bar{y}_i$ となる。従つて特別な場合として方法 I', III 及び V を含む。

比推定は亦それ自身の権利に於て有用である。もしも x_{ij} が以前の調査での y_{ij} の値であるとするれば次のものとして $\bar{y}$ 及び $\bar{x}$ の比推定量を作る事が出来る。

$$\bar{y}_R = \hat{R} \bar{x} : \bar{y}_R = \hat{R} \bar{x}$$

この型の推定量は North Carolina での農場項目に対する前の方法のいかなるものよりも正確である事が見出された。(L.H. Madour 1950, Jebe, 1952)

$V(\hat{R})$ を見出す際に、我々は既に $n=1$ なる場合に作られた分散公式の一つを使用しなければならない。最初に我々はおきかえでの抽出に於ける平均値の分散に対してよく知られた結果を必要とする。結果はむしろ一般的で任意に述べられる。

前提

特別な抽出及び推定方法に対して、標本推定量 u は分散 S_u^2

をもつある母集団特性 U の不偏推定量である。この標本が各抽出後もとにもどして n 回抽出されるとすれば推定量 $u_1, u_2, \dots, u_n$ を得る。従つて もしも $\bar{u}$ をして u_i の算術平均とすれば、

$$V(\bar{u}) = \frac{S_u^2}{n} \quad (11.11)$$

(証明)

$$(\bar{u} - U) = \frac{1}{n} ((u_1 - U) + (u_2 - U) + \dots + (u_n - U))$$

従つて

$$V(\bar{u}) = E(\bar{u} - U)^2 = \frac{1}{n^2} \sum_{i=1}^n E(u_i - U)^2 + \frac{2}{n^2} \sum_{i < j} E(u_i - U)(u_j - U)$$

おきかえをゆるす抽出では u_i に対して得られた値は u_j ($i \neq j$) に対して得られた値によつては影響されない。

もしも u_j を如何なる積でも一定とし u_i 上で平均すれば $E(u_i - U) = 0$ であるからその平均は 0 となる。亦各平方項は S_u^2 となる 従つて

$$V(\bar{u}) = \frac{1}{n^2} n S_u^2 = \frac{S_u^2}{n}$$

前提は如何なる標本が抽出されようとも特殊化されない。即ち抽出が等確率であるうが任意確率であろうが特殊化されない事に注意されたし。

定理 11.1

n 回の単位をもつ標本おきかえを許して抽出され、如何なる抽出でも i 番目の単位に割当てられる選択確率は Σx_i である。各選ばれた単位から大きさ m_i の副次標本が単純無作為標本によって抽出される。推定量は

$$\hat{R} = \frac{\sum_{i=1}^n \frac{M_i \bar{y}_i}{\Sigma x_i}}{\sum_{i=1}^n \frac{M_i \bar{x}_i}{\Sigma x_i}}$$

従って 大標本では

$$V(\hat{R}) \doteq \frac{1}{nX^2} \sum_{i=1}^N \left\{ \frac{1}{Z_i} (Y_i - RX_i)^2 + \frac{M_i(M_i - m_i)}{Z_i} \frac{S_{di}^2}{m_i} \right\} \quad (11.12)$$

此処に

$$S_{di}^2 = \frac{1}{(M_i - 1)} \sum_{j=1}^{M_i} \{ (y_{ij} - RX_{ij}) - (\bar{y}_i - R\bar{X}_i) \}^2$$

(証明)

$n=1$ の場合 (§ 11.3) に対する方法Ⅳの論議に於て、我々は $M_i \bar{y}_i / Z_i M$ は $\bar{Y}$ の不偏推定量を与える事が分った。我々は今おさかえで以て行われたこの様な n 個の抽出をもっている。添字 i をもて標本の i 番目の組を表わすとしよう、此処に同じ単位が一度以上表われる。従って n 個の量

$$\frac{M_i}{Z_i} \bar{y}_i = \hat{Y}_i \quad (\text{とおく})$$

は母集団総計 Y の各不偏推定量である。

$\hat{X}_n$ と同様な定義で以て

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

としよう。従って もしも $R = Y/X$ が母集団比であるとすれば

$$\hat{R} - R = \frac{\hat{Y}_n}{\hat{X}_n} - R = \frac{1}{\hat{X}_n} (\hat{Y}_n - R\hat{X}_n)$$

単純無作為抽出での比推定 (第六章) と同様 我々は近似として分母に於ける標本推定量 $\hat{X}_n$ が X でおさかえる事が出来ると仮定する。そうすれば

* 定理は標本がどの程度の大きさでなければならぬかには関係していない、通常の比推定 (第六章) と同様近似はおそらく もしも $\hat{R}$ の分子及び分母の変動係数が 0.1 以下であれば適當であろう。勿論この点については更に研究の必要はあるが。

$$\hat{R} - R \doteq \frac{1}{X} (\hat{Y}_n - R\hat{X}_n)$$

$$= \frac{1}{nX} \sum_{i=1}^n (\hat{Y}_i - R\hat{X}_i)$$

近似のこのオーターに対して、 $\hat{R}$ は R の不偏推定量である。前提により、変量 $u_i = \hat{Y}_i - R\hat{X}_i$ に適用すると

$$V(\hat{R}) \doteq \frac{1}{nX^2} V(\hat{Y}_i - R\hat{X}_i) \quad (11.13)$$

併し $n=1$ の場合の方法Ⅳの論議に於て、我々は公式 (11.7) 及び (11.7') から $\hat{Y}_i = M_i \bar{y}_i$ であるから

$$V(\hat{Y}_i) = \sum_{i=1}^N Z_i \left(\frac{Y_i}{Z_i} - \bar{Y} \right)^2 + \sum_{i=1}^N \frac{M_i(M_i - m_i)}{Z_i} \frac{S_{di}^2}{m_i}$$

この結果を変量 $\hat{Y}_i - R\hat{X}_i$ に適用しよう。この変量は $M_i \bar{d}_i / Z_i$ に等しい、此処に $\bar{d}_{ij} = y_{ij} - RX_{ij}$ (11.13) に代入すると

$$V(\hat{R}) \doteq \frac{1}{nX^2} \left\{ \sum_{i=1}^N Z_i \left\{ \frac{Y_i - RX_i}{Z_i} - (\bar{Y} - R\bar{X}) \right\}^2 + \sum_{i=1}^N \frac{M_i(M_i - m_i)}{Z_i} \frac{S_{di}^2}{m_i} \right\}$$

$Y = RX$ であるから、これは次の様にかかる

$$V(\hat{R}) \doteq \frac{1}{nX^2} \left\{ \sum_{i=1}^N \frac{1}{Z_i} (Y_i - RX_i)^2 + \sum_{i=1}^N \frac{M_i(M_i - m_i)}{Z_i} \frac{S_{di}^2}{m_i} \right\}$$

此処に

$$S_{di}^2 = \frac{1}{(M_i - 1)} \sum_{j=1}^{M_i} \{ (y_{ij} - RX_{ij}) - (\bar{y}_i - R\bar{X}_i) \}^2$$

系 もしも $m_i = k M_i / Z_i$ として $\hat{R}$ が簡単な Y/X となるとすれば

$$V(\hat{R}) \doteq \frac{1}{nX^2} \sum_{i=1}^N \left\{ \frac{1}{Z_i} (Y_i - RX_i)^2 + \frac{(M_i - m_i)}{k} S_{di}^2 \right\} \quad (11.12')$$

§ 11.9 で注意した様にこの結果は特別な場合として方法Ⅳ、Ⅲ及びⅠ' に対しての分散公式を与える。定理 11.2 は不偏な方法Ⅳ及びⅡに適用する。

定理 11.2

定理 11.1 と同じ選択方法で、 $\bar{y}$ の推定量は

$$\bar{y}_{\pi} = \frac{1}{nM} \sum_{i=1}^n \frac{M_i \bar{y}_i}{Z_i}$$

従って

$$V(\bar{y}_{\pi}) = \frac{1}{nM^2} \sum_{i=1}^N \left\{ \frac{1}{Z_i} (Y_i - Z_i \bar{y})^2 + \frac{M_i(M_i - m_i)}{Z_i} \frac{S_i^2}{m_i} \right\} \quad (11.14)$$

(証明)

前提により

$$\begin{aligned} V(\bar{y}_{\pi}) &= \frac{1}{nM^2} V\left(\frac{M_i \bar{y}_i}{Z_i}\right) \\ &= \frac{1}{nM^2} \sum_{i=1}^N \left\{ \frac{1}{Z_i} (Y_i - Z_i \bar{y})^2 + \frac{M_i(M_i - m_i)}{Z_i} \frac{S_i^2}{m_i} \right\} \quad ((11.7)より) \end{aligned}$$

系 もしも $m_i = k M_i / Z_i$ であれば $\bar{y}_{\pi} = \bar{y} / n k M$ であるので

$$V(\bar{y}_{\pi}) = \frac{1}{nM^2} \sum_{i=1}^N \left\{ \frac{1}{Z_i} (Y_i - Z_i \bar{y})^2 + \frac{(M_i - m_i)}{k} S_i^2 \right\} \quad (11.14')$$

分散公式 (11.12') 及び (11.14') は構造的にはむしろ簡単である。乗数因子を除くと、主な差は比推定では変量 $(y_{ij} - R x_{ij})$ が不偏推定量としてあらわれる変量 y_{ij} でおきかえられるという事である。比推定に対する公式は近似的である。抽出がおきかえを許して行われると仮定すると、不偏推定量に対しての公式は正確である。

§ 11.10 選択の最適確率

Optimum probabilities of selection

この分析は比推定に対して十分実行されるであろう。不偏推定量に対する分析は二つの分散公式の同じ構造のため同様である。それ等の大きさ又は良好な推定量が利用出来るという事を与えると、如何なる選択確率が単位に割当てられねばならぬか？ この問題は Hansen 及び Hurwitz (1949) によって検定された、彼等の分析はその結果及び技術の例として共に重要である。彼等の

考えた費用函数は三つの成分をもっている。

C_u = 一次単位当り一定費用

C_e = 要素即ち副次単位当り費用

C_L = 送ばれた単位で一つの要素をリストする費用及びリストさるべき要素の数と共に変わる他の費用

が三番目の費用因子は抽出者が常に送ばれた単位で要素をリストせねばならず副次標本を抽くためにその数を明らかにせねばならぬために含まれる。

従って

$$\text{費用} = C_u n + C_e \sum_{i=1}^n m_i + C_L \sum_{i=1}^n M_i$$

この公式は我々の目的のためには採用されない、何故ならば量 $\sum m_i$ 及び $\sum M_i$ はたまたま送ばれた単位による無作為変量であるからである。その代りに我々は平均費用を考えよう。単位が任意確率 Z_i で送ばれ、副次標本が $m_i = k M_i / Z_i$ で送ばれる。さて

$$E\left(\sum_{i=1}^n m_i\right) = \sum_{i=1}^N n Z_i m_i = n k \sum_{i=1}^N M_i = n k M$$

同様に

$$E\left(\sum_{i=1}^n M_i\right) = \sum_{i=1}^N n Z_i M_i = n \sum_{i=1}^N Z_i M_i$$

従って 平均費用は

$$C = C_u n + C_e n k M + C_L n \sum_{i=1}^N Z_i M_i \quad (11.15)$$

一定平均費用に対して分散を最小にしようと試みる際に、手元にある変量は n , k 及び確率 Z_i である。

定理 11.1 系により最小になるべき分散は

$$V(\bar{R}) = \frac{1}{nX^2} \sum_{i=1}^N \left\{ \frac{1}{Z_i} (Y_i - R X_i)^2 + \frac{(M_i - m_i)}{R} S_i^2 \right\}$$

微分を簡単にするためにこの表示を少し変化する。我々は亦

$Y_i = M_i \bar{y}_i$, $X_i = M_i \bar{x}_i$ を代入する.

更に m_i は

$$m_i = \frac{k M_i}{Z_i}$$

なる方程式を満足する様に選ばれるから右辺の量 m_i/k は m_i/Z_i でおきかえられる。この変化させたものは

$$V = X^2 V(\bar{R}) = \sum_{i=1}^N \frac{1}{n} \left\{ \frac{M_i^2}{Z_i} (\bar{y}_i - R \bar{x}_i)^2 - \frac{M_i}{Z_i} S_{di}^2 + \frac{M_i}{k} S_{di}^2 \right\}$$

もしも $d_{ij} = (y_{ij} - R x_{ij})$ とかけば量 $(\bar{y}_i - R \bar{x}_i)$ は i 番目の単位に於ける d_{ij} の要素当り平均値でこれを $\bar{d}_i$ で表わす。大括弧内の最初の二項を組合すと

$$V = \sum_{i=1}^N \frac{1}{n} \left\{ \frac{M_i^2}{Z_i} \left(\bar{d}_i^2 - \frac{S_{di}^2}{M_i} \right) + \frac{M_i}{k} S_{di}^2 \right\}$$

最後に n は組合せ $n Z_i$ 及び $n k$ でのみ表わされている事に注意しよう。変量 $Z_i' = n Z_i$ 及び $k' = n k$ を導入すると V はもはや n の明らかな函数ではない。従って

$$V = \sum_{i=1}^N \left\{ \frac{M_i^2}{Z_i'} \left(\bar{d}_i^2 - \frac{S_{di}^2}{M_i} \right) + \frac{M_i}{k'} S_{di}^2 \right\} \quad (11.16)$$

我々は費用が一定であり

$$\sum_{i=1}^N Z_i = 1 \quad ; \quad \text{即ち} \quad \sum_{i=1}^N Z_i' = n$$

なる制限を設けて n , k' 及び Z_i' に関して V を最小にする。未知常数として λ 及び μ をとりそのものを最小にする

$$V + \lambda (Cn + C_e n M + C_e \sum_{i=1}^N Z_i' M_i) + \mu \left(n - \sum_{i=1}^N Z_i' \right)$$

微分すると次のものを得る

$$n : \quad \lambda C + \mu = 0 \quad (11.17)$$

$$Z_i' : \quad \frac{-M_i^2}{Z_i'^2} \left(\bar{d}_i^2 - \frac{S_{di}^2}{M_i} \right) + \lambda C_e M_i - \mu = 0 \quad (11.18)$$

即ち

$$Z_i'^2 = \frac{M_i^2 \left(\bar{d}_i^2 - \frac{S_{di}^2}{M_i} \right)}{\lambda (C + C_e M_i)} \quad (11.19)$$

λ はすべての i に対して同じであるから、これ等の方程式は Z_i' に対して明確な解を与える。従って $Z_i = Z_i'/n$ (11.19) に於ける $Z_i'^2$ の分子は正と仮定される。(もしも負であれば副次抽出は非有効的で一次単位の一様抽出がすぐれている事が分る)。最適 Z_i に対して

$$Z_i = \frac{\frac{M_i D_{iu}}{\sqrt{C + C_e M_i}}}{\sum_{i=1}^N \frac{M_i D_{iu}}{\sqrt{C + C_e M_i}}} \quad (11.20)$$

此処に

$$D_{iu}^2 = \bar{d}_i^2 - \frac{S_{di}^2}{M_i}$$

今や D_{iu}^2 は単位の大きさを M_i によるから検定されなければならない。 D_{iu}^2 に似た量は違つた記号で前に出会った事がある(§8.4)。すべての単位が同じ大きさを M である母集団と考えよう。

変量 $d_{ij} = y_{ij} - R x_{ij}$ について $N \times N$ 表 (P242) と同様に分散分析をしよう。

11.9 表 母集団に対する分散分析

	df	m.s
一次単位間	$(N-1)$	$S_b^2 = M \sum_{i=1}^N \bar{D}_i^2 / (N-1)$
一次単位内要素間	$N(N-1)$	$S_a^2 = \bar{S}_{di}^2$

D_{iu}^2 の定義からすべての単位上でのその平均値は ($M=M$ と仮定)

$$\bar{D}_{iu}^2 = \frac{1}{N} \sum_{i=1}^N \bar{D}_i^2 - \frac{\bar{S}_{di}^2}{M}$$

従つて N が大きければ 11.9 表から D_{iu}^2 の平均値は再び次の様になる。

標にかける

$$\bar{D}_{cu}^2 = \frac{(S_b^2 - S_w^2)}{\bar{M}} \quad (11.21)$$

§ 9.5 に於て、我々は函数 S_b^2 及び S_w^2 が M と何様な関係にあるかを研究した。我々は実験公式として次のものを見出した。

$$S_w^2 = A \bar{M}^g \quad (g > 0)$$

$$S_b^2 = \bar{M} S^2 - (\bar{M} - 1) A \bar{M}^g$$

此処に S^2 は母集団に於けるすべての要素間分散である。方程式 (11.21) によりこれは

$$\bar{D}_{cu} = S^2 - A \bar{M}^g \quad (11.22)$$

を与える。

もしも $\bar{M}$ があまり大きくなければ S_w^2 が一定即ち $g = 0$ 従つて $\bar{D}_{cu}^2 = \text{一定}$ という仮定は § 9.5 で暗示した標にしばしば満足である。その他、方程式 (11.22) は $\bar{D}_{cu}$ が $\bar{M}$ が増すと減少する事が期待される事を示す。Hansen 及び Hurwitz (1943) は経験から $\bar{D}_{cu}^2$ は実際には $1/\bar{M}$ ほど早くは減少しないという事を暗示している。

我々は今や Z_c の最適選択を論ずる立場にある。

(11.20) から

$$Z_c \propto \frac{M_c D_{cu}}{\sqrt{C_u + C_e M_c}}$$

次の如き推論がこの結果からひき出される。

i. 一次単位当りリストする費用及びそれに関連した作業の費用 $C_e M_c$ は一次単位当り一定費用 C_u に比して相対的に小さいと仮定しよう。もしも D_{cu} が一定であれば PPS での選択が最適である。もしも D_{cu} が M_c の増大と共に減少すれば最適確率は $Z_c \propto M_c$ と $Z_c \propto \sqrt{M_c}$ との間にある。

ii もしもリストする費用が主であれば最適確率は $Z_c \propto \sqrt{M_c}$ と $Z_c = \text{一定}$ (等確率) との間にある。

iii もしもリストする費用及び一定費用が大きさが同じオーダーをもつとすれば、 $Z_c \propto \sqrt{M_c}$ が良好な選択策である。

微分により見出される最適値は読者の例題として残しておこう。

その値は

$$k = \frac{\sqrt{\sum M_c S_{c2}^2}}{\sqrt{M C_e} \sum \frac{M_c D_{cu}}{\sqrt{C_u + C_e M_c}}}$$

である。

最適値は k に対する費方方程式 (11.15) を解いて見出される。既に述べた様に、この節での論議はそれ等の大きさ又は良好な推定量が既知であると仮定している。標本を構成する単位に於て必要とされるリストする事に対して以外は、 N 個の単位の大きさについての情報を得る費用は割当ててはいない。

