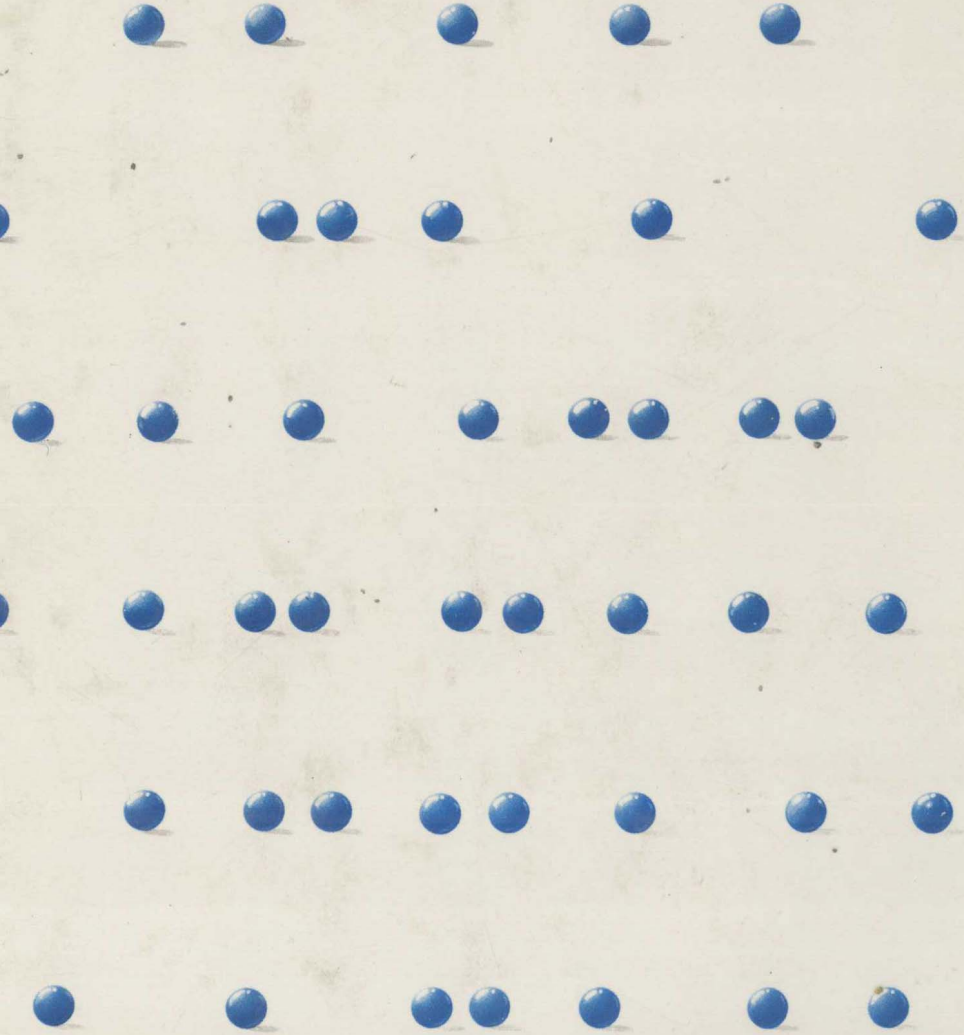
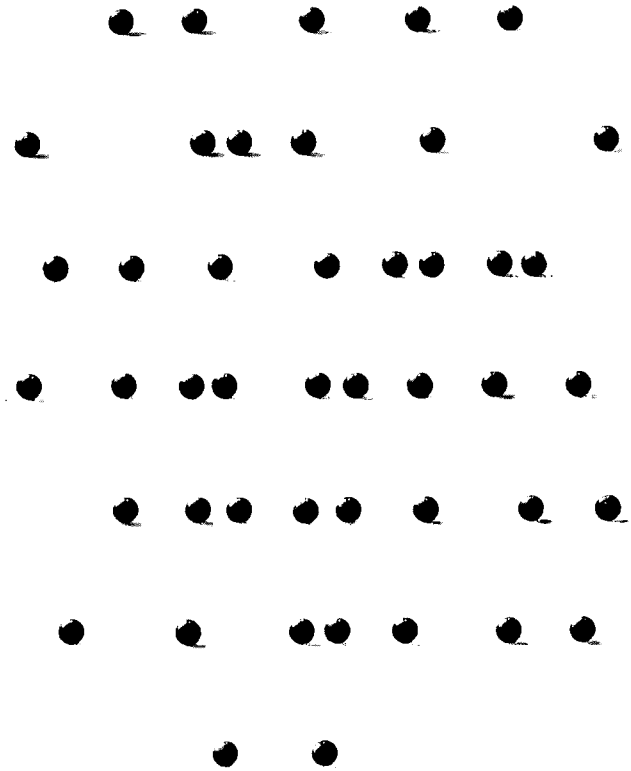
統計学入門

