マッチングとアグリゲーションの関係

前　田　敬四郎

序

科学というのは，絶えざる分析と総合を繰り返すことによって発展されて行くものである。経済学もその例外ではない。経済学の発展は，マクロ分析とミクロ分析の二つによって押し進められて来たことは衆知の処である。

然しながら，不幸なことに，経済学はまだマクロとミクロ分析の間に，コンシステントな関係を樹立することに成功していない。此処から，経済分析は，マクロがよいのか，ミクロの方がよいのかという問題が生じて来る。1970年代に，経済分析はマクロによる方がよいとする *Griliches*[1] とミクロを主張する *Orcutl*[2] による論争が展開されたのは，私の記憶に未だ新しい。現実の経済分析を行うに当って，マクロとミクロ分析の間には，「*Information Loss* と *Specification Error*」の二要因がトレード・オフ関係にある。アグリゲーションの度合が高くなれば，*Information Loss* が大きくなり，デス・アグリゲーションが進めば，*Specification Error* が大きくなるという傾向を持っている。*Optimum* な経済分析を求めるには，*Information Loss* と *Specification Error* の総計が最小になるようなアグリゲーション段階の分析を探がさねばならぬ。[3]

経済分析を行うに当って，最初に遭遇する問題は統計のデータであり，マクロ経済データとミクロ経済データの収集並びに活用の方法である。経済データに溯って，経済理論を考え直して行こうとする *Ruggles* を中心とする *National Bureau of Economic Research* の人達がいる。本稿では，経済のアグリゲーションの理論を，*Matching* という統計データの分野に展開した *Nancy And Richard Raggles* の「*Strategy For Merging and Matching Microdata Sets*」[4] を取り上げることにしよう。

註

(1) *Z. Griliches and Y. Grunfeld, Is Aggregation necessarily bad ? The*

Review of Economics and Statistics, 42 (*February* 1960) *PP*1～13。
(2) *Guy, H. Orcutt and J. B. Edwards, Should Aggregation Prior to Estimate be the Rule ? The Review of Economics and Statistics. Vol. LI, November* 1969. *PP.* 409～420。
(3) *E. Malinvaud, L' Agrégation dan les Modéles Éeconomigues, Cahiers du Séminaire d'Econométrie, No.* 4, *Paris*, 1956, *PP.* 69～146。
前田敬四郎，アグリゲーションの功罪。金沢大学法文学部論集，経済学篇20，1973，*PP.* 1～49。
(4) *Nancy and Richard Ruggles, A Strategy for Merging and Matching Microdata Sets, Annals of Economic and Social Measurement*, 1974. *PP.* 353～371。

マッチングの発展

過去，拾数年の間に，各家計や個人に関する情報の標本であるデータ・セットが，経済分析の主要な道具として出現して来た。これらのミクロデータ・セットは，*natioral accounts*の代替や補完物として考慮され得る。

例えば，アメリカ商務省・経済分析局の研究は[1]，家計部門の所得分配を研究する際に，ミクロデータが*national accounl*の情報を補完するのに，どのように使用されたかを示している。若干，異なった方法ではあるが，租税モデルに関する*Brooking Institution*の研究は[2]，マクロデータだけでは得ることの出来ない主要な問題に対して，ミクロデータ・セットが如何に適切な解答を提供出来るかを示している。更に，所得維持計画[3]，老齢者の所得分布[4]，人口に関する地勢学的，社会学的特性のシミュレーション[5]分析をするためのミクロデータ・セットの使用は，可成りの成功を修めた。

残念なことに，どのようなミクロデータ・セットも，単一では，経済学者が分析しようとする諸問題に対して，必要な各種の情報をすべて含むことはない。色々のミクロデータ・セットが，各種の情報を含む。例えば，租税申告に関する情報を含んでいるミクロデータ・セットは，*Survey of Economic Opportunity*標本で入手出来る家計の社会学的，地勢学的情報を含んでいない。*Brooking Institution* が，これら二つの型の情報を統合して，単一のミクロデータ・セットを作り出したのは，この事実を物語る。理想としては，一つの与えられた家計，更には，その家計内の個人に対して，広範囲の種類で異ったソースで入手出来る色々の型のデータを結合したいであろう。

それで，各家計，各個人に対して，センサス記録，租税記録，社会保証記録を組み合せるのが望ましい。政府外の研究者に対して，斯様なデータのアセンブリは，機密性の問題を生ずる。何故なら，一個人についての情報量が増加するにつれて，特定ケースの確認が可能となることが大いにあり得るからである。それにも関わらず，連邦政府部内では，有意義なデータ体に対して，正確なマッチを作るように相当な努力が向けられている。

然しながら，多くの場合に，正確なマッチということは，理論的に可能ではないであろう。沢山の情報が一つの標本ベースで集められる。二つの標本が関連する時に，同一個人が両方に現らわれる確率は極めて小さい。それで，正確なマッチングは不可能である。異なった二つの標本に含まれる異なった型の情報を，一つのミクロデータ・セットに結合する他の方法が要求されるであろう。

