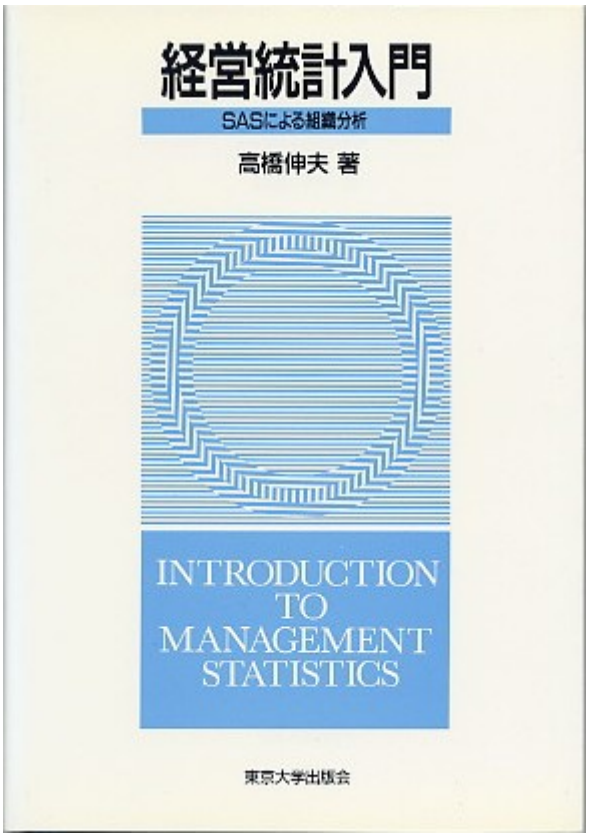
高橋伸夫 (1992)
『経営統計入門: SAS による組織分析』
東京大学出版会.　著者版(全文 PDF)

Takahashi, N. (1992). ***Introduction to management statistics: SAS user's guide for organization analysis*****. Tokyo, Japan: University of Tokyo Press (in Japanese).**

著者版(全文 HTML)　　出版社版



【この著者版(全文 PDF)は著者版(全文 HTML)から作成したものである。全文 HTML 化の際に、誤植を修正するとともに、環境依存の文字を差し替え、図は一部カラー化した。】

目次

まえがき
第 1 章　統計調査データと誤差
第 2 章　SAS 入門: 単純集計
第 3 章　データの記述と平均
第 4 章　相関と回帰
第 5 章　クロス表
第 6 章　調査の手順と実際:「組織活性化のための従業員意識調査」マニュアル
付章　CMS 入門
参考文献

まえがき

章目次
統計パッケージ SAS について
本書の特色と使い方

大学に入ると「統計学」なる科目がある。私も教えている。かなりの人が履修くらいはしてみることになるはずだ。私もそうだった。ところが、多くの人は自分には関係のない科目と早々に決めてしまう。長い人生の中で、この時期を逃すと、統計学を教室で学ぶ機会はもう巡ってこないのに。

今から 8 年ほど前、私が統計学教室のまだ助手をしていた頃のお話。ある業界団体から調査の依頼があり、その下請けで統計処理を引き受けたことがあった。調査結果もまとまり、いざ最終報告という当日、元請けの先生は風邪でダウン。結局、私一人で報告をする羽目になった。クライアントである大手企業の調査部長クラスが顔を連ねる委員会で、1 時間半ほどであったろうか、孤軍奮闘説明を終え、「何かご意見・ご質問等ございませんか？」と聞くと、某社の調査部長がおもむろに手を挙げて質問した。「先生方がよく口にする『相関』って何ですか？ 仕事柄あちこちでよく聞く言葉なんですけど、今までこういう場では偉い先生ばかりで、こんな初歩的なことは気後れして聞けなかったもので……。今日は若い先生なので、良い機会かと……。」

思わず絶句である。これまでの 1 時間半は何だったのだろう。私が意気込んで話していたのは、諸要因間の「相関」関係だったのである(「相関」については本書第 4 章を参照のこと)。現在も統計学を教えているが、これは世を忍ぶ仮の姿。その実態はいわゆる「経営学者」なので、大企業の管理職相手に話をする機会は多い。このときもそうだったが、10 人くらいの企業人を相手に、統計学の話を相関なら相関という概念だけに絞って、反応を確かめながら 1 時間も懇切丁寧に説明してあげると、ほとんどの人はうんうんと嬉しそうにうなずきながら満足して帰っていく。それはそれで教師冥利に尽きる話なのだが……。

《教訓その 1: 初歩的なことになればなるほど、偉くなってからでは聞きにくくなるものである。改まった場ならば、なおのこと。普通の神経の持主は、できるだけお早めに。》

かくいう私も、恥ずかしい思い出がある。12 年前、まだ大学院に入って間もないある日のこと、指導教官に呼ばれた。「予算が残っているので、企業対象に経営についてのアンケート調査をやろうと思う。君にまかせるから好きにやってみなさい。」私は喜び勇んで調査を行い、とりあえず単純集計だけは終えた。さて、後は何をするんだろう?

私は統計学の先生の研究室に相談に駆け込んだ。「そうですか。それでは基本的なところで、こんな本でも読んでみたらどうですか。」と 3 冊ほど見覚えのある統計学のテキストがテーブルの上に並べられた。いずれも読んだことのあるテキストだ。考えてみると、授業だ試験だとこれまでに読んだ統計学のテキストは五指に余るではないか。しかも、当時、大学院の演習で、やたら難解な数理統計学の英文の専門書まで読まされていた。にもかかわらず……。

《教訓その 2: 統計学は、実際に自分で使えなければ、勉強しても何にもならない。》

私は気を取り直して、当時、既にコンピュータ・ソフトや調査の統計処理で立派に生活費を稼ぎ出していた友人数人に頼み込んで、コンピュータの統計パッケージの使い方の手ほどきを受けたり、統計パッケージのマニュアルを手当り次第に読みあさり、そこから遡って、統計学の専門書やテキストを読み直したりし始めた。今にして思えば、自分で苦労して集めたデータを目の前にして、ようやく実用的なツールとしての「統計学」を使う必要に目覚めたのであったが、しかしこうした努力のかいもなく、結局、その調査データの運命は……。

《教訓その 3: 良質なデータは生でもおいしくいただけるが、質の悪いデータは、どんなに高度な統計手法を駆使して料理しても、とてもいただけるものではない。》

統計学のテキストを読むと、データは既に与えられているか、あるいはタダでころがっているかのような印象を受ける。しかし、どんな分野であれ、実際の研究では、良質のデータを集めることに労力のかなりの部分が費やされる。素材を前にして何が良質なのかを見極めるためには、データ収集の段階で既に、統計学をその観点から理解しておかなくてはならなかったのである。

いずれも一昔前の話である。その間に、統計学を巡る状況は一変した。コンピュータと統計パッケージの普及。パーソナル・コンピュータの驚異的高性能化。ワープロ・ソフトの進歩にともなうファイル編集能力の向上。いまや、統計学を使った統計処理は、ほとんど労力を要しない個人作業になっている。私が毎年趣味的に行っている 100 変数×1,000 人程度の従業員意識調査では、かなり分厚い集計表でも一人で週末の 2〜3 日も使えば完成する。一昔前には１〜２カ月はかかっていたのに。統計学はもはや一部の研究者の特権的独占物ではない。多少なりとも知的な仕事をする人にとっては、自分で日常的に使えるツールになったのである。まるでワープロのように。

そんな統計学の新時代が幕を開けようとしていた頃に、一足先に経験させていただいた、貴重ではあったが、今や笑い話のタネとなってしまったエピソードを御披露したわけだが、本書はそんな教訓を生かして、通常の統計学テキストとはまったく異なるコンセプトに基づいて書かれている。

本書は『経営統計入門』と題しているが、正確には「経営学者が自分で使っている統計学と統計パッケージについて書いた入門書」というべきであろう。いわゆる教養書的な入門書ではない。統計学や調査手法、さらにコンピュータや統計パッケージの使い方まで、

ツールとしてパッケージ化してしまおうという発想から生まれてきた統計学ユーザーのための入門書である。本書では最初から統計学教育と情報処理教育は融合してしまっている。取り上げた実例、素材も、私自身が手掛けている経営学分野の調査とそのデータから選ばれたものである。

少なくとも私にとっては、10 年前の統計調査と現在の統計調査とでは、もっている意味も深みも全く異なる別のものである。従来の、そして現在でも大多数の人が信じている統計調査では、大変な時間と労力を注ぎ込んで、統計学やコンピュータの専門家・経験者も交えたチームで調査、集計を行い、集計結果を主体とした報告書がまとめられる。ほとんどの調査はもちろん一回限りで、しかも、標本誤差の評価は多少行っていても、調査のプロセス自体から発生する非標本誤差の管理はほとんど意識されていない。統計分析とは名ばかりで、ほとんど誤差を読んでいたようなものである。

ところが、今日においては、もはや統計的な集計処理は、調査プロセス全体の折り返し「点」にしか過ぎない。集計処理自体は、一人で、ごく短時間のうちに行うことが出来る。集計はゴールなのではなく、きちんとした事例研究を行うための探り針としての役割を果たすことになる。しかも、コンピュータ・テクノロジーの進歩は、全数調査を可能にし、遅かれ早かれ経営学分野で「標本調査」「標本誤差」といった言葉を死語にするだろう。調査では、非標本誤差の管理こそきちんと行われるべきなのである。

統計調査とは事実に肉薄するための手法にほかならない。報告書はもはや集計表で無味乾燥に埋められるようなことはなく、統計調査の前後、特に後に行われたヒアリング調査や事例研究によって得られた知見に溢れている。しかも統計調査は 1 回限りのものではなく、毎年繰り返し繰り返し行われるべき性格のものである。仮説の検定は必ず追試が行われるべきであるし、毎年得られる新鮮な事実発見は、次々と新しい理論や仮説を連鎖的に生み出していく。そこには、事実と論理の絡み合いの中で、調査・分析をする側のオリジナリティーが常に求められているのである。信頼性の高い事例研究をしようとするならば、いまや統計調査は強力な手法なのである。

このことは同時に、組織開発、日本的にいえば、組織活性化の新しい地平を切り開くものである。組織の再開発のためには、あるいは真の問題点を的確に把握・指摘するためには、社内的にも統計数字の裏付けが必要になる。社内での水掛論を排するにも、やはり統計調査の明らかにする事実が必要なのである。その意味からも、本書で調査の実例として取り上げた「組織活性化のための従業員意識調査」のノウハウは「統計的組織活性化」の可能性を示すという意味で、企業の担当者にとって有益なものとなろう。これは、大学の研究者が経営組織の統計調査を行う際にも、必ずやヒントになるはずである。

本書は、大学の学部及び大学院における講義ノートをもとにして作成されている。この講義ノートは、もともとは 1988 年度に学習院大学大学院経営学研究科において非常勤講師として「統計調査論」を担当したことがきっかけとなって書き始められたものである。同時に、当時所属していた東北大学経済学部経営学科で、「経営学」ゼミナールの学部学生対象の卒業論文指導にも 3 年ほど手を加えながら使った。その後、1991 年度から、私の所属が東京大学教養学部社会科学科に変わって、「統計学」「社会調査法」を担当することになったので、この講義ノートの改訂がさらに続けられたのである。この間、(財)日本生産性本部の経営アカデミーで、企業の実務担当者相手に講義する際にも、この講義ノートの一部が使用されてきた。

本書は、その成り立ちからして、実に多くの方々の御指導、御助力に支えられている。まず私の所属する東京大学教養学部社会科学科統計学教室の代々の先生方からは様々な形でお世話になった。特に、大学院時代から御指導をいただき、本書の執筆を勧めてくださった松原望先生と、助手時代から御指導をいただいている林周二先生(現在、明治学院大学

教授)には、この場をお借りして心から御礼申し上げたい。両先生との出会いがなければ、私がこのようなテキストを執筆することはありえなかったであろう。

また本書のもとになった講義ノートを作成するきっかけを作っていただいた、当時の学習院大学教授、現在は筑波大学教授の河合忠彦先生。本書の中で調査の実例として取り上げている「組織活性化のための従業員意識調査」の実施に、1986 年以来、毎年親身になって協力していただいた(財)日本生産性本部の新井一夫氏。SAS に関する最新資料を快く提供していただいた㈱SAS インスティテュートジャパンの竹内清恵さん。そして、忘れてはならないのは、東北大学在職中に、私が SAS ユーザーとして独り立ちするのを多方面からサポートしてくれた東北大学情報処理教育センターのスタッフの方々。本書の執筆を終始暖かく励ましていただいた東京大学出版会の小池美樹彦氏。これらの方々には、この場で感謝の意を表させていただきたい。

1992 年 9 月

高橋伸夫

統計パッケージ SAS について

本書では、統計パッケージとして SAS(サスと読む)を取り上げている。SAS はいまや代表的な統計パッケージであり、多くの大学、企業の大型汎用計算機(これをメインフレーム(mainframe)と呼ぶ)に既に導入されている。またパーソナル・コンピュータ用はもともと廉価で、ワープロ・ソフト並の料金設定（ただしレンタル）である上に、サイト契約を結ぶことで、1 台当りの単位価格を大幅に引き下げることもできる。したがって、本書の想定する読者にとっては、SAS を使用する機会は十分にあると判断した。

SAS とは、もともと Statistical Analysis System (統計分析システム)の略称だったが、現在では、データ管理機能等の重要性が増し、SAS が正式な名称となっている。その原形は、1966 年に米国ノース・カロライナ(North Carolina)州立大学で開発が開始され、1976 年には、SAS Institute Inc.が設立され、以後、SAS の維持、開発、販売、教育等を行うようになった。簡単に言えば、SAS とはデータ・セットと呼ばれるファイルに対して、統計処理・演算処理をはじめとする各種の加工処理を行うパッケージ・プログラムである。各種の統計処理用にプログラムをそれぞれ FORTRAN や BASIC などの言語を使って自分で作成、開発、維持することは、一般には非常に大変な作業である。その点 SAS は、既に完成されている統計用プログラムの集合体、システムであり、利用者が必要なプログラムを指定すると、それらを機能的に結び付けて、統計処理をしてくれるわけである。

SAS はもともと IBM のメインフレーム上で動く統計分析用ソフトウェアとして誕生し、その後もメインフレーム用に進歩を遂げてきたが、最新版である第 6 版(リリース 6)では、メインフレーム用よりもパーソナル・コンピュータ用の方が早く出されている。本書では、この最新版、SAS 第 6 版の使い方とプログラミングについて、パーソナル・コンピュータ版 SAS、いわゆる PC 版 SAS を中心に、IBM メインフレームの代表的オペレーティング・システム(operating system 略して OS) CMS に対応する CMS 版 SAS についても、相違点などにも言及しながら併せて説明していく。PC 版 SAS と CMS 版 SAS の両方を両刀遣い的に取り上げるのは本書が初めてだと思う。これは、最新版である SAS 第 6 版では、メインフレーム用もパーソナル・コンピュータ用もともに C 言語で書かれ、使用方法は基本的に同じであるということに加えて、今後私を含めて多くの人々が、データ・サイズやコスト、利用可能性など様々な要因を勘案して、SAS のプラットフォームとしてのパーソナル・コンピュータとメインフレームを使い分けていくことになると考えたからである。ま

た現在でも一部メインフレームで使用されている CMS 版 SAS リリース 5.18 については、基本的使用方法を概説しておいた。

ちなみに、私はもともと IBM のメインフレームでの CMS 版 SAS ユーザーであるが、本書を書くに当って、もっとも安上がりなシステム構成：

1. ノート型パーソナル・コンピュータ NEC 98NOTE SX (PC-9801NS-20) (32 ビット CPU(386SX)、1MB の RAM ドライブ内臓(増設メモリとして使用)、固定ディスク 20MB 内臓)
2. MS-DOS リリース 5.00
3. SAS リリース 6.04 (プロダクトとしては、BASE SAS と SAS/STAT)

という構成で、実際に PC 版 SAS を比較稼働させながら執筆している。本書の SAS プログラムを実行させる限りでは、実行に数分を要することもあるものの(メインフレームに慣れた者にはややイライラする時間ではあるが)、パーソナル・コンピュータでも十分に稼働することが確認されている。

ただし、MS-DOS リリース 5.00 の EMS や UMB を使っても、まだメモリが十分に確保できない。そこで、日本語フロント・エンド・プロセッサー(FEP)や DOSSHELL、MOUSE 等、メモリに常駐してその分メモリを食ってしまうものについては、AUTOEXEC.BAT や CONFIG.SYS のファイルからあらかじめ削除して使用している。MS-DOS で SAS を動かすときは、このメモリ容量が最大の制約条件のようだ。

また SAS のシステムは、BASE SAS で 7.8MB、SAS/STAT で 6.3MB と合計すると 14.1MB 分のディスクを使う(パーソナル・コンピュータ用の SAS リリース 6.04 の全プロダクトでは 38.2MB のディスク容量が必要になる)。MS-DOS 自体でも、放って置くと 3MB 以上ディスクを使う。本書で使用する SAS プログラムの例では、164 変数×907 件で 1.2MB の永久 SAS データ・セットと呼ばれる MS-DOS ファイルが固定ディスクに常駐することになるので、作業用ファイルのことも考えると、理屈の上では、固定ディスク 20MB は SAS が「統計パッケージ」として稼働可能なぎりぎりのサイズである。それが実際に稼働することを今回確認できたので、パーソナル・コンピュータ用に SAS の導入を考えている読者は、ハードウェア選択の際の参考にしてほしい。

本書の特色と使い方

1. 学生、社会人の初学者を広く対象にする。

本書の執筆にあたって想定している読者は、統計学やコンピュータを専攻しない学生、社会人である。実例や素材が経営学分野から選ばれているので、特に経営学分野の学部学生・大学院生、あるいは企業の人事・教育・能力開発などの担当者には、興味をもってもらえると思う。

2. 統計学、コンピュータの知識を前提としない。

本書は、原則として「高度」「難解」な統計学理論の理解を目指すものではない。統計学、調査、データ解析の考え方、取り組み姿勢を体得することを目指す。統計学、コンピュータの知識は前提にせず、それらの基礎については的を絞って、丁寧に説明する。

3. 調査研究プロセスの全体をカバーし、使える統計学を目指す。

本書の目的は、読者が自ら調査データを集め、自ら統計学やコンピュータ、統計パッケージを駆使して、調査データを分析できるようにすることである。最終的には、卒業論文・修士論文レベルの論文、報告書の作成に「使える」統計学、統計調査法を目指してお

り、特に、参入障壁の高い、コンピュータを使った調査データの統計処理に重点が置かれている。

4. 統計学教育と情報処理教育の融合。

本書は私なりの新しい統計学教育のコンセプトに基づいて書かれている。大学で授業のテキストとして使用するときには、次のような使い方ができる。

a. 《1 年間の「統計学」の講義》従来のテキストのように、各章に 3〜4 回のペースで講義する。その際、希望者には SAS を利用できる機会を与えてやることが望ましい。
b. 《半年の「応用統計学」の講義》従来の統計学教育も重視したければ、1 年の前半は基礎統計学として、他の入門的統計学テキストを使って「統計学」の講義を行っておき、後半の応用統計学相当部分として、本書を第 2 章だけを飛ばして各章を 2〜3 回ペースで、「データ解析」を中心にして講義する。
c. 《週 2 コマ半年間の実習》週 1 コマは本書を使って、データ解析中心の講義として、もう 1 コマはコンピュータ、統計パッケージの実習として本書の演習問題を課題として与え、その都度、レポートを提出させながら進める。
d. 《ゼミナールでの卒業論文指導の一環として 4 日間の講習会》1 日の前半は本書を使った講義あるいは輪読、後半は本書の演習問題を課題として与えた実習の 1 日 2 部構成の講習会形式で第 2〜5 章を 4 日間に 1 日 1 章のペースで集中的にゼミナールを行う。

5. 単独でも使える独立性の高い各章

本書は全体として調査研究プロセスを構成するように書かれているが。その一方で各章は教材としてかなり独立性をもたせている。各章がやや長めなのはそのためで、SAS の部分を除けば、必ずしも 1 冊全部を読まなくても、必要な部分だけを重点的に使用することができる。例えば、本書を企業などでより短期間の研修用テキストとして使用する場合には、「統計・調査概論」ならば第 1 章だけを丁寧に教えればよい。調査の分析結果に接する人に対しては、平均値の利用に関しては第 3 章だけを教えてもよい。相関を中心に第 4 章だけ、あるいはクロス表の読み方に関して第 5 章だけを教えてもよい。SAS 講習会であれば、第 2 章だけを教材に使うこともできる。

6. 本書を片手に自分で調査、分析してみよう。

本書には、実際の統計調査がまるごと掲載されている。第 6 章はもともと統計調査マニュアルを意図して書かれているので、手順にしたがって、自分なりの企画で調査を行うこともできる。

第 1 章　統計調査データと誤差

章目次

1. はじめに
2. 調査のもつ意味
3. 統計学の役割
4. 調査にともなう誤差
5. 無作為標本抽出と確率
6. 有意確率と仮説検定
7. 標本誤差と標本の大きさ
付. 標本比率の分布
演習問題

1. はじめに

統計学は、もともと多くの分野からの小さな支流が、過去 2 世紀以上もかかって合流し、一つの太い流れになった学問である。現在の統計学の流れの中には、自然科学に属する流れも社会科学に属する流れもあるが、人間の生活のあらゆる面とあらゆる科学が統計学と関わってきたといってもいいだろう。いい換えれば、現象の法則性に対する人間のあくなき探求心が統計学を生み出してきたのであり、統計学は科学の文法(the grammar of science)であると呼ばれる由縁でもある。

「統計学」という用語自体は、17 世紀のドイツの国勢学に遡るといわれる。国勢学とは、国家顕著事項の総体の記述に関する学問であり、国情記述学とでも呼ぶべきものであった。今日、統計学を指す Statistik (英語では statistics)は、国家 Staat (英語では state)の状況を歴史的に記述することを意味しているといわれる。英語の statistics には、このように、集団から引き出された数量的情報としての統計数字を指す統計もしくは「統計データ」の意味と、もう一つ、統計データを処理、分析して、集団についての有意義な情報を得ることを主要な課題とする学問である「統計学」の二つの意味がある。二つの異なる概念に同じ単語"statistics"が用いられているのはこうした学説史的事情によるものである。

統計学とは科学する心の結晶であり、社会科学の分野においては、社会、経済、経営現象を含めて、多くの現象がどんなふうに(how)になっているかを調べる学問である。そのプロセスはまず調査(survey)から始まる。この章では、統計学および調査の考え方を概観し、それにもとづいてデータの収集方法や尺度、そして調査にともなう誤差について考えてみることにしよう。

2. 調査のもつ意味

調査および調査結果への接し方に関しては、いくつか注意を要する点があるので、そのことから話を始めることにしよう。

(1)調査結果と実態

まず最初に、調査結果は実態そのものを表しているのではないということには注意する必要がある。調査においては、回答に各回答者の真の意見が表されているかどうかはわからないし、質問の仕方によってはさまざまな統計数字が出てくることになる。つまり、調査結果はそのままで何らかの実態を表したものではない。しかし、調査結果は「ある質問に対して、ある人からある回答が返ってくる」という、まさに「事実」の集団的な積み重ねである。統計的な処理を施された調査結果は、集団を見るための一つの指標(index)なのである。

この指標を実態そのものと受け取るのは誤りである。この事実(＝指標)が何を意味するかは、対象となっている社会あるいは企業の動きと突き合わせることによってはじめてわかってくるものである。例えば、政党支持率は選挙の際の政党の得票率とは一致しないが、政党支持率の推移は政治情勢の行方を見るための一つの指標となっている。そして、政党支持率の推移を他の様々な要因と組み合わせて考えることで、政党の得票率を占う重要な手がかりにもなるのである。つまり、統計処理された調査結果は一つの事実であり、この事実こそ実態を探る手がかりなのである。

(2)調査と因果関係

次に、調査あるいは統計では因果関係は分からないということにも注意する必要がある。いま 2 変数 x と y との間に因果関係が存在するということがいえるためには、次の 3 つの条件が満たされている必要がある。

1. x と y との間の相関が 0 ではなく、
2. x と y との間に時間的順序が存在し、
3. x と y との間に観察し得る関係を生み出す可能性をもつ他の原因が消去できること。

この 3 つの条件を満たすことは可能であろうか。このうち一見難しそうな 2 は実は達成可能である。1 回限りの調査であっても質問文を工夫することで、例えば、変数 x については現時点での値、変数 y については 1 ヵ月前の値を思い出して回答してもらうというようにすれば、変数間に時間的順序を課すことはできる。変数自体の時間的順序と、データ収集に際しての時間的順序とは別ものなのである。問題は、1 と 3 で、他の因果変数の影響を除いた後にも x と y との間に相関が存在することを示さなくてはならない(相関については第 4 章で後述)。しかし、潜在的には、そのような変数は無限に存在しているのであって、どの変数をモデルに入れるかの選択は理論的考察にかかっているのである。したがって、この 1 と 3 の 2 条件を満たすことは本質的に不可能である。2 変数 x と y との間に因果関係が存在することを統計的に立証することはできない。

考えてみれば、調査票を作成する段階で、意識している、していないに関わらず、既に変数の選択を行ってしまっているのであり、暗黙のうちに何らかの因果関係を前提にしてしまっていることになる。もしデータ収集に際して、中心的な変数を落としてしまったり、中心的な概念を妥当性を欠く方法で測定してしまったりしていれば、統計的には正しい分析が実際には誤った結果を導くことにもなる。

さらに、相関は方向性のない対称性のある性質であるから、たとえ相関があるとわかっても、分析者が何らかの理論に基づくか、もしくは自分の頭で考えるかして状況を設定しない限り、変数間の因果の方向を決めることはできない(例えば第 5 章第 6 節を参照のこと)。その意味では、どんな統計的方法も究極の因果関係を扱うことはできないし、そもそも因果関係が存在することさえ教えてはくれないのである。結局、調査あるいは統計では因果関係はわからない。

(3)調査の目的

それでは、調査はどういった目的で行われるのであろうか。まず挙げられるのが、事前の知識の精度を向上させることである。その代表は、通常の研究論文などにみられるように理論に基づいた仮説の構築とその検証である。しかしそれだけには限らない。ただ単に調査を企画、実施し、その結果として得られた統計数字を使うというだけでも、それは自らの知識、常識、理解度を試されていることを意味している。なかには、結果の意外性を期待して調査を行う人がいるが、そうやって調査を考えることは、調査の本質を誤解しているといってよいだろう。調査は仮説とまではいかなくても、たとえ暗黙のうちにでも事実や因果関係についての事前の知識を利用して設計されるものであり、事前の知識が不十分であれば、ヒアリング調査や文献などを通じて、ある程度の知識の獲得が図られるべきである。このようにして、調査の設計が念入りに行われていれば、調査結果に意外性を感じることはほとんどない。

もしも意外な調査結果が得られるようなことがあれば、調査方法等に誤りがなかったか、あるいは自分がとんでもない考え違いをしていなかったかをまず疑ってみる必要がある。経験的には、「意外な」調査結果は「いい加減な」調査設計から生まれることが多い。「常識を覆すような調査結果」は基本的にはあり得ないと心得て、調査あるいは調査結果に接するべきであろう。良い調査では、当り前のことが当り前に結果となって出るものであり、調査をすることにより、それまで漠然とそれらしい傾向があると感じていたことが、より精確に統計数字となって知り得るようになるものである。

調査のもう一つの目的は、その次の調査のためのヒント、ヒラメキを得ることである。常識を覆すような調査結果は基本的にあり得ないと心得て、念入りな調査設計を行ったにもかかわらず、意外な、あるいは、それまで気づかれていなかった新しい事実・関係を発見することがある。いわゆる事実発見である。その時には、もう一度調査を行ってみなくてはならない。つまり、繰り返し調査を覚悟するというのが、調査の本来の姿である。そして、その事実や関係を論理的にどのように説明できるかを常に考えていく姿勢が大切である。

既に述べた通り、統計学とは科学する心の結晶であり、社会科学の分野においては、社会、経済、経営現象を含めて、多くの現象がどんなふうになっているのかを調べる学問である。統計数字で事実を記述することを目指してはいるが、論理を生み出すことはできない。論理は常に人間の頭の中から生み出されるものである。そして、その論理や常識として知っていることの精度を向上させることが、統計調査の主目的なのである。

(4)「組織活性化のための従業員意識調査」

本書では、こうした統計調査の実際の姿をできるだけ具体的かつ明確にイメージしてもらうために、第 2 章以降に登場する統計処理、分析に用いるデータ例をはじめ、SAS プログラムの例なども、すべて一貫して「組織活性化のための従業員意識調査」を素材として、架空ではなく実際のものを用いている。この調査の詳細に関しては、第 6 章で統計調査の例として、その調査手順と方法が、実際に用いられた詳細な資料とともに取り上げられる。第 6 章はもともと統計調査マニュアルを意図して書かれているので、興味のある読者は、第 2 章に進む前に読んでおくと、統計調査の感覚がつかめるだろう。「組織活性化のための従業員意識調査」は、1986 年以来毎年 6 月～翌年 1 月の 8 カ月をかけて、(財)日本生産性本部経営アカデミー「人間能力と組織開発」コースを舞台にして、筆者によって繰り返し繰り返し企画・実施されてきたものである。1991 年までに、のべで約 50 社、約 5,000 人を調べてきた。第 2 章以降では、そのうちある年度の調査データをデータ例として扱うことにした。

この調査では、企業間での横断的な研究グループを作り、このメンバー企業の間で同時に同一の従業員意識調査を行い、企業間・職場間の比較集計を行う。他社の実例や知恵も借りながら、調査結果の企業間比較によって、自社の抱える問題点を探り、事例研究を行うというプロセスを繰り返すことで、研究グループを中心として組織の活性化に関する問題発見を図るのである。この「組織活性化のための従業員意識調査」では、従業員の意識調査を行なうことを契機として問題発見プロセスを進めていくことになっている。

こうして毎年繰り返し繰り返し行なわれるこの調査を通じた調査ノウハウの蓄積で、経営組織を対象とした統計調査は洗練されたものになりつつあり、ヒアリング調査と統計的な比較調査の繰り返しサイクルが、組織開発(organization development)、あるいは日本的にいえば組織活性化(organizational activation)における数量的手法としてかなりの効果をもっていることが経験的に明らかになってきた。特に調査結果の統計数字を前にしての事後的なヒアリング調査はかなり効果的で、的確に核心的な事実について聞き出すことを可能にしてくれる。それは、効果的なヒアリング調査のために統計調査をしているといってもいいほどである。

統計的に処理された結果は、しばしば被験企業にとっては意外な数字であるらしいが、外部から第三者として観察している研究者にとっては、なぜ意外と感じるのか理解に苦しむことがほとんどである。第三者的には、統計数字は事前のヒアリング調査の内容を素直に反映した、経験的、常識的にもっともなものばかりだからである。そして、事実、その後引き続き行われる事後的なヒアリング調査によって、1 企業の内部の人間が抱くこの「意外性」の溝はどんどん埋められていくことになる。その意味では、1 企業という狭くて閉鎖的な世界にしか通用しない前提、常識が、統計数字のもつ明白さをきっかけにして、事実によって棄却を迫られることになるのである。それはこの章の中でこれから扱う仮説検定と同じ論理である。「うちの常識＝よその常識」という仮説が、統計データによって棄却されることを意味している。こうした当該企業のメンバーが共通してもっている前提、常識の客観的な同定と吟味・検討が組織活性化にとっては重要なのである。

3. 統計学の役割

それでは、調査研究プロセスにおいて、統計学がどのような役割を果たしているのかを、調査対象、原データとの関係を考えながら、調査研究プロセスの経過にしたがって、調査対象と調査の種類、原データと統計処理の順番に見ていくことにしよう。

(1)調査対象と調査の種類

調査研究プロセスはまず観測(observation)から始まる。これは自然科学でも社会科学でも同じで、自然科学の分野では、それは一般に実験(experiment)とよばれ、それに関する統計理論としては「実験計画法」がある。他方、社会科学の分野では、それは調査(survey)とよばれ、その統計理論は「社会調査法」である。

社会科学分野での調査においては、例えば、個人、世帯、職場、会社などが調査の単位となる。そして、調査の対象となりうるすべての単位を集めたもののことを母集団(population)とよぶ。母集団とは、調査対象の集まりであって、調査によって、それについて何らかの結論を下そうとしているものである。

調査の対象となる単位のことは要素(element)とも呼ばれる。そういう言い方をすれば、母集団とはまさに集合であり、特定の要素をその中に含ませるかどうかについての明確な規則(ルール)がなくてはならない。このルールには、通常は、単位そのもの、時間、場所な

どに関する記述が含まれている。母集団は調査者から独立に存在しているものではなくて、調査者が具体的かつ明確に設定すべきものである。

調査方法については、統計学の中では既に述べたように「社会調査法」が取り扱うべき課題であるが、ここでおおまかに説明しておこう。一口に調査といっても、これにはさまざまな種類があり、調査対象である母集団の大きさや母集団の数によって異なった種類の調査が行われる。まず最初に、母集団の大きさによって次の二つの調査方法が選択される。

1. 全数調査(census): 母集団を構成するすべての要素を調べる調査
2. 標本調査(sampling survey): 母集団から適当な部分、すなわち標本(sample)を抽出して調べ、その標本について得られた知識に基づいて、母集団に関する推論を行う。

全数調査と標本調査のどちらを選択すべきかは、母集団の大きさ、調査に必要な労力、コストによって決る。例えば、全数調査の代表例としては、国勢調査(人口 census)がある。国勢調査は、決められた年の 10 月 1 日現在の日本の全人口を対象にして行われる全数調査であって、1920 年(大正 9 年)以後 5 年ごとに定期的に行われている。(ただし、戦争の影響のために、1945 年(昭和 20 年)の調査は実施されず、1947 年(昭和 22 年)に臨時調査が実施された。) 一般に、国勢調査に代表される日本全国を対象とする全数調査は、巨額の費用と大量の人員と長い時間を要する。したがって、国勢調査以外には、事業所統計調査、農林業センサス、工業統計、商業統計など、全数調査の数は限られている。実際に、われわれが「日本の○○についての調査」としてよく耳にするのは、そのほとんどが標本調査である。標本調査については、後でこの章の中でより詳しく説明する。

もっとも、このように調査といえば従来はほとんど標本調査を意味していたのであるが、近年の著しいコンピュータ・テクノロジーの進歩は、一つの企業の従業員を調べる程度では、全数調査を十分可能にしてしまった。パーソナル・コンピュータですら、数千人規模の調査をほとんど問題なく集計、分析できる。特に経営分野においては、標本調査が死語になる日が遠からずやってくるだろう。

次に、調査する母集団の数と調査時点によって、調査は表 1.1 のように、横断的調査、比較調査、パネル調査、繰り返し調査の 4 種類に分類される。このうち基本になるのは、一つの母集団に対する 1 回限りの調査である横断的調査である。他の調査方法は、この横断的調査を色々と組み合わせたものである。

表 1.1　調査方法の種類

調査時点	**母集団**	
	同一母集団	**複数母集団**
一時点	横断的調査	比較調査
複数時点	パネル調査	繰り返し調査

1. 横断的調査(cross-sectional survey): 単一の母集団に対して行う、1 回限りの調査。もっとも基本的な調査方法である。他の調査方法は、この横断的調査を、複数時点または複数母集団で組み合せたものである。
2. 比較調査(comparative survey): 複数の母集団の比較を行うための調査。横断的調査を複数母集団に関して同時に行ったものである。

3. パネル調査(panel survey): 単一の母集団に対して 1 回だけ標本抽出を行い、同じ標本に対して複数時点で反復して行う調査。その固定された標本をパネル(panel)とよぶ。法人企業統計調査などはパネル調査として実施されている。ところで、パネルを固定するということは、その間、母集団の時間以外の要因は変動しないと暗黙に仮定している。たとえば、ある年に何人かの従業員をパネルとして固定した場合、パネルの平均年齢は 1 年で確実に 1 歳上昇する。しかし、実際の母集団には、その間に異動、退職等があって、高年齢層は脱落していき、代わって若年層が入ってくるので、母集団の平均年齢は、たとえ上がったとしても、パネルのそれほどには上昇しない。つまり、パネルは老化しやすいのである。そのため、パネルはほぼ毎年、更新をせまられることになる。
4. 繰り返し調査(replicated survey): 複数母集団に対して複数時点で標本抽出を繰り返して行う調査。パネル調査のようにパネルを固定しないので標本は調査のたびに異なる。

調査時点が異なれば、母集団を規定する他のルールが同じでも、厳密には母集団の要素は異なるので、このような場合も、厳密には繰り返し調査になる。しかし一般に、調査時点が異なるだけで、母集団を規定する他のルールが同じならば、そのような調査から得られたデータは、時系列データ(time-series data)とよばれる。時系列データによって、時間の経過による変化や異時点間の関係を知ることができる。パネル調査も、パネルを固定する間、母集団の要素は変動しないと仮定して、時系列データを得るために行われるものである。

いずれにせよ、調査結果や分析結果を正しく理解するためには、その調査がどのようにして行なわれたどんな種類の調査であるかを知らなくてはならない。ややもすると、調査結果の統計数字だけが一人歩きしがちであるが、統計数字は調査方法と常にペアで評価されるべきものなのである。(→演習問題 1.1)

(2)原データの尺度と統計処理

調査の結果は、原データ、もしくは単にデータとよばれ、統計的記述の素材となる。データとは、標本調査ならば標本、全数調査ならば母集団に属する各要素に関する観測値のまとまりを指している。最近のように、コンピュータを使った集計、統計分析が中心になってくると、データの処理上の便宜を考えれば、観測値はカナやアルファベットなどの文字ではなくて、数値で表されていることが望ましい。

ところで、観測値が数値となっていれば、形式的にはそれらの間の演算が可能になる。しかし、注意が必要なのは、観測値がどんな尺度(scale)によって測定されたのかによって、どのような水準の演算を行うことができるかが決まってくるということである。これはコンピュータに判断できることではなくて、人間の側できちんとコンピュータに指示しておくべきことである。よく用いられる尺度の分類としては、次の 4 つがある。

1. 名義尺度(nominal scale): 観測値が単に対象の分類、カテゴリーを示しているものである。例としては、男性に 1、女性に 2 という数値を割り当てた場合や、郵便番号などがある。
2. 順序尺度(ordinal scale): 観測値が序数としての意味をもち、対象間の順序付けを示しているものである。例としては、中学卒に 1、高校卒に 2、大学卒に 3 という数値を割り当てた場合や 1 番、2 番……といった成績順位などがある。

3. 間隔尺度(interval scale): 観測値が任意の単位の何倍という形で示され、測定値の差が意味をもつものである。例としては、時刻や摂氏、華氏の温度などがある。
4. 比率尺度(ratio scale): 間隔尺度と同様に、観測値が任意の単位の何倍という形で示され、さらに原点 0 が絶対的な意味をもっていて、観測値の差だけではなくて、比率も量として意味をもつものである。例としては、時間、質量、個数などがある。

名義尺度、順序尺度にもとづくデータのことを質的(定性的)データ(qualitative data)といい、間隔尺度、比率尺度にもとづくデータのことを量的(定量的)データ(quantitative data)という。上のリストの 1 から 4 へとなるにしたがって、尺度としては上位になっていく。つまり、上位の尺度にもとづく観測値には、下位の尺度にもとづく観測値の意味、およびそれに可能な演算が包含されている。観測値の演算可能性は、

