【論文】

社会生活基本調査ミクロデータにおける平日平均統計量と標本誤差の計測

栗原由紀子*

要旨

社会生活基本調査では，平日特性の代表値として平日平均統計量が多用される。これは，曜日別統計量の単純平均として定義される。しかし，ミクロデータから平日平均統計量とその標本誤差まで算出しようとすると，いくつかの問題点がある。まず，秘匿化のために標本設計情報が一部削除されていることから，提供された情報のみで推定せざるをえない。さらに，2日間固定標本に起因する共分散が存在するため，そのための推定プログラムを独自に作成する必要があり，時間消耗的な作業に労力を費やすことになる。

本稿では，平日平均統計量について，ミクロデータにおける算出方法と推定精度を検討し，平日サンプルについてプールして処理するための調整ウェイトを導入することで，効率的かつ実際的な処理方法を提案した。

キーワード

社会生活基本調査，ミクロデータ，固定標本，標本誤差，ウェイト調整

はじめに

統計法改正によりミクロデータの利用の促進が制度化され，そのためのサテライト機関の設立など環境整備が整いつつある。現在，提供される調査データの種類は拡張されつつあるが，主要な提供データの一つに社会生活基本調査（以下，社会調とも呼ぶ）がある。これは生活時間調査として，特定の時間区分（15分単位）での1日24時間の生活行動を記録することで，人々の生活や行動を時間配分という側面から捕捉しようとするものである。本調査データは個票の80％抽出標本が匿名化処理されて提供されており，これまでの公表集計値だけには縛られない，多様な生活時間分析を可能にするものと期待される。

* 中央大学大学院経済学研究科
〒192-0393　東京都八王子市東中野742-1
ebaku24@gmail.com

しかしながら，ミクロデータの利用においては，匿名化処理に伴う情報損失を含めて，調査データに固有の技術的な問題が浮上する。調査期間を1週間とする社会調の場合，基本統計量である平日平均の算出と標本誤差推定にそれは端的に現れる。特に後者については，標本設計情報が一部秘匿されているため，利用可能な情報にもとづく分散推定量についての考察と処理が不可欠となる。

社会生活基本調査に代表される生活時間調査で多く用いられる分析指標として，総平均時間，行動者平均時間，および行動者率がある。1週間を調査期間とする場合，曜日ごとにそれらの行動特性や時間特性を把握することが可能であるが，通常，平日については月曜から金曜までの各曜日別に算出した統計量の単純平均が平日特性の代表値，すなわち平日平均統計量として利用される。本稿は，こ

の平日平均統計量について，ミクロデータにおける算出方法と推定精度を検討し，効率的かつ実際的な処理方法を提案したものである。

まず，1節では，平日平均統計量と，平日分についてプールしたサンプルによる平日平均（プール平均）との関係について簡単に整理する。2節では社会生活基本調査ミクロデータを用いたときの平日平均とその分散推定量を理論的に示す。3節は，プール平均を平日平均に一致させるための調整ウェイトを提示したうえで，これを用いた平日平均推定量とその分散推定量について議論する。しかし，分散推定量の理論値の算出はミクロデータ・ユーザーにとって負荷が大きいため，4節で実際的な代替分散推定量を提示し，5節ではそれらの利用可能性についてミクロデータを利用して検証する。

1．平日平均とプール平均

生活時間データの基本統計量である平日平均統計量は，社会生活基本調査では月曜から金曜までの平日の統計量についての5日分の単純平均として定義される。以下では，リサンプリングデータを前提に，代表的な統計量として総平均時間を例にとり，平日平均についての標本統計量を定義しておこう。

曜日を$h=1, \cdots, 5$（月－金），ケースの一連番号（曜日別）を$j=1, \cdots, J_h$，曜日hでの個人jの1日のある行動の総時間量をy_{hj}とする。いまミクロデータの抽出ウェイト（復元抽出であれば抽出率の逆数など）をw_{hj}としたとき，個人jが代表する母集団の総時間量は，ウェイト付きの$\check{y}_{hj}=w_{hj}\times y_{hj}$として書ける。また曜日別の推定母人口は$\hat{M}_h=\sum_{j=1}^{J_h} w_{hj}$であり，ミクロデータによる母人口の推定量は，曜日ごとに異なるものと仮定する（$\hat{M}_h \neq \hat{M}_{h'}$）。このとき，曜日別の総平均時間は

$$\hat{\mu}_h=\frac{\sum_{j=1}^{J_h}\check{y}_{hj}}{\hat{M}_h} \qquad (h=1, \cdots, 5)$$

であり，その平日平均統計量（Mean of Weekday，以下MW統計量とする）は次のように書ける[1)]。

$$\hat{\mu}_+^{MW}=\frac{1}{5}\sum_{h=1}^{5}\hat{\mu}_h \qquad (1.1)$$

これは，各曜日の統計量についてすべての曜日の重みを1とする単純平均であり，曜日の水準変化の代表値という機能を果たす。定義式から明らかなように，その計算過程において各曜日の統計量を算出したうえで，さらにその平均を計算するという2段階の作業が必要となり，基本統計量の確認といった目的にさえ，若干煩雑な作業がつきまとう[2)]。

これに対して，調査日が平日であった標本をプールして，そのまま算出する平日平均（プール平均，Mean with the Pooled Data，以下MP統計量とする）がある。平日を通した曜日変動を含む総平均時間の特性を知りたければ，このような平均をとればよい。これは，次のように定義される。

$$\hat{\mu}_+^{MP}=\frac{\sum_{h=1}^{5}\sum_{j=1}^{J_h}\check{y}_{hj}}{\sum_{h=1}^{5}\hat{M}_h} \qquad (1.2)$$

MW統計量に対して，各曜日の重みが異なる加重平均を算出していることになる。明らかに，各曜日の推定母人口が異なるとき（$\hat{M}_h \neq \hat{M}_{h'}$），MW統計量とMP統計量は理論的に一致しない（$\hat{\mu}_+^{MW} \neq \hat{\mu}_+^{MP}$）。その主な原因は，$\hat{\mu}_+^{MW}$では各曜日とも等ウェイトで推定人口の曜日間の差は除かれているのに対して，$\hat{\mu}_+^{MP}$では推定人口の曜日変動分が含まれていることにある[3)]。

そしてMP統計量は，MW統計量の母数（真値）に対してバイアス[4)]をもつが，各曜日統計量を算出し，さらにそれらの平均をとるといったMW統計量の計算ステップに比べて，処理が単純であるという長所をもつ。

2．社会生活基本調査ミクロデータにおけるMW統計量とその理論分散

社会生活基本調査の標本設計は，層化二段抽出法に基づくものであり，第一次抽出単位

には地域（47都道府県）を層とした国勢調査区，第二次抽出単位には世帯を抽出している。さらに，社会調では2日間連続で調査を行うため，調査区をランダムに8区分し，区分された標本をそれぞれ異なる曜日の調査に当てている[5)]。このような標本設計の下で曜日ごとに抽出ウェイトが計算され，世帯単位で80％リサンプリングされた標本がミクロデータとして提供される。ミクロデータに付与される抽出ウェイトは乗率とも呼ばれ，本来調査の抽出ウェイトをリサンプリング率80％で調整すればよい。

いずれにしても，このようなウェイトを使えば，ミクロデータにおいても母集団特性値を推定でき，またその推定誤差（推定量の分散や標準誤差）も求めることができるはずである[6)]。社会調の場合，曜日別にある程度人口数が調整された抽出ウェイトが付与されていることから，各曜日については母集団特性値が比較的容易に求められる。しかしながら，既述のように平日平均は曜日横断的な統計量であることから，MW統計量の分散（あるいは標準誤差）を含めて推定に当たっては特有の工夫が必要となる。その主な原因のひとつは，2日間の固定標本方式で調査されるため，曜日間で相関が生じていることにある。それに加え，ミクロデータでの推定においては，層化変数である地域や第一次抽出単位である国勢調査区に関する情報が削除されており，このことが問題をさらに複雑にする。つまりミクロデータから平日平均統計量の分散を推定する際に利用可能な標本設計情報は，世帯の識別変数とウェイトだけとなる。このような社会調の標本設計とリサンプリングに関する情報は表1のようにまとめられる。

