ミクロ計量経済学入門：

第3回　ミクロ統計データの特性と分析手法

北村行伸
一橋大学経済研究所

2006年4月15日

概 要

連載の第3回はミクロ統計データの特性とそれをつかむための基本的統計的性質の分析手法を中心に論じる。ミクロ統計データを扱う上では、データへの第1次接近である記述統計分析が極めて重要である。ここでデータの性質を完全に把握し、それに応じて最適な分析方法や統計的推定方法を決めることができるようになるのである。

1　はじめに

ミクロ統計データを利用して実証研究する時にまず着手しなければならないことは、データがどのような統計的性質をもっているかを把握することである。これは、簡単そうで実はかなり手間と暇のかかる作業である。しかし、これはすし職人がネタの仕込みに時間をかけたり、大工が木材の選定と加工に時間をかけるのと全く同じで、この作業をおろそかにすると、後でいくら高度な技術を用いても、信頼のおける結果を得ることはかなり難しくなる。

ではデータの性質を把握するとは、具体的にどういうことを意味しているのだろうか。まず、データがどいう単位表示に従っているか、データは数量化できるようなものか、そうではなく質的なものなのか、そして標本のなかに含まれている個々の経済主体は同じような行動様式に従っていると考えていいのか、標本の分布はどのような法則に従っているのか等々を丹念に調べるということである。連載の第1回でミクロ計量経済学の課題として多様性への対応ということを第1に挙げておいたが、ミクロ統計データには必ず異質な経済主体が含まれている。問題はその主体を異質だからということで削除してしまうべきか、それともその異質性からなんらかの意味のある情報を引き出すかということである。もちろん、その答えは分析の目的に応じて変わってくるが、経済理論にあるように、すべての経済主体が同一の目的をもって行動しているということは、ミクロ計量経済学では想定していない。むしろ、その異質性の中にこれまで経済学者が気づかなかった行動原理を発見することに期待をかけているのだと言った方がいいだろう。

2　ミクロ統計データの形態と特性

ミクロ統計データはその特性によって、いくつかのカテゴリーに分けることが出来る。

データは大きく分けて２種類、**量的データ**と**質的データ**に分かれる。**量的データ**は経済学で用いる変数、例えば、消費量や労働人口、株価や金利など実数として表記されるデータをさす。量的データはさらに、**離散データ**と**連続データ**に分けられる。離散データとは 1,2,3 のように実数のうち整数部分のみを扱うデータであり、連続データは数学的に定義されるある区間内でとりうる全ての実数を含むデータを意味している[1]。

量的データは実数の定義域に制約がある場合と、制約がない場合に分けることも出来る。実際の経済変数では負の値をとる変数は、債務超過などの赤字であったり、マイナスの変化率（負の成長率）などが考えられるが、多くの変数は非負という制約がかかっている。この場合、実数の定義域は正の実数であり、それ以外は 0 で表現されるようなデータとなることが多い。また、ある条件を満たす標本のみからデータを集めている場合、例えば、大学入試の合格者の成績は、定義によって、最低点以上の成績を収めた人の成績のみに限定される。また、従業員数 50 名以上、資本金 3000 万円以上の企業に限定した企業調査では、例え、その条件に該当する企業全てが答えたとしても、企業全体の集合からすれば、ごく一部しか捉えていないことになる。これらのデータは**切断データ**あるいは**打ち切りデータ**という。これらのデータをミクロ計量経済学で分析する場合を**制限従属変数**と呼ぶことがある。

量的データの単位については注意する必要がある。同じ貨幣価値で表示されているデータであっても質問項目によって円単位、千円単位, 百万円単位と単位がばらばらであることがあり、それを注意することなく無造作に四則演算したりすると問題が生じる[2]。

量的データと質的データの中間形態として**計数 (カウント) データ**がある。これは、交通事故数であるとか 1 キロ平方あたりの人口など、正の整数で表されるものである。これは計数が多ければ一般の量的データとして扱えるが、計数が少なければ、質的データとして分類するのが適当な場合もあり、どのような分析手法を用いるかには注意を要する[3]。

質的データとは数量化されていないが、ある基準で分類できるデータ、例えば、ある政策に賛成か反対かどちらでもないのか、購入するかしないか、企業業績見通しが良いか悪いかなど、分類可能なデータをさす。質的データは一般には、ある状態に属するとき 1、属さないとき 0 となるようなダミー

[1] もちろん、ほとんど全ての経済変数は測定単位があり、それ以下の規模については測定できないという意味では、厳密には離散データであり連続データではない。例えば、人数は 1 人以下（小数点以下）の人数は考えられないし、1 円以下の貨幣価値は測れない。

