高齢者世帯の消費行動と物価指数*

宇南山　卓（財務総合政策研究所）
慶田　昌之（立正大学経済学部）

要　　旨

本稿では、年齢別の消費者物価指数を構築した。年齢が異なる家計は、ライフサイクルの違いから異なる財・サービスを消費しており、その行動の違いは直面する物価の違いに反映される。さらに、年齢が異なると同じ財を購入するとしても、異なる店舗を利用すると考えられる。店舗が異なれば、提供される小売サービスも異なり、物価動向にも影響を与えると考えられる。ここでは、全国物価統計調査・全国消費実態調査の業態別の情報を用いて、年齢別の購入店舗の違いも考慮した年齢別物価指数を計測した。こうした消費行動の違いを、過去 20 年の物価動向に適用すると、30～39 歳の家計に比べ 70 歳以上の家計が経験したインフレ率は約 4%(年率で約 0.5%ポイント)高かった。

キーワード：高齢者、物価指数、物価スライド、消費行動
JEL classification:　E31、D12

* 本研究の一部は、経済産業研究所のプロジェクトの一部である。吉川洋東京大学教授、森川正之経済産業研究所副所長には貴重なコメントを頂いた。記して感謝したい。

1. はじめに

本稿では、年齢別の物価指数を構築した。年齢が異なる家計は、ライフサイクルの違いから異なる財・サービスを消費しており、その行動の違いは直面する物価の違いに反映される。ここでは、年齢別の選好の違いが過去 20 年程度の長期の物価動向にどのような影響を与えたかを観察する。

年齢別の消費行動の違いの最も重要な部分は、年齢によって消費する財・サービスが異なるという点である。たとえば、高齢になると健康状態が悪化することや、高齢者は相対的には IT 関連の変化に十分に対応できないなどの理由で、保健医療の支出シェアは高めに情報通信関係費の支出は少なめに出る。

こうした財・サービスごとの消費パターンの違いは、公式の消費者物価指数でも「世帯主の年齢階級別指数」として公表されている。しかし、ウエイトの違いは 10 大費目別までしか公表されておらず、長期的に比較可能な時系列にもなっていない。また、研究者レベルでは、北村(2010)が、世帯ごとに支出シェアを計測して、「家計別の物価指数」を計測しているが、必ずしも年齢別の物価指数が目的となっていない。ここでは、年齢別の物価指数に注目し、その時系列的な動向を観察する。

さらに、年齢別の消費行動の違いとして、購入先の業態の違いを考慮する。阿部・外木(2007)は、日経 POS データを用いて、同一商品であっても店舗間で価格の動きが大きく異なることを示している。これは、店舗間で付随する小売サービスが異なることを示唆している。また、Aguiar and Hurst (2007) はスキャナー・データを用いて高齢者の購買行動が若年者と異なることを示している。すなわち、消費者は年齢によって求める小売サービスに違いがあり、小売サービスに対する選好によって購入する店舗を決めていると考えられる。実際、年齢によって購入する店舗の業態は大きく異なる。

これまで、財別のシェアには多くの注意が払われてきたが、購買行動そのものについて考慮して物価指数を計測した研究はない。また、公式の消費者物価指数(CPI)は、品目ごとに「販売数量又は従業者規模等の大きい店舗」を調査対象としており、購買行動の変化については考慮していない。宇南山・慶田(2008)でも指摘されたように、購買行動は経済厚生に大きな影響を与える可能性があり、それを無視しては適切な物価指数の計測はできない。

年齢別の物価指数を計測することは、政策的な意義も大きい。その理由の一つは、公的年金の物価スライド制を適切に運用するために必須の情報となることである。Stephens and Unayama (2011) によれば、引退後の高齢者の所得の 90%以上は公的年金であり、年金支給額は高齢者の生活水準の最も重要な決定要因である。その年金の実質的な価値の維持は重要な政策課題である。そのために物価スライドを適切に運用することは必須であり、高齢者世帯の消費行動を正確に計測して物価に反映させることが必要である。

現実の物価スライドでは、基準となる「物価」は「総務省において作成する年平均の全国消費者物価指数」であり、若年世代も含めた「平均的な消費者が直面する物価」となっている。より適切なスライドには、高齢者の消費する財の種類や購入先を考慮して、高齢者が直面する物価が平均的な物価の動向とどれだけ違うかを計測する必要がある。

財別の支出シェアの違いと購入先業態の違いを考慮して、物価指数を構築した。その結果、高齢者の物価指数は、過去 20 年間の累積で、財別ウエイトの違いが原因で 1.5%、購入先業態別ウエイトの違いが原因で 0.5%、平均的な世帯の物価指数を上回っていた。これは、高齢者でウエイトの大きな保健医療の物価が上昇したこと、一般小売店やデパートといった高齢者が多く支出をする業態の相対価格が上昇

してきたことによってもたらされた。

本稿の構成は以下の通りである。第 2 節では年齢別の物価の比較をすることの理論的な背景を確認した。第 3 節では、年齢別の購入行動がどのように異なるのかが検討された。第 4 節では、これらの違いを考慮した物価指数の動向を示している。第 5 節は、まとめと政策インプリケーションを論じている。

2. 世帯属性別の物価指数

2.1.直面する物価の違いと物価指数の違い

物価指数とは、多くの財・サービスの価格(価格ベクトル)を 1 つの値(スカラー)として表現したものである。ここでは、各財・サービスの価格がどのように集計され、物価として計測できるかを考える。

まず、消費者が効用最大化問題を解いていると考えれば、一定の効用水準 $\bar{u}$ を得るのに必要な支出額 X は、支出関数 e によって次のように書ける。

$$X = e(p, \bar{u}; \delta)$$

ただし、p は財・サービスの価格ベクトルで、δ は効用関数のパラメータであり家計属性による選好の違いである。

通常の物価指数の議論は、さらにδ を固定して、財・サービスの価格が変化した時の「真の物価」を次のように定義する（たとえば、Diewert (1981)を参照)。

$$P_1 = \frac{e(p_1, \bar{u}; \delta)}{e(p_0, \bar{u}; \delta)}$$

ただし、p_0 および p_1 はそれぞれ基準時点・比較時点での各財・サービスの価格ベクトルである。さらに、ラスパイレス指数によって計算される消費者物価指数(CPI)がその上限となることもよく知られている。すなわち、基準時点での各財・サービスの消費量ベクトルをq_0 として、

$$CPI_1 \equiv \frac{p_1 \cdot q_0}{p_0 \cdot q_0} \geq \frac{e(p_1, \bar{u}; \delta)}{e(p_0, \bar{u}; \delta)}$$

となる。

一方で、概念的には支出関数から、同一の財・サービスの価格ベクトルのもとで直面している物価が高いのはどのような家計属性であるかを知ることもできる。たとえば、δ_0 を基準となる家計の効用パラメータ、δ_1 を比較対象の家計の効用パラメータとすれば、

$$\mathrm{R} = \frac{e(p, \bar{u}; \delta_1)}{e(p, \bar{u}; \delta_0)}$$

ただし、R は同じ効用水準を得るのに必要な費用の相対的な高さを示す指数であり、比較対象の属性を持つ世帯が基準となる属性を持つ世帯とを比べたものである。たとえば、肉が安くて魚が高いような状況で、若年者は肉が好きで高齢者は魚が好きであれば、同じ効用水準を得ようとすると高齢者ほど多くの支出をする必要があるというイメージの比較である。

こうした世帯属性ごとの「同一の効用水準を得るために必要な支出額」を計測することは、政策的には重要な役割がある。たとえば、同じ金額でも世帯属性ごとに実質的な価値（その補助金によって得られる効用水準）が異なるのであれば、支給する補助金を調整することができる。また、異なる家計属性を持つ世帯間で効用水準を比較するのにも利用することができる。

しかし、残念ながら、効用そのものが観察可能ではないため、異なる効用関数を持つ家計間で物価を比較することは強い仮定を置かない限り不可能である。異なる家計属性を持つ世帯が、同じ価格ベクトルの下で、異なる行動をとっていたとして、それが効用水準の違いによるものなのか効用パラメータの違いによるものなのかを識別することは不可能なのである。さらに言えば、そもそも効用とは序数的な概念であり、異なる選好を持つ主体間では効用水準の比較は基本的に不可能なのである。言い換えれば、どのようなデータが利用できるにしても、識別は不可能なのである。

そこで、家計属性ごとに直面する物価の絶対水準を比較することではなく、「家計属性ごとの物価の変化」を比較することを考える。すなわち、世帯属性が k である世帯の消費者物価指数を

$$P_1^k \equiv \frac{e(p_1, \bar{u}; \delta_k)}{e(p_0, \bar{u}; \delta_k)} \tag{1}$$

と定義するのである。

この指数によって、たとえば、肉の価格も魚の価格も変化した時に若年者と高齢者のどちらが直面する物価がより大きく上昇したかが比較可能になる。もちろん、物価上昇率が高いからといって、高い物価に直面しているとは限らない。しかし、ある時点での状態を所与とすれば、その後の変化は計測できるのである。この「家計間の物価の絶対水準は比較が不可能」であることと、「家計属性間の CPI の違いによって直面する物価の変化の違いが比較できる」ということの区別は、以下の議論を通じて十分に注意する必要がある。

2.2.年齢階級別の物価指数

ここまで、「家計属性の違い」と一般的な表現を用いてきたが、現実に属性別の物価指数が計測できるケースは限定される[1]。属性別の指数が意味を持つのは、選好が安定的に異なる家計のグループが存在し、さらにその違いが公式統計等で観察可能でなければならい。ここでは、家計属性の違いのうち、特に年齢別の違いに注目する。

年齢別に注目する理由の 1 つは、先験的に消費する財・サービスが年齢によって大きく異なることで

[1] 北村(2010)は、世帯ごとに支出シェアを計測しているが、それが選好の違いによってもたらされているかは検証していない。

ある。高齢者は、平均的には健康状態が良くないため保健医療の支出シェアが大きい。また、子育て世代では極めて大きな支出シェアを占める教育費も、まだ子供のいない世代や子育てを終えた世代にとっては大きなウエイトを持たない。さらに、高齢者は相対的にはIT関連の変化に十分に対応できていないため、通信機器やパソコンなどに対する支出が小さいなど、消費パターンが異なることは十分に予想できる。

また、年齢別のインフレ率の計測が、世代別の再分配政策に重要な含意を持つことも理由の1つとなる。その典型が、「年金給付水準の実質価値の維持を自動的にはかるため、物価上昇分だけ年金額をスライドさせる制度（牛丸, 1996)」である物価スライド制度である。民間保険では回避が難しい「将来のインフレによる年金額の目減りリスク」に対応するための制度であり、高齢者にだけ影響のある制度である。それにもかかわらず、物価スライドの基準となる「物価」は、「総務省において作成する年平均の全国消費者物価指数」であり、若年世代も含めた「平均的な消費者が直面する物価」である。もし年齢別に直面するインフレ率が異なるなら、年金受給者の直面するインフレ率で物価スライドを実施すべきである。その基礎となる年齢別の物価指数の計測はここでの目的の一つである。

年齢別の物価指数を計測するということは、概念的には(3)式を計測することになるが、通常の消費者物価指数の手法を踏襲して、

$$CPI_1^k \equiv \frac{p_1 \cdot q_0^k}{p_0 \cdot q_0^k} \tag{4}$$

を計測することとする。すなわち、家計属性別に基準時点での消費ベクトルq_0^kを作成し、それをウエイトしてラスパイレス指数を計算するだけである。

総務省統計局が公表する公式の消費者物価指数でも、「世帯主の年齢階級別指数」が公表されている。これは、基準年の家計調査に基づき年齢別の支出シェアを計算し、ウエイト作成しているのである。その意味では、年齢別の選好の違いを財・サービスへの支出の違いだけで把握しているのである。

それに対し、ここでは同じ財・サービスへの支出であっても、購入先が違えば異なる消費行動であるとみなしている。阿部・外木(2007)は、日経POSデータを用いて、同一商品であっても店舗間で価格の動きが大きく異なることを示している。市場が十分に競争的であるなら、異なる価格が併存する（価格の高い店舗が存在できる）のは、店舗間で付随する小売サービスが異なるからだと解釈せざるを得ない。つまり、財・サービスの価格の違いとは小売サービスの違いなのである。Aguiar and Hurst (2007) はスキャナー・データを用いて高齢者の購買行動が若年者と異なることを示している。すなわち、消費者は年齢によって求める小売サービスに違いがあり、小売サービスに対する選好によって購入する店舗を決定していると考えられる。

さらに、店舗によってサービスが異なれば価格動向にも違いが生まれると考えられるため、その違いも定量的に把握する必要がある。ここでは、店舗ごとの小売サービスの違いを「業態」としてとらえ、業態ごとの価格の動向と消費者の購入行動を観察する。すなわち、総務省統計局の年齢別物価指数では、財・サービスがN種類あれば、pはN次元のベクトルとなっていたが、ここではK個の購入先業態それぞれで異なる小売サービスが付加されると考え$K \times N$次元のベクトルとして基準となる消費ベクトルを構築する。

3. 年齢別の購買行動の違い

3.1.データ

ここでは、年齢別に購入先業態別がどれほど異なり、業態ごとの価格がどの程度異なるかを検証する。業態ごとの価格は、日本全体に対する代表性を持つ統計として、全国物価統計調査の「業態別小売価格」の情報を用いることができる。全国物価統計調査とは、5 年に一度実施されていた価格の構造調査であり、地域間・店舗間・銘柄間の価格差を把握する統計である。毎月実施され消費者物価指数の基礎統計となる小売物価統計調査と同じ総務省統計局が作成しており、調査対象となる市町村数は 4 倍程度である。ここでは、1987 年調査から 2007 年調査の 5 調査分のデータを用いた[2]。

全国物価統計調査の分類に従い、ここでは業態を「一般小売店」・「スーパー」・「量販専門店」・「コンビニ」・「デパート」・「生協」・「その他」の 7 つに分類した[3]。ただし、1997・2002・2007 年の調査では上記の 7 業態に加えて「ディスカウントストア」調査されているが、ディスカウントストアは量販専門店に類似した業態であると考え、量販専門店とディスカウントストアの単純平均を「量販専門店」の価格として利用した。また、量販専門店・生協は 1997 年調査に導入されていたため、それ以前は欠損となっている[4]。

もちろん、すべての品目の業態別価格が意味を持つわけではなく、実際に全国物価統計調査でも調査されない品目もある。その理由は大きく分けて 3 つあり、業態で分けることができないもの（家賃、電気代、上下水道料など）、業態を定義するのが困難なもの（外食などのサービス関連品目）、価格が全国一律で業態間に差がないもの（教科書・学習参考教材、書籍・他の印刷物、たばこ）である。これらの品目については、業態間で価格差ないものとした。

一方、各家計がどのような業態で財・サービスを購入しているかについては、全国消費実態調査を用いた。全国消費実態調査は、5 年に一度実施される家計の所得・消費の構造統計であり、毎回約 5 万 5 千世帯が調査される。毎月調査されている家計調査よりも詳細な集計が公表されており、品目別・購入先業態別の支出金額が利用可能である。

全国消費実態調査も、全国物価統計調査と同じ総務省統計局の調査であり、業態の区分はほぼ同じである。ただし、単身世帯を含む全世帯ベースで、年齢別・購入先別の集計表は中分類までだけが利用可能であるため、ここでも中分類を最小の財・サービスの最小分類単位とした。

3.2.業態と価格

まず、業態別にどの程度の価格差があるのかを示すために、中分類別・業態別価格指数を作成する。このために、全国物価統計調査で調査された品目別・業態別の価格を品目別に平均が 1 となるような指数に変換し、それの指数を消費者物価指数の品目別ウエイトで集計し、中分類別・業態別指数を作成し

---

[2] 全国物価統計調査は、2007 年調査を最後に小売物価統計調査の一部となっている。

[3] 1997 年調査から、大規模店舗と小規模店舗で別集計となっているが、ここでは両者の単純平均をその業態の価格とした。

[4] 欠損した部分は最も高い業態と同じ価格としている。

た[5]。

表 1 は、2007 年の全国物価統計調査に基づく中分類別・業態別の価格指数である。基本的に、全国物価統計調査では品目ごとに銘柄や属性を細かく指定して調査しており、ほぼ同一の財の価格が調査されている。それにもかかわらず、多くの品目でコンビニ・デパートでは価格が高く、スーパー・量販店の価格が低い傾向がある。一般小売店は、特に食料品を中心に平均価格を下回っているが、スーパー等よりは高めである。

ただし、最も価格差が大きい大分類「被服および履物」に含まれる費目については、価格差の少なくとも一部が、業態ごとの財の品質差である可能性が否定できない。衣類などは、完全に同一の商品を調査するのが困難なためである。たとえば、「洋服」は「スーパー」が 0.52 に対して、「デパート」が 1.87 と 3 倍程度の価格差があるが、スーパーとデパートの洋服が同一商品とは考えられない。ただし、衣類等については販売している業態自体が付加価値を生んでいる可能性もあり、小売サービスの差が価格に反映されていると解釈することもできる。この問題については、次節で考察する。

3.3.年齢と購入先業態

表 2 は、全国消費実態調査から得られる年齢別の購入先業態別の支出シェアを示している。特に消費行動が異なりやすい食料(酒類・外食を除く)、家庭用耐久財、教養娯楽耐久財の購入先業態別の支出シェアを示している。

最も顕著な傾向は、高齢世帯になるほど「一般小売店」で購入する割合が、平均的な世帯よりも高いことである。ここでは全ては示していないが、「食料」ではすべての中分類項目で「一般小売店」での購入割合が平均的な世帯よりも高く、「乳卵類」では 10.6 パーセント・ポイントの差がある。その一方で、「スーパー」で購入する割合がすべての「食料」に含まれる中分類で低い。すなわち、高齢者世帯では食料を「スーパー」で購入せず、「一般小売店」で購入していることが分かる。