1. 名義尺度では計数にもとづく演算だけが意味をもつ。
2. 順序尺度では順位に関する演算も意味をもつ。
3. 間隔尺度では加減の演算も意味をもつ。
4. 比率尺度では加減乗除の演算も意味をもつ。

こうした演算可能性については十分に注意を払わねばならない。

(3)記述統計学と統計的推測

調査によって得られた原データに対して統計処理を施したものが統計資料である。これは、よく調査結果ともよばれる。このとき、与えられた原データを調べ、その規則性から統計的法則を発見する記述統計学(descriptive statistics)が活躍することになる。つまり、図表によってデータを整理し、平均、分散などの特性値によってデータを要約するというように、データを整理・要約して、母集団や標本の集団としての特徴を記述するのである。ただし、データを数値的に要約する場合には、(2)で述べたように、データが測定された尺度が問題になる。データの演算可能性については、データ処理を始める前に、必ず考えておく必要がある。

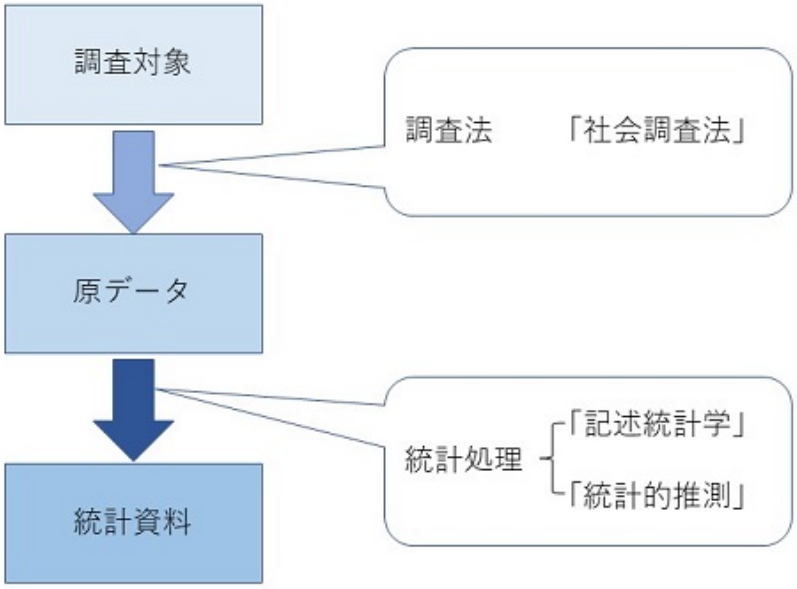
ところで、統計学の役割はデータの整理・要約だけではない。全数調査のように、母集団を構成するすべての要素を調べる調査の場合には、記述統計学だけで十分であり、それによって得られる結果に、それ以上なにも付け加えるべきことはない。ところが、より頻繁に行われる標本調査のように、母集団から適当な標本を抽出して、その標本について得られた知識にもとづいて、母集団についての推論を行う場合、なぜ全数を調べずにその一部の標本を調査するだけで、母集団全体についての推論が可能になるのかという疑問がわいてくる。ここで推測統計学が重要な役割を果たすことになる。

実は、母集団の部分である標本が、無作為抽出によって選ばれていれば、確率という論理装置を通して、部分から全体を知ることは可能なのである。このとき、母集団全体ではなくて、その一部を観察して、その結果にもとづいて、全体の法則性の発見することを統計的推測(statistical inference)とよぶ。統計的推測には、仮説検定(hypothesis testing)と推定(estimation)という二つの柱がある。もちろん、標本調査から導き出された結論は、そこから標本を抽出した母集団についてのみ妥当することになる。そして、確率論および統計学理論によって、標本抽出にともなう誤差を客観的に評価することが可能になるのである。

このように、調査においては、調査にともなって発生する誤差を抑え、あるいはその大きさを評価することが重要になる。実は、調査にともなう誤差は標本抽出にともなうもの

ばかりではなく、たとえ標本抽出を行わずに、全数調査を行った場合でさえ発生する誤差も存在する。そこで、次節では、こうした調査にともなう誤差についてまとめておこう。

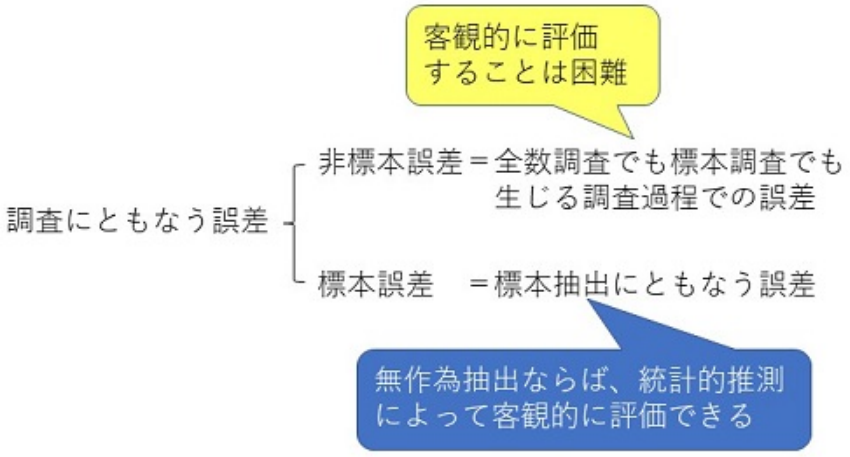
図 1.1　社会科学における統計学の役割



4. 調査にともなう誤差

調査にともなう誤差には、図 1.2 に示すように標本誤差と非標本誤差の 2 種類がある。このうち、非標本誤差(non-sampling error)は、全数調査でも標本調査でも生じる調査過程での誤差であり、これについて客観的に評価することは難しい。統計学理論では標本誤差については、かなり詳細な吟味が行われるが、非標本誤差については、なおざりの感がある。しかし、現実に行なわれる通常の調査では、非標本誤差の存在とその大きさは、かなり深刻な問題となっている。

図 1.2　調査にともなう誤差



(1)非標本誤差の種類

非標本誤差としては

1. 無回答の誤差……調査もれや回収率が低いことにともなう無回答の誤差・偏り。
2. 回答の偏り……回答者が意識的・無意識的に偏った回答をするような場合に生じる偏り。
3. 単純ミス……調査員、回答者による数字や記号のつけ間違いや集計上の転記ミス、データ・エントリー・ミス、計算ミス。

が考えられる。このうち非標本誤差 3 の単純ミスについては、たとえば、できるだけ質問の回答の形式を統一しておいたり、欠損値の指定を厳密かつ周到にしておいたりすることで、転記ミスやデータ・エントリー・ミスの誘発を抑えることができる。またデータ・エントリー・ミスはいわゆる 2 度打ちでほぼ回避できるし、計算ミスは例えば本書でもこれから扱う SAS のような既に広く普及しているコンピュータ用の一般的な統計パッケージを使えば、多数のユーザーが、パッケージ化されたプログラムを(つまり計算手順や計算式を)チェックしていることになるので、ほとんど心配ないといえる。とはいえ、単純ミスを完璧に防ぐことは不可能である。そこで、こうした単純ミスに対しては、調査の各段階で適切できめの細かい管理をすることで、ある程度ミスの発生を抑える努力をすることになる。そのためには調査の規模があまり大きくない方が、管理がしやすいわけで、これが標本調査の最大の利点となっている。特にデータ入力、点検以後をできるだけ少人数で行うことによって、管理上の不手際から生じる単純ミスはかなり回避することができる。

(2)調査方法と非標本誤差

非標本誤差のうち、1. 無回答の誤差、2. 回答の偏りについては、どのような調査方法をとるのかによって、そのおおよその大きさが決ってくる。質問調査票を使った調査方法としては、大きく分けて、表 1.2 のような方法が考えられる。

表 1.2　調査方法の分類

	記入者	留置	配布	回収
面接調査法(interview survey)	他記式	×	調査員	調査員
面前記入法	自記式	×	調査員	調査員
配布回収法(留置(ﾄﾒｵｷ)法)	自記式	○	調査員	調査員
郵送回収法	自記式	○	郵便	調査員
郵送法(mail survey)	自記式	○	郵便	郵便

表 1.2 の中にある他記式(または他計式ともいう)とは、調査員が調査票にしたがって質問し、それに対する調査対象の回答を調査員が調査票に記入する方式である。それに対して、自記式(または自計式ともいう)とは、調査対象が自分で調査票の質問を読みながら、自分で回答を調査票に記入する方式である。

これらの調査方法の比較は表 1.3 のようになる。非標本誤差との関係では、回収率、回答の偏りに注意しながら、調査方法を選択する必要がある。

通常、郵送法の回収率はきわめて低く、表 3 の中の数字よりもさらに低く、10%そこそこのことも多い。郵送法で回収率が高いときには、むしろ、標本抽出の仕方や調査の仕方

を疑ってみた方がよいといわれるほどである。いずれにせよ、一般には、郵送法で収集されたデータは低回収率のため、これから(3)で述べる無回答にともなう非標本誤差が大きすぎ、統計的推測はあてにならないので要注意である。

この表 3 の中で、面接調査法、面前記入法に見られる回答の偏りは、主に回答者の匿名性が回答の際に保たれていないことによるものである。少なくとも回答の際には、回答者は調査員を目の前にしており、調査員の側でも誰が回答しているのかを認識しているからである。したがって、例えば回答の偏りの 1 については、調査員が主婦であるか、男子学生であるか、女子学生であるかによって、見栄や恥ずかしさから回答の傾向が異なってくることが十分に考えられる。

表 1.3　調査方法の比較

	面接調査法	面前記入法	配布回収法	郵送回収法	郵送法
調査費用	• 高い		• それほど高くない		• 安い
調査票の配布回収に要する時間	• 調査対象の協力を得るまで 3〜5 回の訪問が必要 • 数週間かかることもある		• 1 日〜数日留め置く		• 催促を入れ 1 ヵ月程度 • 回収打切りで追加集計リスクがある
調査票の回収率	• 高い(80%前後)		• 高い(80%以上)		• 低い(30%以下)
調査対象本人の確認	• 確認できる		• 確認できない		
質問の意味の理解度	• 調査員が逐一説明できるので一様に高い		• 質問の意味を良く理解しないで回答することがありうる		
回答の偏り	• 調査員の性別、年齢、意見による影響 • 調査員にプライバシーを明かすための偏り • 長い調査期間中の事件や出来事の影響		• 質問によって無記入、あるいは、全く的はずれで事実上の無記入といってよいものが生じやすい。 • ある程度のレベルの国語力が前提になっている。		

また調査員が口にしなくても、調査員の意見によっても影響を受けるかも知れない。「調査方式の比較研究」(杉山 1984, p.66)によると、「奇跡を信じるか？」という質問に対して、表 1.4 のように調査員の意見の影響が見られるといわれる。

表 1.4　調査員の意見と回答

調査員の意見	回答		
	奇跡を信じる	奇跡を信じない	計
奇跡を信じる	25%	75%	119 (100%)
奇跡を信じない	12%	88%	255 (100%)

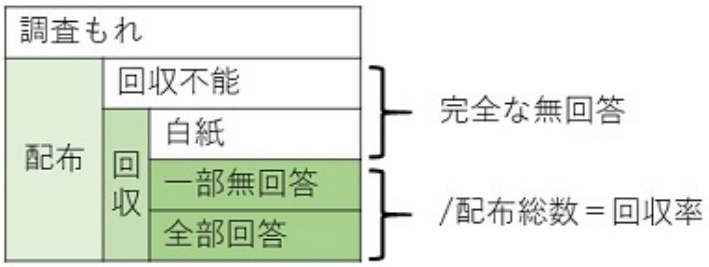
(出典: 杉山(1984) p.66)

本書で取り上げている「組織活性化のための従業員意識調査」の場合には、配布回収法によって、1 週間以内の短期間に、できるだけ高い(90%前後)の回収率で 500 人以上を調査することを目標として行われる。こうした配布回収法が可能なのは、企業側のニーズにある程度合致した内容にすることで、配布回収に際しての協力が得られるという人員的な側面だけではなく、調査をされる側の従業員の高い国語力という側面も見逃せない(表 1.3 の「回答の偏り」の欄も参照のこと)。「組織活性化のための従業員意識調査」では、比較的大きな企業で大卒中心のホワイトカラーを調査対象としていることで、質問の意味の理解度がある程度保証されているのである。

(3)無回答と非標本誤差

ここで、非標本誤差のうち大きな部分を占める「1. 無回答の誤差」について、その大きさがどの程度のものになるのかを考えてみよう。無回答の誤差は、調査もれや回収率が低いことにともなう無回答の誤差・偏りを指している。通常、図 1.3 に示されるように、調査では、母集団のリストの不備、回答者不在のための脱落、回答拒否などによる調査もれを除いて、調査票が配布される。この配布された調査票の中の無回答がここでいう無回答でありる。そのうち、完全な無回答は、配布された調査票のうち回収不能だったもの、回収はしたが白紙だったものとに分けられる。こうした配布されたものの完全無回答であった調査票を除いた残りの部分、つまり回収されかつ一部または全部回答されている調査票数の、全配布数に占める割合を回収率と呼ぶ。

図 1.3　調査もれと回収率



一般に、無回答の誤差が重大になるのは

1. 調査票の回収率(responce-rate)が低く、かつ
2. 無回答者(non-respondent)群の特性と回答者(respondent)群の特性とが著しく異なっている場合

である。1 については、実際には、回収率に算入されている調査票の中にも質問によっては

部分的に無回答が含まれている場合が多く、見かけの回収率以上に質問ごとの有効回答率は低くなる。それだけ無回答の誤差は深刻である。また 2 については、一般に、無回答者群に比べて回答者群の方が質問に対する肯定的な回答が多くなる傾向があること、また企業単位の調査では、大企業に比べて、中小企業の回収率はかなり低くなる傾向があることを十分考えておいた方がよいであろう。

こうした無回答の誤差がどの程度の大きさになるものなのか、いま例として、Yes-No 形式の質問を考えてみよう。調査結果で(すなわち回答者群で) 60%が Yes だったとき、これだけ見ると過半数が Yes と答えたと言えそうである。しかし、本当にそうだろうか。この程度の Yes 比率では、実は無回答者群での Yes が 40%で、Yes 比率と No 比率とが逆転していた……というようなケースがよくあるのである。特定のテーマの調査に対しては、そのテーマに肯定的な人は回答に協力的だが、否定的な、あるいは嫌悪感をもっている人は、回答をめんどうがったり、拒否したりするという傾向があるので、むしろこの程度の両群の Yes 比率の差はごく普通に存在すると考えておいた方がよい。

仮に有効回答率が 50%とすると、全調査対象の Yes 比率は、両者のちょうど中間の 50%だったことになる。つまり、

$$0.6\times0.5+0.4\times0.5=0.5$$

となる。もしも有効回答率が 20%ならば、

$$0.6\times0.2+0.4\times0.8=0.44$$

となってしまい、なんと実際の全調査対象では Yes 比率は過半数を割り、44%しかいなかったことになるのである。回答者群だけを見たときの見かけの Yes 比率とは結論が逆転する。実は過半数が No だったのである。

それでは、こうした考えをもう少し一般化して、このような Yes-No 形式の質問の無回答の誤差がどの程度の大きさになるのかを考えてみることにしよう。

全調査対象の Yes 比率

＝(回答者群の Yes 者数＋無回答者群の Yes 者数)/配布総数

＝(回答者数×回答者群の Yes 者数)/(配布総数×回答者数)

　＋(無回答者数×無回答者群の Yes 者数)/(配布総数×無回答者数)

＝有効回答率×回答者群の Yes 比率＋(1-有効回答率)×無回答者群の Yes 比率

＝回答者群の Yes 比率

　＋(1-有効回答率)×(無回答者群の Yes 比率－回答者群の Yes 比率)　(1.1)

ここで、(1.1)式の下線部分が「無回答の誤差」ということになる。この無回答の誤差を実際に計算してみると、表 1.5 のようになる。

表 1.5　無回答の誤差(網掛け部分は無回答の誤差が 5%以下)

有効回答率(%)	｜無回答者群の Yes 比率(%)－回答者群の Yes 比率(%)｜										
	0	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
100	0	0	0	0	0	0	0	0	0	0	0
90	0	1	2	3	4	5	6	7	8	9	10
80	0	2	4	6	8	10	12	14	16	18	20
70	0	3	6	9	12	15	18	21	24	27	30
60	0	4	8	12	16	20	24	28	32	36	40
50	0	5	10	15	20	25	30	35	40	45	50
40	0	6	12	18	24	30	36	42	48	54	60
30	0	7	14	21	28	35	42	49	56	63	70
20	0	8	16	24	32	40	48	56	64	72	80
10	0	9	18	27	36	45	54	63	72	81	90
0	0	10	20	30	40	50	60	70	80	90	100

↑

そこでさきほどのように、たとえば、回答者群では Yes が 60%で過半数が Yes と見えていても、実は、無回答者群での Yes が 40%だった場合を考えてみよう。このとき、両群の Yes 比率の差は 20%になり、表 1.5 の矢印の列を見るとわかるように、有効回答率が 50%未満のときには、無回答の誤差が 10%を超えるので、全調査対象(標本)の Yes は実は少数派だったという逆転現象が起こってしまう。

後で、第 7 節で述べるように、標本抽出にともなう標本誤差の目標精度を 5%程度に抑えたいと考えているときには、非標本誤差もせめて同程度の 5%くらいには抑えておかないと意味がない。そしてこのときには、調査票の回収率(＞有効回答率)は 80%以上を目標にする必要があることになる。さらに表 1.5 から、調査票の有効回答率が 90%以上あれば、まず逆転現象の心配はないこともわかる。

(4)標本誤差

非標本誤差に対して、標本誤差(sampling error)は標本抽出に伴う誤差であり、標本の抽出方法(sampling method)によっては、客観的に取り扱うことができる。具体的には、次に述べるような有意選出法を用いた場合は困難であるが、無作為抽出法を用いて標本抽出を行った場合には、確率論および統計学理論によって客観的に評価することができる。これが統計的推測の理論である。つまり、統計的推測は、どのような標本に対しても使えるわけではない。このことをもう少し詳しく見てみよう。

(a)有意選出法(purposive selection)

これは、調査者が適当に標本を選ぶ方法である。企業が取り引き先の企業を調査するような方法(機縁法)や、モニターを募集してその意見を聞くような方法(応募法)が有意選出法に含まれる。しかし、もっとも体系的な方法は、割り当て法(quota method)と呼ばれるものであって、

i. まず母集団に関する予備知識をもとに、母集団を特定の標識によって、ある種の属性ごとにグループ分けした上で、
ii. 各グループの大きさに比例した大きさの標本を各グループに割り当て(層別の比例配分)、
iii. 各グループ内での標本抽出は調査員の主観にまかせるという方法である。

もしこのうち iii の各グループ内での標本抽出が無作為に行われるならば、この方法は無作為抽出法の一種の層別抽出法とよばれる方法になる。しかし、この場合には各グループ内での標本抽出は調査員の主観にまかされるために、有意選出法となる。

(b)無作為抽出法(random sampling)

これは、クジ引きの原理で標本となる要素をランダムに(無作為に)選び出す方法である。厳密には、母集団の要素リストで、各要素に通し番号をつけ、この通し番号を、乱数表をもちいて抽出していく。あるいは、さいころや 0 から 9 までの数字が 2 面ずつ刻まれている正 20 面体の「乱数さい」を振って、出た目をもとにして抽出していく。これを簡便化したものに、系統抽出法あるいは等間隔抽出法と呼ばれる方法がある。厳密な意味では無作為抽出法ではないが、母集団の大きさ n を標本の大きさ r で割った n/r 以下の最大の整数 I を抽出間隔として、この I 以下の無作為に選んだ抽出スタート番号から、あとは等間隔 I で番号がなくなるまで選ぶ方法である。いずれにせよ、無作為に確率的に抽出する方法を用いると、母集団を構成する全要素について、それぞれが標本として抽出される確率が一定になるので、標本の性質から母集団の性質を客観的に評価することができるようになる。現在の統計学の推定、検定の理論はこの無作為抽出法を前提としている。

したがって、今日では、有意選出法は「科学的な」調査では本調査の前の予備的調査で利用する程度にすべきであるとされている。しかし

1. 母集団のリストが、完全な形では入手が不可能な場合
2. 無作為抽出法では回収率が著しく低下してしまい、非標本誤差が大きくなりすぎるような場合

には、母集団に関する予備知識を利用して割り当て法で標本抽出を行うことにも、それなりの意味があるとされている。特に、最近は、2 の回収率低下による非標本誤差の増大が、標本誤差に関する統計的推測を無意味なものにするほど深刻なものになっていることを忘れてはならない。

しかし、有意選出法を採用した場合、あるいは有意選出法によって得られたデータを見る場合には、表 1.6 にまとめられた特徴に注意しなくてはならない。すなわち、有意選出法では、無作為抽出法とは異なり、データは偏りをもち、しかも精度は高くてもあてにはならないということを認識する必要がある。そして、有意選出法で得られたデータについては、統計的推測は無力だということである。すなわち、無作為抽出法によって抽出された標本にのみ、統計的推測が可能なのである。それに対して、無作為抽出法では、偏りを回避することができるし、精度は標本の大きさを大きくすることで高められる。しかも、精度の評価を客観的に行うこともできるのである。

表 1.6　標本から母集団の特性を推論する場合の有意選出法と無作為抽出法の比較

	有意選出法	無作為抽出法
偏り	回避できない	回避できる
精度	意識的に粒のそろった要素を選んで標本にすることで高められる	標本を大きくすることで高められる
精度の評価	困難	確率論・統計学で可能

それでは、無作為抽出の場合には、どのようにしてこのような一連の統計的推測が可能なのであろうか。そのことをこれから第 5 節～第 7 節にかけて、確率、仮説検定、推定の順に見ていくことにしよう。ただし、無作為抽出法は、標本誤差に対しては、実に有効な抽出方法であるが、非標本誤差に対しては、なんら保証するものではない。このことは十分に認識しておく必要がある。

5. 無作為標本抽出と確率

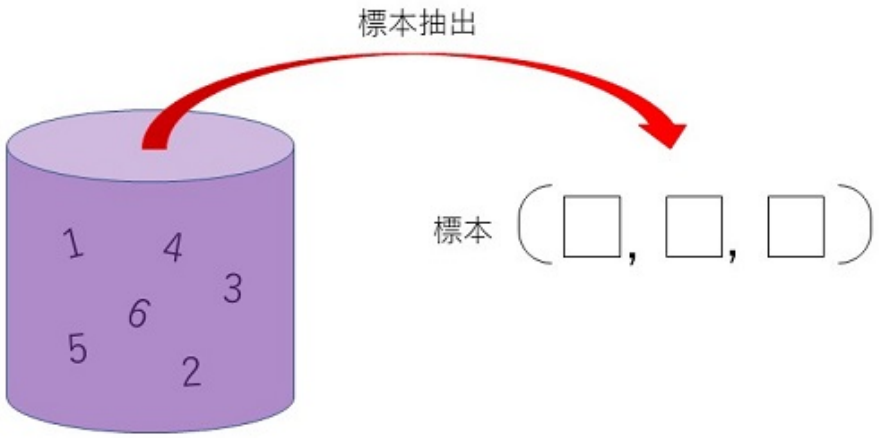
これ以降、確率の話が少々続くが、これは二つの目的のために必要なものである。一つは、通常、統計処理を行った際によく用いられる「有意水準」または「有意確率」の概念を正しく理解してもらうためである。もう一つは、標本の大きさのもつ意味を理解してもらうためである。これは、調査を企画する際に、適切に標本の大きさを決めるために重要になってくるだけではなく、既に行われてしまった調査の標本誤差を評価する際にも基本的な情報となるからである。しかし、全数調査を前提にする場合には、これ以降の節は必要がなくなる。

(1)標本空間

いま、大きさ n の母集団から標本として r 個の要素を非復元抽出(sampling without replacement)で無作為に抽出することを考えよう。つまり、r 個のものを取り出すときに、1 個ごとにいちいち元に戻さずに、続けて r 個を取り出すことを考える。この標本抽出によって得られる結果(outcome)、すなわち標本のことであるが、これは標本点と呼ばれる。あらゆる可能な標本点の集合は標本空間(sample space)と呼ばれて、Ω で表す。

例) いま母集団として、「応用統計学」を履修する学生の集合を考えてみよう。学生は履修番号で表されるものとする。「応用統計学」を履修する学生が 6 人だった($n=6$)とすると、母集団は集合｛1,2,3,4,5,6｝で表すことができる。この母集団から、図 1.4 のように、標本として 3 人の学生を抽出する場合を考えてみよう($r=3$)。この場合、標本は母集団から抽出された 3 人の学生の組、たとえば(1,2,3), (2,3,6)のように表される。

図 1.4　大きさ 6 の母集団から大きさ 3 の標本の抽出



このとき、標本点を書き出すと、標本空間 Ω は、次のような 20 個の標本点からなる集合になる。

Ω＝{(1,2,3), (1,2,4), (1,2,5), (1,2,6),
(1,3,4), (1,3,5), (1,3,6),
(1,4,5), (1,4,6),
(1,5,6),
(2,3,4), (2,3,5), (2,3,6),
(2,4,5), (2,4,6),
(2,5,6),
(3,4,5), (3,4,6),
(3,5,6),
(4,5,6)}

したがって、標本となる学生の組合せとしては、20 通りが考えられる。つまり大きさ 6 の母集団から大きさ 3 の標本を抽出するときの標本点の個数は 20 個というわけである。この場合、標本空間 Ω の要素をすべていちいち書き出さなくても、標本点の個数は、母集団の 6 個の要素から標本として 3 個の要素をとってくる組合せの数なので、次のようにも計算できる。

$${}_6C_3=(6\cdot5\cdot4)/(3\cdot2\cdot1)=20$$

この計算方法について概説しておこう。まず最初に、一般に、順列(permutation)とは n 個のものから r 個とって 1 列に並べたもののことであり、n 個のものから r 個とる順列の数は、次のように与えられる

$${}_nP_r=n(n-1)(n-2)\cdots(n-(r-1))=n!/(n-r)!$$

ここで、$n!=n(n-1)(n-2)\cdots3\cdot2\cdot1$ であり、これを n の階乗(factorial)とよぶ。$0!=1$ ということも約束しておこう。つまり、列の先頭に何がくるのかは n 通り考えられ、先頭の 1 個を決めれば、次の先頭から 2 番目には、先頭に定められた 2 個を除いた残りの $n-1$ 個の中からどれか一つが選ばれることになるので $n-1$ 通り……そして列の最後の 1 個は残っている $n-(r-1)$個の中から選ばれるので、$n-(r-1)$通り、それらをすべて掛け合わせるとこのような式になるのである。

さらに、組合せを考えることもできる。いま、とってきた r 個がどのような順序に並んでいても、組合せとしては同じものと考えられるので、${}_rP_r = r!$個の順列は一つの組合せとして数えられるべきものということになる。したがって、n 個のものから r 個とる組合せ(combination)の数は

$${}_nC_r = {}_nP_r / r! = n!/\{r!(n-r)!\}$$

となる。

(2)事象と確率

事象(event)とは、標本空間 Ω の部分集合のことであり、なんらかの意味で関心がもたれる結果の集合である。さきほどの 6 人の学生からなる母集団から 3 人の学生を抽出する例では、たとえば、次のような事象を考えることができる。

$A = \{(1,2,6), (1,3,5), (2,3,4)\}$

$B = \{(1,3,5)\}$

事象 A は学生の履修番号の和がちょうど 9 になる事象を表しており、事象 B は奇数の履修番号の学生のみであるという事象を表している。

標本抽出(より一般的には「試行」)を行うと、必ず標本空間 Ω の中のただ一つの標本点 ω が実現することになる。この実現した標本点 ω が事象 A に属するとき、この事象 A が生起したという。例えば、6 人の学生から 3 人を標本抽出すると、学生 1、学生 3、学生 5 が抽出されると、3 人の履修番号の和は 9 になり、事象 A が生起したことになる。実は事象 B も同時に生起している。

いま標本が無作為抽出される場合を考えよう。例えば、さいころを振って、出た目の履修番号の学生を標本として抽出するのである。このとき、標本抽出の結果は抽出過程の偶然性に影響されることになる。統計的状況で考えられる唯一の偶然的原因は、標本の母集団からのランダムな抽出だけである。このとき「事象 A の生起することの確からしさ」つまり、確率をはじめて考えることができる。一般に、生起する確率が考えられる事象を確率事象(probability event)という。$Pr(A)$で、事象 A の生起する確率を表すことにしよう。

それでは、標本が無作為抽出される場合に、どのように確率 $Pr(A)$を付与することができるのであろうか。いま n 個の標本点からなる標本空間 Ω を考える。無作為抽出によって標本を抽出する限り、n 個の標本点のどれもが、生起することに関しては「同程度に確からしい(equally likely)」。したがって、このとき、事象 A の生起する確率 $Pr(A)$は、事象 A に属する標本点の個数 k を $m(A)$で表すと、

$$Pr(A) = m(A)/m(\Omega) = k/n$$

と定義できる。さきほどの例では、

$m(\Omega) = 20$

$m(A) = m(\{(1,2,6), (1,3,5), (2,3,4)\}) = 3$

$m(B) = m(\{(1,3,5)\}) = 1$

なので、$Pr(A) = 3/20$、$Pr(B) = 1/20$ となる。こうして、この無作為抽出のときに定義される確率は、標本点の個数、つまり、起こり方の場合の数の数え上げに帰するわけである。

(3)確率変数

確率変数(random variable)とは標本空間 Ω の上で定義された実数値をとる関数 $X(\omega)$, $\omega \in \Omega$ のことである。標本抽出の結果として標本点 ω が定まると、実現値 $x = X(\omega)$が定まる。逆に、$\{\omega: X(\omega) = x\}$と集合を定義すると、これは標本空間の部分集合となる。もちろんこれは確率事象となるので、確率 $Pr(\{\omega: X(\omega) = x\})$ を与えることができる。一般に、この $Pr(\{\omega: X(\omega) = x\})$ を ω を省略して、$Pr(X = x)$と書く。つまり、確率変数は、それがとる各値

に対して、それぞれ確率が与えられている変数なのである。ちなみに、確率変数は X のようにローマ字の大文字で表す。

例の続き) さきほどの例と同様に、「応用統計学」を履修する学生の集合｛1,2,3,4,5,6｝を母集団とする。このうち、履修番号 1,2,3 の学生は「基礎統計学」の試験に合格し、単位を取得しているが、他の 3 人の学生は「基礎統計学」の単位を取得していない。そこでいま、3 人の学生を標本として抽出することを考え、確率変数 X でその標本中の「基礎統計学」の単位取得者数を表すものとする。標本の大きさは 3 だから、確率変数 $X(\omega)$は 0,1,2,3 のいずれかの値をとることになる。

$\{\omega: X(\omega)=3\}=\{(1,2,3)\}$
$\{\omega: X(\omega)=2\}=\{(1,2,4), (1,2,5), (1,2,6), (1,3,4), (1,3,5), (1,3,6), (2,3,4), (2,3,5), (2,3,6)\}$
$\{\omega: X(\omega)=1\}=\{(1,4,5), (1,4,6), (1,5,6), (2,4,5), (2,4,6), (2,5,6), (3,4,5), (3,4,6), (3,5,6)\}$
$\{\omega: X(\omega)=0\}=\{(4,5,6)\}$

無作為抽出によって標本を抽出する限り、20 個の標本点のどれが生起するのも同程度に確からしいので

$Pr(X=3)=Pr(\{\omega: X(\omega)=3\})=1/20$
$Pr(X=2)=Pr(\{\omega: X(\omega)=2\})=9/20$
$Pr(X=1)=Pr(\{\omega: X(\omega)=1\})=9/20$
$Pr(X=0)=Pr(\{\omega: X(\omega)=0\})=1/20$

である。ところで、標本中の「基礎統計学」の単位取得者の割合を P で表すと、P は

$P=X/3$

で定義され、0, 1/3, 2/3, 1 のいずれかの値をとる。すると、上の式から

$Pr(P=1)=Pr(X=3)=1/20$
$Pr(P=2/3)=Pr(X=2)=9/20$
$Pr(P=1/3)=Pr(X=1)=9/20$
$Pr(P=0)=Pr(X=0)=1/20$

となる。つまり、P も確率変数となる。より一般的な比率についての確率の計算については、付節を参照のこと。

6. 有意確率と仮説検定

例をさらに続けよう。「応用統計学」の授業を始めるに当って、授業のレベルを設定するために、「応用統計学」を履修している 6 人の学生の中から、さいころを使って、3 人を標本として抽出(非復元抽出)し、この 3 人について各々入念に口頭試験を行った。その結果、3 人とも「基礎統計学」終了時と同等以上の知識をもっていることがわかった。しかし、仮に、「基礎統計学」と同等以上の知識をもっている者が、「応用統計学」の履修者 6 人のうち、ちょうど半数の 3 人だけだったとしても、抽出のプロセスで、偶然、その 3 人を抽出してしまい、こうした調査結果が得られることもありうる。この場合、無作為抽出で標本抽出したので、各標本点のどれもが抽出されるのは同程度に確からしく、その確率は、容易に計算することができる。すなわち

$Pr(P=1)=Pr(X=3\})=1/20=0.05$

となる。したがって、仮に「基礎統計学」と同等以上の知識をもっている者が、「応用統計学」の履修者 6 人のうち、ちょうど半数の 3 人だけだったということが本当だとすると、標本全員が「基礎統計学」終了時と同等以上の知識をもっていたという確率は、わず

か 5%しかないということになる。この低確率では、母集団で 3 人だけが「基礎統計学」と同等以上の知識をもっていたとは考えにくいであろう。

このことを、仮説検定の言葉を使うと、次のように記述することができる。まず最初に、

- 仮説: 『母集団である「応用統計学」の履修者全体の 50%が「基礎統計学」と同等以上の知識をもっている。』

という仮説を立てる。ところが、この仮説の下で、標本調査の結果のように、$P=1$ すなわち、標本として抽出された 3 人がいずれも「基礎統計学」と同等以上の知識をもっている確率は 5%しかない。したがって、実際の標本調査の結果から、この仮説は棄却される。つまり、標本となった学生を調べた結果にもとづいて、母集団である「応用統計学」の履修者全体の過半数が「基礎統計学」と同等以上の知識をもっていると結論づけることができる。(ということは、「基礎統計学」の単位を取得していない学生の中にも、単位取得同等以上の知識をもっている者がいたということになる。結構なことである。)

ところで、母集団である「応用統計学」の履修者全体のうち、「基礎統計学」と同等以上の知識をもっている者の比率を π で表すと、さきほどの仮説は、

- $H_0: \pi=0.5$

とも書き表せる。統計学においては、仮説は前述のような文章形式よりも、式を使ったこの形式の方がよく用いられる。ちなみに、母集団の特性を表すにはギリシャ文字の小文字、標本の特性を表すにはローマ字を用いる。したがって、いまの例では、母集団、標本における「基礎統計学」と同等以上の知識をもっている者の比率はそれぞれ π、P と表されているわけである。

母集団では $\pi=0.5$ だったのに、偶然、標本では $P=1$ になってしまったという場合の「偶然」の部分を確率で表現したものを有意確率という。したがって、有意確率が大きいときには、標本での $P=1$ は「よくある偶然」と片付けてしまってかまわない。このような偶然は、標本抽出のプロセスで発生しうるもので、統計的にはよくあることで、意味のないものである。つまり、標本誤差の範囲内というわけである。しかし、先ほどのように、5%と有意確率が小さければ、標本での $P=1$ は「めったにない偶然」である。このときは、標本誤差では片付けられず、統計的にも無視できない意味をもつことになる。このことを「統計的に有意」あるいは単に「有意」(significant)という。この場合には、逆に、母集団で $\pi=0.5$ とした仮説の方を疑うべきであると結論づけるのが仮説検定の基本的考え方である。つまり、$\pi=0.5$ ではなくて、$\pi>0.5$ であろうと判断するのである。

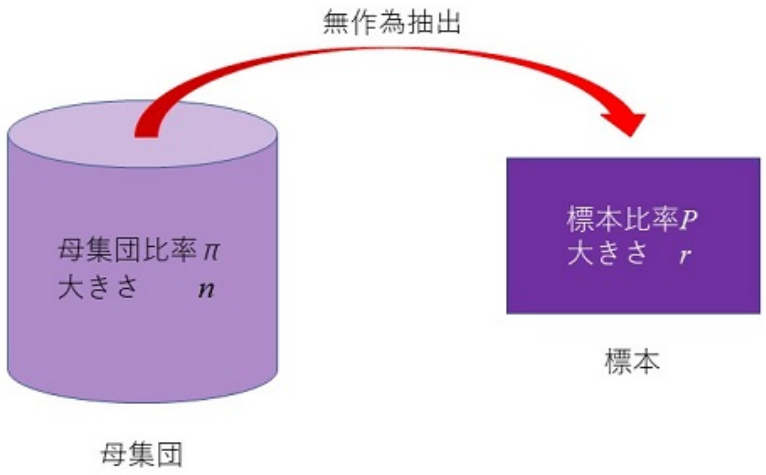
以上のことを整理すると、統計的仮説とは、母集団の特性値、たとえば、比率、あるいは、第 3 章以降で説明する平均、分散、相関係数などについて立てられる仮説のことである。そもそも仮説 H_0は「棄却する」(＝無に帰す)ために設定されたものであり、このことから帰無仮説(null hypothesis)とよばれる。このように仮説検定では、母集団の特性値に関する帰無仮説が真であるときに、それに対応する標本の値の生起する確率が、有意水準(level of significance; 5%がよく使われる)に満たないとき、それは帰無仮説が真でないと判断するに十分な証拠と考えるのである。(→演習問題 1.2)

7. 標本誤差と標本の大きさ

無作為抽出法をとるとき、標本の大きさ(sample size)は、標本誤差をどの程度に抑えたいかで決められるべきものである。標本調査の場合、誤差は、逆に精度という概念で考察される。標本調査において、そのようにして事前に決めた精度のことを目標精度と呼ぶ。

いま比率の推定を考えよう。大きさ n の母集団のうち A 群である母集団比率を π とする。この母集団から無作為抽出法によって、大きさ r の標本を抽出し、そのうち A 群である標本比率を P としよう。無作為抽出を行っているので、標本比率 P は確率変数となる。

図 1.5　母集団比率と標本比率



標本の大きさ r が、ほぼ 50 を超える大きさであるならば、中心極限定理によって、標本比率 P は次の期待値、分散をもつ正規分布にしたがう確率変数となることが知られている。

平均: $E(P)=\pi$

分散: $V(P)=\{(n-r)\pi(1-\pi)\}/\{(n-1)r\}$

したがって、一般に正規分布に従えば、

$$Pr(\{\pi-1.96(V(P))^{1/2}\leqq P\leqq\pi+1.96(V(P))^{1/2}\})=0.95$$

がいえるので(第 3 章第 7 節参照のこと)、この式を変形して

$$Pr(\{P-1.96(V(P))^{1/2}\leqq\pi\leqq P+1.96(V(P))^{1/2}\})=0.95$$

となる。つまり、母集団比率 π が標本比率 $P\pm1.96(V(P))^{1/2}$以内に収まる確率が 95%だということになる。いい換えれば、信頼度 95%で母集団比率を推定することができるのである。(このとき、区間 $[P-1.96(V(P))^{1/2}, P+1.96(V(P))^{1/2}]$ は信頼係数(confidence coefficient) 95%の P の信頼区間(confidence interval)とよばれる。)

