『経済研究』（明治学院大学）第 140・141 合併号　2008 年

グル″[1]ープデータに基づく所得分布の推定

西　尾　　敦

1. はじめに

近年，格差の問題が注目を集め，その拡大が社会問題としてマスコミ等で取り上げられる頻度が増している。「格差」は様々な観点から論じ得るが，所得と資産の格差は比較的客観的すなわち定量的な扱いになじむと思われる。

定量的な格差の分析は，しばしばジニ係数を用いて行われるが，これは所得／資産の分布から計算される。したがってジニ係数は，平均等の統計量と同じように，分布の裾の形状の影響を受けやすいといわれている。所得分布の推定に当たっては，特に「裾の」状態を的確に推定することが重要である。

本稿では，主として公表された家計調査データに基いた所得分布，特に裾すなわち最大所得階層に関連したの推計の問題を扱う。

1.1　データ

本稿では主として，WEB（http://www.stat.go.jp）上で集計公表されている家計調査データを用いる。このうち農林漁家世帯を含む二人以上の世帯を年間所得階層別に分類集計した結果表から，年間収入の項目を分析の対象にした。家計調査では，全国の世帯から層化三段抽出により約 8000 世帯を抽出し毎月ごとに主として消費支出額に関わる事項を調査している。調査報告では，世帯を勤労者世帯とそれ以外の世帯に類別して，全世帯にわたる集計値と勤労者世帯のみの集計値とを求め公表している。標本中の総世帯数およびそれぞれの類型に属する世帯数は事後的に決まるが，その比率はほぼ一定しておよそ半々である。世帯の所得すなわち収入に関わる項目では，「年間収入」と各月の「実収入」がこれに該当する。前者は世帯票に記載される申告額である。一方後者は記録に基づいて得られる統計情報であり，より正確さが期待できるが，勤労者世帯のみにつき得られる情報である。本稿では，年間収入を，年間収入によってグル″ープ化した月次集計表を，2000～2007 年のそれぞれ 1 月および 7 月（以後

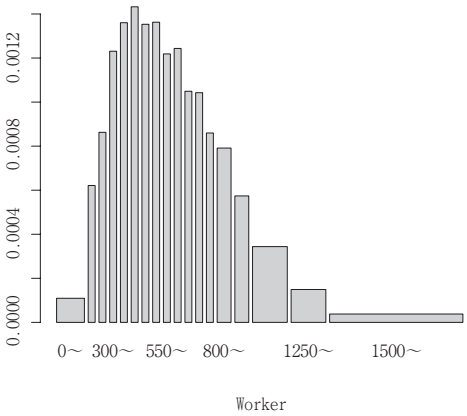

図 1.1　勤労者世帯

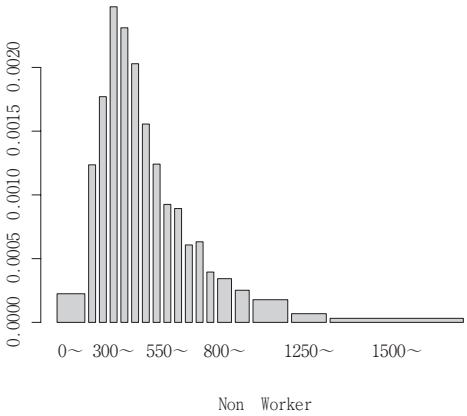

図 1.2　非勤労者世帯

これを「各月」という），計 16ヶ月のデータを特に取り上げて分析する。6ヶ月ごとのデータを用いるのは，同調査では毎月標本の 1/6 づつを入れ替えるため，5 連続月にわたって標本抽出に起因する系列相関が生じるが，半年ごとに標本変動に関して完全に独立な標本となるためである。

図 1.1 および図 1.2 は，それぞれ勤労者世帯[2]，非勤労者世帯[3]の各所得階層の度数を全 16ヶ月にわたりプールして得られた分布（以後累計分布）のヒストグラ″ムである[4]。この 2 つの図には，明らかな差がある。累計分布の中央値[5]は，勤労者世帯でおよそ 610 万円，非勤労者世帯では 395 万円である。また勤労者世帯では年収 500 ないし 600 万円から 1250 万円までの階層（以後中流階層）が厚いのに対して非勤労者世帯では，比較的薄くかつ幾何級数的に減少している。また，最大所得階層（以後裾階層あるいは裾クラス）の級内平均[6]は勤労者，非勤労者世帯でそれぞれ，1765 万円，2214 万円である。ジニ係数は，それぞれ，0.252,0.340 である。非勤労者世帯に類別される世帯には，会社経営者から無職まで様々な世帯が含まれるので当然であるが，非勤労者世帯では勤労者世帯に比べ「格差」がかなり大きい。

1.2　分布変化の検定

2000 年から 2007 年までの 8 年間で所得分布が変化したか否かは，通常の分割表の独立性の検定[7]を行えばよい。実際，対数尤度比検定統計量は，全世帯，勤労者世帯，非勤労者世帯それぞれ，545.7(0),256.6(2.4e-12),379.1(0)である[8]。つまり所得分布の変化は，データ期間中で有意である。

所得分布の変化を表 1 および表 2 に示す。表の数値は，各所得階層の全 16ヶ月の平均（以後平均分布）と各セルの比（の対数）である。つまり数値がプラスならば，そのセルの度数は平均分布より大きく，マイナスならば小さいことを示す。表 1（勤労者世帯）では，表の左上から右下にかけてマイナス符号が目立つ。2000 年から 2007 年にかけて高所得の階層の世帯数が減少し，低所得層が増えていることが読み取れる。非勤労者世帯（表 2）でも傾向は同様であるが，右下すなわち分析期間後半 2005～2007 年の高所得世帯数の落ち込みが勤労者世帯に比べて大きい。

表 1　$\log\hat{p}_{ki}/\hat{p}_{k0}$（勤労者世帯）

年収＼年	2000	2001	2002	2003	2004	2005	2006	2007
0～	0.12	0.05	0.07	−0.18	−0.39	−0.14	0.32	−0.01
200～	−0.41	−0.18	0.06	0.04	0.02	0.00	0.12	0.25
250～	−0.15	−0.12	−0.15	−0.05	0.13	0.15	0.10	0.14
300～	−0.16	−0.07	−0.12	0.03	−0.07	0.13	0.10	0.13
350～	−0.08	−0.06	0.01	0.04	−0.05	−0.02	0.12	0.04
400～	−0.13	−0.08	−0.12	−0.01	0.13	0.02	0.19	−0.03
450～	−0.04	−0.05	0.05	−0.08	0.09	0.01	−0.01	0.01
500～	0.00	0.03	−0.01	0.05	−0.02	−0.04	−0.07	0.04
550～	−0.07	−0.04	0.00	−0.02	0.04	0.09	−0.04	0.03
600～	−0.16	0.00	−0.06	0.12	0.01	0.04	0.04	0.00
650～	−0.03	0.00	−0.02	0.05	−0.12	0.02	0.09	0.00
700～	−0.01	−0.13	0.05	0.02	0.06	0.05	−0.07	0.01
750～	0.08	0.01	−0.08	0.00	0.05	0.04	−0.15	0.02
800～	0.04	0.04	0.09	−0.02	−0.07	0.03	−0.09	−0.03
900～	0.08	0.05	0.04	−0.05	−0.03	0.00	−0.03	−0.07
1000～	0.11	0.05	0.03	−0.01	0.04	−0.13	−0.08	−0.04
1250～	0.26	0.25	−0.02	−0.07	−0.11	−0.19	−0.04	−0.25
1500～	0.22	0.05	0.05	−0.05	−0.10	−0.17	−0.02	−0.06