[2] 前回議論したように、ミクロ統計データの利用者は質問票を良く見て、それぞれの変数の単位については細心の注意を払う必要がある。

[3] 計数（カウント）データの分析手法に関しては Cameron and Trivedi(1998) を参照されたい。

変数として表されることが多い[4]。このようなデータを二項分類に基づく**二項（二値）変数**と呼ぶ。3 つ以上の相互に排他的なカテゴリーに分類されるデータを多項分類に基づく**多項（多値）変数**とよぶ。例えば、ある人の雇用形態に応じて、(1) 常勤、(2) 非常勤、(3) 失業中、(4) 非労働力、に分類されるケースなどがこれに入る。ここでは (1)-(4) が何か意味のある順序に従っている訳ではない。質的データが順序に従って並んでおり、相互に重複しない場合を**順序変数**と呼ぶ。このタイプには例えば、企業格付け（AAA, AA+,A など）のようにランクされているデータ、あるいは満足度調査などで、(1) 大変満足、(2) まあ満足、(3) 中立、(4) やや失望、(5) かなり失望、などのように、(1)-(5) の順に満足度が並べられているデータが考えられる。それぞれのデータに応じて、その統計的性質を分析するための手法が考案されている。

データの分類に関してはその他沢山の見方が出来る。データセットに含まれるものが 1 変数の場合、**１次元データ**と呼び、2 変数以上の場合、**多次元データ**と呼ぶ。データの採集の仕方によって、同じ対象を一定期間毎に調査したデータであれば**時系列データ**、一時点の横断的なデータであれば**クロスセクションデータ**、時系列と横断データを併せ持ったデータ、すなわち、多数の対象に対して繰り返し調査を行う場合は**パネルデータ**と呼ぶ。また、厳格な数値データに限らず、意識や判断も含めたデータは**サーベイデータ**と呼び、これも面白い経済情報を含んでいるので広く用いられいている[5]。また一定期間累積したデータである**ストックデータ**と一定期間に流通したデータである**フローデータ**を区別することも重要である。

3　ミクロ統計データの記述統計

さて、ミクロ統計データを実際に分析し始める段階にたどり着いた。まず利用するミクロ統計データがどのような性質を持ったデータであるかを知る必要がある。そのための簡便な分析方法を**記述統計学**と呼ぶ。我々が通常利用するのは複数の変数を含んだ多次元データであるが、まず、それぞれの個別変数についてその統計的性質を調べよう。

まず、個別のデータの観測値をいくつかの**階級**に分けて、その階級にいくつの観察値が含まれるか（これを**度数**と呼ぶ）を数えて表にしたものを**度数分布表**といい、これをグラフ化したものを**ヒストグラム**と呼ぶ。データの性

[4]質的データの分析手法としては林（1974）, 岩坪 (1987) や西里 (1982) を参照されたい。またデータをカテゴリーに分類して分析する手法をカテゴリカルデータ解析と呼ぶ。この手法については Agresti(2003) を参照。

[5]我が国で最も有名なサーベイデータは日本銀行短期経済観測調査（日銀短観）であろう。ここには業況判断ディフュージョン・インデックス (DI) が含まれており、そこでは各企業に対して景気が「良い」、「さほど良くない」、「悪い」について回答をしてもらい、そこから「良い」－「悪い」の社数構成比% ポイントを計算することで、景気判断をしている。これらのサーベイデータの経済学への利用については加納 (2006) を参照されたい。

質を一目で理解するにはこのヒストグラムを描くのが最も良い方法である。ヒストグラムを描くときに問題になるのは、階級をどれぐらいの数とるかということであるが、これには厳密なルールはなく、観察値の多寡によって試行錯誤しながら決めればよい。同様に階級幅も区切りの良い値をとるのが一般的であるが、これも試行錯誤的に決めるしかない。

ヒストグラムを描くことで変数の度数分布を直感的に知ることができる。多くの経済変数は峰が 1 つの**単峰型**の左右対称な分布をしている。多くの場合、このような変数は正規分布に従っていると仮定できる。単峰型であっても一方に歪んだ分布であることもあるが、この変数を対数変換すれば対称分布に変わることもある。峰が 2 つ以上の**双峰型**分布の場合、性質の異なる標本が混合されていることが考えられる。この場合、データをいくつかの単峰型分布に分離する作業を行うこと（これを**層別**という）が有益である[6]。

度数分布表からはヒストグラムの他にも、累積度数や累積相対度数のグラフも有用な情報源となり得る。また 2 次元データに対して、それぞれの変数の累積相対度数を縦軸と横軸にとったものを**ローレンツ曲線**と呼ぶ。これは、所得や資産が人口にどのように分布しいてるかを見る時によく用いられる[7]。