このとき、推定の誤差の絶対値を

$$\varepsilon=1.96(V(P))^{1/2} \qquad (1.2)$$

とおくと、ε は絶対精度(absolute precision)とよばれるものになる。精度の高い調査あるいは推定とは、推定の幅の 1/2 に相当する ε の小さいものをいう(ただし、単に「精度」といった場合には、標本分散の逆数で定義されるので注意せよ)。このように、母集団の特性値がある一定の確率以上で収まる区間を求める推定法は、区間推定法(interval estimation)とよばれる。

標本の大きさ r が大きくなるほど $V(P)$そして ε は小さくなるので、必要な標本の大きさ s は、目標精度 d に対して、$\varepsilon\leqq d$ となるような最小の標本の大きさである。s は、(1.2)式を r

について解いた

$r=n$/分母

分母$=(\varepsilon/1.96)^2[(n-1)/\{\pi(1-\pi)\}]+1$　　(1.3)

で$\varepsilon=d$とおいて求めたr以上の最小の整数である。

一般に、母集団比率πは未知なので、$\pi=0.5$と仮定する。なぜならば、このとき$(V(P))^{1/2}$は最大になるので、$\pi=0.5$と仮定しておけば、誤差を最大に見積もったことになるからである。(1.3)式の中の$\pi(1-\pi)$も、$\pi=0.5$のとき最大となる。目標精度を$d=0.05$ (5%)程度にすることを考えると、必要な標本の大きさsは、(1.3)式を用いれば、表 1.7(a)のようになる。もし、目標精度を$d=0.025$ (2.5%)に上げようとすると、表 1.7(b)のようになる。

表 1.7　目標精度と標本の大きさ

(a)目標精度 $d=0.05$ (5%)のときに必要な標本の大きさ(近似値)

n	50	100	1,000	2,000	5,000	10,000	20,000	50,000	∞
s	45	80	278	323	357	370	377	382	384

(b)目標精度 $d=0.025$ (2.5%)のときに必要な標本の大きさ(近似値)

n	50	100	1,000	2,000	5,000	10,000	20,000	50,000	∞
s	49	94	607	870	1176	1332	1427	1492	1537

表 1.7 の(a)と(b)を比較すると、精度を上げてその幅を 1/2 にするのには、標本の大きさを約 4 倍にしなければならないことがわかる。一般に、目標精度dを$1/k$倍にしようとすると、必要な標本の大きさは約k^2倍になる。なぜなら(1.3)式の分母にε^2があるからである。また調査の精度は、母集団の大小にほとんど関係なく、標本の大きさの平方根にほぼ反比例する。したがって 50 人程度の母集団では、標本調査といっても、ほとんど全数調査が必要になる一方で、どんなに母集団が大きくても、目標精度が 5%ならば 400 人程度、目標精度が 2.5%でも 1600 人程度の大きさの標本であれば、十分に目標を達成する。

ところで、標本の大きさが 400 程度あれば、絶対精度を 5%以下に抑えられるということは、逆にいえば、たとえば標本比率が 56%であれば、母集団比率πが 50%である確率は 5%(＝1－信頼係数)以下ということである。つまり、H_0: $\pi=0.5$ という仮説は、仮説検定を行えば棄却されることになる。いい方を変えれば、5%程度の比率の差が統計的に有意になり、意味をもつように、標本の大きさを決めてやっていることになる。このように、標本の大きさが十分に大きくて、精度が十分に向上してくれば、もはやいちいち仮説検定に気を回す必要はない。たしかに、小さい標本のデータを与えられているときには、仮説検定で母集団についての知識を検討・吟味する必要がある。しかし、十分な大きさの標本を確保することができたならば、仮説検定が実際上必要ないほど、推定の精度を確保できたことになるのである。

これまでの議論を、実際に調査を設計し、管理するという観点からまとめると、次のようになる。調査にともなう誤差を考え、標本誤差(絶対誤差)と非標本誤差をともに 5%以内に抑えるということをめざすのであれば、標本の大きさとしては 500 程度確保し(したがって、数百程度の大きさの母集団でほとんど全数調査が必要になる)、回収率 80%以上をめざして、きめの細かい管理をすることを考えるべきである。

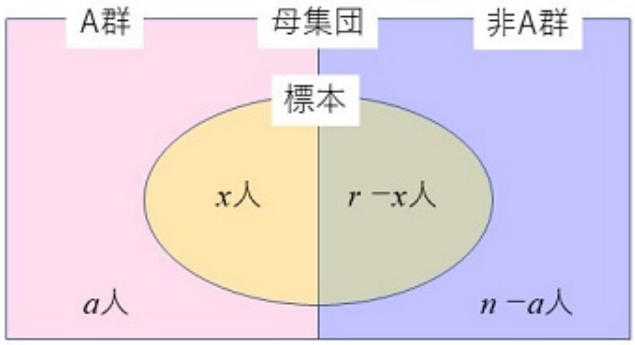
付. 標本比率の分布

比率についての有意確率の計算は、一般的にはどのように行われるのであろうか。いま、n 個の要素からなる母集団を考える。この母集団の要素は、A 群、非 A 群の 2 群に分類され、各要素がどちらの群に属するかは簡単に区別できるとする。

(1)有限母集団の比率: 超幾何分布

この母集団から非復元抽出(sampling without replacement)で、r 個を標本として抽出することを考えよう。つまり、r 個のものを取り出すときに、1 個ごとにいちいち元に戻さずに、続けて r 個を取り出すことを考える。

図 1.6　有限母集団からの抽出



確率変数 X で大きさ r の標本のうち、A 群に属する要素の数を表すものとする。すると

1. 母集団の n 個の要素から標本となる r 個の要素を選ぶ方法は、${}_nC_r$通りで、これは標本空間 Ω の要素の数である。
2. 母集団の A 群の a 個の要素の中から x 個を選ぶ方法は、${}_aC_x$ 通り
3. 母集団の非 A 群の $n-a$ 個の要素の中から残りの $r-x$ 個を選ぶ方法は、${}_{n-a}C_{r-x}$ 通り

したがって、

$$Pr(X=x)=h(x)={}_aC_x \cdot {}_{n-a}C_{r-x} / {}_nC_r$$

これを超幾何分布(hypergeometric distribution)とよぶ。A 群が母集団に占める比率、母集団比率 $\pi=a/n$ を使うと

$$Pr(X=x)=h(x)={}_{n\pi}C_x \cdot {}_{n(1-\pi)}C_{r-x} / {}_nC_r \qquad (1.4)$$

と表すこともできる。さらに、標本として抽出されたもののうち A 群の占める比率、標本比率 $P=X/r$ を用いると

$$Pr(P=p)=h(p)={}_{n\pi}C_{rp} \cdot {}_{n(1-\pi)}C_{r(1-p)} / {}_nC_r \qquad (1.5)$$

と表すこともできる。第 5 節、第 6 節の例の場合には $\pi=0.5$ と仮定しているので、たとえば

$$Pr(P=1)=h(1)={}_3C_3 \cdot {}_3C_0 / {}_6C_3=1/20$$

となる。

ところで、標本でも A 群の比率 $P=x/r$ は、母集団で A 群の比率が $\pi=a/n$ に等しくなりそうなものだが、無作為抽出をすれば、偶然のいたずらで、P は 0〜1 までの値をとりうる可能性がある(ただし、標本の大きさに比べ、母集団が十分に大きいとき)。しかし、直感もまた正しく、確率変数である標本比率 P のとりうる値を、確率をウェイトとして加重平均した値(これを期待値という)は母集団比率 π に等しくなる。実際、(1.5)式で表した場合の超

幾何分布の平均(＝期待値)、分散は次のようになる。

平均: $E(P)=\Sigma_p\, p\cdot h(p)=\pi$

分散: $V(P)=\{(n-r)\pi(1-\pi)\}/\{(n-1)r\}$

このように、標本比率の期待値が母集団比率と一致する場合、標本比率は偏りをもたない(＝不偏)といわれる。

(2)無限母集団の比率: 二項分布

これまでは、n 個の要素からなる有限母集団から非復元抽出を行うことを考えてきた。いまもし、n が非常に大きいとき、極端には無限に大きいと考えられるときはどうなるであろうか。(1.4)式から

$$h(x)=[(n\pi)(n\pi\text{-}1)\cdots(n\pi\text{-}(x\text{-}1))/x!]\times[(n(1\text{-}\pi))(n(1\text{-}\pi)\text{-}1)\cdots(n(1\text{-}\pi)\text{-}(r\text{-}x\text{-}1))/(r\text{-}x)!]\times[r!/\{n(n\text{-}1)\cdots(n\text{-}(r\text{-}1))\}]$$

となり、この式の分母分子を n^r で割ると

$$h(x)=[r!/x!(r\text{-}x)!]\times[\{\pi(\pi\text{-}1/n)\cdots(\pi\text{-}(x\text{-}1)/n)\cdot(1\text{-}\pi)((1\text{-}\pi)\text{-}1/n)\cdots((1\text{-}\pi)\text{-}(r\text{-}x\text{-}1)/n)\}/\{1\cdot(1\text{-}1/n)\cdots(1\text{-}(r\text{-}1)/n)\}]$$

となる。したがって、超幾何分布は $n\to\infty$のとき

$$\lim_{n\to\infty} h(x)={}_rC_x\pi^x(1\text{-}\pi)^{r\text{-}x}=b(x)$$

となる。この分布 $b(x)$は二項分布(binomial distribution)とよばれる。この分布の名前のいわれは、二項式

$$(\pi+(1\text{-}\pi))^r$$

を展開したとき、$\pi^x(1\text{-}\pi)^{r\text{-}x}$の項の係数が ${}_rC_x$になっていることによる。以上のことから、母集団の大きさ n が大きいときには、超幾何分布を二項分布で近似することができる。

二項分布は、無限母集団からの抽出だけではなく、たとえ有限母集団であっても復元抽出(sampling with replacement)、つまり 1 個取り出すごとにいちいち元に戻しながら r 個の標本を取り出すときにも得られる。なぜならば、どちらの場合にも、母集団から 1 個要素を取り出したときに、それが A 群である確率は π、非 A 群である確率は 1-π で、それは抽出回数にかかわらず一定だからである。したがって、A 群が x 個、非 A 群が r-x 個である列、たとえば

AA(非 A)A……(非 A)A

が得られる確率は

$$\pi\pi(1\text{-}\pi)\pi\ldots\ldots(1\text{-}\pi)\pi=\pi^x(1\text{-}\pi)^{r\text{-}x}$$

である。この式からもわかるように、A 群と非 A 群の要素の並ぶ順序には関係なく、A群が x 個、非 A 群が r-x 個である標本点は、同じ確率をもつことになる。したがって、これと同じ確率をもつ標本点は、標本空間の中に ${}_rC_x$個あるので、結果として

$${}_rC_x\pi^x(1\text{-}\pi)^{r\text{-}x}=b(x)$$

という二項分布が得られるわけである。

演習問題

1.1 調査の種類　第 3 節(1)で説明した分類にしたがって、本書で統計調査の実例として取り上げている「組織活性化のための従業員意識調査」の種類を分類してみよ。必要であれば、第 6 章も読んでみること。

1.2 仮説検証　第 6 節の例で、もし口頭試験の結果、標本 3 人のうち 2 人だけが「基礎統計学」終了時と同等以上の知識をもっていることがわかったならば、そのとき仮説 $H_0: \pi = 0.5$ の検定を行なってみよ。

1.3 報道された調査結果の評価　最近、新聞で報道された調査結果を二つ以上探し、その調査結果について、(1)母集団の定義、(2)全数調査か標本調査か、(3)標本調査の場合には標本抽出の方法、(4)調査時点、(5)調査方法、(6)回収率を調べよ。

1.4 調査の企画　「新人類世代の意識を探る」というテーマで、従業員規模(正社員のみ)１万人程度の企業の意識調査をしたい。どのような調査をすべきか企画を立てよ。

第 2 章　SAS 入門: 単純集計

章目次

1. はじめに
2. PC 版 SAS のための MS-DOS 入門
3. SAS の基本的な使用方法
4. SAS プログラムの基本
5. 基本型 SAS プログラム
6. DATA ステップでのデータの加工
付．CMS 版 SAS リリース 5.18 の基本的な使用方法
演習問題

1. はじめに

第 1 章では統計学および調査の考え方を概観したので、この章からはさっそく「組織活性化のための従業員意識調査」によって得られたデータを素材にして、コンピュータを使って統計処理することを始めよう。そのために、この章ではまず統計パッケージ SAS の基本的な使い方を、最初にすべき基本的統計処理である単純集計を例に説明しておくことにする。

SAS はデータに対して、統計処理・演算処理を始めとする各種の加工処理を行うパッケージ・プログラムである。もしわれわれが各種の統計処理用にそれぞれのプログラムを自分で作成・開発・維持しようとすると、これは非常に大変な作業となる。その点 SAS は、既に完成されている統計用プログラムの集合体、システムであり、利用者が必要なプログラムを指定すると、それらを機能的にデータに結び付けて、統計処理をしてくれるのである。

この章では、最新版である SAS 第 6 版の使い方とプログラミングを SAS やコンピュータを全く知らない読者でも理解できるように解説することにしよう。もちろん、単なる SAS 入門ではなく、「組織活性化のための従業員意識調査」のプロセスの 1 ステップとして、調査によって得られたデータを単純集計し、後の章での統計分析のためにデータをデータ・ベース化しておくのが、この章の本来の役割である。

本書ではパーソナル・コンピュータ版 SAS、いわゆる PC 版 SAS とメインフレーム(mainframe)版 SAS の両方について、いわば両刀遣い的に相違点などにも言及しながら説明していく。これは、今後筆者も含めて多くの人が、SAS を使う際、処理時間やコスト、利用可能性等を考慮した上で、パーソナル・コンピュータとメインフレームを使い分け、同じプログラム、データをフロッピィ・ディスク 1 枚で移植し、どちらの機械でも使うことになっていくと考えるからである。とはいうものの、どちらかといえば、PC 版 SAS を中心にして、その使い方を操作の順序にしたがって説明する。より具体的には、MS-DOS を搭載したパーソナル・コンピュータに、既に SAS の第 6 版(リリース 6.04)がインストールされていることを前提にして、話を進めることにしよう。

メインフレーム利用者に対しては、SAS 第 6 版が IBM 及び IBM 互換の国産のメインフレームでも利用可能なので、メインフレーム版 SAS として IBM メインフレームの代表的な OS である CMS に対応している CMS 版 SAS (SAS リリース 6.06)についても、PC 版 SAS

との相違点などを指摘しながら併せて説明していく。また、現在でも、メインフレームの一部では、CMS 版 SAS リリース 5.18 がそのまま使用されているようなので、付節を設け、CMS 版 SAS リリース 5.18 の基本的な使用方法について、第 6 版との違いを概説しておいた。

もっとも、SAS 第 6 版では、メインフレーム用もパーソナル・コンピュータ用もともに C 言語で書かれ、使用方法は基本的に同じである。ファイル名の形式と画面のウィンドウの数が異なることに注意すれば、PC 版 SAS も CMS 版 SAS も同様に使用することができる。しかも、CMS 版 SAS は、他のメインフレーム用 SAS と比較しても、ハードウェアやソフトウェアについての余計な知識や配慮を必要としないという点では、パーソナル・コンピュータ並の使いやすさをもったシステムである。例えば、通常、メインフレームでは SAS のようなパッケージ・ソフトであっても、実際に動くようにするためには、ハードウェアの仕様、ソフトウェアの仕様、さらに動かそうとしているプログラムの仕様やデータのサイズなども配慮した上で、さまざまな環境設定を行う必要がある。この設定を行うのが JCL (job control language、「ジョブコン」と言うこともある)と呼ばれるものであるが、このように、ソフトの使用に際して、利用者がいつも頭を悩まされてきた JCL についても、CMS 版 SAS では一切気にする必要はない。CMS では、ログオン時に PROFILE という名前の初期設定ファイルが自動的に起動され、利用者の環境設定が自動的に行われてしまうのである。

それでは、PC 版 SAS を使用する上で必要になる最小限の MS-DOS の基礎知識についての説明から始めることにしよう。IBM メインフレームと CMS については、付章を参照してほしい。CMS 版 SAS の利用者は、次の第 2 節の代わりに、その付章を読んでから、第 3 節に進むこと。メインフレームを CMS 以外の OS で使用する場合には、センター等に問い合わせて、SAS 利用にどのような JCL が必要になるのかをあらかじめ確認してほしい。

2. PC 版 SAS のための MS-DOS 入門

(1) OS

OS (Operating System)とは、簡単に言えば、外部記憶装置や入出力装置といったハードウェアを操作するソフトウェアのことである。利用者、あるいは利用者が直接触れる高級言語(例えば C 言語)、パッケージ(例えば SAS)が OS の使い方を分かっていれば、外部記憶装置や入出力装置といったハードウェアを物理的に直接操作する必要はなく、OS というソフトウェア上の論理的装置(logical device)の簡単な操作をするだけで、実際の外部記憶装置や入出力装置の複雑な操作は OS が代わりにやってくれる。

もし我々が複雑なコンピュータ・システムの全体を構成するそれぞれの外部記憶装置や入出力装置といったハードウェアを物理的に直接操作しようとするならば、各装置の機械的な仕組みや原理を理解し、なおかつどのように操作するのがもっとも効率的で間違いがないのかを習得しておかなくてはならない。考えただけで気の遠くなるような話である。パーソナル・コンピュータですら、MS-DOS のような OS なしでハードウェアを物理的に直接操作することは不可能である。それだけ OS は便利であり、かつコンピュータ・システムの使い勝手、利用効率を決めるという点で決定的に重要なものである。

そこでここでは、パーソナル・コンピュータのハードウェアについては触れず、OS である MS-DOS について、SAS 利用のために最小限必要なことがらについて説明しておこう。

(2)ファイル

ファイル(file)またはデータ・セット(data set)とは、物理的にはコンピュータ・システムの固定ディスクやフロッピィ・ディスクに格納されたプログラムやデータの集合のことである。より具体的には、パーソナル・コンピュータやメインフレームの端末のディスプレイ画面の上に入力、表示された行の集合をディスクに格納したものがファイルである。ディスクに保存された情報は、このファイルという単位で管理される。(詳細は付章第 1 節を参照されたい。)

ファイルにはそれぞれ名前が付けられ、他のファイルと識別される。MS-DOS のファイル名は、例えば次のような形をしている。

FRQ.LOG
JPCDATA.SAS

ファイル名は、一般的には

ファイル名本体.ファイル名拡張子

という形式をとっている。ファイル名の本体と拡張子の間には必ずピリオド "." を挟むことになっている。MS-DOS のファイル名の拡張子は必ず付けなくてはならないというものではないが、ファイルの内容の分類のためには付けておくと便利である。拡張子を付けない場合にはピリオド "." は付けない。

ファイル名の具体的な表記方法としては、ファイル名本体は 8 文字以内、拡張子は 3 文字以内となっており、英数字 A～Z、0～9、および$、#、@、-(ハイフン)、_(下線)の記号類からなる文字列である。このほかにも使える記号もあるし、ひらかな、カタカナ、漢字も使えることになっているが、メインフレームで使用する場合もあることを考えると、トラブルの原因になるので、あまり勧められない。その意味では、ファイル名の 1 文字目は英字にしておいた方が無難であろう。また、英字に大文字、小文字の区別はなく、MS-DOS では、すべて大文字に変換されて処理される。

新たにファイル名を作る場合に、もし、既に存在しているファイルと同じファイル名を付けようとすると、元からあるファイルの内容が失われたり、エラーが出たりすることになるので、注意が必要である。

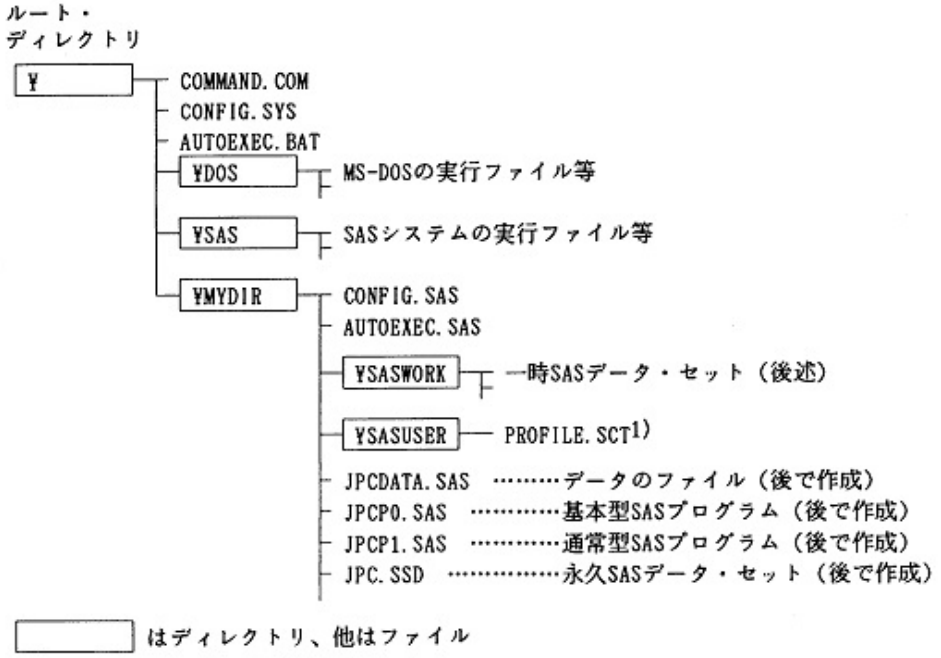
(3)ディレクトリ

固定ディスクや取り替え可能なフロッピィ・ディスクといった「1 枚」のディスクの中には、複数のファイルを格納することができる。そこで、MS-DOS ではディスクの中のどこにどんなファイルが格納されているか、ディレクトリ(directory=住所録)を作って管理している。ただし、詳しい「住所」は MS-DOS のシステムにとっては必要な情報だが、利用者にとっては必要ないので、利用者には表示されない。このディレクトリの中には、ファイルだけではなく、別のディレクトリも入れることができる。元のディレクトリを親とすれば、その一覧表の中に表記され子の関係にあるディレクトリをサブディレクトリと呼ぶ。そのサブディレクトリもそれ自体が親ディレクトリとなって、サブディレクトリをもつことができる。このようにして、ちょうど家系図のように階層ディレクトリ構造が形成されるのである。つまり、複数のディレクトリが、それぞれ他のディレクトリとの親子関係を結んで 1 枚の家系図に収まることで「1 枚の」ディスク上に存在しうるわけである。

例えば、本書では SAS の『導入及び運用と保守の手引き』通りのディレクトリ構造を前提にしているが、その構成は図 2.1 のようになる。固定ディスク(ドライブ A)の内容は SAS の稼働に必要な最小限のファイル構成になっている。SAS の『導入及び運用と保守の手引き』にしたがって、SAS の実行ファイルなどを納めた SAS ディレクトリからユーザー・プログラム、データなどを納めたディレクトリを区別、分離して、「￥MYDIR」という私用

ディレクトリを作成している。これは、SAS のバージョンアップの際の作業を楽にし、複数の人が同じ固定ディスクを使用する際に独立性を確保することなどを目的にしている。

図 2.1 固定ディスクのディレクトリの構成図



1) ディスプレイ・マネージャ(後述)の各種設定保存ファイルだが、本書では出荷時の設定を前提にしているので、これ以上触れない。

しかし、利用者が一度に使用できるディレクトリは一つと決められている。現在使用中のディレクトリをカレント・ディレクトリと呼ぶが、現在どのディレクトリを使用しているのか、カレント・ディレクトリを確かめるには、chdir コマンドを使えばよい。

A>ChDir

と入力すると、カレント・ディレクトリが表示される。例えば、表示が

A:¥

ならば、ルート・ディレクトリがカレント・ディレクトリであり、もし表示が

A:¥MYDIR

ならば、MYDIR がカレント・ディレクトリである。

カレント・ディレクトリは変更することもできる。カレント・ディレクトリから 1 階層上の親ディレクトリにカレント・ディレクトリを変えるときには、やはり chdir コマンドを使い、次のように入力する。

A>ChDir ..

逆に、カレント・ディレクトリのサブディレクトリの一つにカレント・ディレクトリを変えるときには、

A>ChDir サブディレクトリ名

と入力すればよい。例えば、いまルート・ディレクトリを使用しているとすると、

A>ChDir MYDIR

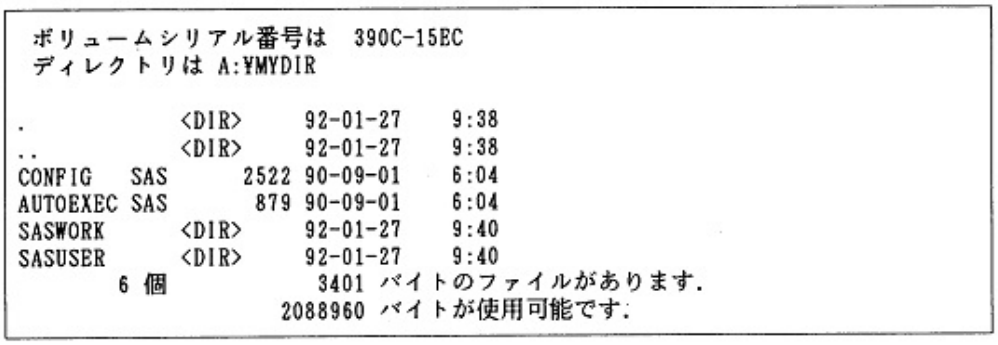
と入力すると、サブディレクトリ MYDIR に移ることができる。

カレント・ディレクトリの中に、どんなファイルやサブディレクトリが存在しているか、その一覧表を表示させるときには、

A>DIR

と入力すればよい。するとカレント・ディレクトリが MYDIR のときは、SAS インストール直後では、本章の中でこれから作成する SAS プログラム等はまだ存在しないので、例えば図 2.2 のようになる。

図 2.2 MYDIR のディレクトリ画面の例(SAS インストール直後)



(4)コマンド入力の表記方法

既に(3)で行っているようなキーボードからのコマンド入力の説明の際には、本書では次のような表記方法をとっている。これを読んだ後で、もう一度(3)を見直してみるとよい。

1. 下線部分はキーボードを使って入力する文字や記号の列、すなわち文字列を表している。表記通りにキー入力し終ったことをディスプレイ画面上で確認したら、最後にリターン・キーを押す。このリターン・キーでディスプレイ画面上の文字列がコマンドとして MS-DOS や SAS のシステム側に送られることになる。リターン・キーを押すことはコマンド入力には当然なので、表記上省略するが、コマンド入力には常に必要となることを忘れないでほしい。
2. 例えば

   A>ChDir ..

   と表記している場合には、英小文字の部分は省略することができ、

   A>CD ..

   と入力しても同じ機能を果たす。
3. 表記上 1 文字分を空けている所は、実際の入力の際にも表記通りスペース・キーを押して、空白(ブランク)を入れること。
4. 大括弧[ ]を用いている場合には、その中の文字列を入力してもよいし、または入力を省略してもよいことを表している。省略した場合には、システムの側で、あらかじめ決められたパラメータ値(これを既定値という)が指定されたものと仮定して処理が行われる。したがって、当然、[ ]を入力した場合と省略した場合とでは機能が異なってくる。

3. SAS の基本的な使用方法

(1) SAS の起動

SAS 起動の仕方は、PC 版 SAS も CMS 版 SAS も基本的に同じである。MS-DOS あるいは CMS のコマンドを入力できるモードから、SAS とコマンド入力の形で入力してやればよいのである。ただし、これは標準的にインストールした場合であり、SAS のインストールの仕方が違ってくれば、もちろん起動の仕方も当然違ってくる。特にメインフレームで

は、管理しているセンターごとにかなり個性があるので、ホスト計算機の管理者に必ず確認すること。実は、SAS の使用方法のうち、ハードウェア、ソフトウェア、管理運営システムで差異が出るのは、この SAS 起動の部分だけである。SAS を一旦起動してしまいさえすれば、こうした動作環境とは一切関係なく、どの動作環境でも同じ使用方法で SAS を使うことができる。ここでは PC 版 SAS を標準的にインストールした場合の起動の方法を代表例として説明しておこう。

それでは、さっそく SAS を起動してみることにしよう。まずパーソナル・コンピュータの電源スイッチである POWER スイッチを入れて、MS-DOS を起動させる。ハード・ディスク・ドライブが A であるとき、プロンプトが「A>」となっていること、そして、私用ディレクトリ MYDIR がカレント・ディレクトリになっていることを確認した上で、

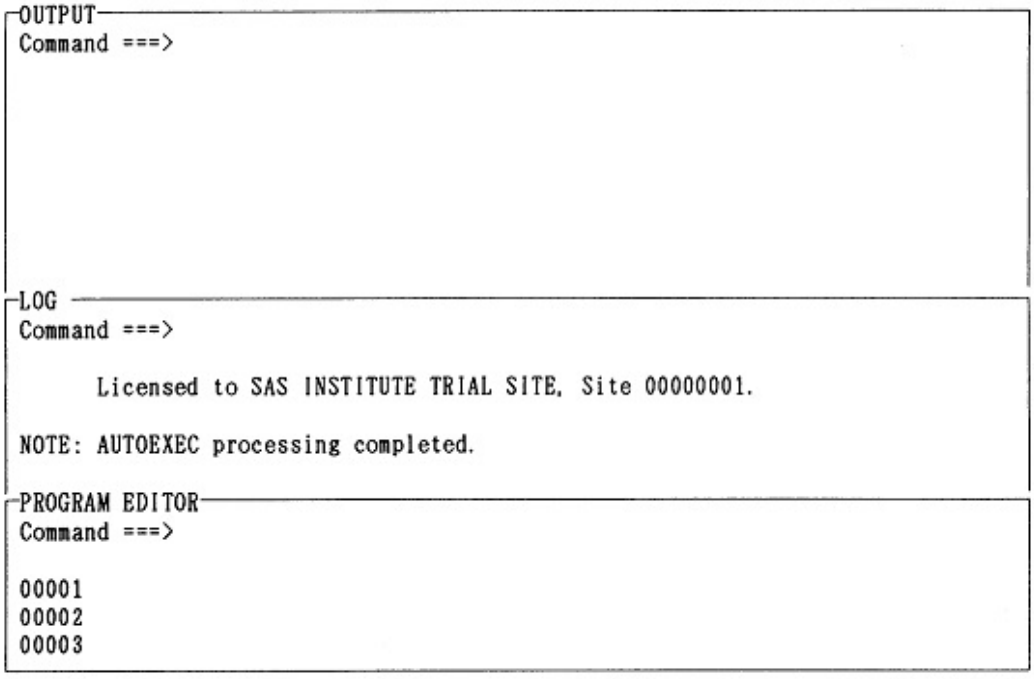
A>SAS

と入力すると SAS が起動される。

(2)SAS のディスプレイ・マネージャ・システムの画面

SAS の起動によって、プログラム作成、実行のための SAS ディスプレイ・マネージャ・システム(SAS Display Manager System)の画面が表示されることになる。この画面は、PC 版 SAS では図 2.3 のようになる。

図 2.3 PC 版 SAS のディスプレイ・マネージャ・システムの画面(SAS 起動時)



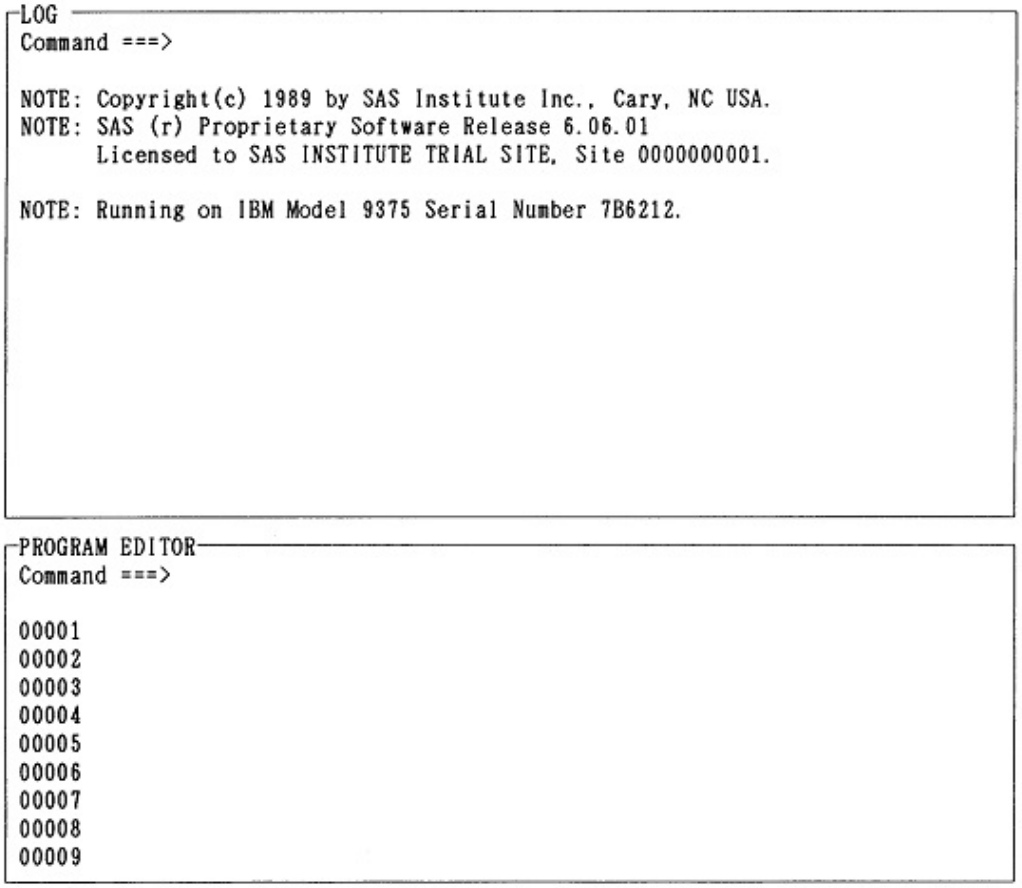
PC 版 SAS のディスプレイ・マネージャ・システムの画面は図 2.3 を見てわかるように 3 つのウィンドウに分割されている。各ウィンドウの名称と役割は、SAS プログラムの編集、実行の手順にしたがえば、下から順に、

1. PROGRAM EDITOR ウィンドウ: SAS プログラムの編集を行うための編集画面で、フル・スクリーン・エディター((4)で後述)となっている。

2. LOG ウィンドウ: SAS プログラムの実行結果(出力ではない)を表示する画面。例えば、プログラムの指示に基づいて SAS の行った処理内容や、その過程で発見されたエラーの種類などについて表示するための画面。
3. OUTPUT ウィンドウ: 実行された SAS プログラムの指示にしたがって行われた出力を表示するための画面。

CMS 版 SAS のディスプレイ・マネージャ・システムの画面は図 2.4 に示されているが、PC 版 SAS と異なり、二つのウィンドウに分割されている。OUTPUT ウィンドウは SAS 起動時には表示されない。もっとも、SAS 起動時には OUTPUT ウィンドウは表示すべき内容もないし、使う用事もないので、そのことで不便はない。

図 2.4　CMS 版 SAS のディスプレイ・マネージャ・システムの画面(SAS 起動時)



PC 版 SAS であれ、CMS 版 SAS であれ、SAS のディスプレイ・マネージャ・システムの画面のファンクション・キーには、利用者が比較的よく使用するコマンドが設定されている。ファンクション・キーに設定されているコマンドは、よく使うコマンドなので、コマンドとしても、その機能をよく理解しておこう。ファンクション・キーの設定は利用者の手で変えることもできるが、表 2.1 は PC 版 SAS での出荷時の設定である。この章では、表 2.1 の設定を前提にして話を進める。また、各ウィンドウでは、表 2.2 のような編集キーもそのまま使える。

表 2.1　ディスプレイ・マネージャ・システムの画面のファンクション・キー
(PC 版 SAS 出荷時)

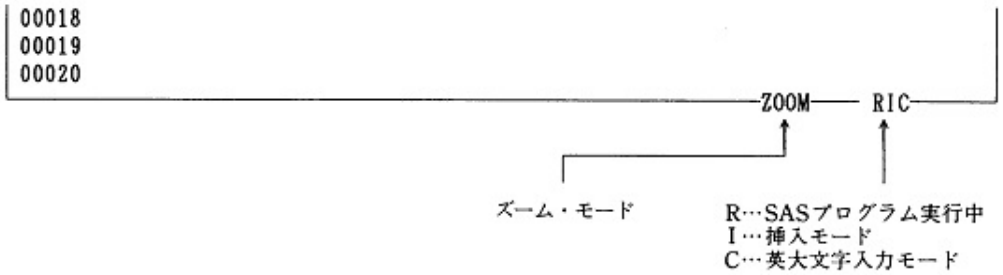
キー	設定コマンド	機能
(f･1)	help	コマンド、機能について説明する HELP ウィンドウを表示し、そのメイン・メニューを表示する。
(f･2)	keys	ファンクション・キーの設定を示す KEYS ウィンドウを表示する。
(f･3)	log	LOG ウィンドウを表示する。既に表示されている場合には、カーソルが LOG ウィンドウに移動する。
(f･4)	output	OUTPUT ウィンドウを表示する。既に表示されている場合には、カーソルが OUTPUT ウィンドウに移動する。
(f･5)	next	ウィンドウを順番にカーソルが移動する。
(f･6)	prg	PROGRAM EDITOR ウィンドウを表示する。既に表示されている場合には、カーソルが PROGRAM EDITOR ウィンドウに移動する。
(f･7)	zoom	カーソルのあるウィンドウだけを画面一杯に表示するズーム・モードにするか、通常のマルチ・ウィンドウ・モードにするかを切り替える。
(f･8)	subtop	PROGRAM EDITOR ウィンドウの第 1 行の SAS 文だけを実行させる。このことでその第 1 行は消え、第 2 行以降が 1 行ずつ繰り上がることになる。
(f･9)	recall	PROGRAM EDITOR ウィンドウにカーソルがあるときに有効で、直前に実行したプログラムを PROGRAM EDITOR ウィンドウ上に呼び出す。既に PROGRAM EDITOR ウィンドウに内容があるときには、その内容の始めに挿入される。recall をさらに実行すると、前々回、前々々回、……と次々と遡って始めに挿入される。
(f･10)	zoom off; submit	マルチ・ウィンドウ状態にしてから、PROGRAM EDITOR ウィンドウ上のプログラムを実行する。

表 2.2　編集キー(PC 版 SAS)

キー	機能
(INS)	挿入モードと重書きモードを切り替える。挿入モードのときは、画面右下端に "I" の文字が表示される。
(DEL)	カーソル上の文字を削除する。
(BS)	カーソルの 1 字前の文字を削除する。
(ROLL UP)	画面を上方にスクロールし、次画面の内容を表示する。
(ROLL DOWN)	画面を下方にスクロールし、前画面の内容を表示する。
(CAPS)	アルファベットの大文字入力モードと小文字入力モードを切り替える。大文字入力モードのときは、画面右下端に"C"の文字が表示される。

ズーム・モード、挿入モード、英大文字入力モードといった、それぞれ対応するキーによってスイッチのように on/off を繰り返すものについては、ディスプレイ・マネージャ・システム画面の右下端のウィンドウの枠上に、図 2.5 のように、現在のモードの状態が表示される。何も表示されていないときには、いずれの状態でもないということを意味することになる。

図 2.5　SAS の状態表示(ディスプレイ・マネージャ・システム画面の右下端に表示)



ファンクション・キーの設定は通常は画面には表示されないので、よく使用するものについては覚えておけば便利である。理由はよくわからないが、CMS 版 SAS では、同じホスト計算機に接続された端末でも、例えば、専用端末とパソコン端末のように、端末の機種によって出荷時の設定も異なっているようなので、CMS 版 SAS のファンクション・キーの設定は、ここで一覧表にすることはしない。使用する際に、自分でまず Command 行に

Command ===> KEYS

と入力して、確認してみること。少なくとも、表 2.1 にあるコマンドに対応するファンクション・キーについては、使い始めのときに自分でチェックしておこう。特に重要なのは、CMS 版 SAS では、編集キーの(ROLL UP)キー、(ROLL DOWN)キーが使えないので、これらの編集キーと同じ機能を果たす forward コマンド、backward コマンドの設定されているファンクション・キーを探して確認しておくことである。

PC 版 SAS では、(f･2)キーに keys コマンドが設定されているので、わざわざ Command 行に入力しなくても、(f･2)キー(=keys)を押せばよい。試しに、(f･2)キーを押してみて、KEYS ウィンドウが表示されるのを確認してみよう。KEYS ウィンドウを閉じるためには、KEYS ウィンドウの Command 行に

Command ===> END

と入力すればよい。HELP ウィンドウについても同様に、end コマンドによってウィンドウを閉じることができるので、(f･1)キー(=help)を押してみて、HELP ウィンドウを表示させてみてから試してみよう。

カーソル・キーである(→)(←)(↑)(↓)によって、カーソルはディスプレイ・マネージャ・システムの画面上を、ウィンドウの境界線に関係なく自由に移動させることができるが、(f･5)キー(=next)を押すことで、ウィンドウ間を順番にカーソル移動させることもできる。また、(f･3)キー(=log)、(f･4)キー(=output)、(f･6)キー(=prg)によって、狙ったウィンドウに、一発でカーソルを移動させることもできるので、試してみよう。

もっとも、実際には、SAS プログラムを作り始めるときは、PROGRAM EDITOR ウィンドウしか使わないので、(f･7)キー(=zoom)を使って、PROGRAM EDITOR ウィンドウだけを画面一杯に表示するズーミングを行い、十分広くとった方が作業しやすい。この zoom コマンドは表 2.1 にもあるように、on/off を繰り返すスイッチのように機能するので、通常のマルチ・ウィンドウ・モードの状態に戻すには、もう一度(f･7)キー(=zoom)を押せばよい。

(3)ウィンドウの内容のファイルへの保存と読み込み

各ウィンドウの内容は、SAS プログラムでも、データでも、出力結果でも、そしてログでも、いずれも同様に、ただそれらが存在している各ウィンドウの Command 行に

Command ===> FILE 'ファイル名'

と入力すれば、ファイルに保存することができる。この file コマンドを使用するときには、ファイル名は必ず、' 'で囲むようにする。

PC 版 SAS では、MS-DOS のファイル名が

[ドライブ名:][￥パス名￥]ファイル名[.拡張子]

のように指定される。ここでパス名とは、ルート・ディレクトリから目的のディレクトリに達するまでのディレクトリの道順(パス)を示すために、各ディレクトリ名を￥で区切って並べて表したものである。例えば、次のようにである。

Command ===> FILE '￥MYDIR￥JPCDATA.SAS'

もっとも、カレント・ディレクトリが MYDIR であれば、この例にあるようなパス名は省略することができて、

Command ===> FILE 'JPCDATA.SAS'

と入力しただけでも、同様にディレクトリ MYDIR のもとに保存される。もし

Command ===> FILE 'B:JPCDATA.SAS'

とすると、こう指定するだけで、ドライブ B になっているフロッピィ・ディスクにもファイルが保存できるので、プログラムやデータの保存や移植、出力の保存などには非常に便利である。

MS-DOS のファイル名の拡張子は必ず付けなくてはならないというものではないが、ファイルの内容の分類のために、筆者は PROGRAM EDITOR ウィンドウから保存したプログラムやデータには SAS、LOG ウィンドウから保存したログには LOG、OUTPUT ウィンドウから保存した出力結果には LST というファイル名拡張子を付けることにしている。拡張子を使ったとき、例えば JPC、JPC.SAS、JPC.LOG、JPC.LST はそれぞれまったく別のファイルを意味することになる。

CMS 版 SAS でも、ファイル名の形式が異なるだけで、PC 版 SAS とまったく同様に、各ウィンドウの内容は、それらが存在している各ウィンドウの Command 行に file コマンドで CMS のファイル名を

Command ===> FILE 'fn ft [fm]'

と入力すれば、CMS ファイルに保存することができる。ファイル・モード fm は省略すると A が既定値となっている。

もし既に同じファイル名 fn ft の CMS ファイルが存在している場合には、この file コマンドの実行にともない、

WARNING: The file already exists. Enter R to replace it, enter A to append to it or enter C to cancel FILE command.