§ 11.11 偏倚対不偏推定量

Biased versus unbiased estimates

この節では我々は標本平均に基いた偏倚推定量がそれに応ずる不偏な方法よりも分散が小さい条件を検定する。一次単位に対する任意の選択確率 Z_c を仮定すると、その比較は 11.8 表の方法 V と方法 IV の内で行えばよい。推定量は

$$\bar{y} = \frac{\sum \frac{M_i}{Z_i} \bar{y}_i}{\sum \frac{M_i}{Z_i}} = \bar{y} \quad (\text{もしも } m_i = k M_i / Z_i \text{ であれば})$$

$$\bar{y}_{IV} = \frac{1}{nM} \sum \frac{M_i}{Z_i} \bar{y}_i = \frac{\bar{y}}{n k M} \quad (\text{もしも } m_i = k M_i / Z_i \text{ であれば})$$

構造的に $\bar{y}_{IV}$ は比推定 $\bar{y}'/k$ と同様である。此処に $y'_i = M_i \bar{y}_i / Z_i$ 及び $Z'_i = M_i / Z_i$ である。一方 $\bar{y}_{IV} = \bar{y} / M'$ “単位当り平均値” の推定量と同じである。通常の比推定の場合と同様に従つて我々は $\bar{y}_i$ と Z'_i との間で相関の大きさが $\bar{y}_{IV}$ がより正確であるかどうかを決定するであろう事を期待する。簡単を為め、二つの方

法で $m_i = k M_i / Z_i$ と仮定する。定理 11.1 系から

$$V(\hat{R}) = \frac{1}{nX^2} \sum_{i=1}^N \left\{ \frac{1}{Z_i} (Y_i - R X_i)^2 + \frac{(M_i - m_i)}{k} S_{di}^2 \right\}$$

分散 $V(\bar{y}_v)$ は $X_{ij} \equiv 1$ とおいた この結果の特別な場合として見出される。これは

$$X = M : X_i = M_i : R = \frac{Y}{X} = \frac{Y}{M} = \bar{y}$$

$$S_{di}^2 = S_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_i)^2$$

を与える。従って

$$V(\bar{y}_v) = \frac{1}{nM^2} \sum_{i=1}^N \left\{ \frac{1}{Z_i} (Y_i - M_i \bar{y})^2 + \frac{(M_i - m_i)}{k} S_i^2 \right\}$$

$V(\bar{y}_{lv})$ に対しては 定理 11.2 系から

$$V(\bar{y}_{lv}) = \frac{1}{nM^2} \sum_{i=1}^N \left\{ \frac{1}{Z_i} (Y_i - Z_i \bar{y})^2 + \frac{(M_i - m_i)}{k} S_i^2 \right\}$$

単位内への貢献は二つの推定量に於て同じである。

従って $\bar{y}_v$ は もしも

$$\sum \frac{1}{Z_i} (Y_i - M_i \bar{y})^2 < \sum \frac{1}{Z_i} (Y_i - Z_i \bar{y})^2$$

であれば小さい分散をもつ

これは次の採にかける

$$\sum Z_i \left(\frac{Y_i}{Z_i} - \bar{y} \frac{M_i}{Z_i} \right)^2 < \sum Z_i \left(\frac{Y_i}{Z_i} - \bar{y} \right)^2 \quad (11.23)$$

$Y'_i = Y_i / Z_i$, $X'_i = M_i / Z_i$ をして各一次単位に対して定義された変量としよう。単位は確率 Z_i で選ばれるから

$$E Y'_i = \sum Z_i \frac{Y_i}{Z_i} = Y$$

$$E X'_i = \sum Z_i \frac{M_i}{Z_i} = M$$

従って $\bar{y} = Y/M$ は Y'_i 対 X'_i の母集同比 R' である。

(11.23) の右辺は Y'_i の母分散であり、一方左辺は変量 $(Y'_i - R' X'_i)$ の母分散である。

$(Y'_i - R' X'_i)$ の分散を展開すると通常の比推定 (§ 6.7) と同様採もしも次の如くであれば $V(\bar{y}_v)$ は $V(\bar{y}_{lv})$ より小さい事が分る。

$$P > \frac{1}{2} \frac{M_i/Z_i \text{ の変動係数}}{Y_i/Z_i \text{ の変動係数}}$$

此処に P は Y_i/Z_i 及び M_i/Z_i との間の相関係数である。

一次単位が等確率で選ばれるときにはこの結果は次の如き簡単な条件となる。

$$\rho_{Y_i, M_i} > \frac{1}{2} \frac{M_i \text{ の C.V.}}{Y_i \text{ の C.V.}} \quad (11.24)$$

この条件は偏倚と不偏推定量の相対的正確度は一次単位総計 Y と一次単位の大さ M_i との間の相関係数の大きさによる事を示す。併し比較は大標本に關するものであり、それに於ては方法 V の推定量に於ける偏倚は無視された。

§ 11.12 推定分散

Estimated variances

既に分った採に、大抵の推定量は § 11.7 までの推定比から法 V の不偏推定量の何れかの特別な場合である。それ等に対する母の分散は定理 11.1 及び 11.2 に示されている。これ等の分散の標本推定量を今与えよう。

定理 11.3

n 個の一次単位の標本が確率 Z_i (おきかえで以て) で抽出される。各選ばれた単位から副次標本が単純無作為抽出で抽出される。推定量は

$$\hat{R} = \frac{\sum_{i=1}^n \frac{M_i}{Z_i} \bar{y}_i}{\sum_{i=1}^n \frac{M_i}{Z_i} \bar{x}_i}$$

である。

$\hat{R}$ の近似分散の不偏推定量は

$$v(\hat{R}) = \frac{1}{n^2 X^2} \sum_{i=1}^n \left\{ \frac{M_i}{Z_i m_i} (y_i - R x_i) \right\}^2 \quad (11.25)$$

此処に R は母集団比 Y/X である。

(証明)

n 個の連続抽出に対する一次単位標本推計を $y_1, y_2, \dots, y_n$ で表わす。同じ単位は一度以上表われるけれども我々はそれが表われる各場合に別個な添字をつける事に注意しよう。

$$d_{ij} = y_{ij} - R x_{ij} \quad \text{としよう。}$$

従って

$$\begin{aligned} \frac{M_i}{Z_i m_i} (y_i - R x_i) &= \frac{M_i}{Z_i m_i} d_i & d_i &= \frac{M_i}{Z_i} \bar{d}_i \\ &= \frac{M_i}{Z_i} \bar{D}_i + \frac{M_i}{Z_i} (\bar{d}_i - \bar{D}_i) & \bar{D}_i &= \frac{D_i}{Z_i} + \frac{M_i}{Z_i} (\bar{d}_i - \bar{D}_i) \end{aligned}$$

$$\text{此処に } \bar{D}_i = \bar{y}_i - R \bar{x}_i \quad \text{及び} \quad D_i = M_i \bar{D}_i = y_i - R x_i$$

さて i 番目の単位からのすべての副次標本上で平均して平均しよう。

$$E \left(\frac{M_i d_i}{Z_i m_i} \right)^2 = \left(\frac{D_i}{Z_i} \right)^2 + \frac{M_i (M_i - m_i)}{Z_i^2} \frac{S d_i^2}{m_i}$$

従って n 個の単位の一様選択に対して

$$\begin{aligned} E \sum_{i=1}^n \left(\frac{M_i d_i}{Z_i m_i} \right)^2 &= E \left[\sum_{i=1}^n \left(\frac{D_i}{Z_i} \right)^2 + \sum_{i=1}^n \frac{M_i (M_i - m_i)}{Z_i^2} \frac{S d_i^2}{m_i} \right] \\ &= \sum_{i=1}^N x_i \left(\frac{D_i}{Z_i} \right)^2 + \sum_{i=1}^N \frac{x_i M_i (M_i - m_i)}{Z_i^2} \frac{S d_i^2}{m_i} \end{aligned}$$

此処に x_i は i 番目の単位が特別な標本に於て抽出された回数とする。 n 個の単位の上で平均するとき、

$$E(x_i) = n Z_i$$

従って

$$E \sum_{i=1}^n \left(\frac{M_i d_i}{Z_i m_i} \right) = n \sum_{i=1}^N \left\{ \frac{D_i^2}{Z_i} + \frac{M_i (M_i - m_i)}{Z_i} \frac{S d_i^2}{m_i} \right\}$$

併し 定理 11.1 公式 (11.12) によって

$$D_i = y_i - R x_i \quad \text{であるから}$$

$$V(\hat{R}) = \frac{1}{n X^2} \sum_{i=1}^N \left\{ \frac{D_i^2}{Z_i} + \frac{M_i (M_i - m_i)}{Z_i} \frac{S d_i^2}{m_i} \right\}$$

証明終り ($v(\hat{R})$ に於ける分母は $n^2 X^2$ である事に注意) 系

定理 11.3 は、 R 及びある応用では X も亦未知であるから、そのまゝでは用いられない。 R の代りに標本推定量 $\hat{R}$ 、又オム章での通常の比推定の場合行った様に分母の n を $(n-1)$ で置きかえる。 X の不偏推定量は

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n \frac{M_i \bar{x}_i}{Z_i}$$

従って我々は次のものを用いる

$$v(\hat{R}) = \frac{1}{n(n-1)X^2} \sum_{i=1}^n \left\{ \frac{M_i}{Z_i m_i} (y_i - R x_i) \right\}^2 \quad (11.26)$$

(例)

11.7 表 (P.265) の方法 I' によって推定平均年令の標準誤差を計算しよう。この例では2の頁が等確率で $m_i = 2$ で以て選ばれた推定量は

$$\bar{y}_I = \frac{\sum M_i \bar{y}_i}{\sum M_i} = \frac{17.121.5}{359} = 47.6922 \text{ 年}$$

余計な術数は後の計算で正確を期すために出して置いた。

公式 (11.26) を適用するため

$$Z_i = \frac{1}{N} : m_i = 2 : \hat{R} = \bar{y}_I : x_i = m_i = 2 : \hat{X} = \frac{N}{u} \sum M_i$$

とおく (11.26) に代入すると、

$$V(\hat{R}) = \frac{n}{(n-1)(\sum M_h)^2} \sum_{h=1}^n \left\{ M_h (\bar{y}_h - \bar{y}_{..}) \right\}^2 \quad (11.27)$$

平方和は P(265) 11.7 表から次の形で容易に計算される

$$\begin{aligned} \sum (M_h \bar{y}_h)^2 - 2 \bar{y}_{..} (\sum M_h \bar{y}_h) + \bar{y}_{..}^2 \sum M_h^2 \\ = 15,375,020 - (95,3844)(309,725.5) + (2274.55)(6481) \\ = 571,300 \end{aligned}$$

$$V(\hat{R}) = \frac{(20)(571,300)}{(19)(259)^2} = 4.67$$

$$S(\hat{R}) = 2.16 \text{ 年}$$

公式 (11.27) は一般抽出での比推定の分散を計算するのに用いた P 127 の公式 (6.14) と同じである事が分る (f.p.c. $(N-n)/N$ を除いた)。従って我々は比処で $y'_h = M_h \bar{y}_h$, $x'_h = M_h$ において比 y'_h/x'_h に対する分散を計算する事によって公式 (6.14) を用いる事が出来るであろう。これは亦第10章で注意した一般結果の特別な場合である。もしも二様抽出で n/N が無視出来れば推定分散は一般抽出に対して用いた公式によって見出し得る。この事は二様抽出の相対的複雑性にも拘わらず、推定分散に対する公式が簡単なまゝであるから都合のよい事である。

次の結果は方法 IV に於ける不偏推定量に対する分散の標本推定量を与える。

定理 11.4

定理 11.3 と同じ抽出方法では、 $\bar{y}_{..}$ の推定量は

$$\bar{y}_{..} = \frac{1}{nM} \sum_{h=1}^n \frac{M_h \bar{y}_h}{Z_h}$$

従って $V(\bar{y}_{..})$ の不偏標本推定量は

$$V(\bar{y}_{..}) = \frac{1}{n(n-1)M^2} \sum_{h=1}^n (\hat{y}_h - \bar{y}_{..})^2 \quad (11.29)$$

$$\text{比処に } \hat{y}_h = \frac{M_h \bar{y}_h}{Z_h} : \hat{\bar{y}}_{..} = \frac{1}{n} \sum_{h=1}^n \hat{y}_h = M \bar{y}_{..}$$

証明は定理 11.3 と同じ方法で得られる。

この章で論じたすべての方法に対する分散の推定量はこれ等の定理から導く事が出来る。それ等は推定量の自己荷重の形で 11.10 表に示す。

11.10 表 母平均の推定量に対する標本分散

方法	単位確率	m_h	$\bar{y}$ の推定量	分散の標本推定量 (自己荷重標本に対する)
I	等確率	kNM_h	$\bar{y}$	$\frac{n}{(n-1)n^2} \sum \{m_h(\bar{y}_h - \bar{y})\}^2$
II	等確率	kNM_h	$\frac{y}{nM}$	$\frac{1}{n(n-1)(kM)^2} \sum (y_h - \bar{y})^2$
III	$\frac{M_h}{M}$	m	$\bar{y}$	$\frac{1}{n(n-1)m^2} \sum (y_h - \bar{y})^2$
IV	Z_h	$\frac{kM_h}{Z_h}$	$\frac{y}{nM}$	$\frac{1}{n(n-1)(kM)^2} \sum (y_h - \bar{y})^2$
V	Z_h	$\frac{kM_h}{Z_h}$	$\bar{y}$	$\frac{n}{(n-1)n^2} \sum \{m_h(\bar{y}_h - \bar{y})\}^2$

もしも自己荷重の条件が適用されなければ、定理 11.3 及び 11.4 から直接に分散の標本推定量を得なければならぬ。11.10 表での公式はもとにもどす場合の抽出を仮定し、単位がもとにもどされる事なしに送られるときは方法 I 及び II に適用する f.p.c. は無視する公式は仮りに n/N が小であっても適當である。

S 11.13 層化抽出の拡張

Extension to stratified sampling:—

不偏な方法 II、III 及び IV に対して、層化拡張は容易である添字 h は層 E を表す； M_h は層 h に於ける要素の総数、 M は

母集団に於ける要素の総数である。推定母平均は

$$\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h, \quad W_h = M_h/M \quad (11.30)$$

此処に $\bar{y}_h$ は $\bar{y}_h$ なる要素当り層平均の推定量を表わす。

更に

$$V(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 V(\bar{y}_h) \quad (11.31)$$

及び推定分散は

$$v(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 v(\bar{y}_h) \quad (11.32)$$

単一層に対しての $v(\bar{y}_h)$ の値を与える。11.10表はこれら分散の推定量を作る際に有用である。

層化抽出の 11.8表の一部の拡張である 11.11表は三つの不偏推定量の代数的形式を表わす。

欄頭 m_{hi} は層内で推定量を自己荷重させる副次標本の大きさを示す。推定量の自己荷重の形は一番右の欄に与えてある。もしも m_{hi} が推定量を層内で自己荷重させる様に選ばれたかつたならば一般形を用いる。表内での和は i は 1 から L 迄, j は 1 から n_h 迄である。

11.11表 層化二段抽出に於ける不偏推定量

方法	層内単位確率	m_{hi}	$\bar{y}$ の標本推定量	
			一般形式	自己荷重形式
II	等確率	$n_h N_h M_{hi}$	$\frac{1}{N} \sum_{h=1}^L \sum_{i=1}^{n_h} M_{hi} \bar{y}_{hi}$	$\frac{1}{N} \sum_{h=1}^L \frac{M_h}{n_h} \bar{y}_h$
III	$\frac{M_{hi}}{M_h}$	$\overline{m_h}$		$\frac{1}{N} \sum_{h=1}^L \frac{M_h}{n_h} \bar{y}_h$
IV	$\frac{M_{hi}}{Z_{hi}}$	$\frac{n_h M_{hi}}{Z_{hi}}$	$\frac{1}{N} \sum_{h=1}^L \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{Z_{hi}} \bar{y}_{hi}$	$\frac{1}{N} \sum_{h=1}^L \frac{1}{n_h} \bar{y}_h$

11.11表の一番右の欄は本推定量が完全な自己荷重となる条件を示す。 m_{hi} が層 h 内のすべての抽出比である事を思いおこすであろう。従つてもしもすべての抽出比をすべての層内で一定にあれば、推定量 II 及び IV は完全な自己荷重である。同じ結果が方法 III に於ける PPS に対しても成立する。何故ならばこの方法では一次単位当り選ばれた要素の数 m_{hi} は層内で一定であるから

$m_{hi} = \bar{m}_h / M_h$ は又層 h でのすべての抽出比であるからである。

偏倚推定量 I' 及び V 及び比推定では、公式 (11.30) と同様の層毎の層に対して推定量の荷重平均を用い得るであろう。併しすべて三つの方法は本質的には比推定であるから、偏倚はすべての層内で同じ符号をもち、もしも層の数が多いければ各層からの小標本では、荷重平均に於ける偏倚は大きいものとなる。この種の推定量は各層で大標本をもつ小数の層であれば又はすべての偏倚が無視出来るという証拠があればおすすめてもよい。

その他は最も一般的な場合に示される組合せ比推定を作る事である。

$$\hat{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{Z_{hi}} \bar{y}_{hi}$$

としよう。 $\hat{X}_h$ についてはそれに応じて定義しよう量 $\hat{Y}_h, \hat{X}_h$ は夫々層総計 Y_h, X_h の不偏推定量である。

組合せ比推定量は次の様に定義される

$$\hat{R}_c = \frac{\sum_{h=1}^L \hat{Y}_h}{\sum_{h=1}^L \hat{X}_h}$$

$\hat{R}_c$ の近似分散は通常の方法で次の様につく事によつて算出される。

$$\hat{R}_c - R \approx \frac{1}{X} \sum_{h=1}^L (\hat{Y}_h - R \hat{X}_h) \quad (11.33)$$

量 $(\hat{Y}_h - R \hat{X}_h)$ は 変数 $d_{hij} = y_{hij} - R x_{hij}$ の層総計 $(Y_h - R X_h)$ の不偏な方法 IV の推定量である。従つて $V(\hat{Y}_h - R \hat{X}_h)$

は、定理 11.2 の特別な場合として見出される。これは次の様な結果を導く。

$$V(\hat{R}_c) = \frac{1}{X^2} \sum_{k=1}^L \frac{1}{n_k} \sum_{i=1}^{N_k} \left\{ \frac{1}{Z_{ki}} (D_{ki} - Z_{ki} \bar{D}_k)^2 + \frac{M_{ki}(M_{ki} - m_{ki})}{Z_{ki}} \frac{S_{d_{ki}}^2}{m_{ki}} \right\} \quad (11.34)$$

