日本統計学会誌
第32巻，第3号，2002年
303頁～314頁

複数の項目をもつ標本調査のための最適配分†

西郷　浩*，美添泰人**，竹村彰通***，坂巻敏夫****

Optimum Allocation for Multivariate Surveys

Hiroshi Saigo*, Yasuto Yoshizoe**, Akimichi Takemura***
and Toshio Sakamaki****

抽出単位が複数の計量値をもち，それらの計量値ごとに推定の目標精度を定めた場合，層別抽出においてどのように標本を配分すべきかを論じる．この問題は，実際の標本調査において，（1）事業所ないし企業において複数の計量値を調査する場合，（2）調査区調査において事業所の規模・業種別に販売額を集計する場合，（3）事業所ないし企業の規模・業種の移動を考慮して標本を配分する場合，などに生じる．具体例として，平成6年の「商業統計調査」にもとづいて，調査区調査を想定したときの標本配分を考察する．複数の目標精度を同時に満足しながら標本の大きさを最少にするという最適配分が，非線形計画法の問題に還元できることを述べ，先の具体例において最適配分を計算して，われわれの提案する手法が十分に実用的であることを示す．

1. はじめに

抽出単位がただひとつの計量値をもつ場合，層別抽出における標本の最適配分は，ネイマン配分などによって求められる．これに対し，抽出単位が複数の計量値をもち，そのおのおのに目標精度が定められている場合，簡単な配分公式は存在しない．便宜的に，ひとつの計量値だけに着目すれば，あるいは，複数の計量値を適当な重みで加重平均すれば，ネイマン配分が適用できる．しかし，複数の計量値の中からひとつの計量値を優先する根拠に乏しい場合や，販売額と従業者数のように計量値の加重平均を考えることに無理がある場合には，別の解決方法が必要になる．以下で述べるように，複数の目標精度を満足しながら必要な標本の大きさを最少にするという最適配分は，非線形計画法の問題に還元できる．本稿では，標本の最適配分における非線形計画法の利用を実用的な観点から考察する．

複数の計量値がある場合の最適配分に非線形計画法が利用できることは，早くから指摘されている．たとえば，Kokan［8］は，層別無作為抽出や多段抽出における非線形計画法の利用を検討している．Huddleston, *et al.*［5］は，非線形計画法による最適配分とその他の便宜的手法による配分とを現実のデータによって比較し，非線形計画法の優位性を強調している．逆に，Kish［7］は，非線形計画法の利用について，むしろ否定的な見解を示している．最適配

† 本稿は，通商産業省（現経済産業省）における研究会，第7回日中統計シンポジウム（2000年，東京）および第69回日本統計学会（2001年，福岡）での報告にもとづいている．報告に対して多数の有益なコメントをいただいた．また，本誌レフェリーからも有用なコメントをいただいた．記して謝意を表したい．

* 早稲田大学政治経済学部　〒169-8050 新宿区西早稲田1-6-1, saigo@waseda.jp
** 青山学院大学経済学部　〒150-8366 渋谷区渋谷4-4-25, yasuto_yoshizoe@post.harvard.edu
*** 東京大学大学院情報理工学系研究科　〒113-0033 文京区本郷7-3-1, takemura@stat.t.u-tokyo.ac.jp
**** 経済産業省経済産業政策局調査統計部　〒100-8902 千代田区霞ヶ関1-3-1

分のための最適値探索のアルゴリズムは，Bethel [1], Chromy [3], Hughes and Rao [6], Kokan and Khan [9] などによって考案されている．以上の成果は，標準的な教科書，たとえばCochran [4] やSärndal, *et al.* [11] にも，概略が紹介されている．

しかし，以上の文献には応用面について十分な記述がなく，非線形計画法の適用可能性が統計作成者に利用しやすい形で整理されていない．さらに，わが国の官庁統計への非線形計画法の応用をあつかった論考は少ない．わずかに，「商業統計調査」をもとに，調査区調査を想定して近似的な最適配分を考察したYoshizoe, *et al.* [12] や，同じ想定で非線形計画法を使用したSaigo, *et al.* [10] があげられる．本稿の目的は，それらの成果をふまえて，最適配分問題への非線形計画法の利用を，統計調査の実情に即して考察することにある．

本稿の構成は以下のとおりである．まず，次節において，統計調査のどのような場面で，抽出単位が複数の計量値をもつと解釈できるのかを述べて，商店を対象とする調査区調査の標本設計を具体例として紹介する．第3節において，標本の最適配分が非線形計画法によって求められることを示し，それを実行するにあたっての初期値の設定方法を述べる．第4節では，第2節で紹介した実験例に非線形計画法を適用して，われわれの提案する手法が十分に実用的であることを示す．第5節で，今後の課題についてふれる．

2. 抽出単位が複数の計量値をもつ調査

2.1 抽出単位が複数の計量値をもつ調査の例

本節では，抽出単位が複数の計量値をもつと解釈できる調査を具体的に検討する．計量値の定義を工夫することによって，次節で紹介する手法が広範に応用できることがわかる．

説明のため，ここでいくつかの記号を導入しよう．抽出単位が，あらかじめL個の層に分けられているとする．第h層の大きさをN_h，そこにおける第i抽出単位の計量値ベクトルを$\boldsymbol{y}_{hi}=[y_{hi1}, y_{hi2}, \ldots, y_{him}]$と記す．$\boldsymbol{y}_{hi}$の次元$m$は，すべての抽出単位について同一であるとする．ただし，実際に計量値の数や種類が異なっている場合でも，該当する計量値をもたない抽出単位の計量値を形式的に0として集計できるのであれば，この点は本質的な制約にならない．以下で，抽出単位が複数の計量値をもつとみなせる調査の例をあげよう．