食料については、単身者の多い 30 歳未満の家計では「コンビニ」のシェアが高い。これは、中分類レベルでも顕著である。30 歳以上の家計についても、年齢が高いほどコンビニの割合が低いことは明らかであり、年齢別の購入先業態の違いを象徴するものとなっている。

一方で、耐久財については、65 歳以上の世帯では、「家庭用耐久財」では 9.4 パーセント・ポイント、「教養娯楽耐久財」については 16.7 パーセント・ポイント、平均的な世帯よりも「一般小売店」での購入割合が多い。これらの財では、若年層が「量販専門店」での購入割合が高く、「家庭用耐久財」では 6.4 パーセント・ポイント、「教養娯楽耐久財」については 13.1 パーセント・ポイントの差がでている。つまり、耐久財については、高齢者世帯では「量販専門店」で購入せず、「一般小売店」で購入する傾向があると言える。

このような購入行動の違いは、財に対するアクセスのしやすさに影響を受けたものと解釈される。たとえば、高齢者では一般小売店での購入が他の世代に比較して多い。これは自動車などの交通手段に制約があり、住居に近い一般小売店を利用している結果であるかもしれない。また、デパートなどの利用が多いのは、数少ない外出の機会を効率的に利用してまとめて購入活動をしている可能性がある。また、

[5] 1987 年と 1992 年は平均価格が報告されていないため、全業態の平均価格が計算されていないため、各業態の欠損値を除く単純平均を平均価格とした。

30 歳未満の若年者では単身者も多く、通常の時間外で買い物をする必要があり、コンビニの利用頻度が高いと考えられる。データとしては業態別の支出シェアの違いが、店舗への物理的・時間的なアクセスのしやすさの影響で決まっているなら、それこそ「小売サービスの違い」であり、価格差の源泉であると考えられる。

4. 年齢別の物価指数

4.1. 財別・業態別・年齢別のウエイトの構築

ここでの最初の目的は、年齢別に「財別・業態別」の消費ベクトル（すなわち、(4)式におけるq_0^k）を構築することである。これは、通常の CPI で「ウエイト」を作成することに相当するが、年齢別に構築しなければならない点と、購入先業態別にも分割する必要がある点で異なる[6]。

具体的な手順としては、まず財別にウエイトを分割しそれをさらに業態別に分割するという手順で構築する。すでに上で述べたように、購入先業態別の支出額は中分類レベルまで利用可能であるので、財別にも中分類レベルでウエイトを作成する。

年齢別・財別の支出額については、2004 年全国消費実態調査の家計収支編「世帯主の年齢階級別支出金額」から取った。公式の年齢別物価指数の元データである「消費者物価指数年報」の「世帯主の年齢階級別ウエイト」を使うこともできるが、消費の 10 大費目だけが公表されており、購入先業態別データと中分類レベルでは接続ができないため、ここでは全国消費実態調査を用いた。

図 1 は、ここで構築された年齢別・財別の支出シェアである。この図より、多くの世帯が無職となる 60 歳以降に、消費の構造に大きな変化が発生することが分かる。高齢者世帯は、「食費」と「光熱・水道」に対する支出が若年者世帯よりもの数%ずつ多い。また、「保健医療」のウエイトは、60 歳未満の世帯はで 3%前後であるのに対し、60 歳以上では 6%と倍増する。それに対し、「教育費」のウエイトは、40～49 歳階級でピークの 10%程度であるのに対し、60 歳以上の世帯では実質的に 0%である。さらに、「交通・通信」は、60 歳大きな断絶があるわけではないが、高齢者世帯になるほどシェアが低い。これは、携帯電話の通信料が、高齢者層ではそれほど大きくないことに起因する。図 1 にはないが、中分類レベルでの詳細な内訳を見ると、例えば、「食料」のうち「肉類」に対する支出は若年層で多く、「魚介類」に対する支出は高齢者層で多い。また、「被服および履物」に含まれる「洋服」に対する支出は若年層で多く、老年層で少ない。

この財別のウエイトを、購入先業態別に分解する。全国消費実態調査では年齢階級別・購入先業態別の支出金額も公表しており、その割合を財別シェアにかけることで分解する。すでに表 2 で見たように、一般に 65 歳以上の世帯では「一般小売店」で購入する割合が高く、「スーパー」や「量販専門店」の割

[6] ただし、世帯主の年齢で分類されていることには注意が必要である。世帯調査によるデータで年齢別の消費を観察する場合、世帯は世帯主の年齢で分類せざるを得ない。しかし、たとえば 3 世代同居のケースでは、祖父・祖母が世帯主であれば若年世代の消費が高齢者の消費として、逆に息子・娘が世帯主であれば高齢者の消費が若年世代の消費に見なされてしまう。つまり、同居高齢者の消費行動が別居高齢者と大きく異なる場合には、世帯主が 65 歳以上の世帯だけを見ても「代表的な高齢者」を観察することはできない。ここでは、世帯主が 65 歳以上の世帯の世帯員は平均で 1.95 人でありそのうち 65 歳以上の世帯員が 1.45 人であり、世帯主が 65 歳以下の世帯の平均世帯員数 2.65 人のうち 65 歳以上の世帯員は 0.57 人であることから、世帯主の年齢で分類することの影響は小さい。

合は低い。

この財別のシェアと購入先業態別のシェアの 2 つをかけ合わせ、財別・購入先業態別にウエイトを構築した。ただし、購入先に分割できない中分類項目(家賃・光熱水道・サービスなど)は、購入先別の分割はしなかった。その意味で、年齢別の消費パターンに大きな違いがある、医療保健と教育については業態の違いが反映されていないことには注意が必要である。

4.2.財別・業態別の価格指数の構築

基準となる消費ベクトルが財別・購入先業態別になっているため、もちろん価格ベクトル（すなわち(4)式の p）も財別・業態別に構築する必要がある。価格ベクトルについては、基準時点と比較時点のデータが必要なため、時系列方向と業態別のクロスセクション方向で別のデータを用いた。時系列的には、全購入先業態の平均の財別価格指数を作成し、各時点での各購入先業態の価格の平均価格との比をかけることで購入先業態別の時系列比較可能な価格指数を構築するのである。

時系列方向には、総務省統計局の公表する消費者物価指数(CPI;2005 年基準)を利用した。CPI の価格データは小売物価統計調査で収集されるが、その調査店舗は「調査品目ごとに販売数量又は従業者規模等の大きい店舗の順に」選定されており、業態別の価格動向は考慮されていない。ただし、実際には主要な購入先業態は含まれるように調査されており、時系列的な動向は全購入先業態の平均とみなすことができると考えられる。

一方、クロスセクション方向には、全国物価統計調査で調査された品目別・業態別価格を業態別の平均が 100 となるように換算したものを用いる。全物価統計調査は、すでに見たように、1987 年から 2007 年までの 5 年毎に調査されており、ここでも購入先業態別の指数を 5 年毎に作成する。また、1987 年と 1992 年の全国物価統計調査では、全業態の平均価格が計算されていないため、各業態の欠損値を除く単純平均を平均価格とした。さらに、大規模店舗と小規模店舗に分けて価格が調査されている場合には、単純平均をその業態の価格とした。

4.3.年齢別の価格指数の構築

上で構築した財別・購入先業態別の消費ベクトルをウエイトとして、財別・購入先業態別の価格ベクトルを指数化した年齢別の物価指数を示すのが表 3 であり、メインの結果である。使用した消費者物価指数が 2005 年のものであり、2005 年が 100 となるように基準化されているが、計算された指数を定数倍することで 1987 年を 100 となるように換算している。

第 1 列の「平均」は、財別のウエイトも購入先業態のウエイトも全世帯平均のものを用いて計算された結果である。平均の物価指数は 2007 年に 108.50 で、20 年間で 8.5%の物価上昇があったことになる。公式の CPI は、1987 年を 100 とすれば 2007 年は 113.08 であり、同時期に約 13%の物価上昇となっている。これは、ここでの計算の元になるウエイトが、2004 年の業態別支出シェアで固定されている影響であり、相対的に価格が上昇した業態ほど 2004 年時点での支出シェアは低いことを意味している。

その横からの各列、30 歳未満、30～39 歳、・・・が年齢別の指数であり、一番右側の列は 65 歳以上の世帯を再集計したものである。65 歳以上の世帯の物価指数は、1987 年を 100 として、2007 年が 110.5

であり、平均的な世帯と比較すると 2.0%ポイント高い。また、30 歳未満を除くと、年齢が高いほどインフレ率も高くなっている。言い換えれば、高齢になるほど高いインフレを経験してきたのである。

この乖離がどのように発生したかを見ているのが、表 4-1 と表 4-2 である。表 3 は、財別ウエイトも購入先業態別ウエイトも年齢別のものを用いており、財別ウエイト・購入先業態別ウエイトを全年齢の平均シェアを用いた物価指数との差は、2 つのウエイトの差の合計となっているのである。表 4-1 及び表 4-2 では、それぞれ財別ウエイトのみ、購入先業態別ウエイトのみを年齢別のものとして、他方は全年齢の平均を用いている。もちろん、2 つのウエイトのクロス項も存在するが、1 次近似としてはこの分解が意味を持つ。

表 4-1 の財別ウエイトのみ年齢別としたものは、家計は財に関しては若年者と異なる選好を持つが、購入先の小売サービスに対しては同一の選好を持つと仮定したケースとみなすこともできる。上で見たように、年齢ごとの財別ウエイトは、10 大費目の教育・保健医療および IT 関連の耐久財で大きく異なっていた。保健医療のウエイトは 60 歳未満の世帯はで 3%前後であるのに対し、60 歳以上では 6%と倍増する。一方で、保健医療は制度的な要因によって特に価格上昇率が高い。また、IT 関連の耐久財は大幅に価格が低下したが、高齢者の支出シェアは低い。こうした要因によって、高齢者の経験したインフレ率は平均よりも約 1.5%ポイント高くなっている。一方で、財別に物価上昇率の高かった保健医療・教育のどちらも支出シェアが小さく、IT 関連の支出が多い 30 歳代は、平均よりも 1.8%インフレ率が低かった。

表 4-2 は、財別ウエイトは平均の支出シェアを用いるが、購入先業態別は年齢ごとのウエイトを用いた指数を作成し、平均との差をとったものである。これは、高齢者が若年者と財に対しての選好は同じで、小売サービスに対する選好だけが異なるケースに該当する。購入先業態別のウエイトが 65 歳以上では、平均よりも約 0.5%ポイント高くなっている。これは、高齢者が選好する一般小売店のような伝統的な業態が、量販専門店やスーパーと比較して物価の上昇率が高かったことを示している。年齢別に見れば、若年者ほどインフレ率の低い業態で購入していたことが分かる。

5. まとめとインプリケーション

本稿では、高齢者と若年層の消費行動の違いを考慮することで、年齢ごとに直面する物価がどのように異なるかを検討した。高齢者と若年層では、支出する財も異なり購入先も異なり、同一の消費行動をとるわけではない。支出する財の違いを考慮すれば、過去 20 年間の高齢者の直面する物価は、平均的な消費者の直面する物価よりも 1.5%程度高い率で上昇していた。また、購買行動の違いによって、0.5%程度高い率の物価上昇に直面していた。

財別の支出シェアで大きな違いが生まれたのは、IT 革命によってハイテク家電やパソコンなどの電子部品集約的な財に対するシェアの違いが原因であった。これらの製品では品質向上が大きく、物価指数が大幅に低下した一方で、高齢者これらの IT 関連の消費のシェアが若年層よりも低かった。言い換えれば､高齢者が IT 革命の恩恵を若年層ほどには受けることができなかったために生じた差である｡一方で、購買行動による違いは、ディスカウント店の一般化の影響である。過去 20 年に日本では紳士服や家電の量販専門店、ドラッグストアなどが急激にシェアを伸ばした。一方で、高齢者は、価格の低下幅の小さかった、いわゆる「一般小売店」や「デパート」などで購入することが多かった。すなわち、高齢者は、

IT 革命のみならず「流通革命」にも十分に対応できなかったのである。

補足として、ここで財別・購入先別の物価指数を構築するのに、基本的にラスパイレス指数算式を用いていることには注意が必要である。すなわち、ウエイトとして、2004 年の全国消費実態調査の支出シェア・購入先別支出シェアを固定して用いている。こうした特定の消費パターンや購買パターンに基づいた物価指数は、財の選択や購入先の選択といった代替行動は考慮していないため、経済学的に適切な「直面する物価の違い」と解釈することはできない。

これは、物価指数の上方バイアスの一部であり、代替バイアスとして知られる問題である。物価指数の上方バイアスとは、現実に計測される CPI は、効用水準を一定に保つために必要な支出を示す「真の物価指数」に比べ、インフレ率が高めに出るという性質である。Unayama (2004; 2008)は、日本の公式統計の作成手順や物価の動向から、年当たりのバイアスを 0.07%程度と推計している。すなわち、物価指数の上方バイアスのうち少なくとも「上位代替バイアス」はそれほど大きくない。

さらに、より一般に、CPI による物価スライドの問題点を考えるためには、物価指数の上方バイアスについてより詳細に検討することは必要である。アメリカでは、1996 年に「ボスキンレポート」と呼ばれる消費者物価に関する諮問委員会の答申(Boskin, et al., 1996)が出されており、年 1.1%の上方バイアスがあることが指摘されている。このバイアスは、基本的に毎年発生するものであり、1.1%のバイアスを持てば、10 年後には年金額が「実質的な価値を維持する水準」よりも 11%過剰になることを意味する。言い換えれば、CPI を適正化すると年金支給を大幅に抑制できることを意味する。[7]

ただし、日本の CPI の上方バイアスは、アメリカほどは大きくないと考えられる。白塚(1998)は、ボスキンレポートに準じて日本でのバイアスの大きさを計測し、年 0.9%程度と結論付けていた。それに対し、総務省統計局(1999)では、アメリカでの問題が必ずしも日本には当てはまらないことを指摘し、CPI の上方バイアスが十分に小さいと反論している。実際、バイアスの半分以上が品質調整によって生じているが、もともと価格調査の銘柄管理の方法が日米で異なっており、アメリカでの問題の大部分は日本では生じていなかった。さらに、2000 年には品質調整の問題が特に重要となるパソコンとデジタルカメラについては、より洗練された手法であるヘドニック法が導入されている。また、サービス関連についても品質調整の問題が多く指摘されている。本稿では、「真の物価指数」と「計測された物価指数」の差という技術的な面については分析の対象とはしなかったが、結論には大きな影響はないと考えられる。

現在、物価指数は、年金の物価スライド制の基準としてではなく、デフレーションを背景とした金融政策の文脈でも注目されている。物価指数の政策的な位置づけがこれまでになく高まっており、経済学的な観点から物価指数とは何かを明らかにし、正確に把握することは重要な課題であろう。

---

[7] ボスキンレポートに対する批判や、統計作成当局の対応については Gordon (2006)を参照。

<参考文献>

牛丸 聡 (1996) 『公的年金の財政方式』、東洋経済新報社.

宇南山 卓・慶田 昌之(2008)、「流通業における規制緩和の効果:少子高齢化社会へのインプリケーション」、RIETI Discussion Paper Series 08-J -047

北村 行伸(2008)「家計別物価指数の構築と分析」『金融研究』第 27 巻第 3 号、2008 年 8 月、pp. 91-150.

白塚 重典 (1998) 『物価の経済分析』東京大学出版会

総務省統計局(1999)「消費者物価指数の精度について」総務省統計局ホームページ「消費者物価指数に関する Q&A」(http://www.stat.go.jp/data/cpi/3.htm)

Aguiar、 M. and E. Hurst (2007) "Lifecycle Prices and Production、" American Economic Review、 Vol.97、 No.5、 pp.1533-1559.

Boskin、 M.、 et al. (1996) "Toward a More Accurate Measure of Cost of Living: Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index." Washington DC: U.S. Govt. Print Office.

Cutler、 David、 and Ernest R. Berndt (2001) *Medical Care Output and Productivity.* NBER Studies in Income and Wealth vol 62. University of Chicago Press.

Diewert, W. E. (1981) "The Economic Theory of Index Numbers: A Survey." A. Deaton ed. *Essays in the theory and measurement of consumer behaviour*. pp.163-208; Cambridge University Press: London.

Gordon、 R. J. (2006) "The Boskin Commission Report: A Retrospective One Decade Later、" *International Productivity Monitor* (Centre for the Study of Living Standards)、 vol. 12、 pages 7-22、 Spring.

Stephens、 M. and T. Unayama (2011) "The Consumption Response to Seasonal Income: Evidence from Japanese Public Pension Benefits" Forthcoming in *the American Economic Journal: Applied Economics.*

Triplett、 Jack E. (1999) *Measuring the Prices of Medical Treatments*、 The Brookings Institution.

Triplett、 Jack E.、 and Barry P. Bosworth (2004) *Productivity in the U.S. Service Sector: New Sources of Economic Growth*、 The Brookings Institution.

Unayama、T.、(2004) "Upward Bias in the Consumer Price Index Under the Zero Inflation Economy、" *Economics Letters*、 vol. 85、 pp. 139-144.

Unayama、 T.、 (2008) "The Demand System and the Substitution Bias in the CPI: Evidence from the Japanese Household Survey Data、" *Applied Economics*、 vol. 40、 pp. 1795-1806.