此処に

$$D_{ki} = Y_{ki} - R X_{ki}$$

$$S_{d_{ki}}^2 = \frac{1}{(M_{ki} - 1)} \sum_{j=1}^{M_{ki}} \left\{ (y_{kij} - R X_{kij}) - (\bar{Y}_{ki} - R \bar{X}_{ki}) \right\}^2$$

方法Ⅰ及びⅠ'に対する分散を導くために 公式(11.34)に

於て $X_{kij} = 1$ とおく。すると

$$X = M : R = \bar{Y}$$

$$D_{ki} = Y_{ki} - \bar{Y}_{ki}$$

$$S_{d_{ki}}^2 = S_{Y_{ki}}^2 = \frac{1}{(M_{ki} - 1)} \sum_{j=1}^{M_{ki}} (y_{kij} - \bar{Y}_{ki})^2$$

(11.34)と同様その結果の分散公式は此の形の近似であり

n_k/N_k は小さいと仮定している。他の特別な場合は次の通りである。

方法Ⅱ。

層内の大きさに比例した確率 $z_{ki} = M_{ki}/M_k$

推定量 : もしも $m_{ki} = \bar{m}_k$ であれば 11.11 表 のものと同じ

分散公式は (11.31) か (11.34) の何れかから導く事が出来る。

$$V(\bar{y}_{mise}) = \frac{1}{M^2} \sum_{k=1}^L \frac{M_k}{n_k} \sum_{i=1}^{N_k} \left\{ M_{ki} (\bar{Y}_{ki} - \bar{Y})^2 + \frac{(M_{ki} - m_{ki})}{m_{ki}} S_{Y_{ki}}^2 \right\}$$

層内の単位に対する最適抽出比 n_k/N_k の選択は第五章で与える方法で出来るが此処では論じない。

$\hat{R}$ の推定分散に対しては、方程式 (11.33) から

$$V(\hat{R}_c) = \frac{1}{X^2} \sum_{k=1}^L V(\bar{Y}_k - R \bar{X}_{k1})$$

旧々の層に於て変量 d_{kij} に定理 11.4 を適用して

$$V(\hat{R}_c) = \frac{1}{X^2} \sum_{k=1}^L \frac{1}{n_k(n_k-1)} \sum_{i=1}^{n_k} (d_{ki}' - \bar{d}_k')^2 \quad (11.35)$$

此処に

$$d_{ki}' = \frac{M_{ki} \bar{d}_{ki}}{Z_{ki}} : \bar{d}_k' = \frac{1}{n_k} \sum_{i=1}^{n_k} d_{ki}'$$

$$\bar{d}_{ki}' = \bar{Y}_{ki} - \hat{R}_c \bar{X}_{ki}$$

もしも X が未知であれば、標本推定量 $\Sigma \hat{X}_k$ をその代りに用いる。

§ 11.14 むすびの註釈

Summary comments

この章で論議が示す様に、単位の大きさが等しくない場合の二段抽出の有効な計画は理想的な前以ての仕事が必要である。この節は簡単に主な論文を紹介しよう。

1. 大きさが分っているが、近似的に分っているか又は未知であるかを見出しなさい。未知である場合は大きさについての情報が容易に得られるかどうかを考えなさい。例えば Jellison 等 (1947) は、あるギリシヤの町でブロックの二段抽出を指導した。それに於てブロック当り家族数の有用な推定量が利用出来なかった。彼等は三つの案口を考えた。即ち (i) 等確率でブロックを抽出する。(ii) 大ざっぱに同じ家族数であると思わしめる様な人工的ブロックを作る様に小さなブロックを集めるためジープで町を急いでめぐった。明かに家族のないブロックはこの操作でなくなつた。標本ブロックがそれから等確率で選ばれた。(iii) 各ブロックの家族数を推定するため各分る建こで1つくり面をまわつたところからブロックを推定大きさに比例した確率で選んだ。

2 層化に対する変量の一つとして単位の大きさを用いるかどうかを考えなさい。これは正確度がものすごく増大するある他の

変量を用いる事の障害とならねば推奨出来る。

3. 如何なる単位が層内で送ばるべきかをきめなさい。もしも大きさが少くとも近似的にでも分つておれば PPS 抽出又はその平方根での抽出がたとえこれは調査費用の性質如何によるけれども最良の方法であろう。

4. 推定方法を選択しなさい。母集団平均及び総計を推定するためには、最近の調査での同じ項目の値を用いる比推定が利用出来れば非常にうまくゆく。標本平均又は荷重標本平均に基いた推定量は不偏推定量よりも正確な事がある。

5. 層内抽出比及び副次抽出比をきめなさい。副次抽出比は推定量が層内で自己荷重をもつ様に送ばれる事をおすすめる。標本が完全自己荷重である様な更に進んだやり方はそれが正確度を着しく減する事がなければおすすめる出来る。一次単位の大さが等しくない二段抽出を行う実際の調査計画並に指導の説明は論文 *A Chapter in population sampling* (1950) 及び *Gray and Corlett* (1950) 及び *Gates* (1949) の印刷物に示されている。

第12章 二重抽出法

Double Sampling

§ 12.1 方法の記述

Description of the technique.

今迄分つた様に、多くの抽出方法は任意変量 x_i についての前以ての情報の獲得如何による。比及び回帰推定は、母平均 $\bar{y}$ についての知識が必要である。もしも x_i の値によって母集団を層化し得れば、それらの度数分布が分らなければならぬ。

この様な情報がない場合には、 x_i のみを測定する大きな一次標本をとる事があまり費用がかゝらない場合がある。この標本の目的は、 $\bar{y}$ の良好な推定量をあたえる事か又は x_i の度数分布を与える事である。その函数がある他の変量 y_i に対して推定出来る様な調査では、たとえこれは y_i についての主調査に於て標本の大きさを減じなければならぬという意味もあるけれどもこの一次標本に大部分の資源とせよこむものである。この方法は二重抽出又は二段抽出法として知られている。論議から分る様に、この方法は、もしも比又は回帰推定又は層化からの正確度の利益が、主標本の大きさの減少による正確度の損失をより以上相殺するとすればそのときのみ利益がある。

二重抽出は、 x_i についての情報が今迄表にされなかった綴り込みカードにある場合などは非常に有用である。例えば 1945 年のドイツの市民人口調査に於て、町からの標本は通常食糧配給通帳から抽出された。それに対しての資料が通常利用出来たところの町内の地理的層化に加えて、性及び年齢による層化が要求された。標本は急いで抽出されねばならず、表はたえず使用されていながら、完全な年齢及び性の分布の表は実際には用いる事は出来なかった。併し適当な大きい組織標本が迅速に抽出出来た。抽出された人を使用してゐる年齢-性クラスに分類した。これ等の資

料から訪向さるべきごく少数の人が選ばれた。

§ 12.2 層化に対する二重抽出

Double sampling for stratification :-

理論は最初に Neyman により与えられた (1938)。

母集団は X_i の値に従って多くのクラスに層化される。最初の標本は大きさ n の単純無作為標本である。

次の様にしよう。

$w_h = N_h/N$ = 層 h に属する母集団の割合

$w'_h = n'_h/n$ = 層 h に属する最初の標本の割合。

従って w'_h は w_h の推定量である。

第二番目の標本はそれに於て y_h が測定される大きさ n_h なる層化無作為標本である： n_h なる単位が層 h から抽出される。第二の標本はしばしば第一の標本から副次標本であるかもしもより便利であれば独立に抽出される。

二つの標本の費用は次の型であると仮定される。

$$C = nC_n + \sum_{h=1}^k n'_h C'_h \quad (12.1)$$

此処に C_n は通常 C'_h に比して大きい。

問題は与えられた費用に対して推定量の分散を最小にする様に n 及び n'_h (従って n) を選ぶ事である。従って我々は最小分散がそれに於て y_h のみが測定される単一の無作為標本に得られ得るものよりも小さいかどうかを明らかにしなければならぬ。

第一段階は推定量を作りその分散を決定する事である。

母平均は

$$\bar{Y} = \sum_{h=1}^k w_h \bar{Y}_h$$

不偏推定量として我々は次のものを用いる

$$\bar{y}_{st} = \sum_{h=1}^k w'_h \bar{y}_h$$

新しい標本が選ばれる時は何時でも、これは第一及び第二の標本

共に新しい抽出を意味する。従って w_h 及び標本平均 $\bar{y}_h$ は共に誤差をうける無作為変量である。従って問題はそれに於て層統計が未だ正確に分らない層化の問題である。層統計はくりかえしの抽出に於て一定と仮定される。

定理 12.1

推定量 $\bar{y}_{st}$ は不偏である。

(証明)

$$w'_h = \bar{w}_h + u_h \quad ; \quad \bar{y}_h = \bar{Y}_h + l_h \quad \text{とかこう}$$

従って推定量の誤差は次のものとして表わされる。

$$\begin{aligned} \bar{y}_{st} - \bar{Y} &= \sum_{h=1}^k (w'_h \bar{y}_h - w_h \bar{Y}_h) \\ &= \sum_{h=1}^k (w_h l_h + \bar{Y}_h u_h + u_h l_h) \quad (12.2) \end{aligned}$$

単純無作為標本の特性によって量 u_h 及び l_h はすべて平均値は0である。更に抽出の方法によつて u_h 及び l_h は独立に分布している。従って

$$E(\bar{y}_{st} - \bar{Y}) = 0$$

$V(\bar{y}_{st})$ に対する定理に於いて抽出比 n/N , n_h/N_h は無視出来ると仮定する。何故なら、これ等の仮定は多くの応用面に於て妥当と考えられるからである。層 h に於ける y_{hi} の分散は通常の様子 S_h^2 で表わす。

定理 12.2

もしも n/N , n_h/N_h が無視出来れば

$$V(\bar{y}_{st}) = \sum_{h=1}^k \left\{ \left\{ w_h^2 + \frac{w_h(1-w_h)}{n} \right\} \frac{S_h^2}{n_h} + \frac{w_h(\bar{Y}_h - \bar{Y})^2}{n} \right\} \quad (12.3)$$

(証明)

(12.2) から

$$V(\bar{y}_{st}) = E(\bar{y}_{st} - \bar{Y})^2 = E\left(\sum_{h=1}^k (w_h l_h + \bar{Y}_h u_h + u_h l_h)\right)^2 \quad (12.4)$$

まず平方の項を考えよう。これ等は (12.4) のすべての他の項は期待値をとると0となるから

$$E\left\{\sum_k (w_k l_k + \bar{y}_k u_k + u_k l_k)\right\} \\ = \sum_k (w_k^2 E(l_k^2) + \bar{y}_k^2 E(u_k^2) + E(u_k^2)E(l_k^2)) \quad (12.4)$$

n/n' は無視出来ると仮定するから変量 u_k は多項分布に従う。従って $E(u_k^2) = n' w_k (1 - w_k)$ 。従って平方の項は次の如くなる。

$$\sum_k \left\{ \frac{w_k^2 S_k^2}{n_k} + \frac{\bar{y}_k^2 w_k (1 - w_k)}{n'} + \frac{w_k (1 - w_k)}{n'} \cdot \frac{S_k^2}{n_k} \right\} \quad (12.5)$$

さて方程式 (12.3') に於ける聳つた層間の積の項を考えよう。もしも k 及び j が二つの層に関するものとするれば、抽出比は違つた層では独立であるから $l_k l_j$ なる形の項からの貢献はなくなる。0でない貢献は $\bar{y}_k \bar{y}_j u_k u_j$ に於ける項からのもののみである。多項分布に対して

$$E(u_k u_j) = -\frac{w_k w_j}{n'}$$

であるから積の項は

$$-2 \sum_{k < j} \frac{\bar{y}_k \bar{y}_j w_k w_j}{n'} \quad (12.6)$$

に貢献する

もしも (12.5) の真中の項を (12.6) と組合せると、これを一緒にして

$$\sum_k \frac{w_k (\bar{y}_k - \bar{y})^2}{n'}$$

となる事は明らかである。

従って

$$V(\bar{y}_{st}) = \sum_k \left\{ \left\{ w_k^2 + \frac{w_k (1 - w_k)}{n'} \right\} \frac{S_k^2}{n_k} + \frac{w_k (\bar{y}_k - \bar{y})^2}{n'} \right\}$$

n' と含まない項は層の大きさが正確に分っている時の分散に対してと同様な表示である。最初の標本に於ける誤差の影響は従って分散に対しての層内の貢献とかなり増大する事であり層間の成分と導く事である

系。

もしも我々が第二の標本に於ける割合を推定したければ

$$S_k^2 = \frac{(N_k)}{(N_k - 1)} p_k Q_k \doteq p_k Q_k$$

$$(\bar{y}_k - \bar{y})^2 = (p_k - p)^2$$

又定理 12.2 より

$$V(p_{st}) \doteq \sum_k \left\{ \left\{ w_k^2 + \frac{w_k (1 - w_k)}{n'} \right\} \frac{p_k Q_k}{n_k} + \frac{w_k (p_k - p)^2}{n'} \right\} \quad (12.7)$$

比述に k は層 k に於ける割合である。

§ 12.4 最適割当て

Optimum allocation

最小分散を導く n_k 及び n の値はかなり複雑である。公式 (12.3) から n_k は

$$S_k \sqrt{w_k^2 + \frac{w_k (1 - w_k)}{n'}}$$

に比例する事は明らかである。

平方根内の第二番目の項は通常第一番目の項に比して小さいので Neyman (1938) は n_k を $w_k S_k$ に比例してとる事を暗示した。従って

$$n_k = \frac{n w_k S_k}{\sum (w_k S_k)}$$

これ等の値を分散 (12.3) に代入し $w_k (1 - w_k)$ の項と無視すると

$$V_{opt} \doteq \frac{(\sum w_k S_k)^2}{n} + \frac{\sum w_k (\bar{y}_k - \bar{y})^2}{n'} \quad (12.8)$$

$$= \frac{V_n}{n} + \frac{V_{n'}}{n'} \quad (\text{とおく}) \quad (12.8')$$

分散に対する近似的表示は以前与えた

$$C = n C_n + n' C_{n'} \quad (12.1)$$

なる形の与えられを費用に対して n 及び n' の選択によって今や最小にとれる。

次の事は容易に分る

$$\frac{n}{\sqrt{V_n C_n}} = \frac{n'}{\sqrt{V_{n'} C_{n'}}} \quad (12.9)$$

この方程式及び(12.1)は n 及び n' を決定する

最小分散に対する表示は二重抽出の後の応用に対して必要となる
う。(12.9) から

$$\frac{n}{\sqrt{V_n C_n}} = \frac{n'}{\sqrt{V_{n'} C_{n'}}} = \frac{n C_n + n' C_{n'}}{\sqrt{C_n C_{n'}} (\sqrt{V_n C_n} + \sqrt{V_{n'} C_{n'}})} \\ = \frac{C}{\sqrt{C_n C_{n'}} (\sqrt{V_n C_n} + \sqrt{V_{n'} C_{n'}})} \quad (12.9')$$

V_{opt} に対して方程式(12.8')にこれ等の解を代入しよう。すると

$$V_{opt} = \frac{(\sqrt{V_n C_n} + \sqrt{V_{n'} C_{n'}})^2}{C} \quad (12.10)$$

(例)

この例は人工的ではあるがそこに含まれる計算を示す。以前
考えた() Jefferson の郡の資料を用いる。 x_i なる変量即ち
農場の大きさは母集団を二つの層に分けるのに用いられる。即
ち二つの層は180エーカーまでの農場及び160エーカー以上の農場
である。穀物の面積(y_i)の標本をとるのに農物の大きさ(x_i)
をとるの、10倍の費用がかゝると仮定すると費用は次の形と
なる。

$$C = 100 = n + 0.1n' \quad (12.11)$$

これはもしも二重抽出が用いられないとすれば($n' = 0$)、
穀物面積推定に対して100の農場の標本をとる事が出来る事
を意味する。

母集団に関する資料は次の通りである。

層	W_h	S_h^2	S_h	$\bar{Y}_h$
1	0.786	312	17.7	19.404
2	0.214	922	30.4	51.626
母集団		620		26.297

公式(12.10)により直ちに V_{opt} を計算出来る。併し中間の段
階が与えられるであろう。即ち

$$V_n = (\sum W_h S_h)^2 = 417$$

$$V_{n'} = \sum W_h (\bar{Y}_h - \bar{Y})^2 = 175$$

従って、公式(12.9)により

$$\frac{n}{n'} = \sqrt{\frac{417}{175} \cdot \frac{1}{10}} = 0.488$$

費用方程式(12.11) から

$$n' = \frac{100}{0.588} = 170 ; \quad n = 170 \times 0.488 = 83$$

この例でこの例に於ける資料から分散公式(12.3)に於ける
 $W_h(1-W_h)$ なる無視した項は事実無視出来る事が分るのである
う。公式(12.8) から

$$V_{opt} = \frac{417}{83} + \frac{175}{170} = 5.02 + 1.03 = 6.05$$

二重抽出を用いない大きさ100なる単純無作為標本に対して

$$V = \frac{620}{100} = 6.20$$

二重抽出からわずかな利益しかない様である。

§ 12.4 層化に対する二重抽出に於ける推定分散

Estimated variance in double sampling for stratification.

n'/N n_h/N_h が無視出来ると仮定すると、 $\bar{y}_{st}$ の真の分散は
(12.3) から

$$V(\bar{y}_{st}) = \sum_{h=1}^L \left[\left\{ W_h^2 + \frac{W_h(1-W_h)}{n'} \right\} \frac{S_h^2}{n_h} + \frac{W_h (\bar{Y}_h - \bar{Y})^2}{n'} \right]$$

もしも標本からの推定量とこの量に代入するとその結果の表示
は $V(\bar{y}_{st})$ の過大推定量となる。不偏推定量はさして困難もな
く作る事が出来る。

定理 12.3

$V(\bar{y}_{st})$ の不偏推定量は

$$v(\bar{y}_{st}) = \frac{n'}{(n'-1)} \sum_h \left[\left\{ w_h^2 - \frac{w_h}{n_h} \right\} \frac{s_h^2}{n_h} + \frac{w_h (\bar{y}_h - \bar{y}_{st})^2}{n'} \right] \quad (12.12)$$

(証明)