香川大学教授 木村 等
香川大学教授 大薮和雄 共著
香川大学助教授 石川 浩



香川大学教授 木村　等
香川大学教授 大薮和雄　共著
香川大学助教授 石川　浩



実教出版株式会社

著者略歴

木村　等（きむら　ひとし）
1924年2月15日生
1947年9月　北海道帝国大学理学部数学科卒業
統計数理研究所等を経て
1959年4月　香川大学教授経済学部，現在に至る．

大藪和雄（おおやぶ　かずお）
1937年10月1日生
1966年3月　一橋大学大学院経済学研究科修士課程修了
香川大学講師，助教授を経て
1980年10月　香川大学教授経済学部，現在に至る．

石川　浩（いしかわ　ひろし）
1941年11月29月生
1969年3月　京都大学大学院工学研究科博士課程修了
米国コロムビア大学客員研究員等を経て
1982年4月　香川大学助教授経済学部，現在に至る．
工学博士

統計学入門　　NDC 417

1982年10月20日　第1刷発行　　定価 1900円

著者　木村　等
大藪和雄
石川　浩
発行者　宇野豊蔵
印刷　中央印刷株式会社
製本　株式会社　若林製本所

発行所　実教出版株式会社
東京都千代田区五番町五番地 〒102
電話 東京(263)0111(大代表)振替東京 4-183260

3041－2228－3205

まえがき

本書は大学初年級の学生を対象とする統計学の入門書として執筆したものである．草稿の段階で本学部学生の多くに目を通してもらい，彼らの初学者としての立場からの率直な意見を取り入れて稿を改めた．社会科学系の学部学生は数学をあまり得意とはしないが，本書の程度まで懇切に書けば，数理的な面も理解できるようである．しかしながら，勿論，標本分布の導出の積分などは，統計学自体の理解のためには必ずしも必要ではないので，数学の不得意な読者はその部分を飛ばしてもらっても支障はない．要するに統計的な考え方を理解して欲しいのである．また，その意味で，各章末には基本的な演習問題を配し，巻末にその全解を入れた．読者は，この問題を，まず自ら解き，解答と参照することによってより深い理解を得ることと信ずる．

本書の構成は2部構成とした．第Ⅰ部は記述統計に関するものであり，統計データの取り扱いの基本的手法について述べた．一方，第Ⅱ部は統計的推論に関するもので，初めに確率を Frequency Theory としてとらえ，これから確率分布，期待値，分散の概念を構成し，さらに母集団と標本の概念，推定および検定の諸手法について述べた．

初学者に対する統計学の入門書としての著者らの目論見がどの程度達成されているかは，大方の厳正な御批判に待つ他はないが，多少なりとも学習の一助になりうるものとすれば望外の喜びである．

本書の出版に当たり，資料の準備段階から原稿の浄書，通読など，本学部学生諸君には多大の労をわずらわした．とくに本常功君，小出智子嬢には記してその労を多としたい．また実教出版の桜井瑞穂氏には，本書の企画段階から校了に至るまで，一貫して多大の御尽力を賜わった．同氏の御力添えがなければ本書が日の目をみることはなかったであろうと思われる．ここに記して深甚の謝意を表する．

昭和57年6月

著　者　一　同

目　　次

第Ⅰ部　記　述　統　計

第Ⅱ部 統計的推論

第Ⅰ部　記　述　統　計

現実の数量的な側面をとらえるために，統計調査を行い，その結果を統計表の形にまとめる．統計表にまとめられたデータを用いて，そのデータの性質を直観的にとらえるためにグラフを描く．またそのデータを縮約した数値を求めて分析を進める．このような目的に用いられる方法が記述統計である．第Ⅰ部では，この記述統計について基本的事項を概説する．