曜日を$h=1,\cdots,7$（$\sum_{h=1}^{\alpha}1=\alpha$，平日平均のとき$\alpha=5$），世帯の一連番号を$i=1,\cdots,m_h$，世帯員番号を$j=1,\cdots,n_{hi}$とする。世帯主など世帯の代表者を示す世帯代表ダミーを$\eta_{hij}=\{0,1\}$，世帯員ダミーを$\gamma_{hij}=1$とおいて，地域（都道府県）・男女・年齢別を示すダミー変数をδ^*，またδ^*に属する基準人口[7)]を$B_{h\delta^*}$とする。リサンプリングの世帯の抽出率を$f^{re}=4/5$，曜日別での世帯単位の線形推定用乗率をw_{hkgi}，この乗率w_{hkgi}を用いた属性δ^*の人口を$\hat{N}_{h\delta^*+}$としたとき，ミクロデータに付与される抽出ウェイトは(2.1)として表すことができる。ここで，シャープ(♯)は秘匿処理のために情報が一部削除されたデータセットの変数，もしくはこれを利用した統計量であることを意味する。

$$w_{hij}^{\sharp}=\lambda_{hij(j\in\delta^*)}\cdot w_{hkgi}\cdot\left(\frac{1}{f^{re}}\right) \qquad (2.1)$$

$$\text{ただし}\quad \lambda_{hij(j\in\delta^*)}=\frac{B_{h\delta^*}}{\hat{N}_{h\delta^*+}}=\frac{B_{h\delta^*}}{\Sigma_{kgij(j\in h)}w_{hkgi}\delta^*}$$

このとき，ある行動に関する1日の行動時間の総計を改めてy_{hij}とおけば，ウェイトで膨らませた曜日別人口と曜日別総時間量の推定量は，

$$\hat{N}_{h+}^{\sharp}=\Sigma_{ij}w_{hij}^{\sharp}\gamma_{hij}$$
$$\hat{Y}_{h+}^{\sharp}=\Sigma_{ij}w_{hij}^{\sharp}y_{hij}$$

となる。したがって，MW統計量は，

$$\tilde{\mu}_{+}^{\sharp}=\frac{1}{\alpha}\Sigma_h\frac{\hat{Y}_{h+}^{\sharp}}{\hat{N}_{h+}^{\sharp}} \qquad (2.2)$$

と書ける。以下ではMW統計量をチルダ(~)付きで表すことにする。

それではミクロデータにおいて，MW統計量の推定誤差はどのように評価すればよいのであろうか。すでに述べたように，社会調は層化二段抽出でデザインされており，本来の推定誤差の算出には層化情報（地域）と2つのクラスター情報（調査区と世帯）[8)]を必要とするが，秘匿処理のためミクロデータでは地域と調査区情報が削除され，世帯情報（世帯の一連番号）しか残されていない。このような状況では，ミクロデータの枠組みの中で忠実に推定量の分散を計算しておき，それを評価の目安とするしかない。いまの場合，ミクロデータは世帯クラスター(i)を無作為抽出した結果として，いわば集落抽出したかの

表１　社会調（2001年）の標本設計およびリサンプリングの基本情報

	層	抽出単位と関連事項	抽出率
第一次抽出単位	地域：$k=1\cdots K$	調査区（1995年国勢調査区）：$g=1\cdots G_k$ G_k：第k地域の標本調査区数	確率比例抽出：$f^1_{kg}=\frac{G_k C_{kg}}{C_k}$ C_{kg}：第k地域，第g調査区の国勢調査人口 C_k：第k地域の国勢調査人口
第二次抽出単位		世帯：$i=1\cdots m_{kg}$ m_{kg}：第k地域，第g調査区の標本世帯数	無作為抽出：$f^2_{kg}=\frac{m_{kg}}{M_{kg}}$ M_{kg}：第k地域，第g調査区の世帯数
調査日割当のための再抽出	調査グループ：$q=1\cdots Q(Q=8)$	調査区：$g=1\cdots G_{qk}$（$=G_{hk}$）	調査区を8区分するときの抽出率（無作為抽出）：$f^3_q=1/8$ ※各曜日の抽出率：$f^3_{h(=1-4)}=f^3_q$，$f^3_{h(=5)}=2f^3_q$，$f^3_{h(=6,7)}=5f^3_q$
［ミクロデータのリサンプリング］			
抽出単位	−	世帯	無作為抽出　$f^{re}\cong 4/5$
［リサンプリング後のデータとウェイト］			
	曜日：$h=1\cdots L$	世帯：$i=1\cdots m_h$ 世帯代表ダミー：$\eta_{hij}=\{0,1\}$ 世帯員：$j=1,\cdots n_{hi}$ 世帯員ダミー：$\gamma_{hij}=1$	ウェイト：$w^{\#}_{hij}=\lambda_{hij}\cdot w_{hkgi}\cdot(1/f^{re})$ $\lambda_{hij(j\in\delta^*)}=\frac{B_{h\delta^*}}{\hat{N}_{h\delta^*+}}$，　$w_{hkgi}=\frac{1}{f^1_{kg}f^2_{kg}f^3_h}r_{kg}$ δ^*：地域・性別・年齢識別ダミー $B_{h\delta^*}$：調査グループの基本人口 r_{kg}：第k地域，第g調査区の修正項

注：本表は総務省統計局（2003，pp.911-913）をもとに独自に作成した。なお，リサンプリング後の世帯数m_hはリサンプリング前の標本世帯数m_{kg}の約8割に減少している。また，ウェイト内のλ_{hij}・w_{hkgi}はリサンプリング前のデータで比推定用乗率として作成されたものである。

ように仮定して分析を進めることになる。このように求めたものを推定量の本来的な分散と区別して，以下では世帯クラスター（Household Cluster）分散（HC分散）と呼ぶことにする。

一般に，層化二段抽出での推定誤差は，層化による縮小効果の下で，第一次抽出単位（調査区）の分散と第二次抽出単位（世帯）の分散の和として概念的には捉えられる。これを上記のように集落抽出と想定したときのデザイン（標本設計）の誤った特定によるバイアスについて，実際のミクロデータから定量的にその近似度を評価することは困難である。他方で，世帯の識別変数情報さえも無視して，個人単位の単純無作為抽出という想定で推定量の分散を簡易計算することも可能ではあるが，これではいわば2重にデザインバイアスを重ねることになる。すなわち本来の標本設計である層化二段抽出を集落抽出とみなさざるを得ないミクロデータ固有の歪みに，さらにミクロデータから世帯という標本設計情報を捨て去る歪みが加わり，誤差評価の理論的解釈はさらに曖昧となる[9)]。このように，ミクロデータの情報形式に忠実な，一種の疑似的な分散推定という方針が現在取り得る最良の選択肢と考えると，ミクロデータから算

出できるMW統計量のHC分散の推定量は次のように書ける[10),11)]。

$$\hat{V}(\bar{\mu}_{+}^{\#}) = \frac{1}{\alpha^2}\Big[\Sigma_h \hat{V}(\hat{\mu}_{h+}^{\#}) + 2\Sigma_q \widehat{Cov}(\hat{\mu}_{h(\in q)}^{\#}, \hat{\mu}_{h'(\in q)}^{\#})\Big] \quad (2.3)$$