これは(12.12)に $W_h = W_h + u_h$, $\bar{Y}_h = \bar{Y}_h + e_h$ を代入

すると得られる。大括弧内の次々の項の期待値は次の様にかかる。

$$E \sum_k W_k^2 \frac{s_k^2}{n_k} = \sum_k \left\{ W_k^2 + \frac{W_k(1-W_k)}{n'} \right\} \frac{s_k^2}{n_k} \quad (12.13)$$

$$-E \sum_k \frac{W_k}{n'} \cdot \frac{s_k^2}{n_k} = -\sum_k \frac{W_k}{n'} \cdot \frac{s_k^2}{n_k} \quad (12.14)$$

$$E \sum_k \frac{W_k (\bar{y}_k - \bar{y}_{st})^2}{n'} = E \left\{ \sum_k \frac{W_k \bar{y}_k^2}{n'} - \frac{\bar{y}_{st}^2}{n'} \right\} \\ = \sum_k \frac{W_k \bar{y}_k^2}{n'} + \sum_k \frac{W_k}{n'} \frac{s_k^2}{n_k} - \frac{\bar{y}^2}{n'} - \frac{V(\bar{y}_{st})}{n'} \quad (12.15)$$

(12.13) (12.14) 及び (12.15) を加えて $(n'-1)E \psi(\bar{y}_{st})/n'$ に対して次のものを得る。

$$\sum_k \left\{ \left\{ W_k^2 + \frac{W_k(1-W_k)}{n'} \right\} \frac{s_k^2}{n_k} + \frac{W_k (\bar{y}_k - \bar{y})^2}{n'} \right\} - \frac{V(\bar{y}_{st})}{n'} \\ = \frac{(n'-1) V(\bar{y}_{st})}{n'}$$

証明終り

もしも n' が n_k に比して相対的に大きいと $\psi(\bar{y}_{st})$ は次の如くなる。

$$\psi(\bar{y}_{st}) \doteq \sum_k W_k \frac{s_k^2}{n_k} \quad (12.16)$$

この表示は層荷重 W_k に於ける誤差が無視出来ると仮定したのと同じである。

系

もしも p_k がある定まったクラスに属する層 k に於ける単位の観測される割合とすれば、又 $p_{st} = \sum W_k p_k / \sum W_k$ が母集団割合の推定量とすれば $V(p_{st})$ の推定量は

$$V(p_{st}) = \frac{n'}{(n'-1)} \sum_k \left\{ \left(W_k^2 - \frac{W_k}{n'} \right) \frac{p_k q_k}{n_k - 1} + \frac{W_k (p_k - p_{st})^2}{n'} \right\}$$

殆んどすべての場合にこれは次の様に簡単に出来る。

$$\psi(p_{st}) \doteq \sum_k \left\{ \frac{W_k^2 p_k q_k}{n_k - 1} + \frac{W_k (p_k - p_{st})^2}{n'} \right\}$$

時には $1/n'$ の項も亦除く事が出来る。

(例)

(例) 374 家族の単純無作為標本に於て 292 が白人の家族で 82 が白人でない家族で占められている。4 家族の中約 1 家族の副次標本が所有者に対する次の資料を与える。

	自己所有	借家	總計
白人	31	43	74
白人でない	4	14	18

標本が抽出された場所での借家家族の割合を推定し推定量の標準誤差を見出せ。もしもオーの層が白人家族から成立っているとするば、

$$w_1 = \frac{292}{374} = 0.78 \quad : \quad w_2 = \frac{82}{374} = 0.22$$

$$p_1 = \frac{43}{74} = 0.60 \quad : \quad p_2 = \frac{14}{18} = 0.78$$

$$p_{st} = w_1 p_1 + w_2 p_2 = 0.64$$

$$n' = 374, \quad n_1 = 74, \quad n_2 = 18$$

$\psi(p_{st})$ の初項だけが重要である事は直ちに分る。従つて

$$\psi(p_{st}) = \sum_k \frac{W_k^2 p_k q_k}{n_k - 1} = \frac{(0.78)^2 (0.60)(0.40)}{73} \\ + \frac{(0.22)^2 (0.78)(0.22)}{17} = 0.00248$$

$$\delta(p_{st}) = 0.049$$

借家家族の推定割合は 0.64 ± 0.049 である。読者は大きさ 92 なる一段単純無作為標本よりもわずかな正確度における利益がある事は証明出来るであろう。白人でない層が相対的に大きさが小さいという見地から白人と白人でないものに対する借家家族の割合間の差が大きくなれば二重抽出と利益あらしめる効果が大きくなるであろう。

§ 12.5 回帰推定

Regression estimates :-

二重抽出の多くの応用に於て、任意変量 x_i は y の回帰推定量

として用いられてきた。我々は母集団が無限で y_i と x_i との関係が直線であると仮定しよう。一つのモデルとして次の様にかく。

$$y_{id} = \bar{y} + B(x_i - \bar{x}) + e_{id} \quad (12.17)$$

比巡に二番目の添字 d は一定 x_i に対して無作為変量 e_{id} が平均値の分散 $S_e^2 = S_y^2(1-\rho^2)$ なる度数分布に従うという注意として導入した。

第一の標本(大標本)に於て、その大きさは n で x_i のみを測定する。二番目の標本は大きさは n で x_i 及び y_{id} 共に測定する。

$\bar{y}$ の推定量は

$$\bar{y}_{er} = \bar{y} + b(\bar{x}' - \bar{x})$$

比巡に $\bar{x}'$, $\bar{x}$ は夫々オ一及びオニの標本に於ける x_i の平均値であり、 b はオニの標本から計算された x_i への y_{id} の最小自乗回帰係数である。

さて推定量 $(\bar{y}_{er} - \bar{y})$ の誤差を検定しよう。(12.17から)

$$\bar{y} = \bar{y} + B(\bar{x} - \bar{x}) + \bar{e} \quad (12.18)$$

$$b = \frac{\sum_{i=1}^n (y_{id} - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= B + \frac{\sum_{i=1}^n e_{id}(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (12.19)$$

(12.18) 及び (12.19) から推定量の誤差に対して、 $\bar{y}$ 及び b に対して代入して

$$\begin{aligned} \bar{y}_{er} - \bar{y} &= (\bar{y} - \bar{y}) + b(\bar{x}' - \bar{x}) \\ &= B(\bar{x} - \bar{x}) + \bar{e} + B(\bar{x}' - \bar{x}) + (\bar{x}' - \bar{x}) \frac{\sum e_{id}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \\ &= \bar{e} + (\bar{x}' - \bar{x}) \frac{\sum e_{id}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} + B(\bar{x}' - \bar{x}) \quad (12.20) \end{aligned}$$

通常の回帰理論に於て $\bar{x}' = \bar{x}$ では標準の特性は x_i の値が一定である。くりかえしの抽出に於て推定量 $(\bar{y}_{er} - \bar{y})$ の誤差の条件付度数分布を論ずる事である。もしもこの糸口が現在の問題に採用されるとすれば、オ一及びオニの標本共に x_i の値を一定に

して、我々は

$$E(\bar{y}_{er} - \bar{y}) = B(\bar{x}' - \bar{x})$$

であるからこの推定量は条件付分布で偏倚しているのが分る。もしも偏倚があまり大きくなければ、それは $\bar{y}_{er}$ の分散にその平方を加える事によって考慮に入れられる事が分った(§1.5) 従って我々は

$$V_0(\bar{y}_{er}) = S_y^2(1-\rho^2) \left(\frac{1}{n} + \frac{(\bar{x}' - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) + B^2(\bar{x}' - \bar{x})^2 \quad (12.21)$$

として $\bar{y}_{er}$ の条件付分散 V_0 を考える。

この表示は分散は二つの標本にあらわれる x_i 組によるから、抽出の他の方法との比較に対して適当でない。その代りに我々はオ一及びオニのすべての可能な抽出上での平均分散を必要とする。平均分散に対する簡単な結果は (i) オ一の標本は無作為に抽出される、(ii) オニの標本はオ一の標本からの無作為標本である、(iii) x_i は正規に分布するという仮定の下に得られる。この場合に、平均分散は次のものである事が分る。 $B^2 S_x^2 = \rho^2 S_y^2$ であるから

$$V(\bar{y}_{er}) = S_y^2(1-\rho^2) \left(\frac{1}{n} + \left(\frac{1}{n} - \frac{1}{n'} \right) \frac{1}{n-3} \right) + \frac{B^2 S_x^2}{n} \quad (12.22)$$

$$= \frac{S_y^2(1-\rho^2)}{n} \left(1 + \frac{(n'-n)}{n'} \frac{1}{(n-3)} \right) + \frac{\rho^2 S_y^2}{n} \quad (12.23)$$

もしも x_i が正規に分布していなければその値が変化する唯一の項は $\frac{1}{n} - \frac{1}{n'}$ で論じた様に $\frac{1}{n(n-3)}$ に於ける項である。仮定 ii を考えると、小標本が大標本から無作為に抽出されない； x_i の値に於て広い範囲を得る様に小標本を選び従って b の抽出誤差を減した方がよい。その効果は是らかなり $\frac{1}{n(n-3)}$ の項を減少する事である。

ある應用に於てはオニの標本はオ一の標本とは独立に抽出される。この場合に、この節に於て与えられた論議は方程式 (12.21) は変らないままに残る。方程式 (12.22) に於て項

$$\left(\frac{1}{n} - \frac{1}{n'} \right) \quad \text{は} \quad \left(\frac{1}{n} + \frac{1}{n'} \right) \quad \text{によつておきかえられる。}$$

この独立標本のこの場合はまた Chameli Bore (1943) によって考えられた。要約すると平均分散に於て $(n-1)$ に於ける項の正確な値についてある疑がある。併し $1/n$ が無視出来るとすればこの項にも無視出来る。これは次の定理と与える。

定理 12.4

もしも α の標本が大きさ n であり β の標本が n' であり、 $1/n$ が無視出来るは、二重抽出に於ける回帰推定量 $\bar{y}_{er}$ の分散は次により近似的に与えられる。

$$V(\bar{y}_{er}) = \frac{S_y^2(1-p^2)}{n} + \frac{p^2 S_y^2}{n'} \quad (12.24)$$

§ 12.6 回帰での二重抽出対単一抽出

Double sampling with regression versus single sampling:-

分散公式 (12.24) から、回帰推定での二重抽出は単一の単純無作為標本と (i) α の標本は単純無作為標本である。(ii) $1/n$ が無視出来る。(iii) β の標本は本単純無作為標本であるといふ仮定の下に比較出来る。この場合に於ける結果は他の場合へのラフな手引と与える。次の様に書く

$$V(\bar{y}_{er}) = \frac{V_n}{n} + \frac{V_{n'}}{n'}$$

此処に $V_n = S_y^2(1-p^2)$: $V_{n'} = p^2 S_y^2$

$$\text{費用} = c = n c_n + n' c_{n'}$$

最適 n 及び n' 及び最小分散を見出す問題は正確に層化に対する二重抽出に於けるものと同じである。(§ 12.3) 方程式 (12.10) は次のものと与える。

$$\begin{aligned} V_{opt} &= \frac{(\sqrt{V_n c_n} + \sqrt{V_{n'} c_{n'}})^2}{c} \\ &= \frac{S_y^2 (\sqrt{(1-p^2)c_n} + p\sqrt{c_{n'}})^2}{c} \end{aligned} \quad (12.25)$$

此処に、 p は正の値とる

もしもすべての資源が回帰に対する補正なしに単一標本に捧げら

れるとすれば、この標本は大きさ $n_s = c/c_n$ をもち、そしてその平均値の分散は

$$V(\bar{y}) = \frac{S_y^2}{n_s} = \frac{c_n S_y^2}{c} \quad (12.26)$$

従って二重抽出はもしも

$$c_n > (\sqrt{(1-p^2)c_n} + p\sqrt{c_{n'}})^2$$

であれば小さい分散と与える。

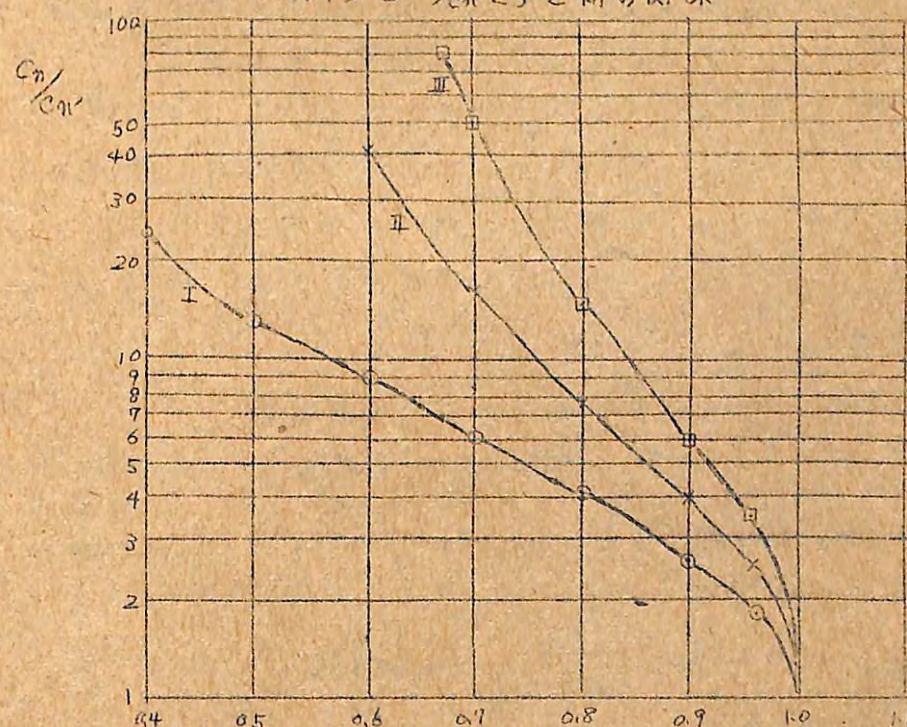
この不等式は次の二つの方法で表わされる

$$\frac{c_n}{c_{n'}} > \frac{(1 + \sqrt{1-p^2})^2}{p^2} = \frac{p^2}{(1 - \sqrt{1-p^2})^2} \quad (12.27)$$

又は

$$p^2 > \frac{4 c_n c_{n'}}{(c_n + c_{n'})^2} \quad (12.28)$$

12.1 図 二重抽出と単一抽出との相対的正確度の三つの一定値に対する $c_n/c_{n'}$ と p との関係



$p = 0.5$ と 0.7 との相関係数

曲線 I: 二重抽出と単一抽出の正確度は等しい, 曲線 II: 二重抽出の正確度は 25% 大
曲線 III: 二重抽出の正確度は 50% 大

方程式(12.27)は与えられた p の値に対して、第一の標本に於ける単位当り費用に対する σ^2 の標本に於ける単位当り費用の比は二重抽出が正確度を増加する前にある限界値を越えねばならないという争を示す。 C_n と C_n' を与えると方程式(12.28)は二重抽出が利益あるために p^2 が越えねばならぬ限界値を示す。

12.1 図は p の値に対して比 C_n/C_n' (対数尺度で)の値がプロットしてある。

曲線Iは二重抽出と単一抽出の正確度が等しい時の関係を示す。曲線IIは $V_{opt} = 0.8 V(\bar{y})$ 即ち二重抽出が25%正確度が大である争に成立する。曲線IIIは50%大なる正確度の場合を示す。例えば $p=0.8$ であれば二重抽出はもしも $C_n/C_n' = 4$ であれば単一抽出と正確度が等しく C_n/C_n' が約1/2であれば25%正確度大で C_n/C_n' が約1/3であれば50%正確度大である。実際の使用に対しては曲線は二重抽出から得るべき利益の過大推定量である。何故ならば、 n 及び n' の最良の値は以前の資料又は推測から推定されるからである。この推定量に於ける誤差をいくらか許容する事は二重抽出を採用する前に行われなければならない。

如何なる p に対しても、二重抽出からの正確度に対する利益に対して上限がある。これは $\bar{x}$ についての情報が自由に得られる($C_n' = 0$)時に起る。相対的正確度に対する上限は $1/(1-p^2)$ である。

§ 12.2 回帰に対する二重抽出に於ける推定分散

Estimated variance in double sampling for regression.

もしも $1/n$ 項が無視出来れば $V(\bar{y}_{er})$ は方程式(12.24)によって与えられる。

$$V(\bar{y}_{er}) = \frac{S_y^2(1-p^2)}{n} + \frac{p^2 S_y^2}{n'}$$

§ 12.3 によって量

$$S_{yx}^2 = \frac{1}{(n-2)} \left\{ \sum_{i=1}^n (y_i - \bar{y})^2 - b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right\}$$

は $S_y^2(1-p^2)$ の不偏推定量である。此処に添字は今は除いて置く。

$$S_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

は S_y^2 の不偏推定量であるから $(S_y^2 - S_{yx}^2)$

は $p^2 S_y^2$ の不偏推定量である

従って $V(\bar{y}_{er})$ の不偏推定量は

$$v(\bar{y}_{er}) = \frac{S_{yx}^2}{n} + \frac{S_y^2 - S_{yx}^2}{n'} \quad (12.29)$$

もしも σ^2 の標本が非常に小さく $1/n$ の項が無視出来れば、分散の暗示推定量は

$$v(\bar{y}_{er}) = S_{yx}^2 \left\{ \frac{1}{n} + \frac{(\bar{x}' - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\} + \frac{(S_y^2 - S_{yx}^2)}{n'}$$

これは条件付分散及び平均分散の一種の混成物である。

§ 12.3 比推定

Ratio estimates: -

もし σ の標本が τ の比推定に対して $\bar{x}$ を得るために用いられるとすれば、推定量は

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}} \bar{x}' \quad (12.30)$$

近似分散を見出すために、次の様にかく

$$\begin{aligned} \bar{y}_R - \bar{y} &= \frac{\bar{y}}{\bar{x}} \bar{x}' - \bar{y} \\ &= \left(\frac{\bar{y}}{\bar{x}} \bar{x} - \bar{y} \right) + \frac{\bar{y}}{\bar{x}} (\bar{x}' - \bar{x}) \\ &= \frac{\bar{x}}{\bar{x}} (\bar{y} - R \bar{x}) + \frac{\bar{y}}{\bar{x}} (\bar{x}' - \bar{x}) \end{aligned}$$

σ の成分は通常の比推定の誤差である(§ 6.3)。§ 6.3 に於て近似誤差分散を得るに際して我々は因子 $\bar{x}$ をこの項に於て1によって置きかえる。近似の同じオーダーに対して、我々は σ の成分に於ける因子 $\bar{y}$ を母集団比 $R = \bar{y}/\bar{x}$ によって置きかえる。

従って

$$Y_R - Y = (\bar{y} - R\bar{x}) + R(\bar{x}' - \bar{x}) \quad (12.31)$$

もしも $\bar{y}$ 及び $\bar{x}$ の標本が独立に抽出されたとすれば

$$V(\bar{y}_R) = \frac{S_y^2}{n} - 2R \frac{S_{yx}}{n} + \frac{R^2 S_x^2}{n} + \frac{R^2 S_{x'}^2}{n'} \quad (12.32)$$