表 2　$\log\hat{p}_{ki}/\hat{p}_{k0}$（非勤労者世帯）

年収＼年	2000	2001	2002	2003	2004	2005	2006	2007
0～	0.12	0.21	0.16	−0.35	−0.24	−0.19	0.13	0.02
200～	−0.09	−0.07	−0.09	0.05	0.03	0.07	0.00	0.06
250～	−0.10	−0.09	−0.08	0.01	0.06	0.05	0.07	0.03
300～	−0.11	−0.14	−0.08	−0.04	0.07	0.00	0.08	0.17
350～	−0.11	−0.12	−0.10	0.02	0.10	0.03	0.08	0.06
400～	−0.12	−0.01	−0.17	0.06	−0.03	−0.02	0.09	0.14
450～	−0.03	−0.12	−0.03	0.02	−0.08	0.15	0.03	0.02
500～	−0.15	−0.01	0.02	−0.01	0.08	0.11	−0.01	−0.06
550～	−0.09	0.02	0.11	0.03	0.00	0.08	−0.06	−0.10
600～	−0.01	0.09	−0.02	0.00	0.04	−0.07	0.11	−0.15
650～	0.08	0.17	0.17	−0.07	−0.13	−0.01	−0.17	−0.07
700～	−0.02	0.01	0.14	−0.06	0.01	−0.03	−0.05	−0.01
750～	0.11	−0.01	0.08	0.13	−0.04	0.06	−0.02	−0.34
800～	0.08	0.06	0.09	0.11	−0.02	0.02	−0.20	−0.17
900～	0.06	0.09	0.15	0.08	−0.02	−0.14	−0.04	−0.21
1000～	0.20	0.31	−0.06	−0.11	−0.08	−0.03	−0.17	−0.12
1250～	0.47	0.14	0.14	0.02	−0.02	−0.38	−0.48	−0.14
1500～	0.44	0.00	0.26	0.05	−0.14	−0.29	−0.30	−0.22

1.3 ジニ係数

ある正値分布の密度関数が $f(x)$，その期待値を $E(X)$，$G(x)=\int_0^x zf(z)dz/E(X)$ とすると，この（理論）分布のローレ″ンツ曲線は，$y=G(F^{-1}(q)),0\leq q\leq 1$と定義される。ジニ係数は，$X,Y$ が独立で $f(x)$ にしたがう確率変数のとき$GI=\frac{E|X-Y|}{E(X)}$と定義されるが，これはローレ″ンツ曲線と直線 $y=q$（45° 線）で囲まれる図形の面積の 2 倍に一致する。

家計調査の収入階層別集計表から，収入階層 $k=1,\ldots,K$ ごとに度数 f_kとクラス平均$\bar{y}_k$が得られる。このとき，ローレ″ンツ曲線，ジニ係数は，理論分布を経験分布で置き換えて推定する。実際には累積世帯数を基準化した$q_k=\frac{1}{n}\sum_{j=1}^{k}f_k$，累積所得を基準化して$Y_k=\frac{1}{n}\sum_{j=1}^{k}f_k\bar{y}_k$（ただし，$n=\sum_{k=1}^{K}$は総世帯数数，$\bar{y}=\frac{1}{n}\sum_{k=1}^{K}f_ky_k$は平均所得）から，座標 $(0,0),(q_1,Y_1),\cdots,(q_{K-1},Y_{K-1}),(1,1)$ を順に結んで得られる曲線（折れ線）によって，理論曲線が推定される。またジニ係数は，この曲線と完全平等線（45°）とで囲まれる面積の 2 倍によって推定できる。図 1.3 は，2000～2007 年のジニ係数の変化である。勤労者世帯（実線）では変化はわずかであるのに対して，非勤労者世帯（破線）では，2003～2004 年の落ち込みが大きい。この期間不況の深刻化の影響で，高額所得階層の所得が減少した結果であろう。

図 1.3 ジニ係数の変化

2. 所得分布

ジニ係数などの分布の特性値は，特定の分布型を前提せずに定義される統計量である。しかし，とくに分布の裾に大きく依存する場合，裾の分布に関して何らかの仮定を置かずに解析することは難しい。このため所得の分析では，少数の母数によって確保される自由度のみを持つ母数型分布が用いられる。所得は代表的な正の値をとる変数であるから，その分析には代表的な正値分布すなわちガンマ分布，べき指数（ワイブル）分布，対数正規分布，指数分布とパレ″ート分布などである[9]。このうち対数正規分布は反射壁を持つ拡散過程によって説明できる等の理由で，初期の分析では好まれたようであるが，データおよび解析手段の発展とともに，実際のデータによって支持されないことが明らかになっている。ここでは，やや恣意的であるが，分布の裾が指数関数的に減少する分布（広い意味ではガンマ分布もこれに属する）を代表して指数分布を，べき乗関数的に減少する分布を代表してとパレ″ート分布を取り上げ，本稿の主題であるグル″ープ化データに基づく推論について考察する。ワイブル分布は，指数関数のべき乗型の裾を持ち，異なった裾の型に分類されるが，両者が 1 母数型であるのに対し，ワイブル分布は 2 母数型であるので，とくにグル″ープ数が少ない場合，比較が容易でなく本稿では取り上げない。

なお，ここで考える指数分布，パレ″ート分布はいずれも単調減少密度関数をもつ。このことは実際

の分布と明らかに整合しない。本稿では所得分布の「右裾」の分析が目的であるので，ある値以上の分布（条件付き分布）について上述の分布型を仮定し議論を進める。

2.1 指数分布

裾の上側分布関数が指数分布，すなわち

$$\overline{F}(x)=P(X\geq x)=ce^{-\beta x} \tag{1}$$

であるとき，(1)の両辺の対数をとれば $\log\overline{F}(x)=c'-\beta x$ すなわち直線になる。家計調査 A の年間収入 500 万円以上の所得階層について，縦軸（対数目盛）に上側累積相対度数，横軸に所得 x をとってプロットし各月ごとに結ぶと図 2.1，2.2 を得る。これをみると，勤労者世帯では年収 800 万円以上の階層ではほぼ直線状である。これに対し，非勤労者世帯では，右端でやや直線の傾きが小さく（すなわち裾が厚く）なる傾向が観察される。そこで，勤労者世帯では 800 万円以上の全クラス，非勤労者世帯では比較的直線的であるように見える 700 万円以上 1250 万円以下のクラスについて，指数分布(1)モデルを