第1章 統計データ

1·1 はじめに

統計データは通常，次のような過程を経て作成される．

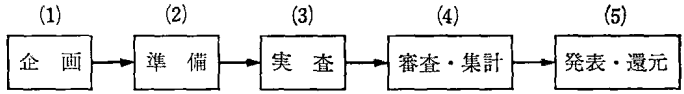


図 1·1 統計調査の作成過程

(1)の企画は，最も重要な過程である．調査目的をはっきりさせ，調査対象を定義し，具体的調査方法を選択し，調査の時と場所を決めなければならない．換言すれば，いわゆる統計調査の 5W を決めなければならない．ここに 5W とは，カウフマン(A. Kaufmann)によれば，調査対象(was)，調査方法(wie)，調査時点(wann)，調査の場所(wo)，調査の主体(wer)をいう．なお，人によってはさらに調査目的(worum)を加えて 6W とするときもある．

(2)の準備には，調査予算の獲得，調査員の選任，訓練，調査票の作成，調査用品の準備が含まれる．

(3)の実査は，調査対象と接触して情報を集める段階である．

(4)の審査・集計は，調査票の内容を検討し，内部矛盾や不備を見出し，調査対象に再度問い合わせて修正する審査の段階と，審査の終わった個票を統計表にまでまとめあげる集計の段階とを含んでいる．

(5)の発表・還元は，内容の公表と，印刷公刊，調査対象に対する得られた情報の周知還元をする段階である．

以上のような統計調査の実際を個別に知ることによって初めて，分析の対象となるデータの性格，利用の仕方がわかるのである．

統計データの性格，利用の仕方，データの体系，各種統計の比較などを，より一般的に考察する「統計データ論」というものが考えられるが，ここでは，統計データの各種の分類について述べ，データの性格についての一面にふれることに留める．

1·2 質的データと量的データ

統計データの分類の1つとして，質的データと量的データという分類がある．**質的データ**とは，男女別人口，職業別就業者数などのように，分類の標識――集団を構成する個々の単位は，多種多様な性質をもっているが，これらの性質のうち，統計的観察の対象となるものを**標識**という――が，質的である場合であり，**量的データ**とは，世帯人員別世帯数，住居の室数別世帯数，世帯の収入階級別世帯数，身長階級別学生数などのように分類の標識が量的である場合である．

量的データは，離散的なものと連続的なものとに分かれている．離散的とは1, 2, 3, …のように，とびとびの数として表される場合であり，連続的とは本来実数として表される場合である．世帯人員別世帯数，住居の室数別世帯数などは標識が離散的であるので**離散的量的データ**と呼ばれる．一方，世帯の収入階級別世帯数，身長階級別学生数などは標識が連続的であると考えられるので**連続的量的データ**と呼ばれる．数学的にいえば，世帯の収入は円の単位で測られており，本来は離散的であるが，1円は何百万円，何千万円に比べて小さいものであるから，統計的な取り扱いにおいては連続的なものとみなしている．

なお，これに関連して，**計数統計**と**計量統計**の区分がある．計数統計は質的データと離散的量的データを合わせたものをいい，計量統計は，連続的量的データに対応したものである．

質的データと量的データの区分については，上述の通りであるが，実際には，質的データを量的データで規定する場合もある．例えば，大企業，中小企業の区分は本来質的なものであるが，中小企業基本法によれば，鉱工業においては資本金1億円以下の会社，あるいは従業員300人以下の企業としている．この場合は質的な中小企業という概念を資本金あるいは従業員という量的なデータで規定している．

質的データの分類については，次の各種の統計用分類が用いられている．

(1) 日本標準職業分類
(2) 日本標準産業分類
(3) 日本標準商品分類

(4)　輸出入統計品目表

(5)　日本標準建築物用途分類

(6)　疾病，傷害および死因の統計分類

このうち産業分類を例にとって，分類の問題を考えてみよう．産業分類では，分類の単位として事業所をとっている．事業所(establishment)とは，「物の生産またはサービスの提供が業として行われている個々の物理的な場所」であり，一般には工場，製練所，鉱山，商店，農家，病院，事務所などと呼ばれる，「一区画を占めて，経済活動を行っている場所」のことである．事業所の産業は，そこで行われている主要な業務によって決定することが原則であるが，事業所によってはいくつかの異なる経済活動を行っている場合もあり，このような場合には，原則として1カ年間の総収入額または総販売額の最も多い産業をその事業所の産業とする．