ただし

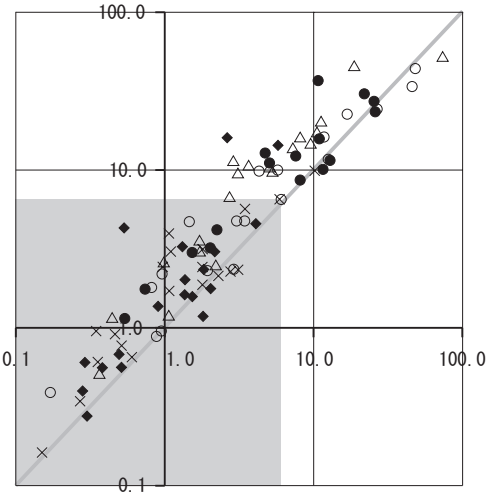
$$\hat{V}(\hat{\mu}_{h+}^{\#}) = \frac{m_{h+}}{m_{h+}-1}\cdot\frac{1}{\hat{N}_{h+}^{\#\,2}} \cdot\Sigma_{i\in h}\Big[\Sigma_{j\in i} w_{hij}^{\#}(y_{hij}-\hat{\mu}_{h+}^{\#})\Big]^2 \quad (2.4)$$

$$\widehat{Cov}(\hat{\mu}_{h(\in q)}^{\#}, \hat{\mu}_{h'(\in q)}^{\#}) = \frac{m_{q+}}{m_{q+}-1}\cdot\frac{1}{\hat{N}_{h+}^{\#}\hat{N}_{h'+}^{\#}} \cdot\Sigma_{i\in h}\Big\{\Big[\Sigma_{j\in i} w_{hij}^{\#}(y_{hij}-\hat{\mu}_{h}^{\#})\Big] \cdot\Big[\Sigma_{j\in i} w_{hij}^{\#}(y_{hij}-\hat{\mu}_{h'}^{\#})\Big]\Big\} \quad (2.5)$$

分散式(2.3)から明らかなように，MW統計量は各曜日の平均値の分散と曜日間の平均値の共分散から構成されている。曜日ごとに独立な標本であれば共分散は考慮しなくともよいが，すでに触れたように，標本が火水と木金で固定されているため，その間の共分散(2.5)も平日平均の分散に影響することになる[12)]。

本節最後に，本来の層化二段抽出による平日平均の推定誤差と，ミクロデータに対して集落抽出と想定したHC分散による推定誤差との数値的な近似の度合いについて触れておこう。前者については，社会生活基本調査報告書（以下，報告書）に標準誤差率が掲載されており，これは調査区を抽出単位とした副標本法に基づく推定結果である。後者については，ミクロデータからHC分散(2.3)を求め標準誤差率を計算したものである（付表1）。これらを用い，図1には報告書の数値とこれに対応する(2.3)式による誤差率との散布図を描いている。

データが45°線に沿って分布していることから(2.3)式による近似が報告書の標準誤差率の特徴を比較的よくトレースしていること，またHC分散による誤差率が45°線の上部に分布していることから報告書の数値より大きめに誤差を見積もっていることがわかる[13)]。とはいえ，いずれも標準誤差率で数パーセント程度のレベルが平日平均の推定値として有効と考えるならば，報告書の数値がその範囲にあるものは(2.3)式による誤差率もほぼ同じレベルに収まっており，HC分散でもミクロデータによる推定値を適切に評価できることが示されている（網掛け内）。この意味において，ミクロデータに対して集落抽出とみなして求めた分散推定量は，社会調本来の層化二段抽出による分散推定量の実際的な近似を与えるものと考えられる。



◆総数 × 男 ○ 男(10－14歳) △ 男(40－44歳) ● 男(80歳－)

注：付表1（平日）の副標本法（調査区抽出）の数値とHC分散の数値を用い，対数軸（底10）を使用して作成している。網掛け箇所は標準誤差率5％以内の領域である。

図1　副標本法（調査区抽出）とHC分散の散布図（標準誤差率：％）

3．調整ウェイトを利用したMP統計量とその分散推定

探索的に分析対象の基本統計量を算出する

とき，各曜日の統計量を算出したうえでその単純平均を計算し，さらには副標本法などでその分散推定量を計算するといった作業の繰り返しではミクロデータの長所は半減される。新統計法の下でのミクロデータ提供は，利用者による分析の自由度を大幅に高めるはずのものだからである。できれば，平日平均統計量の算出とともに，その分散推定量も同時に得られるようなプロセスが望ましい[14]。そこで，作業効率の改善を図るために，MP統計量をMW統計量と一致させるようにウェイトを調整し，平日平均の算出を容易にするとともに，これを用いてMP統計量のHC分散を推定するアプローチが考えられる[15]。

MP統計量がMW統計量に対してバイアスをもつ要因は，曜日間で推定人口が変動することにあるから，ウェイトを曜日間で不変となるように定義すればよい。これを調整ウェイトと呼ぶことにする。

$$v^{\sharp}_{hij}=\frac{w^{\sharp}_{hij}}{\hat{N}_{h+}} \tag{3.1}$$

ただし，$\Sigma_{ij(j\in h)}v^{\sharp}_{hij}=1$，$\Sigma_{ij(j\in h)}w^{\sharp}_{hij}=\hat{N}_{h+}$とする。これは，各曜日のウェイト合計が1となるように調整したものであり，いわば調査日に関して標本設計を事後的に再構成したと考えればよい。なお，調整ウェイトは分析に利用する変数や部分母集団を考慮して，分析の都度作成する必要があるが，MW統計量の計算手順に比べれば極めて容易である。

調整ウェイトを用いたMP統計量を改めて$\hat{\mu}^{v\sharp}_{+}$と表すことにすれば，これは目標であるMW統計量$\tilde{\mu}^{\sharp}_{+}$と当然一致する。以下では，調整ウェイトを用いた推定量にはv♯を付している。

$$\hat{\mu}^{v\sharp}_{+}=\frac{\Sigma_{hij}v^{\sharp}_{hij}y_{hij}}{\Sigma_{hij}v^{\sharp}_{hij}y_{hij}}=\frac{1}{\alpha}\Sigma_h\frac{\hat{Y}^{\sharp}_{h+}}{\hat{N}_{h+}}=\tilde{\mu}^{\sharp}_{+} \tag{3.2}$$

さらに，調整ウェイトを用いたMP統計量のHC分散の推定量は，各曜日の平均値の分散と，曜日間の共分散の和で示されるが，それはMW統計量のHC分散の推定量(2.3)式と一致する。

$$\begin{aligned}\hat{V}(\hat{\mu}^{v\sharp}_{+})&=\frac{1}{\alpha^2}\Big[\Sigma_h\hat{V}(\hat{\mu}^{v\sharp}_{h+})\\&\quad+2\Sigma_q\widehat{Cov}(\hat{\mu}^{v\sharp}_{h(\in q)},\hat{\mu}^{v\sharp}_{h'(\in q)})\Big]\\&=\hat{V}(\tilde{\mu}^{\sharp}_{+})\end{aligned} \tag{3.3}$$

ただし

$\hat{V}(\hat{\mu}^{v\sharp}_{h+})=\hat{V}(\hat{\mu}^{\sharp}_{h+})$

$\widehat{Cov}(\hat{\mu}^{v\sharp}_{h(\in q)},\hat{\mu}^{v\sharp}_{h'(\in q)})=\widehat{Cov}(\hat{\mu}^{\sharp}_{h(\in q)},\hat{\mu}^{\sharp}_{h'(\in q)})$