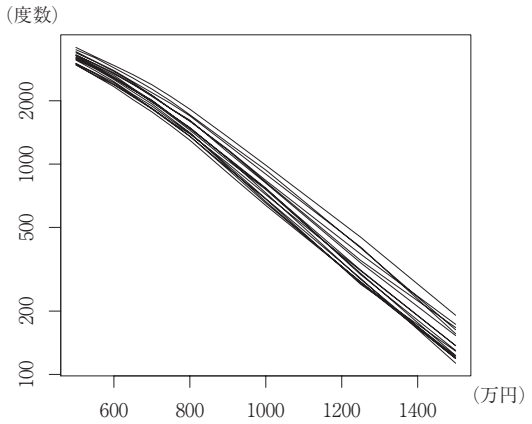

図 2.1 勤労者世帯

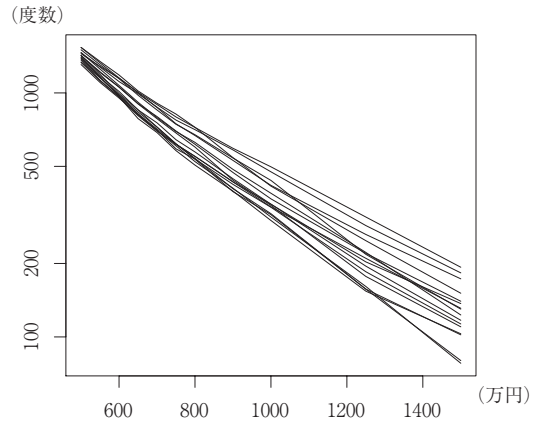

図 2.2 非勤労者世帯

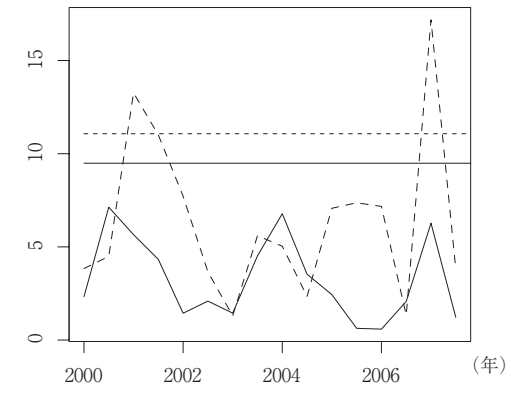

図 2.3 適合度検定

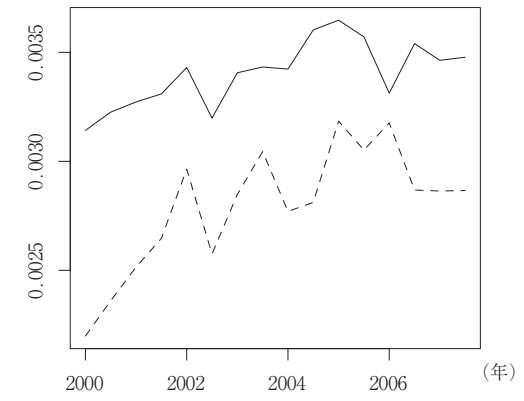

図 2.4 β の変化

仮定し，各月ごとに β を最尤法で推定した。図 2.3 は，指数分布を検定する適合度検定（尤度比検定）の各月ごとの統計量の時系列プロットである（実線は勤労者世帯，破線は非勤労者世帯，以下同様）。水平線（実線，破線）はそれぞれ検定の 5%臨界値を示している。勤労者世帯では，年収 800 万以上の高収入世帯で，指数分布がほぼ当てはまっていると結論できる。一方，非勤労者世帯では，検定統計量は自由度の違いを考慮してもやや大きく，指数分布は全体として十分な説明力を持つとはいえない。

図 2.4 は，$\widehat{\beta}$の時系列プロットである（実線は勤労者世帯，破線は非勤労者世帯。）。一般に非勤労者世帯は勤労者世帯より，β が小さい，すなわち裾が厚いことがわかる。

全 16 期の β の共通性すなわち β が全期間に渡って一定であることの検定（尤度比検定）統計量は，勤労者世帯，非勤労者世帯でそれぞれ，39.97（P- 値 0.04%），41.21（P- 値 0.03%），でいずれも高度に有意である。この検定結果と図 2.4 から β は増加傾向にあることがわかる。すなわち，この 8 年間で高収入階層で均等化が進んだと見なすことができる。

2.2 パレ″ート分布

パレ″ート分布は，次のような上側分布関数を持つ分布である：

$$\overline{F}(x)=\left(\frac{x}{L}\right)^{-\alpha},\ x>L>0 \qquad (2)$$

ここで，L は正の数で，分析に当たっては最低所得とみなし得る。$\alpha>0$ は正の値をとる母数で裾の厚さを規定し，パレ″ート指数とも呼ばれる。一般に α が小さい（0 に近い）ほど裾が重く，$\alpha\leq 2$ のとき分布の分散が（発散し）存在せず，$\alpha\leq 1$ ならば期待値も存在しない。

図 1.1，1.2 に典型的に見られるように所得／資産分布では，全階層にわたって，分布がこの形状を持つことは希である。ここでは，分布がパレ″ート型の裾を持つ，すなわち一定額（x_0 以上の所得の分布について(2)が成り立つと仮定して，該当するセルにのみ（すなわち $X>x_0$ の条件付分布に）注目して議論を進める。簡単にわかるように，(2)の分布について，$X>x_0\ (>L)$ の条件付き上側分布関数は，L を x_0 で置き換えて得られる（(3)式）。本稿の分析では，L は分析対象となる所得階層の最低所得額