図 1 年齢階級別の財別支出シェア(10 大費目)

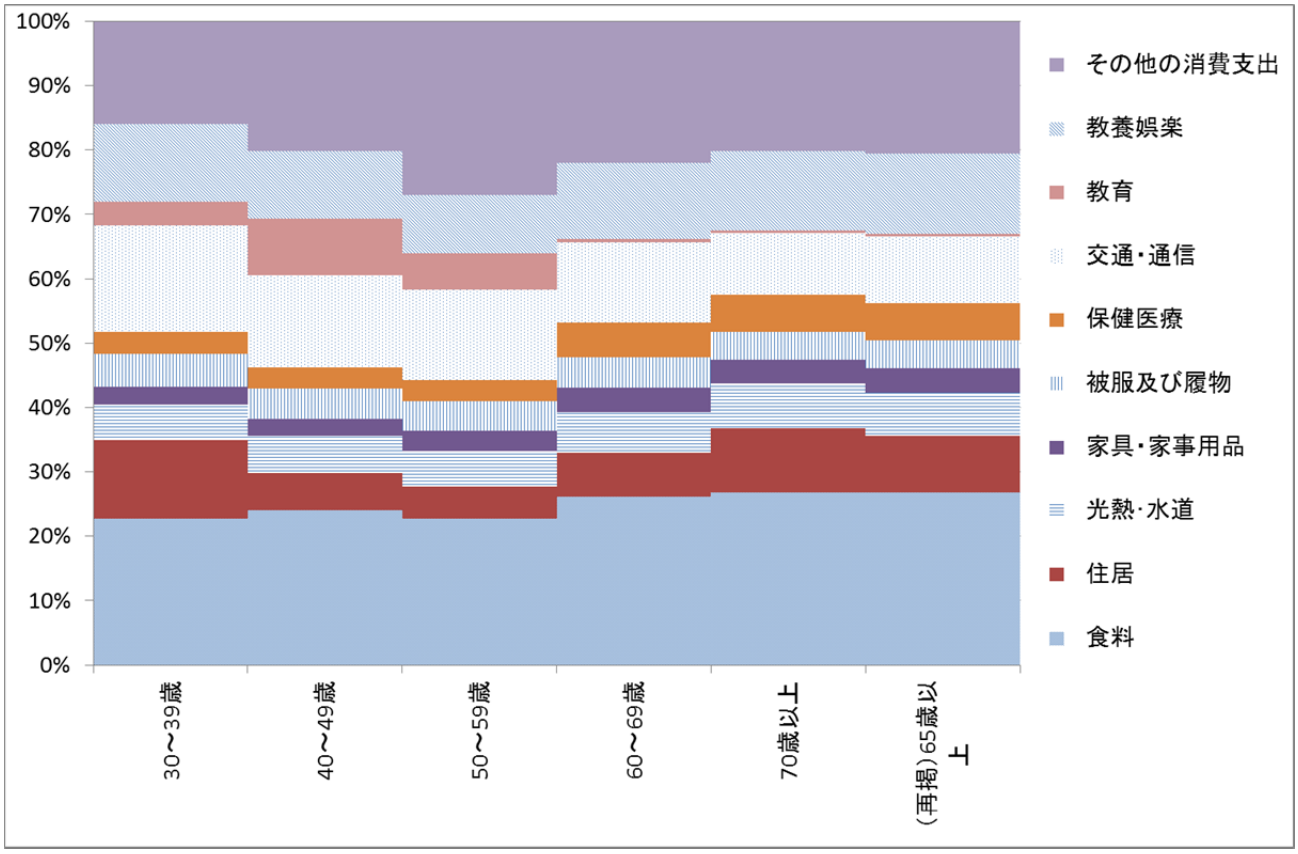
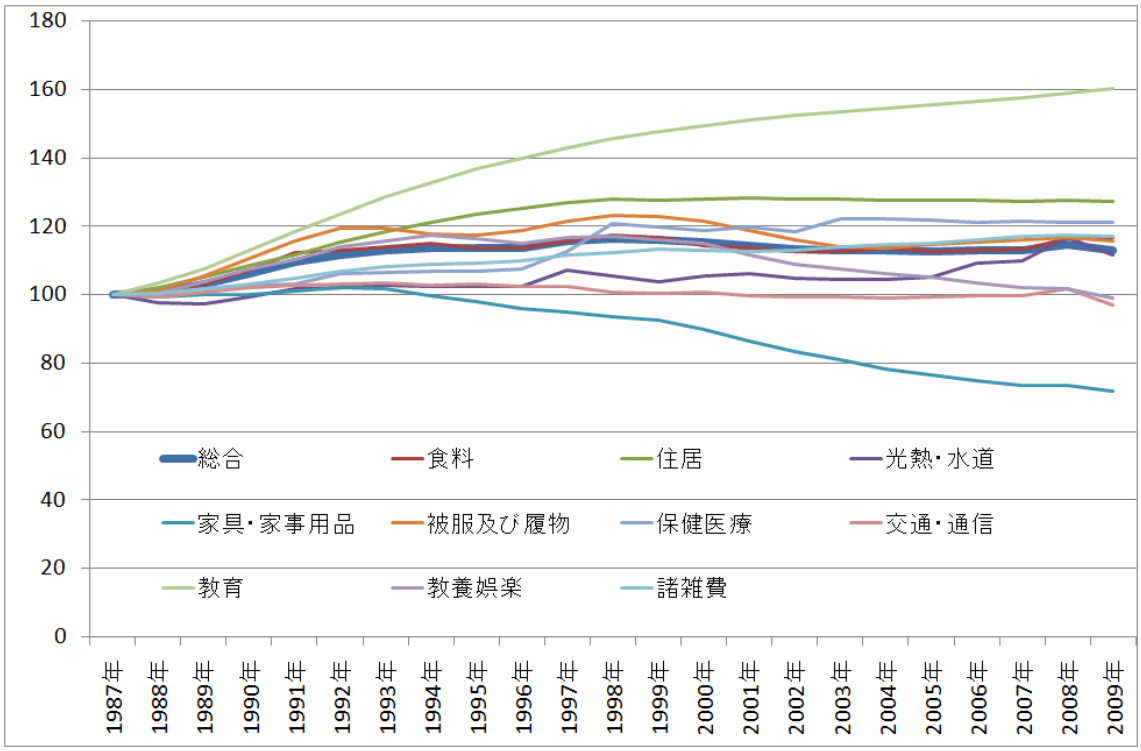


図 2　財別の消費者物価指数の動向(2005 年基準)



(消費者物価指数年報より筆者作成・1987 年=100 に換算)

表１　2007年 費目別業態別価格指数

		一般小売店	スーパー	量販専門店	コンビニ	デパート	生協	その他
食料	穀類	0.970	0.991	0.899	1.132	1.131	0.974	0.968
	魚介類	1.030	0.965	0.925	1.027	1.351	1.000	1.040
	肉類	1.014	0.988	0.947	0.962	1.276	0.978	0.977
	乳卵類	1.020	0.970	0.865	1.083	1.086	0.997	0.993
	野菜・海藻	0.997	0.991	0.830	1.127	1.351	1.029	0.892
	果物	0.984	0.993	0.934	1.182	1.375	1.041	0.837
	油脂・調味料	1.039	0.983	0.826	1.068	1.264	0.979	1.041
	菓子類	1.058	0.927	0.912	1.013	1.260	0.942	1.061
	調理食品	1.017	0.970	0.972	1.063	1.271	0.976	0.981
	飲料	1.013	0.948	0.880	1.149	1.144	0.954	1.092
	酒類	1.014	0.962	0.931	1.061	1.088	0.954	1.033
光熱・水道	ガス代	1.003	1.000	0.982	1.000	1.000	0.985	0.998
家具・家事用品	家庭用耐久財	1.022	0.953	0.974	1.000	1.090	0.984	0.907
	室内装備品	1.158	0.695	0.595	1.000	1.768	0.801	0.841
	寝具類	0.995	0.774	0.661	1.000	2.257	0.854	1.471
	家事雑貨	0.965	0.873	0.862	1.053	1.699	0.850	0.967
	家事用消耗品	0.991	1.005	0.946	1.030	1.104	1.005	1.045
被服および履物	洋服	1.195	0.521	0.605	1.000	1.867	0.633	0.895
	シャツ・セーター類	1.135	0.541	0.495	0.738	2.329	0.600	0.774
	下着類	1.062	0.864	0.760	1.085	1.704	0.889	0.844
	履物	1.248	0.706	0.533	1.000	1.548	0.652	0.605
	他の被服類	1.151	0.975	0.657	0.885	2.039	0.762	0.783
保健医療	医薬品・健康保持用摂取品	1.001	0.988	0.991	1.003	1.063	1.012	1.015
	保健医療用品・器具	0.953	0.821	0.817	1.033	1.265	0.925	1.026
交通・通信	自動車等関係費	1.000	0.996	0.996	1.000	1.002	0.996	0.979
教養娯楽	教養娯楽用耐久財	1.020	1.006	0.976	1.000	1.039	1.020	1.000
	教養娯楽用品	1.011	0.944	0.870	1.059	1.278	0.997	0.981
	教養娯楽サービス	1.109	0.924	1.022	1.020	1.513	0.767	1.058
諸雑費	理美容用品	0.991	1.018	0.970	1.013	1.056	1.020	1.023
	身の回り用品	1.107	0.843	0.799	0.958	1.576	1.178	1.233

表 2 年齢階級別の業態別支出シェア

大分類	中分類		一般小売店	スーパー	量販専門店	コンビニ	デパート	生協	その他
食料	（酒類、外食を除いたすべて）	平均	0.163	0.560	0.025	0.041	0.051	0.091	0.069
		30歳未満	0.128	0.575	0.034	0.134	0.033	0.054	0.042
		30～39歳	0.117	0.597	0.036	0.067	0.036	0.101	0.045
		40～49歳	0.121	0.601	0.030	0.048	0.033	0.118	0.049
		50～59歳	0.153	0.576	0.025	0.030	0.046	0.099	0.071
		60～69歳	0.190	0.538	0.020	0.021	0.064	0.081	0.085
		70歳以上	0.222	0.512	0.017	0.021	0.077	0.068	0.083
		65歳以上	0.213	0.521	0.018	0.020	0.074	0.070	0.085
家事家具用品	家庭用耐久財	平均	0.359	0.057	0.383	0.000	0.047	0.022	0.131
		30歳未満	0.261	0.038	0.524	0.000	0.065	0.008	0.104
		30～39歳	0.190	0.052	0.495	0.000	0.059	0.007	0.198
		40～49歳	0.301	0.058	0.431	0.000	0.031	0.020	0.159
		50～59歳	0.375	0.067	0.391	0.000	0.042	0.025	0.100
		60～69歳	0.379	0.049	0.342	0.000	0.062	0.032	0.135
		70歳以上	0.486	0.063	0.293	0.000	0.034	0.013	0.112
		65歳以上	0.453	0.059	0.319	0.000	0.039	0.026	0.105
教養娯楽	教養娯楽用耐久財	平均	0.289	0.026	0.534	0.000	0.016	0.007	0.127
		30歳未満	0.165	0.028	0.615	0.000	0.017	0.031	0.145
		30～39歳	0.242	0.017	0.507	0.000	0.028	0.002	0.203
		40～49歳	0.263	0.017	0.598	0.000	0.006	0.003	0.113
		50～59歳	0.273	0.031	0.577	0.000	0.014	0.007	0.097
		60～69歳	0.362	0.026	0.487	0.000	0.017	0.001	0.106
		70歳以上	0.434	0.050	0.409	0.000	0.014	0.022	0.070
		65歳以上	0.456	0.047	0.403	0.001	0.022	0.014	0.059

表 3-1 財のウエイトと購買行動を考慮した年齢階級別物価指数

	年齢別・財ウエイト + 年齢別・購入先業態別ウエイト（1987 年=100）							
	平均	30 歳未満	30～39 歳	40～49 歳	50～59 歳	60～69 歳	70 歳以上	65 歳以上
1992 年	108.9	108.7	108.0	108.8	108.8	109.2	109.9	109.7
1997 年	111.0	110.8	109.3	110.7	110.9	111.3	113.0	112.5
2002 年	107.8	108.0	105.9	107.5	107.8	108.0	109.9	109.4
2007 年	108.5	108.4	106.1	107.8	108.7	109.1	111.0	110.5

表 4-1 財ウエイトのみを年齢別にした場合の物価指数の乖離幅

	年齢別・財ウエイト + 全年齢平均・購入先業態別ウエイト （1987 年=100）						
	30 歳未満	30～39 歳	40～49 歳	50～59 歳	60～69 歳	70 歳以上	65 歳以上
1992 年	−0.04	−0.76	−0.07	−0.09	0.18	0.84	0.66
1997 年	−0.35	−1.59	−0.23	−0.04	0.26	1.87	1.40
2002 年	0.02	−1.80	−0.05	0.10	0.08	1.73	1.27
2007 年	−0.35	−2.25	−0.17	0.24	0.35	1.87	1.49

注）財ウエイトを各年齢ごとのものを使った場合の乖離幅(パーセントポイント)

表 4-2 購買行動のみを年齢別にした場合の物価指数の乖離幅

	全年齢平均・財ウエイト + 年齢別・購入先業態別ウエイト （1987 年=100）						
	30 歳未満	30～39 歳	40～49 歳	50～59 歳	60～69 歳	70 歳以上	65 歳以上
1992 年	−0.10	−0.11	−0.06	0.00	0.10	0.19	0.18
1997 年	0.14	−0.11	−0.10	−0.03	0.02	0.15	0.13
2002 年	0.10	−0.10	−0.25	−0.07	0.11	0.36	0.27
2007 年	0.26	−0.17	−0.54	−0.08	0.20	0.64	0.49

注）購入先業態別ウエイトを各年齢ごとのものを使った場合の乖離幅(パーセントポイント)

高齢者世帯の消費行動と物価指数

宇南山　卓（財務省財務総合政策研究所）
慶田　昌之（立正大学）

年齢別の物価の比較

- 年齢別の物価指数の構築:
  - 年齢別の消費する財の違い
  - 年齢別の購入先の違い
    - 店舗形態による価格の違い
    - 年齢による購入先店舗形態の違い

    ＝　全国消費実態調査・全国物価統計調査で店舗形態別の物価が観察可能
- 年齢別物価指数の重要性
  - 公的年金の実質化に不可欠
  - 年齢別の経済厚生の計測
  - 学術的な関心
    - Amble and Stewart (1994)：高齢者ほど高い物価上昇率に直面
    - Goda, Shoven, and Slavov (2011) 医療費について高齢者ほど物価上昇率が高い

3

真の物価指数と定義と計測

- 家計の支出最小化問題

$$X = e(p, \bar{u}; \delta)$$

  - X: 家計支出
  - p: 価格ベクトル
  - : 効用水準
  - δ: “taste” パラメータ

- 真の物価指数の定義

$$P_1 = \frac{e(p_1, \bar{u}; \delta)}{e(p_0, \bar{u}; \delta)}$$

  - 物価指数とは、異なる物価ベクトルの下で、同じ効用水準を得るのに必要な支出額の比率
    - 物価指数には必ず「基準」が存在する
  - 物価指数の性質については、例えば Diewert (1981)を参照

4

ここでやっていないこと: 選好と物価

- 概念的には「同一価格ベクトルのもとで、異なる選好を持つ家計が、同じ効用水準を得るのに必要な支出額」によって直面する物価の比較が可能であるように見える

$$\mathrm{R} = \frac{e(p, \bar{u}; \delta_1)}{e(p, \bar{u}; \delta_0)}$$

  - 選好が異なるため、同じ価格ベクトル（すべての財の価格が同一）であっても、直面する「物価」が異なる可能性がある
    - たとえば、高齢者と若者では、同じ支出でどちらがより高い効用が得られるか？
- しかし、この比較は理論的にも実証的にも不可能
  - 理論的には、序数的効用の仮定より個人間での効用水準の比較は不可能
  - （基数的効用を仮定したとしても）実証的に との識別は不可能
    - 消費行動が異なっても、その理由が「選好が異なる」からなのか「効用水準が異なる」からなのかは区別できない

5

選好とインフレ率

- 物価水準ではなく、物価の上昇率であれば、選好が異なる個人間で比較が可能

$$P_1^k \equiv \frac{e(p_1,\bar{u};\delta_k)}{e(p_0,\bar{u};\delta_k)}$$

  - 選好のパラメータは固定
  - もし$P_1^k > P_1^l$ であれば
    - 選好が“k”であるような家計は、選好が“l”である家計より直面するインフレ率が低い
    - 選好が“k”であるような家計が高い物価水準に直面しているとは限らない

6

年齢別の物価指数の構築

- 構築されるべき真の物価指数は

$$P_1^k \equiv \frac{e(p_1,\bar{u};\delta_k)}{e(p_0,\bar{u};\delta_k)}$$

  - 需要システムの推計が必要
- 実務的に標準的な物価指数の作成方法
  - Laspeyres 指数の構築
  - 真の物価指数の上限となる

$$CPI_1^k \equiv \frac{p_1 \cdot q_0^k}{p_0 \cdot q_0^k} \geq \frac{e(p_1,\bar{u};\delta_k)}{e(p_0,\bar{u};\delta_k)}$$

7

ここでやったこと

- 以下の物価指数を構築:

$$CPI_t^k \equiv \sum_i w_{i0}^k \left(\sum_s w_{is0}^k \frac{p_{ist}}{p_{is0}} \right)$$

- 価格データ
  - 小売物価統計調査
    - 全業態の平均的な価格の動向
    - CPIの元データ
    - 月次統計
  - 全国物価統計調査
    - **業態間の価格差をとらえることのできる統計**
    - 5年に一度 (1987, 1992, 1997, 2002, and 2007→廃止)
- ウエイトデータ
  - 全国消費実態調査
    - 財別・業態別・年齢別の支出額が分かる
  - 5年に一度（ここでは2004年のデータを利用）