あるデータ・セットと他のデータ・セット間で情報を移す伝統的な方法の一つは，回帰分析の使用である。標本Aの各ケースに対して，標本Bに含まれる1変数の推定値を予測する多重回帰モデルを作ることによって，一つのデータ・セットから他のものえと情報が帰属される。物論，この方法が成功するためには，二つの標本が，回帰方程式の独立変数として役立つ共通の変数を含むことが必要である。例えば，一つの標本が，賃金労働者の *Union Status* と年令，性，人種，職業，産業，所得の特性値を示していたならば，同じ年令，性，人種などを含む，もう一つ別のファイルのそれぞれの賃金労働者に，*Union Status* の情報が帰属されたであろう。物論，斯る帰属の有効性は，帰属される変数（*Union Status*）が共通の変数（特性値）によって如何にうまく説明されるかに依存する。多くの分析目的にとって，推定値が個々の観察水準に於て，正確にならねばならぬことはないだろう。推定値が，既存の変動範囲に関して，平均に於て，満足に作用することが必要であるだけである。回帰のフィットが非常に接近しているならば，現実値の代りに回帰値を代入しても，その後の分析を無効にすることはない。

回帰による帰属法は，雑雑な情報セットを移転するのには，満足すべきものとはならない。例えば，家計情報が，社会学的，地勢学的情報をより豊富に含んだ一つの標本に帰属されるならば，家計支出がすべて高度に内部相関するという一つの問題が生ずる。各々の支出に対する個別推定値は，ある特定個人に対して矛盾した家計形態を生ずるであろう。猶，家計情報を集めるこ

との主要目的の一つは，家計項目の間における内部相関関係——各家計支出が独立に帰属されたならば失なわれてしまう内部相関関係——の研究である。已に帰属された要素を各支出項目に対して考慮し，元の標本における情報を留どめるようなモデルを考えることは可能であるけれども，現実の関係が線型又は対数線型モデルによってうまく近似化出来ないならば，斯様なモデルは非常に複雑なものとなる。単純でより満足出来る処理方法は，家計情報の完全なセットを，マッチングの方法によって，一つの標本の観察値から他の標本の観察値へと移転し，両標本に於ける情報セットの完結性を留めることである。

マッチング処理の使用は，方法論的に重要な意味を持つ。回帰による帰属は，平均値を割り当てることに帰する。処が，マッチングの技術は，元のデータ・セットの値の分布を再生する。単一の帰属に対して，平均値は望ましいが，繰り返しの帰属に対する平均値の使用は，観察される分散を破壊してしまう。

マッチング技術の成功は，相似なケースが両データ・セットのなかに見出されるように，データが全く密であることに掛かっている。マッチングの目的に対して，予め，如何なる特定の関数関係も決める必要がないということは特筆すべきことで，その関係が非線型であるという陽表的認地なしに，非線型関係が線型関係と同じように自動的に有効に処理し得る。このことは，予め正確な関数形態の決定を必要とする回帰技術と著しく対蹠的である。関数形がよく知られ，データが散らばっていてマッチングが困難な場合には，回帰分析がより妥当な帰属を提供するが，相似したケースが存在する大きなデータ母体については，マッチングによる帰属が元の標本の分布特性を留め，基本的関係をより正確に反映するという利点を持っている。

註

(1) *Edward C. Budd, The Creation of a Microdata File for Estimating the Size Distribution of Income, Review of Income and Wealth, Series* 17, *No* 4, *December* 1971, *PP.* 317～334。

(2) *Benjamin Okner, Constructing a New Data Base from Existing Microdata Sets: the* 1966 *Merge File, Annals of Economic and Social Measurement, Vol.* 1. *No.* 3, *July* 1972, *PP.* 325～342。

(3) *Nelson Mcclung, John Moeller, and Eduardo Siguel, Transfer Income*

Program Evaluation, The Urban Institute, Washington, D.C. Working Paper 950－3, *September* 2, 1970。

(4) *James H. Schulz, Comparative Simulation Analysis of Social Security Systems, Annals of Economic and Social Measurement, Vol.* 1, *No.* 2, *April* 1972, *PP.* 109～128。

(5) *Harold W. Guthrie, Guy H. Orcutt, Steven Caldwell, Gerald E. Peabody, and George Sadowsky, Microanalytic Simulation of Household Behavior, Annals of Economic and Social Measurement, Vol.* 1, *No.* 2, *April* 1972, *PP.* 141～170。

マッチング問題の明細

二つのデータ・セットが統合され，お互に，それらの間の観察値がマッチされ得るならば，マッチを作るための客観的且つ妥当な規準が存在するように，形式的方法を創設すべきであろう。例えば，二つのデータ・セット，(A) 1970年の *Public Use Sample*（*PUS*）と(B), *the Social Security Longitudinal Employer-Employee Data*ファイル（*LEED*）を統合の候補として考えることにしよう。これらは，ある共通変数x_1，x_2 ……x_nを持つとする。*Public Use Sample*の中には，*LEED*ファイルから入手出来ないy_1 ……y_n変数があり，逆に，*LEED*ファイルの中には，*PUS* ファイルで得ることの出来ないz_1 ……z_n変数が存在する。下の表Ⅰは，これらの変数が何であるかを正確に示している。マッチングが有効となるためには，共通のx変数が観察値を解析的に意義あるグループに分離せねばならぬ。yやz 変数の何れとも関連のない自明のx変数は，単に，確率的なマッチングを生ずることになるだろう。