ところで，産業と商品とは密接に関連しているが，工業統計表をまとめる場合に，以下に示すように『産業編』と『品目編』の違いが生じる．仮想例として，いまM町に事業所がA, B, C, Dの4つあり，それぞれ，みそ，しょう油，清酒などを作っているものとし，ある年の出荷額が表1·1のようであったとする．表1·1から，前述の事業所の産業分類の考え方に従えば，A事業所は産業としてはみそ製造業として格付けされ，B事業所はしょう油製造業，C, D事業所は清酒製造業として格付けされる．したがって，この町の工業統計の『産業編』の数字は表1·2のようになる．その理由は，A事業所の主要な業務はみその製造であるから，A事業所の出荷額の総額はみそ製造業の出荷額に分類され，B, C, D事業所についても同様に取り扱われるからである[1]．また，表1·1から『品目編』は表1·3のように作られる．例え

表 1·1　産業と品目の違い　　(単位 万円)

事業所＼生産物	みそ	しょう油 食用アミノ酸	清酒	計
A	4 600	250	—	4 850
B	500	9 450	1 200	11 150
C	400	—	3 750	4 150
D	—	—	1 500	1 500
計	5 500	9 700	6 450	21 650

1)　統計法による秘密の保持のため，工業統計表ではある分類に1あるいは2事業所しか存在しない場合は x で表示することになっているので，上記の数字は全部 x となる．

表 1·2 工業統計表『産業編』(M町)

産業	出荷額 (単位万円)
みそ製造業	4 850
しょう油・食用アミノ酸製造業	11 150
清酒製造業	5 650

表 1·3 工業統計表『品目編』(M町)

品目	出荷額 (単位万円)
みそ	5 500
しょう油・食用アミノ酸	9 700
清酒	6 450

ばこの2つの表からみその出荷額についてみれば，みそ製造業の出荷額は現実のみその出荷額とは異なるものとなっていることがわかる．これは単なる仮想例ではあるが，実際の統計にもこのようなことがあるので統計表を見る際には十分注意しなければならない．このような問題は，現実にある事業所を1つの産業に分類することは本来不可能であるにもかかわらず，それをあえてすることから生ずるのであって，質的な分類でさえもこのような困難さを含んでいるのである．分類に関するもう1つの問題として，分類が現実の1つの側面をとらえてなされるために，他の側面を捨象してしまうということがある．少数の対象であれば，多方面から1度にとらえることもできるかも知れないが，統計調査が取り扱うような大量な対象については，人間の能力をもってしては，限られた側面からしかみることができず，その取り上げた側面についての分類を行うことによって情報を得ようとするのが統計学であって，この際他の面の情報が捨てられるのはやむをえないこととしているのである．

1·3 静態統計と動態統計

統計データの分類の1つとして，静態統計と動態統計の区別がある．対象が時間的に持続性をもつ場合，例えば，人口，資本ストック，国富などは，一時点でとらえざるをえない．このような統計を**静態統計**という．他方，ある種の対象は瞬間的な出来事として存在し，持続性をもっていない．この種のものをとらえるためには，ある一定の期間を区切って観察しなければならない．例えば，出生，死亡，結婚，離婚，転入，転出，設備投資，生産高といったものは，このような性質をもっており**動態統計**と呼ばれている．

いま，静態統計と動態統計の違いを人口と出生・死亡を例にとって示すために生命線の図，図1·2を作る．○印は出生，●印は死亡を表すものとし，それらを結ぶ線分を**生命線**という．出生数をとらえるためには，出生が瞬間

的な出来事であるから，ある期間を設けて，その期間内にある出生を数えなければならない．例えば，t 年 10 月 1 日から$(t+1)$年 9 月 30 日までに含まれる○印の数を数えれば，4 であるから，この 1 年間の出生数は 4 となる．他方，死亡は，この同じ期間に，2 人について起こっているから，この期間の死亡数は 2 である．また，人口を把握するためには，人口というものが持続性をもっていることから，例えば，t 年 10 月 1 日を示す縦の線と生命線との交点を数えて，この場合 3 が t 年 10 月 1 日現在の人口であるとする．