複数の調査項目

抽出単位と調査対象とが一致していて，複数の計量値が調査されるとする．たとえば，層別された企業を標本抽出して，総売上高，材料費，人件費，減価償却費，営業損益，金融費用，経常損益，設備投資額，ソフトウェア投資額などを調査する．調査項目別に集計・表章して，それぞれについて目標精度を設定する．

この場合には，y_{hi1}を総売上高，y_{hi2}を材料費，などと定義すればよい．

このような調査の例として，日本銀行「企業短期経済観測調査（短観）」があげられる．「短観」では，総売上高に目標精度を設け，ネイマン配分をもちいている．もし，調査項目の優先順位が自明ではなく，複数の調査項目に目標精度を設定するのであれば，次節の方法をもちいる方がよい．

集落抽出

調査費用の軽減などのために，官庁統計において調査区調査が利用される例は多い．調査対象は事業所や世帯などであるから，調査区調査は集落抽出の一種である．いま，調査区が層別されており，おのおのの層において調査区を標本抽出して，標本調査区内のすべての事業所の販売額を調査することを考えよう．規模・業種別に総販売額を表章し，目標精度は表章に合わせて設定する．

この場合は，調査区が抽出単位であり，y_{hij}が，第h層第i調査区において第j規模・業種に

属する事業所の販売額の総和と定義される．

こうした調査区調査は，総務省「サービス業基本調査」における従業者規模10人未満の事業所および新設事業所の調査や，経済産業省「商業動態統計調査」における従業者数19人以下の小売商店を対象とした指定調査区調査などに見られる．どちらの調査においても，次節の手法とは異なる方法で標本が配分されている．

規模・業種移動を考慮した最適配分

事業所や企業の調査において，抽出された事業所・企業の規模・業種が，調査前のそれと異なっていることが調査後に判明することがある．規模・業種別の集計は，調査時点で把握したものにもとづいておこなわれる．このような移動を考慮した上で，規模・業種別集計の目標精度を確保する問題も，抽出単位が複数の計量値をもつ場合の一種とみなすことができる．いま，過去の調査にもとづいて，規模・業種別に企業が層別されているとしよう．おのおのの層において企業を抽出し，規模・業種の移動が認められれば，事後的にそれを変更する．事後的な規模・業種によって販売額を集計し，表章される規模・業種ごとに目標精度を設定する．

この場合には，y_{hij} をつぎのように定義する．第 h 層第 i 企業の事前の規模・業種は h に対応する．この企業の事後の規模・業種が k に対応しているとする．$h \neq k$ であれば，規模・業種が移動したことを意味する．この企業の計量値ベクトル $\boldsymbol{y}_{hi}$ において，y_{hik} には販売額を代入し，$j \neq k$ については $y_{hij}=0$ とする．この定義は人工的に見えるけれども，事後的な規模・業種にもとづいて集計するということは，この定義を採用することと同じである．

平成10年の通商産業省「商工業実態基本調査」では，事後的な規模・業種にもとづいて販売額が集計・表章されている．達成精度も事後的な規模・業種にもとづいて評価されている．標本の配分は，ネイマン配分によって求められた．もし，規模・業種の移動があったとすると，その配分が最適であるとはかぎらない．

アクティビティ・ベースの集計

官庁統計においては，通常，事業所ないし企業がただひとつの規模・業種に属し，その所属にもとづいて規模・業種別の集計がなされる．もし，販売額をアクティビティごとに把握し，アクティビティ別に集計するのであれば，抽出単位が複数の計量値をもつことになる．目標精度もアクティビティごとに定めるのが自然である．

その場合に，y_{hij} は，第 h 層第 i 事業所・企業における第 j アクティビティによる販売額となる．

この例は，業種の移動を考慮した最適配分の一般化と見ることもできる．すなわち，現行において，ただひとつの業種（アクティビティ）に一括して計上されている事業所・企業の販売額を，活動内容別に分解して計上するのである．もし，将来的に，事業所や企業のアクティビティ別に販売額の調査・集計が可能となって，目標精度を集計値ごとに決めるのであれば，次節の手法によって最適配分が求められる．

2.2 調査区調査の実験例

ここで，問題を具体的に検討するために，Yoshizoe, *et al.* [12] が考察した，商店を対象とする調査区調査の標本設計を要約する．この設計は，もともと，平成11年に実施を予定されていた「商業統計調査中間年補完調査」のモデルとして考案されたもので，母集団情報として，平成6年に実施された「商業統計調査」を使用している．想定された標本設計は以下のとおりである．

- 従業者規模1-9人の商店を標本調査の対象とする．従業者規模10人以上の商店は悉皆調査すると想定しているので，標本調査の対象とならない．
- 抽出単位は調査区であり，抽出調査区内の商店がすべて調査される．