このように，調整ウェイトを利用することで，平日に該当する曜日データをプールした標本に対して，通常の平均を求める作業でMW統計量と同値の平日平均統計量を算出することができる。ただし，分散推定量については，その一部を構成する共分散の算出に独自のプログラムを作成する必要があり，実は作業負荷はさほど軽減されない。(2.3)あるいは(3.3)のHC分散の推定には別の角度から検討を加えなければならない。

4．MP統計量の分散推定のための代替アプローチ

調整ウェイトを利用しても，平日平均統計量のHC分散の推定については，2日間固定標本による共分散が存在するため，作業プロセスの軽減は原理的に期待できない。自然な帰結として，HC分散の値とは完全に一致しなくとも，それをよく近似する代替的な分散推定量を利用する方法が考えられる。そしてこのような分散推定量は，平日平均統計量を計測する過程の延長上で算出できることが作業上望ましい。そのためには平日サンプルをプールし，調整ウェイトを用いることで算出可能な統計量が代替分散の候補となる。

このような分散推定量の主な候補として，以下の4種類の推定量が考えられる。その特性をまずは簡単に整理しておくことにしよう[16]。

(i) 平日プールサンプルに対する世帯クラスターの単純無作為抽出（HP：Household cluster for the pooled data，以下HP近似と呼ぶ）

世帯をクラスター単位に単純無作為抽出したものとして，標本調査データの平均値の分散推定量を算出する。この場合，世帯の識別変数（世帯の一連番号）も計算に利用するため，(3.3)式の共分散に相当する分散も計測される。

$$\hat{V}(\hat{\mu}_{+}^{v\sharp})_{HC}=\frac{1}{\alpha^2}\cdot\frac{m_{+}}{m_{+}-1}\cdot\Sigma_i\left[\Sigma_{hj\in i}\frac{w_{hij}^{\sharp}}{\hat{N}_{h+}^{\sharp}}(y_{hij}-\hat{\mu}_{+}^{v\sharp})\right]^2 \qquad (4.1)$$

(ii) 平日プールサンプルに対する曜日層化，世帯クラスターの単純無作為抽出（SHP：Household cluster stratified by days for the pooled data，以下SHP近似と呼ぶ）

(i)と同種であるが，曜日をさらに層化情報として用いたときの計算式である。そのため世帯変数は曜日ごとに切り離されてしまい，実際にはHC分散の共分散部分をゼロとおいた分散を計測していることになる。また，これには曜日別統計量$\hat{\mu}_{h+}^{v\sharp}$の分散に相当する部分も含まれる。

$$\hat{V}(\hat{\mu}_{+}^{v\sharp})_{SHC}=\frac{1}{\alpha^2}\Sigma_h\frac{m_h}{m_h-1}\cdot\Sigma_{i\in h}\left[\Sigma_{j\in i}\frac{w_{hij}^{\sharp}}{\hat{N}_{h+}^{\sharp}}(y_{hij}-\hat{\mu}_{+}^{v\sharp})\right]^2-\frac{1}{\alpha^2}\Sigma_h\frac{1}{m_h-1}(\hat{\mu}_{h+}^{v\sharp}-\hat{\mu}_{+}^{v\sharp})^2 \qquad (4.2)$$

(iii) 副標本法（RG：Random groups）[17]

一般に，調査法や推定すべき統計量が複雑である場合に用いられ，総務省統計局(2003)に掲載されている標準誤差率も副標本法によるものである。調査区情報がないミクロデータの場合，分析対象の個人または世帯をランダムにk区分し(k=4)，それぞれ4グループの統計量と，対象とする全標本データによる統計量を用いて分散推定量を定義することになる。サンプルサイズの小さい部分母集団を対象とした場合には，その推定量は副標本の取り方に大きく左右される欠点をもつ。

$$\hat{V}(\hat{\mu}_{+}^{\sharp})_{RG}=\frac{1}{(4-1)}\Sigma_k(\hat{\mu}_{k}^{\sharp}-\hat{\mu}_{+}^{\sharp})^2$$

$$\widehat{SE}(\hat{\mu}_{+}^{\sharp})=\sqrt{\hat{V}(\hat{\mu}_{+}^{\sharp})_{RG}/4} \qquad (4.3)$$

(iv) ジャックナイフ法（JK：The delete-one jackknife）[18]

標本要素1個を除いた推定を，すべての標本要素について繰り返し，それらの推定値に基づく分散推定値である。除外する標本単位を世帯とすれば，近似的に2日連続調査による共分散の影響を含む推定量が得られる。推定時には，i番目の世帯のウェイトをゼロ，それ以外の世帯の調整ウェイトを$\frac{m_{+}}{(m_{+}-1)}v_{hij}^{\sharp}$とする。このウェイトを用いて，i番目の世帯を除いた推定値を$\hat{\mu}_{i}^{\sharp}$，この推定値について全てのiに関する平均を$\bar{\hat{\mu}}=\Sigma_i^m\hat{\mu}_i^{\sharp}/n$とすれば，分散推定量は以下のように算出される。

$$\hat{V}(\hat{\mu}_{+}^{v\sharp})_{JK}=\frac{n-1}{n}\Sigma_i^m(\hat{\mu}_{i}^{\sharp}-\bar{\hat{\mu}})^2 \qquad (4.4)$$

5．HC分散に対する代替分散の利用可能性

4種類の代替分散は，ミクロデータによる平日平均統計量のHC分散に対して，実際に利用可能なほどの近似値を与えてくれるのであろうか。その検証結果を図2に示している。これは社会生活基本調査ミクロデータ(2001年)から各分散推定値を算出し，標準誤差率[19]を整理したものである。縦軸が総平均時間の平日平均に関する標準誤差率，横軸が標本世帯数$m_{h\kappa+}$である。比較の対象とした変数は「休養･くつろぎ」の一日の合計時間(分)であり，また部分母集団κは「子ども」とした[20]。標本世帯数$m_{h\kappa+}$の増減による影響を

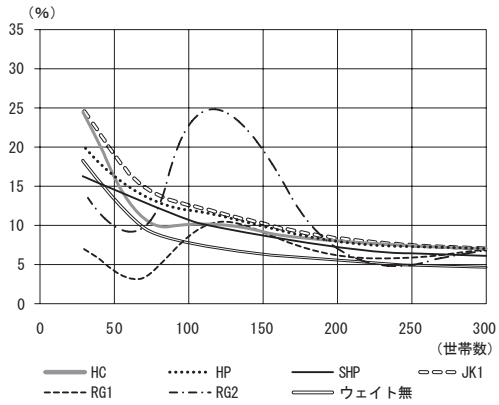


図２　平日平均に関する標準誤差率

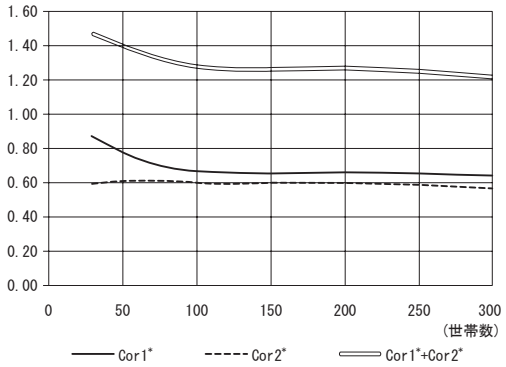


図３　曜日間の相関係数

測るため，最初に抽出した標本から次々と継続的にランダムに標本世帯を抽出することで，同一の部分母集団について世帯数のみを減少させたときの分散推定量の効果を計測した[21]。算出した分散推定量は，HC分散(3.3)，HP近似(4.1)，SHP近似(4.2)，副標本法(4.3)，ジャックナイフ法(4.4)であり，順に凡例の「HC」，「HP」，「SHP」，「RG1，RG2」，および「JK1」に対応している。なお，副標本法は抽出される標本に大きく依存するため，安定性などを考慮して，異なる副標本による2通りの推定値RG1とRG2を計算している。