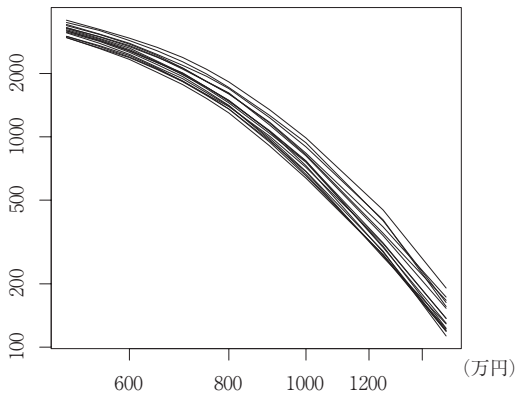

図 2.5　勤労者世帯

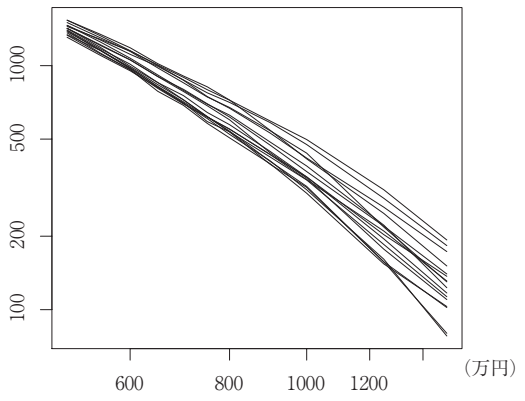

図 2.6　非勤労者世帯

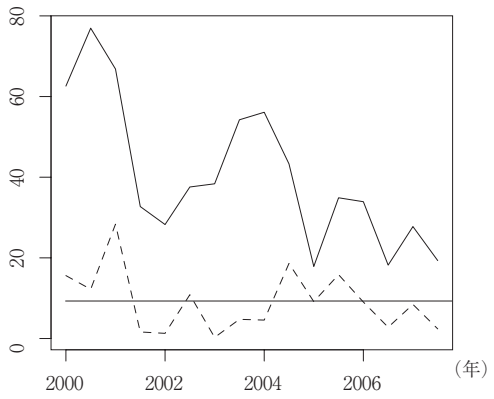

図 2.7　適合度検定

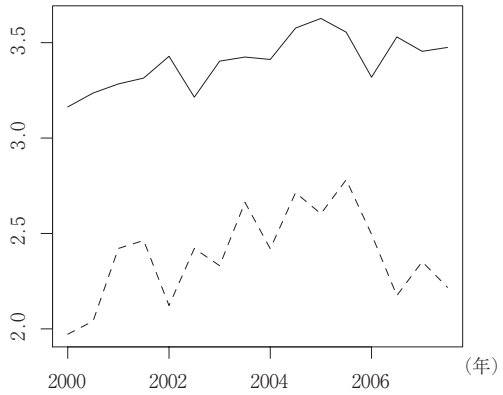

図 2.8　α の変化

をあらわすものと解釈すればよい。

$$\overline{F}(x)=\left(\frac{x}{x_0}\right)^{-\alpha},\ x>x_0>0,\ \alpha>0 \tag{3}$$

(3)の両辺の対数をとると，直線の式を得る。前節と同様の所得 x の範囲について，所得，上側累積相対度数を両対数目盛にプロットすると，図 2.5，2.6 を得る。勤労者世帯では直線からはずれて，右下がりの曲線を描いている。一方，非勤労者世帯では，月によって高収入側が直線状である場合もあるように観られる。年収 800 万円以上のクラス（以後中流層という）について，各月ごとにパレ″ート指数を最尤法で推定した。図 2.7 は適合度の尤度比検定統計量である（実線が勤労者世帯，破線が非勤労者世帯)。指数分布を考えたときとは逆に，勤労者世帯の適合度は非常に悪い，一方非勤労者世帯では全く悪いというわけではない。ただし，5%水準で棄却される月数はかなり多く，やはり指数分布（図 2.3）のほうが当てはまりが勝る。

当てはまりは悪いが，あえてパレ″ート指数 α を示すと，図 2.8 のように変化する。これは，指数分布の β の動きによく似ており，2000 年から 2008 年の 8 年間で，中流層の所得分布の裾が軽くなっている様子が観察できる。

3. 裾クラスの平均収入

前節の分析は，所得階層の度数分布のみの情報に基づくものであった。家計調査報告書では，各クラスごとの平均年間収入が公表されている。特に所得分布の裾の分布がわれわれの主たる関心事である。このため，裾クラスすなわち最大所得階層のクラス平均，言い換えるとその区間の条件付き期待値を検討し，裾の分布に関する推論を行ってみよう。

3.1　指数分布

指数分布(1)では，条件 $X>x_0$ の下での $X-x_0$ の条件付き分布は，無条件分布(1)と一致する。(1)

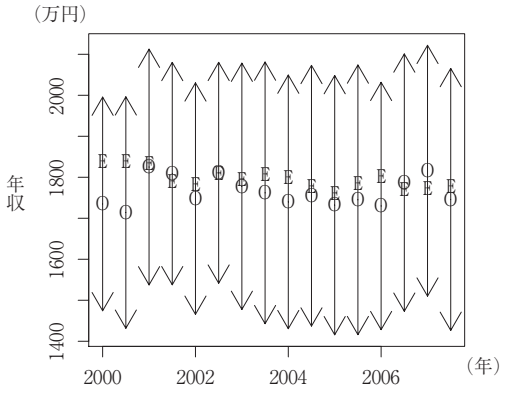

図 3.1　勤労者世帯

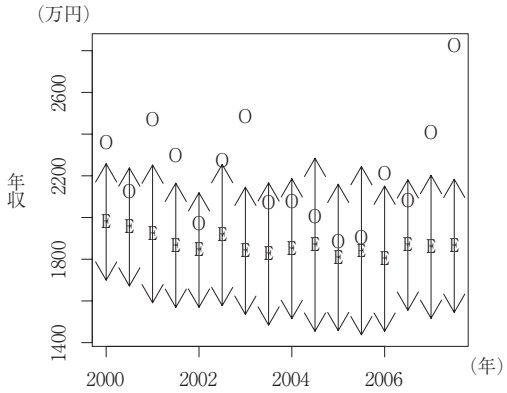

図 3.2　非勤労者世帯

の期待値は$\frac{1}{\beta}$であるから，条件付期待値は

$$E(X|X>x_0)=x_0+\frac{1}{\beta} \tag{4}$$

である。$x_0=1500$（万円）とし，βに前節で推定した値を用いて得られる，各月の（最大所得階層の）クラス平均の理論値は，図 3.1，図 3.2 の“E”印で示した値である[10]。実際の平均年収（“O”）を，図に重ねてプロットした。勤労者世帯では，指数分布ときわめてよく整合している。前節の分析と併せて，中流階層以上の勤労者世帯の所得分布は指数分布でよく説明できることがわかる。

一方，非勤労者世帯では，実際の平均年収は指数分布に基づく理論値を上回っているように見られる。少し詳しく見ると“O”は，分析期間の初めから“E”を上回り，2003 年から 2005 年にかけて，両者は接近したが，2006 年の調査以降再び顕著になりつつある。非勤労者世帯については，裾クラスの所得は（指数分布を仮定した）中流層の分布の延長では説明できないことが明らかとなった。

3.2　パレ″ート分布

前節の議論で，勤労者世帯の分布は，指数分布で十分説明できることが示された。この節では非勤労者世帯のみを扱う。パレ″ート分布(2)は，$\alpha>1$ のとき期待値が存在し $E(X)=\frac{\alpha L}{\alpha-1}$である（Johnson et al., 1994 を参照）。またこのとき，右半開区間(x_0,∞)の条件付き期待値は

$$E(X|X>x_0)=\frac{\alpha}{\alpha-1}x_0 \tag{5}$$

で，Lに無関係である。前節で推定したαをこの関係式に代入すると，裾クラスの平均収入（理論値）が得られる。この値を“P”印で，実際の平均年収を“O”印で重ねて図 3.3 に示す。矢印で示した区間は，モンテカル″ロ法によって得られた観測平均年収の 95％予測区間である。予測区間の作り方その他，標本平均に基づくパレ″ート指数αの推定については，4 節で論じる。