- 調査区を，その中に含まれる商店の従業者規模(5-9人と1-4人との2種類)およびその業種(46種類)の稀少性に応じて92の層に分ける．具体的には，5-9人規模の商店を含んでいる調査区のうち，もっとも稀少な業種に属する商店を含んでいる調査区を第1層に分類する．つぎに，残りの5-9人規模の商店を含む調査区のうち，二番目に稀少な業種に属する商店を含んでいる調査区を第2層に分類する．この手続きを業種の数（46）だけ繰り返せば，5-9人規模の商店を含む調査区はすべて第1層から第46層までに分類され，残りの調査区には1-4人規模の商店だけが含まれていることになる．残りの調査区について，上と同様に業種の稀少性にもとづいて層別する．
- おのおのの層において，調査区内の従業者数に比例する確率で調査区を系統抽出する．
- 調査された商店の販売額を，従業者規模・業種別に集計する．推定には，Horvitz-Thompson（HT）推定量を使う．
- 規模・業種別の総販売額に目標精度を設定する．推定量の変動係数を，5-9人規模（46業種)については，それぞれ0.05以内とし，1-4人規模(46業種)については，それぞれ0.1以内とする．

稀少な業種を優先して調査区を層別したのは，稀少な業種に属する商店が，より高い確率で標本に含まれるようにするためである．こうした層別が，販売額全体の推定効率を向上させるとはかぎらない．しかし，稀少業種を優先することによって，稀少業種の推定精度が確保しやすくなり，また，地域別の表章の際に，稀少業種の販売額推定値が0となる事態をできるだけ回避できるという効果がある．以上の標本設計は平成元年の「第1回サービス業基本調査」の考えに近いものであり，官庁統計調査における調査区調査として典型的なもののひとつと考えられる．

3. 非線形計画法による最適配分

3.1 非線形計画法の利用

本節では，抽出単位が複数の計量値をもつときの最適配分が，非線形計画法によって求められることを示す．前節と同じく，y_{hij}を第h層第i抽出単位の第j計量値とする．第j計量値の母集団合計 $Y_j=\sum_{h=1}^{L}\sum_{i=1}^{N_h} y_{hij}$ について目標推定精度が定められているとする．第h層における標本の大きさをn_hとすれば，最適配分は，制約条件

$$V(\hat{Y}_j)\leq c_j, \quad (j=1, 2, ..., m) \tag{1}$$

$$1\leq n_h\leq N_h, \quad (h=1, 2, ..., L) \tag{2}$$

のもとにおける最小化問題

$$\min \sum_{h=1}^{L} n_h \tag{3}$$

の解である．ここで，$\hat{Y}_j$はY_jの推定量であり，c_jは所定の目標精度から定められる，n_hに依存しない非負の値である．なお，この最適配分問題の若干の拡張が補論Aに示してある．

ここで，推定量 $\hat{Y}_j$，その分散 $V(\hat{Y}_j)$，および制約条件(1)におけるc_jについて詳しく述べておこう．まず，異なる層において標本が独立に抽出され，推定量 Y_j が，おのおのの層における合計 $Y_{hj}=\sum_{i=1}^{Nh} y_{hij}$ の推定量 $\hat{Y}_{hj}$ の総和 $\hat{Y}_j=\sum_{h=1}^{L}\hat{Y}_{hj}$ と定義されるのが通常であるので，$V(\hat{Y}_j)=\sum_{h=1}^{L} V(\hat{Y}_{hj})$ が成立する．

つぎに，$V(\hat{Y}_{hj})$の値は，標本抽出と推定量 $\hat{Y}_{hj}$ とに依存する．この点について，典型的なふ

たつの場合をあげよう．なお，記法の簡便のため，第 h 層において $i=1, 2, ..., n_h$ が標本に抽出されたとしている．

- 層別非復元無作為抽出のもとで，標本平均拡大型推定量 $\hat{Y}_{hj}=N_h\bar{y}_{hj}$（ただし，$\bar{y}_{hj}=n_h^{-1}\sum_{i=1}^{n_h} y_{hij}$）をもちいるとする．このとき，$V(\hat{Y}_{hj})=(n_h^{-1}-N_h^{-1})N_h^{-2}(N_h-1)^{-1}\sum_{i=1}^{Nh}(y_{hij}-Y_{hj}/N_h)^2$ となる．
- 層別確率比例系統抽出のもとで，HT 推定量 $\hat{Y}_{hj}=n_h^{-1}\sum_{i=1}^{n_h}(y_{hij}/p_{hi})$（ただし，確率比例抽出のための補助変数を z_{hi} と記し，$p_{hi}=z_{hi}/\sum_{i=1}^{N_h}z_{hi}$）をもちいるとする．このとき，補論 B に述べる条件のもとで，近似的に $V(\hat{Y}_{hj})=(n_h^{-1}-N_h^{-1})\sum_{i=1}^{Nh} p_{hi}(yh_{ij}/p_{hi}-Y_{hj})^2$ となる．

以上のいずれの場合においても，

$$V(\hat{Y}_{hj})=d_{hj}n_h^{-1}-e_{hj} \qquad (4)$$

と書き表せることに注意する．ただし，d_{hj}，e_{hj} は n_h に依存しない非負の値である．第 1 の例においては，$d_{hj}=N_h^2(N_h-1)^{-1}\sum_{i=1}^{N_h}(y_{hij}-Y_{hj}/N_h)^2$，$e_{hj}=d_{hj}/N_h$，第 2 の例においては，$d_{hj}=\sum_{i=1}^{N_h} p_{hi}(y_{hij}/p_{hi}-Y_{hj})^2$，$e_{hj}=d_{hj}/N_h$ となる．補論 C にあるとおり，層別非復元無作為抽出ないし層別確率比例系統抽出において比推定量をもちいた場合にも，$V(\hat{Y}_{hj})$ を式（4）によって近似的に表現できる．比推定量まで含めれば，現行の官庁統計における標本抽出と推定量との組み合わせのかなりの部分がカバーされる．以下，本稿では式（4）が成り立つものとする．