さて、分布に関してはヒストグラムで概略はつかめたとして、次ぎに統計的な代表値を求めよう。最も一般的な統計量は**平均**あるいは**算術平均**である。これは観察値の総和を観察値の総数で割ったものである。多くの場合、平均値は変数の統計情報として有益ではある。先に見たようにヒストグラムが対称分布をしていれば平均値をもってこの変数の代表値としてもいいだろう。しかし、ヒストグラムが左右に歪んでいる場合は平均値だけを見ても統計情報としては不十分である。それに代わる情報としては分布のちょうど中位 (50% 地点）にあるデータ、**中位値**(median あるいは**中央値**) を用いることもある。これは、平均値では極端な値をとる異常値（外れ値）に数値が引っ張られてしまうが、中位値は、純粋に数値を低い順位に並べて 50% の位置にいるサンプルの値を表しているにすぎないので、異常値に左右されることはない。中位値の考え方を拡張して、 4 分位 (quartile：25% 、50% 、75% でデータを区切る) や 5 分位（quantile:20% 、40% 、60% 、80% でデータを区切る）などの情報を用いることもある[8]。しかし、分布の中央に位置していることが、そのデータを代表していることには必ずしもならない。例えば、所得分布のようにほとんどの人が 1000 万円以下の所得であるが、それでも 15% を越える人に 1000 万円以上の所得があれば、平均値も中位値もかなり高所得者に

[6]例えば、東京大学教養学部統計教室（編）(1991、pp.21-22) で論じられているように、住宅面積の分布に持家と借家が混在している場合、双峰型分布になるが、持家、借家に別ければ単峰型分布になる。

[7]不平等度の指標である**ジニ係数**はローレンツ曲線と 45 度線との間の弓形の面積と正方形との比率を 2 倍したものとして計算できる。

[8]実際それぞれのデータの値を分位ごとの平均からの差を最小にするように回帰に行って、パラメータを求める手法を Quantile Regression 推定というが、これもデータの分布が非対称な場合には有効な方法である。Koenker(2005) 参照。

引っ張られる。しかし最も多くの人の得ている所得は 500 万円台であるという場合、この最も多くの人が得ている所得を表すのは度数分布が最大である階級を指している。これを統計学では**最頻値**(mode) と呼んでいる[9]。一般に、データの分布で峰が左に偏っているような分布では平均値 > 中位値 > 最頻値と並び、分布が右に偏っている場合には逆に並ぶことが知られている。また記述統計として、最大値、最小値あるいはその拡張である、最大 5 つの値、最小 5 つの値などの情報も有益である[10]。

次に、データのばらつき度合いを測る指標としては、**分散**と**標準偏差**を用いることが多い。時間を経て規模が拡大するような変数の場合には、標準偏差を異時点間で比較することはあまり意味がない。その場合には、標準偏差を平均で割った**変動係数**を用いて比較することがある。

次に、２変数以上の多次元データの記述統計を考えよう。多次元データの分析で最も大切なことは、変数間の関係を見出すことにある。関係と言っても変数 x から変数 y への**因果関係**（例えば、猛暑と電力消費）を示すこともあれば、単なる**相関関係**（例えば、体重と身長）であることもある。あるいは同時に第 3 の変数に影響を受けて変動している場合（例えば、ビールとコーラの消費の相関は気温に影響を受けている）もある。このような、様々な関係をデータを丹念に見ながら確定していくことができれば分析は半分終わったも同然である。

最も簡単な方法は２変数の**散布図**を描くことである。２変数の間に強い関係があれば、散布図上で何らかの規則性が見出されるはずである。そのような規則性が全く見出されなければ、２変数を結びつける関係はあまりなさそうだと言える。では、どの２変数を選んで関係を調べればいいのだろうか。第一に経済理論に基づいて、理論上の関係を確認するという方法があるだろう。これは、最も一般的な方法であり、時間の無駄を省くことが出来る。第二に、しかし、全ての経済関係に対応した経済理論が存在している訳ではないし、経済理論も実証データに裏づけされているものもあれば、されていないものもある。もし理論的手がかりがなければ、手元にある多次元データから２変数のペアを網羅的に選び出して散布図を描いてみればよい[11]。多次元データには量的データだけでなく質的データも含まれている。質的データの場合は散布図を描くことはできないので、その場合には**分割表**(**クロス表**) を用いて相対度数などを表示することで、関係を見つければよい。具体的には、