此処に $f \cdot p \cdot c$ の項は無視出来できると仮定する。

もしも $\bar{x}$ の標本が $\bar{y}$ の標本の無作為標本とすれば、方程式(12.31)を次の形に再び並べる。

$$\bar{y}_R - \bar{y} = (\bar{y} - R\bar{x}) + R(\bar{x}' - \bar{x}) = (\bar{y} - \bar{y}) + R(\bar{x} - \bar{x})$$

$f \cdot p \cdot c$ を無視すると次の様になる事は容易に証明される

$$V(\bar{y} - \bar{y}) = \frac{S_y^2}{n}$$

$$\text{Cov}\{(\bar{y} - \bar{y}) R(\bar{x}' - \bar{x})\} = -R S_{yx} \left(\frac{1}{n} - \frac{1}{n'} \right)$$

$$V\{R(\bar{x}' - \bar{x})\} = R^2 S_{x'}^2 \left(\frac{1}{n} - \frac{1}{n'} \right)$$

従って $V(\bar{y}_R)$ は次の形となる。

$$V(\bar{y}_R) = \frac{S_y^2}{n} - 2R \frac{S_{yx}}{n} + \frac{R^2 S_x^2}{n} + \frac{2R S_{yx}}{n} - \frac{R^2 S_{x'}^2}{n'} \quad (12.33)$$

公式(12.32)及び(12.33)は共に次の形となる事に注意され

$$V(\bar{y}_R) = \frac{V_{11}}{n} + \frac{V_{kk'}}{n'}$$

従って n 及び n' の最適選択及び単一抽出との比較に対する最小分散は層化及び回帰推定と同じ方法によって見出される。詳細は此処では述べてない。

分散の標本推定量に対しては、量 S_y^2 , S_{yx} , S_x^2 及び $\hat{R}$ が (12.32) 及び (12.33) に代入される。その結果の推定量 $v(\bar{y}_R)$ は不偏ではなくて分散に示された近似のオーダーで適当である。

§ 12.9 同じ母集団のくりかえし抽出

Repeated sampling of the same population :-

標本抽出の信頼度が増すにつれて、規則正しい間隔で印刷されている一連の重要な資料を集めるために標本に頼って実行する事がごく普通の事となってくる。幾分か、これは動的な母集団で

の稀な間隔での調査が限られた使用をもっているという現実による。1945年7月と1950年7月に於ける母集団特性についての非常に正確な情報は、1952年に於ける母集団の知識を要求する計画には大して助けにならない。早とか又はより短かい間隔での相対的に小さい間隔のシリーズがより役立つものとなる。

同じ母集団が(時間の経過がもたらす変化とは別に)くりかえし抽出された時、我々は費用及び分散の現実的な推定量を作り、抽出の最適有効性を導く方法を適用する理想的な立場にある。一つの重要な問題はいかにしばしば又どの様に標本が時がたつにつれて変化せねばならぬかという事である。多くの考え方がその決定に影響する。人々は時が経つて同じ型の情報を与える事については心よしとしなり。応答は訪問で受け入れる情報によつて影響される。そしてこれは急速に時がたつにつれて代表性を失ってくる。併し時には彼等は一番始めの訪問のときよりも二番目の訪問のときの方がよりよくなり情報が技術的なものか信頼性のあるものである時は、二番目の訪問が一番目よりは正確な資料をもたらす。

この章の終りで我々は標本のおきかえの問題と一連のくりかえしの標本から推定量を作るそれに関連した問題を考えよう。二重抽出の技術が利用されるから、この話題をとりあげるのである。一連の標本から資料が与えられると我々が推定したい、三種類の量がある。

- i 一つの場合から他の場合への $\bar{y}$ の変化
- ii すべての場合の $\bar{y}$ の平均値
- iii 最も最近の場合の $\bar{y}$ の平均値

大抵の調査では、興味は、特にもしも母集団特性が時と共に急速に変化する様であれば、最後の平均値(iii)に集中される。一方時の変化が遅い様な母集団では十二ヶ月標本又は6ヶ月標本でとられた、年平均(ii)が大抵の使用に対しては適当である。これは長期間の慢性病の流行の研究に於てとられる立場であろう。その慢性病が著しい四季の変動を示す様な病気では最近の資料が主として興味を呼ぶ

が、年平均は異なる地方及び異なる年の比較に有用である。変化の推定(i)は母集団に於てはたゞくという事が分った力の影響を研究するのに主として望まれる。例えば家の建造をすゝめたい予算が通過したとすれば新しい家の建つ割合が年々増加したかどうかを知る事に關心がある。(増加は全く予算にはすぎないという理想を以て)

我々は標本の構成を変えたりそのまゝとしたりする事に自由であるとして又標本の全部の大きさはすべての場合に同じであると仮定しよう。もしも我々は正確度を最大にしたければ、次の論述を方策をたてるために行ふ事が出来る。

- i. 変化を推定するためすべての場合を通じて同じ標本を保持する事が最良である。
- ii. すべての場合に於ける平均を推定するためには各場合新しい標本を抽出する事が最良である。
- iii. 最近の推定量には、あらゆる場合に同じ標本をとるか又は標本を変えるかによって等しい正確度が得られる。各場合標本の一部をおきかえる事は、これ以外のものよりも良好である。

論述 i 及び ii は二つの連続事象で同じ単位の測定間に正の相関があるために成立する。一つの単位での、推定された変化の分散は $(S_1^2 + S_2^2 - 2\rho S_1 S_2)$ である。此処に添字はその場合に属するものである。もしも変化が二つの違った単位から推定されるとすれば、分散は $(S_1^2 + S_2^2)$ である。二つの場合に対してすべての平均を推定する際には分散は、もしも同じ単位分とられたとすれば $(S_1^2 + S_2^2 - 2\rho S_1 S_2)/2$ であり新しい単位が選ばれたとすれば $(S_1^2 + S_2^2)/2$ である。

論述 iii は明確ではないが、次の節で研究される。

§ 12.10 二つの事象での標本抽出

Sampling on two occasions

両方の事象で標本が同じ大きさ n であり、最近の推定量にまず

* 原本では n になっている。

関心をもっていると仮定しよう。おきかえの方策は *Jessen* により検定された(1942)。簡単なため、我々は、単純無作為抽出が用いられ、 y_i の母分布 S^2 は二つの事象において共に同じであると仮定しよう。

オ一の標本の平均値は以前の情報を利用されなから、 $3/2n$ なる分散をもつ、オ二の標本を選ぶ際にオ一の標本の単位の中 n 個が又とられた、(即ち n 個は同じである)そして残りの n 個の単位 n は用いてない)がすべて新しく選択でおきかえられる。

記 事

$$\begin{aligned}\bar{y}_{Au} &= \text{事象 } u \text{ に於ける同じでない部分の平均} \\ \bar{y}_{Am} &= \text{同じ部分の平均} \\ \bar{y}_n &= \text{全標本平均}\end{aligned}$$

オ二の標本の同じでない部分及び同じ部分は 12.1 表に示す様に、 $\bar{y}_2$ の独立な推定量 $\bar{y}'_{2u}, \bar{y}'_{2m}$ を与える。

12.1 表 同じでない部分及び同じ部分からの推定量

	推定量	分散
同じでない部分	$\bar{y}'_{2u} = \bar{y}_{2u}$	$S^2_u = \frac{1}{W_{2u}}$
同じ部分	$\bar{y}'_{2m} = \bar{y}_{2m} + b(\bar{y}_u - \bar{y}_m)$	$\frac{S^2(1-\rho^2)}{n} + \rho^2 \frac{S^2}{2} = \frac{1}{W_{2m}}$

同じ部分に於ては二重抽出回帰推定量を用いる。此処に大標本がオ一の標本で、任意変量、 x_i はオ一の事象での y_i の値である。 $\bar{y}'_{2m}$ の分散は 12.24 (12.24) から得られる。この場合 m 及び n は、夫々公式 (12.24) の n 及び n' に対応する事に注意。

$\bar{y}_2$ の最良組合せ推定量はそれ等の分散と逆に重みを二つ独立変量につけて得られる。もしも W_{2u}, W_{2m} が逆分散であれば、この推定量は

$$\bar{y}'_2 = \rho_2 \bar{y}'_{2u} - (1 - \rho_2) \bar{y}'_{2m} \quad (12.24)$$

此処に $\rho_2 = \frac{W_{2u}}{W_{2u} + W_{2m}}$

最小自乗理論により、 $\bar{y}_2'$ の分散は

$$V(\bar{y}_2') = \frac{1}{W_{2u} + W_{2m}}$$

12.1 表から簡単にすれば、これは次の如くなる。

$$V(\bar{y}_2') = \frac{S^2(n - \mu p^2)}{(n^2 - \mu^2 p^2)} \quad (12.35)$$

もしも $\mu = 0$ (完全に同じ) 又は $\mu = n$ (全く一致しない) であるればこの分散は同じ $\frac{S^2}{n}$ になる事に注意。 μ の最適の値は μ の変化に關して (12.35) を最小にする事によって得られる。即ち

$$\frac{\mu}{n} = \frac{1}{1 + \sqrt{1 - p^2}} \quad ; \quad \frac{m}{n} = \frac{\sqrt{1 - p^2}}{1 + \sqrt{1 - p^2}} \quad (12.36)$$

最適 μ を (12.35) に代入すると最小分散は次の如くして求まる。

$$V_{opt}(\bar{y}_2') = \frac{S^2}{2n} \left\{ 1 + \sqrt{1 - p^2} \right\} \quad (12.37)$$

12.2 表は p の値のシリーズに対して、同じでなければならぬ最適割合及び同じでないものと比較したときの相対的利益が示してある。同じであるべき最良百分率は 50% 以上には決してならないし、 p が増すと共に減少する。 $p = 1$ のとき、公式は $m = 0$ を示し、 m が合理的に大と仮定したから我々の仮定の範囲外にある。この場合の正しい方法は $m = 2$ をとる事である。

正確度には於て到達し得る最大の利益は $p = 1$ のときの 100% である。 p が大でなければ利益は目立たない。

たとえ同じであるべき最適百分率が p と共に変化しても、たゞ一つの百分率が調査に於けるすべての項目に対して実際に用いる事が出来る。12.2 表の右欄は '各' 及び '各' 単位が同じである場合の百分率の利益を示している、共に p が 0.95 以上である項目以外はよくあっている。

12.2 表 同じである最適百分率

p	同じである最適百分率 %	正確度における利益 %	利 益 (%)	
			$\frac{m}{n} = \frac{1}{3}$	$\frac{m}{n} = \frac{1}{2}$
0.5	46	7	7	6
0.6	44	11	11	9
0.7	42	17	17	15
0.8	38	25	25	23
0.9	30	39	39	39
0.95	24	52	50	52
1.00	0	100	67	75

§12.11 ニつ以上の継時事象の標本抽出

Sampling on more than two occasions :-

おきかえの一般問題は Yates (1949) 及び Patterson (1950) により最近の推定量及び変化の推定量に關して研究された。二つ以上の継時事象がある時には、資料を自由に用いる機会が増してくる。事象 k に於て我々は事象 $(k-1)$ と同じ標本をもち、 $(k-1)$ 及び $(k-2)$ の両方で同じである部分をもつ等々とする。最近の推定量と改善する際に我々は以前の場合にすべて同じである。様々な多項回帰を試みる。亦事象 k に対する資料が分つた後に事象 $(k-1)$ に於ける最近の推定量を修正する事も可能である。修正推定量に於て事象 $(k-2)$ 及び事象 $(k-1)$ の回帰は標本の適宜に同じである部分が利用出来ると仮定すると用い得るであろう。

この節ではその問題を招かす事にする。注意は直ちに前の標本での回帰だけが利用出来る最近の推定量に限定する。これは正確度を幾分損ずるが相関 p が事象間の間隔が増すにつれて減少するから正確度の損失はそう大でもない。分散 S^2 及び二つの継時事象に於ける同じ単位上での項目数値間の相関係数 p はすべてを通じて一定と仮定する。

三番目の事象に於て m 及び u は夫々同じである単位の数及び同じでない単位の数としよう。作る得る $\bar{y}$ の二つ推定量は 12.3 表・あたえてある。才二の事象からの方法に於ける変化のみ (12.1 表) が同じ部分からの推定量の自帰調整に於て標本平均 $\bar{y}_2$ のかわりに改善された推定量 $\bar{y}_{2m}$ を用いるといふ事である

12.3 表	三番目の事象での $\bar{y}$ の推定量	推定量	分散
同じでない部分	$\bar{y}_{2u} = \bar{y}_{2u}$	$\frac{S_u^2}{u}$	$= \frac{1}{W_{2u}}$
同じ部分	$\bar{y}_{2m} = \bar{y}_{2m} + b(\bar{y}_2 - \bar{y}_{2m})$	$\frac{S^2(1-p^2)}{m} + p^2 V(\bar{y}_2)$	$= \frac{1}{W_{2m}}$

12.3 表の同じ部分の推定量 $\bar{y}_{2m}$ の分散は記号を少し変えて §12.5 (12.22) 式から導かれる。方程式 (12.22) は $1/n^2$ の項を除いて次の様になる。

$$V(\bar{y}_{2m}) = V(\bar{y}_{2m}) = \frac{S^2(1-p^2)}{m} + \frac{B^2 S^2}{m}$$

この表示に必要な変換は次の如くである。

m = 小標本の大きさ = m

$S_y^2 = S^2$

$B = p$ 何故なら二つの事象に於て S^2 は等しいと仮定されるから。

$S_{\bar{y}_2}^2$ = 大標本からの推定平均値の分散 = $V(\bar{y}_2)$

これ等の立場がとられるとき 12.3 表に示した公式が得られる。

三番目の事象でもしも添字 u を u とおきかえ添字 m を $(n-1)$ とおきかえれば二つの推定量は 12.3 表と同じになる。

さて各事象での m 及び u の最適の値を見出そう。最適 m は継時で事象では一定に増加し、むしろ名の極限値に急速に近づく事を示そう

前の様に、分散の逆数を重みとして、 $\bar{y}_h$ の最良推定量

$$\bar{y}_h = p_h \bar{y}_{hu} + (1-p_h) \bar{y}_{hm}$$

此処に

$$p_h = \frac{W_{hu}}{W_{hu} + W_{hm}}$$

$$V(\bar{y}_h) = \frac{1}{W_{hu} + W_{hm}} \quad (12.39)$$

であるから次のものを最小にする事によって最適 m を見出す事が出来る。

$$(W_{hu} + W_{hm})$$

次の様にかくと助けとなる

$$V(\bar{y}_h) = g_h \frac{S^2}{n} \quad (12.40)$$

$V(\bar{y}_h) = S^2/n$ であるから量 g_h は事象 h と才一の事象との分散比である。もしも継時推定量が一樣に正確となれば g_h は h の減少函数となる。

さて 12.3 表から u の代りに n とおいて

$$\frac{1}{W_{hu}} = \frac{S^2}{u} ; \quad \frac{1}{W_{hm}} = \frac{S^2(1-p^2)}{m} + \frac{S^2 p g_{h-1}}{n}$$

これは次のものを与える

$$W_{hu} + W_{hm} = \frac{1}{S^2} \left(u + \frac{1-p^2}{m} + \frac{p^2 g_{h-1}}{n} \right) \quad (12.41)$$

従って

$$S^2(W_{hu} + W_{hm}) = (n-m) + \frac{1-p^2}{m} + \frac{p^2 g_{h-1}}{n}$$

m について微分して最適 m_h は次の方程式を満足する様に見られる

$$\frac{\sqrt{1-p^2}}{m_h} = \frac{1-p^2}{m_h^2} + \frac{p^2 g_{h-1}}{n}$$

$$\text{即ち } \frac{m_h}{n} = \frac{\sqrt{1-p^2}}{g_{h-1}(1+\sqrt{1-p^2})} \quad (12.42)$$

この方程式は (12.36) を一般化したものである。(12.36) は二番目の事象での最適 m を与えるのであるが。方程式 (12.42) は同じあるべき最適百分率は時と共に一定して増加する事を暗示している。何故ならば我々は $V(\bar{y}_{h-1})$ 従って g_{h-1} は時と共に減少すると期待するのであるからである

解を完全に求めるために g_h の値を見出す事が必要である。定義により (12.40) は (12.41) より

$$\frac{1}{g_h} = \frac{S^2}{n V(\bar{y}_h)} = \frac{S^2}{n} (W_{hu} + W_{hm}) = \frac{1}{n} \left(u + \frac{1-p^2}{m_h} + \frac{p^2 g_{h-1}}{n} \right)$$

方程式 (12.42) からの最適 m 及び u をこの方程式に代入して

g_n と g_{n-1} との間の再帰関係を得る代数的操作によつてその関係は次の如くなる

$$\frac{1}{g_n} = 1 + \frac{(1 - \sqrt{1 - p^2})}{g_{n-1}(1 + \sqrt{1 - p^2})} \quad (12.43)$$

$g_1 = 1$ であるから g_n の次々の値従つて最小分散及び最適 m が p の与えられた値によつて計算出来る。期待される様に g_n の次々の値は一定に減少し、 m_n のそれは一定に増加する事が分る。 m が増すと量 g_n が極限になる事を示す事は容易である。その値は (12.43) で $g_n = g_{n-1} = \infty$ とおいて g_{∞} について解けば分る。これは

$$g_{\infty} = \frac{2\sqrt{1-p^2}}{1+\sqrt{1-p^2}}$$

従つて g_n の分散は次の如くなる

$$\sqrt{(g_{\infty})} = \frac{S^2}{n} \left(\frac{2\sqrt{1-p^2}}{1+\sqrt{1-p^2}} \right) \quad (12.44)$$

最後に m_n の限界値は p の値に拘わらず (12.42) から次の如く得られる

$$\frac{m_{\infty}}{n} = \frac{\sqrt{1-p^2}}{g_{\infty}(1+\sqrt{1-p^2})} = \frac{1}{2}$$

12.4 表は同じであるべき最適百分率を示す — 方程式 (12.42)

12.4 表 同じである場合の最適百分率及正確度に対する利益

n	同じである場合 %				正確度に対する利益 %			
	p =				p =			
	0.7	0.8	0.9	0.95	0.7	0.8	0.9	0.95
2	42	38	30	24	17	25	39	52
3	48	47	30	26	19	31	55	80
4	50	48	47	43	20	33	61	94
5	50	50	49	46	20	33	63	102
6	50	50	50	48	20	33	64	106
∞	50	50	50	50	20	33	65	110

から見出しを $100 m/n$ — 及び $p = 0.7, 0.8, 0.9$ 及び 0.95 と n の値のシリーズに対して方程式 (12.42) から計算した正確度の相対的利益を示す。

これらの結果はくりかえしの抽出では、良好な実行規則は π の事象では π の標本の多か少の割合をなもつ事を暗示している。更に標本の $1/2$ が各事象で保持されねばならぬし、新しく n 抽出しなければならぬ。