最後に，制約条件（1）における c_j について述べる．c_j は推定量の目標精度によって定められる．官庁統計では，推定量の目標精度を「誤差率」で指定することが多い．しかし，その定義には，以下で述べる 2 通りの流儀があり，注意を要する．

- 目標精度を「推定量の変動係数が α_j 以下，すなわち，$CV(\hat{Y}_j)\leq\alpha_j$」と定めるなら，$c_j=(\alpha_j Y_j)^2$ である．
- 目標精度を「相対誤差率が α_j 以下となる確率が 0.95 以上になる，すなわち，$\Pr\{|\hat{Y}_j-Y_j|/Y_j \leq\alpha_j\}\geq 0.95$」と定めるなら，$c_j=(\alpha_j Y_j/1.96)^2$ である．

目標精度が別の形であたえられたときには，それに合わせて c_j を計算する．

さて，以上の準備のもとに，最適配分問題（3）を非線形計画問題として書き換える．見通しをよくするため，$x_h=n_h^{-1}$，$a_{jh}=d_{hj}$，$b_j=c_j+\sum_{h=1}^{L} e_{hj}$ とおけば，最適配分（3）は，制約条件

$$\sum_{h=1}^{L}a_{jh}x_h\leq b_j, \quad (j=1, 2, ..., m) \qquad (5)$$

$$N_h^{-1}\leq x_h\leq 1, \quad (h=1, 2, ..., L) \qquad (6)$$

のもとでの最小化問題

$$\min \sum_{h=1}^{L} x_h^{-1} \qquad (7)$$

の解である．ここで $a_{jh}\geq 0$, $b_j\geq 0$ であることに注意しよう．さらに，$b_j=0$ となるのは，$c_j=0$, $e_{hi}=0(h=1, 2, ..., L)$ となる極端な場合，たとえば，復元の無作為抽出において $c_j=0$ とする場合，であるので，通常は $b_j>0$ を仮定してよい．x_h を連続量とみなせば，最適化問題（7）において，制約条件（5），（6）が空でない凸多面体をなし，目的関数（7）が厳密な凸関数であるから，この問題は一意的な最適解をもつ (Kokan and Khan [9])．本来なら，n_h が自然数であるから，x_h が連続量とならないので，最適化問題（7）の厳密な解法は複雑である．しかし，x_h を連続量とみなして最適化問題（7）の解を求め，$n_h=x_h^{-1}$ が直近の整数となるように

調整したとしても，目標精度(1)はおおよそ達成されるであろうし，全体の標本の大きさ(3)も最適値とそれほど変わらないであろうから，実用上問題はない(Särndal, *et al.*[11], p. 469)．実際，この近似はネイマン配分においても使われている．このような x_h^{-1} の整数値への調整を前提とすれば，非線形計画法によって最適配分が求められる．

3.2 初期値の設定

最適解を効率的にえるためには，初期値の設定が重要な問題となる．ここでは，Chromy[3]のアルゴリズムによってえられる解を初期値として使用することを考えよう．

第1節で述べたように，最適化問題（7）の数学的特性に着目して，最適解をえるためのアルゴリズムがこれまでに考案されている．しかし，それらの手続きはきわめて複雑なものが多い．たとえば，Bethel[1]のアルゴリズムは，収束が保証されているものの，実行には面倒な数値探索を必要とする．例外的に，Chromy[3]のアルゴリズムは，数値探索をまったく必要とせず，機械的に実行できる．ただし，収束に関して理論的な証明がなされておらず，制約条件（6）が $x_h>0$ に置き換えられているので，最適値探索のアルゴリズムとしては問題が残る．実際，制約（6）を満足しない解が求まることがある．しかしながら，Chromy[3]によれば，最適値にきわめて近い解がえられることが多い．したがって，第1段階でChromy[3]のアルゴリズムによって標本配分をいったん求め，第2段階で，それを初期値として，制約条件を満たさない初期値をあつかえる非線形計画法のアルゴリズムを実行するのが，容易で安全な解決策であろう．

Chromy[3]のアルゴリズムでは，非線形計画問題（7）において，Lagrange乗数を所与として解 $x_h(h=1, 2, ..., L)$ を求め，つぎにそれを所与としてLagrange乗数を更新する，というプロセスを反復する．初期条件はLagrange乗数にあたえらえれ，収束条件が満足されるまで計算が繰り返される．アルゴリズムの詳細はChromy[3]に譲る．ここでは，Bethel[1]の解説を参考に，表現を改めて計算手順を示す．まず，必要な記号を準備する．$b_j>0$ と仮定できるので，制約条件(5)の両辺を b_j で除して右辺を1に基準化したときの係数 $\tilde{a}_{jh}=a_{jh}/b_j$ を第 (j, h) 要素とする係数行列を $\tilde{A}$ と記す．$\boldsymbol{f}, \boldsymbol{x}^{(k)}$ を L 次元の列ベクトル，$\boldsymbol{g}, \boldsymbol{\lambda}^{(k)}$ を m 次元の列ベクトル，ϵ を収束の判定に必要な微小正数とする．$\boldsymbol{\lambda}^{(k)}$ は，第 k ステップにおける基準化されたLagrange乗数ベクトル，$\boldsymbol{x}^{(k)}$ は第 k ステップにおける解を表している．sum(•)はベクトルの要素の合計を，$*$ は行列の要素ごとの積(Hadamard積)を表す．また，スカラーを変数とする関数がベクトルに適用されたときは，要素ごとにその関数を計算するものとする．←は代入文を意味する．