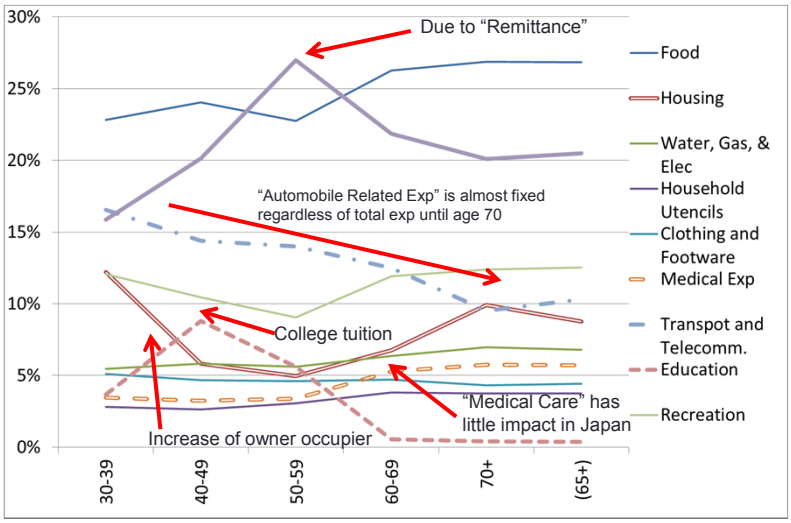
8

年齢階層別の「嗜好」の違い

- 高齢者と若年者では消費行動が大きく異る.
  - 高齢者は若年者とは異なる財を購入する.
    - 価格は財によって異なる動きをする
    - 通常は「選好」とは個々の違いを指す
      - 総務省統計局も「年齢階級別」の物価指数を公表している
  - 高齢者は若年者と異なる店舗で購入する.
    - 店舗によって価格の動きは異なる.
    - **この論文の貢献!**
- 年齢別・財別・購入先業態別の支出額・価格動向を把握
  - 日本のデータであれば可能

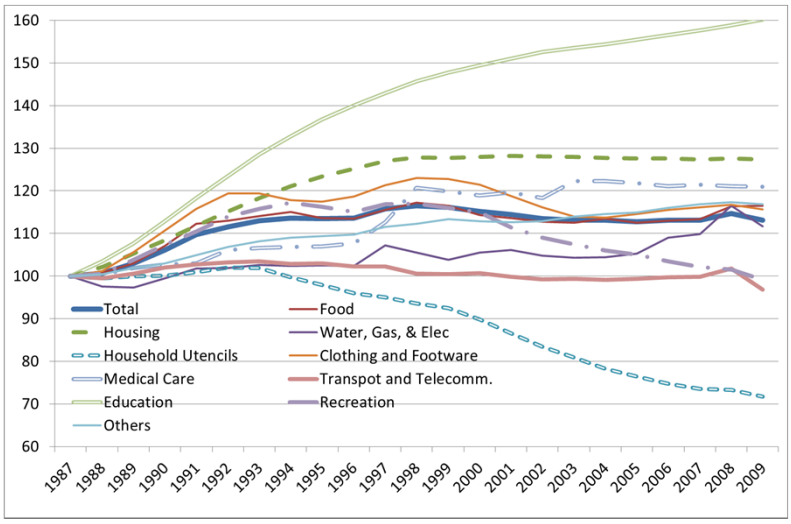
9

年齢による消費パターンの違い



10

財別のインフレ率



11

購入先業態別に関する選好

- 業態によって価格水準が異なる
  - Scannerデータを用いた検証(for example, Abe and Tonogi, 2007)
  - ただし小売店舗全体に対する代表性はない
- 異なる業態は異なるサービスを提供.
  - 一般小売店(Mom-Pop store):カスタマイズされたサービス
  - スーパー: 食料品については安価で画一的なサービス
  - ディスカウントストア: 特定の品目が安価だが市街立地も多い
  - 百貨店: 多種多様な商品を高い品質で提供
  - コンビニ: 営業時間が長いが一般に割高
  - 生協: メンバーシップが必要だが比較的安価で宅配などもある
- 消費者は価格とサービスを比較して購入先を選択
  - 基本的には異なる業態での商品は「異なる財」と考えるべき

12

年齢による購入先業態の違い

They sometime offer delivery service

		Retail Store	Supermarket	Discount Store	Convenience Store	Department Store	Cooperative Store	Others
Food (Excl. "Alcohol" and "Eating Out")	Ave	16%	56%	3%	4%	5%	9%	7%
	30-	13%	58%	3%	13%	3%	5%	4%
	30-39	12%	60%	4%	7%	4%	10%	5%
	40-49	12%	60%	3%	5%	3%	12%	5%
	50-59	15%	58%	3%	3%	5%	10%	7%
	60-69	19%	54%	2%	2%	6%	8%	8%
	70+	22%	51%	2%	2%	8%	7%	8%
	(65+)	21%	52%	2%	2%	7%	7%	8%
Recreational Durable Goods	Ave	29%	3%	53%	0%	2%	1%	13%
	30-	17%	3%	62%	0%	2%	3%	15%
	30-39	24%	2%	51%	0%	3%	0%	20%
	40-49	26%	2%	60%	0%	1%	0%	11%
	50-59	27%	3%	58%	0%	1%	1%	10%
	60-69	36%	3%	49%	0%	2%	0%	11%
	70+	43%	5%	41%	0%	1%	2%	7%
	(65+)	46%	5%	40%	0%	2%	1%	6%

Don't like a "mass sales" outlet?

Prefer more "traditional" or customized outlet?

Online Shopping?

13

業態による価格水準の違い

- 業態による価格水準の違い

	Retail Store	Supermarket	Discount Store	Convenience Store	Department Store	Cooperative Store	Others
Meat	101.4	98.8	94.7	96.2	127.6	97.8	97.7
Snacks	105.8	92.7	91.2	101.3	126.0	94.2	106.1
Kitchen Items	99.1	100.5	94.6	103.0	110.4	100.5	104.5
Household Appliances	102.2	95.3	97.4	NA	109.0	98.4	90.7
Recreational Durables	102.0	100.6	97.6	NA	103.9	102.0	NA
Shampoo and Cosmetics	99.1	101.8	97.0	101.3	105.6	102.0	NA

"Cheapest" outlet　　"Most expensive" outlet

  - ここでは「固定ウエイト」指数を構築しており、業態ごとのサービスに変化がなければ、サービスが異なること自体は問題にならない
  - 業態による価格水準ではなく「価格動向」が重要

14

業態によるインフレ率の違い

- どの業態が最も価格が上昇したか?

$$P_{st} \equiv \sum_i \overline{w}_{i0} \frac{p_{ist}}{p_{is0}}$$

	1992	1997	2002	2007
Ave.	108.9	111.0	107.8	108.5
Retail Store	110.6	111.7	108.4	109.5
Supermarket	110.0	111.2	106.0	105.0
Discount Store	105.1	108.3	104.4	100.8
Convenience Store	107.5	112.3	108.6	108.3
Department Store	110.5	113.0	112.7	119.8
Coop	104.5	109.4	107.2	104.9
Others	108.0	107.5	104.3	102.1

15

メインの結果：年齢別の物価指数

	Age Specific Goods-Outlet Weights（1987=100）							
	Ave	30-	30-39	40-49	50-59	60-69	70+	(65+)
1992	108.9	108.7	108.0	108.8	108.8	109.2	109.9	109.7
1997	111.0	110.8	109.3	110.7	110.9	111.3	113.0	112.5
2002	107.8	108.0	105.9	107.5	107.8	108.0	109.9	109.4
2007	108.5	108.4	106.1	107.8	108.7	109.1	111.0	110.5

2 percent point more
(0.2 percent point per year between 1987-97)

The older experienced slightly smaller deflation

16

直面するインフレ率の差の要因分解

- 構築される指数

$$CPI_t^k \equiv \sum_i w_{i0}^k \left(\sum_s w_{is0}^k \frac{p_{ist}}{p_{is0}} \right)$$

- 消費される財だけが異なる場合の指数

$$\widehat{CPI}_t^k \equiv \sum_i w_{i0}^k \left(\sum_s \bar{w}_{is0} \frac{p_{ist}}{p_{is0}} \right)$$

  - 購入する業態は年齢によって差がないケース

- 購入する業態だけが異なる場合の指数

$$\widetilde{CPI}_t^k \equiv \sum_i \bar{w}_{i0} \left(\sum_s w_{is0}^k \frac{p_{ist}}{p_{is0}} \right)$$

  - 消費する財は年齢によって差がないケース

17

要因分解の結果

Differences are mainly caused by consumption pattern

	Age Specific Goods Weight (Deviation from the Average)						
	30歳未満	30～39歳	40～49歳	50～59歳	60～69歳	70歳以上	65歳以上
1992年	-0.04	-0.76	-0.07	-0.09	0.18	0.84	0.66
1997年	-0.35	-1.59	-0.23	-0.04	0.26	1.87	1.40
2002年	0.02	-1.80	-0.05	0.10	0.08	1.73	1.27
2007年	-0.35	-2.25	-0.17	0.24	0.35	1.87	1.49

	Age Specific Outlet Weight (Deviation from the Average)						
	30歳未満	30～39歳	40～49歳	50～59歳	60～69歳	70歳以上	65歳以上
1992年	-0.10	-0.11	-0.06	0.00	0.10	0.19	0.18
1997年	0.14	-0.11	-0.10	-0.03	0.02	0.15	0.13
2002年	0.10	-0.10	-0.25	-0.07	0.11	0.36	0.27
2007年	0.26	-0.17	-0.54	-0.08	0.20	0.64	0.49

The outlet that the older prefer experienced higher inflation rate

18

結論

- 高齢世帯ほど直面したインフレ率が高かった
  - 1987-97の年平均で0.2 パーセントポイントほど公式のCPIよりも高い
- 乖離の75%程度は消費する品目の違いで説明できる
  - 高齢者は「住居」および「保健医療」に多く支出をしている
  - ただし、これは「バイアス」ではなく経済構造の変化による結果
- 高齢者が利用する購入先形態はより大きく物価が上昇した
  - 一般小売店や百貨店などが相対的に物価上昇率が高かった
  - 若者の支出が多いコンビニは相対的に物価上昇率は低かった

補足

- Boskin et al. (1996)が 指摘するCPIのバイアスについては考慮していない
  - Laspeyres 指数は真の物価上昇率を過大評価
  - バイアスの大きさが年齢によって異なる可能性はある
- 教育費と保健医療は、物価上昇率が高い
  - 年齢によって支出シェアが大きく異る財でもある
  - 品質調整が困難な財でもある
  - 品質調整の影響は年齢よって異なる

エビデンスに基づいた匿名化

星野　伸明*

平成 26 年 1 月 21 日

Evidence Based Anonymization

Nobuaki Hoshino*

概 要

匿名データや個人情報は、個体識別が可能か否かで法律上区別される。しかしこの区別の方法は不明確で、改善のための明示的議論の対象になっていない。従って本論文は、個体識別可能性の判定方法を明確化する。このような判定に関する既存研究は、個体識別可能性の定量評価について閾値を定める理論を欠く。この点について本論文では、個体識別が起きていないという観測可能な事実に基づいて閾値を推定する。また部分的にしか観測されず定量評価できない情報も、事例間で等しいか否かという判断しやすい方法で利用する。このような観測に基づいて意思決定する態度は、エビデンスに基づいた匿名化と呼ぶのがふさわしい。この立場から、匿名データ審査体制の改善点が指摘できる。

Japan Law discriminates Anonymized Data or personal information by discerning that a specific individual is identifiable. The state of being identifiable, however, is not defined, and thus we can not explicitly improve the evaluation of identifiability. Therefore the present article explicates a method to decide whether given data are identifiable or not. The existing literature on this issue lacks the theory of deciding the critical value of measured identifiability; we estimate it based on a fact that identification has not been observed. Also partially observed factors are compared in our decision, which is called evidence based anonymization. This theory leads to institutional improvements on Anonymized Data.

キーワード: 母集団一意, プライバシー, 統計的開示制限.

*金沢大学経済学類, 〒 920-0927, 石川県金沢市角間町, E-mail: hoshino@kenroku.kanazawa-u.ac.jp

1　はじめに

匿名データは、平成 21 年度に四調査（全国消費実態調査、社会生活基本調査、 就業構造基本調査、 住宅・土地統計調査）から提供が開始された。平成 23 年度末現在、国民生活基礎調査や労働力調査の匿名データ提供も決まっている。新しい制度がこのように実績を重ねてきたことは喜ばしい。ただ今後は実績という経験を活かし、制度を継続的に改善する道筋をつけるべきである。特に利用者からのデータ改善要求に応える必要があろう。

匿名データは元の個票を変換（匿名化）して作られる。例えば全国消費実態調査等の匿名データでは、15 歳から 84 歳までの年齢を 5 歳階級別に変換している。また地域情報は「3 大都市圏」及び「その他の地域」の 2 区分に変換している。このような変換により、各歳別の分析や詳細な地域別分析は不可能となる。データ分析において、匿名化は明らかに望ましくない。故に匿名化の緩和は利用者の典型的な要求である。

しかし全ての匿名化を外せるわけではない。匿名データの定義（統計法第 2 条第 12 項）を引用すると、「一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む。）ができないように加工したもの」である。元の個票（調査票情報）はこの定義を満たすように匿名化されなければ、匿名データとして提供不可能[1]である。従って匿名化は、個体識別が不可能な範囲で少ない方がよい。つまり匿名データの改善の多くは、個体識別が可能か否かという判断を必要とする。

この判断について、総務省政策統括官（統計基準担当）(2011) による「匿名データの作成・提供に係るガイドライン」（以下、ガイドライン）には、審査用資料として「チェックリスト」を作成することが定められている。そして「チェックリストに記載された内容等を基に」、「匿名化処理の妥当性等に係る審査を実施する」とある。参考として世帯調査のチェックリスト（H23/3/28 改正版）の要約を付録 A に収めた。チェックリストは個体識別に関係する要因を記載しているはずである。しかしその使い方は説明されていない。

結局「一律に匿名化の基準を設定することは困難」なので「一橋大学における匿名標本データの試行的提供の事例[2]及び諸外国の統計機関における同様の提供の事例等を参考に」匿名化せよとガイドラインは書く。同様とはどのような事例で、それをいかに参考にしたらよいか。この点についての判断は審査担当者の見識に委ねられている。個体識別可能と不可能の区別は、不明確である。

この区別の明確化、精密化は匿名データに関してだけの課題ではない。いわゆる個人情報保護法において個人情報は「生存する個人に関する情報であって、当該情報に含まれる氏名、生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ、それにより特定の個人を識別することができることとなるものを含む。）」と定義される（第 2 条）。このように個体識別が可能か否かを区分の基準とする例は外国法[3]でも見られる。いか

[1]本論文において統計法改正は手段として除外する。しかし個体識別性をデータ提供の基準とするのは必ずしも望ましくない。例えば個体識別されてもデータが悪用されなければよいという主張は妥当かもしれない。

[2]試行的提供の詳細については山口 (2008) を見よ。

[3]例えば U.S. Privacy Act など。U.S. Office of Federal Statistical Policy and Standards (1978, pp.3–5) の解説を見よ。

に匿名化すれば個体識別が不可能かという問題は普遍的である。

ところがこの基本的な問題がないがしろにされている。匿名化についての多くの研究は個体識別の危険性（開示リスク）の測り方は定める。しかし開示リスクの目標値について、せいぜい危険選好に応じてデータの分析価値（有用性）とバランスさせよ[4] としか言わない。つまり現実の匿名化が許容範囲か否かは、実務家の価値判断の問題とされる。このような態度は匿名化の技術開発には好都合である。しかし結果として実務家は途方に暮れ、判断を何らかの権威に投げているのが現状であろう。

現行法の下での匿名化の実務的判断について、客観化の余地が残っていると著者は考える。本論文では諸要因と個体識別行為の関係をモデル化し、個体識別が可能という状態を要因と関係づける。このようなモデル化は、開示リスクを具体的に定めることに他ならない。従って、評価される開示リスクの目標値が問題となる。この点について本論文は、個体識別が不可能な状態を過去の事例を基に決める方法を提案する。過去に公開されたデータについて個体識別が観測されていないとすれば、その事実は個体識別不可能ということについて情報を持っている。故に個体識別が観測されることの確率モデルを構成し、既公開の匿名データ等を匿名化の程度判断についての統計的証拠に転ずる。

このような理論なくして、明確な個体識別性の審査はあり得ない。そもそもチェックリストの記載事項は、個体識別と理論的に関係する要因であるべきだ。またリストの使い方も理論が定める。それから本論文の理論は観測結果と関係を持ち、実証の対象である。いかなる理論も実証を経ることで継続的に改善される。従って本論文は、個体識別に関する統計的証拠—エビデンスに基づいた匿名化 (Evidence Based Anonymization, EBA) を主張する。

本論文の構成は以下の通りである。2 章は全体で、個体識別が可能か否かの判定方式を明らかにする。まず 2.1 節において、個体識別の観測と可能性の関係を確率モデルで表す。次に 2.2 節では、個体識別行為を所与の要因についての確率モデルで表す。2.3 節では、個体識別の要因について計量可能な方式を考察する。ここまでの議論で、情報不足により定性評価で満足せざるをえない個体識別の要因があることになる。2.4 節ではそのような要因を明らかにする。2.5 節では、個体識別の行為と観測の関係について考察する。3 章では 2 章の理論を基に、匿名データ審査体制の改善点を指摘する。