これらの推奨はすべてのおきかえ方策は全標本の大きさと一定にしたまゝで同じ費用が同じ自由性を仮定している。実際には費用及び自由性の問題はみすゞされてはならない。新しい標本を抽出したり調したりする余分の費用がかかるから費用の考えは次につづく記事でおそまきなが指摘されるであろう。更に 12.4 表は p が 0.7 以下であれば完全に一致した場合に於ける正確度の利益の n は大したものでなく最適の同じである場合からかけはなれると正確度に対する n どの損失についての一般的な結論はないであろう事を明らかに示している。

第 13 章 調査に於ける誤差の原因

Sources of errors in surveys§ 13.1 緒言 *Introduction*

前章で示した理論はすべてを通じてある種類の確率抽出が用いられ、 i 番目の単位での観測値 x_i はその単位に対して正しい値であると仮定している。推定量の誤差は n 個の単位が N 個の単位の完全母集団の代りに測定される時に示される無作為抽出変動から単独に生ずる。

これ等の仮定は測定の手が正確で仕事の質が高潔である調査のより、簡単な型に於てよく合理的に成立する複雑な調査に於ては時に測定の困難な問題が含まれる場合にはその仮定は真とはかけはなれて、三つの附加的な誤差の原因は次の通りである。

- i. 送ばれた標本に於ける単位のあるものと測定するのに失敗する事。これは見落としをするか又は人間母集団の場合に起る、といふのは訪問した時不在である場合とか、在宅していても質問に答えるのを拒否したりする事があるからである。
- ii. 単位上の測定誤差。測定方法が偏倚しているか又は不正確である場合である。人間母集団では応答者は正確な情報をもっていないか又は偏倚した答をする。
- iii. 結果を編集したり表示したりする場合に生ずる誤差。

これ等の誤差の原因は標準の抽出理論の変形を必要とさせる。この様な変形の主目的は無作為抽出誤差の原因と他の誤差との間に関する誤差の割当てについての手引を与える事であり、他の誤差が存在する時にも妥当である。

標準誤差及び信頼限界の計算に対しての方法を展開する事である。最近迄、多くの抽出理論の仕事は無作為抽出誤差の減少に関するものであつた。そして理論に於ける必要な変形は現在尚不完全である。併し、多くの困難が残っているにしても、已に良好な端緒が開かれている。

§ 13.2 無回答の影響

Effect of non-response :-

我々は選ばれた標本に於ける単位のあるものを測定するに失敗するに因して無回答なる言葉を 用いる。無回答の研究に於ては母集団が二つの層に分けられると考えると便利である。即ち第一の層は測定がもしもまたその単位が標本に属しておれば得られるのであろう。すべての単位から成立っており、第二の層は測定が得られない単位から成立っている。二つの層の構成はまづ最初単位を見出し資料を得るのに用いる方法如何による。少なくとも三つの要求が必要ならば、各家で行われるとか、資料を与える事を拒否するすべての人に創外的とも思える説得力で当業者が要求する様な調査はたゞ一回しか各家で試みられたものよりは小さい無回答の層を形成する である。

勿論この二つの明らかな層に母集団を分割する事は簡易に過ぎるものである。与えられた試みの数に於て単位が見出され測定されるかどうかを決定する際に偶然性がその役割りを演ずる。

より特理化された問題に於ては、我々は各単位に対してもしもそれが標本に属するならば与えられた野外の方法によつて測定される機会を表わす確率を得るであろう。併し、二つの層への分割は此処に示される分析に対して適當である。

標本は無回答の層2に対しては情報を与えない。これはもしも層2の特性が層1のものと同じであると仮定出来れば問題はない。併しチェックが行われると、無回答の層に於ける単位は測定される単位と違うといふ事がしばしば判明する。その例を心ノ表に示す。資料は1946年 North Carolina に於ける果樹園の実験的抽出からとったものである。同じ質問の三つの次々の郵便が所有者に送られた。

質問の一つ即ち果樹の数に対しては完全な資料が母集団に対して利用された。(Finckel, 1950)

心ノ表 郵便調査に於ける三つの質問に対する回答

	所有者の数	母集団の%	所有者当り果樹の平均数
最初の郵便に対する答	300	10	456
二番目の郵便に対する答	543	17	382
三番目の郵便に対する答	434	14	340
三番目の郵便後の無回答	1839	59	290
合 計	3116	100	329

次々の回答に於ける所有者当り果樹本数に於て一定の減少がある事が明らかに分る。即ち第一回郵送ではこの数は456、二回目では382、三回目では340。三回やつて拒否されたものに対して290である。全回答は貧弱であつた即ち三回も試みてなお資料を得るのに失敗したのは母集団の半分以上であつた。さて我々は標本推定量に対する無回答の影響を考えよう。 N_1 及び N_2 をして二つの層に於ける単位の数とし、 $W_1 = N_1/N$ 、 $W_2 = N_2/N$ としよう。そのとき N_2 は母集団に於ける無回答の割合である。単純無作為標本が母集団から抽出されたと仮定しよう。実際の仕事は完全である時我々は層1から単純無作為標本に対してこの資料を得、層2からは資料を得ない。

従つて標本平均に於ける偏倚の量は

$$E(\bar{y}) - \bar{Y} = \bar{Y}_1 - \bar{Y} = \bar{Y}_1 - (W_1 \bar{Y}_1 + W_2 \bar{Y}_2) \\ = W_2 (\bar{Y}_1 - \bar{Y}_2) \quad (13.1)$$

偏倚の量は無回答の割合と二つの層の平均値間の差との積に等しい。標本は $\bar{Y}_2$ についての情報を与えないから偏倚の量はその範囲を標本資料より以外のある原因から $\bar{Y}_2$ にふく事が出来なければ未知である。連続変量では確實に割当てる事が出来る範囲のみがしばしば利用されない程広い事がある。従つて連続資料では無回答の如何なる大きさの割合でも標本の結果から $\bar{Y}$ に対して有用な信頼限界を割当てる事を不可能にする。我々は推測を実証する事なしに偏倚の量についてある推測にたよる立場が幾されている。

割合に対する標本抽出では層2に於ける未知の割合は0と1

との同になければならぬから、少しく容易になる。もしも W_2 が分つておれば、 P_2 に対する範囲は母集団割合 P に対する信頼限界を作る事を可能にする。 n 個の単位の単純標本が抽出され、測定が標本に於ける n 個の単位について得られたと仮定しよう。 n を充分大きいと仮定すると、 P に対する 95% 信頼限界は次により与えられる。

$$p \pm 2 \sqrt{\frac{p(1-p)}{n}}$$

此処に p は標本の割合で P を無理する。

我々が P についての信頼論述をしようとする時、 $\hat{p}_L$ を見出すときに $P_2 = 0$ 、 $\hat{p}_U$ を見出すとき $P_2 = 1$ と仮定するならば安全側である。従つて 95% 限界は次の如くなる。

$$\hat{p}_L = W_1(p - 2\sqrt{\frac{p(1-p)}{n}}) + W_2(0) \quad (132)$$

$$\hat{p}_U = W_1(p + 2\sqrt{\frac{p(1-p)}{n}}) + W_2(1) \quad (133)$$

これ等の限界は論述

$$\hat{p}_L \leq P \leq \hat{p}_U$$

に対して控え目であつたという事を証明する事は容易である。即ち

$$W_1(p - 2\sqrt{\frac{p(1-p)}{n}}) \leq p \leq W_1(p + 2\sqrt{\frac{p(1-p)}{n}}) + W_2$$

は論述

$$p - 2\sqrt{\frac{p(1-p)}{n}} - \frac{W_2}{W_1}(1-P_2) \leq p \leq p + 2\sqrt{\frac{p(1-p)}{n}} + \frac{W_2}{W_1}P_2 \quad (134)$$

と同じである。

P_2 の値がどうであろうと向隔 (134) は常に向隔

$$p - 2\sqrt{\frac{p(1-p)}{n}} \leq p \leq p + 2\sqrt{\frac{p(1-p)}{n}}$$

を含む。

$$\text{従つて } P\{\hat{p}_L \leq P \leq \hat{p}_U\} \geq 0.95$$

もしも母集団に於ける無応答の割合 W_2 が分つておればこの様にして限界を見出し得るけれども W_2 が非常に小さくなれば限界は非常に広くなる。13.2 表は標本の大きさ $n=1000$ 及び W_2 及

が p の値のシリーズに対しての平均限界を示す。方程式 (132) 及び (133) は n の値によるので (標本に於ける応答の数) 我々は 13.2 表の計算に於てその平均値 $n_e = nW_1$ をとつた。

W_2 が増すと信頼間隔の幅が急速に増す事は明らかである。もしも W_2 が 0 であれば信頼間隔の同じ幅を与えるにはどの位の n の値が必要であるかを検定する事は興味がある。

13.2 表 P に対する 95% 信頼限界 (%)

$n=1000$

新答の % $100W_2$	標本百分率 $100P$			
	5	10	20	50
0	(36, 64)	(81, 119)	(175, 225)	(467, 532)
5	(34, 111)	(76, 163)	(165, 265)	(444, 556)
10	(32, 158)	(72, 208)	(156, 304)	(420, 580)
15	(30, 205)	(68, 252)	(147, 343)	(396, 604)
20	(28, 252)	(63, 297)	(137, 383)	(372, 628)

これは p が 50% であるときに容易に行われる。 $W_2=5\%$ に対して 13.2 表は信頼間隔の半分幅が 5.6 であることを示している。無応答がないと仮定すると等量の標本の大きさ n_e は次の方程式から見出される。

$$5.6 = 2\sqrt{\frac{(50)(50)}{n_e}}$$

$$n_e = 320$$

$W_2=10\%$ 、 15% 及び 20% に対して n_e の値は夫々 155, 90 及び 60 である。明らかに無応答の減少には資源の膨大な割合を捧げなければならないのである。

13.2 表に於ける限界は最も悪い可能な場合が次第に起つたと仮定しているのであまりに控え目すぎるくらいがある。更に限界はあまり広すぎるから用いられないので常にその中に P_2 が存

在する範囲についてある推測をしようと試み、これ等の範囲に基いた P に対するより狭い信頼限界を作ろうと試みている。もしも範囲が正しいならばこの方法では誤っていないが、我々は方法は客観的な証拠に対しての推測の代用を表わしているという事を認識しなければならぬ。

無応答が存在するとき標本の大きさを見出す興味ある方法が *Birnbaum* 及び *Sirken* (1950) によつて与えられた。無応答の割合 w_2 は調査の特別な型に於て以前の経験から分つていたり仮定される。 P_1 , P_2 又は P の前以ての知識はないと仮定される。従つてもしも無応答があり我々が標本の割合に於ける絶対誤差を d 以下にしたければ (8.4.4 により) 次の様になる。

$$n = \frac{t_{\alpha}^2 P Q}{d^2}$$

此処に t_{α} は誤差が d 以上になる限界 d に応ずる正規偏差である。 P についての前以ての情報が無いと我々は少し都合のよい場合として $P=0.5$ をとつて

$$n = \frac{t_{\alpha}^2}{4d^2} \quad (13.5)$$

偏倚 $w_2 (P_1 - P_2)$ 及び P_1 の値の少しく都合のよい組合せをとる事によつて *Birnbaum* 及び *Sirken* は危険率 α で誤差が d 以下を保証する n の値は

$$n = \frac{t_{\alpha}^2}{4d(d-w_2)w_1} - 1 \quad (13.6)$$

という事を示している。

もしも $w_2 > d$ であれば n の値は満足でない事は注意されよ。もしも $w_2 = 0$ であれば、この方程式は分析に於ける近似から項 -1 を除いた (13.5) となる。*Birnbaum* 及び *Sirken* の方法によつて与えられた n の値を 13.3 表に示す。

13.3 表 危険率 α で誤差 d の与えられた限界に対しての最小の値

無応答の % $100w_2$	d (%)			
	20	15	10	5
0	24	43	96	384
2	27	50	122	483
4	31	60	166	2000
6	36	75	255	--
8	43	99	521	--
10	53	142	--	--
15	112	--	--	--

表は 13.2 表と同じ非観的な結果を示している。もしも我々がうな推定量 ($d=20$) で満足すれば 10% 迄の無応答の量は標本の大きさを倍にして取扱う事が出来る。併し無応答のかなりな大きさの割合は応答者内の標本の大きさを増す事によつて非常に明らかな正確度を得る事を不可能にするか又は非常に費用がかかる事になる。

§ 13.3 無応答者間最適抽出化

Optimum sampling fraction among the non-respondents:-

調査に於ける無応答者は二つのクラスに分けられる。その一つは一これはひどい核心だが一資料を得る事が不可能である単位から成立つ。人間母集団ではこのクラスは見出す事が出来ない人、資料を与えるのを頑強に拒否するか又はある理由で与える事が出来ない人及びその人に対して他の人によつて資料を与える事が出来ない人を含む。二番目のクラスはその資料に対しては調査に対して最初に企画したよりも一層ひどくて費用がかかる野外の方法によつて得る事が出来る人を含む。このクラスは最初調査したとき不在である人及び気のすまない人を含むが資料を得るためには

説得出来る人である。

Hansen 及び Hurwitz (1946) はこの節に於ける結果は彼等によるのであるが、このクラスは層ノの単位よりも小さい抽出比で抽出する事を主張している。

第一段階は通常の野外の方法を用いて n 個の単位の単純無作為標本をとる事である。 n をして求めている資料を与える標本に於ける数とし n_2 をして無回答グループに於ける数としよう。よりひどい努力によつて資料は後に n_2 から n_2 の単純無作為標本から得られる。

$$n_2 = k(r_2) \quad (k > 1) \quad (13.7)$$

としよう。

そのとき第一の層に於ける平均抽出比は二番目の層の抽出比の k 倍である。これはもし前以て k が一定であれば

$$E\left(\frac{n_1}{N_1}\right) = E\left(\frac{n_2}{N_2}\right) = k E\left(\frac{r_2}{N_2}\right)$$

であるからである。

n (最初の標本の大きさ) 及び k の値は最小費用で特別な正確度を与えるように送られる。

標本をとる費用は

$$C = C_0 n + C_1 n_1 + C_2 r_2$$

此処に C 等は単位当り費用である。即ち C_0 は最初の試みに要する費用で、 C_1 及び C_2 は夫々二つの層で資料を得て処理する費用である。

n 及び n_2 は最初の試みが行われるまで分らないから、標本の計画には期待費用が用いられる。 n 及び r_2 の期待費用は夫々 $W_1 n$ 及び $W_2 n/k$ である。従つて期待費用は

$$C_0 n + C_1 W_1 n + \frac{C_2 W_2 n}{k} \quad (13.8)$$

$\bar{y}$, $\bar{y}_{2r}$ をして二つの層に於ける標本平均としよう。添字 r , 第二期の層に於ける標本が大きさ n_2 からなるという事を注意するためにつけたものである。母平均の推定量として

$$\bar{y}' = \frac{1}{n} (n_1 \bar{y}_1 + n_2 \bar{y}_{2r}) \quad (13.9)$$

二番目の層は標本は n_2 の大きさのみからしかなつていないけれども荷重 n_2 をつける。大きさ n なる第一の即ち大標本は層の相対的大さの推定量 n_1/n_2 を与える。第二の即ち小標本は第一の標本では大きさ n_1 で第二の標本では大きさ n_2 である。悪い事に、 $\bar{y}$ の分散は層化した場合の二重抽出に対しての §12.2 に与えた分散公式から導く事は出来ない。§12.2 では二番目の標本の大きさ n_2 は一定と仮定するが現在の問題では n_1 及び n_2 は無作為変量である。

$V(\bar{y}')$ と見出すため次の様にかく

$$\bar{y}' = \frac{1}{n} (n_1 \bar{y}_1 + n_2 \bar{y}_{2n}) + \frac{n_2}{n} (\bar{y}_{2r} - \bar{y}_{2n}) \quad (13.10)$$

此処に $\bar{y}_{2n}$ は層 2 からの大きさ n_2 なる全標本の平均値である。

右辺の第一項は全母集団からの大きさ n なる無作為標本の平均値である。従つてその分散は

$$\frac{(N-n)S^2}{Nn}$$

此処に S^2 は全母集団の分散である。更に我々が $\bar{y}'$ の分散を見出すとき第一項と第二項間の積からの貢献はなくなる。何故ならば n_2 なる一定大きさの標本から抽出出来る大きさ r_2 なるすべての無作為標本上で

$$E\{\bar{y}_{2n}(\bar{y}_{2r} - \bar{y}_{2n})\} = 0$$

であるからである。

(13.10) の右辺の第二項を考えよう。もしも $\bar{y}_2$ が「無回答」層の母平均であれば

$$(\bar{y}_{2r} - \bar{y}_2) = (\bar{y}_{2r} - \bar{y}_{2n}) + (\bar{y}_{2n} - \bar{y}_2)$$

従つて前と同じ理由で積の項はなくなるから

$$E(\bar{y}_{2r} - \bar{y}_2)^2 = E(\bar{y}_{2r} - \bar{y}_{2n})^2 + E(\bar{y}_{2n} - \bar{y}_2)^2$$

さて $\bar{y}_{2r}$ は第二番目の層からの大きさ n_2 なる単純無作為標本の平均値であり、 $\bar{y}_{2n}$ は同じ層からの大きさ n_2 なる単純無作為

標本の平均値である。従つて一定 n_2 及び r_2 に対して

$$\frac{(N_2 - r_2)}{N_2} \frac{S_2^2}{r_2} = E(\bar{y}_{2r} - \bar{y}_{2n})^2 + \frac{(N_2 - n_2)}{N_2} \cdot \frac{S_2^2}{n_2}$$

此処に S_2^2 は「無応答」層内分散である。これは $n_2 = k_2$ であるから

$$E(\bar{y}_{2r} - \bar{y}_{2n})^2 = S_2^2 \left(\frac{1}{r_2} - \frac{1}{n_2} \right) = S_2^2 \frac{(n_2 - r_2)}{n_2 r_2} = S_2^2 \frac{(k_2 - 1)}{n_2}$$

従つて (13.10) に於ける二つの項の分散を加えて一定 n_2 に対して

$$\begin{aligned} V(\bar{y}') &= \frac{(N - n)}{N} \frac{S^2}{n} + \left(\frac{n_2}{n} \right)^2 \cdot \frac{(k_2 - 1)}{n_2} \cdot S_2^2 \\ &= \frac{(N - n)}{N} \frac{S^2}{n} + \frac{(k_2 - 1)}{n_2} \cdot n_2 S_2^2 \quad (13.11) \end{aligned}$$

$E(n_2) = nW_2$ であるからこれは期待分散に対して次のものを与える。

$$V(\bar{y}') = \frac{(N - n)}{N} \frac{S^2}{n} + \frac{(k_2 - 1)W_2}{n} S_2^2 \quad (13.12)$$