図 3.3 では実際の収入“O”がパレ″ート分布に基づく理論値“P”あるいはその予測区間を下回るケースが多い。これは，指数分布に基づく図 3.4（図 3.2 再掲）と対照的である。指数分布の裾は指数関

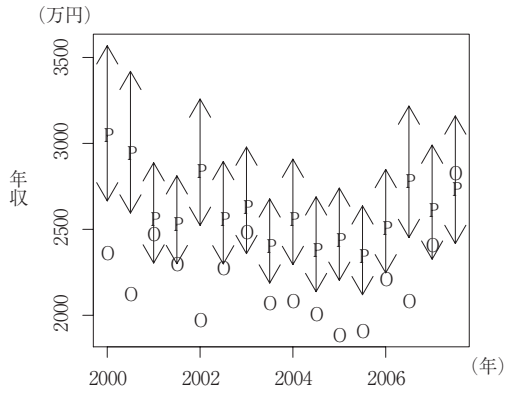

図 3.3　パレ″ート分布

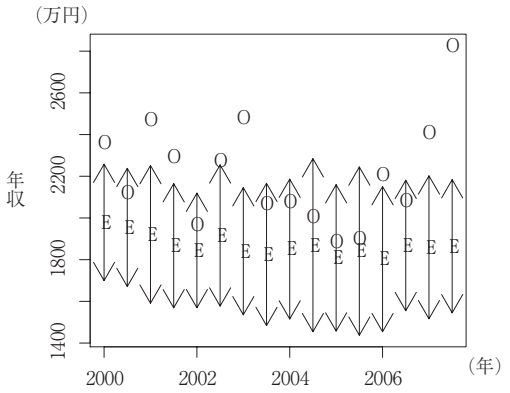

図 3.4　指数分布

数的に減少し，パレ″ート分布ではべき乗関数的減少であるから，前者のほうが裾が軽いことを勘案すれば，実際の分布（の裾）は両者の中間に位置するものと考えるのが妥当である。

3.3　混合分布

前 2 節の考察を承け，ここでは，非勤労者世帯全体を異なる 2 集団，すなわち所得が指数分布したがう集団（以後「指数型集団」）とパレ″ート分布にしたがう集団（以後パレ″ート型集団）が一定の割合で混合しているとしよう。また，パレ″ート集団の所得は最大所得階層にのみ属すると仮定する。このとき，次のように 2 集団の混合比率などを求めることができる。

まず，裾クラスを除く中流階層（$x_0 = 800 \leq X < 1500 = x_1$）のデータから指数分布を推定する（3.1 節）。ここで推定した$\hat{\beta}$と中流層の度数を用いて指数型集団の裾クラスの度数を

$$\hat{f}_K^{(e)} = (\sum f_k)\frac{\exp\left(-\hat{\beta}x_1\right)}{\exp\left(-\hat{\beta}x_0\right)-\exp\left(-\hat{\beta}x_1\right)}$$

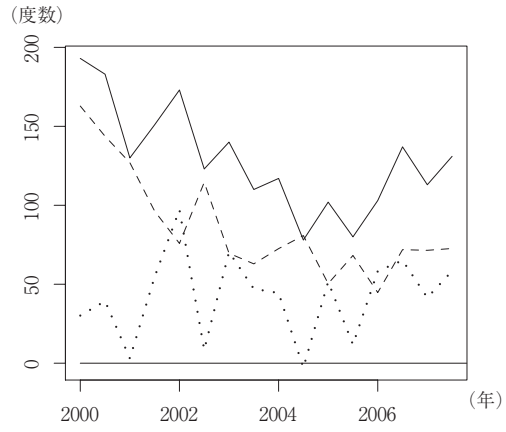

図 3.5　世帯数（最大所得階層）

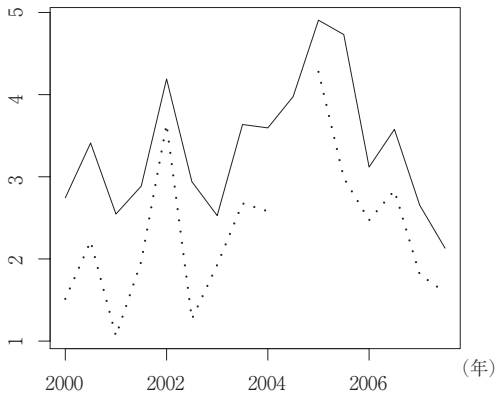

図 3.6　パレ″ート指数

と推定する（ここで $(\sum f_k)$ は中間層の世帯数である）。これを裾クラスの総度数 f_K から差し引けばパレ″ート型集団の度数 $f_K^{(P)}$ も推定できる。度数 $f_K^{(e)}$ に(4)によって推定された裾クラスの平均所得をかけて，裾クラスにおける指数型集団の所得総額を求め，さらに所得総額 $f_K \times m_K$ から差し引いてパレ″ート集団の所得総額を得る。これをパレ″ート型集団の度数 $f_K^{(P)}$ で割れば，パレ″ート型集団の裾クラスの平均所得 $m_K^{(P)}$ が得られる。

図 3.5 は，裾クラスの総世帯数（実線），上のようにして求めた指数型集団の世帯数（破線），パレ″ート型集団の世帯数（点線）である。図 3.6 はモーメント法で推定した α の値である（破線は，混合分布を前提として得たパレ″ート集団の平均 $m_K^{(P)}$ を次節(7)式の $\overline{X}$ に代入して得た推定値，実線は全体をパレ″ート集団と仮定してクラス平均 m_K から求めた）。2004 年 7 月はパレ″ート集団の世帯数の推定値はマイナスになったのでこれを除外した。パレ″ート集団の比率が 0 すなわち分布が指数分布であれば，誤差の影響で $f_K^{(P)}$ がマイナスになることもあり得る。このような月は非勤労者世帯も分布が指数分布でよく説明できていると解釈できる。この他の月も，パレ″ート型集団の世帯数 $\hat{f}^{(p)}$ が，正ではあるが小さいケースが見られる（2002 年 1 月，2004 年から 2005 年，これは不況が深刻化した期間に（少し遅れて）対応すると思われる。これは，図 3.4 を参照すると，指数分布が比較的よく当てはまっている月に一致することがわかる。

パレ″ート集団を分離することで指数 α の値は大きく減少する場合がある（図 3.6）。α の減少は分布の裾が厚くなることを意味するから，混合分布を仮定することにより超高額所得者を含む集団（自営業者層か）の存在が示唆される。上述の不況期ではパレ″ート指数 α の推定値が大きい。一方，図 3.4 で実際の収入が指数分布を大きく上回っている月では α の推定値が小さく，先行研究（Johnson et al., p. 575）でも指摘されているように，2.0 前後の値をとっている。これは諸外国で報告されている所得分布の法則が日本にも当てはまることを示唆している。