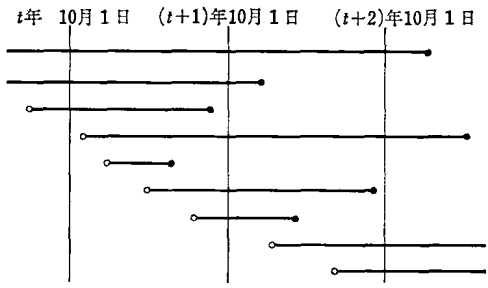


図 1·2　生命線と人口（静態統計と動態統計）

さて，次に，静態統計と動態統計との関連を考えてみよう．

P_t：ある県の t 年 10 月 1 日の人口，

P_{t-1}：ある県の$(t-1)$年 10 月 1 日の人口，

B_t：ある県の$(t-1)$年 10 月 1 日から t 年 9 月 30 日までの出生数，

D_t：ある県の$(t-1)$年 10 月 1 日から t 年 9 月 30 日までの死亡数，

I_t：ある県の$(t-1)$年 10 月 1 日から t 年 9 月 30 日までの転入者数，

O_t：ある県の$(t-1)$年 10 月 1 日から t 年 9 月 30 日までの転出者数，

とすれば，これらの間には

$$P_t = P_{t-1} + B_t - D_t + I_t - O_t \qquad (1\cdot1)$$

の関係が成り立つ．$B_t - D_t$ は**自然増加**，$I_t - O_t$ は**社会増加**と呼ばれている．ある県について，ある年の国勢調査人口を出発点として，将来の人口を推計しようとする場合を考えよう．初めに，P_{t-1} にその年の国勢調査人口を代入する．B_t は母親の年齢階級別特殊出生率（いわば，女性が子供を産む確率である）からある程度正確に推計することができるし，D_t は生命表の死亡率を用いて相当正確に推計できる．難しいのは，転入・転出の推定である．過去の動向から，それらがどのように変動するかを予測する．この部分がうまく予測できさえすれば，P_t の予測が可能となる．このようにして 1 年ごとに P_t を推定することを繰り返して将来の予測を行うことができる．逆に，

現時点から過去にさかのぼって考えると，P_{t-1} は，P_{t-2}, B_{t-1}, D_{t-1}, I_{t-1}, O_{t-1} などから決まっている．このように，時点をさかのぼって考えると，結局，現在の人口は，過去の出生，死亡，転入，転出によって決定されていることがわかる．このような意味で，静態統計は動態統計の積み上げによって成り立っているということができよう．静態統計は経済学でいうストックに，動態統計は経済学でいうフローに対応している．

1·4 官庁統計と民間統計

調査主体によるデータの区分として，国や地方公共団体およびその他の公的機関が作成する統計を**官庁統計**，それ以外の調査主体が作成する統計を**民間統計**と呼んでいる．わが国の大部分の統計は官庁統計であり，民間統計には，労働組合の統計，民間業界の統計，新聞社・放送局などの統計，専門調査機関の統計などがある．調査主体によって調査目的，回収率なども異なり，結果の数字の正確さや意味が異なる場合がありうる．

1·5 第1義統計と第2義統計

これは調査の目的と調査方法に関する区分といってよいであろう．**第1義統計**は，本来統計をとる目的でとられた統計であり，本格的な統計といってもよいであろう．これに対して，**第2義統計**は**業務統計**ともいわれるように，他の目的，例えば，行政目的のために集められた資料をまとめて統計にしたものである．

建築基準法に基づく届出書から作成される「建築着工統計」，戸籍法に基づく各種届出書を調査票に転記することにより作成される「人口動態統計」，税関に提出される輸出入申告書などから作成される「貿易統計」，警察活動の各段階で作成される犯罪原票の結果を集計した「犯罪統計」，職業安定所の「職業紹介の統計」，日本国有鉄道の「輸送統計」などは，第2義統計の例である．

最近，統計調査の数が増えたこともあって，調査される側の負担が増していること，および，プライバシーを重視する傾向が強くなったために，調査される側がより慎重に対応するようになったこと，などのために，調査が困難になりつつある．このような情況の下で，業務統計の重要性が再認識され