1. 第1ステップ：$\boldsymbol{\lambda}^{(1)}\leftarrow[1/m, 1/m, \cdots, 1/m]'$, $k\leftarrow 2$.
2. 第 k ステップ：$\boldsymbol{\lambda}^{(k-1)}$ を所与として，以下を計算する．

(a) $\boldsymbol{f}\leftarrow\sqrt{\tilde{A}'\boldsymbol{\lambda}^{(k-1)}}$

(b) $\boldsymbol{x}^{(k-1)}\leftarrow 1/\{\mathrm{sum}(\boldsymbol{f})\boldsymbol{f}\}$

(c) $\boldsymbol{g}\leftarrow\tilde{A}\boldsymbol{x}^{(k-1)}$

(d) $\boldsymbol{g}\leftarrow\boldsymbol{g}*\boldsymbol{g}*\boldsymbol{\lambda}^{(k-1)}$

(e) $\boldsymbol{\lambda}^{(k)}\leftarrow\boldsymbol{g}/\mathrm{sum}(\boldsymbol{g})$

(f) もし，$\max_j|\lambda_j^{(k)}-\lambda_j^{(k-1)}|<\epsilon$ であれば，計算を終了し，$\boldsymbol{x}^{(k-1)}$ を最適解とする．そうでなければ，$k\leftarrow k+1$ として，つぎのステップに移る．

4. 実験例における最適配分の算出

この節では，第2節で述べた実験例について，最適配分を計算する．まず，Chromy[3]のアルゴリズムによって初期値を設定し，つぎに，制約条件を満たさない初期値も処理できる，

SAS の NLP プロシジャによって最適解を求める．非線形計画法のアルゴリズムは，デフォルトで選択させているが，制約条件の数から，双対擬似 Newton 法が選択され，双対 Broyden-Fletcher-Goldfarb-Shanno 公式が使われている．解の安定性を確かめるため，求めた解を初期値として再度非線形計画法を実行している．SAS プログラムとデータセットの構成を補論 D に示す．

比較のため，Yoshizoe, *et al.* [12] が考察した簡便法による配分も示す．これは，ここでの層別の特徴を生かして制約条件（5）を満足する解を導こうとするものである．層別の仕方から，たとえば，第 1 層形成の基準となった稀少な規模・業種に属する商店は第 2 層以降の調査区には存在せず，第 2 層形成の基準となった稀少な規模・業種に属する商店は第 3 層以降の調査区には存在しない．したがって，制約条件(5)において，$a_{jh}=0(h>j)$ となることがわかる．このことから，$j=1$ のとき，$a_{11}x_1\leq b_1$ から x_1 を最大（x_1^{-1} を最小）にする $x_1=b_1/a_{11}$ が求められる．つぎに，$j=2$ のとき，$a_{21}x_1+a_{22}x_2\leq b_2$ に $x_1=b_1/a_{11}$ を代入し，x_2 を最大（x_2^{-1} を最小）にする $x_2=(b_2-a_{21}b_1/a_{11})a_{22}$ が求められる．以下，逐次的に解を求められる．言い換えれば，簡便法は不等式制約（5）を等式制約とみなして連立方程式の解を求めているのと同じであり，不等式制約（5）の端点のひとつを探そうとするものと解釈できる．もし，逐次的に求められる解のすべてが範囲（6）に収まれば，実行可能解が求まる．層が少数であれば，この方法によって実行可能解がえられる公算は大きいであろう．

しかし，簡便法には少なくともふたつの欠点がある．第 1 に，たとえ首尾よく解が求まったとしても，それが最適である保証がない．第 2 に，逐次的な解が範囲（6）から逸脱し，場合によっては $x_h<0$ となることすらある．このような事態が起こりうる理由は，つぎのように説明できる．簡便法においては，h の小さい層から順番に，その層において優先された規模・業種の販売額の推定精度を確保するのに必要な，最少の配分が決まっていく．しかし，その配分が，その層よりも後で優先される規模・業種の販売額の推定精度を確保するのに十分であるとはかぎらない．もし，第 h 層にいたった時点で，それまでに決められた配分が第 h 層で優先される規模・業種の販売額の推定にとって過少であれば，$x_h=N_h^{-1}$ すなわち第 h 層を悉皆調査としても，推定精度が確保できなくなるため，$x_h<0$ という無意味な解がえられるのである．便宜的な処置方法として，$x_h<N_h^{-1}$ となった場合には $x_h=N_h^{-1}$ とし，これを「悉皆ルール」と名づける．また，$x_h>1$ となった場合には，各層から少なくとも 1 つの単位は抽出することとして $x_h=1$ と定め，これを「最少ルール」と名づける．これらはあくまでも便宜的な処置であり，完全な解決策ではない．とくに，「悉皆ルール」が適用された層において優先された規模・業種の販売額の推定精度は，目標精度を下回る．

表 1 には，最適解，Chromy [3] のアルゴリズムによる初期値に対応する解，簡便法による解が示されている．ただし，$n_h=x_h^{-1}$ は小数点第 1 位を四捨五入して整数化している．

表 1 から，最適値と初期値とはほとんどの層で一致しており，Chromy [3] のアルゴリズムが初期値の設定方法として機能していることがわかる．また，簡便法によると，調査区総数が 43,490 となり，最適解よりも 5 割弱多くなっていることから，簡便法には大幅な改善の余地があることがわかる．とくに，簡便法を使ったときには，第 30，41，43，47，48，52，53，56，58，63，68，71，72 層において「悉皆ルール」が適用されているので，これらの層で優先された規模・業種の推定精度が確保できない．その原因は，それらの層よりも番号が若い層，たとえば第 20 層などで，標本の大きさが過少なことにある．さらに，表 1 から，「悉皆ルール」が適用される層においては，最適配分と簡便配分との相違が極端に大きくなることもわかる．