[9] この統計値の問題点は全く同じ度数のピークが２つ以上ある場合には、１つの代表値として表現できないことである。

[10] データの分布が対称分布をしていて正規分布に近似できると判断できるときは、平均値 ±3× 標準偏差は 0.9973 の範囲に収まっているはずである。すなわち、標本の 0.3% 程度が平均値 ±3× 標準偏差を超える範囲にいれば、正規分布に従っていると言えるが、それをかなり超える標本が範囲外に存在していれば、それは正規分布よりも裾野の厚い (fat tail という）分布であることを意味する。金融データの多くはそのような fat tail 分布に従っていることが知られている。

[11] もちろん、常識的にあり得ないペアまで考慮する必要はない。事前の情報でかなりのペアは削除できるはずである。

質的データに応じて、量的データも適当な階級に場合分けし、項目化することによって、2 変数の関係を分かりやすく表現することが求められる。

２変数の関係を統計的に測る方法としては、変数 x と変数 y の相関係数 r_{xy} を用いるのが最も一般的である。これはより厳密にはピアソンの**積率相関係数**と呼ばれているもので、次のように定義されている。

$$r_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \tag{1}$$

ここで $\overline{x}$、$\overline{y}$ はそれぞれの平均値、$-1 \leqslant r_{xy} \leqslant 1$ である。分母は x、y の標準偏差の積を表し、分子は x と y の**共分散**を表している。

ただ注意しなければならないのは、相関が高くても**見せかけの相関**であることが往々にしてある。先に示したように、ビールとコーラの消費の相関は気温上昇という第 3 の要因によって説明されているのであって、コーラとビールの消費には強い因果関係も、補完関係もないと考えられる。このような変数の取り扱いには注意を要する[12]。

質的データであっても順位を決めることのできる順序変数であれば、そのような２変数の順位の相関をしめす**順位相関係数**も利用されることがある[13]。

ここまでの作業は、ミクロ統計データを本格的に計量経済学を用いて利用する前の下準備にあたる。これは丁度、すし職人がネタの入手と仕込みをしている段階に相当する。実際のすし職人は仕込みに長い時間をかけ、すしの握り自体には 10 秒もかけないというのに似て、良い結果を出すためには、質の良いデータの性質を徹底的に分析し、最適な処理を行った上で、分析に入るというのが、理想的な実証分析のあり方だと筆者は考えている。

そういう意味で、この段階でデータの性質をほぼ完全に掌握しておくことが肝心である。いったん実証分析を始めて、おかしな結果に気づき、またデータに戻って整理をしなおす事も間々あるが、事前のデータ加工の段階に全研究期間の 50%-70% をかけることで、後の実証分析が格段に容易になることは筆者の経験が教えてくれることである。また、この段階で、漠然とデータを処理しているだけではなく、データから何が言えるのか、言えないのかを徹底的に考えるべきである。このデータ処理の段階で手を抜くと、後でいくら高度な分析手法を用いても誤った結果を得る危険性が高くなることも認識しておくべきである。

筆者の個人的経験では、ミクロ統計データを探索しているこの段階が最も楽しく、新しい発見が多い。ミクロ統計データ自体はマクロ経済データや金融データと比べれば、外れ値や欠損値、脱落サンプルなど様々な予想外の事

[12]統計的には偏相関係数を計算することもできるが、この段階で偏相関を求めてもあまり意味がないことが多いので、見せかけの相関かどうかを確定することが重要である。また、全体として相関は見られないが、標本をグループに分割（層別化）すると、特定のグループでは強い相関が見られることもある（例えば、高齢者の医療支出と他の支出との負の相関）。

[13]順位相関係数にはスピアマンの提示した指標とケンドールの提示した指標がある。詳細については例えば、東京大学教養縛部統計教室（編）(1991、pp.54-55) を参照されたい。

態に直面する可能性が高く、実際の計量分析を行う前の段階で悩むことが多い[14]。例えば、標本数 5 万を越すデータであれば、個々の識別に使う**識別番号**（id）にさえミスや混乱が含まれていることがある。それらを手際よく処理していくにはある程度の熟練が必要になるが、ミクロ統計データを使う人のための工房や職人養成所がある訳ではないので、個々人がデータにじっくりと向き合って、データと格闘し、データの扱いに慣れていくしかない。

4 確率変数と確率分布

ミクロ統計データを統計学の枠組みで分析するためには、用いるデータが確率に従って発生していると仮定し、確率変数が特定の分布に従っていると考えるのが一般的である。以下では、その考え方を簡単に紹介したい。

確率変数とはそれがとる値に対して確率が与えられている変数をさす。変数の取りうる値はサイコロの数字のように離散型であったり、実数全てを含む連続型であったりする。ここでは簡単化のため連続型の確率変数を考える。