始めている．例えば，行政管理庁では，サービス業に関する業務統計の整備を手がけようとしているが，今後，各方面での業務統計の発掘・整備が強く望まれる．

1·6　1次統計と2次統計

これは調査方法による区分である．**1次統計**がもともと得られた統計データであるのに対して，**2次統計**は**加工統計**といわれるように，1次統計を加工して作ったデータである．2次統計は各種の経済指数を含んでいる．例えば，総理府統計局が発表している「消費者物価指数」P_{ot} は，p_{oi} を i 品目の基準時の価格，p_{ti} を i 品目の比較時の価格，$p_{oi}q_{oi}$ を基準時の i 品目への支出金額，n を採用品目の数として，次式

$$P_{ot}=\left(\sum_{i=1}^{n}\frac{p_{ti}}{p_{oi}}\cdot p_{oi}q_{oi}\Big/\sum_{i=1}^{n}p_{oi}q_{oi}\right)\times 100 \qquad (1\cdot 2)$$

で計算されているが，p_{ti}/p_{oi} の部分は「小売物価統計調査」という1次統計のデータを用いており，ウェイト $p_{oi}q_{oi}$ の部分は「家計調査」という1次統計のデータを用いている．このように，消費者物価指数は2次統計である．また経済企画庁が公表している「景気動向指数」は「鉱工業製品在庫率指数」，「卸売物価指数」，「原材料在庫指数」，「鉱工業生産指数」などの2次統計と，機械受注，輸入通関実績，全銀預貸率などの1次統計を用いて作成した2次統計である．また，国民所得勘定，産業連関表，資金循環表，国民貸借対照表，国際収支表といった社会会計上のデータは，非常に多くの1次統計を総動員して作成される2次統計の例である．

2次統計の正確さやデータの性質を知るためには，そのもとになっている1次統計の正確さや性質を知る必要がある．また，2次統計に1次統計を組み込む方法の吟味が必要であるのはもちろんである．

1·7　全数調査データと標本調査データ

これも調査方法についての区分である．**全数調査データ**は**悉皆調査データ**とも呼ばれており，調査対象をことごとく調査して得られたデータであり，**標本調査データ**は，調査対象の一部を調査して得られたデータである．

この2つのデータには一長一短がある．全数調査データは，調査対象全体

を調査して得られたものであるから，誤差を含まず正確なものであると思われている．しかしながら，調査の実施に当たっては，調査項目に関する概念規定の誤り，調査票などの設計上の誤り，調査対象名簿のもれや重複による誤差，申告者の意識的無意識的誤りなどを避けることができない．また，調査員を数多く動員しなければならず，主旨の徹底を欠くことなどに起因した調査員の誤解による誤差，調査員の記入誤りによる誤差なども生じる．その他，集計の際にも誤記・転記などの誤差，計算上の誤差などを生じる．以上述べたようないわゆる**非標本誤差**は，全数調査においても避けることができず，かえって誤差が大きくなる場合もあり得る．また，調査費用も膨大になり，集計時間も多くかかる．しかしながら全数調査は，標本調査のための基礎資料を与えるものであり，また数が多いことから精密な分析が可能であるという利点をもっている．

他方，標本調査は，調査対象の一部を調査するものであって，調査数も比較的少ないことから，調査費用も安く，調査員の訓練も十分に行き届き，集計時間も少なくてすみ，**標本誤差**は存在するが，その大きさをある程度管理できることもあって，多くの統計が標本調査によって作られている．

全数調査のデータとしては，国勢調査，事業所統計調査，農(林)業センサス，工業統計調査，商業統計調査などがある．この中で，国勢調査は，特に調査が膨大であって，多大の費用がかかることから5年ごとに行われている．その間の時期における労働力の状況を把握するために，労働力調査，就業構造基本調査といった標本調査が実施されている．

1·8 タイム・シリーズ・データ(時系列データ)とクロス・セクション・データ(横断面データ)

調査の「時」に関連する区分である．**タイム・シリーズ・データ**(time-series data)は，時間の順序に並べられたデータを意味している．これに対して，1つの時点あるいは1つの期間について，階層別あるいは地域別などに分割されたデータを**クロス・セクション・データ**(cross-section data)という．

図1·3はこの関係を示すために，可処分所得を横軸にとり，消費支出金額を縦軸にとって作ったグラフである．実線は昭和45年と昭和55年のクロス・セクション・データであり，A点は昭和45年のクロス・セクション・