と聞いてくるので、CMS ファイルの内容を更新する場合には R、付加する場合には A、file コマンドを取り止める場合には C を入力してから(実行)キーを押す。

ところで、こうしたプログラム、ログ、出力内容は、従来の「常識」からすると、必ず紙の上にプリント・アウトし、ハード・コピーとしてとっておくことになる。実際、PC 版 SAS では、各ウィンドウで file コマンドを用い、あたかもファイル名'PRN'の MS-DOS ファイルに出力するように、

Command ===> FILE 'PRN'

と入力するとプリンタに出力し、印刷できることになっている。(ただし、これにはプリンタの設定が適切に行われている必要があるので、注意を要する。) しかし、これはワー

ド・プロセッサーのパソコン・ソフトやワープロ専用機がこれだけ普及した今日にあっては、ほとんど用紙の無駄遣いにすぎない。こうした出力内容は、いきなり紙の上に印刷せずに、file コマンドによって、MS-DOS ファイルとして固定ディスクやフロッピィ・ディスクに書き出し、その上で、この MS-DOS ファイルを使い慣れたワード・プロセッサーによって編集し、必要なものだけを印刷する習慣をつけるべきであろう。

既に MS-DOS ファイルや CMS ファイルとして保存してあった SAS プログラムやデータは、PROGRAM EDITOR ウィンドウに限っては読み込むことができる。その場合には、MS-DOS ファイルや CMS ファイルに保存しておいた SAS プログラムのファイル名を PROGRAM EDITOR ウィンドウの Command 行に

Command ===> INClude 'ファイル名'

と入力して、画面上に読み込めばよい。もっとも、読み込んだからといっても、元の MS-DOS ファイルや CMS ファイルが失われるわけではない。いわば画面上にコピーしてくるわけである。SAS プログラムだけではなく、あらゆるファイルが読み込み可能である。したがって、SAS プログラムはもちろんデータのファイルの編集、変更に PROGRAM EDITOR ウィンドウが使える。さらには他のファイルのエディターとして使用することも可能である。ただし CMS 版 SAS では、データのファイルを読み込むときには、

WARNING: The file contains sequence numbers. Enter R to remove the sequence numbers or K to keep the sequence numbers.

と聞いてくるので、K と入力してから(実行)キーを押す。

ところで、PC 版 SAS でも、CMS 版 SAS でも、既に PROGRAM EDITOR ウィンドウ画面にプログラム、データ等の内容が表示されている場合には、その内容の最後部に続けてファイルを読み込むことになる。PROGRAM EDITOR ウィンドウに表示されている内容が消去された上で、全面的に新たなファイルに置き換えられて更新されることにはならないので注意が必要である。ファイル内容の全面的更新をしたいときには、後述する clear コマンドを使って、一旦、PROGRAM EDITOR ウィンドウの内容を消去しておく必要がある。

(4)SAS プログラムの編集

新規に SAS プログラムを作成する場合には、PROGRAM EDITOR ウィンドウの中の行番号のついている行ならば、どこからでも自由に文字を書き込んでいけばよい。カーソル・キーを使わなくても、リターン・キー(←)を使えば、改行することが出来る。ただし、行番号の次の 1 字分は保護フィールドと呼ばれ、文字を入力しようとすると「ピッ」と警告音がして、入力は受け付けられないので注意がいる。

PROGRAM EDITOR ウィンドウに表示されている最終行まで使い切った場合でも、リターン・キー(←)を押すと 1 行上にスクロールして、新たに 1 行表示されるので入力が続けられる。(ROLL DOWN)キーを押して、一気に次の画面に進むこともできる。前の行を見たり直したりしたいときには、(ROLL UP)キーを押して、前の画面に戻せばよい。CMS 版 SAS では、 (ROLL DOWN)キー、(ROLL UP)キーが使えないので、これらのキーの代わりに、forward コマンド、backward コマンドの設定されているファンクション・キーを用いる。

SAS の PROGRAM EDITOR ウィンドウの編集機能を司るエディターはフル・スクリーン・エディターと呼ばれ、書き込まれた文字が画面のイメージ通りに SAS プログラムとして扱われることになる。(ちなみに、これが仮にフル・スクリーン・エディターではなく、ライン・エディターであれば、入力し変更を加えた行は、コマンド入力のように、各行でリターン・キー(←)を押して、計算機にその旨を確認しないといけない。画面上の変更はあくまで画面上でのみの変化を意味し、ホスト計算機の記憶装置上の変化を意味しないの

で、編集は常に行単位で行わなければならないのである。その点、フル・スクリーン・エディターは非常に便利である。)

PROGRAM EDITOR ウィンドウでは次のような SAS 編集コマンドが使用できる。

(a)スクリーン・コマンド

SAS 編集コマンドのうちでも、スクリーン・コマンドはコマンド行に入力して使用するコマンドである。そこで、スクリーン・コマンドを使用する際にはカーソル・キーなどを使って、カーソルをコマンド行に移動させてからコマンドの入力を行う。あるいは PC 版 SAS では、二つのキーを同時に

(SHIFT)+(HOME CLR)

と押せば、一発でカーソルをコマンド行に移動させることができるので便利である。CMS 版 SAS では、これと同じ機能をもつ home コマンドを設定したファンクション・キーが普通はあるはずなので、それを KEYS ウィンドウで探して使えばよい。

ファンクション・キーで設定されているコマンドももともとはスクリーン・コマンドなので、例えば、HELP ウィンドウを表示させる際には、PC 版 SAS では(f･1)キーを押してもいいし、コマンド行に

Command ===> HELP

と入力してもいい。機能としては同じことになる。ファンクション・キーで設定されていないスクリーン・コマンドもあるが、よく使うものとしては、例えば次のようなものがある。

1. SAS セッションを中断し、MS-DOS コマンドまたは CMS コマンドを入力できるモードになる。

   Command ===> X

   SAS ディスプレイ・マネージャ・システムの画面に戻るには、MS-DOS では EXIT コマンドを、CMS では RETURN コマンドを実行する。
2. ウィンドウに表示されている内容を消去する。

   Command ===> CLEAR
3. カーソルを移動する。

   Command ===> TOp　……テキストの先頭へカーソルを移動する。

   Command ===> BOTtom　……テキストの最後へカーソルを移動する。
4. ウィンドウに表示されている以降のテキストで、指定の文字列を検索・置換する。

   Command ===> FINd 文字列　……文字列を検索する。

   Command ===> CHANge 文字列 1 文字列 2　……文字列 1 を文字列 2 に置換する。

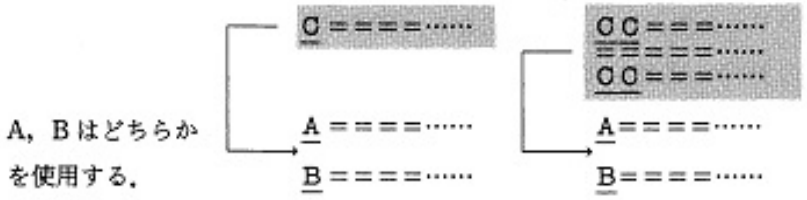
(b)行コマンド

SAS 編集コマンドのうちでも行コマンドと呼ばれるコマンド群は、これまでのスクリーン・コマンドとは異なった使い方をする。行コマンドは、編集画面左端に並んでいる 5 桁の行番号部分(これを=====で示すことにしよう)に、行番号の上に(1 桁目からの必要はない)重ね書きする形で入力して、リターン・キー(←)を押し、使用するコマンドである。次のようなコマンドが利用可能である。

1. 挿入(Insert): その行の後に n 行を挿入する。n を省略した場合には、1 行挿入する。

   In===　……

2. 複写(Copy): C で指定した行、または CC で指定した 2 行とそれに挟まれた行を、A で指定した行の直後、または B で指定した行の直前に複写する。



3. 移動(Move): M で指定した行、または MM で指定した 2 行とそれに挟まれた行を、A で指定した行の直後、または B で指定した行の直前に移動する。

4. 削除(Delete): D で指定した行、または DD で指定した 2 行とそれに挟まれた行を削除する。

ここで、2 と 3 で用いる A、B はそれぞれ After、Before の意味で、どちらか一方を用いて、複写先、移動先の指定を行うものである。A を入力した場合には、その行の直後に複写もしくは移動が行われ、B を入力した場合には、その行の直前に複写もしくは移動が行われる。

(5)SAS プログラムの実行と修正

SAS プログラムができたら、(f･10)キー(=zoom off; submit)を押して、PROGRAM EDITOR ウィンドウ上のプログラムを SAS に提出し(submit)、実行させてみる。ただし、処理を異常終了することもあるので、submit する前に SAS プログラムを MS-DOS ファイルに保存しておかないと、SAS プログラムは失われてしまうことがある。submit の前にこまめに file しておく方が安全である。

うまく SAS プログラムの実行に成功すると、出力があれば OUTPUT ウィンドウに表示されることになる。CMS 版 SAS では、うまく SAS プログラムの実行に成功して、出力があれば、この段階で、ディスプレイ画面全面が OUTPUT ウィンドウに切り替わる。

いずれにせよ、SAS プログラムの実行が失敗し、うまくいかなかった場合には、LOG ウィンドウ上に展開された実行結果を検討し、うまくいかなかった原因をチェックする必要がある。うまく実行されなかった原因は、ほとんどがプログラムの文法上の誤りやコマンドの綴り等の誤りで、これらはエラーとして、LOG ウィンドウに表示される。こうしたプログラム上のエラーはバグ(bug＝虫)と呼ばれ、エラーだらけで虫食い状態のプログラムは動かない。LOG ウィンドウに表示されているメッセージを注意深く検討して、エラーを取り除き、プログラムを修正して実行可能なプログラムに仕上げることを虫を取り除くという意味でデバッグ(debug)という。

初心者がよくやる誤りで、しかも本人がなかなか気がつかない誤りは、次のような見た目には似ている文字、記号の打ち間違いである。

- , (カンマ)と . (ピリオド)
- ; (セミコロン)と : (コロン)
- 0 (ゼロ)と O (オーの大文字)と o (オーの小文字)
- l (エル)と 1 (いち)

人間の目には大差なくても、コンピュータにとってはまったく異なる文字、記号なので、入力のときはもちろん、デバッグのときも細心の注意を払ってチェックしなくてはいけない。

デバッグするためには、まず実行に失敗した SAS プログラムを呼び戻さなければならない。そこで、(f･9)キー(＝recall)を押して、直前に実行したプログラムを PROGRAM EDITOR ウィンドウ上に呼び戻してくる。そこで、この呼び戻されてきたプログラムの修正を行い、できたらまた実行させてみる。ごく短いプログラムを除いて、プログラムが一度で正常に実行されることはめったにない。通常は何度も実行、修正のサイクルを繰り返して、ようやくまともに動くプログラムができるものである。それだけに、自分で組んだプログラムが実際に動いたときには、それなりの感激があるものである。その感激を味わうためにも、ここでくじけずに、こまめにデバッグするしかないのである。

(6)SAS プログラムの保存との終了

PROGRAM EDITOR ウィンドウ上の SAS プログラムは、そのまま SAS を終了してしまうと失われてしまう。そこで、完成した SAS プログラムは、ファイルとして保存しておこう。人間の手作業とは違い、プログラムとデータさえ保存しておけば、どんなに長い出力結果であっても、コンピュータの処理には 100%の再現性がある。SAS プログラムをファイルとして保存するためには、既に、(3)でも述べたように、

Command ===> FILE 'ファイル名'

と入力する。こうすることで、PROGRAM EDITOR ウィンドウ上にある SAS プログラムが、PC 版 SAS では指定ファイル名の MS-DOS ファイルに、CMS 版 SAS では指定ファイル名の CMS ファイルにそれぞれ書き出され、保存される。まさに、ファイルされるのである。

以上のような作業を行って、SAS を終了する際には、

Command ===> BYE

と入力すればよい。ただし、この SAS 終了によって、PROGRAM EDITOR ウィンドウの SAS プログラムが失われるだけではなく、LOG ウィンドウ、OUTPUT ウィンドウの内容も失われる。実は、LOG ウィンドウ、OUTPUT ウィンドウの内容は、途中 clear コマンドを実行しない限り、SAS を起動してから終了するまで(これをセッションと呼ぶ)の間、失われることなく蓄積、保存されているのである。したがって、画面をスクロールさせることで、そのセッションでそれまでに実行されていた諸結果を参照することも可能である。このことがあるために、セッションをあまり長く続けて、プログラムを流し続けていると、記憶、蓄積される各ウィンドウの内容が膨大になりすぎて、プログラム実行の際の異常終了の原因ともなるので、ときどきはウィンドウの内容を clear するか、あるいはいっそのこと時々SAS セッションを終了させた方がよい。

このように、bye コマンドによって、LOG ウィンドウ、OUTPUT ウィンドウの内容が失われるが、既に述べたように、それぞれのウィンドウで file コマンドを使えば、MS-DOS ファイルあるいは CMS ファイルとして保存することは可能なので、必要ならばそうするとよい。

4. SAS プログラムの基本

(1) SAS プログラムの基本的構成

SAS プログラムは、SAS 文と呼ばれる文によって構成される一種の文章である。SAS 文は SAS システムに対してある処理をさせるための命令文だから、SAS システムが読んで理解できるような形式、文法で書かれたものでなくては意味がない。各 SAS 文は形式的には「;」 (セミコロン)で終わる空白を含めた文字列であるが、

1. 一つの文が複数行にわたってもかまわない。
2. 一つの行にいくつもの文を書いてもかまわない。
3. 行の初めやキーワード間に空白を自由に入れてよい。

というように、形式的には極めて自由度が高い。

ただし、"*" で始まる SAS 文はコメントとみなされ、SAS システムは読み飛ばすことになっている。つまり、"*" で始まるコメント文は、SAS システムのためにではなく、プログラム作成者自身を含めた人間のために書かれるものである。実際、誰か他人のためというよりも、将来の自分がプログラムの内容を理解できるようにするために、プログラムの解説をコメントの形で入れておいた方がよい。

SAS は、利用者が作成したデータ・セットに対して、利用者が指定し、呼び出した各種統計用プログラム(SAS ではこのようにあらかじめ用意されたそれぞれのプログラムをプロシジャ(procedure)と呼んでいる)を機能的に結び付けるためのシステムである。利用者は SAS に対して、データ・セットの作成と希望するプロシジャの指定さえ行えばよく、あとは SAS が統計処理を行ってくれるのである。したがって、SAS プログラムは、基本的には次の２種類のステップの繰り返しとして構成される。

1. DATA ステップ: SAS データ・セットの作成・加工をする SAS 文の集合である。このステップで SAS データ・セットを次々と作成しながら加工処理していく。
2. PROC ステップ: SAS データ・セットに対する統計処理をするプロシジャを呼び出す SAS 文の集合である。

したがって、通常書かれる一番単純な SAS プログラムは一つの DATA ステップと一つの PROC ステップとから構成される。DATA ステップ、PROC ステップはそれぞれ形式的には、DATA 文、PROC 文で始まり、次の DATA 文、PROC 文または RUN 文までの SAS 文から構成される。

(2)SAS データ・セット

DATA ステップで作成・加工の対象となり、PROC ステップで統計処理の対象となる SAS データ・セットは、特殊な形式のファイルである。より具体的に、質問票調査をした場合を想定するとわかりやすいが、調査対象である各個人、これを SAS ではオブザーベーション (observation)と呼んでいるが、このオブザーベーションごとに、各質問に対する回答を表形式にまとめて整理したものとなっている。SAS 流にいえば、SAS データ・セットはオブザーベーション番号 1, 2, 3,...ごとに、いくつかの変数(variable)、例えば X1, X2, X3,... に対するデータ値を書いた表 2.3 のような表形式のファイルということになる。いわば表形式のデータ・ベースである。

表 2.3　SAS データ・セットの構造

オブザーベーション番号	変数			
	X1	X2	X3	・・・・・・
1				
2				
3				
: :				

SAS データ・セットには、

1. SAS セッションが終ると消去される一時 SAS データ・セットと
2. SAS セッションが終っても保存され残っている永久 SAS データ・セット

の 2 種類のデータ・セットがある。せっかく生成した SAS データ・セットを SAS のジョブごとに作り直さないで保存しておき、データ・ベースとして何度も再利用しようというのが、この永久 SAS データ・セットである。この 2 種類の SAS データ・セットは、PC 版 SAS と CMS 版 SAS とでは、やや扱いが異なる。

(a) PC 版 SAS の SAS データ・セット

1. 一時 SAS データ・セット: このファイルは SAS システムが使用する作業用の一時ファイルで、ディレクトリ SASWORK の下に SAS が自動的に作成する MS-DOS ファイルであるが、SAS 終了時に自動的に消去される。
2. 永久 SAS データ・セット：このファイルは SAS システムが使用する作業用ファイルで、利用者が指定するディレクトリの下に SAS が作成する MS-DOS ファイルで、SAS 終了後も MS-DOS ファイルとして保存され残る。

(b) CMS 版 SAS の SAS データ・セット

1. 一時 SAS データ・セット: ファイル・ネーム fn だけからなる 1 レベル名の SAS データ・セットである。このファイルは SAS システムが使用する作業用の一時ファイルで、次の SAS ジョブの開始時または TSS のログオフ時に消去される。
2. 永久 SAS データ・セット: ファイル・ネームとファイル・タイプの 2 レベル名をもつ CMS ファイル fn ft である。SAS プログラムの中で表記する場合には、ft.fn と日本風に、姓が前、名が後ろになる。その際に "." を間に入れるのは、英語で名前を姓、名の順に書き表すときに、間に "," を入れるのと同じ発想である。この 2 レベル名をもつ SAS データ・セットは、他の操作、指定等なしで、自動的に永久ファイルとなり、CMS ファイルとして保存される。ただし、ft を "WORK" とすると、一時ファイルになってしまうので、それ以外のものにすること。

SAS データ・セットは元のデータ・ファイルに比べるとかなり大きな容量を必要とする。PC 版 SAS では、この章で後述する基本型プログラムでも 82.5KB 程度のデータ・ファ

イルから 607.5KB 程度の SAS データ・セットが生成され、7.4 倍にもなる。通常はもっと多種多様な変数が作成されるので、例えば、この章の通常型プログラムでは、同じデータ・ファイルから 1.2MB 程度の SAS データ・セットが生成される。実に 14.5 倍である。一時 SAS データ・セットでも、SAS 終了時まで消去されないので、1 回の SAS セッションの間に、異なる名前の一時 SAS データ・セットをたくさん作ると、大容量の固定ディスクでも、あっという間に空き領域を使い尽くすことになるので、特に、パーソナル・コンピュータで SAS を使う場合には、できるだけ同じ名前の SAS データ・セットを繰り返し繰り返し使用した方が、ディスク容量の節約になる。また、SAS データ・セットの生成には、その都度時間がかかるので、メインフレームならばともかく、パーソナル・コンピュータでは同じ一時 SAS データ・セットを何度も生成するのは時間の無駄である。そのため、本書では、一度、永久 SAS データ・セットを生成しておき、後のプログラムはそれを繰り返し再利用することを原則としている。

(3)変数

SAS では変数名は英字で始まる 8 文字以内の英数字とすることになっている。最大 4,000 個まで変数を定義することが可能である。また、変数には数値データも文字データも与えることができる。ただし、変数に文字データを与える際には、それが文字データであることを明示する必要がある。具体的には、文字データを ' ' で囲むことで、例えば

'123'　'ABC'　'ASA100'

のように、それが文字データであることを明示する。したがって、123 は百二十三という「数値」であるが、'123' は 123 という「数字」つまり文字列であることを意味している。変数のとりうる値は、数値データの場合には $\pm 10^{-73} \sim 10^{73}$ の範囲、文字データの場合には 1～200 文字(バイト)である。

質問票調査の場合を考えるとわかるように、調査を実際に行うと、一部無回答の質問票というのが、ある程度の割合で出てくることになる。つまり、オブザーベーションによっては、実際には値の存在しない変数がいくつか存在することになる。こうした場合には、SAS システムに変数の値が「欠損している」ということ知らせるために、データ・エントリーを行う際に、何らかの「値」を与えなくてはならない。この「値」のことを欠損値と呼び、SAS では変数に表 2.4 の欠損値を与えておくと、SAS が統計処理を行なう際に、処理対象の変数が欠損値であるオブザーベーションについては、あらかじめそれを除いた上で各種処理が行われることになっている。

表 2.4　欠損値

	数値変数	文字変数
欠損値を示すデータ	• ピリオド "."	• 空白 " "
他に欠損値として扱われる入力データ	• 数値データ以外 • フォーマット形式[1)]での空白 • フォーマット形式に合わない数値データ	• 入力形式に合わないもの

1) フォーマット形式については、次の第 5 節で説明する。

また、SAS プログラムの中では変数名をリスト状に並べて用いたり、同じような変数名がいくつも並んだりするようなケースが起こる。実は、その方がプログラムを書く際には楽なのである。なぜなら、例えば、五つの変数を並べた

X1 X2 X3 X4 X5

のような場合には、

X1-X5

という省略形が使えるからで、これはとても便利である。一般的には、Xm-Xn と記すと、Xm から Xn までのすべての変数(m

ALL ……すべての変数

NUMERIC ……すべての数値変数

CHARACTER……すべての文字変数

のような省略形も用意されている。

ところで、変数名を考え出すというのは、意外と大変な作業である。変数の数が増えるほどその大変さは身に染みてくる。実際、変数名を考えた本人ですら、変数名を見ただけでは、その変数がどの質問項目に対応していたのかがわからなくなってしまう。しかも、変数名が複雑で錯綜してくればくるほど、さきほどのような変数名の省略形を使いにくくなってくるのである。そこで、お勧めなのは、筆者が実行しているちょっとした工夫である。第 6 章でも触れるが、まず、変数名リストの省略形を使いやすいように、変数名の末尾は数字を質問の順に付けたものがよい。また変数名は英字から始めなくてはいけない。したがって、質問票の設計段階で、事前に質問番号を

アルファベットまたはローマ数字による大分類［＋中分類］＋数字による小分類

という構成にしておくのである。例えば、

質問番号	II.3	IV.4	A.III.5	D.IX.9
変数名	II3	IV4	AIII5	DIX9

というようにすると、簡単に変数名が付けられるし、逆に変数名を見れば、元の質問番号がすぐにわかるのである。SAS プログラムで変数名を定義するときに質問票の質問番号をそのまま変数名にすることができて、便利で間違いも少ない。ただし、変数名は 8 文字以内なので、質問番号を決めるときにはなるべくシンプルにつけた方がよいことに変わりはないが……。変数の具体的な意味内容については、変数名に反映させるのではなく、後述するような変数ラベルを用いた方がよい。これで出力時に表示することができ、その際には字数制限も 40 文字以内と大幅に緩くなっており、日本語、漢字なども使え、はるかに実用的である。

5. 基本型 SAS プログラム

この節では、実際に簡単な SAS プログラムを作成し、データ・エントリーと単純集計及び永久 SAS データ・セットの生成を行ってみよう。この一連のプロセスと SAS プログラムの基本を理解してもらうために、この節では、もっともシンプルな基本型 SAS プログラムを取り上げる。この基本型プログラムを進化させる形で、次の節で、筆者が実際に調査のときに使用する通常型 SAS プログラムへと発展させる。もっとも、基本型プログラムでも欲を出さねば、十分実用になる。

(1) DATA ステップでのデータ入力

例えば、第 6 章で調査の実例としてあげる「組織活性化のための従業員意識調査」で得られた回答は、調査票のままでは、人間には読めても、コンピュータや SAS は読み取ることができない。そこでまず最初に、コンピュータや SAS が読み取り可能な形のデータに書き直してやる必要がある。具体的には、回答をデータ値として、ある規則に則ってディス

プレイ画面上やファイル上に書き並べてやるのである。これをデータ・エントリーという。よく用いられる代表的なデータ値の書き並べ方には次の二つの形式がある。

1. リスト形式……各変数のデータ値は、1 個以上の空白によって分離されて並べられている。
2. フォーマット形式……各変数のデータ値がどのカラムに書き込まれているかがあらかじめ指定されている。

それぞれに対応する形で、後述する INPUT 文の変数名の並びの形式がある。これらの形式がよく使われる入力の仕方について、それぞれ具体的に説明しておこう。

(2)リスト形式で SAS プログラム内部からのデータの読み込み

これは SAS プログラム内部に直接データを書き込む方法である。リスト形式では、リスト入力が行われ、変数名の並びに対応させながらデータ値を読み込む。一般的には、変数名の並びは、INPUT 文で

INPUT 変数名 1 変数名 2 ……;

とそのまま変数名を書き並べる形式をとる。リスト入力の場合、データ値は空白によって区切られるということの他に、

1. 必要な小数点はデータの中に含まれていること。
2. データの欠損値は "." で示されていること。
3. 変数に文字データを与えるときには "変数名 $" と変数名の後に$を必ず付けること。
4. 一つのオブザーベーションに対するデータが 2 行以上にわたるときには、行の区切りを示すために "/" を変数名の並びの中に入れること。

などに注意しなくてはならない。

いま、この方式でデータを入力し、永久 SAS データ・セットを生成することを考えよう。PC 版 SAS では、次のようにする。

PC 版 SAS

```
LIBNAME libname '￥パス名';
DATA libname.永久 SAS データ・セット名本体;
INPUT 変数名の並び;
CARDS;

                データ行

;
```

例)

```
LIBNAME SAVE '￥MYDIR';
DATA SAVE.JPC;
INPUT KC SC IC I1-I4;
CARDS;
5 7 01 1 8 1 4
5 7 02 1 3 1 4
5 7 03 1 3 1 3
5 7 04 1 3 1 4
5 7 05 1 5 1 2
5 7 06 1 5 2 3
5 7 07 1 3 1 2
5 7 08 1 5 1 2
;
```

PC 版 SAS では、永久 SAS データ・セットのファイル名の拡張子は必ず "SSD" として生成される。このプログラム例の場合には、"JPC.SSD" というファイル名の永久 SAS データ・セットが MS-DOS ファイルとして生成されることになる。LIBNAME 文では、この永久 SAS データ・セットがどこのディレクトリの下に生成されるのかを示すためのパス名が指定される。LIBNAME 文でこのパス名と対応づけられた libname (ライブラリ参照名)が、パス名の代わりに、その SAS セッションの間、用いられることになる。

このプログラム例の場合では、MYDIR というパス名を "SAVE" という libname に結び付けて定義している。この指定により、永久 SAS データ・セット JPC.SSD はディレクトリ MYDIR の下に生成されることになるとともに、"SAVE.JPC" で "￥MYDIR￥JPC.SSD" を意味することになる。

CMS 版 SAS では、PC 版 SAS のように libname の指定がいらない。既に、第 4 節(2)(b)で述べたように、CMS 版 SAS では、永久 SAS データ・セット名＝ft.fn と ft が先にくるが、この 2 レベル名の指定だけで自動的に永久 SAS データ・セットが生成され、libname は必要がない。したがって、CMS 版 SAS では PC 版 SAS では必要だった 1 行目の LIBNAME 文はつけず、2 行目の DATA 文から始める。その際、形の上では PC 版 SAS の DATA 文中の libname を指定していたところで、CMS 版 SAS では永久 SAS データ・セット名の ft を指定することになる(ft を WORK にすると、一時 SAS データ・セットとして認識されてしまうので、それ以外のものにすること)。CMS 版 SAS では、このように指定するだけで, SAS データ・セットは永久 SAS データ・セットとなり、DATA ステップで行った加工処理は、そのまま CMS ファイルとして保管されることになる。したがって、形式的には、PC 版 SAS のプログラム例の 1 行目を削除したものが、そのまま CMS 版 SAS のプログラム例となる。すなわち、

CMS 版 SAS

```
DATA  永久 SAS データ・セット名;
INPUT  変数名の並び;
CARDS;

                データ行

;
```

例)

```
DATA SAVE.JPC;
INPUT KC SC IC I1-I4;
CARDS;
5 7 01 1 8 1 4
5 7 02 1 3 1 4
5 7 03 1 3 1 3
5 7 04 1 3 1 4
5 7 05 1 5 1 2
5 7 06 1 5 2 3
5 7 07 1 3 1 2
5 7 08 1 5 1 2
;
```

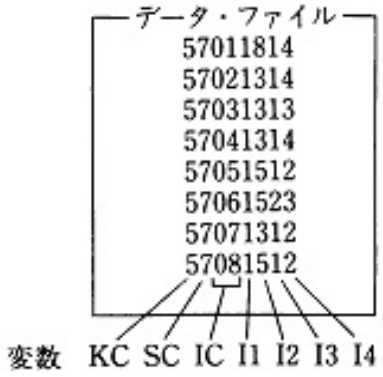
いずれにせよ、リスト形式はデータ値を空白によって区切っていくので、変数が多くなると、空白だらけになる。しかも、空白もまた 1 文字であり、キーの 1 タッチを要する。したがって、データを大量にエントリーする場合には、効率やファイル容量の点では問題があることには注意がいる。また、SAS プログラム内に直接、データを書き込んでおくことで、大量のデータ、例えば 1,000 オブザーベーションのデータを扱うと、プログラムの長さも、最低でも 1,000 行を超えることになり、データの編集、デバッグ等を考えると実用的ではない。そこで考えられるのが、プログラムとデータをファイルとして分離しておく次のような方式である。

(3)フォーマット形式でデータ・ファイルからのデータの読み込み

フォーマット形式では、リスト形式のように、データ値の間に空白の区切りがないので、データ・ファイルを見ただけでは、どこからどこまでが一つのデータ値になっているのかがわからない。そこで、データ・ファイルを作成した人間の側で、そのデータ・ファイルの中でデータ値がどのような書式で並べられているのかを指示してやる必要がある。

いま、MS-DOS ファイル JPCDATA.SAS、あるいは CMS ファイル JPCDATA SAS からデータを読み込むことを考えよう。このファイルには、(a)のプログラム例と同じデータが図 2.6 のように入っているとする。つまり、変数 KC, SC, IC, I1, I2, I3, I4 に対応するデータ 8 人分が、図 2.6 のように変数に対応させながら入力してある。

図 2.6　データ・ファイル(JPCDATA.SAS あるいは JPCDATA SAS)の中のデータと変数の対応



入力した人間の側では、図 2.6 で示したような対応をつけながら入力したわけであるから、その対応関係のことを SAS 側に教えてやらなければならない。そのとき、各変数に対応するカラムを指定するために用いられるのが書式であり、この場合には、例えば単純に変数とカラムの対応を付けて、

INPUT KC 1. SC 1. IC 2. I1 1. I2 1. I3 1. I4 1.;

あるいは、変数名の並びと、書式のまとまりを考えて、

INPUT (KC SC IC I1-I4)(2*1. 2. 4*1.);

というように指定してやればよい。つまり、リスト形式のように、データ値の間に、空白の区切りがないので、書式の上で区切りをつけてやろうというのである。一般的には、INPUT 文で

INPUT (変数名 1 変数名 2 ……)(対応する書式)[(変数名 3 ……)(対応する書式)……];

という形式をとる。下線部全体をさして、フォーマット形式では「変数名の並び」と呼ぶ。つまり、フォーマット形式では「変数名の並び」には書式まで含まれていることになる。ここで、書式とは次のものを空白で区切って並べたものをいう。

w.　……フィールドの長さ w(1≦w≦32)の数値データを変数に与えるとき
$w.　……フィールドの長さ w(1≦w≦32)の文字データを変数に与えるとき
m*w.　……同じ w の数値データのフィールドが m 個並んでいる場合
m*$w. ……同じ w の文字データのフィールドが m 個並んでいる場合
+n　……n カラム読み飛ばすとき

また、一つのオブザーベーションに対するデータが 2 行以上にわたるときには、変数名とそれに対応する書式は 2 行分以上にわたらないようにまとめて分け、その間に行の区切りを示すために "/" を入れる。

PC 版 SAS では、このような書式を用いて、図 2.6 の MS-DOS ファイル JPCDATA.SAS からのデータを読み込み、永久 SAS データ・セットを生成するプログラムは次のようになる。

PC 版 SAS
LIBNAME libname '￥パス名'; DATA libname.永久 SAS データ・セット名本体; INFILE データ・ファイル名; INPUT 変数名の並び;

例)

```
LIBNAME SAVE '￥MYDIR';
DATA SAVE.JPC;
INFILE 'JPCDATA.SAS';
INPUT KC SC IC I1-I4;
```

このプログラム例のように、データ・ファイル名(例では JPCDATA.SAS)の指定の際にドライブ名、パス名を省略した場合は、データ・ファイルはカレント・ディレクトリ(本書では MYDIR)の中になければならない。

CMS 版 SAS では、永久 SAS データ・セットの使用の際に、libname の指定がいらないが、その代わり、CMS ファイルからのデータの読み込みの際には ddname の指定が必要になる(詳しくは付章第 2 節を参照のこと)。CMS では、データ入力機器の既定値は端末になっているので、CMS ファイル(fileid = fn ft [A])からデータを読み込む際には、"X CMS コマンド"でデータ入力機器の設定を変更する必要がある。

CMS 版 SAS
X FIledef ddname DISK fileid; DATA 永久 SAS データ・セット名; INFILE ddname; INPUT 変数名の並び;

例)

```
X FI IN1 DISK JPCDATA SAS;
DATA SAVE.JPC;
INFILE IN1;
INPUT (KC SC IC I1-I4)(2*1. 2. 4*1.);
```

ddname として筆者は"18"または"IN"などをよく使う。ここでのプログラム例は、PROGRAM EDITOR ウィンドウで作成したデータ・ファイルである CMS ファイル JPCDATA SAS を読み込むことにしている。

いずれにせよ、筆者が大量データの際に勧めるのは、このフォーマット形式である。この形式は、多くの変数と多数のオブザーベーション、つまり大量のデータを扱うような場合、特にデータ・エントリーを外注して、フロッピィ・ディスクや磁気テープでデータが納入されるような場合に適している。PC 版 SAS では、データ・ファイルをいちいち固定ディスクにコピーしておかなくても、INFILE 文のデータ・ファイル名の指定を、例えば

```
INFILE 'B:JPCDATA.SAS';
```

とドライブ名つきで指定すれば、直接、ドライブ B のフロッピィ・ディスクからデータを読み込むこともできる。固定ディスクから読み込むよりは、やや遅くなるとはいうものの、本書で扱うようなデータではせいぜい秒単位の話である。

ところで、CMS 版 SAS では、文字変数を使用する場合には、意外なトラブルに巻き込まれることがある。それは、メインフレームの端末やプリンタによっては、"$" を "￥" に置き換える必要があるためである。したがって、なるべく"$"を使用しないような SAS プログラムを作った方が無難である。例えば、データはすべて数字で作成し、数値変数で入力できるようにしたり、それでも、文字変数をどうしても使用したければ、'文字列'代入方式をとるというような工夫をしてみた方がよい。