図2が示すように，標本世帯数が200以上の場合には，HP近似およびジャックナイフ法はほぼ一致している。それ以外はすべて過小推定の傾向がみられる。しかし，標本サイズが大きければ，いずれの方法でも理論値の比較的近傍に位置することがわかる。また100から200世帯（1曜日当たり40－20世帯）の標本サイズであれば，HC分散の標準誤差率は先の例より若干上昇し，代替分散の推定値の過大・過小傾向が顕在化し始めている。その中でSHP近似だけはHC分散の直ぐ傍に位置している。これに対して，世帯数100から50の標本では過大に，50以下では過小に推定されるようになり，標本サイズが小さいとき各代替分散の過大・過小傾向も不規則に変動する。全体的にはジャックナイフ法，HP近似，およびSHP近似がHC分散の近傍値として代替可能であると考えてよい。

また副標本法では，標本サイズが小さい（200以下の）部分母集団を対象とするとき，副標本の取り方によってバイアスの方向も異なり安定した結果は得られない。なお，抽出ウェイトを頻度的に解釈して分散を計算することも考えられるが，これでは大幅な過小推定となる。その代わりに，ウェイトを使用しない（すべてのウェイト＝1）分散推定量の計算も考えられる。図2の「ウェイト無[22]」がその動きを表しいているが，いずれにしても過小推定の傾向にあり，標本サイズが小さいときにはとくに注意を要する。

共分散部分に影響を及ぼす曜日間の相関係数と標本世帯数の関係を図3に示している。これをみると，火水の2日連続調査グループ（Cor1*）は標本世帯数によらず相関係数0.6付近を推移しているが，木金の調査グループ（Cor2*）は標本世帯数100以下で相関係数が上昇している。標本サイズが小さく曜日間の相関が高い状況では，曜日間の共分散部分をゼロと仮定するSHP近似ではHC分散との大きな乖離が生じる危険性が伴う。

このように標本サイズが大きければ，HP近似およびジャックナイフ法による推定量が

よい近似を与えており，標本サイズが小さい場合にはHP近似，SHP近似およびジャックナイフ法での代用が考えられる。すなわち，標本サイズに依らずHP近似とジャックナイフ法のパフォーマンスが高いと言えるが，ジャックナイフ法での推定では標本サイズが大きい場合には計算に多くの時間が必要となる。これらの点を考慮したとき，どのような標本サイズでも効率的で安定的な概算値を提供するHP近似が代替分散の推定に適していると結論づけられる。

おわりに

社会調ミクロデータの平日平均の計測には，常に計算のための時間消耗的な作業に労力を費やすことになる。これを回避するため，データを平日に関してプールして平均値を算出しようとすれば，その推定値にはバイアスが生じる。これらの点を考慮した上で，社会調の標本設計方式に基づいて平日平均を算出するには，プール平均のバイアスを修正し，各曜日で推定母人口が全て1となるように調整したウェイトを用いることが考えられる。

また，その分散推定量を理論式に基づいて得ようとすれば，標本設計情報の一部秘匿や2日間固定標本に起因する共分散が存在するため，方法的には，そのための推定プログラムを独自に作成するしかない。しかし，理論式による推定値と大きな乖離がなければ，例えば世帯クラスターを想定した分散など，その他の分散推定量による近似推定も有効であり，これにより推定作業の簡便化と推定精度の評価が図られる。ただし，分析ごとに調整ウェイトを算出する作業は不可避であるが，これには調整ウェイトのプログラムを数行作成し，各分析の前に実行させるだけで十分であり，全体の作業効率は大幅に改善される。

社会生活基本調査は，綿密かつ効率的な標本設計に基づいて，大規模標本調査として実施されており，データの情報価値は極めて高い。そのミクロデータの二次利用においては，秘匿のために標本設計情報が一部制限されるが，提供された情報の積極的活用と推定方法の柔軟な工夫により，十分実用的な精度で必要とされる統計量を獲得できる。それには，政府統計レベルで調査・収集・作成された秘匿処理済みミクロデータについて，それぞれの特性を十分に加味した推定技法の検討とその蓄積が不可欠であるように思われる。

謝　辞

本稿では，平成13年社会生活基本調査（総務省統計局）の匿名データ（申請年度2009年，申請者：中央大学・坂田幸繁，共同利用者：栗原由紀子）の利用による分析を行った。本研究の結果数値は総務省統計局が作成・公表している統計量とは異なることを明記しておく。なお，秘匿処理済みデータの提供時には，総務省統計局，統計センター，および一橋大学社会情報研究所にはお世話になりました。記して感謝します。