Nirei and Souma (2007) は，納税額データを用いた分析を行い所得分布に本稿に類似の混合分布を考えている。彼らはこれを，勤労所得と資産所得の混合と見なしている。いずれにせよ，所得分布に関して複数の母集団の混合であることが，異なるデータに基づいて得られることは興味深い。

4. 標本平均に基づくパレ″ート指数の推論

4.1 モーメント法と最尤法

パレ″ート分布(2)からの標本 $X_1 X_2,\ldots,X_n$ に基づくパレ″ート指数 α の推定は，最尤法，あるいはこれを分布の裾に限定した Hill の推定（Hill, 1975）が知られている。最尤推定量は次のように陽に表現することができる：

$$\hat{\alpha} = n\left\{\sum_{i=1}^{n} \log\left(\frac{X_i}{L}\right)\right\}^{-1}$$

これに対し本稿では，グル″ープデータ，とくに裾クラスの分布あるいは条件付き分布(3)におけるパレ″ート指数 α の推定に関心がある。ここでは，個別の標本値は利用できず，クラス平均 $\overline{X} = \sum_{i=1}^{n} X_i/n$ の

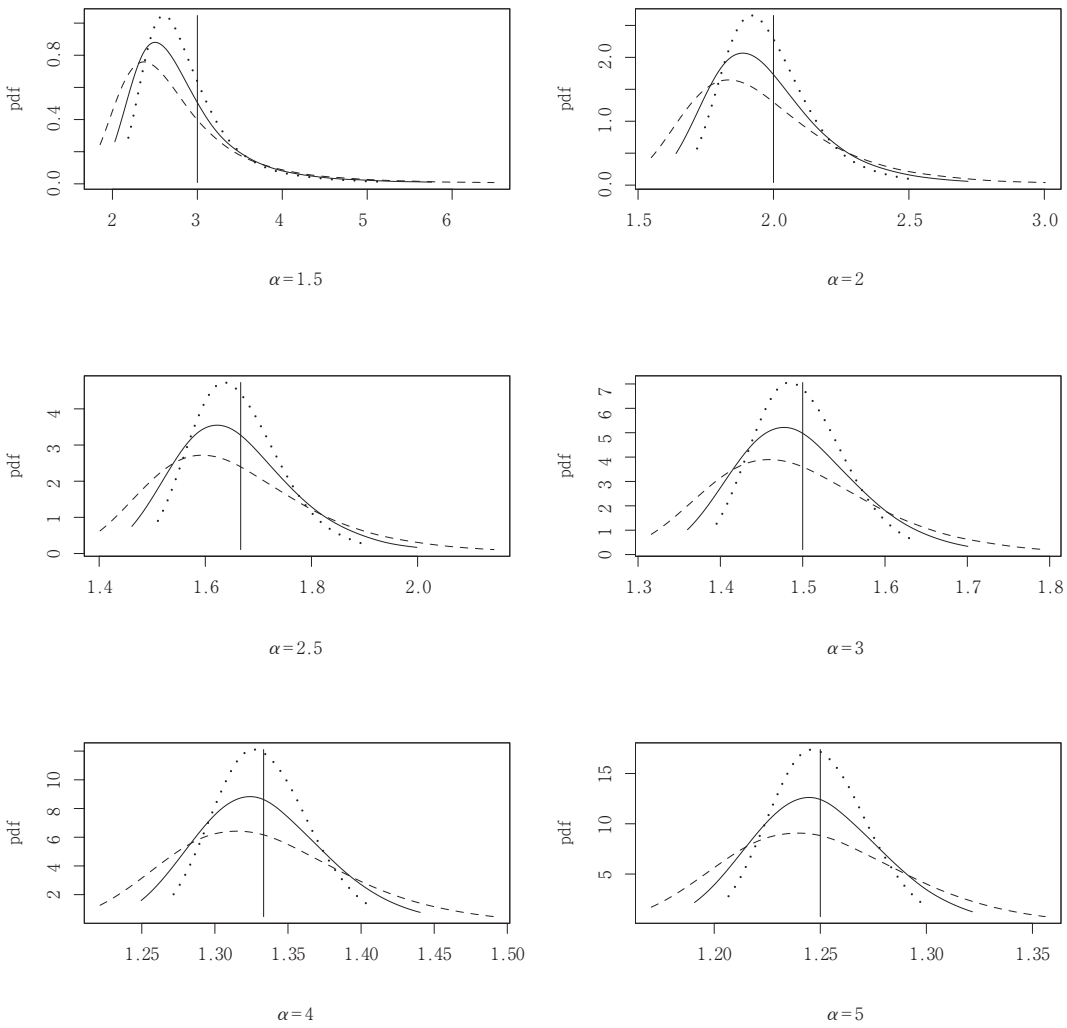

図 4.1　パレ″ート母集団からの標本平均の分布

みが与えられる。この場合には，モーメント法と呼ばれる次のような方法が適用できる：すなわち，$E(\overline{X})=E(X_1)=\frac{\alpha}{\alpha-1}x_0$であるから，この左辺を実現値で置き換えて，さらに α について解けば次のような推定量が得られる：

$$\hat{\alpha}^*=\frac{\overline{X}}{\overline{X}-x_0} \tag{7}$$

4.2　推定量の分布

モーメント法は広く用いられる手法であるが，パレ″ート分布への適用には，他の多くのケースとことなり，標本平均の漸近正規性が保証されないという難点がある。実際，パレ″ート分布の分散は $\alpha\leq 2$ のとき存在せず，$\alpha>2$ のとき，$V(X)=\frac{\alpha L^2}{(\alpha-1)^2(\alpha-2)}$である。また，一般に α 次以下のモーメントが