2　個体識別可能性の判定方式

2.1　個体識別の観測モデル

個体識別が可能か否かは、明らかにデータの表現に依存する。ここで匿名化による表現の変化は滑らかだが、個体識別が不可能と可能の差は不連続である。これをモデル化する場合、データ表現の適当な実数特性値が閾値を超えれば個体識別が可能とみなすのが定石であろう。本論文で

[4]例えば Duncan et al. (2001) や Domingo-Ferrer and Torra (2001) など。伊藤 (2012) のサーベイを参照せよ。

もこのように考え、個体識別の難易度とみなせる特性値に注目する。この難易度が閾値より高ければ個体識別が不可能とみなすのは自然[5] である。

具体的な難易度測度の設計については後で考察するとして、モデル化をすすめよう。難易度 δ を引数とする関数 f は個体識別が可能なら 1、不可能なら 0 を返すとする。つまり閾値が α として

$$f(\delta) = \begin{cases} 1 & \delta < \alpha \text{ の場合} \\ 0 & \delta \geq \alpha \text{ の場合} \end{cases} \tag{1}$$

ということになる。もし所与のファイルが個体識別可能か否か判定したいなら、その難易度を求めて α と大小を比較すればよい。ただ α が既知となるような難易度測度の設計は難しい。後述されるように、事実上観測可能でない個体識別の要因が残るはずだ。では α が未知の場合にどうするべきか。

まず α が推定可能か考えてみよう。統計的に未知母数 α を推定するには、観測値が必要になる。しかし個体識別が可能か否かは、観測されることではない。観測可能な事実は、個体識別が起きたか否かである。モデルを用いて説明しよう。確率変数 X が 1 なら個体識別が観測され、0 なら観測されないこととする。個体識別が不可能なら必ず $X = 0$ である。個体識別が可能の場合、難易度 δ に依存する確率 $p(\delta)$ で識別が観測されると考えよう。すなわち $\Pr(X = 1; \delta < \alpha) = p(\delta), \Pr(X = 0; \delta < \alpha) = 1 - p(\delta)$ とする。個体識別は決して起きないと考えれば危機管理にならないので、$p(\delta)$ は正と想定するべきだ。

このような状況で閾値が共通する n 件の事例が存在するとしよう。$i, i = 1, 2, \ldots, n,$ 番目について観測されるのは、少なくとも難易度 δ_i と識別の有無 x_i である。単純化のため $\delta_1 < \delta_2 < \cdots < \delta_n$ としよう。この中で個体識別が観測された（$x_i = 1$ となる i が存在する）場合は $\delta_i < \alpha$ と分かる上、実務的に重要でない。故に個体識別がこれまで[6] 起きていない（全ての i について $x_i = 0$）として考察を続ける。この場合モデルの尤度 ℓ は、$\delta_i < \alpha \leq \delta_{i+1}$ の時 $\ell(\alpha) = \prod_{j=1}^{i}(1 - p(\delta_j))$ となる。そして全ての δ について $0 < p(\delta) < 1$ なら、α の最尤推定値 $\hat{\alpha}$ は δ_1 以下である。つまり過去の事例で個体識別が観測されていなければ、その最も低い難易度 δ_1 以下と閾値 α は推定される。このように $\hat{\alpha}$ が一意に定まらないのは、情報が無いので区別出来ないことを意味する。

情報を補うため α の事前分布を用いることは考えられるが、それよりも観測情報の増加を工夫する方が健全だろう。つまり X を $\{0, 1\}$ の二値とするのではなく、α と δ の距離に依存する連続量と出来ればよい。これは治験薬の用量反応関係の推測と考え方が同じである。薬の臨床試験では、人体に決定的な悪影響を及ぼしてはならない。このような制約下では、薬剤の投与を少量から始めて徐々に増やし、危険な兆候が見られれば中止する。生死の二値ではなく、投与量が死亡の閾値と近いことを示す兆候（心拍や呼吸の異常等）を観測するのである。

[5] 一般目的汎用ファイル (Public Use File, PUF) の作成においても整合的な考え方である。星野 (2010) の考察の通り、PUF は匿名データよりも強い匿名性を必要とする。しかし統計法には個体識別が可能か不可能かという二区分しか存在せず、PUF と匿名データを区別できない。ところが個体識別の難易度という概念を用いれば、個体識別が不可能という状態の中で匿名データと PUF を区別出来る。

[6] データの公開直後に識別が起きなくても、ある程度後で識別が起きることはあり得る。閾値の推定をする時点に依存して各 $p(\delta_i)$ は変化するかもしれない。しかし $0 < p(\delta_i) < 1$ なら $\hat{\alpha} \leq \delta_1$ という結論は変わらない。

問題は、α と δ の距離に依存する観測可能な事象として何を用いるかである。例えば個体を識別できたと誰かが誤って主張することなどが、個体識別発生の兆候として考えられる。この場合、個体識別を試す気にさえならない水準よりは難易度が下がっていることがわかる。一般に 1 件の重大事故の陰には 300 件のヒヤリ・ハットが起きているという。匿名データ提供に対する社会的反応が、警鐘になる可能性はある。また攻撃者の動機を考えることで、ある程度 $p(\delta)$ を定めることができるかもしれない。2.5 節でそのような考察を部分的に行うが、実務への反映は拙速と思われる。

では現時点で、新しく公開するファイルの難易度 δ_{n+1} をどのように決めたらよいか。一つの考え方は $\delta_{n+1} = \delta_1$ とすることだろう。強い仮定を置かずに推測出来るのは、難易度の閾値が δ_1 以下ということまでである。これは δ_1 未満の難易度について個体識別が不可能な証拠がないということだ。個体識別が可能になるという過誤の可能性[7]を考えれば、慎重な判断は正当化されるだろう。また $\delta_{n+1} = \delta_1$ として難易度 δ_1 における観測が蓄積されることは、将来的に意味を持つ。同難易度の複数のファイルについて個体識別が観測されなければ、その難易度が個体識別不可能な確率は高まる。また観測情報が増えないと、閾値との近さについて確かな判断は出来ない。

このように個体識別の難易度という概念を用いれば、個体識別が可能か否かの判断において過去の事例を統合して利用できる。しかし例えば国や公開時期が違う事例において、個体識別が可能となる難易度の閾値 α は同じだろうか。

閾値 α が共通する事例の範囲は δ の具体型に依存する。個体識別に関する要因を全て勘定する理想的な δ を用いる場合、全事例で閾値が共通するとみなしてよい。逆に閾値の変動は、δ が考慮しない要因の変化から生ずる。データ表現の実数特性値として導入した δ だが、その他の要因を算出に用いるべきかもしれない。以下では良い δ の構成を考察しよう。

2.2 個体識別行為の確率モデル

前節では個体識別の難易度に依存して個体識別が確率的に観測されるモデルを考察した。具体的に個体識別の難易度を定めるには、個体識別行為をモデル化する必要がある。本節ではそのようなモデルを構成し、個体識別の難易度の定式化をすすめる。

個体識別行為の確率モデルに関する先行研究としては、英国国勢調査匿名化標本の開示リスクを評価した Marsh et al. (1991)、及びこの論文を再考した Dale and Elliot (2001) が挙げられる。Marsh 等は個体識別が起きる条件を具体的に挙げ、それらが満たされる確率を個別に評価することで、個体識別が起きる確率を計算しようとした。まずこの試みを検討しよう。

最もあり得る個体識別の形態は、識別を試みる者（「攻撃者」）が素性を知る個体を（匿名化された）公開ファイルの中に見つけることだと言われている。このような行為を竹村 (1997) は「順攻撃」と呼ぶ。一方、公開ファイル中の特定個体を母集団に探す行為は「逆攻撃」と呼ばれる。この区別が意味を持つのは、攻撃者が探して素性を知ることが出来る個体群が、（攻撃用ファイルの）

[7]個体識別が可能であるにも関わらず識別が起きない ($p(\delta) < 1$) と過誤が生ずる。故に過誤の可能性を減らすには、δ_1 の例だけでも攻撃実験をするなどして識別が可能か確認することも役立つ。2.5 節の議論も参照のこと。

素性を知る個体群と異なる場合である。これがどのような場合に異なるかは後で考察するとして、順攻撃の考察を続ける。

既知の個体について攻撃者が知る属性（「キー変数」）を並べたファイルを「攻撃用ファイル」と呼べば、「攻撃」とは公開ファイルと攻撃用ファイルでキー変数が同じレコードを探すことと言える。しかしそのような個体が見つかったとしても、それは母集団に複数存在する属性が同じ個体のうちの一でしかないかもしれない。故に統計当局は、母集団に一しか存在しない個体（「母集団一意[8]」）を公開ファイルの中に攻撃者が発見することを警戒しなければならないと言われている。

このような背景の下、Marsh 等は個体識別が実際に起きる確率を以下のように分解する。

$$\Pr(\text{識別が実際に起きる}) = \Pr(\text{識別が起きる} \mid \text{識別を試みる}) \Pr(\text{識別を試みる}) \tag{2}$$

更に識別を試みた時にそれが成功する事態は、以下の 4 つの条件が成立する場合だという。

(a) 攻撃用ファイルと公開ファイルのキー変数が同じ（時点や分類の）基準で記録されている。

(b) 公開ファイルに個体が含まれている。

(c) 個体が母集団一意である。

(d) 個体が母集団一意と確証出来る。

これらの条件が満たされる事象をそれぞれ a から d と書けば

$$\Pr(\text{識別が起きる} \mid \text{識別を試みる}) = \Pr(a) \Pr(b|a) \Pr(c|a,b) \Pr(d|a,b,c) \tag{3}$$

ということになる。右辺の確率を個別に評価できれば、左辺の確率が求められる。

このような分解により、個体識別という漠然とした行為は直感的に解釈できる事象の積となる。(3) 式の分解で鍵となる母集団一意概念は、この種の議論では珍しく非専門家でも理解が可能であり、よく知られている。

ただし母集団一意は特殊な匿名化[9] において無意味な場合がある。また普通の匿名化手法だけ用いるとしても、個体識別が論理的に可能なのは母集団一意に限らない。母集団二意の個体も、自レコードが分かるなら、もう片方の個体のレコードが識別できる。そして三意以下でも、個体間で結託すれば識別できる。しかし母集団一意数と二意以下の珍しい個体数は経験的に比例する。故に母集団一意数を管理すれば、二意以下の識別可能性も同時におさえられる。母集団一意は実際の値に意味があるというより、管理対象のリスク測度として分かり易い点が望ましい。

また母集団一意数は、匿名化の程度についてある種の単調性を持つ。主に使われる匿名化手法は「再符号化[10]」と呼ばれ、個体属性を粗く分類する。再符号化で分類を併合してより粗い分類に

[8] 公開ファイル中で所与のキー変数の組み合わせの条件を満たす個体数が 1 の場合、そのような個体は「標本一意」と呼ばれる。標本一意でも母集団一意とは限らないが、母集団一意なら標本一意である。

[9] Hoshino (2009, Section 6) で議論したとおり、匿名化された表現が互いに排他的でないと母集団一意は一意にならない。通常用いられる大域的再符号化なら、表現は互いに排他的となる。

[10] トップコーディングや削除 (suppression) も再符号化の特殊ケースである。

変換する場合、母集団一意数は非増加である。つまり匿名化が施されて直感的により安全なデータの母集団一意数は、匿名化が施される前のデータの母集団一意数より多いことはない。EBA では匿名化事例に順序をつけるので、このように直感と矛盾しない順序を得られる方法が重要である。

母集団一意数は使用に異論[11] もあるが、以上の議論のような望ましい性質を持つ。Marsh 等の分解を活かして個体識別の難易度を構成出来ないだろうか。

問題は、Marsh 等が分解した要因 (Pr(識別を試みる), $\Pr(a)$,$\Pr(b|a)$,$\Pr(c|a,b)$,$\Pr(d|a,b,c)$) はそれぞれ評価できるとは限らないことである。まず Pr(識別を試みる) の評価は難しいと Marsh 等も認めており、定性的に議論[12] した上で「識別を試みた例を知らないので識別を試みる確率の最良の推定値は経験からゼロ」と述べている。また $\Pr(d|a,b,c)$ についても分からないので、「非常に多くのキー変数について事前情報が無いはずなのでゼロと信ずるが 0.001 と仮定」している。これでは数値評価が出来ているとは言えない。これらの確率評価は出来るとしても膨大な情報を必要とする。現実的に Marsh 等の方法では、リスク測度 (2) と (3) のいずれも数値評価できない要因を抱える。

もちろん Pr(識別が起きる | 識別を試みる) の条件付き確率の積による分解は一意ではない。故にこれを全て評価できる要因の積に書ければ、評価出来ない問題は解決する。しかしそのような分解は不可能であろう。最大の困難は識別が観測されないことである。個票の公開で先行する海外でも識別が起きないように匿名化しているので、ほとんど観測されない事象[13]の確率の推定を強いられる。それにも関わらず、個体識別という事象を細かい要因の積に分解すれば、要因毎に十分な観測数が得られるはずがない。目的の事象の観測が限られる以上、要素の分解に依存したアプローチは、どこかで情報不足の壁に阻まれるであろう。個体識別について数値評価出来ない要因の存在を前提とするべきである。

実は数値評価する要因 $\boldsymbol{y}_1$ と評価しない要因 $\boldsymbol{y}_2$ が分かれても、ある程度は個体識別の難易度を相対比較できる。数値評価した結果 $g(\boldsymbol{y}_1)$ について、個体識別の難易度 δ が $h(g(\boldsymbol{y}_1),\boldsymbol{y}_2)$ と書けるとしよう。ここで $\boldsymbol{y}_2$ は数値評価できないので、難易度関数 h の具体型は分からない。しかし h は以下のような単調性を持つとする。

$$g(\boldsymbol{y}_1') \geq g(\boldsymbol{y}_1) \Rightarrow h(g(\boldsymbol{y}_1'),\boldsymbol{y}_2) \geq h(g(\boldsymbol{y}_1),\boldsymbol{y}_2) \tag{4}$$

つまり事例 $(\boldsymbol{y}_1',\boldsymbol{y}_2)$ と $(\boldsymbol{y}_1,\boldsymbol{y}_2)$ では、数値評価部が低いほうが難易度が低いということである。単調性 (4) さえ成り立てば、$\boldsymbol{y}_2$ が共通する複数の事例から、最も個体識別の難易度が低いものを選べる。そして 1 節のように $n+1$ 番目の新しい匿名データを公開するとして、数値評価値 g を $\boldsymbol{y}_2$ が共通する過去最低の事例に合わせればよい。このように g の達成目標値は $\boldsymbol{y}_2$ に依存して決まる。

[11] 例えば同じ母集団一意でも、似た個体が居ない方が目立って識別し易いだろう。故に母集団一意のレコードの中で、似た属性の個体が多いか少ないかで開示リスクを変える考え方を「レコードレベルリスク」と呼ぶ。例えばより低次元の周辺分割表で一意になる個体の方が危険とみなす “Special Unique” は比較的計算しやすい (Elliot et al., 1998)。このような議論は一理あるが、リスク管理の対象として複雑な測度は望ましくない。また匿名化の程度についての単調性が崩れるかもしれない。

[12] 曰く Pr(識別が起きる | 識別を試みる) が減少すれば Pr(識別を試みる) も減る。またデータの観測と公開の時点が離れれば Pr(識別を試みる) も減る、等。

[13] 開示制限を失敗して個体識別が可能な例は Sweeney (2002) が報告している。

次に Marsh 等の方法と 1 節の個体識別モデルとの関係を整理しよう。まず個体識別が可能ということは、識別を試したときに識別が起きる確率が正ということと同じである。故に個体識別が可能かの判断は、(3) 式が正かの判断と同じである。そして識別が実際に起きた場合に必ず観測されるなら、(2) 式の Pr(識別が実際に起きる) は、1 節の $p(\delta)$ と同じ概念となる。しかし攻撃者が識別に成功しても、黙っていれば観測されるか分からない。故に識別が実際に起きることと観測されることは区別した方がよいかもしれない。この議論は 2.5 節へ先送りする。

結局 (3) 式の右辺の要素のどれかが 0 なら、個体識別が不可能と言える。しかし公開される母集団一意が皆無になるのは例外的で、普通は $\Pr(a,b,c)$ は正となる。(3) 式の右辺を書き換えると

$$\Pr(\text{識別が起きる} \mid \text{識別を試みる}) = \Pr(a,b,c)\Pr(d|a,b,c) \tag{5}$$

であり、$\Pr(d|a,b,c)$ が 0 なら個体識別が不可能と考えられよう。つまり Marsh 等の枠組みにおいて通常の場合、個体識別が可能か否かは $\Pr(d|a,b,c)$ が 0 か否かという問題に縮退する。しかし Marsh 等は $\Pr(d|a,b,c)$ の評価に失敗している。

我々の議論に沿って $\Pr(d|a,b,c)$ が 0 か否かの判別方式を構成しよう。これまでの議論では、個体識別の難易度 δ が (1) 式のように閾値 α 未満なら個体識別が可能ということであった。また δ は (4) 式の関数 h で表されると考えていたので、

$$\delta = h(g(\boldsymbol{y}_1), \boldsymbol{y}_2) < \alpha \Rightarrow \Pr(d|a,b,c) > 0 \tag{6}$$