確率変数 X のとる確率は関数 $f(x)$ によって次のように定義できる時、X は連続型確率分布を持つという。

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{2}$$

ここで確率の定義により、全ての x に対し、$f(x) \geqq 0$、$\int_{-\infty}^{+\infty} f(x)dx = 1$ である。ここで $f(x)$ を X の**確率密度関数**という。これは前節で論じたヒストグラムの厳密な数学的表現であると考えてよい。ある値以下の確率を知りたいときには、確率変数 X に対して、x 以下の確率を X の**累積分布関数**と呼び、次のように定義する。

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du \tag{3}$$

ここで $F'(x) = f(x)$ である。この累積分布関数は次の 3 つの性質を持っている。(1)F は単調非減少 ; $x_1 < x_2$ なら $F(x_1) < F(x_2)$、(2) $\lim_{x \to -\infty} F(x) = 0$、$\lim_{x \to \infty} F(x) = 1$、(3)$F$ は右連続 ; $\lim_{x \to a+0} F(x) = F(a)$. また、以上の 3 つの条件を満たす任意の実数値関数 $F(x)$ は分布関数と呼ばれている。

前節で記述統計として平均、分散、標準偏差などについて論じたが、ここでも確率変数の代表値として次のものを考えてみよう。

[14]例えば、欠損値がいくつかの変数についてあったとしても、その経済分析にとってそれほど重要な変数でなければ、それは無視して標本に残しておくべきである。すなわち、どの変数が最終的に重要になるかは、事前には分からないので、出来るだけ多くの標本をできるだけ最後まで残しておくべきである。また外れ値の処理も同じで、外れ値をとる標本が分析の中で、明らかに異質であり、統計推測上バイアスを生じさせることが確定されれば、削除してもかまわないが、外れ値をとるような標本が社会に存在していることは統計上重要な情報であるので、容易には削除すべきではない。

確率変数 X の加重平均を確率変数の**期待値**$E(X)$ と呼び次のように定義する。

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{4}$$

次に**分散**$V(X)$ は期待値 $E(X) = \mu$ とおくと、次のように定義できる。

$$\begin{aligned} V(X) &= \int_{-\infty}^{\infty} (X-\mu)^2 f(x) dx \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - (E(X))^2 \end{aligned} \tag{5}$$

標準偏差$D(X)$ は分散の平方根で定義される。$D(X) = \sqrt{V(X)}$

さらに、分布の非対称性を表す指標として**歪度**（skewness）SK を次のように定義する。

$$SK = E(X-\mu)^3/\sigma^3 \tag{6}$$

$SK > 0$ がならば、右すそが長く、$SK \leq 0$ ならば、左すそが長い。

もう一つの指標として分布の**尖度**（kurtosis）KT も次のように定義できる。

$$KT = E(X-\mu)^4/\sigma^4 \tag{7}$$

正規分布のとき $KT = 3$ となることが知られているので、**過剰尖度**(EK=excess kurtosis) を定義して、使うことが多い。

$$EK = KT - 3 \tag{8}$$

$EK > 0$ なら正規分布より尖っており、$EK \leq 0$ なら正規分布より丸いことを意味している。

統計学や経済学でよく用いる標準的な分布には次のようなものがある。

１回の試行で特定の事象 A が起こる確率を p とし、A の起こる回数を X とすれば、X は確率変数と考えられる。その確率密度関数 $f(x)$ は次のように表せる。

$$f(x) =_n C_x p^x (1-p)^{n-x}, \quad x = 0, 1, 2, ...n \tag{9}$$

この確率分布を**二項分布**という。この分布の期待値は $E(X) = np$、分散は $V(X) = np(1-p)$ となる。二項分布において np を一定の正数 λ に保ちながら $n \to \infty, p \to 0$ となる極限では二項分布 $f(x)$ は $e^{-\lambda}\lambda^x/x!, x = 0, 1, 2..$ に近づく。これを**ポアソンの小数法則**という。

$$f(x) = e^{-\lambda}\lambda^x/x!, \quad x = 0, 1, 2.. \tag{10}$$

この確率分布をポアソン分布 $Po(\lambda)$ と呼ぶ。ポアソン分布の期待値は $E(X) = \lambda$、分散は $V(X) = \lambda$ であり、分散と期待値が等しい。ポアソン分布は小数法則の条件から明らかなように、多くの事象の中でめったに起こらない事象の確率分布を表しており、交通事故や破産件数、火災件数などリスクや安全性に関する分析の時に用いられる。またポアソン分布はスポーツ統計にもよく当てはまることが知られている[15]。