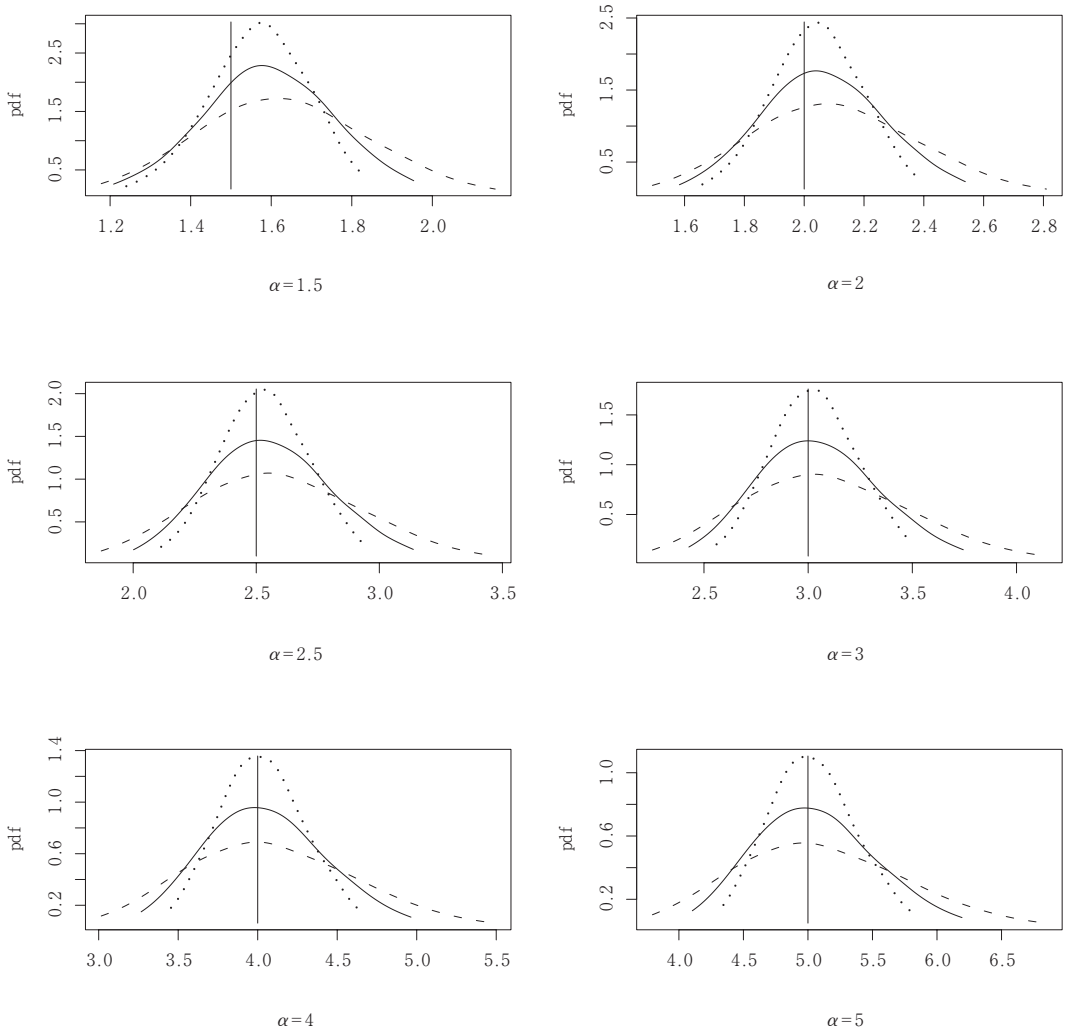

図 4.2　モーメント法による推定量

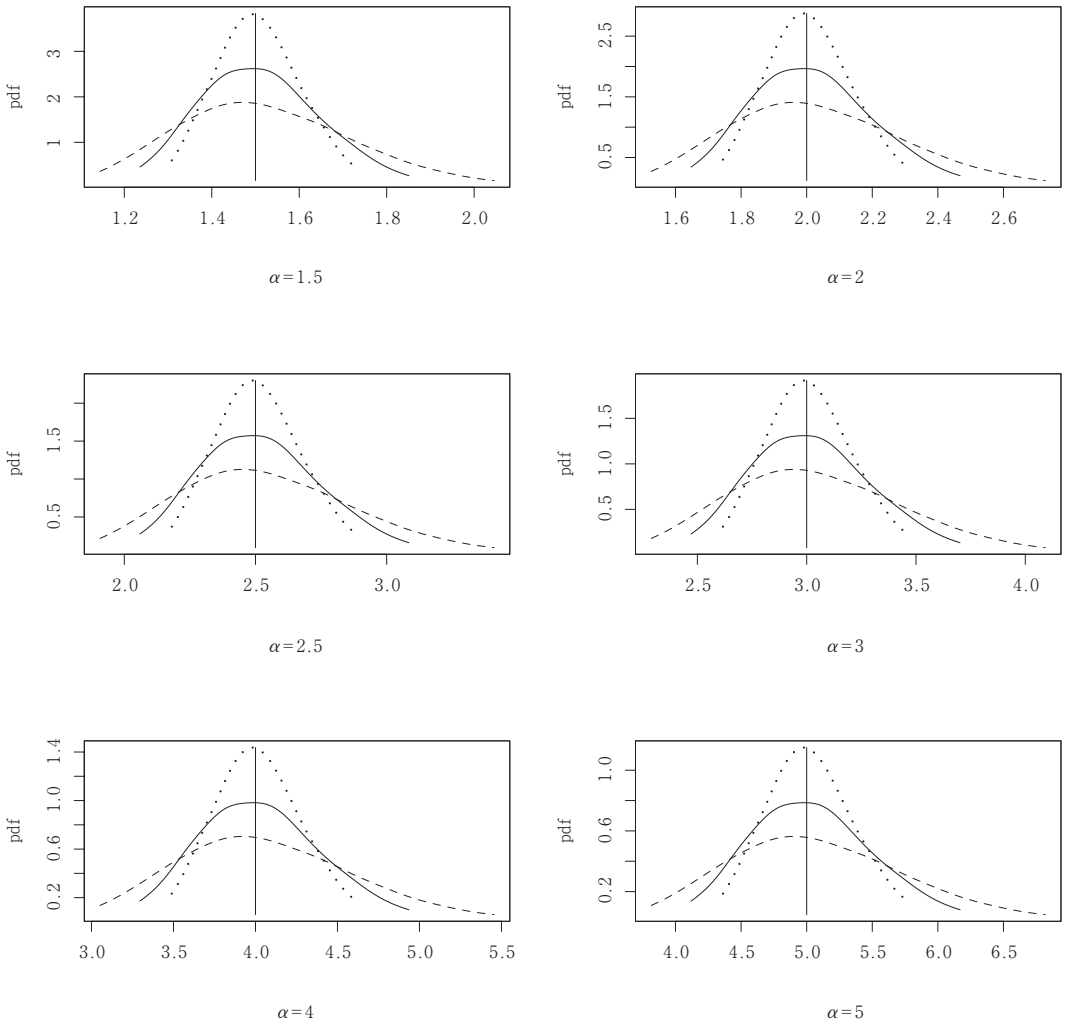

図 4.3　最尤法による推定

存在しない。このため中心極限定理の条件が満たされない。また $\alpha>2$ の場合でも，α が 2 に近ければ正規分布への収束は遅い。図 4.1 に，$\alpha=1.5,2,2.5,3,4,5$ について，$n=50$（破線），$n=100$（実線），$n=200$（点線）のときの，$\overline{X}$の分布を示す[11]。n の増加とともにゆがみが少なくなるが，$n=200$ でもなお $a=3$ 程度以下であると分布のゆがみは消えていない[12]。

図 4.2 は，上の実験で得られた標本平均から(7)によって求めた$\hat{\alpha}^*$の分布である。図 4.1 とは逆に，分布は左にゆがんでおり，$\alpha\leq 2$ では，モーメント法による推定は α を過大推定する傾向がある。