とすればこれまでの議論と整合する。つまり個体識別の難易度 δ が閾値 α を下回れば、個体識別が可能ということである。

このように考えると、関数 h は $\Pr(d|a,b,c)$ が正という判定とできるだけ直接関係するのが望ましい。そして事象 (a,b,c) が条件の確率を判定するなら、h は (a,b,c) を要因とするべきだろう。これを基準化して $-\Pr(a,b,c) = g(\boldsymbol{y}_1)$ とすれば、g が (4) 式の単調性を満たして都合がよい。何故なら確率 $\Pr(a,b,c)$ は、正確に表現されて公開される母集団一意数と比例する。そして正確に表現されて公開される母集団一意数の増加は、母集団一意の確証をより容易にすると考えられる。故にあとは $\Pr(a,b,c)$ が数値評価可能であれば、その評価値に基づいて匿名化を管理できる。

これまでの議論では $\Pr(a,b,c)$ の意味が曖昧だったが、計算方法を定めれば概念は限定される。また関数 g の具体型と必要な情報 $\boldsymbol{y}_1$ も、計算方法に依存して定まる。そして $\boldsymbol{y}_1$ が決まらなければ、$\boldsymbol{y}_2$ も定まらない。これらは理論モデルとは異なる次元の問題なので、節を改めて考察しよう。

2.3 匿名性の計測—実質と下限

EBA において匿名性の評価値を相対比較する際、評価手法のゆれは望ましくない。また出来るだけ多くの事例を統計的証拠として用いるには評価が名人芸であってはならず、形式的な手続きでなければならない。そのように匿名性の計算手法は具体的に定めておくべきである。本節では前節のモデルに沿って匿名性の評価値 $g(\boldsymbol{y}_1) = -\Pr(a,b,c)$ の計算を考察する。

匿名性の評価をする際、実質か下限のいずれを求めるのか意識的でなければならない。実質とは実際の攻撃者の能力に合わせた評価という意味であり、下限とは統計当局と同じ情報を持つ「最強」の攻撃者を想定するということである。

両者の違いを母集団一意数を例にとって説明しよう。母集団一意数は、キー変数群の多元分割表における度数 1 のセル数と形容することも出来る。この母集団一意を計算する多元分割表で、各変数の区分（カテゴリー分類）は公開データの区分と一致させるのが常識的である。ただ攻撃用情報の精度が公開データより粗ければ、公開データの区分方法で算出した母集団一意は、攻撃者にとっての母集団一意にならない。例えば攻撃者が五歳階級のデータしかもっていなければ、各歳別でデータが公表されていても階級内で識別できない。故に実質的な母集団一意数の評価では、公開表現と攻撃用情報の粗い方に各変数の区分を合わせる。常識的な方法では公開表現の方が攻撃用情報より常に粗いので、最強の攻撃者が想定されている。

実質的な匿名性評価では、現実の攻撃者の能力を知る必要がある。そして攻撃者の能力を知るための情報収集体制については、Elliot et al. (2010) の重要な議論が存在する。この議論は 2.4 節で紹介するが、そのような情報の完全な収集は資源の制約等から無理であろう。つまり実質的な匿名性評価の問題は、必ずしも評価に必要な情報を得られないことである。

部分的な情報から実際にありそうな攻撃方法（シナリオ）を推定し、匿名性を評価することはできる。このようなシナリオ依存のリスク評価は、例えば Paas (1988) が採用している。しかし想定した攻撃者より強い攻撃者が存在した場合、個体識別の可能性は管理されない。

一方、下限の匿名性は後で確認するように、公開データ表現とその元データから評価する。これらの情報は統計当局にとって常に入手可能であり、シナリオ選択に起因する評価のゆれが起きない。また最強の攻撃者より弱い攻撃者についても、個体識別の可能性は（過剰だが）管理できる。ただ問題は、過去の事例における個体識別の有無が現実の攻撃者の能力を反映していることである。

この問題を一般的に考えよう。匿名性の数値評価値の要因 $\boldsymbol{y}_1 = (\boldsymbol{e}_1, \boldsymbol{e}_2)$ について $\boldsymbol{e}_1$ は公開データ表現とその元データと考える。そして $\boldsymbol{e}_2$ は、必ずしも観測されない攻撃者の能力とする。$\boldsymbol{e}_2$ が観測されるとして、実質的な匿名性の数値評価値が $g(\boldsymbol{e}_1, \boldsymbol{e}_2)$ で表される。一方、最強の攻撃者にとっての匿名性の数値評価値を

$$\inf_{\boldsymbol{e}_2} g(\boldsymbol{e}_1, \boldsymbol{e}_2) =: \underline{g}(\boldsymbol{e}_1)$$

で表そう。ここで $\underline{g}$ を用いて過去の事例で計算した匿名性の最低数値評価値を $\underline{\gamma}_1$ と書く。新規に公開するデータの匿名性を $\underline{g}$ で計算して $\underline{\gamma}_1$ としてよいだろうか。

所与の $\boldsymbol{y}_2$ について、匿名性の数値評価値 g が β 未満なら個体識別が可能としよう。過去最低の実質的な匿名性 γ_1 は $\underline{\gamma}_1$ 以上である。故に閾値 $\beta \leq \gamma_1$ が正しいとしても、$\underline{\gamma}_1 < \beta$ となる場合があり得る。このとき新規に公開するデータの実質的匿名性が例えば $\underline{\gamma}_1$ と等しければ、個体識別は可能となってしまう。

このような事態は、匿名性の実質と下限の差が変動する場合に起こりうる。新規に公開するケー

スについて匿名性の要因を $(\boldsymbol{e}'_1, \boldsymbol{e}'_2)$ と書く。ただし $\inf_{\boldsymbol{e}'_2} g(\boldsymbol{e}'_1, \boldsymbol{e}'_2) = \underline{\gamma}_1$ となるように匿名化がなされているとしよう。そして $\underline{g}$ を用いて評価した過去最低の匿名性のケースの要因を $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ と書く。つまり $\inf_{\boldsymbol{e}'_2} g(\boldsymbol{e}'_1, \boldsymbol{e}'_2) = \inf_{\boldsymbol{e}_2} g(\boldsymbol{e}_1, \boldsymbol{e}_2)$ が成立している。ここで匿名性の実質と下限の差を過去のケースは $c = g(\boldsymbol{e}_1, \boldsymbol{e}_2) - \inf_{\boldsymbol{e}_2} g(\boldsymbol{e}_1, \boldsymbol{e}_2)$ 、新規のケースは $c' = g(\boldsymbol{e}'_1, \boldsymbol{e}'_2) - \inf_{\boldsymbol{e}'_2} g(\boldsymbol{e}'_1, \boldsymbol{e}'_2)$ で表す。攻撃者の能力が向上して $c > c'$ の時、新規ケースの実質的匿名性は過去最低の実質的匿名性を下回る。そして $\beta = g(\boldsymbol{e}_1, \boldsymbol{e}_2)$ なら、新規ケースの実質的匿名性は $g(\boldsymbol{e}'_1, \boldsymbol{e}'_2) = \underline{\gamma}_1 + c' < \beta$ であり、過去のケースでは不可能だった個体識別が可能となる。

なお上の考察で c は下限評価の歪みを含む。下限評価の真値からのずれは、一定なら問題にならないことは重要だ。つまり実質 g と下限 $\underline{g}$ の差 c がケース毎に変化しなければ、過去最低の匿名性の下限 $\underline{\gamma}_1$ を与えるケースでは実質的な匿名性も過去最低になる。そして $\underline{\gamma}_1 > \beta - c$ なら $\gamma_1 > \beta$ なので、$\underline{g}$ を用いて匿名化の程度を決めれば実質も管理される。言い換えれば、EBA は匿名化を相対比較するので、匿名性の絶対値に意味は無いということである。

匿名性の実質と下限の差 c が一定という重要な条件を満たすには、たとえ歪んでいても同じ方法で測ることが重要である。また $\boldsymbol{e}_2$ は無視できず、変化を確認するべきだ。しかし情報 $\boldsymbol{e}_2$ は入手性に問題があるので、$\boldsymbol{y}_2$ の一部として定性評価するしかないだろう。

このような前提で、匿名性の数値評価は $\underline{g}(\boldsymbol{e}_1)$ を用いるのが望ましい。つまり公開データ表現とその元データから匿名性の下限を求めるということである。そのように $\Pr(a, b, c)$ が計算できるか要素毎に確認しよう。

$\Pr(a)$ の評価　Marsh 等は誤分類や誤記が公開ファイルと攻撃用ファイルのキー変数で起きていない確率を $\Pr(a)$ とした。1981 年の英国センサスの事後調査 (Post Enumeration Survey) で求めた変数の誤分類率を参照して、5 つのキー変数が全て正確に分類されている割合は 0.8 程度と見積もられている。この場合に誤分類が公開ファイルと攻撃用ファイルのキー変数で独立に起きているなら、$\Pr(a) = 0.8^2 = 0.64$ である。なおキー変数が増えれば、全てのキー変数が正確に分類されている確率は減少する。しかし母集団一意数は増えることになる。

実際には、公開ファイルと攻撃用ファイルで調査時点の差や変数の定義の差も存在するだろう。これらの差は $\Pr(a)$ を低下させる。1971 年の英国センサスの 1 年後に再調査した結果、同じ職業だった人の割合が 61%でしかない例を Marsh 等は挙げている。1991 年の英国センサスについては、Dale and Elliot (2001) が各キー変数が経時変化する程度を調べている。ただ Dale and Elliot も述べているように、本気の攻撃者は特定の調査が数年後に公開されることを見込み、同時点に調査した攻撃用ファイルを準備しておくだろう。この場合は、調査時点や変数の定義の差に多くの保護効果を期待出来ない。このように攻撃のシナリオに依存して、$\Pr(a)$ はかなり変化する。

我々は匿名性の下限を評価したいので、最強の攻撃者を想定する。このシナリオでは、匿名化される前のキー変数が全て攻撃者にばれていると考える。この場合キー変数の精度を評価するに

は、匿名化されていない元ファイル[14] と公開ファイルのキー変数を比較する。そして近いレコードが同個体（のペア）と判定し、正しく判定された割合[15] を $\Pr(a)$ と考える。このような手法は開示リスク評価によく用いられるので、研究蓄積が利用可能である。例えば伊藤他 (2009) を見よ。なお我々のシナリオでは、公開ファイルと攻撃用ファイルで調査時点の差は存在しない。また両者で定義の差は、匿名化によるもののみである。そして元ファイルのキー変数がどれほど誤分類されていたとしても、個体と正しく対応可能である。

このようなシナリオの非現実性は、現実に用意可能な攻撃用データの質と量に依存する。この情報が $\boldsymbol{e}_2$ であり、$\boldsymbol{y}_2$ の一部と考える。一方 $\boldsymbol{y}_1$ は元ファイルと公開キー変数データだが、これらを比較し、正確にマッチされたレコードの割合が $\Pr(a)$ として計算可能であった。注意すべきなのは、毎回同じ方法でマッチさせることである。

$\Pr(b|a)$ の評価　Marsh 等の議論で確率 $\Pr(b|a)$ は、公開個体数が母集団サイズにしめる割合である。例えば 1991 年の英国センサス匿名化標本では 2%となる。全数調査から等確率でサブサンプリングした公開ファイルなら、個体は等確率で公開ファイルに含まれる。その場合に Marsh 等の方法は妥当である。

しかし現実の標本調査では不等確率の複雑な抽出が行われる。また一部の個体について、調査されたか否かを攻撃者が知っているかもしれない。例えば集落抽出を行う調査では、被調査者は隣家も調査されたと推測できる。従って一般に真の $\Pr(b|a)$ は個体毎に異なる。

ただ我々は個体毎（いわゆるレコードレベル）の確率評価をしているのではなく、ファイルレベルの評価が目的である。ファイルレベルでは公開の平均的な可能性を評価すると考えて、Marsh 等の方法を用いることにしよう。

$\Pr(c|a,b)$ の評価　母集団一意数が母集団サイズにしめる割合を求めればよい。なお本節冒頭で議論したように、母集団一意数を計算するための変数の区分は公開表現に従うべきである。また世帯単位のファイルでは、世帯毎の固まりを「レコード」として母集団一意を計算するのが筋である。具体的には、世帯員のレコードを年齢順に連結したまとまりを一レコードとして扱えば良い。この場合、世帯人数が異なればレコード長も異なる。

ただし全数調査でない限り、母集団一意数は推定しなければならない。そして星野 (2003) で説明したように、母集団一意数の推定は単純ではない。

Marsh 等は英国センサスの全数データが使えなかったため、イタリアのセンサスデータで母集団一意を数えて外挿している。キー変数が 8 つで 10 万人レベルの地域区分を公開するとして、

[14] 攻撃者は補定、エディットのルールを知らないはずなので、補定等を施す前のデータを元ファイルとする方が現実に近いかもしれない。ただそのようなデータが常に利用可能とは限らない。相対比較可能性を考えれば、補定等を施した後のデータを元としてよいだろう。実質と下限の差があるとしても、補定等の割合が小さかったり調査毎に大きく変動しない場合は無視できる。

[15] 何を分母とするかは議論の余地がある。本当に評価したいのは、母集団一意レコードについてのキー変数の精度である。しかし全数調査でないと、母集団一意のレコードを決めるのは難しい。そして評価の歪みより方法の変動を避けたいので、標本一意数を分母とするのが一案である。近さの計算方法によるが、一意にペア相手が見つかるレコード数と標本一意数はほぼ同じである。なお分母が 0 の場合は $\Pr(a) = 0$ とみなして差し支えないだろう。

$\Pr(c|a,b)$ は 2.4%程度とされた。なおこの値は世帯単位ではなく個人単位で評価されている。1991 年の英国センサスデータについては、Dale and Elliot (2001) によるとキー変数が 7 つで 12 万人レベルの地域区分を公開する前提で、$\Pr(c|a,b)$ は 4.8%であった。

母集団一意数評価は Marsh 等の時代に比べてかなり進歩しており、（母集団サイズが所与で）公開ファイルの情報だけから推定できる。しかし評価手法による結果の違いが大きいので、同一手法によって評価することの重要性も大きい。

幅広い母集団について一意数の推定精度をルーチンワークとして確保するには、ピットマンモデル（付録 B を参照のこと）の使用を推奨する。この方法においてデータは、無限母集団すなわちピットマン分布からの標本とみなされる。そして母集団[16]も同一無限母集団からの標本とみなすので、データからピットマン分布の母数を最尤推定し、推定値の下で母集団一意数の挙動を求める。より具体的には、付録 C の手順書を参照されたい。

開示リスクを評価するファイルのレコード数は、母集団個体数のせいぜい一割程度であろう。この場合に安定的な母集団一意数の推定量は、全てバイアス[17] を持つ。手順書の推定量も例外でなく、おそらく過大に一意数を推定する。しかし既に考察したように、バイアスは一定であれば問題にならない。

なお特定のモデルと決めつけるよりも、モデル集合からデータに良く当てはまるモデルを選択し、そのモデルで一意数を推定する方が正確になろう。しかしモデル集合の空間をうまく張らないと、リスク評価値がぶれる。また経験的に多くの場合、ピットマンモデルが選択[18] される。故に手間や精度及び様々な結果の整合性を勘案すれば、母集団一意数は常にピットマンモデルによって推定するのが最善と思われる。

一点つけ加えておくと、母集団一意数の推定改善にセルの番地情報を使うアプローチはあり得る。しかし大規模かつ疎な分割表では絶対的に情報が不足しているので、うまくいかないであろう。またそのようなアプローチは高度なモデリングが要求され、開示リスク評価の試行錯誤にも向かない。故に実務への採用は難しいはずだ。

このように $\Pr(a,b,c)$ の下限評価に必要なのは、

$$\boldsymbol{e}_1 = (\text{元ファイル}, \text{公開ファイルのキー変数}, \text{母集団サイズ})$$

である。これらの情報が数値評価の対象となり、$\boldsymbol{y}_2$ には含まれないと考えるべきだろう。次節では数値評価しない要因 $\boldsymbol{y}_2$ を確定しよう。

---

[16] 一部が観測されている現実の母集団について推定するのではなく、同サイズの母集団を新たに発生させる場合の挙動が推定される。

[17] 有限母集団から非復元単純無作為抽出する場合、一意数の不偏推定量は一意に存在する。しかしこの不偏推定量は標準誤差が大きく、標本抽出率がかなり高くないと実用に耐えない。そして一意な不偏推定量なので、推定を安定させるためのいかなる工夫もバイアスを生む。

[18] 裾の長いモデルとして代表的な負の二項分布は、統計的開示制限の分野ではポアソン＝ガンマモデルとして知られている。このモデルは基本的に広義のピットマンモデルの特殊ケース ($\alpha \leq 0$ に対応）である。故にピットマンモデルのデータへのあてはまりは、基本的にポアソン＝ガンマモデルを下回らない。そしてポアソン＝ガンマモデルによる母集団一意数の推定値は、必ず Pitman モデルの推定値より（かなり）小さくなると考えて良い。

2.4 定性評価の要因

本節では個体識別について数値評価しない要因 $\boldsymbol{y}_2$ を定める。これまでの議論より、我々は (6) 式に基づいて母集団一意の確証の可能性 $\Pr(d|a,b,c)$ を判断するので、h の引数 $\boldsymbol{y}_2$ は母集団一意の確証にかかる要因である。そして前節では $\boldsymbol{e}_2$、すなわち攻撃用データの質と量が、$\boldsymbol{y}_2$ の一部ということであった。

Marsh 等は母集団一意の確証手法として、全数名簿と公衆の目の利用[19] を検討している。全数名簿の利用とは、職業人名簿等で母集団一意が分かる場合を指す。特定の条件を満たす集団について全数の名簿があれば、その集団内の一意[20] は母集団でも（特定の条件を満たす）一意である。そのような個体について、Marsh 等は特に強い匿名化を求めている。それから公衆の目とは、珍しくて目立つ個体が有名な場合を言う。例えば職業が現職の首相であれば、母集団一意を確証可能である。昨今ではソーシャルネットワークの拡大により、公衆の目は無視できないように思う。