ある変数に対しては，ファイルの一つのなかに派生値が創出されねばならぬことになるかも知れない。例えば，最後に働いた年は，LEEDファイルのなかで陽表的に与えられないが，長さを示す労働歴から導出される。そこには，また，一つのデータ・セットの一つのx変数が，他のデータ・セットの対応するx変数に，正確に対応しない時には，非常にシリアスな整頓の問題が存在する。例えば，*Bureau of the Census*によって集められた賃金情報が，定義上，並びに，統計的理由の両者に対して，*Social Security Administration*に報告される賃金情報と対応しないかも知れない。一方で，*Public Use Sample* の賃金情報は，社会保証制度でカバーされるかどうか

表Ⅰ　1970年のPUSとLEEDファイルの変数

A. Public Use Sample.	*B. LEED* ファイル
***x* 変数**	
x_1 年令	x_1 年令
x_2 性	x_2 性
x_3 人種	x_3 人種
x_4 州	x_4 州
x_5 労働時間	x_5 （労働時間―導出）
x_6 最終労働年	x_6 （労働の最終年―導出）
x_7 現在の産業	x_7 現在の産業
x_8 労働の種類	x_8 労働の種類
x_9 雇傭の地位	x_9 （雇傭の地位―導出）
x_{10} 昨年の労働	x_{10} （昨年の労働―導出）
x_{11} 労働した週	x_{11} （労働した週―導出）
x_{12} 賃金	x_{12} 賃金
x_{13} 5年前の労働の種類	x_{13} （5年前の労働の種類―導出）
x_{14} 5年前の州	x_{14} （5年前の週―導出）
x_{15} 5年前の産業	x_{15} （5年前の産業―導出）
***y* 変数**	***z* 変数**※
y_1 基本的家族関係	z_1 従業員の確認者
y_2 詳細な関係	z_2 ファイルの従業員年数
y_3 サブ家族の数	z_3 雇傭された年数
y_4 グループの居所形態	z_4 z_3に対する雇傭者の数
y_5 スパニシュの名前	z_5 雇傭者の確認者
y_6 出生の地区	z_6 賃金項目の数
y_7 婚姻状態	z_7 年間の賃金
y_8 出生地	z_8 第1・4半期の賃金
y_9 最高学歴	z_9 第2・4半期の賃金
y_{10} 最終学歴	z_{10} 第3・4半期の賃金
y_{11} 出産した子供	z_{11} 第4・4半期の賃金
y_{12} 現在の職業	z_{12} 全体の推定賃金
y_{13} 5年前に軍隊	
y_{14} 5年前に大学	
y_{15} 事業所得	
y_{16} 農業所得	
y_{17} 社会保証所得	
y_{18} 福祉所得	
y_{19} 他の所得	
y_{20} 個人総所得	
y_{21} 貧困状態	
y_{22} 家族数	

	y 変 数
y_{23}	サブ家族関係
y_{24}	家族単位のメンバシップ
y_{25}	スペイン系の子孫
y_{26}	市民権
y_{27}	移入して来た年
y_{28}	結婚回数
y_{29}	初婚の年令
y_{30}	初婚の場所
y_{31}	職業訓練
y_{32}	訓練の分野
y_{33}	不具
y_{34}	5年前の職業

※ 12年間 *FICA* 税を支払った各雇傭者別に4半期毎に各個人に関して入手し得る。

と関係なく，あらゆる賃金に対して言及する。他方，*Social Security* の賃金報告は，それが与えられた場合には，正確さの水準で *PUS* の対応する情報より統計的に勝れている。時折，定義上の違いが考慮される。それで，例えば，*PUS*で示されるような，ある人の職業，又は，雇傭形態が，明きらかに社会保証体系によってカバーされないならば，*LEED* ファイルのなかに一つのマッチを見出す試みはなされないであろう。適用範囲の相違を調整した後に，二つのファイルにおける賃金分布が，未だ，著しく異っているならば，二つの情報セットを整頓するのに，更に，統計的調整が必要となるであろう。この特定ケースにおいて，その整頓は，*Public Use Sample* が*LEED*ファイルの賃金情報によりよく一致するように，*PUS* の賃金情報を調整することに関係する。

マッチング目的に対する x 変数の定義並びに整頓の問題は極めて重要であり，マッチング努力のエネルギーの大部分をそれに費すかも知れない。確かに究極的なマッチの質は，異なったデータ・セットにおける x 変数の定義の調整，並びに，整頓が，どのように徹底して行われたかに依存する。このトピックは，それ自身で一つの論文に値するが，此処での仕事ではない。この論文の残余は，已に整頓され終えた x 変数を含んでいるミクロデータ・セットを統合し，マッチングを行う戦術についての検討に集中する。

マッチングの過程は，二つのデータ・セットの観察値を一緒にするために，一つの x 変数の値を，もう一つの別のデータ・セットの値と比較することに関係する。この過程の中心問題は，一つのマッチを決めるための規準の選択ということに帰着する。標本 A における x 変数の値が，標本 B の x 変数の値に正確にマッチする場合は問題はない。斯様な場合には，x 変数に対して同一の値を持つファイル A，B の観察値が，一つの確率的基盤でマッチされる。新しく余分の情報が加わわらない限り，よりよく改善することは出来ない。二つのデータ・セットにおける x 変数の値が，幾らか異なった時に，現実の問題が生じ，x 値のどのコンビネーションが，最も満足なマッチを生ずるかを決める必要が生まれる。