付表１　属性・行動種類別総平均時間に対する標準誤差率の比較表

行動種類	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
［平日］																				
＝ 副標本（調査区）[a] ＝																				
総数	0.09	0.38	0.30	0.29	0.28	1.53	0.53	5.78	2.62	1.31	0.90	0.51	0.49	1.37	1.81	1.36	4.08	2.03	2.16	1.83
男	0.15	0.60	0.27	0.35	0.36	2.30	1.81	10.20	3.48	3.13	1.06	0.52	0.47	1.09	1.80	0.97	5.86	2.77	1.06	1.79
男（10－14歳）	0.17	0.82	0.89	0.96	46.23	0.95	11.89	146.93	49.03	12.57	6.14	2.93	1.94	1.47	3.04	3.46	17.10	5.76	27.25	4.32
男（40－44歳）	0.36	2.20	1.06	0.97	0.44	72.99	5.02	18.63	9.59	3.65	2.71	1.73	1.71	8.15	5.22	7.27	11.25	3.12	10.56	2.88
男（80歳－）	0.54	2.03	0.75	22.29	11.80	166.55	8.31	26.00	130.96	7.65	8.15	1.53	2.30	26.61	3.27	13.03	10.79	4.80	5.09	11.09
＝ HC分散[b] ＝																				
総数	0.12	0.56	0.28	0.61	0.40	1.59	4.31	14.37	16.04	3.29	1.37	0.56	0.68	2.03	1.19	1.62	4.59	1.78	3.05	2.36
男	0.16	0.65	0.35	0.96	0.61	2.12	2.42	9.92	5.70	2.35	1.72	0.78	0.91	3.01	1.88	2.44	6.58	2.29	3.96	3.14
男（10－14歳）	0.39	1.80	0.88	2.18	34.07	0.95	16.24	51.43	44.02	11.73	6.60	2.34	2.27	4.71	4.78	4.78	22.72	9.88	24.41	9.89
男（40－44歳）	0.51	2.51	1.21	2.60	1.17	52.35	10.50	46.09	14.75	10.69	6.71	3.07	3.60	16.18	9.84	13.75	20.33	9.58	17.43	11.34
男（80歳－）	1.15	3.24	1.77	30.37	10.13	81.17	8.70	27.51	76.31	12.40	8.67	3.01	4.25	23.42	9.87	11.68	37.16	12.75	10.96	15.77
＝ 副標本（世帯）[c] ＝																				
総数	0.16	0.35	0.36	0.51	0.46	1.47	0.59	8.02	2.94	0.88	0.79	0.58	0.49	1.38	1.13	1.81	3.84	0.48	1.08	2.37
男	0.23	0.32	0.24	0.71	0.41	1.71	1.99	10.71	4.93	2.54	0.86	0.89	0.55	3.74	0.55	1.64	5.51	2.20	4.24	3.26
男（10－14歳）	0.28	1.10	0.40	1.94	41.69	3.14	14.85	33.60	81.51	21.86	5.98	4.75	2.37	0.91	11.01	5.67	47.33	13.92	13.00	6.95
男（40－44歳）	0.66	1.18	0.75	1.64	1.89	46.74	14.09	53.67	7.62	13.81	3.67	2.29	5.07	19.49	7.92	17.47	36.25	15.61	29.27	11.00
男（80歳－）	1.41	3.30	2.50	26.19	13.72	85.57	6.85	22.66	83.92	15.77	13.06	3.01	1.64	27.43	5.00	15.39	41.58	15.19	14.14	11.15
［日曜］																				
＝ 副標本（調査区）[a] ＝																				
総数	0.09	0.36	0.40	0.50	1.06	3.41	0.50	2.69	2.74	1.14	0.77	0.43	0.98	1.21	1.22	1.79	3.39	1.61	4.25	1.67
男	0.08	0.31	0.37	1.24	1.16	1.72	2.27	7.00	2.37	1.97	1.37	0.29	0.62	1.94	1.07	1.28	2.14	1.92	8.55	0.56
男（10－14歳）	0.24	1.11	0.65	11.33	111.54	7.81	7.99	58.69	57.61	6.35	3.47	1.47	3.36	3.00	4.05	6.16	6.76	5.22	27.41	7.92
男（40－44歳）	0.24	0.59	1.00	4.60	3.98	74.90	8.36	16.86	6.14	2.15	1.53	2.25	2.55	6.16	3.24	3.51	6.92	3.98	22.56	3.16
男（80歳－）	1.05	2.99	1.44	35.26	7.74	－	9.24	28.85	46.60	12.80	6.49	1.49	3.29	2.53	7.14	5.09	14.87	5.64	5.23	12.26
＝ HC分散[b] ＝																				
総数	0.12	0.45	0.25	1.72	1.15	3.05	0.52	4.21	2.57	0.93	0.97	0.45	0.65	2.21	1.00	1.86	2.95	1.40	4.19	1.74
男	0.17	0.65	0.32	2.24	1.51	4.48	1.78	8.16	3.88	1.42	1.18	0.58	0.89	2.98	1.23	2.02	3.66	1.90	6.18	2.35
男（10－14歳）	0.40	1.98	1.05	10.35	60.57	6.32	9.92	78.31	40.58	5.52	3.83	1.96	3.29	6.05	3.58	3.88	14.40	6.91	39.54	10.74
男（40－44歳）	0.79	2.69	1.03	7.57	7.72	84.74	5.87	28.30	9.60	4.37	3.62	2.34	2.92	12.85	4.06	7.45	9.60	6.48	23.79	6.93
男（80歳－）	0.86	3.00	1.27	28.62	10.47	－	8.54	26.38	45.97	9.17	9.76	2.33	3.65	21.28	8.17	10.81	19.95	10.60	16.29	11.74
＝ 副標本（世帯）[c] ＝																				
総数	0.14	0.33	0.14	1.70	1.17	2.72	0.19	7.26	0.89	0.94	1.29	0.28	1.05	1.87	1.16	1.78	2.92	0.49	3.42	0.94
男	0.16	0.63	0.20	0.93	1.15	3.56	2.02	8.21	1.80	0.95	0.99	0.42	1.39	1.49	1.56	2.09	5.06	0.48	8.55	1.00
男（10－14歳）	0.32	1.30	0.55	3.63	55.39	3.63	14.91	64.18	31.70	4.77	5.43	0.96	4.99	2.35	5.54	2.34	12.69	2.21	28.55	6.79
男（40－44歳）	0.52	1.11	0.57	3.72	6.11	81.54	5.15	32.21	5.97	3.55	4.36	1.54	2.45	13.52	1.17	10.34	6.37	5.61	31.35	3.75
男（80歳－）	0.72	3.25	1.37	20.59	11.15	－	5.53	14.91	42.31	8.13	11.56	1.50	2.48	7.20	6.59	14.99	21.90	9.83	20.34	10.64

注：a）副標本（調査区）は，総務省統計局(2003) pp.800-805より抜粋したものであり，全調査データにより調査区を事後的に４区分し副標本法で算出した各行動種類別総平均時間に対する標準誤差率である。b）HC分散は80％抽出のミクロデータを用い第３節の分散式(3.3)に基づいて算出した標準誤差率である。c）副標本(世帯)は参考数値として，ミクロデータから世帯を事後的に４区分した副標本法での標準誤差率を示している。なお，ハイフン(－)は行動した標本数がゼロのケースを示している。行動種類とその符号は次のように対応している。1. 睡眠 2. 身の回りの用事 3. 食事 4. 通勤・通学 5. 仕事 6. 学業 7. 家事 8. 介護・看護 9. 育児 10. 買い物 11. 移動（通勤・通学を除く）12. テレビ・ラジオ・新聞・雑誌 13. 休養・くつろぎ 14. 学習・研究（学業以外）15. 趣味・娯楽 16. スポーツ 17. ボランティア活動・社会参加活動 18. 交際・つきあい 19. 受診・療養 20. その他

付表2　平日平均の理論分散とその代替分散

抽出ステップ	対象世帯数	対象標本数	推定人口	総平均	標本調査データの標準誤差率						JK1	JK2	RG1	RG2
					HC	(ΣV_h)	(COV_1)	(COV_2)	HP	SHP				
0	17244	24752	20778815	79.33	1.19	16.14	2.09	1.03	1.19	1.02	−	−	1.17	0.92
1	15478	22235	18681895	79.31	1.26	17.91	2.42	1.17	1.26	1.07	−	−	1.17	0.96
2	12284	17688	14765859	78.88	1.41	22.40	2.83	1.48	1.41	1.20	−	−	1.40	0.45
3	8709	12589	10475878	78.73	1.66	30.70	4.22	1.86	1.67	1.41	−	−	1.33	0.95
4	5309	7675	6331961	78.27	2.07	47.17	6.47	2.61	2.07	1.76	−	−	1.76	0.78
5	2666	3874	3227908	77.56	3.00	95.02	14.30	5.94	3.01	2.52	−	−	2.05	2.73
6	2558	3726	3118343	77.89	3.07	99.50	15.53	6.21	3.08	2.57	−	−	2.05	2.66
7	2359	3439	2885721	78.20	3.15	106.62	16.10	6.63	3.16	2.65	−	−	2.29	2.98
8	2085	3025	2528570	78.95	3.38	125.30	19.90	6.63	3.42	2.85	−	−	2.11	3.10
9	1741	2533	2107352	79.13	3.73	148.48	24.90	9.92	3.74	3.09	−	−	3.11	3.34
10	1389	2041	1713581	80.58	4.18	184.37	35.68	13.89	4.18	3.37	4.20	3.38	3.52	4.60
11	1060	1557	1298632	81.31	4.86	251.25	48.78	20.46	4.89	3.92	4.92	3.93	3.72	6.53
12	744	1086	886209	79.51	4.99	278.23	43.92	13.49	5.00	4.20	5.02	4.22	5.30	5.46
13	516	765	613903	77.97	6.09	404.22	60.67	19.22	6.12	5.17	6.17	5.20	4.73	7.19
14	337	525	413201	77.33	6.73	521.90	46.06	31.39	6.77	5.92	6.85	5.97	5.43	7.72
15	209	332	279925	80.64	7.90	794.48	62.94	47.47	8.06	7.08	8.21	7.19	5.37	6.13
16	118	199	161931	76.25	10.03	1085.74	95.56	93.10	11.37	9.97	11.84	10.25	8.33	25.03
17	68	104	88742	66.72	11.11	1023.13	141.72	33.21	14.04	13.05	15.25	13.87	14.00	9.59
18	29	43	34751	61.72	24.46	2357.88	1653.81	15.41	20.19	16.24	24.75	17.91	30.78	13.79