全数名簿が利用出来たり、属性が公衆に知られていたりする個体については、詳しい個体情報が社会に流通しているということだ。母集団一意の確証可能性及び実質的な $\Pr(a,b,c)$ は、そのような個体情報の環境に依存するだろう。個体情報環境を知るため、Elliot et al. (2010) は (i) アクセス制限付きデータベースの調査項目、(ii) 公知の個体データの形態、(iii) ネットショッピング等での web 上データ収集項目、(iv) 商業データベースの情報、(v) 個人情報の収集実験結果、(vi) 情報保有組織における個体情報の取り扱い慣行、(vii) ソーシャルネットワークでの個体データの形態、を調べることを提案している。またそこで現れる様々な変数間の関係を、ツリー構造を用いて記録することとしている。これらの要因は調査できたとしても、定量評価は（識別成功が希なので）難しい。ただこれらについての理解から、現実的な攻撃用データとして

$$\boldsymbol{e}_2 = (\text{外部データに含まれる個体数、変数の種類、精度})$$

を想定するべきだろう。なお世帯データの $\boldsymbol{e}_2$ は、事業所データの $\boldsymbol{e}_2$ と明らかに異なる。従って世帯データの匿名化事例は、事業所データの匿名化のエビデンスとして直接使えないということになる。個体情報環境は個体単位（個人、世帯あるいは事業所等）毎に集約するべきだ。

外部データに含まれる個体数は、実質的に攻撃可能な母集団一意数と比例するだろう。なお外部データに含まれる個体数増加の効果は、サブサンプリングにより $\Pr(b|a)$ を下げれば打ち消すことができる。公開個体率 $\Pr(b|a)$ も攻撃可能な母集団一意数と比例すると考えられるからだ。

外部データの変数の種類は、キー変数の決定に用いる。前節ではキー変数が所与であったが、実際はキー変数を選択しなければリスク評価が出来ない。そしてキー変数の選定基準の揺れは避けた方がよい。これを念頭におき、過去の事例で用いたキー変数の種類は、キー変数の選択で考慮すべきである。キー変数に相当すると判断した根拠の外部データの状況が変わらなければ、同じ変数はキーとして用いなければならない。根拠が変われば、キー変数も変えるべきだろう。なお

[19] 他に母集団一意の確率を統計モデルで求めることを挙げているが、それでは母集団一意の確証にならない。Dale and Elliot (2001) による Marsh 等の議論の再評価でも、統計的推測は母集団一意の確証として扱われていない。

[20] 全数調査において低次元クロス集計の結果の度数が 1 と分かるような場合も該当する。

Elliot et al. (2011) がキー変数の選択基準を考察している。彼らの議論では、変数のアクセス容易性を定性評価した上でキー変数が選択される。

外部データの変数の精度は、$\Pr(a)$ と $\Pr(c|a,b)$ の実質的な値と関係する。なお変数の精度上昇の効果は、匿名化を強く施せば無効化出来る。何故なら匿名化で定まるデータの粗さ以上に変数の精度が上昇しても、開示リスクは変化しない。

$\boldsymbol{e}_2$ 以外の定性評価要因として、匿名化の「曖昧さ」を検討しておこう。ここでは匿名化に用いたデータ変換 m の形を攻撃者が完全には知らない場合を曖昧と呼ぶ。例えば米国センサスミクロデータのように、スワッピングが施されているがその割合やスワップ相手の選択方法などが未公開な状態は曖昧である。他方、労働力調査等の匿名データでは、符号表を読むことで匿名化が施されている変数や程度が完全に分かる。この状態は曖昧ではない。

曖昧さは余り研究されておらず、その効果[21]に定説はない。一つの理由として、計算機科学では曖昧さによる安全性を認めないことが挙げられる。その前提で設計した匿名化は統計当局が隠した情報が漏れても[22] 破られないので、保守的と言える。しかしこのような態度は最強の攻撃者を想定することと同じである。従って下限と実質の差の問題が起きる。

曖昧さが母集団一意の確証に影響する例を挙げよう。年齢と性別の二キー変数について、元ファイルが {(110,M),(120,F)}、公開ファイルが {(120,M),(110,F)} だとする。年齢をスワップしたと考えれば第一レコード同士が同一個体であり、性別をスワップしたと考えれば元ファイルの第一レコードと公開ファイルの第二レコードが同一個体となる。この場合は m について何も知らないと、公開ファイルのレコードが元ファイルのどちらのレコードか分からない。ところが年齢変数に適当なノイズを付加したという情報が有れば、元ファイルと公開ファイルで同一個体のレコードが判明する。そして年齢が 120 歳の母集団一意な個体は、公開ファイルの第二レコードと確証される。

このように母集団一意の実質的確証可能性は、曖昧さの程度に依存するかもしれない。故に曖昧さは匿名化設計の一部として、明示的に考察した方がよい。現実には、攪乱的手法の詳細を公開する程度を $\boldsymbol{y}_2$ の一部として管理するということになろう。なお補定やエディットの母数を明らかにしないことは、曖昧と同じことになる。

ここまでの議論で、既存の情報に基づく順攻撃による個体識別はある程度管理されるだろう。しかし情報が追加できるなら、これまでの枠組みでは管理されない事態が起きる。例えばあるレコードの識別が既存情報から確証できないにせよ、可能性が高いとしよう。この場合に追加の情報を詐取などすれば、確証できるかもしれない。追加情報を想定しての匿名化はあり得るが、詐取の可能性を際限なく考慮すると、有用なファイルの提供は不可能だろう。それよりも追加情報取得の可能性を低く保つ工夫をする方が良い。

[21]特定の曖昧さの効果は、例えば以下のように評価できる。保守的な攻撃者なら、曖昧な部分に自分に不利な事前分布を入れる。このように評価される攻撃の難易度と真の難易度の差が、曖昧さの効果である。

[22]関係者による情報漏洩だけ考慮すれば良いわけではない。攻撃者が攪乱の母数を推定できる可能性がある。例えば匿名データと匿名化されていないデータの分析結果を比較することで、攪乱の率の見当をつけられるかもしれない。攻撃者本人が 33 条申請による目的外使用でデータを手に入れなくても、他人が書いた論文や公の集計表が比較対象になり得る。

重要な追加情報を得るには、当該個体に接触する必要があるのではないか。そして接触するには、広い意味[23]での個体の位置（住所、職場など定期的に訪れる場所、電話番号等）を知らなければならない。故にそのような接触可能性に係る条件で母集団一意な個体は、そうでない母集団一意よりも追加情報を得やすいので、確証の可能性が上がる。また逆攻撃は、接触可能な範囲で行われる。従って広い意味での位置情報の精度は、一定以上にならないように管理するべきである。

本節の議論をまとめておこう。定性評価の対象 $\boldsymbol{y}_2$ として過去の事例と比較されるのは、以下の 3 要因である。

1. 同種の母集団についての $\boldsymbol{e}_2$：民間データベース等に含まれる個体数、変数の種類、精度
2. 匿名化の曖昧さ
3. 接触を可能とする情報の精度

本節で考察したように、匿名化手法の変更により攻撃者の能力向上を無効化できる場合がある。故に $\boldsymbol{y}_2$ が過去と同じかどうかの判断は、データ表現にある程度依存してしまう。匿名性の数値評価値 $\underline{g}$ を変えるために匿名化手法を変更すると、定性評価も変わるかもしれないことは注意するべきである。

2.5 識別を試みる確率の決定要因

これまでの議論で後回しにされた、個体識別が実際に起きる確率と観測される確率 $p(\delta)$ の差について本節では議論する。つまり $p(\delta) = \Pr(a, b, c, d) \cdot \Pr(\text{識別を公表するつもりで試みる})$ という関係が (2) 式の関係 $\Pr(\text{識別が実際に起きる}) = \Pr(a, b, c, d) \Pr(\text{識別を試みる})$ と違うかもしれないので、識別を公開すること、公開しないことについて要因の考察を行う。

識別を試みるという意思決定は、識別成功の損得や容易性に依存すると考えられる。Marsh et al. (1991) が指摘するように、$\Pr(\text{識別が起きる} \mid \text{識別を試みる}) = \Pr(a, b, c, d)$ の減少は $\Pr(\text{識別を試みる})$ を減少させるだろう。他に Elliot et al. (2010) は、もっともらしい攻撃シナリオの考察こそが、$\Pr(\text{識別を試みる})$ の妥当なモデル化につながると主張している。

攻撃者が真に識別を成功させた場合、その事実を公表して得られる利益と、識別を隠して得る利益がある。まず識別成功を公表した場合、攻撃者は有名になるだろう。そして識別された個体は情報の漏洩を知ることになり、識別によって入手した情報を用いた詐欺、ストーキング等は難しくなる。そのように識別で得た情報を実用するには、識別成功は公表しない方がよい。また識別成功を公表すれば法的、社会的制裁の対象[24] になるかもしれない。故に例えば商業目的なら識別成功を公表せず、攻撃者は精度の良いマーケティングの利益を享受するだろう。

[23] 狭義の地理情報が強力なキー変数であることは良く知られている。

[24] 匿名データの利用者については、統計法第 43 条第 2 項に「当該匿名データをその提供を受けた目的以外の目的のために自ら利用し、又は提供してはならない」とある。個体識別の成功を公表することは（識別目的でのデータ提供は行われないので）本条に違反するが、直ちに罰則が適用されるわけではない。匿名データの利用者についての罰則は「匿名データを、自己又は第三者の不正な利益を図る目的で提供し、又は盗用した者」に対して「五十万円以下の罰金に処する」(61 条 3 項）とだけ定められている。例えば匿名データの不備を指摘するための個体識別の公表は不正な利

このように考えれば、公開ファイルが含む実用（隠れて悪用）可能な情報が多ければ、Pr(識別を公表するつもりで試みる) を増加させる。また識別を試みるという事象は識別を公表するつもりで試みる事象を包含するので、Pr(識別を試みる) も増加する。

ただファイルが実用可能な情報を含まなくても、識別成功の公表により統計当局の面目を失わせ有名になることを、魅力的に感じる人間が居ないとは言えない。故に Pr(識別を公表するつもりで試みる) は正のはずで Pr(識別を試みる) $= 0$ にはならない。しかし実用の帰結は多様なのに対し、識別成功の公表は帰結が同じである。実用できない情報は公表することによってしか利益を得られないので、攻撃の誘因として全て等価ということになる。

では実用可能な情報とは何か。多様な犯罪を想像して判断するしかないが、匿名化によって実用性は変えられることを指摘しておく。例えば病歴という情報は、削除したり罹患時期を区間表示したりすることで、実用困難にできる。多くの統計調査は適切に匿名化すれば実用可能な情報を含まず、攻撃の誘因は識別成功の公表による利益のみとなる。そしてこの場合 $p(\delta) \doteq$ Pr(識別が実際に起きる) と考えて良いはずだ。

なお調査客体が秘密にしたい調査項目（変数）を「センシティブ変数」と呼ぶ。秘密でない情報は保護に値しないので、実用を妨げる目的での匿名化の対象は、センシティブ変数の一部と考えられる。重要なのは、センシティブか否かは調査客体の主観に依存[25] するということである。故にセンシティブの程度は、攻撃の動機付けの程度と必ずしも一致しない。

3　おわりに—匿名データの審査体制について

最初にこれまでの議論をまとめる。個体識別が不可能かつ有用なデータを統計的根拠に基づいて作成する手順は以下のようになる。

1. $\boldsymbol{y}_2$ が共通する過去の事例をリストアップする。
2. それらの事例について $\Pr(a, b, c)$ の下限 $\underline{g}$ をそれぞれ計算する。
3. その中で最も高い $\underline{g}$ を g^* と書く。
4. データの匿名性の評価値が g^* となるように匿名化する。

なお匿名性の評価値が同じになる複数のデータ表現では、データの有用性が高いものを選べばよい。本稿で有用性の評価は議論しないが、例えば星野 (2010) は基本的な考え方を説明している。以下ではこのような立場から、望ましい匿名データの審査体制を考察する。

益を図る目的と必ずしも言えないので、罰則は適用できないのではないか。なお 33 条申請によって調査票情報を手に入れた者が個人又は法人の秘密を漏らした場合は「二年以下の懲役又は百万円以下の罰金」(57 条 2 項 3 号)、自己又は第三者の不正な利益を図る目的で提供又は盗用した場合は「一年以下の懲役又は五十万円以下の罰金」(59 条 2 項)、と罰に差がつけられている。ところが匿名データの利用者が（個体識別によって入手した）秘密を漏らした場合の罰則規定はなく、そのような事態を統計法は想定していないように思われる。

[25] 全ての調査客体の判断を聞くのは非現実的なので、リスクの評価者がセンシティブ変数を定める際に保守的であれ、ということになる。

まず審査用資料（チェックリスト）は、蓄積して参照するものだということをはっきりさせておきたい。EBA は過去の経験をエビデンスとして用いる。故に過去のチェックリストを、経験の要約として用いたい。この事情は、将来的に α と δ の関係をモデル化する場合も変わらない。

従ってチェックリストの記載事項は、事例の十分統計量であるべきだ。つまり個体識別が可能か否かの判断に用いる情報を（過）不足なく記入するということである。このような観点から、現行のチェックリストは日本の制度にふさわしいだろうか。世帯調査のチェックリスト（H23/3/28 改正版）についてのみ、改善できる点を指摘したい。

チェックリストに記載するべき項目で漏れているのは、まず $\Pr(a,b,c)$ の下限 $\underline{g}$ である。付録の手順書に従えばそれほど計算に手間がかかるとは思えず、匿名化表現の要約として費用対効果が高い情報と考える。またキー変数の情報を持つ部分集団の全数名簿は、母集団一意の確証について重大な影響がある。故に質問項目として特に欄を設けるべきである。そのような名簿が存在するなら、名称、部分集団の種類、個体数、含むキー変数の種類、精度を記述させるとよい。他に狭義の地理情報については記入欄が存在するが、接触を可能とするような広義の地理情報の有無を確認するべきだ。

それからチェックリストに存在する項目で、記入の焦点をしぼるべき箇所がある。まず「ミクロデータを特定できる可能性のある外部ファイル」の存在を記入することになっているが、どのようなファイルが該当するのか明確化するべきである。具体的には、匿名データが含む変数と同じ情報（これがキー変数ということである) をもつ外部ファイルの有無を問うべきだ。そしてそのファイルの名称、含むキー変数の種類、精度、及び個体数を分けて記述してもらう方がよい。また「秘密の情報」（センシティブ変数）のうち、「特に秘匿する必要性の高い調査項目」の有無を聞いているが、必要性の意味を明白にしたほうがよい。個体識別の可能性を制限するための必要性ではなく、実用性を限るための匿名化の必要性を聞かなければならない。またチェックリストには「誤差（ノイズ)」を聞く項目が存在する。誤差の付加は「攪乱」手法の例なのだが、用語の問題は別にして、この項目には匿名化の曖昧さを評価するための情報を記入させるべきだ。具体的には、攪乱手法のパラメータと、その公開方針を分けて書かせるということになる。

チェックリストに記入される情報の使われ方は、説明書を用意するべきであろう。現行のチェックリストも冒頭で匿名化の考え方などが書かれているが、やや説明不足に見える。個体識別可能性の判定方式を明示すれば、焦点がずれたチェックリスト記入の恐れは減る。

Acknowledgements

本研究は科学研究費及び統計数理研究所の共同研究経費の補助を受けている。以下の付録 B,C は星野 (2012) の一部を改訂したものである。

付録

A　世帯調査のチェックリスト（H23/3/28 改正版）要約

1. 地理的情報
   (a) 地理情報のレベル、加工の有無
   (b) 地理情報以外の地理的情報の有無
   (c) 地域分析用の地理情報提供の有無
   (d) 特定の種類の施設の情報の有無
2. 世帯の識別情報
   (a) 世帯のキー変数
   (b) キー変数への匿名化及び分布
   (c) 世帯のまとまりへの匿名化の有無
3. 個人の識別情報
   (a) 個人のキー変数
   (b) キー変数への匿名化及び分布
4. 攪乱の有無
5. サブサンプリングの有無
6. 外部の情報
   (a) 個人・世帯の特定に使える外部情報の存在
   (b) 母集団情報として利用している情報
7. その他
   (a) データの並び順についての匿名化措置
   (b) サンプリング情報により特定の地域や集団であることが明らかになる可能性
   (c) センシティブ変数への匿名化
   (d) 提供時期と調査時点との差
   (e) その他の匿名化処理の有無

B　ピットマンモデルについて

自然数 $n \in \mathbb{N} := \{1, 2, 3, \ldots\}$ を自然数の和で表す事を分割と呼ぶ。この和の中で自然数 i が足される回数を s_i で表せば、$\boldsymbol{s}_n := (s_1, s_2, \ldots, s_n)$ は（順序無しの）分割を表す。非負整数の集合を $\mathbb{N}_0$ で表すと、n の全ての分割の集合は $\mathcal{S}_{|n} := \{\boldsymbol{s}_n : s_i \in \mathbb{N}_0, i \in \{1, 2, \ldots, n\}, \sum_{i=1}^n i s_i = n\}$ で表される。この集合上の分布が自然数の確率分割である。以下では $u := \sum_{i=1}^n s_i$ とする。

Pitman 分布 (Pitman, 1995) は自然数の確率分割であり、母数 $0 \leq \alpha < 1, \theta > -\alpha$ について確率関数は以下のように書ける。

$$p(s_1, s_2, \ldots, s_n) = n! \frac{\theta^{[u:\alpha]}}{\theta^{[n]}} \prod_{j=1}^n \left(\frac{(1-\alpha)^{[j-1]}}{j!} \right)^{s_j} \frac{1}{s_j!}, \quad \boldsymbol{s}_n \in \mathcal{S}_{|n}, \tag{7}$$