表1 最適配分，初期配分，および簡便配分

h	1	2	3	4	5	6	7	8	9	10
最適配分	168	186	317	317	364	406	1058	981	467	1012
初期配分	168	186	317	317	364	406	1058	981	467	1012
簡便配分	168	186	317	317	364	406	1058	981	467	1012
h	11	12	13	14	15	16	17	18	19	20
最適配分	1565	314	402	436	251	951	1715	289	620	1536
初期配分	1565	314	402	436	252	951	1715	289	621	1537
簡便配分	1565	204	248	458	257	554	1841	171	670	160
h	21	22	23	24	25	26	27	28	29	30
最適配分	258	796	293	683	532	617	1761	194	1346	442
初期配分	258	796	293	683	532	616	1761	194	1346	442
簡便配分	126	636	244	1147	693	125	2202	112	1946	*2736
h	31	32	33	34	35	36	37	38	39	40
最適配分	292	547	1065	567	387	524	324	137	200	296
初期配分	292	547	1065	567	387	523	324	137	200	295
簡便配分	156	128	131	120	63	742	598	12	926	44
h	41	42	43	44	45	46	47	48	49	50
最適配分	2012	334	315	133	229	453	116	59	28	17
初期配分	2010	334	315	133	228	453	116	59	28	17
簡便配分	*3465	69	*3370	76	48	26	*238	*440	46	19
h	51	52	53	54	55	56	57	58	59	60
最適配分	21	50	37	24	417	121	20	197	4	9
初期配分	21	50	37	24	417	121	20	197	5	9
簡便配分	38	*1209	*817	11	11	*1086	35	*907	10	5
h	61	62	63	64	65	66	67	68	69	70
最適配分	10	17	8	6	6	17	14	34	5	2
初期配分	10	17	8	6	6	17	14	34	4	2
簡便配分	14	35	*1531	3	9	1	19	*1978	8	1
h	71	72	73	74	75	76	77	78	79	80
最適配分	8	19	5	1	1	1	3	1	1	1
初期配分	8	19	5	1	1	1	3	0	0	0
簡便配分	*1916	*1726	9	†1	†1	2	3	†1	†1	†1
h	81	82	83	84	85	86	87	88	89	90
最適配分	1	1	1	1	1	1	1	1	1	1
初期配分	0	2	0	0	0	0	0	0	0	0
簡便配分	†1	†1	2	†1	†1	†1	†1	†1	†1	†1
h	91	92	合計							
最適配分	1	1	29384							
初期配分	0	0	29369							
簡便配分	†1	†1	43490							

* と † とは，それぞれ，「悉皆ルール」と「最少ルール」とが適用されたことを意味する．

5. 結　　び

抽出単位が複数の計量値をもつとみなしうる統計調査について考察した．具体的に検討する

ため，商店を対象とした調査区調査を想定した．この場合の最適配分が非線形計画法によって求められることを示し．非線形計画法の実行にあたってしばしば問題となる，初期値の設定方法も検討した．例示した調査区調査における最適配分を，平成6年の「商業統計調査」を母集団情報として試算し，本稿で提案した方法が実用的であることを検証した．

今後の課題として，つぎのことがあげられる．実際の標本設計においては，前回の大規模調査の結果や，他の調査から利用できる代理変数などを使って標本を配分するので，最適化問題の解が真の最適配分とは多少とも乖離する．とくに，両者の乖離が，目標精度と達成精度とにどの程度の差をもたらすかは重要な問題である．差異の大きさは，調査の種類や代理変数の性質などによって異なると考えられる．実際の調査データにもとづいてこの点を確認することが望まれる．

補論 A：最適配分問題の若干の一般化

最適配分問題（3）において，変更可能な箇所を述べる．

- 単位調査費用が層ごとに異なるならば，第 h 層における単位調査費用を γ_h と記して，目的関数（3）を $\sum_{h=1}^{L}\gamma_h n_h$ で置き換える．
- 推定量の分散の評価が必要な場合には，おのおのの層から少なくともふたつの単位を抽出しなければならないので，制約式（2）を $2\leq n_h\leq N_h$ とする．
- 制約式の数は必ずしも計量値の個数 m と等しくなくてもよい．たとえば，$Y\equiv\sum_{j=1}^{m}Y_j$ の推定量 $\hat{Y}=\sum_{j=1}^{m}\hat{Y}_j$ に目標精度 $V(\hat{Y})\leq c$ を定めるのであれば，制約条件にこれをふくめればよい．本質的な点は，推定量に関する制約条件が式（5）の形に書き直せることである．

これらの拡張を考慮して本論を適宜変更することは容易である．

補論 B：確率比例系統抽出における Horvitz-Thompson 推定量の分散の近似式

ここでは，確率比例系統抽出のもとでの HT 推定量の分散の近似式について検討する．結論から述べるなら，それは，ある一定の条件が成り立つとき，確率比例復元抽出のもとでの Hansen-Hurwitz（HH）推定量の分散に有限母集団修正を乗じて近似できる．