概念的には，データ・セット A，B の各観察対に対して，すべての x 変数の値の間における差を表わすのに距離関数が構成され得る。

斯る方法の目的は，データ・セット A における各観察値に対して，最も小さい距離測度を持つデータ・セット B の観察値を見出すことである。斯様な距離関数を構成するために，x 変数のお互の差が何を意味するかについての解析的測度が必要になる。

原則として，x 変数は，それらの関数が y，z 変数を総合的に一緒に齎らすという意味において仲介的である。x を条件とする y，z の同時分布に関する外部情報が入手出来るならば，それはマッチングの規準の一部として導入され得る。この可能性は，此処では考慮されない。特殊な解析目的に対してマッチングが行われるならば，ある y，z 変数は他のものよりも，もっと重大となり得る。それで，各変数にそれぞれのウェイトがつけられる。例えば，二つのデータ・セットをマッチする目的が，地勢学的，経済学的変数間の相関関係を分析しようとするならば，これらの変数が重視されるであろう。然し，国民経済勘定が多くのアグリゲート型の分析に対するデータを提供するように，その目的が種々の用途に役立つように工夫したデータ・セットを作ることにあるならば，より一般的方法が必要とされる。斯様な目的に対しては，y，z 変数，それ自身が，二つの観察値が相似であるか，どうかを決めるための一般基準として使用される。

マッチを決める諸方法

距離関数を展開する一つの方法は，多変量回帰分析を使用することである。

そのなかで，従属変数は，y，z の*non-matching*変数，独立変数は x 変数で，y，z 変数の最良の説明を得るために，x 変数の各々に所属されるウエイトを決める。斯様な情報から，一つの距離関数が構成される。*Horst Adler*による論文は，*Statistics Canada*[(1)]による斯る方法の使用を説明している。

実施中の *tax model* ファイルに *Survey of Economic Opportunity* ファイルを統合する *Okner* の研究は一つの距離関数を作り出した。

それは，種々の基準に対する *consitency score*（一致の得点）を割当て，これらの一致得点に応じてマッチングが行われることを要求する。この処理の第一段階は，各ファイルの諸単位を，マッチング処理に対して非常に重要であると考えられる広いカテゴリ，すなわち〝*equivalence classes*〟と呼ばれるものにグループ化することであった。これらの *equivalence classes*〟内には，より狭い所得層の帯が定義され，これらの帯のなかでは，許容され得るマッチを定義するための一致得点が使用された。許容し得るマッチは，標本確率を基礎に行われた。

Current Population Survey ファイルと *tax model* ファイルを統合する *BEA* における *Edward Budd and Daniel Radner* による研究は，*Okner* の方法と多小異なっていた。

Budd-Radner の方法は，広い *equivalence classes* 内の二つのファイルの観察値の順位の序列に依存する。事実，その処理は，両ファイルを相当広い賃金の順位層のなかに，営業所得と資産所得別に順位をつける。一つの特定部分集合の同じ順位を持つ二つの記録に対するウエイトが同じになるように，各ファイルの記録を分割することによって現実のマッチが達成される。二つのファイルのランク・オーダを使用するこのマッチング技術は，二つのファイルにおける情報の一般序列が正確で，しかも整頓問題が一つの水準であるという仮定に立って，整頓問題を取り扱うことを留意すべきである。

1970 *Public Use Sample* を *Survey of Economic Opportunity* ファイルにマッチするのに，少し異なった方法が *Richard Rockwell* によって展開された。このマッチにおいて，データを 288 個のセルの中にクロス分類するために，非常に広い区間に分類する五つの変数が使用された。これらのセルのなかで，最後のマッチに到達するために，新らたに三つの変数を次々と加えることによって，幾つかのマッチが達成された。クロス表のセル・マッチは，三つの附加的変数の一連の順序付けに基づいているから，純粋な *sort*

と *merge* の処理によって *Rockwell* の結果が達成されたであろう。

註

(1) *Horst Adler, Creation of a Synthetic Data by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey*, 1970。

マッチングに関するクロス表技術の検討

マッチング処理は，x変数のすべてを使用するn次元のクロス表によって，実際に行われる。そこでは，幾つかのマッチが，同じセルのなかに落ちて来る観察値の間で確率的に作られる。一つのセルの正反対の境界に存する二つの観察値が，お互にマッチされることは可能であるから，この処理は距離関数を使用して得られるものとは異なった諸結果を生ずる。ところが，距離関数が使用されたならば，一つのセルの境界近くにある観察値が隣り合ったセルの境界値近くの観察値にマッチされる。クロス表による方法のもう一つの観点は，与えられた何れのクロス表に対しても，あるセルの観察密度が非常に高くなり得るので，より正確なマッチングは，距離関数を使用するか，より細かなクロス分類の何れかによって達成されるということである。更に，クロス表は，標本 A に関し，一つ以上の観察値を含むが，標本 Bに関して，一つも観察値を含まないセルを生ずるかも知れない。また，その逆のケースもあり得る。