ただし $\theta^{[u:\alpha]} = \theta(\theta+\alpha)\cdots(\theta+(u-1)\alpha)$, $\theta^{[n]} = \theta(\theta+1)\cdots(\theta+n-1)$ である。

個票データとの対応を述べておこう。匿名化の程度を定めることで、キー変数に関する分割表が出来る。分割表の情報のうち度数を、第 j セルについて f_j と書く。ただしセル総数が J として $j \in \{1, 2, \ldots, J\}$ である。ここで $i = 1, 2, \ldots, n$ について度数 i のセルの数を s_i と表す。つまり指示関数 $1(\cdot)$ を使えば、$s_i = \sum_{j=1}^J 1(f_j = i)$ である。例えば s_1 は標本で一意なレコード数となる。このように作られる $\boldsymbol{s}_n$ を「寸法指標」と呼び、$n \ll J$ なら Pitman 分布の標本とみなせる。母集団サイズ $\tilde{n}$ も J よりかなり小さいなら、確率変数 $\boldsymbol{S}_{\tilde{n}} := (S_1, S_2, \ldots, S_{\tilde{n}})$ が Pitman 分布に従う場合、S_1 で母集団一意数の挙動が表せる。

経験ベイズ的に母集団一意数推定の論理を説明すると、以下の通りになる。$\boldsymbol{S}_{\tilde{n}}$ の事前分布が Pitman 分布であり、その母数 (α, θ) は超母数である。超母数はデータ $\boldsymbol{s}_n$ により（最尤）推定される。推定したい母数は S_1 であり、$(S_2, S_3, \ldots, S_{\tilde{n}})$ は局外母数である。母数 S_1 の周辺分布については Hoshino (2012, Theorem 3) を見よ。

Pitman 分布に従う $\boldsymbol{S}_n$ の任意の周辺階乗モメントは、Yamato and Sibuya (2000) が与えている。特に

$$\mathrm{E}(S_i) = \frac{(1-\alpha)^{[i-1]} n^{(i)}}{i!} \theta \left(\frac{(\theta+\alpha)^{[n-i]}}{\theta^{[n]}} \right), \tag{8}$$

である。

超母数の最尤推定量を $(\hat{\alpha}, \hat{\theta})$ と書けば、(8) 式に $\alpha = \hat{\alpha}, \theta = \hat{\theta}, n = \tilde{n}, i = 1$ を代入して母集団一意数の点推定量が得られる。すなわち

$$\hat{S}_1 = \tilde{n} \frac{(\hat{\theta}+\hat{\alpha})(\hat{\theta}+\hat{\alpha}+1)\cdots(\hat{\theta}+\hat{\alpha}+\tilde{n}-2)}{(\hat{\theta}+1)(\hat{\theta}+2)\cdots(\hat{\theta}+\tilde{n}-1)}. \tag{9}$$

なお Hoshino (2001, Proposition 3) によれば、$\alpha \geq 0$ について $\lim_{n\to\infty} \mathrm{E}(S_1)/\mathrm{E}(U_n) = \alpha$ であ

る。ただし $U_n := \sum_{i=1}^{n} S_i$ は度数が 0 でないセルの総数なので、母集団で空でないセルのうち一意のセル数の割合は α と解釈出来る。

次にフィッシャー情報量を確認しておこう。まず対数尤度関数を

$$L(\alpha,\theta) = \sum_{i=1}^{u-1} \log(\theta + i\alpha) - \sum_{i=1}^{n-1} \log(\theta + i) + s_1 + \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \log(j - \alpha) + \text{Const.} \tag{10}$$

で表す。二次の微分係数は

$$\frac{\partial^2 L(\alpha,\theta)}{\partial \theta^2} = -\sum_{i=1}^{u-1} \frac{1}{(\theta + i\alpha)^2} + \sum_{i=1}^{n-1} \frac{1}{(\theta + i)^2}, \tag{11}$$

$$\frac{\partial^2 L(\alpha,\theta)}{\partial \alpha^2} = -\sum_{i=1}^{u-1} \frac{i^2}{(\theta + i\alpha)^2} - \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \frac{1}{(j - \alpha)^2} < 0, \tag{12}$$

$$\frac{\partial^2 L(\alpha,\theta)}{\partial \theta \partial \alpha} = -\sum_{i=1}^{u-1} \frac{i}{(i\alpha + \theta)^2} < 0 \tag{13}$$

である。(13) 式より $\hat{\alpha}$ と $\hat{\theta}$ は負の相関を持つ。情報量はこれらの式について u を U_n に、s_i を S_i に置き換えて期待値をとる。$\mathrm{E}(S_i)$ は (8) 式で与えられているので、あとは

$$\mathrm{P}(U_n = u) = \frac{\theta^{[u:\alpha]}}{\theta^{[n]}} (-1)^{n-u} C(n,u,\alpha) \alpha^{-u}, \quad u \in \{1,2,\ldots,n\}. \tag{14}$$

を利用して数値的に評価できる。$C(\cdot,\cdot,\cdot)$ は C-ナンバーと呼ばれ、一般化されたスターリング数である。C-ナンバーについては Charalambides and Sing (1988) を参照のこと。Sibuya and Yamato (2001, Proposition 5) がフィッシャー情報量行列のオーダーを評価しており、$n \to \infty$ の時 $I_{\theta\theta} = O(1), I_{\theta\alpha} = O(\log n), I_{\alpha\alpha} = O(n^{\alpha})$ である。特に θ の推定精度は悪い。

C　母集団一意数の推定手順

以下では標本サイズを n、母集団サイズを $\tilde{n}$ と記す。

1. 評価するキー変数とその精度を決める。
2. 決められたキー変数全てについてクロス集計する。つまり（高次元の）分割表を作り、各セルに所属するレコード数（度数）を数える。
    - セル総数 J は、全てのキー変数のカテゴリー数の積である。連続変数でも現実には有限個の表現しかとらず、その表現の数をカテゴリー数と考える。

- 第 j セルの度数を $f_j, j = 1, 2, \ldots, J,$ と書く。以下の結果はインデクス j の付け方に依存しない。

3. 空でないセルの度数の度数（寸法指標）を数える。

   - $i = 1, 2, \ldots, n$ について度数 i のセルの数を s_i と表す。つまり指示関数 $1(\cdot)$ を使えば、$s_i = \sum_{j=1}^{J} 1(f_j = i)$ である。
   - 最大のセルの度数が m ならば、$m < i$ について $s_i = 0$ である。

4. データを生成した構造（確率分布）を推定する。

   - 現実の母集団を無限母集団（超母集団）からの標本とみなす。この場合、手元の標本から超母集団の分布を推定すれば、母集団の挙動も推定される。
   - 超母集団の分布として広義の Pitman モデルを仮定し、その母数を最尤推定する。
   - Pitman モデルは 2 母数 (α, θ) を持ち、α が負の場合と正の場合で分けて考えた方が良い。どちらの場合も $u = \sum_{i=1}^{n} s_i$, $n = \sum_{i=1}^{n} i s_i$ である。
     (a) $0 \leq \alpha < 1, \theta > -\alpha$ について Pitman モデルの確率関数は (7) 式で表される。
     (b) $\alpha < 0$ の場合は (7) 式で $\theta = -J\alpha$ とおき、さらに $-\alpha = \gamma$ とおく。すると一母数の確率関数を得る：

$$p(s_1, s_2, \ldots, s_n) = \frac{n! J! \Gamma(J\gamma)}{\Gamma(J\gamma + n)} \prod_{i=0}^{n} \left(\frac{\Gamma(\gamma + i)}{\Gamma(\gamma) i!} \right)^{s_i} \frac{1}{s_i!}. \tag{15}$$

     ここで $\gamma > 0$ であり、$s_0 = J - u$ である。

   - モデル (7) を「(狭義の)Pitman モデル」と呼ぶ。モデル (15) を「多項ディリクレモデル」と呼ぶ。本来は AIC 等によりデータ依存でいずれかをモデル選択するのが良いが、ここでは簡易的な選択基準を示す：
     (a) 母集団サイズ $\tilde{n}$ が総セル数 J より大の場合、多項ディリクレモデルを用いる。
     (b) その他の場合は Pitman モデルを用いるが、尤度の最大化に失敗する（繰り返し計算が収束しない）場合、多項ディリクレモデルを用いる。なお初期値をランダムに変えていると、そのうち収束することもある。正確に判断するには、尤度の等高線図を見れば良い。
   - 狭義の Pitman モデルの最尤推定は以下のように行えば良い。
     (a) 対数尤度関数は (10) 式で与えられている。

(b) 最尤推定量は以下の同時方程式の解である.

$$\frac{\partial L(\alpha,\theta)}{\partial \theta} = \sum_{i=1}^{u-1} \frac{1}{\theta + i\alpha} - \sum_{i=1}^{n-1} \frac{1}{\theta + i} = 0,$$

$$\frac{\partial L(\alpha,\theta)}{\partial \alpha} = \sum_{i=1}^{u-1} \frac{i}{\theta + i\alpha} - \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \frac{1}{j - \alpha} = 0.$$

(c) $L(\alpha,\theta)$ の最大化は汎用最大化ルーチン (R の optim() 関数等) に任せても良いだろう。

(d) 最尤推定値を自前で評価するなら、二次の微分係数 (11),(12),(13) 式を用いたニュートン＝ラフソン法が適当である。

(e) $c = s_1(s_1 - 1)/s_2$ として、以下の近似的なモメント推定量を得る。これらをニュートン＝ラフソン法の初期値として使うことが考えられる。

$$\hat{\theta} = \frac{nuc - s_1(n-1)(2u+c)}{2s_1u + s_1c - nc}, \quad \hat{\alpha} = \frac{\hat{\theta}(s_1 - n) + (n-1)s_1}{nu},$$

- 多項ディリクレモデルの最尤推定は以下のように行えば良い。

(a) 対数尤度関数は定数を除いて

$$L(\gamma) = -\sum_{i=0}^{n-1} \log(J\gamma + i) + \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1} \log(\gamma + j).$$

(b) 最尤推定値は尤度方程式

$$\frac{dL(\gamma)}{d\gamma} = -\sum_{i=0}^{n-1} \frac{J}{J\gamma + i} + \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1} \frac{1}{\gamma + j} = 0$$

の解である。

(c) $L(\gamma)$ の最大化は汎用最大化ルーチン (R の optimize() 関数等) に任せても良いだろう。

(d) 最尤推定値を自前で評価するなら、二次の微分係数

$$\frac{d^2L(\gamma)}{d\gamma^2} = \sum_{i=0}^{n-1} \frac{J^2}{(J\gamma + i)^2} - \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1} \frac{1}{(\gamma + j)^2}$$

を用いたニュートン＝ラフソン法が適当である。

(e) 尤度関数は単峰であり、それほど初期値に依存せず最大化が可能である。ただ最尤推定値が無限大に発散する事はあり得て、それは確率関数が等確率 J 項分布である事を意味する。また最尤推定値が 0 の場合、狭義の Pitman モデルの方が適切と思われる。

- 狭義の Pitman モデルと多項ディリクレモデルの境界 ($\alpha = 0$) のモデルを Ewens モデルという。Ewens モデルの確率関数は以下の通り：

$$p(s_1, s_2, \ldots, s_n) = n! \frac{\theta^u}{\theta^{[n]}} \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{s_j} \frac{1}{s_j!}. \tag{16}$$

  - 同じデータについて Ewens モデルの最尤推定値を $\hat{\theta}_E$ と書き、Pitman モデルの最尤推定値を $(\hat{\alpha}, \hat{\theta}_P)$ と書く。もし $\hat{\alpha} > 0$ ならば $\hat{\theta}_E > \hat{\theta}_P$.
  - 上の結果は Pitman モデルのチェックに使える。また最尤推定の繰り返し計算の範囲を限定できる。
  - Ewens モデルの尤度関数は単峰であり、最大化は容易である。

5. 同定されたデータ構造の下で母集団一意数の推定値 $\hat{S}_1$ を求める。

(a) 狭義の Pitman モデルの場合、母数の最尤推定値を $\hat{\alpha}, \hat{\theta}$ と書けば (9) で推定される。

(b) 多項ディリクレモデルの場合、母数の最尤推定値を $\hat{\gamma}$ と書けば

$$\hat{S}_1 = \tilde{n}(J-1)\hat{\gamma} \frac{((J-1)\hat{\gamma}+1)((J-1)\hat{\gamma}+2)\cdots((J-1)\hat{\gamma}+\tilde{n}-2)}{(J\hat{\gamma}+1)(J\hat{\gamma}+2)\cdots(J\hat{\gamma}+\tilde{n}-1)}.$$

- これらの推定値はモデルの下での度数 1 のセル数の期待値である。
- 注 1）Ewens モデルの母集団一意数推定式は、Pitman モデルの推定式に $\hat{\alpha} = 0$ を代入して得られる。
- 注 2) 等確率 J 項分布の母集団一意数推定値は $\tilde{n}(1 - 1/J)^{\tilde{n}-1}$ である。

参考文献

[1] Charalambides, C.A. and Singh, J. (1988) A Review of the Stirling Numbers, Their Generalizations and Statistical Applications. *Communications in Statistics*, Thoer. Meth., **17**, 2533–2595.

[2] Dale, A. and Elliot, M. (2001) Proposal for 2001 Samples of Anonymized Records: An Assessment of Disclosure Risk. *Journal of the Royal Statistical Society*, Series A, **164**, 427–447.

[3] Domingo-Ferrer, J. and Torra, V. (2001) A Quantitative Comparison of Disclosure Control Methods for Microdata. *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Doyle et al. (Eds.), Elsevier, Amsterdam, 111-133.

[4] Duncan, G., Keller-McNulty, S.A. and Stokes, S.L. (2001) Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 121, National Institute of Statistical Sciences, Durham, North Carolina.

[5] Elliot, M. J., Skinner, C. J., and Dale, A. (1998) Special Uniques, Random Uniques, and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk. *Research in Official Statistics*, **1**, 53–67.

[6] Elliot, M., Lomax, S., Mackey, E. and Purdam, K. (2010) Data Environment Analysis and the Key Variable Mapping System. *Privacy in Statistical Databases*, Domingo-Ferrer, J. and Magkos, E. (Eds.), LNCS 6344, 138–147, Springer-Verlag, Berlin Heidelberg.

[7] Elliot, M., Mackey, E. and Purdam, K. (2011) Formalizing the Selection of Key Variables in Disclosure Risk. *Int. Statistical Inst.: Proceedings of the 58th World Statistical Congress*, 2777–2784.

[8] Hoshino, N. (2001) Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment, *Journal of Official Statistics*, **17**, 499–520.

[9] 星野伸明 (2003) 「超母集団モデルによる個票開示リスク評価」, 統計数理, **51**, 297–319.

[10] Hoshino, N. (2009) The Quasi-multinomial Distribution as a Tool for Disclosure Risk Assessment, *Journal of Official Statistics*, **25**, 269–291.

[11] 星野伸明 (2010)「公的統計ミクロデータ提供制度の課題」, 日本統計学会誌, **40**, 23–45.

[12] 星野伸明 (2012)「公的統計の開示リスク評価—労働力調査の論点」, 『経済統計・政府統計の数理的基礎と応用-I 』, 国友直人・山本拓共編, CIRJE 研究報告書シリーズ, CIRJE-R-10, 40–56.

[13] Hoshino, N. (2012) On the Marginals of a Random Partitioning Distribution. 研究集会「数理統計学の沃野」予稿集, 78–86.

[14] 伊藤伸介 (2012) 「政府統計ミクロデータの提供における匿名化措置—イギリス統計法における法制度的措置と攪乱的手法の適用可能性を中心に—」, 明海大学経済学論集, **24**, 1–14.

[15] 伊藤伸介・磯部祥子・秋山裕美 (2009) 「秘匿性の評価方法に関する実証研究—全国消費実態調査のミクロアグリゲートデータを用いて—」, 統計センター製表技術参考資料, **11**, 12–14.

[16] Marsh, C., Skinner, C., Arber, S., Penhale, P., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991) The Case for a Sample of Anonymized Records from the 1991 Census. *Journal of the Royal Statistical Society*, Series A, **154**, 305–340.

[17] Paass, G. (1988) Disclosure Risk and Disclosure Avoidance for Microdata. *Journal of Business and Economic Statistics*, **6**, 487–500.

[18] Pitman, J. (1995) Exchangeable and Partially Exchangeable Random Partitions. *Probability Theory and Related Fields*, **102**, 145–158.

[19] Sibuya, M. and Yamato, H. (2001) Pitman's Model of Random Partitions. 数理解析研究所講究録, **1240**, 64–73.

[20] 総務省政策統括官（統計基準担当）(2011). 「匿名データの作成・提供に係るガイドライン（平成 23 年 3 月 28 日改正版）」

[21] Sweeney, L. (2002) k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, **10**, 557–570.

[22] 竹村彰通 (1997) 「個票データ開示の理論」, 科学研究費補助金（課題番号 08209102）報告書, 2–25.

[23] U.S. Office of Federal Statistical Policy and Standards (1978). *Report on Statistical Disclosure and Disclosure Avoidance Techniques.* Statistical Policy Working Paper 2, U.S. Department of Commerce, Washington DC.

[24] Yamato, H. and Sibuya, M. (2000). Moments of Some Statistics of Pitman Sampling Formula. *Bulletin of Informatics and Cybernetics*, **32**, 1–10.