(4)単純集計の PROC ステップ

以上のような DATA ステップに続いて、PROC ステップに様々な統計処理を実行させるための SAS 文を書くことになるが、一般には、この PROC ステップは DATA ステップと比べても簡単で短い行数で済むことが多い。ここでは、一番最初にするべき統計処理である単純集計のみを例にして考え、他の様々な統計処理についてはそれぞれ対応する章において、SAS 文の使用について取り上げることにしよう。

単純集計は、初心者はややもすると軽視しがちであるが、統計処理として一番基本的でかつ欠かしてはならない重要なものである。まず、この単純集計結果は報告書等の作成の際には、必ず掲載しなくてはならない情報の一つである。また、それだけではなく、それによって、データ・エントリーのミスや SAS プログラムの DATA ステップでの様々な設定が正しくなされているのかをチェックするという意味でも、欠かすことの出来ない重要な初期処理である。この段階で、次の第３章で述べる階級の設定なども実際の度数分布を参考にしながら行ってしまうのである。この単純集計を行うというステップで念入りにチェックを行っておかないと、後で統計処理のやり直しを迫られる羽目になるので十分注意深く行ってほしい。この単純集計を見ながら、これから行う統計処理についての様々なアイデアを練るのである。

統計学の用語でいえば、単純集計をとるということは、全変数の度数分布表を作成するということである。それには次のようなごく簡単な SAS 文だけで済む。

```
PROC FREQ;
[TABLES  変数名の並び;]
[OPTIONS NOCENTER;]
```

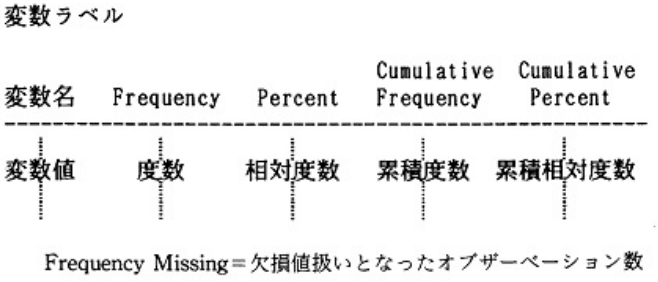
例)

```
PROC FREQ;
        TABLES KC SC I1--VI15;
        OPTIONS NOCENTER;
```

ここで、TABLES 文を省略すると、既定値通り、全変数について度数分布表が作成される。もし全変数の度数分布表が必要ないときには、プログラム例のように必要なものだけを変数名の並びで TABLES 文の中に指定することもできる。しかし、単純集計では、プログラム、データのチェックをするという意味からも、全変数についての度数分布表を作成するのが原則である。

また、OPTIONS NOCENTER; を指定すると出力は左端に寄せて行われる。この指定を省略すると出力は画面の中央に寄せて行われる。出力を MS-DOS ファイルや CMS ファイルに書き出して、ワープロで編集する際には、OPTIONS NOCENTER; を指定して、出力を左端に寄せて行っていた方が、ファイル容量的にも、作業的にも無駄がない。中央寄せとは、結局、各行の先頭部分にたくさんの空白(ブランク)を挿入することだからである。この PROC FREQ によって作成される度数分布表は、次のような形式になっている。

図 2.7　PROC FREQ によって作成される度数分布表の一般形式



度数分布表の形式や読み方については、次の第 3 章で説明するので、ここではとりあえず、単純集計とは、この度数分布表、その中でも特に、度数と相対度数を求めることだということだけを言っておこう。このような度数分布表の形で得られた単純集計結果で確認すべきことを、もう一度まとめておこう。

1. 変数値にあり得ない値が存在していないかを確認する。もし、ありえない値が存在している場合は、コーディング・ミスやエントリー・ミスといったデータのエラーの場合と、入力のフォーマット形式の誤りや欠損値処理のエラーなどのプログラムのエラーの場合が考えられる。後者の場合もあることを念頭に置いて、エラーの可能性はしらみつぶしにし、これによって SAS プログラムが期待通りに機能しているかのチェックを行い、SAS プログラムのデバッグを行う。
2. 1 と同時にコーディング・ミスやエントリー・ミスのチェックを行い、もしこうしたミスが存在しているときは、データ・ファイルを PROGRAM EDITOR ウィンドウに呼び出して、データを直接訂正する。
3. 階級の分け方が適切に行われているかどうかを吟味・検討する。必要ならば、新たな階級を定義し、階級用の文字変数を定義する。このことについては、次の第 3 章を参照のこと。

(5)基本型 SAS プログラム例とその実行

ここでは、質問調査票の回答のデータ・エントリーと単純集計の作成、そして今後の統計処理に再利用する永久 SAS データ・セットの生成を作業の順に取り上げる。まず、データ・エントリーから始めよう。

(a)データ・エントリー

基本型 SAS プログラムの中でデータ・ファイルとして指定することになる MS-DOS ファイル JPCDATA.SAS、あるいは CMS ファイル JPCDATA SAS を自分で作成する場合を考えてみよう。データ・ファイルは自分の使いやすいワープロ・ソフトで作ってもよいが、ここでは、せっかく PROGRAM EDITOR ウィンドウの使い方を説明したので、この PROGRAM EDITOR ウィンドウで作成してみよう。データ・ファイル作成はいたって簡単である。図 2.7 では、データ例として、第 6 章で実例として取り上げる「組織活性化のための従業員意識調査」の実際の回答のうち、20 人分のデータを PROGRAM EDITOR ウィンドウで入力、作成しているところを 2 画面で例示している。

この PROGRAM EDITOR ウィンドウでの入力作業が終ったら、コマンド行に、PC 版 SAS のときは

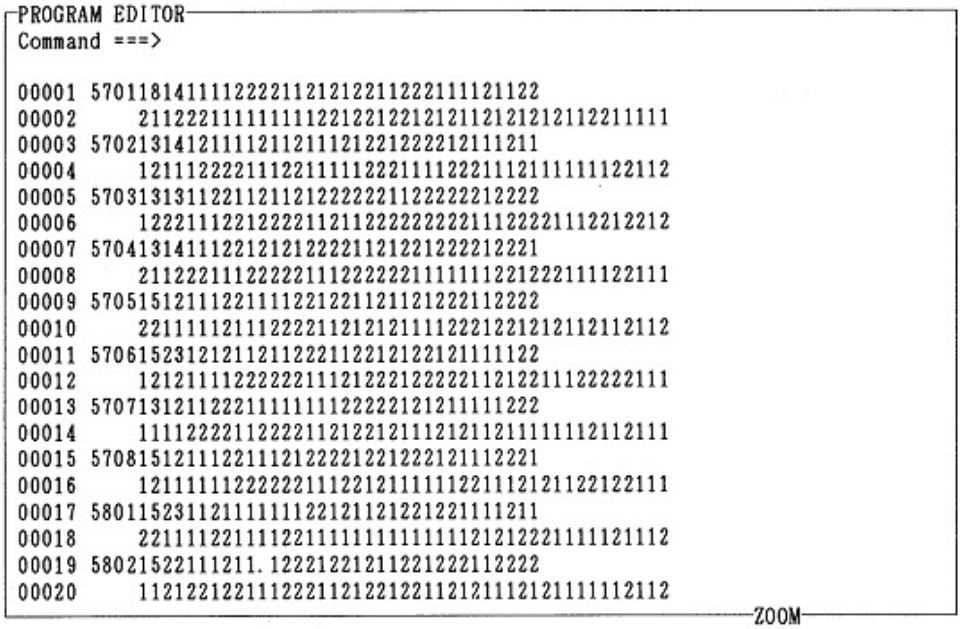
```
Command ===> FILE 'JPCDATA.SAS'
```

CMS 版 SAS のときは

```
Command ===> FILE 'JPCDATA SAS'
```

と入力して、MS-DOS ファイル JPCDATA.SAS (ディレクトリは自動的にカレント・ディレクトリ MYDIR になる)または CMS ファイル JPCDATA SAS の中に保存しておこう。もっとも、入力し終るまで保存しないというのは、もし何がトラブルがあったときに、それまでのデータ・エントリーの苦労が水の泡となってしまい、危険なので、作業途中に何度も、この file コマンドを使って、こまめにファイルへの保存作業を行っておいた方が無難である。

図 2.8　PROGRAM EDITOR ウィンドウで作成するデータ・ファイルの一部(20 人分)
(MS-DOS ファイル JPCDATA.SAS、あるいは CMS ファイル JPCDATA SAS)



実際には、データ・ファイル JPCDATA.SAS あるいは JPCDATA SAS の中には、907 人分 1814 行のデータを入れることになるわけだが(ファイル容量としては 82,538 バイト)、これだけの量になると、信頼できるデータ・エントリー業者(昔はカード・パンチを外注するということから「パンチ業者」といっていた)に外注した方が、タイプ・ミスもなく安全である。しかし、このわずか 20 人分でも実際に自分の手で入力してみると、

1. データがすべて数字で、しかもほとんど 1 と 2 ばかりである。
2. フォーマット形式にして、数字の間に空白(ブランク)を入れない。

という「データ・エントリー2 原則」が、入力をいかに容易にかつスピーディにしているか

が実感できるだろう。仮に、データがアルファベットで、しかもリスト形式で空白で区切って並べる必要があるとすると、時間がかかって面倒なだけではなく、プロがやってもタイプ・ミスを誘発することになる。われわれのような素人では、目もあてられない結果に終る。つまり、非標本誤差の増大を招くのである。したがって、簡単なことではあるが、この二つの原則は重要なのである。

(b)基本型 SAS プログラムの作成

最もシンプルで簡単な基本型 SAS プログラムの例は図 2.9 のようなものになる。図 2.9 はこの基本型プログラムを PROGRAM EDITOR ウィンドウで作成した状態を示している。これは基本型とはいうものの、第 6 章で調査の実例として挙げる「組織活性化のための従業員意識調査」によって得られたデータを実際に単純集計し、永久 SAS データ・セットを生成することができる SAS プログラムである。シンプルではあるが、これでその役目を十分に果たすことになる。

図 2.9　PROGRAM EDITOR ウィンドウで作成した基本型 SAS プログラム(PC 版 SAS)
(MS-DOS ファイル JPCP0.SAS、あるいは CMS ファイル JPCP0 SAS に保存)

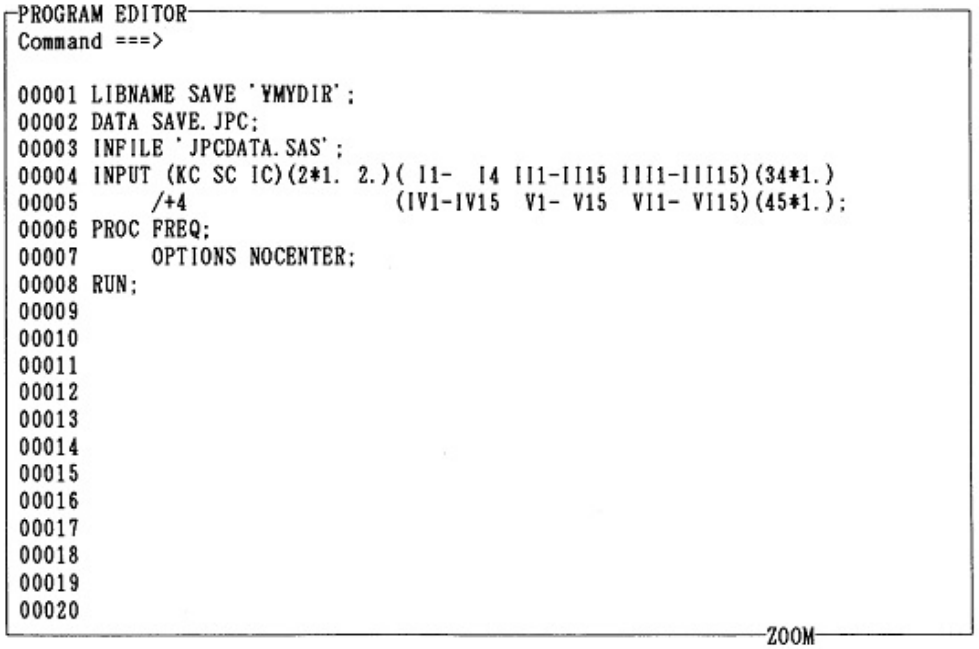


図 2.9 は PC 版 SAS 用のプログラムであるが、CMS 版 SAS 用は 1 行目、3 行目をそれぞれ次のように置き換えればよい。

00001 X FI IN1 DISK JPCDATA SAS;

00003 INFILE IN1;

この基本型 SAS プログラムは、PROGRAM EDITOR ウィンドウで作成した後、やはり file コマンドで、MS-DOS ファイル JPCP0.SAS、あるいは CMS ファイル JPCP0 SAS に保存しておく。

(c) SAS プログラムの実行

さて以上の準備が整ったら、さっそく SAS プログラムを実行してみよう。うまくいけば、この SAS プログラムの実行に伴って、次々と LOG ウィンドウに表示が現れて、実行終了時には、図 2.10 のような表示が出されているはずである。

図 2.10　基本型 SAS プログラムを実行したときの LOG ウィンドウの表示(PC 版 SAS)

```
LOG
Command ===>

NOTE: Copyright(c) 1985,86,87 SAS Institute Inc., Cary, NC 27512-8000, U.S.A.
NOTE: SAS (r) Proprietary Software Release 6.04
      Licensed to SAS INSTITUTE TRIAL SITE, Site 00000001.

    1    LIBNAME SAVE '¥MYDIR';
    2    DATA SAVE.JPC;
    3    INFILE 'JPCDATA.SAS';
    4    INPUT (KC SC IC)(2*1. 2.)( I1-  I4 II1-II15 III1-III15)(34*1.)
    5         /+4                (IV1-IV15  V1- V15  VI1- VI15)(45*1.);
    6    PROC FREQ;
NOTE: The infile 'JPCDATA.SAS' is file A:¥MYDIR¥JPCDATA.SAS.
NOTE: 1814 records were read from the infile A:¥MYDIR¥JPCDATA.SAS.
      The minimum record length was 38.
      The maximum record length was 49.
NOTE: The data set SAVE.JPC has 907 observations and 82 variables.
NOTE: The DATA statement used 3.13 minutes.
    7         OPTIONS NOCENTER;
    8    RUN;
NOTE: The PROCEDURE FREQ used 6.33 minutes.

                                                                    ZOOM
```

このログの内容からもわかるように、PC 版 SAS では、この SAS プログラムの実行時間も表示される。この例では、実行には計 10 分弱を要したことがわかる(もっとも、まったく同じ SAS プログラムでも、この実行時間は実行の度に数秒程度のばらつきが出る)。その結果、生成される永久 SAS データ・セット JPC.SSD は 607,528 バイトつまり約 600KB の大きさの MS-DOS ファイルとなる。CMS 版 SAS では、実行時間は表示されないが、はるかに短時間で実行が終り、永久 SAS データ・セット JPC SAVE が生成される。

単純集計の結果は、OUTPUT ウィンドウに次々と表示されるが、そのうちの一画面分を抜き出すと、図 2.11 のようになる。

図 2.11　基本型 SAS プログラムを実行して得られた出力の一部を表示する OUTPUT ウィンドウ

```
OUTPUT
Command ===>

SAS                                        10:58 Monday, January 27, 1992     9

                          Cumulative  Cumulative
I1   Frequency   Percent   Frequency    Percent
-----------------------------------------------
 1        765      87.2         765       87.2
 2        112      12.8         877      100.0

Frequency Missing = 30

                                                                    ZOOM
```

質問 I 1 は性別に関するもので、この場合、1 は男で、2 は女を表している。こんな簡単な単純集計であっても、従業員（正社員）の多くが男性であるという日本の大企業におけるホワイトカラーの職場の実態が垣間見える。

6. DATA ステップでのデータの加工

以上の基本型 SAS プログラムは、単純集計用プログラムの中でも最も簡単なものである。これでも十分に実用になるが、筆者が実際に調査の際に用いる通常型 SAS プログラムでは、出力結果をわかりやすくするために変数にラベルをつけたり、後々行う統計処理のことなども考えて、これらの変数をもとにして様々なデータの加工、処理を行ったりしておくとともに、出力結果をワード・プロセッサーで編集する際の作業をやりやすくするような様々な工夫を施してある。そこで、この節では、この最もシンプルな基本型 SAS プログラムをベースにして、実際の調査に用いられる単純集計用の通常型 SAS プログラムへと進化させてみよう。

(1)変数のラベル

変数には変数名だけではなく、変数の説明のためにラベルをつけることができる。変数のラベルは 40 文字(バイト)以内で、次のような形式で指定することが出来る。

```
LABEL  変数名 1='ラベル 1'  変数名 2='ラベル 2' ……;
```

例)

```
LABEL KCODE='COMPANY' SCODE='UNIT'
      I1='SEX'   I2='AGE'   I3='OCCUPATION'   I4='RANK';
```

ラベルそれ自体の中に「'」を含めたいときは「''」と 2 個続けて書くと、SAS では 1 個と認識され、表示されることになる。

変数のラベルに日本語データ(全角文字)を使うことも出来る。しかし、筆者としては、特に、CMS 版 SAS では、SAS プログラムの段階で全角文字を使用することは勧められない。実際の経験からすると、プログラム段階で日本語データを含めておくと、OS やメインフレームの端末の種類、日本語変換プロセッサーのくせや性能によって、処理内容が影響を受け、思いがけないトラブルの原因になるからである。しかも、IBM のメインフレームでは、シフト・コードで全角文字をはさむ必要があるので、そのときにはラベルは 19 文字以内ということになるだけではなく、そのシフト・コードがデータの転送やワード・プロセッサーによる編集の際にトラブルの原因になるのである。出力結果の図表でラベルとして日本語を用いたいのであれば、出力を書き出した MS-DOS ファイルを編集する際に、日本語ワード・プロセッサーの編集機能を使って、日本語ラベルに置換していった方がはるかに確実、実用的でかつ速い。

(2)変数への代入

変数に変数値を代入することは、割り当て文(assignment statement)によって行われる。割り当て文といっても、形式はいたって簡単で、

```
変数名=式;
```

と "=" を使って代入することを示すだけでよい。この割り当て文を使えば、INPUT 文を使わなくても、つまり、データ入力の形式によらなくても、変数に値を与えることが出来る。変数に与えることの出来る値は次のようなものである。

(a)定数

定数は数値データでも文字データでもよい。数値データの場合には、直接数値を代入すればよいが、文字データの場合には、その文字データを'　'で囲んで、それが文字データであることを明示する必要がある。具体的には次のような形式になる。

変数名=数値;	**例)** TI1=1;
変数名='文字列';	**例)** I1='1.MALE ';

(b)同一オブザーベーションの変数値の演算値

同一オブザーベーション内での変数値の演算値を与える場合には、具体的には変数名を使った数式の形で示せばよい。

変数名=数式;	**例)** SCODE=KC*10+SC;

という形式になる。そうすると各オブザーベーションの変数値を使って演算が行われ、その演算結果の値が変数に代入される。例では、各オブザーベーションで、変数 KC, SC の値を使って演算が行われ、その演算結果の値が変数 SCODE に代入されることになる。つまり、会社コードを十の位に、会社ごとに付けられた組織単位コードを一の位にもつ、全会社を通して使える 2 桁の「組織単位コード」を与えているのである。

算術演算子の優先順位は

1. べき乗**
2. 乗算*、除算/
3. 加算+、減算-

ということになっていて、これはわれわれが計算するときにしたがっている通常の優先順位と同じである。

(3)変数値による条件別の処理

いつも単純に同じ処理を繰り返すのではなく、変数値によって条件別に処理を選択することもできる。これには SELECT/WHEN 文が用いられる。SELECT/WHEN 文では、WHEN で条件を示し、その条件が真のとき処理が実行される。処理 3 は実際にはなくても、OTHERWISE は必ず入れておくことに注意してほしい。

```
SELECT;
WHEN(条件式 1)処理 1;
WHEN(条件式 2)処理 2; ……
OTHERWISE[処理 3];
END;
```

例)

```
SELECT;
   WHEN(TI1=1) I1='1.MALE ';
   WHEN(TI1=2) I1='2.FEMALE';
   OTHERWISE I1=' ';
   END;
```

このプログラム例は変数値ラベルの定義をしているものである。プログラム例の中で、TI1=1 とか TI1=2 とかあるのは、値の代入式の意味ではなくて、条件を示す論理式である。この例のように、条件式は次のような演算子を使って書かれた真偽の判定できる論理式である。

比較演算子	EQ	NE	GT	GE	LT	LE
または	=		>	>=	<	<=
論理演算子	AND	OR	NOT			
または	&		|		

しかし、先ほどのプログラム例のように、条件式がいずれも同一の変数に関するもので、その変数が特定の値と等しいことを意味しているような場合には、やや煩わしいので、次のような形式で書くことも許されている。

```
SELECT(変数名);
WHEN(変数値 1)処理 1;
WHEN(変数値 2)処理 2; ……
OTHERWISE[処理 3];
END;
```

例)

```
SELECT(TI1);
   WHEN( 1 ) I1='1.MALE ';
   WHEN( 2 ) I1='2.FEMALE';
   OTHERWISE I1=' ';
   END;
```

この SELECT/WHEN 文の処理はこのような変数値の代入だけではなく、様々な処理が可能である。具体的に多く用いられる処理のケースとしては次のようなものがある。

(a)変数値の代入

もっとも一般的でよく使われる。既に見てきた変数値ラベルの定義もこれにあたる。ただし、注意しなければならないのは、文字変数の値は、出力の表中に表示されるときには昇順に並べられるということである。それために、英字から始まる変数値を文字データとして与え、ラベル代わりに使用する際には辞書順序となってしまう。それを回避するためには、先ほどのプログラム例にあった '1.MALE', '2.FEMALE' のように、最初に数字を入れると良い。

また第 3 章でも述べる階級の定義も変数値ラベルの定義と同様に次の例のように行う。

例)

```
SELECT;
   WHEN (20<=I2<25) LI2='20-24';
   WHEN (25<=I2<30) LI2='25-29';
   WHEN (30<=I2<35) LI2='30-34';
   WHEN (35<=I2<40) LI2='35-39';
   WHEN (40<=I2<45) LI2='40-44';
   WHEN (45<=I2<50) LI2='45-49';
   WHEN (50<=I2<55) LI2='50-54';
   WHEN (55<=I2<60) LI2='55-59';
   OTHERWISE LI2=' ';
   END;
```

階級用の文字変数に与える文字データは、同一文字変数であれば、空白を入れてでも同じ文字数にしておいた方がトラブルを避けられる。出力の表中に表示されるときには、8 文字(バイト)ずつで区切られて改行した形で表示されるので、そのことを念頭に置いて文字データを与える。

また変数値ラベルは 10 文字(バイト)以下に抑えた方が、PROC でトラブルが起きない。10 文字を越えた分は、クロス表などでは無視される。

(b)オブザーベーションの削除

```
DELETE;
```

SELECT/WHEN 文の処理として、オブザーベーションの削除を用いると、条件にあったオブザーベーションだけを残して統計処理をすることが可能になる。例えば、調査対象全体をいくつかのサブグループに分けてそれぞれの特性値を求めるような際には必要となる。ただし、この DELETE 文を永久 SAS データ・セットに用いると、オブザーベーションは永久に消えてしまうので、サブグループごとの統計をとるためには一時 SAS データ・セットについて DELETE 文を使っていることを確認しながら使用した方がよい((6)の(b)も参照のこと)。

(c)DO グループ

DO グループは、DO;と END;ではさまれた SAS 文群で、SAS では DO-END 文で DO グループを一つの SAS 文のように扱う。実は SELECT/WHEN 文で WHEN 文に続いて書かれる処理は、形式的には一つの処理しか書けないようになっている。そこで、複数の SAS 文を実行させたいときに、DO グループの形をとって、形式的に一つの処理として収めてしまおうというわけである。次の例では、下線部が DO グループになっている。

```
DO; SAS 文群  END;
```

例)

```
SELECT;
   WHEN(KC=1) KCODE='1.A ';
   WHEN(KC=2) KCODE='2.B ';
   OTHERWISE DO; KCODE=' ';SC=.; KC=.; IC=.; END;
   END;
```

(d)SELECT/WHEN 文

これは SELECT/WHEN 文を入れ子の形で使用することが可能だということである。

(4)異なる変数に対する同じ処理の繰り返し

異なる変数に、何度も同じ処理を繰り返す場合は、いちいちそれを書いていては、プログラムもいたずらに長くなってしまうし、分かりにくくなってしまう。そうした場合には、次のように ARRAY 文を用いると便利である。

```
ARRAY  変数名{n}  変数名の並び;
DO I=1 TO n;

        変数名{I}を含んだ処理

END;
```

例)

```
ARRAY TYN{15} TII1-TII15;
ARRAY YN{15} $ II1- II15;
DO I=1 TO 15;
   SELECT; WHEN(TYN{I}=1) YN{I}='1.YES';
           WHEN(TYN{I}=2) YN{I}='2.NO ';
           OTHERWISE YN{I}=' ';END;
END;
```

ここで n には、変数名の並びにある変数の個数を数値として入れる。プログラム例の ARRAY 文の 2 行目にもあるように、変数に文字定数を代入する際、ARRAY 文の中で、その変数の前に文字変数であることを示す"$"をつけることを忘れないようにする。

(5)通常型 SAS プログラム例

それでは、以上に述べてきたような様々な SAS 文を駆使して、データの入力や変数名、変数ラベルの設定を行い、永久 SAS データ・セットを作成するとともに、単純集計を行う通常型 SAS プログラムの例を、まず PC 版 SAS 用に示すことにしよう。

```
*-------------------------------------------------------;
*-------------------------------------------------------;
*JPCP1.SAS ;
* Permanent SAS data set JPC.SSD will be made. ;
*-------------------------------------------------------;
*-------------------------------------------------------;
LIBNAME SAVE '¥MYDIR';                                    (i)
DATA SAVE.JPC;
*-------------------------------------------------------;
* Raw data will be read from MS-DOS file JPCDATA.SAS. ;
*-------------------------------------------------------;
INFILE 'JPCDATA.SAS';                                     (ii)
INPUT (KC SC IC)(2*1. 2.)
(TI1-TI4 TII1-TII15 TIII1-TIII15)(34*1.)
/+4 (TIV1-TIV15 TV1-TV15 TVI1-TVI15)(45*1.);
*-------------------------------------------------------;
* KCODE SCODE will be made. ;
*-------------------------------------------------------;
SELECT;
WHEN(KC=1) KCODE='1.A ER ';
WHEN(KC=2) KCODE='2.B STORE ';
WHEN(KC=3) KCODE='3.C RAIL ';
WHEN(KC=4) KCODE='4.D EP ';
WHEN(KC=5) KCODE='5.E SER ';
WHEN(KC=6) KCODE='6.F BANK ';
WHEN(KC=7) KCODE='7.G HI ';
WHEN(KC=8) KCODE='8.H OIL ';
OTHERWISE DO; KCODE=' '; SC=.; IC=.; END;
END;
SCODE=KC*10+SC;
*-------------------------------------------------------;
* Classification data will be labeled. ;
*-------------------------------------------------------;
SELECT;
WHEN(TI1=1) I1='1.MALE ';
WHEN(TI1=2) I1='2.FEMALE';
OTHERWISE I1=' ';
END;
SELECT;
WHEN(TI2=1) I2='1.20-24';
```

```
WHEN(TI2=2) I2='2.25-29';
WHEN(TI2=3) I2='3.30-34';
WHEN(TI2=4) I2='4.35-39';
WHEN(TI2=5) I2='5.40-44';
WHEN(TI2=6) I2='6.45-49';
WHEN(TI2=7) I2='7.50-54';
WHEN(TI2=8) I2='8.55-60';
OTHERWISE I2=' ';
END;
SELECT;
WHEN(TI3=1) I3='1.STAFF';
WHEN(TI3=2) I3='2.E & P';
WHEN(TI3=3) I3='3.R & D';
WHEN(TI3=4) I3='4.SALES';
OTHERWISE I3=' ';
END;
SELECT;
WHEN(TI4=1) I4='1.GEN ';
WHEN(TI4=2) I4='2.MGR ';
WHEN(TI4=3) I4='3.HEAD ';
WHEN(TI4=4) I4='4.OTHERS';
OTHERWISE I4=' ';
END;
ARRAY TYN{75} TII1--TVI15;
ARRAY YN{75} $ II1-II15 III1-III15 IV1-IV15 V1-V15 VI1-VI15;
DO I=1 TO 75;
SELECT; WHEN(TYN{I}=1) DO; TYN{I}=1; YN{I}='1.YES'; END;
WHEN(TYN{I}=2) DO; TYN{I}=0; YN{I}='2.NO '; END;
OTHERWISE DO; TYN{I}=.; YN{I}=' '; END; END;
END;
LABEL
KCODE='COMPANY'
SCODE='UNIT'
I1='SEX'
I2='AGE'
I3='OCCUPATION'
I4='RANK';
*-----------------------------------------------------;
* Check the labels and the frequencies. ;
*-----------------------------------------------------;
PROC FREQ;
OPTIONS NOCENTER;
RUN;
```

このプログラムは MS-DOS ファイル JPCP1.SAS に保存しておこう。これが、第 6 章で調査の実例としてあげる「組織活性化のための従業員意識調査」の際に単純集計と永久 SAS データ・セット生成に実際に用いられた SAS プログラムである。

先ほどの基本型 SAS プログラムと同じ、データの入った MS-DOS ファイル JPCDATA.SAS から、永久 SAS データ・セットが生成される。永久 SAS データ・セット名に同じ JPC を使っているので、MS-DOS ファイル JPC.SSD は重ね書きされ、内容は更新されることになる。処理時間としては、DATA ステップで 6 分強、PROC ステップでは約 9 分を処理に要する。したがって、全体では 15 分強ということになる。この結果、生成される永久 SAS データ・セット JPC.SSD は 1,211,212 バイトつまり約 1.2MB の大きさの MS-DOS ファイルとなる。基本型 SAS プログラムの 82 変数に対して、この通常型 SAS プログラムでは 164 変数が定義されることになるので、基本型 SAS プログラムの約 2 倍のファイル容量を必要とすることになる。

CMS 版 SAS 用のプログラムは、この PC 版 SAS 用のプログラムのうちのわずか 2 行、(i)と(ii)を次のように置き換えればよい。

i. X FI IN1 DISK JPCDATA SAS;
ii. INFILE IN1;

(6)永久 SAS データ・セットの再利用

永久 SAS データ・セットは一度作成してしまえば、もうデータの読み込みや変数の定義、データの加工を繰り返す必要はなく、あとはそれを別のプログラムの中で呼び出してやるだけで、SAS はその SAS データ・セットを参照して、PROC ステップを行うことができる。したがって、既存の永久 SAS データ・セットを使う SAS プログラムは、簡潔かつ簡単なもので済ませることができる。次の第 3 章以降で取り上げる SAS プログラムは、そのために、ごく簡単なものになる。ここでは永久 SAS データ・セットを繰り返して再利用する際の一般的な方法についてまとめておこう。永久 SAS データ・セットの呼び出しは次のように行う。

(a)永久 SAS データ・セットの読み込みと書き出し

PC 版 SAS で、永久 SAS データ・セット名 1 の永久 SAS データ・セットの読み込みを行い、また新たに定義した変数や、新たに加工されたデータを含めて、永久 SAS データ・セット名 2 の永久 SAS データ・セットに書き出すには次のようにすればよい。

PC 版 SAS

```
LIBNAME libname '￥パス名';
DATA libname.永久 SAS データ・セット名 2 本体;
SET libname.永久 SAS データ・セット名 1 本体;

        データの加工・処理
```

例)

```
LIBNAME SAVE '￥MYDIR';
DATA SAVE.JPC;
      SET SAVE.JPC;
SELECT; WHEN(TI1=1) I1=1;
            WHEN(TI1=2) I1=0;
            OTHERWISE I1=.; END;
END;
```

こうすることで、SAS データ・セットは永久 SAS データ・セットとなり、DATA ステップで行った加工処理は、そのまま MS-DOS ファイルとして保管されることになる。このプログラム例のように二つの永久 SAS データ・セット名を同じにすると、プログラム実行の都度、永久 SAS データ・セットは書き換えられ、更新されることになる。DATA 文の SAS デ

ータ・セット名 2 を永久 SAS データ・セット名 1 と別のものにすれば、新たにもう一つ別の永久 SAS データ・セットが自動的に生成される。

CMS 版 SAS では、PC 版 SAS のように libname の指定がいらない。つまり CMS 版 SAS では 1 行目の LIBNAME 文はつけずに、2 行目の DATA 文から始めるわけだが、その際、PC 版 SAS の DATA 文中の libname に相当するものとして、CMS 版 SAS では永久 SAS データ・セット名の ft を指定することになる。そのため、永久 SAS データ・セット名＝ft.fn と ft が先にくることに注意する。CMS 版 SAS では、このように指定するだけで，SAS データ・セットは永久 SAS データ・セットとなり、DATA ステップで行った加工処理は、そのまま CMS ファイルとして保管されることになる。したがって、形式的には、プログラム例で、1 行目を削除したものが、そのまま CMS 版 SAS のプログラム例となる。すなわち、永久 SAS データ・セット名 1 の永久 SAS データ・セットの読み込みを行い、また永久 SAS データ・セット名 2 の永久 SAS データ・セットに書き出すには次のようにすればよい。

CMS 版 SAS
DATA 永久 SAS データ・セット名 2; SET 永久 SAS データ・セット名 1; データの加工・処理

例)

```
DATA SAVE.JPC;
   SET SAVE.JPC;
SELECT; WHEN(TI1=1) I1=1;
        WHEN(TI1=2) I1=0;
        OTHERWISE I1=.; END;
END;
```

ここで、プログラム例では ft を SAVE にしているが、これは他のものにしてもいい。ただし、もし ft を WORK にすると、一時 SAS データ・セットとして認識されてしまうので、それ以外のものにすること。

(b)永久 SAS データ・セットの読み込み

永久 SAS データ・セットの読み込みだけを行い、永久 SAS データ・セットを書き直したり、書き出したりするようなことはしたくないときには、次のように DATA 文で一時 SAS データ・セットを指定して、その中に、希望する永久 SAS データ・セットを読み込んでくる形をとればよい。(a)との違いは、PC 版 SAS では DATA 文の中の libname の有無だけ、CMS 版 SAS では同じく DATA 文の SAS データ・セット名の ft の有無だけである。

PC 版 SAS

```
LIBNAME libname 'ディレクトリ名';
DATA 一時 SAS データ・セット名;
SET libname.永久 SAS データ・セット名本体;

        データの加工・処理
```

例)

```
LIBNAME SAVE '￥MYDIR';
DATA            JPC;
     SET SAVE.JPC;
SELECT; WHEN(TI1=1) I1=1;
          OTHERWISE   DELETE;
          END;
```

CMS 版 SAS

```
DATA 一時 SAS データ・セット名;
SET  永久 SAS データ・セット名;

        データの加工・処理
```

例)

```
DATA            JPC;
     SET SAVE.JPC;
SELECT; WHEN(TI1=1) I1=1;
          OTHERWISE   DELETE;
END;
```

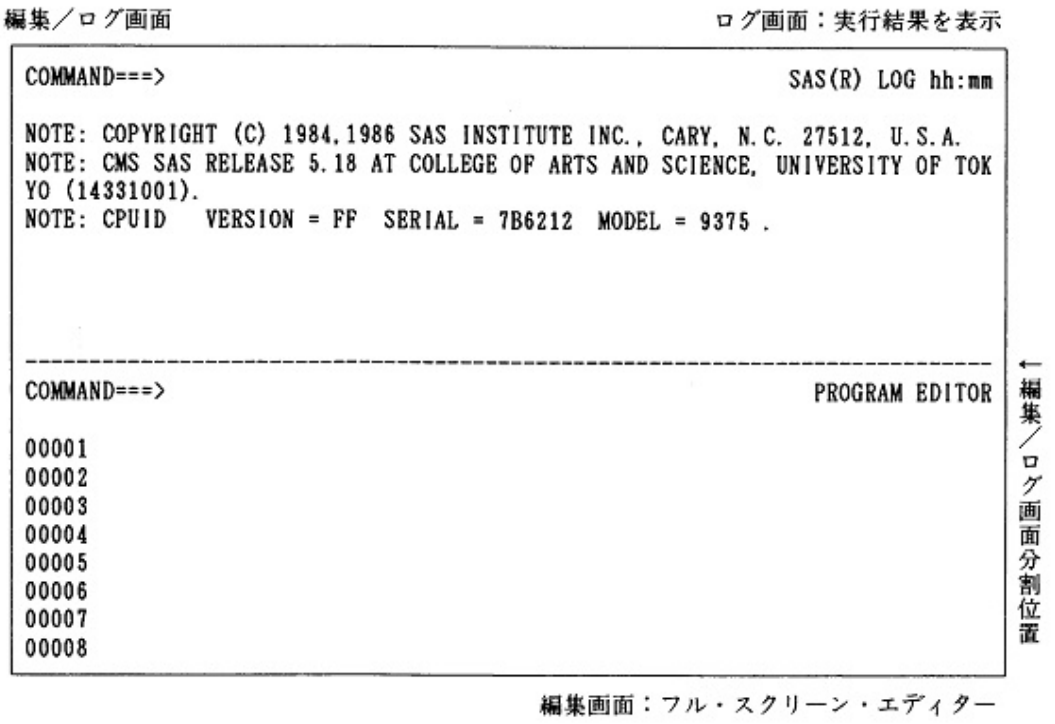
付．CMS 版 SAS リリース 5.18 の基本的な使用方法

この章では SAS 第 6 版の使用方法を解説してきたが、最新版の第 6 版がリリースされてまだ間もないし、現在でも IBM のメインフレームと IBM 互換国産メインフレームの一部では、SAS 第 5 版の最新版であったリリース 5.18 がそのまま使用されているところもあるようなので、この節では、CMS 版 SAS リリース 5.18 の基本的な使用方法について概説しておこう。もっとも、CMS 版 SAS の第 6 版との相違点は、ディスプレイ・マネージャ・システムの使用方法に関するものが中心である。

(1)SAS プログラムの編集／ログ画面

SAS を起動すると、プログラム作成、実行のためのシステムが起動されて、CMS 版 SAS の第 5 版では、端末のディスプレイ画面に図 2.12 のような編集／ログ画面が表示される。これは構成的には CMS 版 SAS の第 6 版のウィンドウと基本的には同じである。

図 2.12　CMS 版 SAS リリース 5.18 のディスプレイ・マネージャ・システムの画面



CMS 版 SAS 第 6 版と同様に、keys コマンドが使えるので、この編集／ログ画面の PF キー(ファンクション・キー)の設定は、自分で確認すること。CMS 版 SAS の第 6 版との違いを次に挙げておく。その他についてはあまり違いはない。

1. CMS 版 SAS5.18 では、zoom コマンドが存在しない。その代わり、(PF2)キー(=split)を使って、編集／ログ画面分割位置をカーソル行に変更する方法がとられる。
2. うまく SAS プログラムの実行に成功すると、出力があれば出力画面に変わるという点は CMS 版 SAS の第 6 版と同じであるが、通常は出力画面にコマンド行は表示されないので、もし PF キーを使わずにコマンドを使う際には、まず(PF2)キー(=COMMAND)を押して、コマンド行を設定してからコマンドを入力しなければならない。
3. CMS 版 SAS 第 6 版では各ウィンドウで使えた file コマンドが存在しない。唯一、編集画面でのみ SAS プログラムやデータを CMS ファイルに保存することができるが、そのときには、