この難問は，先づ，非常に細かいクロス分類の碁盤目のようなものを使用し，マッチし得るケースはマッチを行い，遂次的にすべての観察値の完全なマッチングが達成されるまで，セルの大きさを増加することによって解決される。この方法は，他の方法にあると同様な基本問題につきあたる。クロス表を展開するのに使用されるx変数の区間を決めるのに，幾つかの客観的基準が必要とされる。猶，x変数の区間は，x変数のy，z変数に対する関係に依存する許りでなく，変数空間についての観察値の密度に依存する。大きな標本に対する，より細かなクロス階級化は，妥当且つ可能であり，過剰費用なしに，より高い品質のマッチが達成される。

最良のマッチを決めるために，二つの観察値の比較を要するマッチング技術が用いられるならば，非常に大きいデータ・セットを処理する費用は莫大なものとなるので，そのことも留意すべきである。従って，大きなサンプル

については，マッチングに関するクロス表技術の採用は，魅力のあるものである。

マッチングに対する選別—統合の戦術

クロス表セルの階層別枝分れを生ずるファイルの単一選別によって，逐次的クロス表処理と同じ結果を達成することが可能である。*sorting* というのは，実際にクロス表を作り出す伝統的な方法である。セルの階層別枝分れを作るために，各々が階層別枝分れの一水準を示す一連の選別標識セットが，各観察値につけられる。選別標識の最初（最も左）のセットは，使用予定の最も広いセルを決める。選別標識の最初のセットを作るために，ある基準で，各x変数が広い区間に分割される。すべての変数に対するこれらの区間仕様書は，クロス表に対するセルの境界を定義する一組の選別標識を構成する。最初の広いセルのなかに，より細かい分類を導入するため，第二番目の選別標識集合が作られる。x変数のそれぞれを，より狭い区間のなかに分割することで，これが達成される。必要があれば，x変数の生の値が到達されるまで，この過程が繰り返えされる。換言すれば，その処理は，可成り広いセルを最初に取り，それをより小さなセルに分け，更に，小さなセルへと分割して行く。そして，極度に小さいセル・サイズが到達されるまで続く。これらの枝分れ選別標識集合に応じて選別を行い，その後，二つのデータ・セットを統合すれば，一つの統合されたデータ・セットを生ずるであろう。そのなかにおいて，少くとも，第Ⅰの階層水準（最も広いセル）のなかに，一つのA観察値と一つのB観察値がある限り，お互に最も近い観察値が，ある階層水準の共通セルのなかに，定義上，落ちて来るであろう。斯くて，あるセル・サイズに於て，一つのマッチが保証されるであろう。そのマッチが起るセルの大さは，二つの標本の観察密度に依存する。非常に大きい標本のなかでは，x変数の本当に細かい分類が，最も詳細な階層別水準で使用される。その理由は，その規定水準では，相当に沢山のマッチが起ると期待されるからである。マッチが起る可能性が少ない小標本では，より広い分類が使用されねばならぬ。このことは，より高い密度の標本は，より沢山のマッチを生ずるということでもある。

マッチング変数の諸区間における差の統計的測定

セルの堺界として使用されるx変の諸区間の決定は，マッチングの中心問題である。理想的には，我々は，与えられたx変数の規定された区間内で，yとz変数の分布が不変という保証を持ち度いであろう。換言すれば，規定された区間が，あるy，z変数の有意に異なった分布を生ずるならば，唯，xの一つの区間と別の区間に区別することが必要となる。

これを検定するために，x変数の二つの異なった指定区間のなかに落ちる観察値は，異なった標本として取り扱われる。これらの標本が異なった母集団から来る確率が低いならば，マッチング目的のこれらの区間に，区別をするための統計的基礎が，そこにないことを意味する。逆に，特定区間の標本が，異なった母集団から生ずる確率が高いならば，マッチングの規準を展開するのに，この情報を利用することが重要である。

観察される差が有意であるかどうかを決定するために，x変数のそれぞれ異なる区間に対して，y，z分布へのカイ二乗検定が適用される。観察数が小さい場合には，相違が実際に存在する時でも，x変数の区間の間における差を見出すことは可能でないかも知れない。他方，観察数が非常に大きい場合は，異なった区間に対するyとz分布の観察値の差が比較的小さくても，高度に有意なカイ二乗値を生ずるであろう。観察される差の有意性を検定するには，できるだけ大きな標本が使用されるべきである。このことは，ある場合には，x変数のおのおの値に対して充分な観察値が入手出来るように階層別標本が求められることを意味する。

xの異なった区間に対するy，z分布の有意差が見出されるならば，x変数の異なった区間に基づくセルの階層別枝分れに対する基礎を提供するために，これらの差の相対的重要性に関する一層の評価をなす必要がある。y，z変数に対する百分率の分布が，xのある二つの特定区間に対して，如何に密接に相関しているかどうかを測定することによって行われる。二つの百分率の分布が同じであるならば，それらは45%の回帰直線上にあり，相関係数は1.00となるであろう。二つの百分率の分布が異なるならば，相関係数が，この差の大きさを示すであろう。指定された幾つかのx区間に対して相関が高い場合には，マッチング目的のこれらの区間を単一区間に崩壊することは，低い相関が存在する場合よりも，y，z変数により小さいゆがみを生ずる。求められていることは，xの結合された区間が，何れか一つだけのよき代理であ