第一項はむしろ無応答グループに於けるすべての n_2 が抽出されるならば得られるであろう分散である。二番目の項は n_2 の中から k_2 のみの抽出からの分散に於ける増加である。

量 n 及び k は従つて期待分散 (13.12) の前以て割当てられた値に対して平均費用 (13.8) を最小にするため選ばれる。

解は

$$k_{opt} = \sqrt{\frac{C_2(S^2 - W_2 S_2^2)}{S_2^2(C_0 + C_1 W_1)}} \quad (13.13)$$

$$n_{opt} = \frac{N\{S^2 + (k_2 - 1)W_2 S_2^2\}}{N\bar{V} + S^2} \quad (13.14)$$

此処に $\bar{V}$ は推定母平均の分散に対して特殊化された値である。

解は W_2 の知識を必要とする。しばしばこれは以前の経験から推定出来る。標本の大きさの問題に於て前以て推定されねばならない S^2 に加える際に解は亦無応答層での分散 S_2^2 を含む。

S_2^2 は通常予測困難である。それは恐らく S^2 とは同じで は ない

い であろう。例えば大極の種類の新着企業の郵送によつて行われる調査に於ては応答者は無応答者よりもより大きい単位間分散をもつより大きい仲買人になる傾向がある。

この方法はまず手紙に答えなかつた人の副次標本に対しての訪問によつて行われた郵送により行われた調査に対して示された。郵送調査では W_2 は大きくその値は n_{opt} を決定するため予測する事は困難である。この場合には満足な近似はりと安全な上限との間の W_2 の仮定した値の範囲に対して n_{opt} の値を見出す事である。このシリーズに於ける最大 n_{opt} は最初の標本の大きさ n として採用される。郵送調査に対する回答をうけとつたとき n_2 の値が分かる。そのとき分散公式 (13.11) は欲する分散 V を決定する k の値を見出すため解かれる。この方法に対する費用は通常もしも W_2 が分つていれば適用されたとであろう最適費用よりもかなり高いものである。

(例)

この例は Hansen Hurwicz (1946) の論文から引用したものである。最初の標本は郵送により行われ、応答割合 W_2 は 50% であると期待される。欲する正確度はむしろ無応答がなければ大きき 1000 なる単純無作為標本によりあたえられるであろうものである。質問も郵送する費用は 10 セントで完全な訪問を処理する費用は 40 セントである。個人的な訪問を行うには 4.10 ドルかかる。どの位の質問が送られねばならぬか？ 又無応答のどの位の割合が訪問されねばならぬか？

費用函数 (13.8) の項に於てドルでの単位の費用は次の如くである。

C_0 = 最初の試みの費用 = 0.1

C_1 = 応答者に対して処理する費用 = 0.4

C_2 = 無応答者に対して資料を得て整理する費用 = 4.5

最適 n 及び k は (13.13) 及び (13.14) なる方程式から見出す事が出来る。もしも分散 S^2 及び S_2^2 が等しいと仮定し N が非常に

大きいと仮定すると

$$k_{opt} = \sqrt{\frac{C_0(1-\pi_2)}{C_0 + C_1\pi_1}} = \sqrt{\frac{(2.5)(0.5)}{0.1 + (0.4)(0.5)}} = \sqrt{2.5} = 2.239$$

$$n_{opt} = \frac{S^2 \{1 + (k-1)\pi_2\}}{V} = 100 \{1 + (2.239)(0.5)\} = 1870$$

我々はこれは標本平均がもしも1000の標本がとられて完全に答を得られたとすれば得られるであろう値であるから $S^2/V = 1000$ 即ち $V = S^2/1000$ とおいた事に注意する。

従つて1870の質問が郵送されねばならぬ。もとつてこなかつた935の中我々は935/2239 即ち34%の無作為副次標本を訪問する。費用は2095ドルである事が分るのである。

層化抽出では個々の層に於ける n_k 及び k_k の最適の値はむしろ複雑である。良好な近似はまづ § 13.5 及び § 13.6 の方法によつて無回答がもしもなければ層に於て必要である標本の大きさ n_0 を推定する事である。さてももしも $\pi_2 = 0$ であれば方程式 (13.14) から

$$n_0 = \frac{NS^2}{N\pi + S^2}$$

を得る。従つて方程式 (13.14) は又次ようにかける。

$$n_{opt} = n_0 \{1 + (k-1)\pi_2 \frac{S^2}{S^2}\} \quad (13.15)$$

各層に別個に適用したこの方程式は最適 n_k に対しての近似を与える。 k_k の値は各層に方程式 (13.13) を適用して見出される。

これ等の方法は比又は回帰推定で用い得る。比推定では S^2 及び S_2^2 及び S_{2d}^2 でおきかえられる。比推定に $d_i = y_i - R x_i$ である。回帰推定では S^2 は $S^2(1-f^2)$ となり S_2^2 は $S^2(1-f^2)$ となる。

§ 13.4 無回答に対する他の方法

Other techniques for non-response:—

時には無回答の問題を解決すると考えられる一つの方法は、無回答に出くわす時はいつでもある隣りの単位に対して資料を求め得る事である。例えばもしも標本に属する家が不在であると分つたならば、その町で隣りの家を訪問する。この様にして資料が得られる標本の大きさは最初に計画したときの大きさと同じ様にする。

この方法が遂行するすべでは第一の層に於て標本の大きさを増すのである。無回答層に対しては資料は得られない。

この方法は全く非有効的なものではない。何故ならば第一の層に於ける標本の大きさの増加から正確度の増加があるからである。併し § 13.2 で分つた様にこの増加はもしも π_2 が全く大であれば些細なものとなる様である。この代替の方法は抽出者が無回答の問題が適当に取扱はれたと述べて考えるならば臨実にそこなはれる。

無回答がまづ第一に不在による時にすぐれたやり方は *Politz* 及び *Lehmmon* によつて行はれた。(1949, 1950) すべての要求が野外調査の期間、例えば一月に 6pm と 9pm との間に行はれたと一つの要求が無回答者あたりに行われると仮定しよう。もしも一人の人がこの月にこの時間に家にいたならばその人が標本に属しておれば確実に分る。この時間に半分家にいた人はもしも我々が要求が期間の無作為の時刻に行われると仮定出来れば訪問者が訪ふる時在定である場合の 50% の機会である。

母集団に於ける人は在宅の時間の割合によつて層に分ける事が出来る。この割合を k 層で π_k としよう。もしも層 k のメンバーが標本に属するならば彼が見出される確率は π_k である。このやり方は我々が § 13.2 及び § 13.3 で採用した数学的モデルの一般化である。前の場合ではこのモデルは夫々 $\pi_k = 1$ 及び 0 なる二つの層のみである場合であつた。

新しいやり方では資料が得られる標本はその時間の大幅の時に在宅している人に重みをつけすぎる様に思はれる。もしもこ

平等の人がしばしば不在の人の特性と違つておれば、無回答の偏倚を生ずる。 π_k の値を与えるときの偏倚の大部分のものは除く事が出来る。人工的資料はこの過程を13.4表に示してある。

13.4表 不在の色々な割合をもつ例示された結果に対しての資料

π_k	n_k	n'_k	$\bar{y}_k = \bar{y}'_k$	y_k	y_k / π_k
1	100	100	1	100	100
0.5	100	50	2	100	200
0.25	100	25	3	75	300
総計	300	175		275	600

$\pi_k = 1, 0.5$ 及び 0.25 なる等しい大きさの三つの層がある。

$n = 300$ なる最初の標本の大きさが計画される(その内100は大体各層に属するであろう)。不在のため実際の標本の大きさ n'_k は夫々100, 50, 25となる。

$\bar{y}_k$ の仮定した値で以て、真の母平均 $\bar{y}$ は2である。我々は $\bar{y}_k = \bar{y}$ として層内抽出分散を無視した。観測標本総計 y は275となり、標本平均は $275/175 = 1.57$ となる。

これは不在が在宅よりも $\bar{y}_k$ の大きな値をもっているから負の偏倚である。

偏倚をうけない推定量は無回答の割合を層からの総計を逆数の重みづけをして得られる。

$$\bar{y}' = \frac{1}{n} \sum \frac{y_k}{\pi_k} = \frac{600}{300} = 2$$

此処に n は最初の標本の大きさである。

実際の施用ではある偏倚が推定量に残る。何故ならば $\pi_k = 0$ なる人は標本では全く示されていないからである。

主なる問題は π_k を推定する事である。Politz & Simon

Simmons により提案された方法に於て訪問者は、その五日前の夜に訪問をする日の特別な時に在宅するかどうかをたづねる。応答者はそのとき $1/6$ から $5/6$ を含む π_k の推定値でもつて6つの層に分けられる。もしも訪問をする時間が無作為であると考えられるとすればこの方法は π_k の不偏推定量を与えるが推定量の分散はむしろ大である。例えば6pmから7pmは規則正しく家にいるか、7pmから9pm 又は外出している人はもしも訪問者が7pm 前に訪問すれば1なる π_k を報告し、7pm以外に訪問すれば標本に属しないとす。その他の方法は6pmと9pmとの間の時間の5つの無作為時間について訪問する事である。

勿論訪問者の要求が無作為であるという仮定が未解決問題である。仮定の合理性についてのある判断は訪問者が訪問する時間の分岐により得る事が出来る。

この方法の分散公式及び無回答を減ずる方法との比較は、1949年に Politz & Simmons の論文に出ている。

その方法を用いる事に関心のある人はこの文献を参考にされたい。ここではあまり簡単に述べすぎたし $\bar{y}$ の不偏推定量を用いるに必要の仮定の注意深い論述が必要である。

§ 13.5 測定誤差 Errors of measurements

この言葉は一番目の単位での項目に対する正しい値 y_i とそれから推定が行われる計算に於てその項目に割り当てられる値 $\hat{y}_i$ との差を示すのに広い意味で用いられる。この意味で測定誤差は測定方法の不備から生ずる誤差と同様に資料を記録し、編輯し表示する際に導入される誤差を含む。測定誤差の問題は物理、化学、生物に於て古いものであつた。非常に多くの情報か測定器具及び多くの測定方法で役割を重なる観測者と共にその行為についてあつめられた。この知識の多くは標本調査における測定誤差に転用されねばならぬ。

情報や訪問に際してき業の意向に対しての答に於て口頭で与えられるときは測定誤差については情報は少ない。二人の間の相互作用が含まれるから誤差模型は複雑となる。

この問題の重要性は規則正しい人エ母集団及び訪問者の誤差の調査研究が増しつつある機関によつて今実現化されている。

§13.6 測定誤差に対する数学的模型 *A mathematical model for errors of measurement*

さて我々は測定誤差の主成分たるいくつかの数学的論述を作る。 i 番目の単位区に非常に多くの測定のくりかえしを考へ、 $y_{i\alpha}$ をして、 α 番目のくりかえしで得られた値としよう。そのとき次のようにかく

$$y_{i\alpha} = \eta_i + g_i' + e_{i\alpha} \quad (13.16)$$

此処に η_i = 正しい値

g_i' = 偏倚成分

$e_{i\alpha}$ = 平均値のなる無作為成分

この模型は、 i 番目の単位のくりかえし測定に於て誤差 ($g_i' + e_{i\alpha}$) は平均値 g_i' なるある度数分布に従うと假定している。よくコントロールされた測定器具では度数分布はしばしば正規型に近くなる。一方新しく又複雑な測定過程ではくりかえしは単なる度数分布に従うとは勿論考へる事は出来ない。即ちこの型は時と共に不定に変化する。この様な場合は我々は $e_{i\alpha}$ は確率の意味で無作為変量の如く作用するを假定するから此処に示される理論は適用されない。

次の段階は我々が一つの単位から他の単位に移動する時これ等の成分が如何に変化するかを考へる事である。色々複雑な事が起る。

偏倚成分 g_i' に対してはすべての単位に同様に影響する一定偏倚 η がある。母集団上ではかなり簡単な度数分布に従う成分 η_i がある。この成分は正しい値 η と相関がある。例えば測定装置は η の高い値をたえず過小評価し、低い値をたえず過大評価する。 g_i' と η_i の間の相関の結果生ずるところの相関には全く別に関連した単位上で η の値向の相互し関係の複雑な模型がある。最も簡単な例は“訪問者の偏倚”である。目立つた違いがしばしば同じ母集団の明らかに比較出来る部分を抽出する。置つた訪問者によつて得られた η の平均値に於て見出された。(Allman, 1941 及び Mahalanabis, 1946 を参照) 同じ影響は数物の

標本が違つた組によつてかりとられた時及び化学的又は生物的分析が違つた研究室で行われた時にあらわれた。人間の因子はほぼ同じ時間に測定される単位間の相関に原因するだけではない。多くの測定装置は天候によつて影響される。又あるものは品質が組毎に違う原料を使用する。

誤差 e_{ix} の“無作為”成分にせよとしてみる。それが従つて假定される度数分布は良好な測定システムではこの様に変化は些細なものであるが単位毎に変化する。 μ_i と同様に違つた単位上で e_{ix} の値間には相関がある。

測定誤差の成分をまとめるに 13.5 表の通りである。

記号	成分の性質
μ	すべての単位上での一定偏倚
μ_i	i が変化する時平均値となる度数分布に従ひ正しい値 μ_i と相関がある 偏倚の“変動”成分
e_{ix}	i が一定 x に対して変化する時 平均値となる度数分布に従う誤差の“無作為”成分。

これ等の度数分布は i が変化する時型が変化する。

我々は更に違つた単位上での μ_i の値は互に相関があり、違つた単位上で e_{ix} の値に対してと同様である事を注意した。

この種のモデル作成の更に詳細な説明は Jansen 等 (1951) によつてあたえられている。この論文は後節に與えた多くの結果を含んでいる。

13.7 一定偏倚の影響 *Effects of constant bias*

すべての単位上での測定 y_i はその大きさと未知なる一定偏倚 μ を

うけると假定しよう。そのとき単純無作為標本の標本平均 $\bar{y}$ は一定偏倚 μ をうける。標本平均につける推定誤差分数に於ては、この推定量は $(y_i - \bar{y})^2$ なる平方和から導かれるから偏倚はうちけされてなくなる。従つて標本資料から $\bar{y}$ に対する信頼限界の通常の計算は偏倚を考慮に入れぬ。同じ結果は層化無作為抽出にと成立する。

この立場は本質的に同帰及び比推定にと亦同じである。同帰推定を考へよう。

$$\bar{y}_{er} = \bar{y} + b(\bar{x} - \bar{x})$$

此処に μ_i 及び x_i 共に夫々一定偏倚 μ_y 及び μ_x をうける。最小自乗推定量 b は不変である。そして風 $(\bar{x} - \bar{x})$ は偏倚 μ_x はうちけされてなくなるから $\bar{y}_{er}$ は μ_y なる偏倚をうける事になる。 $V(\bar{y}_{er})$ の標本推定量は偏倚をうけない事を証明するのは容易である。

此推定では

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}} \bar{x}$$

偏倚は亦大標本では $E(\bar{x}/\bar{x})$ は x_i が一定偏倚をうけると近似的に 1 であるからオ一次近似に対して μ_y である。

大標本では分散の標本推定は

$$v(\bar{y}_R) = \frac{(N-1)}{Nn} \frac{\sum (y_i - \bar{y}_R)^2}{n-1}$$

は $E(\bar{y}_R - \bar{y})^2$ 即ち偏倚平均 $\bar{y}^*$ またりの分散の推定量として殆んど偏倚をうけない。

* 「ある意味では比推定は同帰又は単位当り平均推定量よりは一定偏倚によつてより影響される。もし μ_i と x_i の関係が強弱を通る直線 $\mu_i = \mu_i + \mu_x$ であれば μ_i と x_i の直線関係は原誤差を通らぬ。その結果は偏倚平均 $\bar{y}$ の推定量としての比推定の正確度は偏倚 μ_y 又は μ_x によつて影響されるという事である。この点はとしか比推定の使用についての決定が μ_i 及び x_i

同の観測関係にともなうてをれば實際上は大して重要ではない。とりまきとめるに一定偏倚は標本資料によつては見出し得ない。既に分つた様に(5.15)の信頼確率はとて推定平均値の標準誤差に対する μ の比が μ/σ 以下であれば殆んど影響しないうが比がこの値を越える時は信頼限界の計算はあやまつたものとなる。すべて同じ偏倚をうける多くの独立な推定量が平均されるときこの比は偏倚は一定のまま残つてゐるが平均値の標準誤差はきえてなくなるから増大する事は亦記憶しておかねばならぬ。時の変化の推定又は層毎の変化の推定は假りに偏倚が全般にわたつて一定にすれば不偏である。

5.13.8 単位毎に独立な成分の影響

Effects of components that are independent from unit to unit: -

として一定偏倚を無視すればモデルはその様になる。

$$y_{i\alpha} = \mu_i + \mu + e_{i\alpha}$$

i 番目の単位上で偏倚 μ_i と測定は無作為誤差 $e_{i\alpha}$ との間の区別はとして我々がそれに於て同じ単位が正確度を増すために数回測定される抽出計画を論じてゐるならば心に留めておかねばならない。論議は如何なる単位とたゞ一度しか測定されない場合に限定するから誤差($\mu_i + e_{i\alpha}$)は単一の項 $e_{i\alpha}$ にまきめる事が出来る。モデルは次の如く簡単になる。

$$y_{i\alpha} = \mu + e_{i\alpha}$$

母集団の性質は明らかに実現化されねばならぬ。くりがえし抽出では μ_i はこの単位に対して正しい値であるから i 番目の単位が標本にあらわれるときはいつでも μ_i の値は不変である。併し測定誤差 $e_{i\alpha}$ に対して新しい値をあてて単位の新しい測定が行はれる。 $e_{i\alpha}$ を含む如何なる表示に於て我々は単位のすべての可能な測定及び亦あてられた型のすべての単位又はすべての標本上共に平均する。一定偏倚は無理をされるので、母集団上の $e_{i\alpha}$ の平均値は0となる。これは次の様にかける。

$$E_i E e_{i\alpha} = 0$$

比に E は i 番目の単位の上での測定上の平均を表わし E_i はすべて単位上でのそれに附随する平均を表わす。仮説により $e_{i\alpha}$ は μ_i と相関がありそれは亦単位毎に相関がある。さて $e_{i\alpha}$ と $e_{i\beta}$ が標本の二つの置つた単位上で確率の意味で独立であるとするれば前章で与えた公式は假りに母集団が無限と考える事が出来れば妥当である。これは単純無作為抽出に対して容易に分る。次々の単位を独立に抽く事は標本の μ の次々の値がお互に独立である事と同じである。 $e_{i\alpha}$ の値はお互に独立であるが、我々がこれがそうであるに限定した事と同じである。従つて μ_i は次々の単位で独立であり、無限母集団からの無作為抽出に対する、通常の理論が μ_i に適用出来る。