注：世帯の抽出方法は，1－5回目までは90%，6－18回目までは96%のサンプルを，それぞれ前の抽出ステップのサンプルから抽出している。対象世帯数および対象標本数とは，抽出された子どものいる世帯と子どもの標本サイズを示しており，それぞれ5日調査分の合計となっている。なお，「子ども」とは続き柄が「子」に該当する者を表す。推定人口は対象標本サイズから平日平均として人口数を算出している。また「HC」，「ΣV_h」，「COV_1」，「COV_2」は(3.3)式に対応している。「HP」，「SHP」，「RG1」，「JK1」はそれぞれ4節の(i)，(ii)，(iii)，(iv)に基づく推定値である。「RG2」は「RG1」とは異なる副標本を用いたときの(iii)による分散推定値を示し，「JK2」はジャックナイフ法であるが，1日目と2日目を異なる世帯として扱った推定値である。なお，ジャックナイフ法（JK1，JK2）については標本サイズが大きいとき計算時間を要するため抽出ステップ10回目以降に限り計測した。

付表２（続き）　平日平均の理論分散とその代替分散

抽出ステップ	頻度計算の標準誤差率					ウェイト付		ウェイト無		n1	n2
	ウェイト無	（平均値）	ウェイト付	（平均値）	調整ウェイト付	Cor1*	Cor2*	Cor1	Cor2		
0	0.74	80.41	0.01	79.30	58.28	0.48	0.53	0.28	0.30	4088	4112
1	0.78	80.66	0.01	79.31	58.32	0.48	0.54	0.15	0.31	3667	3698
2	0.88	80.19	0.01	78.88	58.27	0.48	0.55	0.28	0.31	2900	2926
3	1.05	80.70	0.02	78.68	58.61	0.49	0.51	0.30	0.31	2087	2065
4	1.34	80.24	0.02	78.17	57.86	0.47	0.48	0.32	0.33	1255	1278
5	1.89	80.10	0.03	77.46	57.83	0.50	0.49	0.37	0.33	634	661
6	1.91	80.58	0.03	77.77	57.65	0.49	0.50	0.38	0.33	610	639
7	2.00	80.94	0.03	78.01	57.77	0.48	0.48	0.39	0.32	557	597
8	2.12	80.50	0.03	78.68	57.50	0.49	0.50	0.34	0.32	481	525
9	2.33	80.67	0.04	78.84	57.32	0.51	0.49	0.38	0.31	401	440
10	2.63	81.89	0.04	80.29	57.06	0.48	0.50	0.40	0.35	322	348
11	3.02	82.65	0.05	80.94	57.11	0.66	0.52	0.44	0.37	250	266
12	3.58	83.55	0.05	79.42	55.20	0.64	0.44	0.47	0.30	175	191
13	4.31	80.51	0.07	78.11	57.31	0.60	0.50	0.51	0.37	121	133
14	4.71	79.40	0.08	77.21	53.18	0.63	0.56	0.21	0.50	82	91
15	5.49	82.41	0.09	80.18	50.32	0.66	0.60	0.06	0.45	54	61
16	7.42	76.28	0.12	73.83	51.06	0.66	0.60	0.03	0.46	31	40
17	10.10	65.63	0.16	62.77	49.15	0.72	0.62	0.51	0.12	12	20
18	18.46	57.56	0.28	60.16	52.74	0.87	0.60	0.72	0.27	3	8

注：「頻度計算の標準誤差率」は，頻度計算用（SPSSの場合ウェイトつきクロステーブルコマンドなど）の標準誤差から算出した。「ウェイト無」は，ウェイトを付けずに平日平均を推定したものであり，平均値はMW統計量に対してバイアスをもつ。また「ウェイト付」は，調整ウェイトではなく通常のウェイトによる標準誤差率を示しており，その平均値もMW統計量とは一致しない。「調整ウェイト付」は，調整ウェイトを用いて頻度計算用の標準誤差を算出したものである。「ウェイト付」と「調整ウェイト付」の詳細は脚注[15]を参照のこと。さらに，「Cor1*」と「Cor2*」，「Cor1」と「Cor2」，および「n1」と「n2」はそれぞれ火水と木金の調査標本について，ウェイト付きの相関係数，ウェイト無しの相関係数，および標本数を示している。

注

1）下付きのプラス（+）は，該当の属性を合計した統計量であることを意味する。
2）複数の属性や変数で絞り込む，いわばクロスにクロスを重ねるタイプの部分母集団を問題にする場合には，とくにこのような定義での作業は負荷が大きい。
3）各曜日の平均値がほぼ等しい位置にあるときは，人口が曜日間で変動しても，MP統計量はMW統計量に近似する。しかしながら，実際にいくつかの部分母集団について2つの統計量を計算し比較したところ，標本サイズの小さい部分母集団については，MW統計量とMP統計量の間に顕著な差が確認される。
4）本稿では，定義上の平日平均統計量であるMW統計量に対して，ある推定量がそれと一致しないとき，便宜上，「バイアス」と表現している。母数（真値）と推定量の期待値との差という意味でのバイアスと実質的には同じである。
5）調査期間は土曜から開始し次の週の日曜までとしており，8区分した調査グループを，最初の土日に2グループ，日月，火水，木金，金土にそれぞれ1グループ，最後の土日に2グループずつ割り当てている。
6）一般的な標本理論に基づく推定量および推定量の分散については，土屋(2009)，松井(2005)，Cochran, W.G.(1977)，StataCorp.(2009）などを参照のこと。
7）総務省統計局(2003)によれば，国勢調査の結果数値などから，推計した地域，男女，年齢別の人口を基準人口としている。
8）標本設計情報が全て利用できるとき，非復元での層化二段抽出法であることから，理論上は第1次抽出単位（PSU）の分散推定量と第2次抽出単位（SSU）の分散推定量の合計を全体の推定値とすべきである。ここで，総務省統計局(2002，2003)から平均的な調査区抽出率の概算値として地域（都道府県）別の標本調査区数/調査区数を算出し度数分布表（参考表）としたとき，実際にはPSUの抽出率に関する概算値は極めて小さいことが確認できる。このような場合，通常ならPSUの分散推定量のみで十分近似できると考えられるが，ミクロデータにはPSUに関する標本設計情報が付与されていないため，式(2.3)のように，SSUのみの分散推定量を理論分散とする以外にない。

参考表　地域別第1次抽出単位（調査区）抽出率（概算値）の度数分布表

PSUの平均抽出率	～0.004	0.004～0.007	0.007～0.010	0.010～0.013	0.013～0.016	0.016～0.019	0.019～
度　数	3	12	10	8	5	5	4

9）世帯クラスターは標本設計情報の最終抽出単位であることから，その他の情報が利用できないとき，世帯を利用するのが自然である。しかし，世帯クラスターではなく，個人を無作為抽出したものと仮定することも可能である。この場合，世帯クラスターでの分散推定量に対して，若干ではあるが減少するため，過小推定する傾向にあることが分かっている。たとえば，2001年の社会生活基本調査ミクロデータで検証したところ，下二桁以降の数値で違いが出ている。さらに，世帯クラスターとして算出するとき，4節以降の調整ウェイトを用いたMP統計量を算出するための理論と計算操作が容易になるという利点があることを指摘しておく。
10）標本設計情報を全て用いた場合の分散推定量などの詳細はKurihara, Y.(2010）を参照のこと。なお，部分母集団の推定時には，対象を識別するダミーをκ_{hij}として，次のように求めればよい。これは共分散についても同様である。

$$\hat{V}(\hat{\mu}^{\#}_{h\kappa+}) = \frac{m_{h+}}{m_{h+}-1}\cdot\frac{1}{\hat{N}^{\#\,2}_{h\kappa+}}\Sigma_{i\in h}\left[\Sigma_{j\in i}w^{\#}_{hij}\kappa_{hij}\left(y_{hij}-\hat{\mu}^{\#}_{h\kappa+}\right)\right]^2$$