記号の簡単のため，以下では添え字 h, j を省略し，大きさ N の層における標本の大きさが n であるとき，$i=1, 2, ..., n$ が標本に抽出されたものとする．抽出単位が正の補助変数 z_i に比例する確率で系統抽出されたとき，この層における計量値の合計についての HT 推定量を

$$\hat{Y}=\sum_{i=1}^{n}(y_j/\pi_i)$$

で定義する．ただし，π_i は，抽出単位 i が標本に含まれる確率である．確率比例系統抽出の場合には，$\pi_i=np_i$，ただし $p_i=z_i/\sum_{i=1}^{N}z_i$ である．この推定量の分散を $V(\hat{Y}_j)$ と表す．π_{ik} を抽出単位 i, k が同時に標本にふくまれる確率とすれば，

$$V(\hat{Y})=\sum_{i<k}(\pi_i\pi_k-\pi_{ik})(y_i/\pi_i-y_k/\pi_k)^2$$

がえられる（Cochran［4］, p. 260）．

$V(\hat{Y})$ の評価は，このままでは困難である．そこで，当該の層における y_i がつぎのような超母集団からの無作為標本であると仮定しよう．すなわち，$z_i(i=1, 2, ..., N)$ を所与として，

$$y_i=\alpha+\beta z_i+z_i^{g/2}u_i,$$

ただし，$u_i(i=1,2,...,N)$ は，相互に独立に，平均0，分散 σ^2 の分布にしたがうとする．この超母集団モデルは，確率比例抽出のもとで推定量の効率を比較するときによくもちいられており，$1<g<2$ が現実的な想定とされる (Cochran [4], p. 257)．この仮定に加えて，$z_i(i=1,2,...,N)$ の並び方が無作為であることも仮定しよう．これらふたつの仮定は，確率比例系統抽出のもとでの HT 推定量 $\hat{Y}$ と，つぎに説明する Rao-Hartley-Cochran (RHC)推定量との比較を可能とする．

ある層における計量値の合計の RHC 推定量 $\hat{Y}_{RHC}$ は以下のようにして求められる．まず，当該層のN個の抽出単位を無作為に n 個のグループに分割する．$V(\hat{Y}_{RHC})$ をできるだけ小さくするためには，グループの大きさがなるべく均等になるようにする．簡単のため，N/n が整数であると仮定しよう．つぎに，おのおののグループにおいて，z_i に比例する確率でひとつの抽出単位を選ぶ．第 k グループにおいて選ばれた抽出単位を i_k と記す．このとき，RHC 推定量は，

$$\hat{Y}_{RHC}=\sum_{k=1}^{n}(\sum_{[k]}z_i)(y_{j_k}/z_{i_k}),$$

と表される．ただし，$\sum_{[k]}$ は第 k グループに属する抽出単位についての合計を表す．$z_{i_k}/\sum_{[k]}z_i$ が第 k グループにおける抽出単位 i_k の抽出確率であることに注意すれば，$(\sum_{[k]}z_i)(y_{i_k}/z_{i_k})$ は，分割を所与としたときの，第 k グループにおける計量値合計の不偏推定量になることがわかる．RHC 推定量は，確率比例非復元抽出を近似的に実現する手法のひとつと位置づけられる．

さて，以降の説明は2段階で進む．第1に，先の超母集団モデルと抽出単位の並び方が無作為であるという仮定のもとで，$\hat{Y}$ と $\hat{Y}_{RHC}$ との間に

$$E_\xi V(\hat{Y}_{RHC})\gtrless E_\xi V(\hat{Y})\ (g\gtrless 1\text{ にしたがって})$$

が成り立つ (Cassel, *et al.* [2], p. 159)．ただし，E_ξ は，超母集団からN個の抽出単位を無作為に抽出することに関する期待値である．このことから，$g\doteqdot 1$ が想定できて，N がある程度大きければ，$V(\hat{Y})$ を $V(\hat{Y}_{RHC})$ によって近似してもよいであろう．

第2に，N個の抽出単位から z_i に比例する確率で n 個を復元抽出するときの HH 推定量を

$$\hat{Y}_{HH}=n^{-1}\sum_{i=1}^{N}(y_j/p_i)$$

と記す．$p_i=z_i/\sum_{i=1}^{N}z_i$ は，個々の抜き取りにおいて単位 i が選ばれる確率である．$V(\hat{Y}_{HH})$ と $V(\hat{Y}_{RHC})$ との間には，

$$V(\hat{Y}_{RHC})=\{1-(n-1)/(N-1)\}V(\hat{Y}_{HH})$$

が成り立つ．ただし，N/n が整数であることを使っている (Cochran [4], p. 267)．一方，

$$V(\hat{Y}_{HH})=n^{-1}\sum_{i=1}^{N}p_i(y_i/p_i-Y)^2$$

であるから，$(n-1)/(N-1)$ と n/N との差が無視できる程度であれば，$V(\hat{Y}_{\mathrm{RHC}})=(n^{-1}$

$-N^{-1})\sum_{i=1}^{N}p_i(y_i/p_i-Y)^2$ が成り立つ．

以上のふたつの段階を総合すれば，

$$V(\hat{Y})\doteqdot(n^{-1}-N^{-1})\sum_{i=1}^{N}p_i(y_i/p_i-Y)^2$$

が成り立つ．

補論 C：比推定量の場合の分散の評価式

比推定量のための補助変数を u_{hij} と記し，$U_{hj}=\sum_{i=1}^{Nh}u_{hij}$，$U_j=\sum_{h=1}^{L}U_{hj}$ と定義する．また，$\hat{Y}_{hj}=Y_{hj}/N_h$，$\bar{U}_{hj}=U_{hj}/N_h$，$R_{hj}=Y_{hj}/U_{ij}$，$R_j=Y_j/U_j$ と定義する．個別比推定量

$$\hat{Y}_{RS}=\sum_{h=1}^{L}\hat{Y}_{hj}\cdot U_{hj}/\hat{U}_{hj}$$

と総合比推定量

$$\hat{Y}_{RC}=\hat{Y}_j\cdot U_j/\hat{U}_j$$

の場合について述べる．ただし，上記における右辺の推定量は，層別無作為抽出の場合には標本平均拡大推定量を，層別確率比例系統抽出の場合には HT 推定量を表すものとする．抽出方法と推定量との組み合わせに応じて 4 つの場合について表す．