るかどうかである。一つのx区間が，他のx区間と同じy，z変数の分布を生ずることを相関係数が示すならば，一つの結合された区間は満足な代理となるだろう。この統計的測度は，相関係数のそれぞれの水準によって，選別標識の階層別水準を指定することを可能にする。

このようにして，二つの基準が導入されて来た。カイ二乗の基準は，一つのx変数の二つの区間に伴ったy，z変数の分布が，標本の大さと分布における観察された差の両者に基づいて，お互，有意に相違があるかどうかを決めるように意図されている。有意差が見出されない場合には，マッチに対し侵害することなしに，区間が結合される。有意差が見出される場合には，これらの差の重要性が評価されねばならぬ。

相関測度は，y，z変数の分布の全分散のどれほどが説明し得るかによって，分布がどの程度，異っているかを求める。説明されない分散が非常に小さい場合(相関が高い場合)，問題のy，z変数の分布を有意に変えることなしに，x変数の二つの区間が結合される。カイ二乗と相関の両測度は，妥当且つ意味ある区別を提供するために必要である。一方，非常に大きい標本については，カイ二乗値が大きく，且つ，相関係数も大きくなり得る。他方，小さな標本については，低いカイ二乗は低い相関係数を伴う。最初の場合には，分布の間に統計的な有意な差があるが，その差はとるに足らない。それで，区間を結合することは，マッチングの処理に如何なる侵害も与えない。第二の場合には，分布の間に大きな差があり，統計的に信頼出来ない。それで，マッチングの基準として使用されるべきでない。相対的に高いカイ二乗が，相対的に低い相関に結びついている時にのみ，二つの区間の区別が維持されることが望ましい。

カイ二乗並びに相関測度が如何に適用されるかの具体例は，分析を明らかにするのに役立つであろう。表2は，x変数の「雇傭の形態」の二つの区間が，y変数の「家族の規模」にどのように関連しているかを示している。設定された問題は，「仕事に従事」の区間と「仕事に従事していない」の区間の区別が，有意に異なった「家族の規模」の分布を生ずるかどうかである。カイ二乗検定は，分布の観察された差が有意であるということに，非常に低い確率を与える。それで，y変数の「家族の規模」に対する「雇雇の形態」の二つの区間を，マッチングの目的のために一つに結合しないという統計的理由がないことが決定される。

表2　労働者の雇傭状態別の家族規模の分布

y 変数	x 変数：雇傭状態			
	仕事に従事		仕事に従事していない	
家族規模(人数)	観察数	パーセント	観察数	パーセント
1	973	11.9	16	13.2
2	1,602	19.5	26	21.5
3	2,487	30.0	31	25.6
4	1,740	21.2	29	24.0
5	846	10.3	13	10.7
6	329	4.0	5	4.1
7	135	1.6	1	0.8
8	52	0.6		
9	19	0.2		
10，それ以上	12	0 0.1		
総計	8,195	100.0	121	100.0

分布間の比較
カイ二乗確率0.0086（観察数の分布に基づく）
相関係数0.9852（百分率分布に基づく）

表3において，x変数は「勤労者の分類」で，y変数は「事業所得」である。カイ二乗は1.000で，「従業員」と「自営業者」に対する「事業所得」の分布間の差が，統計的に有意であることを示している。低い相関係数は，その差が有意であることを示している。それで「事業所得」がy変数の一つであるならば，マッチングの規準として「従業員」と「自営業者」の区別を維持することが重要である。

表4では，x変数が「勤労者の分類」で，y変数は「家族の規模」である。0.9536のカイ二乗は「家族の規模」に関する二つの分布間の観察される差が，統計的に有意であるという強い確率を示している。然しながら，相関係数は高く，全体の分散によって，二つの分布間の差が小さいことを示している。それで，マッチング目的のために「公務員」と「民間労働者」を別々の区間に保つことは，「家族規模」の属性を，顕著には改善しないであろう。

表3　従業員と自営業者に対する事業所の分布

y　変　数	x　変　数：勤　労　者　の　分　類			
	従　業　員		自　営　業　者	
勤　労　所　得	観　察　数	パーセント	観　察　数	パーセント
−9,900～−100	19	4.4	7	0.7
0～200	74	17.3	32	3.2
210～600	80	18.7	52	5.2
601～1,000	37	8.6	52	5.2
1,001～1,300	10	2.3	40	4.0
1,301～2,000	45	10.5	116	11.5
2,001～2,500	23	5.4	64	6.4
2,501～3,200	26	6.1	87	8.6
3,201～4,100	23	5.4	129	12.8
4,101～5,000	25	5.8	108	10.7
5,001～7,600	40	9.3	152	15.1
7,601～15,500	23	5.4	146	14.5
15,501～24,500	2	0.5	16	1.6
25,501及びそれ以上	1	0.2	6	0.6
総　計	428	100.0	1,007	100.0