11）統計量の平均の分散については，松井(2005) pp.115-117を参照のこと。また集落抽出に基づく平均値の分散については，土屋(2009) pp.139-145，StataCorp.(2009) pp.155-160などを参照。
12）生活時間調査の調査曜日を設定する方法は，平日平均統計量の分散に直接影響するため重要な問題である。欧州各国の調査方法も含めて整理すれば，主に7日間連続調査，2日間連続調査，1日

調査がある。Eurostat(2009) もしくはその翻訳資料である水野谷(2010) の生活時間調査のガイドブックでは，推定値の分散を最小にするため，1 世帯について平日 1 日および週末 1 日の調査を推奨している。

13）この原因の一つにリサンプリングデータであることによる標本サイズの縮小も考えられ，全標本を利用して HC 分散を算出すれば，より 45 度線ラインに接近することが想定される。

14）最近の統計処理用アプリケーションソフトには，標本設計情報（層化やクラスター情報）を考慮した推定を可能にするプログラムが実装されている。固定標本方式も含む複雑な標本設計でなければ，一般利用者でも容易にアウトプットを得ることができる。しかしそのことは逆に，与えられた標本設計の下で適切な推定量の選択や定式化の適否が問われることを意味している。このような問題背景が本節以降での議論の焦点のひとつでもある。

15）全調査データが利用できたとしても，曜日別人口として調整されているのは地域・男女・年齢別の層までであり，その他の変数（配偶関係や就業関係など）については各曜日の人口は変動する。そのため MW 統計量とその標準誤差の算出に伴う煩雑さはリサンプリング率に関係なく発生する。

16）その他の代替分散として，ウェイトを頻度として計算する推定法も考えられる。しかし各曜日の総計が 1 になるように調整されていることから，頻度計算による平均値の標準誤差 $\widehat{SE}_{Freq}$ は，下式のように非常に偏った値となり利用できない。これは付表 2 の「調整ウェイト付」での標準誤差率でも確認できる。

$$\widehat{SE}_{Freq}(\hat{\mu}_{+}^{v\#}) = \sqrt{\frac{\hat{V}_{Freq}(\hat{\mu}_{+}^{v\#})}{\Sigma_{hij} v_{hij}^{\#}}} = \sqrt{\frac{1}{\alpha} \cdot \frac{1}{(\alpha - 1)} \Sigma_{hij} \frac{w_{hij}^{\#}}{\hat{N}_{h+}^{\#}} (y_{hij} - \hat{\mu}_{+}^{v\#})^2}$$

また頻度計算による通常のウェイトを用いたプール平均の標準誤差は

$$\widehat{SE}_{Freq}(\hat{\mu}_{+}^{\#}) = \sqrt{\frac{\hat{V}_{Freq}(\hat{\mu}_{+}^{\#})}{\Sigma_{hij} w_{hij}^{\#}}} = \sqrt{\frac{1}{\hat{N}_{+}^{\#}} \cdot \frac{1}{(\hat{N}_{+}^{\#} - 1)} \Sigma_{hij} w_{hij}^{\#} (y_{hij} - \hat{\mu}_{+}^{\#})^2}$$

であり，分母の推定人口が大きいために標準誤差が非常に小さく算出される。当然，通常のウェイト利用であるため，平日平均推定値 $\hat{\mu}_{+}^{\#}$ は MW 統計量に対してバイアスをもつ。これは付表 2 の「頻度計算ウェイト付」に示されている。

17）副標本法による標本誤差については Wolter, K.M.(2007) pp.22-27 を参照。

18）ジャックナイフ法による標本誤差については，Wolter, K.M.(2007) pp.152-153 を参照。

19）各抽出ステップで MW 統計量の値が変化することもあり，推定量の分散（または標準誤差）ではなく，一般的に標準誤差率（標準誤差を MW 統計量で除した値）を用いて議論している。

20）本稿で算出した数値は同一部分母集団を対象とした試行結果ではあるが，その他の行動種類や部分母集団についても確認したところ，同様の傾向がみられた。

21）世帯の抽出方法やその他の詳細は付表 2 を参照のこと。

22）ウェイト無の平均値は調整ウェイトを利用していないため，MW 統計量に対してバイアスをもつ。その程度は付表 2（続き）を参考のこと。

参考文献

［1］ 坂田幸繁・栗原由紀子（2010），「世帯員間同時分布モデルと生活時間分析の方法 ― 社会生活基本調査の2次利用をめぐって ― 」，『研究所報』，No. 39，pp.67-88，法政大学日本統計研究所.

［2］ 総務省統計局（2002），『平成12年国勢調査，調査区関係資料利用の手引』，日本統計協会.

［3］ 総務省統計局（2003），『平成13年社会生活基本調査報告　第1巻　全国　生活時間編（その1）』，財務省印刷局.

［4］ 高橋雅夫・臼井彩子（2005），「平成13年社会生活基本調査における標本の代表性と調査結果の推定について」，『統計研究彙報』，第62号，pp.23-70.

［5］ 土屋隆裕（2009），『概説標本調査法』，朝倉書店.

［6］ 標本誤差推計研究会（1998），『標本誤差の推計方法 ― 最新時代の理論と実証 ― 』，財団法人統計情報研究開発センター.

［7］ 松井　博（2005），『標本調査法入門』，日本統計協会.

［8］ 水野谷武志（2010），「欧州統一生活時間調査（HETUS）ガイドライン－2008年版（翻訳と解説）」，『統計研究参考資料』No. 107，pp.21-23，法政大学日本統計研究所.

［9］ Cochran, W. G.（1977）, *Sampling Techniques*, Third Edition, John Wiley & Sons.

［10］ Eurostat（2009）, *Harmonised European time use surveys : 2008 guidelines*, pp.16-18, eurostat Methodologies and Working papers.

［11］ Wolter, K.M.（2007）, *Introduction to Variance Estimation Second Edition*, Springer.

［12］ Kurihara, Y.（2010）, “Estimation of Weekday Averages and Their Variance with The Resampled Data from The Survey on Time Use and Leisure Activities”, *The Annual of the Institute of Economic Research Chuo University*, No. 41, The Institute of Economic Research Chuo University.

［13］ Patterson, H.D.（1950）, “Sampling on Successive Occasions with partial replacement of Units”, *Journal of the Royal Statistical Society Series B*（*Methodological*）, Vol. 12, pp.241-255.

［14］ Skinner, C.J.（1989）, *Analysis of Complex Surveys*, ed. C.J. Skinner, D. Holt & T.M.F. Smith, pp.23-58, John Wiley & Sons.

［15］ StataCorp.（2009）, Stata Survey Data Reference Manual Release 11, pp.163-164.

Estimation of Sampling Errors in Measures of the Average of Weekday Using Anonymized Microdata from the Japanese Survey on Time Use and Leisure Activities

Yukiko KURIHARA

（Graduate school of economics, Chuo University ; ebaku24@gmail.com）

Summary

This paper theoretically studies the estimator of the average of weekday and its variance by utilizing anonymized microdata from the Japanese Survey on Time Use and Leisure Activities. It also investigates efficient and practical data handling by calculating the adjusted weight of the pooled data over a weekday.

To examine the basic characteristics of weekday activities on the basis of time use data, we use conventional measures to estimate the average of weekday, such as mean statistics by days. However, there are several issues to be noted for the calculations. First, we need to assume that the household clusters were randomly sampled, because the original sampling information is not available, although we are aware that stratified two-stage sampling was employed. Second, a customized computing program was required in order to exist covariance caused by the survey method that the households were surveyed over two days.

Key Words

Japanese Survey on Time Use and Leisure Activities, anonymized microdata, fixed samples, sampling error, adjusted weight