二項分布もポアソン分布も離散分布で整数値をとる確率変数の分布群であったが、連続分布で最も代表的な確率分布は**正規分布**である。この分布は次のような確率密度関数 $f(x)$ に従っている。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty \tag{11}$$

ここで、正規分布の期待値は $E(X) = \mu$、分散は $V(X) = \sigma^2$ である。正規分布はこの 2 つのパラメータで表現できるので $N(\mu, \sigma^2)$ と表現されることも多い[16]。定数項 $1/\sqrt{2\pi}\sigma$ は次の関係から導かれたもので、この分布を $\int_{-\infty}^{\infty} f(x)dx = 1$ とするための規格化定数である。

$$\int_{-\infty}^{\infty} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx = \sqrt{2\pi}\sigma \tag{12}$$

正規分布は人間の身長や胸囲、試験の成績など、数多くの事象を近似できることが知られている[17]。

ミクロ統計データは多数の変数について調査しており、実証研究では幾つかの確率変数が同時に発生する状況を想定している。ここで k 個の確率変数からなる**同時確率密度関数**を考えよう。

$$f(x_1, x_2, ...x_k) \geqq 0 \ \text{ かつ } \int\int ... \int_S f(x_1, x_2, ...x_k) dx_1 dx_2 ... dx_k = 1 \tag{13}$$

ここで S は標本空間（確率の定義域）を表す。事象 A が起こる確率は次のように表せる。

$$P((x_1, x_2, ...x_k) \in A) = \int\int ... \int_A f(x_1, x_2, ...x_k) dx_1 dx_2 ... dx_k \tag{14}$$

[15] 2002 年 FIFA ワールドカップの 1 次リーグの全 48 試合について各チームが挙げた得点は平均で 1.344 でそれをポアソン分布に当てはめると極めてよく当てはまることが報告されている。http://takanaka-akio.cool.ne.jp/etc/fifa_poisson/を参照。

[16] 確率変数 X を $(X-\mu)/\sigma$ と基準化すると、正規分布も $\mu = 0$、$\sigma = 1$ の基準正規分布に従い、$N(0, 1)$ と表現される。

[17] 経済分析では二項分布、ポアソン分布、正規分布以外の分布の方が実際のミクロ統計データに近似していることもある。例えば、幾何分布、負の二項分布、一様分布、指数分布、ガンマ分布、ベータ分布、ワイブル分布、コーシー分布、パレート分布などについては数学的表現が明らかになっており、それを実証研究で用いることもある。詳しい統計的性質や数学的表現に関しては蓑谷 (2003) を参照されたい。また統計的に分布を推定する方法については Silverman (1986) を参照。

確率変数 X_i のそれぞれ単独の確率分布 $F(x_i)$ は**周辺確率分布**と呼ばれている。

もしここで周辺確率分布の積が同時確率分布と等しくなるならば、X_i と X_j は互いに**独立**であると言える。

$$F(x_1, x_2, ...x_k) = F_1(x_1)F_2(x_2)F_3(x_3)...F_k(x_k) \tag{15}$$

これは、それぞれの確率変数の発生は他の確率変数の発生と全く関係がないことを意味している。独立していれば無相関を意味するが、無相関だからといって独立しているとは限らない。その意味で、独立の方が無相関より強い概念である。

一般的には X_i と X_j の共分散はゼロではなく、ある程度の相関関係が見出される。前節で相関関係を分割表で表す方法を論じたが、その確率論的表現は X_i が x_i で与えられた時の X_j の条件付確率密度関数として次のように表せる。

$$g(x_j|x_i) = f(x_j, x_i)|h(x_i) \tag{16}$$

x_j に関して和をとると、

$$\sum_{x_j} g(x_j|x_i) = \sum_{x_j} f(x_j, x_i)|h(x_i) = h(x_i)/h(x_i) = 1 \tag{17}$$

となり確率分布の条件を満たしている。この条件付確率分布の条件付期待値と条件付分散はそれぞれ次のように表すことができる。

$$E(X_j|x_i) = \int_x x_j g(x_j|x_i) = \mu_{x_j|x_i} \tag{18}$$

$$V(X_j|x_i) = \int_x (x_j - \mu_{x_j|x_i})^2 g(x_j|x_i) dx_j \tag{19}$$

ある事象の発生が他の事象の発生に条件付けられていることが明らかな場合には、実証分析においてもその事実を反映させなければならない。すなわち、これは何らかの因果関係を示しているとすれば、そのようなモデル化が必要になる。あるいは少なくとも場合分けして条件をそろえた上で分析しなければ統計的推測にバイアスがかかる可能性があることは認識すべきである。