分布間の比較
カイ二乗確率 1.000（観察数の分布に基づく）
相関係数 0.1479（百分率の分布に基づく）

表4　民間並びに政府従業員に対する家族規模の分布

y　変　数	x　変　数：労　働　者　の　分　類			
	民間会社の従業員		公　務　員	
家族規模(人数)	観　察　数	パーセント	観　察　数	パーセント
1	869	12.4	186	13.6
2	1,394	19.9	279	20.4
3	2,075	29.6	439	32.1
4	1,445	20.6	288	21.1
5	728	10.4	115	8.4
6	289	4.1	38	2.8
7	124	1.8	13	1.0
8	50	0.7	6	0.4
9, それ以上	17	0.2	3	0.2
	8,537	100.0	1,707	100.0

分布間の比較
カイ二乗確率 0.9536（観察値の分布に基づく）
相関係数 0.9966（百分率の分布に基づく）

一つのマッチング変数を幾つかの区間に分割すること

x変数を分割する規準として，カイ二乗並びに相関測度を適用するには，データを処理し，諸結果を分かり易い形で報告するために，コンピュータ・プログラムに具体化出来るようにアルゴリズムの展開を必要とする。x変数が，(1)「よく順序付けがなされ得る」。又は，(2)「順序付けが行われ得ない」又は「部分的に順序付けがなされ得る」かどうかによって，異なったアルゴリズムが必要である。賃金所得は，「よく順序付けがなされ得る」変数の一例である。労働者の人種，職種は順序付けがなされ得ない，そして，産業，地域，州の如き変数は，階層別集合に部分的に順序付けがなし得る。

相対的に生の値が少なく，各生の値に対して沢山の観察を持ったよく順序付けられた変数に対して，その方法はストレートに運ぶ。

x変数の生の値の隣り合った区間に対するyとz変数の分布が比較され，カイ二乗並びに相関測度が計算される。如何なる有意差も見出されないならば，つまり，差の大さが与えられた水準以下にあるならば，生の値が結合される。それから，新しく結合された区間とそれに隣り合う他の区間の間に，一つの比較が行われる。斯様に，x変数は，カイ二乗並びに相関係数の指定された水準に基づいて，区間の一集合に分割される。

ある場合は，よく順序付けられたx変数が不都合なほど沢山の生の値を持つかも知れない。斯くして，*Public Use Sample* における変数「賃金」は，100ドル刻みの250区間からなり，*LEED*ファイルは一ドル単位で賃金を報告する。各生の値を比較する代りに，一つの異なった方法が使用される。その時，比較されるx変数は，恣意に，相対的に少数の区間に分割され，これらが比較される。有意な差が見出される場合，これらの区間の各々は，二つの区間に分割され，これらが比較される。この過程は，区間の間に如何なる有意差も見出されなくなるか，或は，生の値に到達するまで続けられる。x変数を広い区間に分割するのに色々な技術が使用され得たであろうが，採用されたものは、x変数に関するサンプルを順序府け，それを主要な八つの部分に分割することに基づいている。それらの各部分は同一観察値を持つ。この方法は，それから生ずる区間が，信頼出来る比較を提供するに足る充分な観

察数を含んでおり，然も，最適利用が標本の規模から構成されるのを保証する。

殆んど生の値を持たないよく順序付けられた x 変数と多くの生の値を持ったよく順序付けられた x 変数を分析する方法の唯一の相違は，前者の場合に，より小さい区間がより大きな区間に集計され，後者の場合は，より大きな区間がより小さな区間に分解されるということである。

順序付けられない x 変数に対しては，隣り合う区間という概念は意味を持たない。それで，何れが結合され得るかを決めるために，区間の関係について出来る限り遂次比較をするのが必要であろう。部分的な順序付けがなされている，又は，階層別 x 変数に対しては，先づ，最も広いグループ水準（主要産業又は地域）で比較がなされる。これらのグループに対して，可能な限り遂次比較がなされるべきである。個別グループが識別される場合，その主要グループ内のサブ・グループに対して遂次比較が行われる。この過程は，階層別の順序付けが汲み尽されるまで続けられるべきである。区間を結合するためのカイ二乗，並びに，相関基準の指定が，分割における区間の数を決定するということは，明きらかにすべきである。区間の間の小さな差でも，統計的に有意で，重大であると考えられるならば，そこには，より多くの区間が存在するであろう。大きな差も許容されるものならば，その時は，x 変数の分割される区間の数は減少する。斯くして，カイ二乗並びに相関係数の各水準を，基準として使用することによって，分割の各水準が作り出され，区間の階層別集合を生ずる。

一つの x 変数は，一般に，一つ以上の y，z 変数によって分析される。それ故，一般化された分割が，個々の y，z 変数から生ずる個別的分割から如何に導出されるかを考慮することが必要である。二つの異なった規則が適用され得る。第一に，個々の分割に代表される最も細かな区間を反映するように，一般化分割を構成することが可能である。第二に，すべての y，z 変数に対する百分率の分布をプールし，これらのプールされた分布を基礎に相関係数を計算することが可能となる。

賃金区間の三つの枝分れ集合に分割された x 変数（賃金）の一例が表5に示される。生の賃金値は，1－99バルから25,000ドル乃至それ以上に渡って100ドル刻みの250の賃金層からなった。区間分析をするのに，27個の y 変数が使用された。最も細かな階層別水準（水準III）において，すべての y 分布