なお，本稿が想定する状況では現実的でないが，完全データを用いた最大推定量の分布は図 4.3 のようになる。

4.3　モーメント法の効率

表 3 は，シミュレーションに基づく$\hat{\alpha}^*$の分散と最尤推定量の分散[13]の比較である。表の値は $n=50,100,200$ について，それぞれの α のときの各推定量の分散である[14]。

n が大きくなるにつれ，モーメント法の効率が落ちていることは，注目に値する。いずれの n でも α が大きいとき，モーメント法の効率は高いが，想定される $1.5<\alpha<2.5$ の範囲で，$n=100$ の場合，効率はおよそ 80%程度である。しかしながら，データをグル〝ープ化する前の完全情報に基づく最尤法に対してこの程度の効率を確保しているのであれば，十分な精度ともいえよう。

表 3　推定量の分散の比較

$n=50$ の場合

α	1.5	2.0	2.5	3.0	4.0	5.0
最尤推定 $\hat{\alpha}$	0.048	0.0854	0.1334	0.1921	0.3415	0.5336
モーメント法 $\hat{\alpha}^*$	0.0557 (0.861)	0.0999 (0.854)	0.1508 (0.885)	0.2105 (0.913)	0.36 (0.949)	0.5516 (0.967)

$n=100$ の場合

α	1.5	2.0	2.5	3.0	4.0	5.0
最尤推定 $\hat{\alpha}$	0.0232	0.0412	0.0644	0.0928	0.1649	0.2577
モーメント法 $\hat{\alpha}^*$	0.0326 (0.712)	0.0548 (0.753)	0.0787 (0.819)	0.1068 (0.869)	0.1778 (0.927)	0.2698 (0.955)

$n=200$ の場合

α	1.5	2.0	2.5	3.0	4.0	5.0
最尤推定 $\hat{\alpha}$	0.0114	0.0202	0.0316	0.0455	0.0809	0.1264
モーメント法 $\hat{\alpha}^*$	0.02 (0.568)	0.0307 (0.659)	0.0414 (0.762)	0.0544 (0.836)	0.0885 (0.914)	0.1332 (0.948)

5. まとめ

本稿では，2000 年から 2007 年の家計調査結果の公表値をもとに，勤労者世帯，非勤労者世帯それぞれについて，指数分布とパレ″ート分布を中心に分布の適合度を検討した。この結果，中流階層以上の勤労者世帯の所得分布には指数分布がよく適合していることがわかった。一方，非勤労者世帯ではどちらの分布も余りよく適合しない。非勤労者世帯分布については，さらに両者の混合分布について検討し，やや不十分であるが，最大所得階層については，両者の混合分布で説明できることを示唆した。

注

1　本稿では，カタカナ化した外来語について，ラ行で表記される音のうち原語で“r”に相当するものには日本語の濁点に相当するものとして″をつけ“l”の音と区別した。
2　世帯主の職業等により分類される。
3　本稿では「勤労者以外の世帯」をこのように表す。
4　このうち非勤労者世帯は，全世帯表と勤労者世帯表の差から得た。
5　メディアン（median），官庁統計関係者はしばしば中位数という。
6　全 16ヶ月の裾クラスの単純平均。
7　参考：$\boldsymbol{X}_i=(X_{1i},X_{2i},\ldots,X_{Ki})\sim M_K(n_i,\boldsymbol{p}_i),i=1,\ldots,N$ ここで，$M_K(n_i,\boldsymbol{p}_i)$ は，それぞれ繰り返し n_i各項の確率が $\boldsymbol{p}_i=(p_{1i},p_{2i},\ldots,p_{Ki})$ である K 項分布。
$\boldsymbol{X}_i=$ の $\boldsymbol{x}_i=(x_{1i},x_{2i},\ldots,x_{Ki})$ の時の $\boldsymbol{p}_i,i=1,\ldots,N$ の対数尤度は $\sum_{i=1}^{N}\sum_{k=1}^{K}x_{ki}\log p_{ki}$ 最大尤度は $\hat{p}_{ik}=\frac{x_{ik}}{n_i}$ のとき

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\boldsymbol{x}_{ki}\log\hat{\boldsymbol{p}}_{ik}$$

一方，仮説 $\boldsymbol{p}_i\equiv\boldsymbol{p}_0,i=1,\ldots,N$ のもとでは，$\hat{p}_{k0}=x_{k.}=\sum_{i=1}^{N}\boldsymbol{x}_{ki}$ で最大対数尤度は

$$\sum_{k=1}^{K}\boldsymbol{x}_{k.}\log\hat{\boldsymbol{p}}_{k0}=n\sum_{k=1}^{K}\hat{\boldsymbol{p}}_{k0}\log\hat{\boldsymbol{p}}_{k0}$$

したがって，尤度比検定統計量は

$$\lambda=-2\sum_{k=1}^{K}\sum_{i=1}^{N}x_{ki}\log\hat{\boldsymbol{p}}_{k0}/\hat{\boldsymbol{p}}_{ki}$$

8　（　）内は，自由度 $(18-1)+(8-1)=119$ の χ^2 分布に基づく検定の $p-$ 値。
9　所得分布のモデルについて McDonald（1984）に包括的な説明がある。
10　指数分布(1)の標準偏差が $1/\beta$ であることを用いて得られる，実際の収入“O”の 95％測区間を矢印で表した。クラスの度数が変化するため，これに応じて区間の幅も変動している。
11　各繰り返し 100,000 回の乱数シミュレーション結果。図中，$y-$ 軸に平行な直線はそれぞれ期待値 $\boldsymbol{E}(\overline{X})$ の位置。
12　本稿で用いた家計調査データの裾クラス（年間収入 1500 万円以上）の世帯数は，78 から 193 である。また多くの文献では $1.5<\alpha<2.5$ であると報告されている（Johnson et al. (1994), Nirei and Souma (2007) など）。
13　シミュレーションであるから標本 $X_1,X_2,\ldots,X_n$ をすべて用い(6)によって求める。本稿で想定している実際の場ではこれを求めることはできない。
14　（　）内は，推定量の効率である。推定量の効率は，最尤推定量の分散を当該推定量の分散で割った値と定義される。

参考文献

Hill, B. (1975) “A Simple General Approach to Inference about the Tail of a Distribution,” *Ann. Stat.,* 3, 1163–1174.

Johnson, N. L., S.Kotz and N.Balakrishnan (1994) *Continuous Univariate Distributions, Vol.1,* New York, John Wiley & Sons.

McDonald, J. B. (1984) “Some Generalized Functions for the Size distribution of Income,” *Econometrica,* 52, 647–665.

Nirei, M. and W. Souma (2007) “A Two Factor Model of Income Distribution Dynamics,” *Rev. Income & Wealth,* 53, 440–459.

（2007 年 11 月 20 日　産業経済研究所受理）