   COMMAND===> SAVE fn(実行)

   と入力する。ただし、このときファイル・ネーム fn は' 'で囲まないので注意がいる。こうすることで、編集画面上にある SAS プログラムが CMS ファイル "fn SAS" に書き出され、保存される。
4. 既に CMS ファイルとして保存しておいた SAS プログラムやデータを編集する場合には、ファイル・タイプ ft が SAS である CMS ファイルについてのみ、ファイル・ネーム fn を

   Command ===> INClude fn(実行)

   と入力することで、画面上に読み込むことができる。やはりこのときファイル・ネーム fn は' 'で囲まない。SAS プログラムだけではなく、ファイル・タイプが SAS であれば、固定長のあらゆる CMS ファイルが読み込み可能である。

(2)CMS ファイルへの出力

出力画面で file コマンドが使えないことから、出力を CMS ファイルに保存したい場合には、直接、出力を CMS ファイルに対して行うしかない。ここに述べる方法は、CMS 版 SAS 第 6 版でもそのまま使える(もっともその必要はないだろうが……)。

CMS では、データ入出力機器の既定値は端末になっているので、CMS ファイルに出力する場合にも、CMS ファイルからデータを読み込む際と同様に、"X CMS コマンド" で設定を変更し、ddname を指定することが必要になる。具体的には、CMS ファイル(プログラム例では、fileid=JPC LST [A])へ結果を出力したいと思っている PROC 文の前へ、次の 2 行を挿入するとよい。

```
X FIledef ddname DISK fileid;
PROC PRINTTO UNIT=ddname;
```

例)

```
X FI 18 DISK JPC LST;
PROC PRINTTO UNIT=18;
```

"UNIT=" には装置番号しか使えないので、筆者は ddname として 18 をよく使う(11〜15 は使えない)。同じ fileid の CMS ファイルが既に存在している場合には、重ね書きされていくので注意がいる。また、この PROC 文の後で出力機器の設定を端末に戻すには次の一文を入れる。

```
PROC PRINTTO;
```

この文がなければ同 SAS セッション中は設定変更が生き続けることになる。

これらの SAS 文は第 5 節(5)の基本型 SAS プログラムには、次の下線部のように挿入される。

```
X FI IN1 DISK JPCDATA SAS;
DATA SAVE.JPC;
INFILE IN1;
INPUT (KC SC IC)(2*1. 2.)(TI1 -TI4 TII1-TII15 TIII1-TIII15)(34*1.)
        /+4                    (TIV1-TIV15 TV1 -TV15 TVI1 -TVI15 )(45*1.);
X FI 18 DISK JPC LST;
PROC PRINTTO UNIT=18;
PROC FREQ;
        OPTIONS NOCENTER;
PROC PRINTTO;
RUN;
```

演習問題

2.1 基本型 SAS プログラムの実行　第 5 節(5)に示される通りに SAS プログラムを作成、実行して、単純集計を行ってみよ。その際のデータは図 2.8 に示されている 20 人分をそのまま用いて行え。

2.2 通常型 SAS プログラムの理解　第 6 節(5)に例示されている通常型 SAS プログラムがどんな処理を行っているのか、日本語で説明せよ。ヒントは第 5 節、第 6 節の各所に分散している。

2.3 通常型 SAS プログラムの実行　第 6 節(5)に例示されている通常型 SAS プログラムを自分で入力し、実際に単純集計を行ってみよ。その結果を第 6 章の資料 F の形式を参考にしながら、資料 B の質問調査票に手書きで書き込み、単純集計の結果を報告せよ。

2.4 単純集計の利活用　演習問題 2.1 (または 2.3)で作成した単純集計に基づいて、次のことについて整理してまとめよ。

1. 調査対象の基本特性項目上の特徴。
2. 単純集計を読んで、特徴的なことを 3 つ挙げ、考えられる理由も述べよ。

第 3 章　データの記述と平均

章目次

1. はじめに
2. 図・表による整理
3. 平均による要約
4. 分散による要約
5. 2 値質的データの平均と分散
6. SAS による平均・分散の計算
7. 正規分布と標準得点
8. 2 群の平均値の比較
9. k 群の平均値の差の検定
演習問題

1. はじめに

この章からは、調査によって得られたデータを具体的にどのように統計処理するのかを説明することにしよう。まず調査によって得られたデータについて考えてみよう。例えば「組織活性化のための従業員意識調査」(詳細については第 6 章で解説する)を行えば、個人単位の質問調査票が回収できる。第 2 章第 5 節の図 2.7 には、その調査によって得られた回答の一部 20 人分がデータとして表示されているが、いまその最初の 8 人、すなわち組織単位コード 57 の「E サービス㈱の所内保全サービス部門」について、性別、職種、職位を図に表示されているコードのままで並べれば、表 3.1(A)のようなデータが得られたことになる。これでは人間にはピンとこないので、元のカテゴリーの形に戻して表示すると表 3.1(B)のようになる。

表 3.1 をより一般化して、n 人のデータとして書き表せば、表 3.2 のようになる。ここで、x, y, z はそれぞれ変数(variable)あるいは変量(variate)と呼ばれる。(ただし、統計の分野では因子分析、主成分分析、判別分析等を総称して多変量解析(multivariate analysis)と呼ぶが、これを多"変数"解析とはいわないので注意がいる。) それに対して、x_i, y_i, z_iは観測値、測定値と呼ばれる。観測値は、変数と区別するために添字を付けて表すことにしよう。n はデータのサイズを表している。

表 3.1　E サービス㈱の所内保全サービス部門の性別、職種、職位データ

(A)コード表示

個人番号	性別	職種	職位
1	1	1	4
2	1	1	4
3	1	1	3
4	1	1	4
5	1	1	2
6	1	2	3
7	1	1	2
8	1	1	2

(B)カテゴリー表示

個人番号	性別	職種	職位
1	男	事務	一般
2	男	事務	一般
3	男	事務	係長
4	男	事務	一般
5	男	事務	課長
6	男	技術	係長
7	男	事務	課長
8	男	事務	課長

注）表(B)の中の「事務」は「事務・スタッフ」、「技術」は「技術・製造」、「係長」は「係長・主任クラス」、「課長」は「課長クラス」をそれぞれ省略したもの。

表 3.2　一般の 3 変数の n 人データ

	x	y	z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
:	:	:	:
i	x_i	y_i	z_i
:	:	:	:
n	x_n	y_n	z_n

例えば、性別だけのデータ、実はこの 8 人は全員が男性なのだが、

男, 男, 男, 男, 男, 男, 男, 男

のような一つの変数の観測値からなるデータ、$x_1, x_2, ...$ は 1 変数データあるいは 1 次元データ(1-dimensional data)といわれる。それに対して、例えば、性別と職種からなるデータ、

(男, 事務), (男, 事務), (男, 事務), (男, 事務), (男, 事務), (男, 技術), (男, 事務), (男, 事務)

のように二つの変数の観測値からなるデータ、$(x_1, y_1), (x_2, y_2), ...$ は 2 変数データあるいは 2 次元データ(2-dimensional data)といわれる。この 2 変数の場合も含めて、2 変数以上の観測値からなるデータ、例えば 3 変数からなるデータ、

(男, 事務, 一般), (男, 事務, 一般), (男, 事務, 係長), (男, 事務, 一般),
(男, 事務, 課長), (男, 技術, 係長), (男, 事務, 課長), (男, 事務, 課長)

一般的には、$(x_1, y_1, z_1),(x_2, y_2, z_2),...$ などは、多変数データあるいは多次元データ(multidimensional data)と呼ばれる。

実は、通常得られる調査データは、このように各変数についてのみ見ることで 1 変数データとして扱うこともできるし、複数の変数の組を見ることで、多変数データとして扱うこともできる。1 変数データとして扱うのであれば、各変数単独での特性を調べることになるし、多変数データとして扱うのであれば、変数間の関係を調べることになるが、いず

れにせよ、1 変数あるいは多変数のデータを整理・要約して、標本調査ならば標本の、全数調査ならば母集団の集団としての特徴を記述することが記述統計学(descriptive statistics)の役割である。しかし、記述統計学という名前がついているかどうかにかかわらず、データを整理・要約することは重要なことである。

この章では、1 変数データの整理・要約について取り扱うが、データから事実を探り、知るためにはこのステップは欠かせない。統計的に分析をする際には、ややもするといきなり高度な手法に走りがちであるが、われわれが「分析」と呼んでいる過程は、このデータの整理・要約、特に 1 変数のデータの整理・要約をきちんとしておかないと、しばしば無意味になってしまうことがある。しかも、このデータの整理・要約の過程自体も、

- 《ステップ 1》図・表によるデータの整理
- 《ステップ 2》平均、分散などの数値によるデータの要約

のようなステップを踏んで行っておかないと、後になって後悔することになる。つまり、ステップ 1 のデータの整理を飛ばして、いきなりステップ 2 から始めてデータを要約することは、一見早いようだが、結局は無駄な努力に終ることが多い。経験的には、説得的な事実に肉薄するという意味では、1 変数・多変数のデータの整理・要約で十分なケースが多いものである。

第 1 章で述べたように、名義尺度、順序尺度にもとづくデータのことを質的(定性的)データ(qualitative data)といい、間隔尺度、比率尺度にもとづくデータのことを量的(定量的)データ(quantitative data)という。質的データ、量的データの区別とある意味では混同されがちなものに、離散変数と連続変数の区別があるので、一応説明しておこう。離散(discrete)変数とは、とりうる値が高々可算個、つまり有限個または可算無限個、例えば、とりうる値の集合が、整数、有理数の変数である。それに対して、連続(continuous)変数とは、とりうる値が可算個ではない、例えば、とりうる値の集合が実数全体であるような変数である。実は、ここでいう「連続」は通常の数学的意味で使われているのではなく、「離散」と対比して使用されていると考えた方がよい。

離散変数と連続変数の区別は数理統計学、確率論の上では意味のある区分だが、実際のデータ収集、統計処理上は、厳密に区別することにはあまり意味がない。なぜなら、まず、実際の観測値は正確には常に離散的だからである。理屈の上では連続変数であっても、その観測値は測定機器の精度の限界、分析の目的により、適当な桁数の有効数字に丸められ、離散的となるのである。しかも、逆に、たとえ理屈の上では離散変数の場合でも、とりうる値が多いときには、「試験の点数」のように、連続変数とみなして処理することも多いのである。

そこで、本書では、質的データと量的データの基本的な区別に注意しながら、説明を進めていくことにしよう。

2. 図・表による整理

統計データ処理とは結局、上手に図・表にまとめることである。と言ったら言い過ぎだろうか。実感としては、統計データの本来もっている説得力を上手に引き出す図・表による整理が出来るのであれば、それ以上の高度な統計手法は不要なものに思われる。少なくとも、図・表にまとめて、全体を記述した上で、はじめてそれから先の処理、分析に進むことができるということを肝に命じておくべきであろう。この章でこれから説明していく

が、平均、分散といった非常によく用いられる基本的な統計量ですら、分布の形がなめらかで整っていればこそ、用いることも可能になるのである。

(1)表による記述

表による記述は、質的データに対しても量的データに対しても行われるが、量的データでは、まずデータの値をいくつかの段階に分けた「階級」(class)と呼ばれるものを設定する。この階級は質的データの「カテゴリー」に相当するものであり、そのため形式的には、質的データも量的データもその度数分布表(frequency table)、累積度数分布表(cumulative frequency table)は同じものになる。ここで、度数とは、当該カテゴリーもしくは階級に該当するオブザーベーション数を計数したものである。累積度数とは、量的データの場合には度数を下の階級から順に積み上げたときの度数である。質的データの場合には、カテゴリーの並んでいる順に積み上げたときの度数になる。

実際、質的データの表、量的データの表は、例えば「組織活性化のための従業員意識調査」のデータを用いると、それぞれ表 3.3(A)(B)のようになり、形式的には同じものとなる。

表 3.3　度数分布表・累積度数分布表の例

(A)質的データ(性別)の度数分布表・累積度数分布表

カテゴリー	度数	相対度数	累積度数	累積相対度数
男	765	87.2	765	87.2
女	112	12.8	877	100.0
計	877	100.0	-	-

(B)量的データ（年齢）の度数分布表・累積度数分布表

階級	度数	相対度数	累積度数	累積相対度数
20～24 歳	60	6.7	60	6.7
25～29 歳	169	18.8	229	25.5
30～34 歳	175	19.5	404	45.0
35～39 歳	153	17.1	557	62.1
40～44 歳	159	17.7	716	79.8
45～49 歳	99	11.0	815	90.9
50～54 歳	62	6.9	877	97.8
55～60 歳	20	2.2	897	100.0
計	897	100.0	-	-

表 3.3 の中の各列を度数分布、相対度数分布、累積度数分布、累積相対度数分布ということもある。「相対」とついているのは、全体に対する相対的な百分率を表しているからである。相対度数は、集団の大きさ n が異なる複数の集団の分布を比較するときに有効である。SAS を使って集計処理した場合にも、こうした基本的な度数(frequency)、相対度数(relative frequency)、累積度数(cumulative frequency)、累積相対度数(relative cumulative

frequency)は出力されることになる。いずれにせよ、(A)(B)両者が形式的には同じものであることがわかるだろう。質的データの表、量的データの表をより一般的に対照して示せば、表 3.4 のようになる。

表 3.4　度数分布表・累積度数分布表

(A)質的データ

カテゴリー	度数	相対度数	累積度数	累積相対度数
$\boldsymbol{A_1}$	f_1	$f_1/n\times100$	F_1	$F_1/n\times100$
$\boldsymbol{A_2}$	f_2	$f_2/n\times100$	F_2	$F_2/n\times100$
:	:	:	:	:
$\boldsymbol{A_i}$	f_i	$f_i/n\times100$	F_i	$F_i/n\times100$
:	:	:	:	:
$\boldsymbol{A_k}$	f_k	$f_k/n\times100$	F_k	$F_k/n\times100$

(B)量的データ

階　級	度数	相対度数	累積度数	累積相対度数
$\boldsymbol{a_1\sim b_1}$	f_1	$f_1/n\times100$	F_1	$F_1/n\times100$
$\boldsymbol{a_2\sim b_2}$	f_2	$f_2/n\times100$	F_2	$F_2/n\times100$
:	:	:	:	:
$\boldsymbol{a_i\sim b_i}$	f_i	$f_i/n\times100$	F_i	$F_i/n\times100$
:	:	:	:	:
$\boldsymbol{a_k\sim b_k}$	f_k	$f_k/n\times100$	F_k	$F_k/n\times100$

ここで　$a_1<b_1\leqq a_2<b_2\leqq\cdots\cdots\leqq a_i<b_i\leqq\cdots\cdots\leqq a_k<b_k$

$F_i=f_1+f_2+\cdots+f_i=\Sigma_{j=1}^{i} f_j$, $i=1, 2, ..., k$

こうした表からもわかるように、累積度数分布は度数分布から簡単に計算できるし、逆に累積度数分布から度数分布を算出することもできる。このことは、度数分布と累積度数分布とが同じだけの情報を含んでいることを意味している。したがって、どちらか一方を示せば十分なわけだが、通常は、度数および相対度数の分布だけ示せばよい。ここでは触れないが、累積度数および累積相対度数はやや特殊な用途に用いられる。

ここで、「分布」という用語の使用についてやや注意を述べておくと、母集団の分布の形を正規分布などで表現することがあるが、これは相対度数分布を確率分布に見立てて用いるのである。この場合、「母集団分布」は正確には相対度数分布のことであって、確率分布を意味するものではない。「確率」分布は、標本抽出、正確には無作為抽出という確率的装置を通して、標本分布にはじめて現れるのであって、母集団には論理的には存在しない概念である。ちなみに、統計学でいう「標本分布」とは標本における平均などの統計量の分布のことであって、これは確率分布である。標本における記述統計としての度数分布のことではないので、この点にも注意がいる。

(2)SAS による度数分布表の集計

度数分布表、相対度数分布表、累積分布表、累積相対度数分布表を求めることは、SAS を使えば簡単である。単純集計のときと同様に FREQ プロシジャを使えばよいのである。実際、質的データであれば、単純集計とは度数分布表を作ることと同じことになる。第 2 章で作成した永久 SAS データ・セット(PC 版 SAS のプログラム例では JPC.SSD、CMS 版 SAS のプログラム例では JPC SAVE)を利用すれば、度数分布表作成のための SAS プログラムはわずかに数行で済む。

これには 2 通りの方法があって、まず最初の方法としては、SAS プログラムの基本形である DATA ステップと PROC ステップの 2 ステップ構成にこだわる方法で、このときは次のようなプログラムになる。

PC 版 SAS

```
LIBNAME libname '￥パス名';
DATA libname.永久 SAS データ・セット名本体;
SET libname.永久 SAS データ・セット名本体;
PROC FREQ;
TABLES  変数名の並び;
[OPTIONS NOCENTER;]
RUN;
```

例)

```
LIBNAME SAVE '￥MYDIR';
DATA SAVE.JPC;
   SET SAVE.JPC;
PROC FREQ;
        TABLES I2;
        OPTIONS NOCENTER;
RUN;
```

CMS 版 SAS

```
DATA  永久 SAS データ・セット名;
SET  永久 SAS データ・セット名;
PROC FREQ;
TABLES  変数名の並び;
[OPTIONS NOCENTER;]
RUN;
```

例)

```
DATA SAVE.JPC;
   SET SAVE.JPC;
PROC FREQ;
        TABLES I2;
        OPTIONS NOCENTER;
RUN;
```

CMS 版 SAS のプログラムは、形式的には PC 版 SAS のプログラムの 1 行目を削除したものになる。このプログラムのように DATA ステップをつけると、内容的には何の変更もない全く同じ永久 SAS データ・セットを実質的にまた作り直していることになるので、その分だけ時間がかかる。メインフレームでは気にならないが、パーソナル・コンピュータでは、筆者の実験では DATA ステップに約 2 分ほどかかる。PROC ステップには 20 秒もかかからないので、時間の無駄といえないこともない。

そこで、もう一つの方法としては、DATA ステップをなくして、PROC ステップだけにして、直接永久 SAS データ・セットを参照させる方法がある。この方法では、PROC ステップにかかる時間だけで済むので、20 秒以内に処理が可能である。また DATA ステップがない分だけ、プログラムも簡単にすることができる。それには、次のように PROC FREQ 文の DATA=オプションで永久 SAS データ・セット名を指定してしまうのである。もちろんこの方式では、DATA ステップは不要になる。

PC 版 SAS

```
LIBNAME libname '￥パス名';
PROC FREQ DATA=libname.永久 SAS データ・セット名本体;
TABLES  変数名の並び;
[OPTIONS NOCENTER;]
RUN;
```

例)

```
LIBNAME SAVE '￥MYDIR';
PROC FREQ DATA=SAVE.JPC;
        TABLES I2;
        OPTIONS NOCENTER;
RUN;
```

CMS 版 SAS

```
PROC FREQ DATA=永久 SAS データ・セット名;
TABLES  変数名の並び;
[OPTIONS NOCENTER;]
RUN;
```

例)

```
PROC FREQ DATA=SAVE.JPC;
        TABLES I2;
        OPTIONS NOCENTER;
RUN;
```

さきほどと同様に、CMS 版 SAS のプログラムは、形式的には PC 版 SAS のプログラムの 1 行目を削除したものになる。本書ではこれ以降、特に断らない限り、この後者の方式の PROC 文の DATA=オプションで指定する方法でのみ、SAS プログラムの解説をすることにしよう。

プログラム例の中の変数 I2 は「組織活性化のための従業員意識調査」(詳細は第 6 章)では、年齢を表す変数である。変数 I2 については、これまでの調査経験と事前知識から、既に調査時にプリコーディングしてあるので、実は、単純集計を求めることで、即年齢階層別の度数分布を求めることが出来る。ただし後述するように、通常はどのような階級を設定するのかについては、この度数分布表を作成しながら試行錯誤を重ねて納得のいくような階級を設定しなくてはならない。ここで示した分類は、そうしたこれまでの試行錯誤の一成果であると考えてほしい。

この SAS プログラムを実行させると、どちらのプログラム例でも次のような集計結果が端末の画面に表示されることになる。

実は、この図 3.1 の集計結果から、既に提示した表 3.3(B)が作成されている。ここで、図 3.1 に示した画面の一番下の行には

Frequency Missing = 10

とあるが、これは年齢を表す変数 I2 で欠損値となっている人が 10 人いたことを示している。全体の 1.1% (＝10/(897+10))程度が欠損値だったことになるが、この程度の低水準に抑えられているのであれば、調査データとしては良好なものだといえるだろう。

図 3.1　SAS で求めた度数分布表

OUTPUT
Command ===>

SAS　　13:52 Thursday, January 30, 1992　1

AGE

I2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.20-24	60	6.7	60	6.7
2.25-29	169	18.8	229	25.5
3.30-34	175	19.5	404	45.0
4.35-39	153	17.1	557	62.1
5.40-44	159	17.7	716	79.8
6.45-49	99	11.0	815	90.9
7.50-54	62	6.9	877	97.8
8.55-60	20	2.2	897	100.0

Frequency Missing = 10

ZOOM

↑ 変数値　↑ 度数　↑ 相対度数　↑ 累積度数　↑ 累積相対度数

(3)図による記述

表で記述する場合の度数分布表、累積度数分布表に対応して、量的データ、質的データともに、それぞれ 2 種類の図で記述することができる。

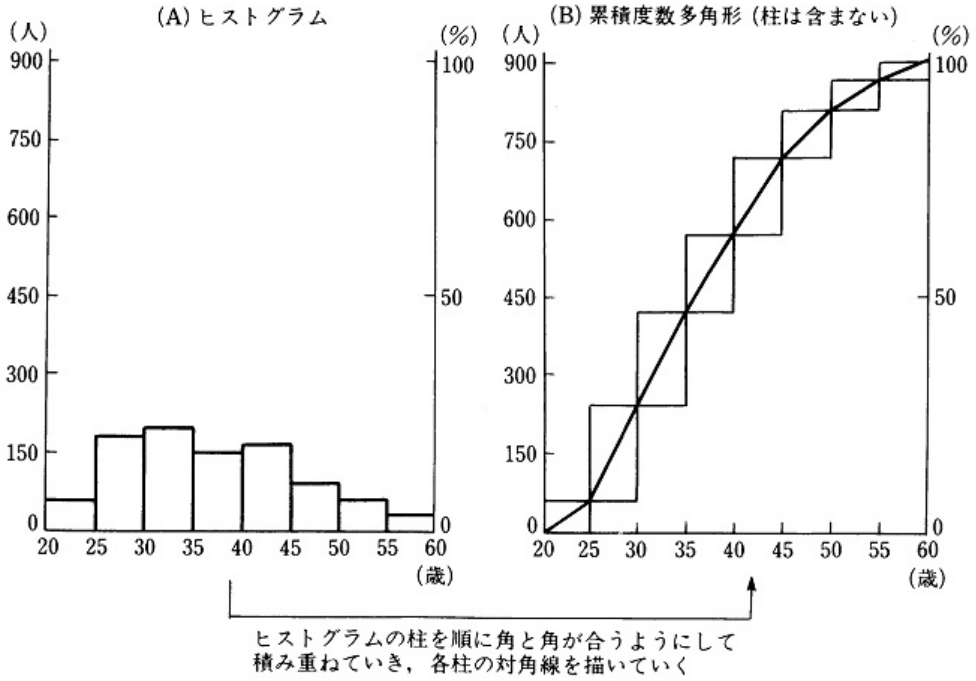
(a)量的データの場合

この場合には、度数分布表に対応してヒストグラム、累積相対分布表に対応して累積度数多角形が描かれる。

1. ヒストグラム(histogram)または柱状図は、長方形をした柱の 面積で度数を示した グラフである。したがって、階級の間隔が等しくないような度数分布からヒストグラムを描くこともできる。この場合、各階級の度数と長方形をした各柱の面積とが対応するようにする。
2. 累積度数多角形(frequency polygon)または累積度数折れ線は、ヒストグラムの柱を順に角と角が合うようにして積み重ねていき、各柱の対角線を描いていく。こうして描いた累積度数折れ線の折れ曲がっているその角の点の横座標は階級の上の境界に対応している。つまり、もし階級で分けない素データをつかった場合であっても、累積度数曲線はこの点を通るはずである。

ヒストグラムと累積度数多角形の例として、表 3.3(B) (あるいは図 3.1)の度数を使って描いてみると、図 3.2 のようになる。このうち、累積度数多角形のグラフでは、説明のためにヒストグラムの柱が重ね合わせて示されているが、累積度数多角形は折れ線の部分だけなので、注意してほしい。

図 3.2　ヒストグラムと累積度数多角形



ところで、この例では、年齢は 60 歳までになっているが、企業のように定年制のあるところは別として、一般的には年齢のような変数の最後の階級は、オープン・エンドの階級、つまり上限がわからない階級となっていることが多い。こうしたオープン・エンドの階級の処置が実はあやしくなってしまう。正確に図を描くことが出来ないのである。ヒストグラムの場合には、多少面積を小さめにして、目分量で記入するしかないし、累積度数多角形の場合には、100%よりやや下のところに目分量で線を結ぶしかないのである。したがって、ヒストグラムや累積度数多角形を正確に描くことを考えているのであれば、オープン・エンドの階級は不用意に作るべきではない。もし作るのであれば、その階級でのデータの平均値、最小値または最大値を注記する心遣いがほしい。

(b)質的データの場合

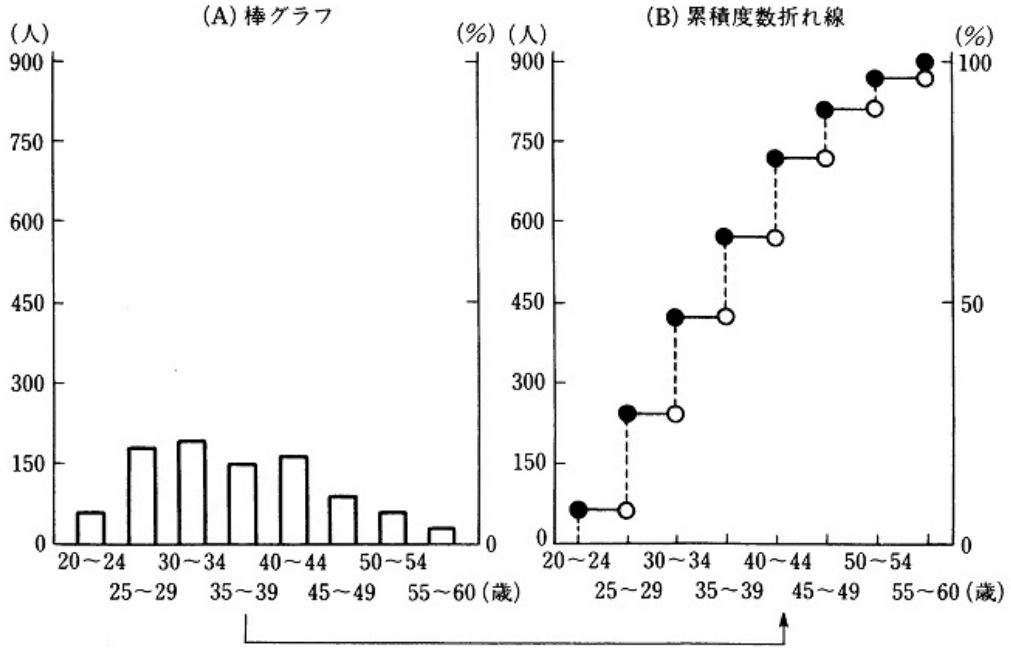
この場合には、度数分布表に対応して棒グラフ、累積相対分布表に対応して累積度数折れ線が考えられる。

1. 棒グラフ(bar graph)は、とりうる値の上に、長さで度数を表すように棒を描いたグラフである。棒は水平方向に寝ている棒でも、垂直の立った棒でもかまわない。ただし、棒には幅があってもいいが、その幅は統一する。また棒同士の間には隙間をあけて棒同士をくっつけないようにして、ヒストグラムではないことを示す。
2. 累積度数折れ線はとりうる値のところでジャンプする階段状のグラフで図示したものである。

こうしたものには量的データで、階級をカテゴリーとして扱う場合も含められる。実は、本書で扱っているような分野の統計では、ヒストグラムや累積度数多角形を描くことはほとんどない。というよりも、SAS などの統計パッケージを使って集計する場合には、量的データであっても、結局は階級をカテゴリーとして扱って、あたかも質的データであるかのように棒グラフを描かせることになってしまう。例えば、次の例は、図 3.2 のヒス

トグラム、累積度数多角形を、階級をカテゴリーとして、(垂直)棒グラフ、累積度数折れ線を描いたものである。

図 3.3　棒グラフと累積度数折れ線



(4)SAS による棒グラフの作成

棒グラフを作成することは、SAS を使えば簡単である。CHART プロシジャを使えばよい。第 3 章で作成した永久 SAS データセット(例では JPC91 SAVE)を PROC 文の DATA=オプションで指定すれば、棒グラフ作成のための SAS プログラムはわずかに 3～4 行で済む。

PC 版 SAS

```
LIBNAME libname '￥パス名';
PROC CHART DATA=libname.永久 SAS データ・セット名本体;
HBAR  変数名の並び;
[OPTIONS NOCENTER;]
RUN;
```

例)

```
LIBNAME SAVE '￥MYDIR';
PROC CHART DATA=SAVE.JPC;
        HBAR I2;
        OPTIONS NOCENTER;
RUN;
```

CMS 版 SAS

```
PROC CHART DATA=永久 SAS データ・セット名;
HBAR  変数名の並び;
[OPTIONS NOCENTER;]
RUN;
```

例)

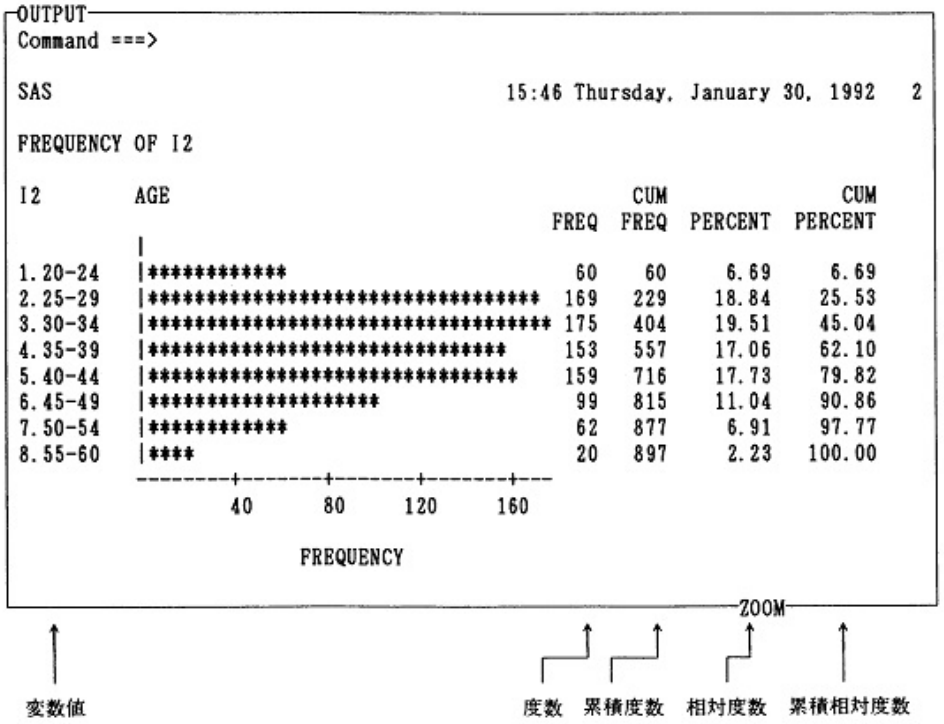
```
PROC CHART DATA=SAVE.JPC;
        HBAR I2;
        OPTIONS NOCENTER;
RUN;
```

これもさきほどと同様に、CMS 版 SAS のプログラムは、形式的には PC 版 SAS のプログラムの 1 行目を削除したものになる。度数分布表を求めるプログラムと比較しても、キ

ーワードの FREQ と TABLES が、それぞれ CHART と HBAR に置き換えられているだけである。

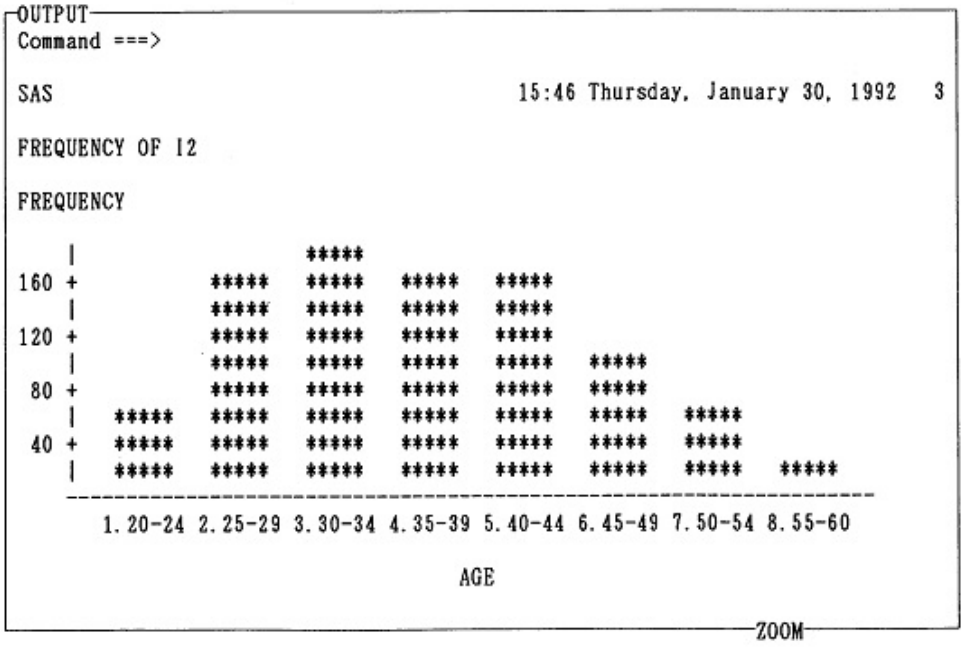
プログラム例の中の変数 I2 は既に述べたように、「組織活性化のための従業員意識調査」の例では年齢を表す変数である。これによって、年齢の度数分布を求めることが出来る。この SAS プログラム例を実行させると、PC 版 SAS では 20 秒弱で図 3.4 のような集計結果が画面に表示されることになる。

図 3.4　SAS で求めた水平棒グラフ



プログラム中のキーワード"HBAR"を"VBAR"に変えると、図 3.4 に例示されるような水平棒グラフの代わりに、図 3.5 に例示されるような垂直棒グラフが表示されることになる。ただし、図 3.5 を見てもわかるように、垂直棒グラフでは度数、累積度数、相対度数、累積相対度数は表示されず、垂直棒グラフのみの表示となる。

図 3.5　SAS で求めた垂直棒グラフ



(5)カテゴリーとしての階級の作り方

階級の設定には、こうしたら良いというような一般的なルールは存在しない。むしろ、試行錯誤を繰り返しながら決めていくべきだといった方がよいだろう。ただし、試行錯誤とはいっても、手作業でやるのは大変である。集計はさておいても、何度も図表を作り直すのは、それだけでも大変な作業量となる。したがって、実用的な意味で、何度も気の済むまで試行錯誤ができるのは、SAS 等の統計パッケージのおかげということになる。その結果として、なめらかな整った形の分布のヒストグラムや棒グラフが得られれば、一応満足して試行錯誤を終了するわけである。しかし、なめらかで整った形であることが、良い階級設定の唯一の基準ではない。ある階級設定が良いものであるかどうかは、統計処理をして分析する側の意図に大きく依存しているのである。このことは、後述の例題を自分で考えてみるとよくわかる。

しかし、次のようなポイントは、一応、頭の片隅に置いて着目してみると役に立つので、例題を考える際にも参考にしてほしい。

- 《ポイント 1》各階級の区間はできるだけ等間隔にとり、階級の境界が切りのいい数値になるようにすること。
- 《ポイント 2》階級の個数 k はデータのサイズ n を考慮して決めること。これには次のスタージェス(Sturges,H.A.)の経験公式が、目安として役に立つ。

$$k \fallingdotseq 1+\log_2 n = 1+(\log n / \log 2) \fallingdotseq 1+3.32 \log n$$

ここで、ポイント 2 のスタージェスの経験公式の中の log は常用対数 $\log_{10}$ のことである。この公式によれば、まずデータが 1 個の場合でも階級は一つは必要である。さらに、データのサイズ n が 2 倍になるごとに階級の個数 k を一つ増やした方がよいというのである。実際にいくつかのデータ・サイズについて、この公式を用いて階級の個数を計算してみると、表 3.5 のようになるので、参考にするとよい。

表 3.5　データのサイズと階級の個数

データのサイズ(*n*)	10	20	50	100	200	500	1,000	2,000	5,000	10,000
階級の個数(*k*)	4.32	5.32	6.64	7.64	8.64	9.96	10.96	11.96	13.28	14.28

例題)　下のデータは、某日、東京都のある小売店のうち、100 軒をくじ引き的に抽出して、そこにおける某品の在庫量(単位は個数)を調べたものである(林, 1973, p.11)。度数分布を図表化してみよ。もちろん SAS 等を使わずに手計算でやってかまわない。その度数分布からどんなことがいえるか。

21	142	0	282	187	100	0	0	225	42
72	159	33	21	61	593	0	47	97	0
0	4	192	0	17	75	133	125	13	163
122	30	0	263	27	186	23	41	18	44
0	8	12	4	47	20	0	16	91	20
0	18	6	0	25	29	178	8	110	0
62	0	98	0	55	0	0	43	75	61
14	0	63	0	25	0	27	12	561	62
14	20	34	64	43	90	169	4	81	44
85	4	10	19	0	35	39	28	0	0

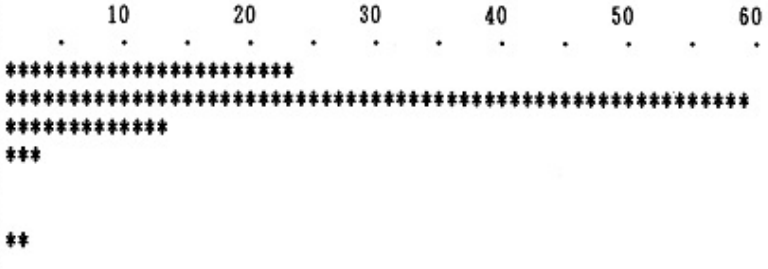
こうした問題にはただ一つの正解というものはない。ポイント 2 については、*n*=100 であるから、スタージェスの公式から階級の数は 7〜8 といったところである。問題はポイント 1 で、これについて二つの解答例を考えてみよう。

《解答例 1》

階級の区間を等間隔にとった場合、次のような度数分布表、棒グラフが描ける。データを一見して、0 が多いことがわかるので、「0」だけで一つの階級にしている。なお棒グラフは度数分布と対照させるために、水平棒グラフにしてある。