表5　賃金分類の区間分割

賃金分類（ドル）	階層別水準					
	水準 I		水準 II		水準 III	
	区間数	観察のパーセント	区間数	観察のパーセント	区間数	観察のパーセント
1～99	1	31.7	1	31.7	1	3.3
100～399					2	9.8
400～599					3	2.1
600～799					4	3.4
800～1,799					5	13.1
1,800～2,299	2	68.3	2	39.6	6	6.9
2,300～2,799					7	6.0
2,800～3,499					8	9.1
3,500～3,899					9	5.1
3,900～4,299					10	6.0
4,300～4,499					11	1.7
4,500～4,899					12	4.8
4,900～5,299			3	14.7	13	6.3
5,300～5,499					14	1.4
5,500～6,299					15	7.0
6,300～7,499			4	5.8	16	5.8
7,500～8,499			5	3.0	17	3.0
8,500～9,099			6	2.0	18	1.3
9,100～9,799				1.5	19	0.7
9,800～11,799			7	1.7	20	1.5
11,800～25,000※			8		21	1.7

※ 最高所得クラスは25,000ドル以上

区間を結合するための仕様書，

カイ二乗が0.00と0.94の値域にあるならば，区間は相関係数と関係なしに結合される。

カイ二乗が0.95と1.00の値域にあるならば，相関係数がそれぞれの階層水準に対して下に示される水準以上にあれば，区間は結合される。

階層別水準	相関係数
1	0.70
2	0.90
3	1.00

に対する区間の間の差のカイ二乗測度が0.95以下である場合の賃金クラスのみが結合された。この規準は，規模が100ドルから13,200ドルの領域にあっ

て，0.7から13.1パーセントまでの観察値を含んだ21区間を生じた。21番目の区間（11,800～25,000乃至それ以上）に対する広い賃金クラスは，大部分が，この範囲の相対的に小さな観察数に依るものであることが指摘されねばならぬ。これらの *run* がなされた標本は，約20,000の観察を含み，約300の観察が21番目の区間にあったことを意味する。標本の規模における増加，並びに，層別標本の使用は，恐らく，21番目の区間が幾つかの区間に分割される結果を生じたであろう。マッチング処理によって，斯様により細かい区間にすることは，マッチされるデータの1.7パーセントのみに対して，マッチングを改善するが，最も高い賃金クラスの分析が重要である場合の研究に対しては，これらの賃金クラスにおけるマッチングを改善する特殊な注意が傾到さるべきである。水準IIに対して，水準IIIに使用された区間を結合するための基準は，カイ二乗が0.95以上で，相関係数が0.9を超えた場合の区間を，加法的に結合するように緩和された。これは，最小賃金クラスの規模が1,000ドルで1.5パーセントの最小カバレージで持って，区間の数を8に減少した。IIIの階層別水準で規定された21区間のうち4つは，そのまま，IIの階層別水準に持ち込まれた。最後に，相関係数基準を0.70に緩和することによって，IIの階層別水準における8つの区間は，水準Iに対する二つの区間に崩壊する。この水準で区別される二つの所得クラスは，1～1,799ドルと1,800ドル乃至それ以上のものである。最初の区間は，32パーセントの観察を含んでいる。物論，望むだけ多くの階層別水準を生ずることは可能である。然しながら，x 変数のあるものに対しては，正確なマッチングが必要であると決定されるかも知れない。正確なマッチに対し起り得る三つの候補は年令，性，人種である。これらの変数に関する正確なマッチは，特定年令，人種，同棲が両ファイルのなかで認識されるという利点を持ち，これら同棲に対する y，z 変数の平均値，分布は，マッチ処理によって影響されないであろう。

要　　約

ミクロ経済データの発展と共に，二つのデータファイルを比較するとか，幾つかのデータファイルを一つに統合するとか言う必要は，最初，統計データを取り扱う実務家の側に起って来たと思う。他方，ミクロ経済分析の計量化への指向は，ミクロ経済データの生産とその処理を益々必要とするようになった。そこで，データファイルの比較，統合を行わんとする際に，如何に

したら元のデータファイルの性格を壊すことなく，統合，分割が出来るかということは，非常に重要な事柄となった。此処にマッチングの問題が出現するゆえんがある。

マッチングについての研究が始まったのは，本文で述べたように，今から拾数年前である。ラグルス夫妻は，これまでのマッチングに関する研究を，距離関数による方法とクロス分類によるものとの二つに大別して，それらに検討を加えながら，新しいマッチングの理論を打立てた。

ラグルス夫妻の理論を振り返えって，一言で表現すれば，統計処理に使われる選別—統合の方法をマッチングに利用したものである。最適マッチを得るためには，データファイルの区間を決定せねばならぬ。そのために，ファイルの区間分割，統合が行われるが，その規準として，カイ二乗検定と相関係数の両測度を使用した。

然し，これまでの研究が，単なる統計処理の域を脱しなかったのに対して，ラグルス夫妻のは，経済理論のアグリゲーション，それに伴うデータのアグリゲーションという視点で捉え，経済学の理論と実証を一体化せんとする試みの一環であった。