ここで確認しておきたいことは、記述統計で探索したことが、ほぼ平行する形で、確率変数の分析として行えるということである。これによって、ミクロ統計データを確率変数を扱う形で展開されるミクロ計量経済学と結びつけることができるのである。

5 最尤法

ミクロ計量経済学では多くの分析が意志決定に関する二項（2 値）変数に関するものであったり、その他の質的データを扱うことが多い。通常の量的

データを線形モデルで推定するのであれば、最小二乗法を用いれば良いが[18]、それ以外の非線形推定に関しては最尤法を用いることが多い。そこで、以下では最尤法の基本的な考え方を概観しておきたい。

標本データ $\mathbf{y} = (y_1, \cdots y_n)\prime$ を所与として、未知母数 $\theta = (\theta_1, \theta_2, \cdots \theta_p)'$ の関数を**尤度**とよび $L(\theta)$ と表す[19]。ここで尤度 $L(\theta)$ を最大にする θ の値 $\tilde{\theta}$ は最尤推定量と呼び、標本データで評価したときに最大確率を起こりうる θ を推定したことになる。$L(\theta)$ の代わりに対数をとった $logL(\theta)$ を最大にしても、最尤推定量 $\tilde{\theta}$ は推定できる。

具体的に接片ゼロの単回帰モデルを考えてみよう。

$$y_i = \beta x_i + \varepsilon_i \qquad i = 1, \cdots, n \tag{20}$$

ここで誤差 ε_i は正規分布 $N(0, \sigma^2)$ に従うとすると、対数尤度は次のように表せる。

$$\begin{aligned} \log L(\beta) &= \log \left[(2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{(\mathbf{y} - \beta\mathbf{x})'(\mathbf{y} - \beta\mathbf{x})}{2\sigma^2} \right\} \right] \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \beta\mathbf{x})'(\mathbf{y} - \beta\mathbf{x}) \end{aligned} \tag{21}$$

これを最大化する β は最小二乗解になる。また最尤推定量は

$$\tilde{\beta} = \hat{\beta} = \mathbf{x}'\mathbf{y}/\mathbf{x}'\mathbf{x} \tag{22}$$

となる。

ここで $\log L(\beta)$ の 2 階微分

$$\frac{\partial^2 \log L(\beta)}{\partial \beta^2} = -\mathbf{x}'\mathbf{x}/\sigma^2 \tag{23}$$

は β^2 の係数であり、対数尤度関数 $\log L(\beta)$ の頂点の曲率を表す量となっている。別の言い方をすれば、$\tilde{\beta}$ の推定量の分散に関する情報を表しており、**フィッシャー情報量**$I(\theta)$ と呼ばれている。これは対数尤度の 2 階微分は標本 $\mathbf{y}$ に依存するので、$\mathbf{y}$ が密度関数 $f_\theta(y)$ に従っているとき、期待値を取ると

$$\begin{aligned} I(\theta) &= -E \left\{ \frac{\partial^2 \log L(\theta)}{\partial \theta^2} \right\} \\ &= -E \left\{ \frac{\partial^2 \log f_\theta(\mathbf{y})}{\partial \theta^2} \right\} \end{aligned} \tag{24}$$

と表わすことができる。

[18] 線形関数の場合、最尤法推計は最小二乗法推計と一致する。その意味では最尤法がミクロ計量経済学の最も基本的な推定法であると言える。

[19] 本節は東京大学教養学部統計学教室（編）（1992、第 4 章）を参照している。

最尤法による不偏推定量の分散の下限は、フィッシャー情報量 $I(\theta)$ を用いて次のように表せる。

$$V\{t(\mathbf{y})\} \geqq \frac{1}{I(\theta)} \tag{25}$$

これを**クラメール・ラオの不等式**と呼び、右辺を**クラメール・ラオの下限**とも呼ぶ。クラメールラオの下限をとる不偏推定量は**有効推定量**という。

最尤法では、パラメータ θ の有意性の検定に Z 値を使う。これはパラメータの分布が帰無仮説 $H_0 : \theta = \theta_0$ の下で $\sqrt{n}(\tilde{\theta} - \theta_0)$ が漸近的に正規分布 $N(0, 1/I_1(\theta_0))$ に従うことを利用して、$\theta_1 > \theta_0$ の場合に棄却域

$$\sqrt{nI_1(\theta_0)}(\tilde{\theta} - \theta_0) > Z_\alpha \tag{26}$$

として計算したものである。ここで $I_1(\theta) = I(\theta)/n$ はデータ 1 個あたりのフィッシャー情報量であり、Z_α は標準正規分布の上側確率が α となる水準を表している。