図 3.6　解答例 1

階級	度数
0	23
1〜 99	59
100〜199	13
200〜299	3
300〜399	0
400〜499	0
500〜599	2



この棒グラフから、比較的きれいに分布している様子がわかる。しかし、実は、こうしたデータに対して、この解答例のように、等間隔原則を形式的に適用すると、観測値が片

方の端のごく狭い部分に具体的には 0〜199 に集中して、集中部分の情報が失われてしまうことになる。このことは、次の解答例 2 と比較するとよくわかる。

《解答例 2》

階級の区間を対数の意味で等間隔にとった場合。これは対数をとれば、

$\log a \cdot b = \log a + \log b$

$\log a \cdot b^n = \log a + n \log b$

という性質があることから、つまり b 倍ごとに対数の意味で等間隔になっていると考えるのである。スタージェスの公式から階級の数は 7〜8 ということを配慮して、いま 2 倍ごとに区切って階級を作ってみると次のようになる。

図 3.7　解答例 1

階級	度数	10 20 30 40 50 60
0	23	**********
1〜 9	7	*******
10〜 19	11	**********
20〜 39	17	**********
40〜 79	18	**********
80〜159	13	**********
160〜319	9	*********
320〜639	2	**

念のため、各階級の区切りの値の対数をとると次のようになっている。

log 10
log 20=log 10+log 2
log 40=log 10+2 log 2
log 80=log 10+3 log 2
log 160=log 10+4 log 2
log 320=log 10+5 log 2
log 640=log 10+6 log 2

こうすると、「0」という階級を除いて、きれいな分布をしていることがわかる。このような考え方は、従業員規模に大きな違いのある企業の集団について、従業員規模別企業数の度数分布を調べるような場合にも同じようにあてはまる。例えば従業員規模 10 人の企業にとって従業員が 10 人増減するということは大変なことである。しかし、従業員規模 10,000 人の企業で従業員が 10 人増減することはあまり問題にはならない。なぜなら、前者ではたった 10 人でも従業員規模は倍増もしくは消滅を意味しているのに、後者では 10 人では 0.1%にしかならず、この程度の従業員数の変動は日常的に起きている可能性がある。このように、絶対数での差よりも、むしろ相対的な比率に意味があると考えられる場合には、等しい比率が等しい間隔になるようにした方がよい。つまり、対数の意味での等間隔である。(→演習問題 3.1)

ところで、この解答例 2 では、階級「0」に集中した分布の山と、階級「40〜79」をピークとしたきれいな分布の山の二つの山のあることがわかる。このように、分布の峰が複数ある場合には、全く特性の異なる複数の母集団からの標本が混じっている可能性がある。こうした場合には、層別(stratification)によって、適切にグループ化すると、それぞれのグループで、峰が一つの単純な分布(単峰型;unimodal)が現れることが多い。

例題の場合もこの作業が必要かもしれない。少なくとも、「0」という値は異常値で、何か特別な理由があるはずである。おそらく、「某品の在庫量」調査とはいうものの、実際には、たまたま調査時に在庫が底をついていて、在庫量が 0 だったという小売店はごく少数で、むしろその多くはずっと恒常的に 0、つまり、そもそも当該商品をまったく取り扱わない小売店がけっこうあるのであろう。他方、当該商品を扱っている小売店では、その規模等で在庫量にばらつきが生まれるために、きれいな分布が現れたものと思われる。当該商品を最初からまったく相手にしていない小売店がけっこうあるということ自体が、その商品の販売担当者にとっては重要かつ有用な情報となるであろう。もっとも、こうした有用な情報も、解答例 2 のように整理したからわかったもので、解答例 1 のように整理していては、気付かれることもなかっただろうが……。

3. 平均による要約

前節の例題のように、全く特性の異なる複数の母集団からの標本が混じっているような場合には、いわゆる「平均」のような代表値に、代表値としての意味はほとんどないということはもうおわかりだろう。これから、この節で扱う平均、分散は単峰型の分布を要約する際に用いられるべきものである。

(1)(算術)平均

通常、平均といえば算術平均(arithmetic mean)のことになるが、算術平均を計算することは、統計学を知らない人でも行うごく普通の操作であるといえる。

サイズ n のデータ $x_1, x_2, ..., x_n$ の平均 x は

$$\bar{x}=(x_1+x_2+\cdots+x_n)/n=(1/n)\Sigma_{i=1}^{n} x_i$$

平均は変数 x の記号の上に横棒を付して $\bar{x}$ (「x バー」または「バーx」と読む)と表記するのが慣例となっている。

任意の定数 a, b, c に対して、$ax+by+c$ の算術平均を考えると、

$$\overline{ax+by+c}=(1/n)\Sigma_{i=1}^{n}(ax_i+by_i+c)=(a/n)\Sigma x_i+(b/n)\Sigma y_i+(1/n)nc=a\bar{x}+b\bar{y}+c$$

つまり、算術平均は次のような性質がある。

$$\overline{ax+by+c}=a\bar{x}+b\bar{y}+c$$

したがって、平均値を使った演算は「常識的」に行うことが出来る。

当然といえば当然のことであるが、平均は実際には観測されていない値になることが多い。特に離散型データの場合には、そもそもありえない値になることが多い。

また $x_i-x, i=1, ..., n$ を平均からの偏差と呼ぶが、その和は次に示すように 0 になる。

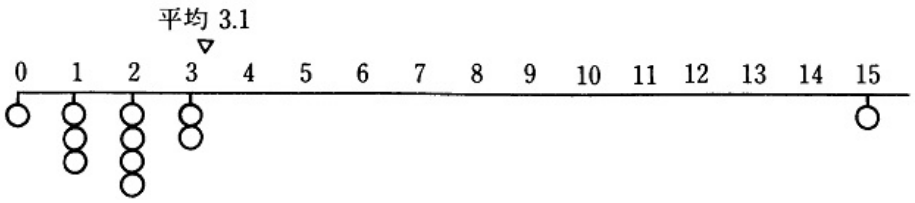
$$\Sigma_{i=1}^{n}(x_i-\bar{x})=\Sigma x_i-\Sigma\bar{x}=n\bar{x}-n\bar{x}=0$$

の性質の意味するところは、仮に各観測値に重さ 1 の分銅を割り当て、数直線にぶら下げたときに、平均の回りのモーメントは 0 になるということである。つまり、平均の所に支点があれば、この天秤はつり合うことになる。このことは、平均は全データの「重心」に当るということを意味している。したがって、全データの「中心」に当たると考えると誤解を生ずることがある。

例えば、10 社から中間管理職 1 人ずつの計 10 人からなるグループで、「自分の直接の部下の数」について調べて、天秤で図示してみると、図 3.8 のような結果になったとしよう。日本の企業では、ポスト不足から、直接の部下数の少ない中間管理職が以外と多く、中には部下のいない「スタッフ管理職」もいるのである。もちろん図 3.8 は仮想例であるが、こうした実態は、この章の冒頭にあげた「E サービス㈱の所内保全サービス部門」のデータを見ても明らかである。なにしろ、8 人の職場で、課長 3 人、係長 2 人、一般 3 人

なのであるから、どう考えても直接の部下数は 1 人いるかいないかである。実は、平均は少数のはずれ値あるいは異常値に大きく影響を受けるので、そうした場合には、平均が集団を代表していると考えるには問題がある。この例では、1 人を除いて全員が平均を下回っており、直接の部下の数の分布を平均 3.1 で代表させると、嘘をつかれたような気になる。

図 3.8　重心としての平均



実は、一般に分布がゆがんでいる場合、算術平均のもつ意味については注意が必要になる。いま、比較のためにメディアンを考えよう。メディアン(median)は中央値、あるいは中位数とも言われ、データ $x_1, x_2, ..., x_n$ を小さい値から順に並べたものを $x_{(1)}, x_{(2)}, ..., x_{(n)}$ とすると、メディアンはその中央の値ということになる。偶数個のデータの場合もあるので、メディアンの定義は次のようになる。

a. $n=2q+1$ (n が奇数)のとき、$M=x_{(q+1)}$
b. $n=2q$ (n が偶数)のとき、$M=\{x_{(q)}+x_{(q+1)}\}/2$

図 3.8 ではメディアンは 2 人になり、これならまだ平均 3.1 人でこの分布を代表させるよりはましのようである。一般に、

1. 度数分布が左右対称ならば、$\bar{x}=M$
2. 度数分布が左にゆがんでいるならば、$\bar{x}<M$ つまり「平均よりも上の人が多い」。
3. 度数分布が右にゆがんでいるならば、$\bar{x}>M$ つまり「平均よりも下の人が多い」。

しかし、分布の形がゆがんでいるときや複数の峰をもつときに、平均であれ、メディアンであれ、その分布をたった一つの値で代表させようと考えること自体が非常識で無謀である。

(2)加重平均

データ $x_1, x_2, ..., x_n$に対して、ウェイト $w=(w_1, w_2, ..., w_n)$ による加重平均は

$$x_w=(w_1/\Sigma_j w_j)\,x_1+(w_2/\Sigma_j w_j)\,x_2+\cdots\cdots+(w_n/\Sigma_j w_j)\,x_n=\Sigma_{i=1}^{n}(w_i/\Sigma_j w_j)\,x_i$$

と定義される。ただし、$w_i \geqq 0,\ i=1, 2, ..., n$ である。$w_1=w_2=\cdots=w_n$ のときの加重平均は算術平均と同じことになる。

この定義だけでは一体何の目的のために加重平均を求めるのかよくわからないだろう。しかし実は、加重平均は暗黙のうちに多用されているのである。例えば、企業の自己資本比率は、自己資本比率＝自己資本／総資産 で定義され、各企業の自己資本と総資産がわかれば、それぞれの企業の自己資本比率を計算することは簡単である。それでは、表 3.6 のような n 社のデータが与えられたときには、全体の自己資本比率はどのように計算するだろうか。各社の自己資本比率を計算して、その平均をとるだろうか。SAS を使えば、そんな処理はいともたやすい。しかし、表 3.6 のような表が与えられていると、自己資本の合計額を総資産の合計額で割ってみたくなる誘惑にかられる。これは無意味なことだろうか。実はこれが加重平均になっているのである。

表 3.6　総資産・自己資本・自己資本比率

企業	総資産	自己資本	自己資本比率
1	x_1	y_1	z_1
2	x_2	y_2	z_2
:	:	:	:
n	x_n	y_n	z_n
全体	x_w	y_w	?

$$
\begin{aligned}
\text{自己資本比率} &= y_w / x_w = (y_1+y_2+\cdots+y_n) / (x_1+x_2+\cdots+x_n) \\
&= \{x_1 / (x_1+x_2+\cdots+x_n)\}(y_1 / x_1) + \{x_2 / (x_1+x_2+\cdots+x_n)\}(y_2 / x_2) \\
&\quad + \cdots\cdots + \{x_n / (x_1+x_2+\cdots+x_n)\}(y_n / x_n) \\
&= \{x_1 / (x_1+x_2+\cdots+x_n)\}z_1 + \{x_2 / (x_1+x_2+\cdots+x_n)\}z_2 \\
&\quad + \cdots\cdots + \{x_n / (x_1+x_2+\cdots+x_n)\}z_n
\end{aligned}
$$

つまり、総資産額による加重平均となっているのである。多くの場合、各社の自己資本比率を計算して、その平均を求めたものよりも、この加重平均の方が実質的には、はるかに意味があるだろう。したがって、コンピュータと SAS を使って、何でもかんでも力まかせに平均を求めればよいというものではない。

(3)度数分布表から求める平均

度数分布表から平均を近似的に求める場合には、この加重平均が使われる。一般に、m 組のデータを込みにした全データの平均は各組の平均をサイズによって加重平均したものとなる。

表 3.7　グループのサイズと平均

グループ	サイズ	平均
1	n_1	x_1
2	n_2	x_2
:	:	:
m	n_m	x_m
全体	n	x

つまり、いまグループ k の i 番目のデータを x_{ki} で表すと、

$$
\begin{aligned}
\bar{x} &= (1/n)(\Sigma_{i=1}{}^{n}{}_{1}\, x_{1i} + \Sigma_{i=1}{}^{n}{}_{2}\, x_{2i} + \cdots\cdots + \Sigma_{i=1}{}^{n}{}_{m}\, x_{mi}) \\
&= (n_1/n)\{(1/n_1)\Sigma_{i=1}{}^{n}{}_{1}\, x_{1i}\} + (n_2/n)\{(1/n_2)\Sigma_{i=1}{}^{n}{}_{2}\, x_{2i}\} + \cdots\cdots + (n_m/n)\{(1/n_m)\Sigma_{i=1}{}^{n}{}_{m}\, x_{mi}\} \\
&= (n_1/n)\bar{x}_1 + (n_2/n)\bar{x}_2 + \cdots\cdots + (n_m/n)\bar{x}_m
\end{aligned}
$$

ここで、グループ 1, 2, ..., m は、カテゴリーでもいいし、階級でもいいのである。したがって、この考え方は、量的データの度数分布表からの全データの平均の算出に適用することが出来る。計算機を使用して平均を計算できる場合には必要のないことであるが、質問票の段階で階級にプリコーディングしてある場合には次のような手続きが必要になる。つまり、

1. 各階級で平均を計算し、
2. 各階級での平均を各階級の度数で加重平均する

のである。2 については、既に述べたので、残る 1 について考えてみよう。

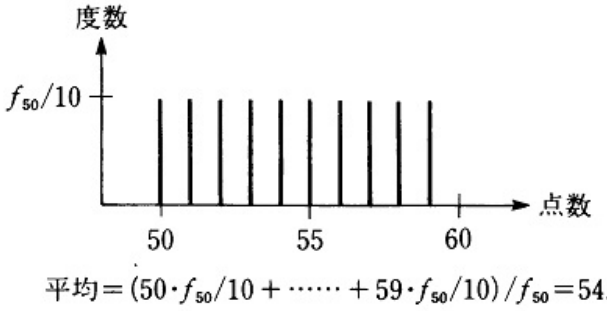
各階級の区間での一様分布を仮定することができれば、各階級での平均、階級値で近似できる。実際には、各階級の区間での一様分布をほぼ仮定できるので、平均を計算するということは、階級値(class value)を計算するということになる。階級の境界 a_i, b_i がわかっていれば、階級値 m_i は次の定義のようにいたって簡単に計算できる。

$$m_i=(a_i+b_i)/2, \quad i=1, 2, ..., k$$

ただし、問題は階級の境界である。階級の境界 a_i, b_i については変数とその観測値の性質を考慮することが必要になる。

a. 端数切捨てによって得た連続変数のデータの場合。例えば、「満年齢 50 歳」は満年齢 50 歳以上 51 歳未満のことなので、階級「50 歳以上～60 歳未満」の階級値は
   (50+60)/2=55 歳
   となる。
b. 四捨五入によって得た連続変数のデータの場合。例えば、「体重 50kg」というのは、実際には、「49.5kg 以上～50.5kg 未満」のことなので、階級「50kg 以上～60kg 未満」は、実際には「49.5kg 以上～59.5kg 未満」を意味している。つまり、0.5kg 下方にシフトしていたはずのものなのである。したがって、階級「50kg 以上～60kg 未満」の階級値は
   (49.5+59.5)/2=54.5 kg
   となる。
c. 離散変数のデータの場合には、例えば、階級「50 点以上～60 点未満」は、実際には「50 点以上～59 点以下」のことであり、59 点より大きく 60 点未満の点数はとりようがない。したがって、この階級の階級値は、図 3.9 からも明らかなように、
   (50+59)/2=54.5 点
   となる。

図 3.9　階級「50 点以上～60 点未満」の点数の一様分布と平均



細かいことのようでも、階級値が 0.5 ポイントずれるということは無視できない大きさである。つまり、不注意に平均を計算すれば、プリコーディングをしたことで、平均年齢や平均体重、平均点がそれぞれ 0.5 歳、0.5kg あるいは 0.5 点もずれて結論を出してしまうのである。

(4)幾何平均

サイズ n のすべて正のデータ $x_1, x_2, ..., x_n$ の幾何平均(geometric mean) G は

$$G=(x_1 \cdot x_2 \cdot \cdots \cdot x_n)^{1/n}$$

で定義される。この幾何平均は時系列的に得られた変化の比率を平均する場合などに用いられる。

例えば、ある企業の 3 年間の売上高の対前年伸び率が、20%、25%、20%だったとき、この 3 年間の年平均の売上高伸び率は次のように幾何平均で求められる。

$$G=(1.20 \cdot 1.25 \cdot 1.20)^{1/3}=1.80^{1/3} \fallingdotseq 1.216$$

ただし、これは 3 年間で売上高が $1.20 \cdot 1.25 \cdot 1.20=1.80$ と 80%伸びているので、次のように考えることと結局同じである。つまり、ある企業の売上高が 3 年間で 1.8 倍になった。毎年同じ伸び率 G で売上高が伸びているとすると、3 年間で売上高が 1.8 倍になるような G を求めると

$$G \cdot G \cdot G=G^3=1.8$$

$$\therefore \quad G=1.8^{1/3} \fallingdotseq 1.216$$

このような G を年平均の売上高伸び率と考えるのである。

しかし、この例からもわかるように、幾何平均では結果的に途中の数字の動向はすべて計算に入らないことになる。このことを知らずにデータを集めると、せっかく集めたデータが生かされないケースも起こりうる。例えば、調査で、ある企業の売上高が、当初 100 億円で、続く 3 年間で 120 億円、150 億円、180 億円と伸びたことがわかったとしよう。このとき、3 年間の対前年売上高伸び率 120/100(=1.20), 150/120(=1.25), 180/150(=1.20) の幾何平均は

$$G=(120/100) \cdot (150/120) \cdot (180/150)^{1/3}=1.8^{1/3} \fallingdotseq 1.216$$

となる。しかし、この式をよく見ると、幾何平均の計算の途中で、分子分母が互いに消去されるので、結局のところ、幾何平均 G は最初と最後の売上高 100 億円と 180 億円しか反映していないことがわかる。つまり、未知数 a, b, c を含んだ次の式でも十分なのである。

$$G=(a/100) \cdot (b/a) \cdot (180/b)^{1/3}=1.8^{1/3} \fallingdotseq 1.216$$

このことは、逆に非常に長期にわたる、例えば 50 年間の平均の売上高伸び率であっても、実は、幾何平均を使う限りは、最初の年と最後の年の売上高データさえあれば、途中のデータがごっそり抜けていても計算できるということになる。

ところで、両辺の対数をとると、幾何平均の対数は、観測値を対数に変換したときの算術平均に等しいので、統計パッケージを使うときに便利である。

$$\log G=(\log x_1+\log x_2+\cdots\cdots+\log x_n)/n=(\Sigma_{i=1}^{n} \log x_i)/n$$

4. 分散による要約

(1)分散

度数分布が左右対称で、したがって平均とメディアンが等しいときでも、次の 4 つの分布の形は明らかに異なり、そのことは図 3.10 のように棒グラフを作るとはっきりする。

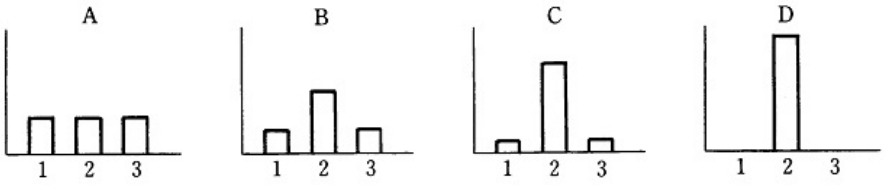
A: 1, 1, 1, 2, 2, 2, 3, 3, 3

B: 1, 1, 2, 2, 2, 2, 2, 3, 3

C: 1, 2, 2, 2, 2, 2, 2, 2, 3

D: 2, 2, 2, 2, 2, 2, 2, 2, 2

図 3.10　平均が等しくて、分散の異なる分布



4 つの分布はいずれも左右対称で、その平均は 2 である。しかし、分布の形は明らかに異なっている。これは実は分布の散らばっている程度(逆に言えば、分布のかたまっている程度)が異なっているからである。その程度を表す代表的なものが分散である。

データ $x_1, x_2, ..., x_n$の分散(variance)は次のように定義される。

$$s_x^2=\{\Sigma_{i=1}^{n}(x_i-\bar{x})^2\}/n=\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2\}/n$$

ここで、分散の分子は平方和(sum of squares)または変動と呼ばれ、

$$S_x=ns_x^2=\{\Sigma_{i=1}^{n}(x_i-\bar{x})^2\}=\Sigma x_i^2-2\bar{x}\Sigma x_i+n\bar{x}^2=\Sigma x_i^2-2n\bar{x}^2+n\bar{x}^2=\Sigma x_i^2-n\bar{x}^2$$

分散はこのように偏差平方和を用いているので、その単位は観測値の単位、つまり変数の測定単位の 2 乗である。そこで、観測値と同一の単位にするために分散の正の平方根をとった標準偏差も広く利用される。きちんと定義しておけば、標準偏差(standard deviation)は

$$s_x=(s_x^2)^{1/2}$$

のように定義される。

任意の定数 a, b に対して、$ax+b$ の分散は

$$s_{ax+b}^2=\{\Sigma(ax_i+b-(\overline{ax+b}))^2\}/n=\{\Sigma(ax_i+b-(a\bar{x}+b))^2\}/n=\{\Sigma(a^2(x_i-\bar{x})^2\}/n=a^2s_x^2$$

となる。つまり、次のような性質がある。

$$s_{ax+b}^2=a^2s_x^2$$

(2)変動係数

分布の散らばっている程度を表すものとして分散や標準偏差を考えたが、実は分散や標準偏差では分布の散らばり具合を形式的に比較できないような場合もある。そのようなときには、変動係数(coefficient of variation)

$$C.V.=(s_x/\bar{x})\times100$$

が用いられる。変動係数はこの定義のように%で表すことが多い。具体的には、変動係数は観測値がすべて正のとき、次のようなケースで相対的な散らばり具合を比較するのに用いられる。

変動係数は、たとえば測定単位の異なるデータ(ただし、m と cm のように原点の一致している測定単位)を比較する場合に用いられる。このとき、平均と標準偏差は同じ単位なので、変動係数は無名数となり、比較が可能になる。つまり、$y_i=ax_i, i=1, 2, ..., n, a>0$ としても

$$s_y/\bar{y}\times100=\{(as_x)/(a\bar{x})\}\times100=(s_x/x)\times100$$

となるので、測定単位が異なっていても変動係数は変わらないという性質を使っているのである。

例えば、地域間県民所得格差のデータによると、1965 年の平均所得が 26.6 万円、標準偏差が 7.5 万円、1975 年の平均所得が 117.5 万円、標準偏差が 23.8 万円であった。標準偏差だけを見ると、1965 年の方が圧倒的に小さい。しかし、1975 年は 1965 年に比べて、平均所得は約 4.5 倍、標準偏差は約 3 倍となっており、相対的な地域間所得格差は 1975 年の方

が小さくなっているといってよさそうだ。このように、測定単位が同じであっても中心の位置が著しく異なるデータの場合には、変動係数が用いられる。1965 年の変動係数は 28.2、1975 年の変動係数は 20.3 となり、確かに、1975 年の方が変動係数は小さくなる。

5. 2 値質的データの平均と分散

質的データがカテゴリーA とカテゴリー非 A しかもたないとき、2 値質的データと呼ばれる。既に第 1 章第 3 節で述べた尺度による観測値の演算可能性を繰り返せば、

1. 名義尺度では計数にもとづく演算だけが意味をもつ。
2. 順序尺度では順位に関する演算も意味をもつ。
3. 間隔尺度では加減の演算も意味をもつ。
4. 比率尺度では加減乗除の演算も意味をもつ。

ということだった。こうした演算可能性については、十分に注意を払わなければならないし、この規則にしたがえば、名義尺度や順序尺度に基づいた質的データの算術平均(arithmetic mean)は通常は無意味なはずである。しかし、質的データであっても、複雑な演算に耐えられるように扱う方法もまた工夫されている。いま、質的データが、カテゴリーA とカテゴリー非 A の二つの値のみしかもたないとき、変数 x_i を

a. カテゴリーA ならば、$x_i=1$
b. カテゴリー非 A ならば、$x_i=0$

とおくと、この変数の平均 $\bar{x}$ は次のようにカテゴリーA の比率 p を示すと解釈することができる。いまカテゴリーA と非 A の度数を、それぞれ f_1、f_0 とすると

$$\bar{x}=(\Sigma x_i)/n = [(1+\cdots\cdots+1)+(0+\cdots\cdots+0)]/n = f_1/n = p$$
$$\qquad\qquad f_1\text{個} \qquad\qquad f_0\text{個}$$

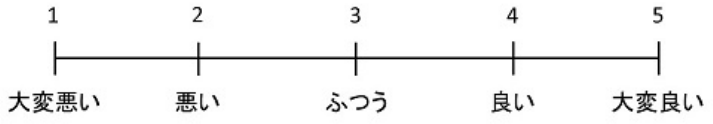
となる。

また、平均が考えられたように、2 値質的データの分散も考えることが出来る。平均がカテゴリーA の比率 p になっているので、分散は次のようになる。

$$s_x^2=\{\Sigma(x_i-\bar{x})^2\}/n = [\{(1-p)^2+\cdots\cdots+(1-p)^2\}+\{(0-p)^2+\cdots\cdots+(0-p)^2\}]/n$$
$$\qquad\qquad f_1\text{個} \qquad\qquad f_0\text{個}$$
$$=[f_1(1-p)^2+f_0(0-p)^2]/n = p(1-p)^2+(1-p)p^2=p(1-p)$$

したがって、分散は $p=1/2$ で最大値 1/4 をとることになる。ところで、このことから、標準偏差は各カテゴリーの比率の幾何平均になっているのである。

このように 2 値質的データの場合には、質的データであっても数量化することが可能となり、そのことにより、より高度の演算を必要とする統計手法を使用することが可能となる。しかし、次のような方法は、似て非なるものなので注意がいる。いま、個人の判断する好き嫌いや良し悪しなどの評価を、たとえば



といった 5 段階くらいで答えさせ、回答を 1〜5 点のスコアに置き換えることを考えよう。このようにして得られた複数の質問のスコアを合算して用いる尺度をリッカート・スケール(Likert scale)という。

こうした 5 点尺度あるいは 7 点尺度などで質問の回答を求めることは、リッカート・スケールという意識がなくても、調査ではよく用いられる方法である。しかし、この方法で得られた観測値間の演算を行うことは、実は理屈の上では問題が多い。なぜならば、それらは、たとえば通信簿の点のように、同一人物が統一的に評価したものではなく、複数の人がまったく独立に主観的に評価したものなので、順序尺度として扱うことにすら疑問があるからである。にもかかわらず、5 点尺度、7 点尺度の観測値は、とりうる値が 5 つ、7 つと多いために、カテゴリーとして扱うには不便であるということもあって、平均値をとったり、より複雑な演算にもとづくデータ処理が行われたりすることが多い。しかし、この種の観測値を間隔尺度以上の尺度にもとづく観測値として扱うことには、前述の演算可能性の規則からいうと根拠がない。

したがって、5 点尺度や 7 点尺度のデータを扱う場合には、無造作な演算は許されない。細心の注意が必要である。もし、その平均値をとる場合には、少なくとも分布の峰が一つで、単峰型(unimodal)をしていることを確認すべきである。峰が二つ以上ある場合には、既に述べたように、平均値で代表させる意味はほとんどない。また、複数の質問のスコアを合算して用いる場合には、項目分析を行うなどして、各質問項目の合算の適否を検討すべきである。

しかし、5 点尺度や 7 点尺度を使って調査することの一番大きな問題点は、5 段階評価や 7 段階評価を集計分析段階で、たとえば「良い」「悪い」の二つのカテゴリーに 2 分割することが、あまりに無造作にあるいは作為的によく行われるということである。この場合には、中立的回答をどちらのカテゴリーに分類するのかで、「良い」「悪い」のどちらを多数意見とするかを調査者側が恣意的に決めてしまう危険性がある。通常は、中立的回答、すなわち尺度のほぼ中央に分布の峰があるので、ほとんどのケースで調査者側に恣意的な判断を求めていることになる。最終的に「良い」「悪い」の二つのカテゴリーにまとめるつもりならば、調査段階で質問の回答の選択肢を「良い」「悪い」にしておくべきであろう。

6. SAS による平均・分散の計算

SAS を使えば、変数の平均、分散を計算することは簡単である。MEANS プロシジャを使えば、わずか数行で済む。

PC 版 SAS

```
LIBNAME libname '￥パス名';
PROC MEANS [ｵﾌﾟｼｮﾝ] DATA=libname.永久 SAS ﾃﾞｰﾀ･ｾｯﾄ名本体;
[VAR  変数名の並び;]
[OPTIONS NOCENTER;]
RUN;
```

例)

```
LIBNAME SAVE '￥MYDIR';
PROC MEANS DATA=SAVE.JPC;
        VAR TII1--TVI15;
        OPTIONS NOCENTER;
RUN;
```

CMS 版 SAS

```
PROC MEANS [ｵﾌﾟｼｮﾝ] DATA=永久 SAS ﾃﾞｰﾀ･ｾｯﾄ名;
[VAR  変数名の並び;]
[OPTIONS NOCENTER;]
RUN;
```

例)

```
PROC MEANS DATA=SAVE.JPC;
        VAR TII1--TVI15;
        OPTIONS NOCENTER;
RUN;
```

CMS 版 SAS のプログラムは形式的には PC 版 SAS のプログラムの 1 行目を削除したものになる。VAR 文を省略すると、すべての数値変数について求められる。PROC MEANS 文のオプションを省略すると、欠損値でないオブザーベーション数(N)、最大値(Minimum)、最小値(Maximum)、平均(Mean)、標準偏差(Std Dev)について求められる。主なオプションとしては次のようなものがある。

CV ・・・・・・ 変動係数を求める
MAX ・・・・・・ 最大値を求める
MIN ・・・・・・ 最小値を求める
MEAN ・・・・・・ 平均値を求める
N ・・・・・・ 欠損値でないオブザーベーション数を求める
STD ・・・・・・ 標準偏差を求める
SUM ・・・・・・ 合計を求める
VAR ・・・・・・ 分散を求める

オプションを一つでも指定すると、オプションで指定したものだけが求められる。いまプログラム例の PROC MEANS 文で

PROC MEANS N MEAN VAR CV SUM DATA=SAVE.JPC;

と N MEAN VAR CV SUM の 5 つのオプションを指定してからプログラムを実行してみると、その結果が、PC 版 SAS では 2 分強で画面に 5 画面分表示される。図 3.11 はそのうちの第 1 画面である。(→演習問題 3.2)

図 3.11　SAS で求めた平均、分散

OUTPUT
Command ===>

SAS　　8:20 Friday, January 31, 1992　1

N Obs	Variable	N	Sum	Mean	Variance	CV
907	TII1	902	738.0000000	0.8181818	0.1489254	47.1666049
	TII2	906	581.0000000	0.6412804	0.2302940	74.8330461
	TII3	904	603.0000000	0.6670354	0.2223451	70.6911313
	TII4	902	378.0000000	0.4190687	0.2437203	117.8041573
	TII5	905	315.0000000	0.3480663	0.2271672	136.9338451
	TII6	902	658.0000000	0.7294900	0.1975533	60.9288302
	TII7	901	543.0000000	0.6026637	0.2397262	81.2423990
	TII8	903	663.0000000	0.7342193	0.1953577	60.1990258
	TII9	901	454.0000000	0.5038846	0.2502627	99.2811904
	TII10	905	593.0000000	0.6552486	0.2261478	72.5754980
	TII11	906	298.0000000	0.3289183	0.2209750	142.9168603
	TII12	904	348.0000000	0.3849558	0.2370270	126.4701734
	TII13	903	436.0000000	0.4828350	0.2499822	103.5513588
	TII14	893	298.0000000	0.3337066	0.2225958	141.3818477
	TII15	904	457.0000000	0.5055310	0.2502462	98.9546032

ZOOM

ここで、変数 TII1～TVI15 は、もともとは「組織活性化のための従業員意識調査」の Yes-No 形式の全質問 75 問に対応する変数であるが、第 2 章第 6 節(5)の通常型 SAS プログラムの中で、Yes-No 形式の質問の回答で、「1. Yes」ならば 1、「2. No」ならば 0 という値をとるように再定義しておいたものである。したがって、この章の前節で扱った 2 値質的データの平均と分散を求めたことになる。つまり、各変数の平均は Yes 比率を示してい

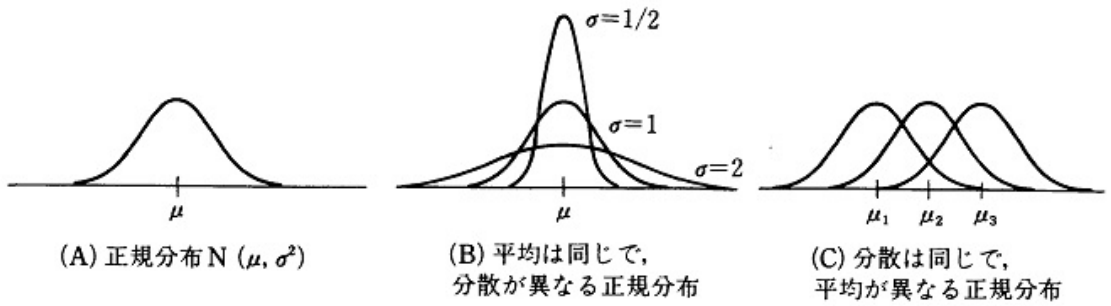
る。ところで、分散は Yes 比率と No 比率(＝1－Yes 比率)の積になっているはずであるが、実際に計算してみると、小数点以下 3 桁までしか一致していないことがわかる。このことからすると、SAS の分散の有効桁数は 3 桁程度のようである。それ以上の桁数を長々と出力通りに書き移しても無意味である。

7. 正規分布と標準得点

正規分布(normal distribution)は、ガウス(C.F.Gauss, 1777-1855)が天文観測データの測定誤差の研究から誤差理論を確立した際の誤差関数(error function)が原型となったものである。そのため、正規分布のことをガウス分布(Gaussian distribution)ということもある。

平均 μ、分散 σ^2の正規分布は $N(\mu,\sigma^2)$ で表す。図 3.12 では様々な正規分布の形を描いているが、正規分布 $N(\mu,\sigma^2)$ は平均 μ を中心にして左右対称の均整のとれたベル型をしている。平均 μ で最大値 $1/\{(2\pi)^{1/2}\sigma\}$ をとり、$\mu\pm\sigma$ が変曲点となっている。平均 μ の値が変わると、分布全体が左右に平行移動し、分散 σ^2の値が大きくなると、分布が全体的に平べったく広がるという性質がある。

図 3.12　正規分布



にもかかわらず、確率変数 X が正規分布 $N(\mu,\sigma^2)$ にしたがえば、例えば次のようなことがわかっている。

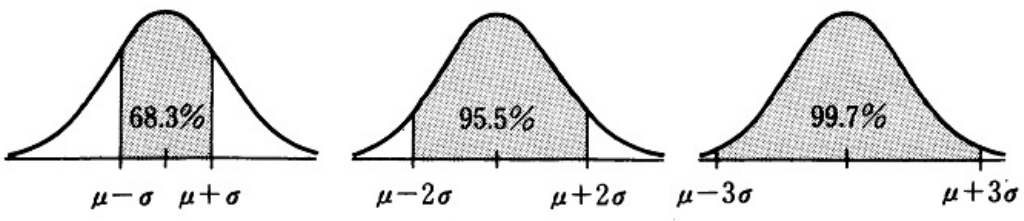
$Pr(\mu-\sigma \leqq X \leqq \mu+\sigma)=0.683$

$Pr(\mu-2\sigma \leqq X \leqq \mu+2\sigma)=0.955$

$Pr(\mu-3\sigma \leqq X \leqq \mu+3\sigma)=0.997$

これを図示すると、図 3.13 のようになる。

図 3.13　正規分布と標準偏差



このことは非常に重要なことである。母集団の相対度数分布を正規分布で近似できるような場合、その母集団を「正規母集団」と呼ぶが、このように、ある集団の特性値の分布が正規分布で近似できるような場合には、分布の平均 x と分散 s_x^2 がわかれば、例えば、区間 $[\bar{x}-s_x,\ \bar{x}+s_x]$ には 68.3%、区間 $[\bar{x}-2s_x,\ \bar{x}+2s_x]$ には 95.5%、区間 $[\bar{x}-3s_x,\ \bar{x}+3s_x]$ には実に全体の 99.7%が属していることがわかるのである。したがって、正規分布は左右対称だから、当然、区間 $(-\infty,\ \bar{x}-s_x]$ と区間 $[\bar{x}+s_x, \infty)$ にはともに (100-68.3)/2＝15.85% が属しているというようなこともわかってしまう。これは、別に標準偏差 s_x の整数倍に限ったことではなく、ある値 x が平均 $\bar{x}$ から標準偏差 s_x の何倍離れているのかさえわかればよいのである。例えば、区間$[\bar{x}-1.96s_x,\ \bar{x}+1.96s_x]$には全体の 95%が属している(第 1 章第 7 節の場合)。そこでいま変数 x を次のように変数 z に変換することを考えてみよう。これを標準化(standardization)という。

$$z=(x-\bar{x})/s_x$$

この z は標準得点(standard score)または z 得点(z-score)と呼ばれる。この標準得点については、定義より、

$$\bar{z}=(\bar{x}-\bar{x})/s_x=0$$

$$s_z^2=s_x^2/s_x^2=1$$

つまり、標準得点 z の平均は 0、標準偏差は 1 という性質がある。言い換えれば、平均 0、標準偏差 1 (したがって、分散も 1)の変数に変換することが「標準化」なのである。したがって、標準得点は平均 0、分散 1 の正規分布 N(0,1) にしたがうことになる。この N(0,1) は標準正規分布と呼ばれる。

ある集団の特性値が正規分布をしていると考えられる場合には、各値の分布全体の中での位置を知るためには、この標準得点が強力な方法となる。つまり、標準得点が z だということは、標準化する前のもとの値が、平均から標準偏差の z 倍離れているということを示している。したがって、標準得点はそれだけで全体の中での位置(例えば、上位○○%、あるいは上から○番目くらい)の情報をもたらしてくれるのである。例えば、いまある人が自分の標準得点 $z=2$ であることがわかれば、標準得点 $z=2$ 以上の者は全体の (100－95.5)/2＝2.25% になるので、自分が上位 2.25%にいることがわかる。仮に総受験者数が 10,000 人であったなら、自分は上から 225 番前後であることまでわかってしまう。考えてみると、通常は順位を知るためには、全個体を得点順に並べ直してみなくてはいけないはずなのに、それをしなくても、平均と標準偏差(あるいは分散)さえわかれば、標準得点を計算してやることで、全体の中でのおおよその位置を知ることが出来るのである。その意味で、標準得点は強力である。

試験結果の通知などによく用いられる「偏差値」(deviation score)は、標準化された変数 z から次のように計算される。

$$\text{偏差値}=50+10z$$

つまり、偏差値は平均 50、標準偏差 10 となるように点数を変換したものである。したがって、標準得点と同様に、全体の中での位置の情報をもたらしてくれることになる。

実際、試験結果に限らず、母集団分布に正規分布を仮定することが一般的に行われている。例えば、身長や測定誤差の分布など、正規分布で表せる分布が多い。その他にも、例えば、所得は対数をとると正規分布で表せることが知られているが、このように適当な変数変換を施すと、正規分布で表せるものも多いということが知られている。