1．層別無作為抽出において個別比推定量をもちいる場合は，$d_{hj}=N_h^2(N_h-1)^{-1}\sum_{i=1}^{Nh}(y_{hij}-R_{hj}u_{hij})^2$，$e_{hj}=d_{hj}/N_h$ となる．
2．層別無作為抽出において総合比推定量をもちいる場合は，$d_{hj}=N_h^2(N_h-1)^{-1}\sum_{i=1}^{Nh}\{y_{hij}-\hat{Y}_{hj})-R_j(u_{hij}-\bar{U}_{hj})\}^2$，$e_{hj}=d_{hj}/N_h$ となる．
3．層別確率比例系統抽出において個別比推定量をもちいる場合は，$d_{hj}=\sum_{i=1}^{Nh}p_{hi}\{(y_{hij}/R_{hi}-Y_{hj})-R_{hj}(u_{hij}/p_{hi}-U_{hj})\}^2=\sum_{i=1}^{Nh}(y_{hij}-R_{hj}u_{hij})^2/p_{hi}$，$e_{hi}=d_{hj}/N_h$ とする．
4．層別確率比例系統抽出において総合比推定量をもちいる場合は，$d_{hj}=\sum_{i=1}^{Nh}p_{hi}\{(y_{hij}/p_{hi}-Y_{hj})-R_j(u_{hij}/p_{hi}-U_{hj})\}^2$，$e_{hj}=d_{hj}/N_h$ とする．

補論 D：SAS プログラムおよびデータセットの構成

第 4 節の実験例でもちいた SAS のプログラムとデータセットの構成を示す．データセットについては，構成のみを示している．データセットの中の /*　　*/ の箇所は注釈を表し，実際のデータセットからは削除しなければならない．

SAS プログラム

```
%let L = 92;
data prcsn(type = est);
  infile 'shogyo.dat' lrecl = 2000;
  input _type_ $ x1-x&L _rhs_;
 run;
proc nlp inest = prcsn outmod = model all;
    min y;
    parms x1-x&L;
    array x[&n] x1-x&L;
```

```
    y=0.;
    do h=1 to &L;
      y=y+1./x[h];
      end;
 run;
proc nlp inest=prcsn model=model all;
    min y;
    parms x1-x&L;
    include model=model;
 run;
```

データセット shogyo.dat **の構成**

```
PARMS x_1 x_2 x_3 ... x_L . /* 初期値，最後の . は必要 */
LE a_11 a_12 ... a_1L b_1 /* 制約式1，LE は less than or equal の意味 */
LE a_21 a_22 ... a_2L b_2 /* 制約式2 */
...
LE a_m1 a_m2 ... a_mL b_m /* 制約式 m */
LOWERBD 1/N_1 1/N_2 ... 1/N_L . /* 下限，最後の . は必要 */
UPPERBD 1 1 ... 1 . /* 上限，最後の . は必要 */
```

参 考 文 献

[1] Bethel, J. (1989). Sample Allocation in Multivariate Surveys, *Survey Methodology*, **15**, 47-57.

[2] Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*, Wiley.

[3] Chromy, J. R. (1987). Design Optimization With Multiple Objectives, *Proceedings of the Section on Survey Research Methods*, the American Statistical Association, 194-199.

[4] Cochran, W. G. (1977). *Sampling Techniques*, third edition, Wiley.

[5] Huddleston, H. F., Claypool, P. L., and Hocking, R. R. (1970). Optimal Sample Allocation to Strata Using Convex Programming, *Applied Statistics*, **19**, 273-278.

[6] Hughes, E., and Rao, J. N. K. (1979). Some Problems of Optimal Allocation in Sample Surveys Involving Inequality Constraints, *Communications in Statistics, Theory and Method*, **A 8**, 1551-1574.

[7] Kish, L. (1976). Optima and Proxima in Linear Sample Designs, *Journal of the Royal Statistical Society*, Ser. A, **139**, 80-95.

[8] Kokan, A. R. (1963). Optimum Allocation in Multivariate Surveys, *Journal of the Royal Statistical Society*, Ser. A, **126**, 557-565.

[9] Kokan, A. R., and Khan, S. (1967). Optimum Allocation in Multivariate Surveys: An Analytical Solution, *Journal of the Royal Statistical Society*, Ser. B, **29**, 115-125.

[10] Saigo, H., Yoshizoe, Y., Takemura, A., and Sakamaki, T. (2000). Optimal Allocation for Statistical Survey of Census of Commerce, *Proceedings of the Seventh Japan-China Symposium on Statistics*, 28 October-1 November: Tokyo, 331-334.

[11] Särndal, C. E., Swesson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer.

[12] Yoshizoe, Y., Saigo, H., Takemura, A., and Sakamaki, T. (1997). An Approximately Optimal Sampling Design for Commercial Statistical Survey, *Bulletin of the International Statistical Institute*, Contributed Paper Book 1, 613-614.