特に、として $E(\mu_i) = \mu$ が真の母平均であれば

$$V(\mu_i) = E(\mu_i - \mu)^2 = \frac{\sigma_y^2}{n} = \frac{1}{n} (\sigma_y^2 + \sigma_e^2 + 2\rho_{ye} \sigma_y \sigma_e) \quad (5.17)$$

比に $\sigma_y^2 = E(\mu_i - \mu)^2$

$$\sigma_e^2 = E E (e_{i\alpha}^2)$$

$$\rho_{ye} \sigma_y \sigma_e = E_i E \{ e_{i\alpha} (\mu_i - \mu) \}$$

更に $\mu_{i\alpha}$ は無限母集団からの単純無作為標本の独立組であるから

$$v(\bar{y}) = \frac{1}{n} \frac{\sum (y_{i\alpha} - \bar{y})^2}{(n-1)} \quad (5.18)$$

は $V(\bar{y})$ の不偏標本推定量である。この場合に於て μ と測定誤差が居る時には推定抽出誤差に対する通常の公式が妥当である。特に二つの測定過程は與えられた費用に対して $v(\bar{y})$ のより小さい値を與えるのを見出す事によつて実験的に比較出来る。

同じ様な方法を假りに μ_i 及び $e_{i\alpha}$ の二つの測定誤差が単位毎に独立であり、 ρ_{ye} が無視出来るとすれば誤差分散の標本推定に対して前章で與えた公式は層化抽出、比及び回帰推定に対して妥当である事を示す事が出来る。

μ_i が無視出来る時公式に少し変化が必要である。これを単純無作為標本の平均値に対して簡単に示してみよう。有限母集団で

は母分散及び共分散は次の様に定義されるであろう。

$$S_y^2 = \frac{1}{(N-1)} \sum_{i=1}^N (y_i - \bar{y})^2$$

$$\sigma_e^2 = \frac{1}{N} \sum_{i=1}^N E(e_i^2)$$

$$\rho_{ye} S_y \sigma_e = \frac{1}{(N-1)} \sum_{i=1}^N E\{e_i y_i - \bar{y} e_i\}$$

σ_e^2 を見出すのに分母が N を用いる事に注意。これは結果を簡単にする助けとなる

標本平均の推定量の誤差は

$$\bar{y} - \bar{H} = (\bar{y} - \bar{H}) + \bar{e} \quad (13.19)$$

定理 2.3 により

$$E(\bar{y} - \bar{H})^2 = \frac{(N-n)}{Nn} S_y^2 \quad (13.20)$$

$\bar{e}$ の平均値に対して

$$E\bar{e} = E\left(\frac{1}{n} \sum_{i=1}^n e_i\right) = 0$$

同数の標本に於ける選った単位上での測定誤差は仮説により独立であるからである。

大きされるすべての単純無作為標本上での平均をとると

$$E\bar{e} = \frac{1}{nN} \sum_{i=1}^N E(e_i^2) = \frac{1}{n} \sigma_e^2 \quad (13.21)$$

積 $E\bar{e}(\bar{y} - \bar{H})$ に対して我々は n 番目の単位と関係のある一定の測定誤差 e_i があるを仮定しよう。そのとき有限母集団からの単純無作為標本に対する通常の理論により (定理 2.3)

$$E\bar{e}(\bar{y} - \bar{H}) = \frac{(N-n)}{Nn} \frac{\sum_{i=1}^N e_i y_i}{(N-1)}$$

此処にこの平均はこの一定の測定誤差の組にとつすべての単純無作為標本上にとられる。これ等の誤差のすべての可能な組について平均して ρ_{ye} の定義により

$$E\bar{e}(\bar{y} - \bar{H}) = \frac{(N-n)}{Nn} \rho_{ye} S_y \sigma_e \quad (13.22)$$

最後に (13.20) (13.21) 及び (13.22) から

$$V(\bar{y}) = \frac{(N-n)}{Nn} \{S_y^2 + 2\rho_{ye} S_y \sigma_e\} + \frac{\sigma_e^2}{n} \quad (13.23)$$

この公式は $f \cdot p \cdot c$ が無抽出でない時の (13.17) の代りである。もし $N=n$ であれば分母は 0 の代りに σ_e^2/n になり測定のある誤差が N 個の独立測定の平均にのこるという事象を反映して

る。このモデルは標本平均からの平均平方偏差

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

は $S_y^2 = S_y^2 + 2\rho_{ye} S_y \sigma_e + \sigma_e^2$

の不偏推定量である事を示す事が出来る。

$\bar{y}$ の推定分散に対する通常の公式 (5.2.6) は

$$v(\bar{y}) = \left(\frac{N-n}{N}\right) \frac{s^2}{n}$$

である。従つて

$$Ev(\bar{y}) = \frac{(N-n)}{Nn} \{S_y^2 + 2\rho_{ye} S_y \sigma_e + \sigma_e^2\}$$

(13.23) と比較する事によつて $v(\bar{y})$ は σ_e^2/n なる量の偏倚をもち通常小さいであろう事が分る。もし $f \cdot p \cdot c$ を $v(\bar{y})$ から除くと過大推定をする事になる。不偏推定量は σ_e^2 の知識なしには作る事は出来ない。

§ 13.9 選った単位上での誤差間の相関の影響

Effects of correlation between errors on different units: -

同じモデル即ち

$$y_{ix} = \bar{y}_i + e_{ix} \quad \text{で}$$

同じ標本に於て異なる単位上で e_{ix} 間に相関があると仮定しよう。無限母集団からの単純無作為抽出が仮定される。

$E(\bar{y} - \bar{H})^2$ を見出す際に前節に於ける結果 (13.17) に付け加へるべき唯一つの項は積の項

$$\frac{2}{n^2} E\bar{e} \left\{ \sum_{i,j} e_{ix} e_{jx} \right\} \quad (13.24)$$

我々は次の方程式により平均標本内相関係数 $\bar{\rho}_{ee}$ を定義する。

$$\bar{\rho}_{ee} \sigma_e^2 = \frac{2}{n(n-1)} E\bar{e} \left\{ \sum_{i,j} e_{ix} e_{jx} \right\}$$

従つて積の項 (13.24) は $\frac{(n-1)}{n} \bar{\rho}_{ee} \sigma_e^2$ となる。

この項を (13.17) に加えると

$$V(\bar{y}) = \frac{1}{n} \{S_y^2 + 2\rho_{ye} S_y \sigma_e + \sigma_e^2\} + (1 - \bar{\rho}_{ee}) \frac{\sigma_e^2}{n} \quad (13.25)$$

$v(\bar{y})$ の平均値は同様な方法で次である事が分る。

$$Ev(\bar{y}) = \frac{1}{n} \{S_y^2 + 2\rho_{ye} S_y \sigma_e + \sigma_e^2\} + (1 - \bar{\rho}_{ee}) \frac{\sigma_e^2}{n} \quad (13.26)$$

従つて標本の推定分散は偏倚がある。\$\bar{P}_{ee}\$ は大抵の測定誤差の型に対して正であるから \$\bar{P}_{ee}\$ は通常四角形である。四角形面ひどいかわが \$\bar{P}_{ee}\$ の値によると同様に \$\sigma_y^2\$ 及び \$\sigma_e^2\$ の相対的大きさによる。

組織抽出又はより一般的には戻還抽出の類似性は明らかである。

§ 13.10 相互貫入副次標本 Interpenetrating subsamples:-

測定誤差が違つた単位上で相関がある時には、不偏な標本分散推定量 \$V(\bar{y})\$ を得る方法及び相関が真の \$V(\bar{y})\$ を増大し従つて推定の正確度を増す程度を見出す方法を知りたいであろう。これは多くのインドの調査にそれを用いた Mahalanabis (1946) の方法によつて行われる。

\$n\$ 個の単位を \$m\$ 個の単純無作為標本が各々のグループが \$m = n/k\$ なる単位を含む \$k\$ 個の単位のグループは無作為に分けられる。標本をとる野外のは事は違つたグループに属する。二つの単位の測定誤差には相関がない様に計算される。例えば我々がうまく取扱はるらぬ相関は訪問者の偏倚から生じたと假定しよう。としとる人の訪問者の各々が違つたグループにわりあてられ違つた訪問者に対しての測定誤差間に相関がないとすれば今提案された方法の例となる。

同じ数学的モデルで二つの添字によつて単位を表わすに便利である。即ち

$$y_{ij} = \mu_j + e_{ij}$$

此処に \$i\$ はグループを表わし、\$j\$ はグループ内のメンバーを表わす。母集団は無限と假定し、母集団中の \$e_{ij}\$ の平均は 0 であると假定する。

グループ \$j\$ は単純無作為標本の無作為副次標本であるからそれその他ののは母集団からの単純無作為標本である。従つて方程式 (13.25) によりグループ平均 \$\bar{y}_j\$ の分散は

$$V(\bar{y}_j) = \frac{1}{m} (\sigma_y^2 + 2\rho_{ee}\sigma_y\sigma_e + \sigma_e^2 \{1 + (m-1)\bar{P}_{ee}\})$$

誤差は置つたグループに於て独立であるから \$n = mk\$ を入れて

$$V(\bar{y}) = \frac{1}{k} V(\bar{y}_j) = \frac{1}{k} \{ \sigma_y^2 + 2\rho_{ee}\sigma_y\sigma_e + \sigma_e^2 \{1 + (m-1)\bar{P}_{ee}\} \} \quad (13.27)$$

前の様にとしと無作為に選ばれた訪問者によつて単位の単一測定分散が

$$\sigma_y^2 = \sigma_y^2 + 2\rho_{ee}\sigma_y\sigma_e + \sigma_e^2$$

で示されるとすれば (13.27) は次の様にかける

$$V(\bar{y}) = \frac{\sigma_y^2 - \bar{P}_{ee}\sigma_e^2}{k} + \frac{\bar{P}_{ee}\sigma_e^2}{k} \quad (13.28)$$

同じ分散公式を導く他のモデルは

$$y_{ij} = \mu_j + \rho_i + e_{ij} \quad (13.29)$$

である。

このモデルに於て \$\rho_i\$ は \$i\$ 番目の訪問者の個人的偏倚である。訪問者の母集団上で \$E(\rho_i) = 0\$ (即ちすべての訪問者には通常偏倚はない) 及び \$E(\rho_i^2) = \sigma_\rho^2\$ と表わす。測定の“無作為”誤差 \$e_{ij}\$ は今 \$\rho_i\$ とは相関はあるが単位毎には相関はないと母集団上で平均すれば 0 になると假定する。

くりがえし抽出に於て我々は訪問者の假想無限母集団から \$n\$ 人の訪問者の無作為標本を各時に選ぶ。各訪問者は単位の単純無作為標本の無作為副次標本にわりあてられる。

\$V(\bar{y})\$ を見出すため通常次の様にかく

$$\bar{y} - \bar{\mu} = (\bar{y} - \bar{\mu}) + \bar{\rho} + \bar{e}$$

平方ですべての標本上で平均すると積 \$\bar{y}(\bar{y} - \bar{\mu})\$ からの貢献はなくなる。何故ならばくりがえし抽出に於ては與えられた訪問者の組はしばしば単位のあたえられた組をばじめに表われるからである。

同じ理由で \$E\{\bar{\rho}\bar{e}\}\$ は 0 である。従つて (13.29) から分つた \$\sigma_y^2\$ を代入して

$$\begin{aligned} V(\bar{y}) &= \frac{\sigma_y^2 + 2\rho_{ee}\sigma_y\sigma_e + \sigma_e^2}{k} + \frac{\sigma_\rho^2}{k} \\ &= \frac{\sigma_y^2 - \sigma_e^2}{k} + \frac{\sigma_\rho^2}{k} \end{aligned} \quad (13.30)$$

これは方程式 (13.28)

この方程式は訪問者の一定数を得る事が出来る。正確度に限界がある事を示す。與えられた k に対して正確度は $k=2$ のとき即ち、各訪問者がただ一つの単位を測定する時が最大である。(この結論は訪問者が等しい能力の訪問者のあつまりから無作為に抽出されると想定しているから、^的実際などのではない。) 実際には平均能力は多くの訪問者が見出されねばならぬときより少数の訪問者しかあつたときの方が高いであろう。

同じ仮定を以て、標本からの $V(\bar{y})$ の不偏推定量は

$$v(\bar{y}) = \frac{\sum_{i=1}^k (y_i - \bar{y})^2}{k(k-1)}$$

である事を容易に証明される。

これは相互貫入副次標本の方法の通常の特性である。さて、推定量が基いている自由度が $k-1$ である事に注意しなさい。これはともし少数の訪問者しか予期されないならば小さいであろう。標本資料の分散分析(3.6表)は亦興味のあるものである。平均平方に対して示された期待値はモデルから証明出来る。

3.6表 標本の分散分析に於ける期待値(単一単位を基礎とした)

	df	$m.s.$	期待値
グループ間	$(k-1)$	$d_b^2 = \frac{m \sum_{j=1}^k (\bar{y}_j - \bar{y})^2}{k-1}$	$(\sigma_y^2 - \sigma_g^2) + m \sigma_g^2$
グループ内	$k(m-1)$	$d_w^2 = \frac{\sum_{j=1}^k \sum_{i=1}^m (y_{ij} - \bar{y}_j)^2}{k(m-1)}$	$(\sigma_y^2 - \sigma_g^2)$

グループ内分散 d_w^2 はともし訪問者の偏倚を除く事が出来たならば得られるであろう単位当りの変動の不偏推定量である。一方 d_b^2 は調査に於ける単位当り実際の分散を推定する。二つの平均平方の比較は偏倚からの正確度に対する損失の程度を表わす。

$V(\bar{y})$ に対する公式は亦與えられた調査費用の立場に対してグループの最適メンバーを定めるのに同じにする事が出来る。適用する費用出数の簡単な型は

$$C = C_n + C_r k$$

ともし k 及び k が一定 C に対して $V(\bar{y})$ を最小にする様に選ば

るとすればグループの最適大きさは
(訪問者当りの訪問の数)

$$\left(\frac{n}{k}\right)_{opt} = \frac{\sqrt{\sigma_y^2 - \sigma_g^2}}{\sigma_g} \sqrt{\frac{C_r}{C_n}}$$

初期標本からの標本推定量は

$$\left(\frac{\bar{y}}{k}\right)_{opt} = \frac{\sqrt{m} d_w}{\sqrt{\sigma_y^2 - \sigma_g^2}} \sqrt{\frac{C_r}{C_n}}$$

構造的にはこれらの方程式は二段抽出に於ける最適抽出比及び最適副次抽出比をきめて3.10.6のどのと同じである。つまり、全体の数学的やり方は二段抽出に対してのどのと同じである。この分析は我々が平均の能力の影響なしに訪問者のメンバーを変える事が出来るという已に述べた仮定をうける。

この方法はともしグループが無作為副次標本であるければ適用出来ない。通常の実行は訪問者あたり旅行費用を減ずるために彼の家附近の小地理的面積に各訪問者を割当ててゐる事である。ともし、選んだ区域に対して平均 $\bar{y}_j$ 、同じ値の差があればこれ等の差はともし訪問者に偏倚があれば分散分析に表われる。従つて、 d_b^2 は単位当りの実際の分散の最大推定量となり、 d_w^2 はともし訪問者の偏倚を除く事が出来れば適用されるであろうところの分散の最小推定量となる。

あまり調査費用をよそないでこの困難性をさける一つの方法は標本を完全な面積(区域)に属化する事である。各層に於ける標本はそのとき無作為副次グループに分けられる。各訪問者は全標本上を巡りて層全体について旅行する事が必要である。ここに示された分析はそのとき個々の層に適用され $V(\bar{y}_{opt})$ の不偏推定量は通常の方法でくみだされる。問題の他の解に対してはHansen等(1951)を参照される。

相互貫入副次標本のこの方法は誤差の不偏推定量をあたえるばかりでなく、選んだ単位上での測定誤差に於ける相関の影響の大きさを探究出来るのでうまい方法である。その利用性はただ訪問者の偏倚を取扱う場合に限らない。即ち、選んだグループが穀物の

標本をとる違つた野外チームである場合、又は診断検査をする
違つた医者チームである場合等に適用される。その方法の限界
はある調査では膨大なものとなる。旅行費用の増大でありその方
法がすべての単位で一定である偏倚は取扱われぬという事実に
ある。

§ 13.11 結 び Summary: -

前章で與えた公式の効果の見地から誤差の附加的な要因は次の
様に分類される。

- i 単位毎に独立であり、全母集団で平均となる測定誤差は假リ
に $F \cdot P \cdot C$ と無視するに推定量の標準誤差計算に対しての通常の
公式に適當に差應が払われている。勿論この様な誤差は正確度
を減じ、その減少の程度がひどいかどうかを見出すためにはそ
の大きさについて研究する事は価値のある問題である。
- ii 無応答では測定された単位から計算した時の標準誤差に対す
る通常の公式はそれ等は応答者と無応答者間の差による偏倚を
無視するから推定量の過小推定量である様である。抽出者はこ
の問題に注意しなからとつて并解してはならない。理由の
あるすべての無応答の完全な記録は実行の本質的な部分である。
としと無応答が母集団のある一部分で多くの努力の犠牲に於て
減ずる事が出来れば *Hansen & Hurwicz (1946)* の方法はこ
の部分への原因をしっかりとわけて示している。
- iii としと測定誤差が単位毎に相関があれば、標準誤差に対する
通常の公式は偏倚がある。標準誤差はあまりに小にすぎ、何
故ならば相関は實際には大抵正にあらわれるからである。この
誤差の型を見出すのが困難であり恐らく注意されぬでみのが
されるのであろう。相互費入副次標本の方法は眞の誤差分散の
不偏推定量と同様に影響の大きさの尺度を與える。(期待され
る一定偏倚)
- iv すべての単位には同様に影響する一定偏倚は見出すのに最も
困難なものである。標本資料の標本ではこの偏倚は表わし得ぬ

いであらう。之等の問題については多くのなすべき仕事があ
る。恐らく殆んど緊急を要するのは、標本調査に於ける測定
誤差の性質及びその大きさについで資料をありゆる事であら
う。多くの場合に通常の公式及び方法は殆んどとるにたぬ程
しが混乱させられぬであらうという事が分るであらう。他
の場合に研究のある型でおこる様に無作為抽出誤差は我々のトラ
ブルの最小のものであり、正確な推定量は他の誤差原因の一つ
に於ける思いきつた減少が行われるに得る事が出来ぬという
事は明らかとなるであらう。亦立派に訓練を受けた人々が監督
をする人によつて及びより経験のある人から単位を再チェック
する事によつて何がなしとげられ得るかを研究する事が一層必
要である。