また最尤法では帰無仮説が複数の制約式からなる場合、Z 値ではなく、カイ二乗分布に基く**尤度比検定**を行う。

ここで帰無仮説 $H_0 : \theta = \theta_0$、対立仮説 $H_1 : \theta \neq \theta_0$ とすると、H_0 下では、最尤推定量を θ_0 と漸近分散で規準化したものの二乗は、漸近的に自由度 1 のカイ二乗分布 $\chi^2(1)$ に従うので次のような関係が成り立つ。

$$2\log\frac{L(\tilde{\theta})}{L(\theta_0)} > \chi^2_\alpha(1) \ \ (= Z^2_{\alpha/2}) \tag{27}$$

これは尤度比検定の棄却域 $\chi^2_\alpha(1) = Z^2_{\alpha/2}$ を表している。

代替的な検定としては**ワルド検定**や**ラグランジュ乗数検定**がある[20]。

最尤法は計量経済学の中では、もっとも広く利用されている推計方法である。確率変数の関数型を特定する必要があり、それが必ずしも現実のデータに当てはまらないという限界はあるが[21]、関数型が特定化されており、パラメータを推計できることは経済学的な解釈が行いやすいことも意味している。

6 おわりに

今回は連載の第 3 回ということで、ミクロ統計データの第 1 次的なアプローチとしての記述統計の方法およびそれに対応した確率論の考え方を紹介した。

[20] これらの検定の概説は北村 (2005、第 1 章) を参照されたい。

[21] この点に関しては最近、分布の関数型を特定化することなくノンパラメトリックに最尤推定を行う Empirical Likelihood 推定の手法が開発されてきている。Mittelhammer, Judge and Miller (2000) や Owen (2001) を参照。

ここでは、ミクロ統計データを利用するにあたって、記述統計のレベルでデータの特性を徹底的に把握しておくことが望ましいことを繰り返し論じた。筆者はデータへの１次的接近をしながら、ああでもないこうでもないと色々な仮説をたてたり、データ処理の方法を考える時間が最も楽しい。ここには、生のミクロ統計データを扱った人だけにわかる苦労と喜びがある。

計量経済学上の分析技術は日々進歩しているが、ミクロ統計データを分析し始める時のなんとも労働集約的な作業は変わりがないようである。まだミクロ統計データを扱ったことのない人は、是非、きれいにクリーニングされたデータを使うのではなく、色々な処理の必要な手つかずのデータから使い始めてみることをお勧めする。それでミクロ計量経済学にすっかりはまりこむか、もう二度の近づこうとしないかの分かれ目になるかもしれないが、かなりの数の人がその楽しさにはまるものと確信している。

参考文献

[1] Agresti, Alan (2003) 『カテゴリカルデータ解析入門』、サイエンティスト社

[2] 岩坪秀一 (1987)『数量化法の基礎』、朝倉書店

[3] 加納悟 (2006)『マクロ経済分析とサーベイデータ』、岩波書店

[4] 北村行伸 (2005) 『パネルデータ分析』、岩波書店

[5] 東京大学教養学部統計教室（編）(1991)『統計学入門』、東京大学出版会

[6] 東京大学教養学部統計教室（編）(1992)『自然科学の統計学』、東京大学出版会

[7] 西里静彦 (1982)『質的データの数量化』、朝倉書店

[8] 林知己夫 (1974)『数量化の方法』、東洋経済新報社

[9] 蓑谷千凰彦 (2003)『統計分布ハンドブック』、朝倉書店

[10] Cameron, A.C.and Trivedi, P.K.(1998) *Regression Analysis of Count Data*, Cambridge University Press.

[11] Cameron, A.C. and Trivedi, P.K.(2005) *Microeconometrics: Methods and Applications*, Cambridge University Press.

[12] Davidson, Russell and MacKinnon, James G.(2004) *Econometric Theory and Methods*, Oxford University Press.

[13] Koenker, Roger. (2005) *Quantile Regression*, Cambridge University Press.

[14] Mitterlhammer, Ron C.,Judge, Gerorge G. and Miller, Douglas, J.(2000) *Econometric Foundations*, Cambridge University Press.

[15] Owen, Art B.(2001) *Empirical Likelihood*, Chapman & Hall

[16] Silverman, B.W.(1986) *Density Estimation for Statistics and Data Analysis*, Chipman & Hall.

[17] Winklemann, Rainer and Boes, Stefan.(2005) *Analysis of Microdata*, Springer.

[18] Wooldridge, Jeffrey. M.(2003) *Econometric Analysis of Cross Section and Panel Data*, The MIT Press