8. 2 群の平均値の比較

いま、企業に対する標本調査を行い、標本について、男女別に平均年齢を求めてみると、男子従業員の平均年齢は 38.2 歳、女子従業員の平均年齢は 28.8 歳であったとしよう。実に 10 歳近くも差があるわけだが、この平均年齢をもって、今回の調査企業では、男女の平均年齢には差があると結論してもよいものだろうか。ひょっとすると、母集団では男女の平均年齢には差がないのに、標本誤差のために、今回の標本では平均年齢に差が出てしまったのかもしれない。もっとも、常識的に考えて、母集団では差のなかったものが、標本誤差だけで、10 歳も差が出るとは考えにくい。その「10 歳も差が出るとは考えにくい」という程度を、確率を使って表現して吟味しようというのが、平均値の差の検定である。2 群の平均値の差の検定は、通称「t 検定」(t-test)としてよく知られているものである。

(1)平均値の差の検定

母集団のもっている分布を母集団分布と呼ぶが、母集団からランダムに大きさ m の標本 $(X_1, X_2, ..., X_m)$ を選ぶと、各 X_i はこの母集団分布に従う確率変数であると考えられる。母集団分布が平均 μ_1、分散 σ_1^2 の正規分布 $N(\mu_1, \sigma_1^2)$ で近似される正規母集団の場合、標本平均

$$\bar{X}=(X_1+X_2+\cdots+X_m)/m$$

は正規分布 $N(\mu_1, \sigma_1^2/m)$ にしたがうことが知られている。

いま、もう一つの母集団からの標本 $(Y_1, Y_2, ..., Y_n)$ についても、標本平均を

$$\bar{Y}=(Y_1+Y_2+\cdots+Y_n)/n$$

とすると、母集団分布が $N(\mu_2, \sigma_2^2)$ のとき、同様に、$\bar{Y}$ は正規分布 $N(\mu_2, \sigma_2^2/n)$ にしたがうことになる。いま仮に、母分散 σ_1^2、σ_2^2 が既知であれば、各群の標本平均だけではなく、2 群の標本平均の差 $\bar{X}-\bar{Y}$ も正規分布 $N(\mu_1-\mu_2, \sigma_1^2/m+\sigma_2^2/n)$ にしたがうことがわかっている。

いま仮説として「2 群の母平均が等しい」つまり、

$$H_0: \mu_1=\mu_2$$

を立てる。この仮説の下では、$\mu_1-\mu_2=0$ なので、$\bar{X}-\bar{Y}$ の標本分布が $N(0, \sigma_1^2/m+\sigma_2^2/n)$ にしたがうことがすぐにわかる。つまり、標準化すると

$$Z=(\bar{X}-\bar{Y})/(\sigma_1^2/m+\sigma_2^2/n)^{1/2} \qquad (3.1)$$

が正規分布 $N(0,1)$ にしたがうことになる。

しかし、一般には、母分散は未知でわからない。そこで、母分散の代わりに、不偏標本分散を使うことを考えよう。ここで、不偏分散(unbiased variance)と呼ばれるのは、その期待値が母分散に一致するからである。母分散の代わりに標本分散を使うことで、分布は正確には正規分布ではなくなってしまう。正規分布と似た形をした t 分布あるいはステューデント(Student)の t 分布と呼ばれる分布になるのである。このことは、ゴセット(William Gosset, 1876-1937)によって見いだされたもので、ステューデントとは彼が論文を書く際に使ったペンネームである。t 分布は、標準正規分布 $N(0,1)$ と同様に平均 0 について左右対称で、標本の大きさが大きい場合には、標準正規分布 $N(0,1)$ とほとんど変わらない。特に $k=\infty$ のときは、標準正規分布 $N(0,1)$ と一致する。

こうしたことがわかっているので、標本分散を使うには次の二つのケースが考えられる。

(a)母分散 σ_1^2、σ_2^2 が未知ではあるが、等しいとき。

このとき、2 群を合併して計算した合併した分散(pooled variance)

$$s^2 = \{\Sigma(X_i-\bar{X})^2+\Sigma(Y_j-\bar{Y})^2\}/(m+n-2) \qquad (3.2)$$

を考えると、(3.1)式での σ_1^2、σ_2^2 の代わりに s^2 を代入した確率変数

$$T=(\bar{X}-\bar{Y})/(s^2/m+s^2/n)^{1/2} \qquad (3.3)$$

が自由度 $m+n-2$ の t 分布 $t(m+n-2)$ にしたがうことがわかっている。したがって、標本平均と分散を(3.3)式に代入して求めた値を t とすると、確率変数 T の絶対値がこの実際の観測された値 t の絶対値以上となる確率

$$Pr(|T|>|t|)=\alpha$$

を計算して求めることができる(実際にはコンピュータが計算してくれる)。つまり、「標本平均の差の絶対値がこれだけ大きくなる確率は 100α%しかない」ということが計算できるのである。第 1 章第 6 節でも述べたように、統計学では、通常この 有意確率が 5%未満 のとき、「もし仮説が正しければ、標本平均の差がこれだけ大きな値をとることは めったに起きない」と判断する。したがって、標本から求めた t の値という事実から、母集団についての仮説 $H_0: \mu_1=\mu_2$ は棄却されることになる。つまり、仮説は正しくなく、母平均は等しくないと判断するのである。

逆に、もしこの有意確率が 5%以上のときは、母集団についての仮説 $H_0: \mu_1=\mu_2$ が正しくても、この程度の t の値は標本誤差として標本抽出上ありうることと考え、仮説は採択される。つまり、母平均は等しいと判断するのである。

(b)母分散 σ_1^2、σ_2^2 が未知であり、等しいとは限らないとき。

この場合は、$\bar{X}-\bar{Y}$ の正確な標本分布を求めることはできない。近似的に求める方法として、ウェルチの近似法(Welch's test)が知られている。各群での標本分散を

$$s_1^2=\{\Sigma(X_i-\bar{X})^2\}/(m-1)$$

$$s_2^2=\{\Sigma(Y_j-\bar{Y})^2\}/(n-1)$$

とすると、確率変数

$$T=(\bar{X}-\bar{Y})/(s_1^2/m+s_2^2/n)^{1/2}$$

が近似的に、自由度が

$$\nu=(s_1^2/m+s_2^2/n)^2 /\{(s_1^2/m)^2/(m-1)+(s_2^2/n)^2/(n-1)\}$$

にもっとも近い整数 ν^*の t 分布 $t(\nu^*)$ にしたがうことが知られている。

(2)分散比の検定

それでは、2 群の母分散が等しいかどうかは、何を基準に判断すればよいのだろうか。それには、仮説として「2 群の母分散が等しい」つまり、

$$H_0: \sigma_1^2=\sigma_2^2$$

を立てる。いまこの仮説: $\sigma_1^2=\sigma_2^2$ が正しければ、母分散比(variance ratio) $\sigma_1^2/\sigma_2^2=1$ となっているはずである。しかし、実際に観測された標本分散比 f は 1 にはならない。問題は、観測された標本分散比の値が 1 からずれた分を標本誤差と考えてしまってよいか、それとも、標本誤差では考えにくい(つまり、確率的には起こりそうもないか)ということである。実は、標本分散比

$$F=s_1^2/s_2^2$$

が自由度 $(m-1, n-1)$ の F 分布 $F(m-1, n-1)$ にしたがうことが知られているので、この標本分散比を使って、その確率を計算して求めてやることができる。実際には、

$$F'=\max\{s_1^2, s_2^2\} /\min\{s_1^2, s_2^2\}$$

と 1 未満の値をとらないように定義された、折り重ね形式の F 統計量(folded F statistics)と

呼ばれるものを使って、

$$Pr(F' > f') > \alpha$$

を計算することになる。この確率が計算できると、標本分散比が 1 からずれた値 f' をとる確率が 100α%であることがわかる。統計学では、通常この 有意確率が 5%未満 のとき、もし仮説が正しければ、標本分散比が 1 からこれだけずれた値をとることは めったに起きない ことだと判断する。したがって、標本から求めた標本分散比 f' の値という事実から、母集団についての仮説 $H_0: \sigma_1^2 = \sigma_2^2$ は棄却されることになる。つまり、母分散は等しくないと判断するのである。

逆に、もしこの有意確率が 5%以上のときは、母集団についての仮説 $H_0: \sigma_1^2 = \sigma_2^2$ が正しくても、この程度の標本分散比 f' の値は標本誤差として標本抽出上ありうることと考え、仮説は採択される。つまり、母分散は等しいと判断するのである。

(3)SAS による 2 群の平均値の比較

2 群の平均値を t 検定(*t*-test)で比較する SAS プログラムは、文字どおり TTEST というプロシジャとして用意されている。その使用法はいたって簡単である。

PC 版 SAS

```
LIBNAME libname '¥パス名';
PROC TTEST DATA=libname.永久 SAS データ・セット名本体;
CLASS  分類変数;
VAR  平均値を求める変数リスト;
RUN;
```

例)

```
LIBNAME SAVE '¥MYDIR';
PROC TTEST DATA=SAVE.JPC;
        CLASS I1;
        VAR AGE;
RUN;
```

CMS 版 SAS

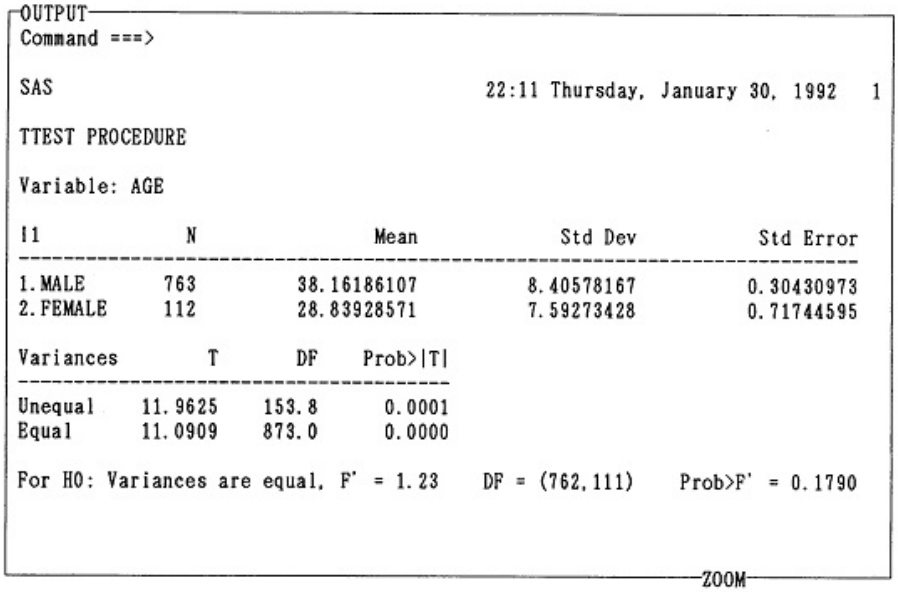
```
PROC TTEST DATA=永久 SAS データ・セット名;
CLASS  分類変数;
VAR  平均値を求める変数リスト;
RUN;
```

例)

```
PROC TTEST DATA=SAVE.JPC;
        CLASS I1;
        VAR AGE;
RUN;
```

CMS 版 SAS のプログラムは、形式的には PC 版 SAS のプログラムの 1 行目を削除したものになる。もともと「組織活性化のための従業員意識調査」の年齢階層を表す変数 I2 はプリコーディングしてあるので、年齢の階級が既に設定されている。プログラム例の中の変数 AGE は、その階級の代わりに、階級値を与えて作ったものである(→演習問題 3.3)。分類変数は数値変数でも文字変数でもかまわない。このプログラムを実行すると、PC 版 SAS では、20 秒弱で図 3.14 のような結果が画面に出力される。

図 3.14　SAS による平均値の差の検定(t 検定)



```
OUTPUT
Command ===>

SAS                                        22:11 Thursday, January 30, 1992   1

TTEST PROCEDURE

Variable: AGE

I1              N              Mean           Std Dev           Std Error
-------------------------------------------------------------------------
1.MALE        763        38.16186107        8.40578167        0.30430973
2.FEMALE      112        28.83928571        7.59273428        0.71744595

Variances         T        DF     Prob>|T|
-----------------------------------------
Unequal      11.9625     153.8       0.0001
Equal        11.0909     873.0       0.0000

For HO: Variances are equal, F' = 1.23    DF = (762,111)    Prob>F' = 0.1790

                                                                      ZOOM
```

この結果は次のように読むことになる。まず、男(1. MALE)、女(2. FEMALE)の年齢の平均(Mean)はそれぞれ、38.16･･･歳、28.83･･･歳で、これだけで見ても、ほぼ 10 歳も年齢差があり、男の方が、年齢が高いと言えそうであることがわかる。標準偏差(Std Dev; 2 乗したものが分散)もほとんど等しいので、平均だけがずれた、ほぼ同じような形の分布になっているといえる。

実際、この標本分散比(F')は 1.23 とほとんど 1 である。より正確に言えば、自由度(DF)が 763-1=762、112-1=111 の F 分布にしたがうことがわかっているので、標本分散比がこれ以上この値(F')からずれる確率(Prob>F')は 0.1790 で、母分散は等しいと判断される。したがって、分散 (Variances)が等しい(Equal)ときの t の値を見ればよい。この値(T)は、11.0909 で、自由度(DF)が 763+112-2=873 の t 分布にしたがうので、標本平均の差の絶対値が $|t|=11.0909$ よりも大きくなる確率(Prob>|T|)は 0.0000、つまり 1 万分の 1 未満と非常に小さくなる。このことから標本平均には有意な差があり、母平均は異なっていると結論づけられる。

もっとも、標本平均の違いがこれだけはっきりしていれば、統計学の知識がなくとも、標本平均を見ただけでも判断ができる。しかし、例えば、両群の分布の形を比較するために標準偏差を比較するというようなプロセスは初心者に限らず、しばしば見落とされがちである。統計学の知識は、こうした当然必要となる処理も含めて、われわれの日常的な判断、いわば直感的、常識的判断に対して、客観的でより精度の高い判断の根拠を提供しているのである。

9. k 群の平均値の差の検定

(1)分散分析

2 群の平均値の比較をしたように、この k 群の平均値の比較もできないだろうか。いま表 3.8 のような k 群のデータが得られ、それぞれの群での平均値が計算できたとしよう。

表 3.8　k 群のサイズと平均値

群	サイズ	平均
1	n_1	x_1
2	n_2	x_2
:	:	:
i	n_i	x_i
:	:	:
k	n_k	x_k
全体	n	$\bar{x}$

実は、2 群の平均値の差の検定を単純に k 群に拡張しようとしても、k 群の平均値の差の検定を想像することは難しい。そこで、k 群の平均値の差の検定の考え方について、簡単に説明しておこう。まず、全平方和

$$S=\Sigma_{i=1}^{k}\Sigma_{j=1}^{n_i}(X_{ij}-\bar{X})^2$$

は群内平方和(within samples sum of squares)の和と群間平方和(between samples sum of squares)に分けられるという性質がある。つまり、

$$\begin{aligned}S&=\Sigma_i\Sigma_j[(X_{ij}-\bar{X}_i)+(\bar{X}_i-\bar{X})]^2\\&=\Sigma_i\{\Sigma_j(X_{ij}-\bar{X}_i)^2+\Sigma_j 2(X_{ij}-\bar{X}_i)(\bar{X}_i-\bar{X})+\Sigma_j(\bar{X}_i-\bar{X})^2\}\\&=\Sigma_i\{\Sigma_j(X_{ij}-\bar{X}_i)^2+2(\bar{X}_i-\bar{X})\Sigma_j(X_{ij}-\bar{X}_i)+n_i(\bar{X}_i-\bar{X})^2\}\end{aligned}$$

ここで、$\Sigma_j(X_{ij}-\bar{X}_i)=0$ であるから、

$$S=\Sigma_i\Sigma_j(X_{ij}-\bar{X}_i)^2+\Sigma_i n_i(\bar{X}_i-\bar{X})^2\} \qquad (3.4)$$

つまり、

全平方和＝群内平方和の和＋群間平方和

となるわけである。

そこで、本代に戻ろう。「k 群の平均値の差の検定」とはより正確には「k 群の標本の母平均間の差の検定」のことである。つまり、母集団の平均についての仮説

$$H_0:\mu_1=\mu_2=\cdots=\mu_k$$

が正しいかどうかが問題となる。この仮説のような関係が標本でも成立していれば、各群の平均値は等しく、全体の平均値とも一致しているはずなので、(3.4)式の右辺第 2 項の群間平方和の部分は 0 となっているはずである。しかし、たとえ仮説が正しくても、標本抽出の際の偶然によって生じる標本誤差のために、これが正確に 0 になるというようなことはめったに起きない。それでは、その標本誤差はどの程度の大きさになるのだろうか。言い方を変えると、群間平方和がどのくらい大きければ、標本誤差の範囲を逸脱して、母平均間に差があると考えざるをえなくなるのだろうか。

まず、(3.4)式から、

群内平方和の和＝全平方和－群間平方和　　(3.5)

となる。ここで、(3.4)式の右辺第 2 項の群間平方和の定義式を見ても分かるように、群間平方和は、各群の平均値と全体の平均値との差の 2 乗を各群のサイズで加重して加えたものになっていることが分かる。つまり、平均値の差の平方和である。さらに、群間で平均値が異なるために生じたこの群間平方和を除いた残りの群内平方和の和は偶然によって生じた誤差平方和(error sum of squares)と考えられるので、(3.5)式は

誤差平方和＝全平方和－平均値の差の平方和

と書くこともできる。この誤差平方和(＝群内平方和の和)を単位、基準にして、平均値の差の平方和(＝群間平方和)を計り、評価してやることを考えればよい。

そこで、表 3.9 のような表が作成される。この表の中にある「平均平方和」(mean squares)とは「平方和」(sum of squares)を自由度で割って平均したもので、不偏分散のことである。したがって、この表によって計算された $F=V_A/V_E$ は、平均値の分散が誤差の分散の何倍あるのかをみたものになっている。実は、この F が自由度 $(k-1, n-k)$ の F 分布にしたがうことが知られている。このことを利用して検定が行われるために、この検定は F 検定とも呼ばれる。

表 3.9　分散分析表

要因	平方和	自由度	平均平方和	F 値
群によるもの	S_A	$k-1$	$V_A=S_A/(k-1)$	$F=V_A/V_E$
誤差	S_E	$n-k$	$V_E=S_E/(n-k)$	
全体	S	$n-1$		

ところで、この表は分散分析表(analysis of variance table)と呼ばれるが、正確にはこうした一連の手続きで、一元配置(one-way layout)の分散分析 (analysis of variance)が行われたことになる。しかし、ここで「一元配置」や「分散分析」の意味をこれ以上述べても、平均値の差の検定という目的にプラスにならないので、これ以上の説明はしないことにしよう。

既に述べたように、k 群の平均値の差の F 検定は、考え方としては、2 群の平均値の差の t 検定の単純な拡張をしたわけではない。しかし、2 群の平均値の差の検定については、F 検定も t 検定も実質的には同じことをしていることになる。実は、2 群の平均値の差の F 検定で求めた F 値は、2 群の平均値の差の t 検定で求めた t 値の 2 乗になっているのである。実際、 $F=V_A/V_E$ のうち、分母 V_E は、

$$V_E=S_E/(n-2)=\{\Sigma(X_{1j}-\bar{X}_1)^2+\Sigma(X_{2j}-\bar{X}_2)^2\}/(n_1+n_2-2)=s^2$$

すなわち、t 検定の際の合併した分散(3.2)式と等しくなる。また

$$\bar{X}=(n_1\bar{X}_1+n_2\bar{X}_2)/(n_1+n_2)$$

に注意すると、分子 V_A については、

$$V_A=S_A/1=n_1(\bar{X}_1-\bar{X})^2+n_2(\bar{X}_2-\bar{X})^2=(\bar{X}_1-\bar{X}_2)^2/(1/n_1+1/n_2)$$

したがって、 $F=V_A/V_E$ は(3.3)式の T の 2 乗に等しいことになる。このように、t 分布と F 分布の間には、「量 T が自由度 $n-2$ の t 分布にしたがうときは、量 $F=T^2$ は自由度 $(1, n-2)$ の F 分布にしたがう」という関係があったのである。つまり、2 群の平均値の差の t 検定を単純に拡張して、k 群の平均値の差の F 検定を思いつくことは難しくても、k 群の平均値の差の F 検定の特別の場合として、2 群の平均値の差の F 検定を考え、その F の定義式の分子分母のルートをとることで、t 検定の t の定義式を導出することはできるのである。

(2)SAS による k 群の平均値の比較

SAS によって k 群の平均値を比較するには、GLM プロシジャを使う方法が考えられる。実は、ANOVA プロシジャを使うこともできるが、GLM プロシジャの方が、汎用性が高く、使い方を覚えていると得である。また本書では扱わないが、実験計画法に基づかない社会科学分野のデータに対しては、特に 2 元配置以上の分散分析には ANOVA プロシジ

ャの使用は勧められないというような事情もある。GLM プロシジャを使うと、次のようになる。

PC 版 SAS

```
LIBNAME libname '¥パス名';
PROC GLM DATA=libname.永久 SAS データ・セット名本体;
CLASS  分類変数;
MODEL  平均をとる変数=分類変数;
MEANS  分類変数;
RUN;
```

例)

```
LIBNAME SAVE '¥MYDIR';
PROC GLM DATA=SAVE.JPC;
        CLASS I4;
        MODEL AGE=I4;
        MEANS I4;
RUN;
```

CMS 版 SAS

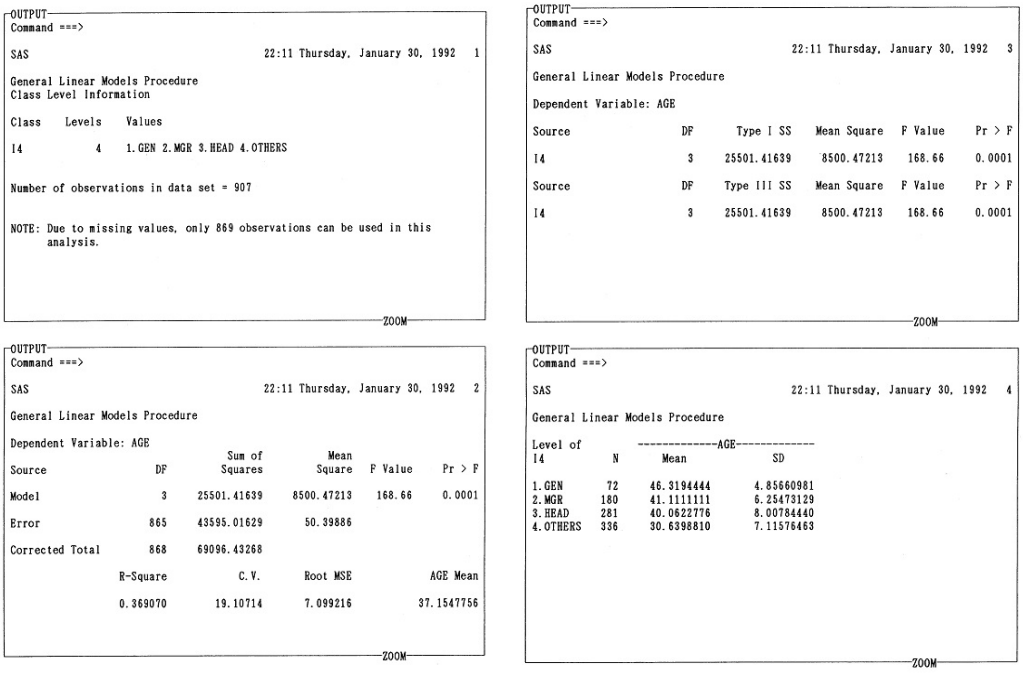
```
PROC GLM DATA=永久 SAS データ・セット名;
CLASS  分類変数;
MODEL  平均をとる変数=分類変数;
MEANS  分類変数;
RUN;
```

例)

```
PROC GLM DATA=SAVE.JPC;
        CLASS I4;
        MODEL AGE=I4;
        MEANS I4;
RUN;
```

CMS 版 SAS のプログラムは、形式的に PC 版 SAS のプログラムの 1 行目を削除したものになる。ここで、プログラム例の中の変数 I4 は職位を表す変数である。変数 AGE は前節で作った年齢を表す変数である。このプログラムを実行すると、PC 版 SAS では約 40 秒で図 3.15 のように 4 画面にわたって結果が出力される。

図 3.15　SAS による　k 群の平均値の比較



この結果の第 2 画面が分散分析表に相当している。第 3 画面も類似の表示がなされているが、今回の比較には関係がないので、第 2 画面をもとにして、自由度(DF)、平方和(Sum of Squares)、平均平方和(Mean Square)、F 値(F Value)に注意しながら、分散分析表の形に対応させて書き直すと、表 3.10 のようになる。ここで、*F* 値につけた「***」は、有意確率(Pr > F)が 0.0001 なので、当然 0.1%水準で有意ということで表示したものである。つまり、分散分析の結果、*k* 群による平均値の差は有意ということになる。変数 AGE の全体の平均(AGE Mean)は第 2 画面に 37.1547756 と表示され、各群での平均は第 4 画面に表示されているので、これらをまとめると表 3.11 のようになる。

第 2 画面に表示されている全体の自由度 868 が、第 1 画面の"NOTE"に表示されている全体の N に相当するオブザーベーション数 869 から 1 引いたものになっていることも確認してほしい。検定結果が示唆していたように、この表 3.11 によって、各群での年齢の平均の違いが明らかにあることがわかる。課長クラスと係長クラスの平均年齢の差は僅差であるが、部長クラスと一般とでは、実に 16 歳も平均年齢が違うのである。(→演習問題 3.3)

表 3.10　分散分析表

要因	平方和	自由度	平均平方和	*F* 値
群によるもの	25501.41639	3	8500.47213	168.66***
誤差	43595.01629	865	50.39886	
全体	69096.43268	868		

表 3.11　カテゴリーごと(各群)の平均値

	N	平均
1. 部長クラス(GEN)	72	46.3
2. 課長クラス(MGR)	180	41.1
3. 係長クラス(HEAD)	281	40.1
4. 一般(OTHERS)	336	30.6
全体	869	37.2

演習問題

3.1 階級の設定と棒グラフ　第 2 節(5)の例題の解答例 1、解答例 2 の棒グラフを SAS を使って描いてみよ。

3.2 分散と変動係数　第 6 節の図 3.11 を参考にしながら、2 値質的変数の変動係数を求めることの意義について、分散と対比しながら述べよ。

3.3 平均値の比較と棒グラフ　プリコーディングしてある年齢の各カテゴリーに対して、その階級値を与えた変数として変数 AGE を SAS プログラム上で定義した上で、次のような統計処理を SAS を使って行なってみよ。

1. 階級を作らずに、そのまま変数 AGE の度数分布表を作成せよ。
2. SAS にこの変数 AGE の棒グラフを描かせてみよ。

3. 第 8 節のプログラム例のように、男・女の 2 群の平均値を比較してみよ。
4. 第 9 節のプログラム例のように、部長クラス・課長クラス・係長クラス・一般の 4 群の平均値を比較せよ。

第 4 章　相関と回帰

章目次

1. はじめに
2. 散布図
3. 相関
4. 単回帰と相関係数

1. はじめに

いま 2 変数 x, y のデータが与えられている場合を考えよう。例えば、一般的には個人ごとの身長と体重のデータ、あるいは前章冒頭の例でいえば、個人ごとの性別、職種、職位のデータ、そして表 4.1 のような企業の資本金と従業員数のデータなどである。

表 4.1　私鉄大手 16 社比較(1990 年 3 月末現在)

会社名	資本金(百万円)		従業員数(人)	
東　急	106,710	(1)	7,028	(6)
近　鉄	89,851	(2)	11,873	(1)
阪　急	68,578	(3)	5,373	(7)
名　鉄	65,664	(4)	8,268	(5)
東　武	62,971	(5)	11,017	(2)
営　団	58,100	(6)	10,572	(3)
小田急	57,606	(7)	3,911	(13)
京　王	48,845	(8)	4,236	(12)
京　阪	40,049	(9)	3,371	(14)
京　急	31,396	(10)	4,741	(11)
阪　神	27,829	(11)	2,230	(16)
相　鉄	27,143	(12)	2,637	(15)
西　鉄	25,589	(13)	8,486	(4)
南　海	22,162	(14)	4,855	(9)
西　武	21,665	(15)	5,239	(8)
京　成	13,583	(16)	4,805	(10)

(　)内は順位を示す。

表 4.1 はもともと「組織活性化のための従業員意識調査」の対象企業の中に、私鉄大手のうちの何社かが含まれていたために、それらの企業の規模を相互に比較したり、あるい

は業界の中での位置づけを行ったりするために集められたデータである。通常、企業規模を表現するためによく用いられる資本金や従業員数をピックアップして、資本金順に並べてみた。しかし、明らかに従業員数順にはなっていない。両者を組み合わせたときに対象企業の位置づけがどういった結論になるのか、この表だけではピンとこない。

こうした 2 変数の間の関係の分析は 1 変数の場合とほぼ同様に次のような手順で行われる。

- 《ステップ 1》図・表によってデータを整理
- 《ステップ 2》次のような方法で、数値によって要約(3 変数以上の場合もほぼ同様)
    1. 相関(correlation): x と y を対等に見て、相互の関係を扱う
    2. 回帰(regression): x から y (y から x)が決定される一方向の関係を扱う

それではまず、ステップ 1 の図・表によってデータを整理する方法についてから解説することにしよう。初心者はこのステップを軽視して、いきなり相関や回帰をしたがるが、こんな雑で荒っぽいことをしていては、分析が進んでしまった後で、後悔することになる。このステップ 1 で図・表によってデータをきちんと整理して、データの特性をおおまかにつかんでおかないと、ステップ 2 でどのような方法を用いたらよいのか、また得られた数値が、本当にデータを要約したことになっているのかについて、基本的な判断を間違えることになる。

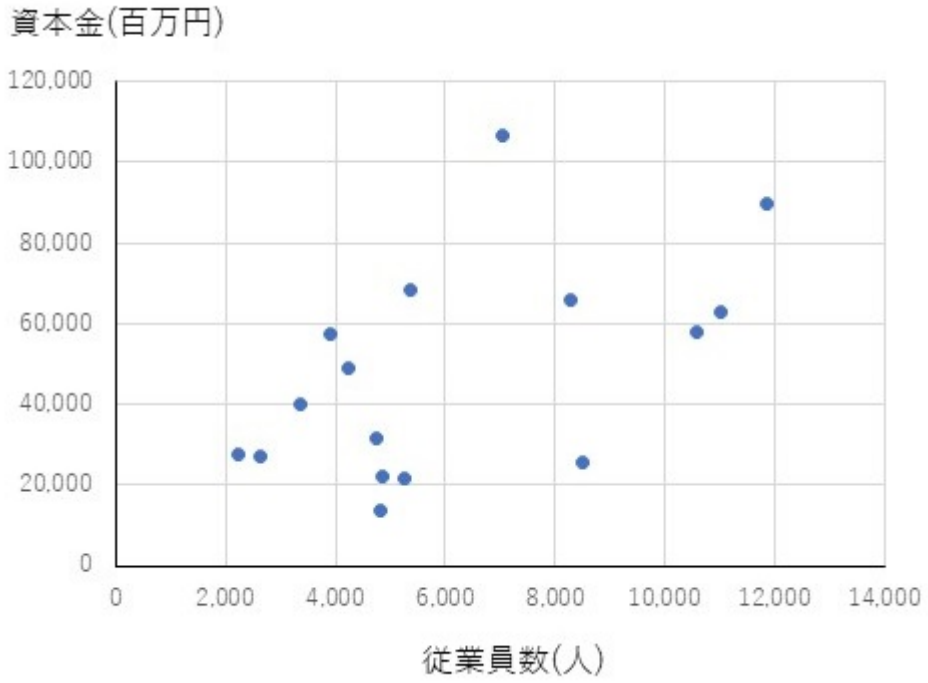
この章では、量的データの整理・要約をまず説明し、それからより頻繁にかつ一般的に用いられる質的データの整理・要約について説明することにしよう。

2. 散布図

量的データの場合には、相関表という表形式、あるいは散布図(scatter diagram または scattergram)という図によってデータを整理する。このうち相関表については、次の第 5 章で取り上げることにして、この章では散布図について説明しよう。散布図とは、例えば図 4.1 は表 4.1 の資本金と従業員数の 2 変数のデータを基にして描いた散布図であるが、このように、平面上に直交座標を定めて、横軸に変数 x (図 4.1 では従業員数)、縦軸に変数 y (図 4.1 では資本金)をとり、2 変数データ (x_i, y_i), i=1, 2, ..., n をこの座標平面上に点(1 点＝度数 1)で記入したものである。

散布図を一度自分で手書きで描いてみるとわかることだが、このようなせいぜい十数個の点をプロットするだけでも、実は大変な作業量である。グラフ用紙に向かって、縦軸、横軸の目盛をどの程度の細かさで設定するかを全て決め、その上で 1 点 1 点をプロットしていくのである。しかも、この程度の個数でも、どの点がどの会社であったか、描き終った時点ではもはやわからなくなっている。

図 4.1　私鉄大手 16 社の資本金と従業員数の散布図



散布図は、SAS を使えば簡単に描くことができる。次のように PLOT プロシジャを用いればよい。

```
PROC PLOT [DATA=SAS データ・セット];
PLOT  変数 1*変数 2[=変数 3];
RUN;
```

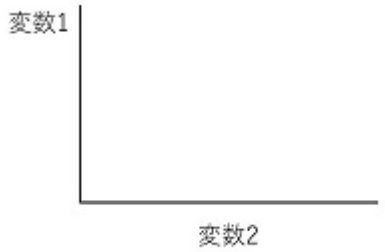
ここで変数 1、変数 2 は変数のリストでもよいことになっている。変数 3 は文字変数でも数値変数でもよく、その変数の値の最初の 1 文字が、「点」の代わりに表示されることになる。変数 3 の代わりに、PLOT X*Y="*"; のように"*"などの文字定数を置くと、やはりその文字定数の最初の 1 文字が表示される。変数 3 が指定されなければ、度数によって、表 4.2 のように表示される。実は、PLOT プロシジャによって作成されるのは、正確には散布図ではなく、後述するクロス表を非常にきめ細かくしたものなのである。したがって、いかにセルが細かく設定されているとはいっても、一つのセルに複数の「点」を「プロット」しなくてはならないような事態も散布図に比べると頻繁に生じることになるので、こうした度数表示が行われるのである。

表 4.2　SAS の散布図における度数の表示

度　数	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	・・・・・・
表　示	A	B	C	D	E	F	G	H	I	J	K	L	・・・・・・

変数 1 と変数 2 の指定により、散布図の軸は図 4.2 のように設定されることになる。

図 4.2　指定された変数と散布図の軸



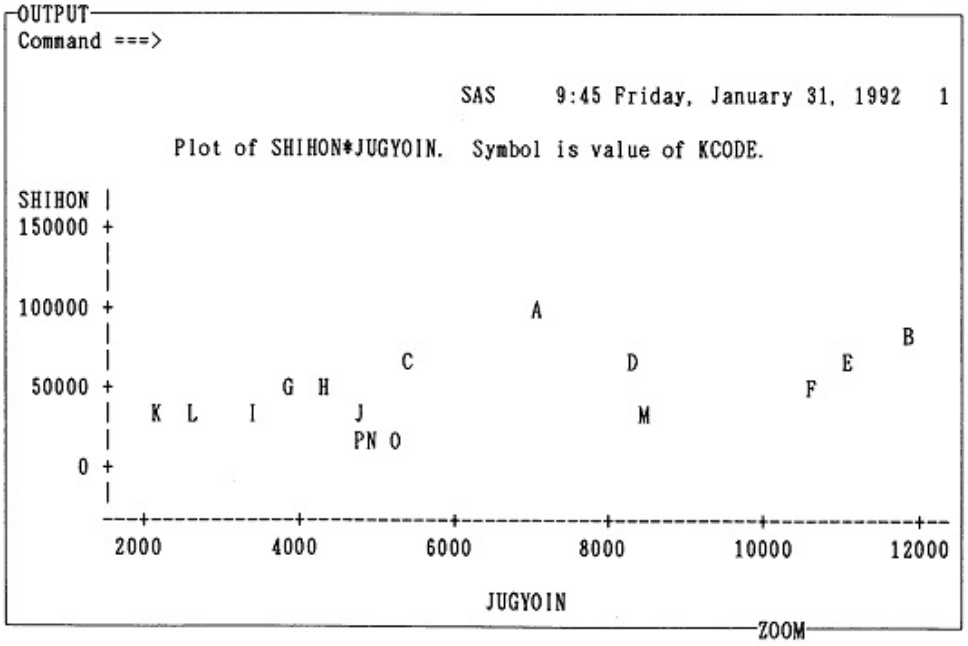
ところで、「組織活性化のための従業員意識調査」の質問票調査では、得られるのは質的データばかりなので、散布図を描くのに適したデータがない。そこで、ここでは、先ほどの 1990 年 3 月末現在での私鉄大手 16 社の資本金と従業員数のデータを使った SAS のプログラム例を考えてみよう。このデータについては、まだ永久 SAS データ・セットが存在していないので、ここでは、SAS プログラム内に直接データを書き込んでおいて、しかも新たに永久 SAS データ・セットをついでに作ってしまう PC 版 SAS のプログラム例を挙げておこう。

```
LIBNAME SAVE '¥MYDIR';
DATA SAVE.SHITETSU;
INPUT KCODE $ SHIHON JUGYOIN;
CARDS;
A 106710    7028
B  89851   11873
C  68578    5373
D  65664    8268
E  62971   11017
F  58100   10572
G  57606    3911
H  48845    4236
I  40049    3371
J  31396    4741
K  27829    2230
L  27143    2637
M  25589    8486
N  22162    4855
O  21665    5239
P  13583    4805
;
PROC PLOT;
     PLOT SHIHON*JUGYOIN=KCODE;
RUN;
```

このプログラムの 1 行目を削除すれば、そのまま CMS 版 SAS のプログラム例となる。PC 版 SAS では、このプログラム例を実行すると、20 秒弱で永久 SAS データ・セット SHITETSU.SSD がディレクトリ MYDIR の下に生成されるとともに、図 4.3 のような結果が

出力される。CMS 版 SAS で生成される永久 SAS データ・セット名は SHITETSU SAVE となる。

図 4.3　SAS による私鉄大手 16 社の資本金と従業員数の散布図



3. 相関

(1)相関関係

二つの変数の間で、一方の値が決まると他方の値が一意に定まる関係があるとき、両変数の間には関数関係があるという。それに対して、ここでいう相関関係とは、二つの変数の間で、一方の値が決まると他方の値が一意に定まるというわけにはいかないまでも、両者の間になんらかの関連性が認められる関係をさしている。例えば、通常、人間の身長と体重の間には、身長が高いほど体重も重く、体重が重いほど身長も高いという関係があるということは常識的に理解できる。しかし、身長が決まると体重も一意に決まるというような関数関係があるわけではない。同じ身長でも、細目の人から太目の人まで体重には大きな差があるものである。

統計学で用いられる相関関係は、特に直線的な関係を指しているものである。つまり、次のような関係を指している。

1. 正の相関であれば、二つの変数の間に、一方が増加すれば他方もおおむね増加するという傾向があること。
2. 負の相関であれば、二つの変数の間に、一方が増加すれば他方もおおむね減少するという傾向があること。

もちろん、「相関」は統計学以外でも使われる言葉であるが、統計学における相関(correlation)は、あくまでも変数間の単調な増減関係に限定して用いられることに注意してほしい。

また、相関関係と因果関係とは別の概念であるということにも注意がいる。相関関係は観測されたデータから、表面的に認められる事実関係のみを意味するのに対して、因果関係は論理的に考えられるものである。統計学では事実としての相関関係を扱うことはでき

ても、それが本当に因果関係を意味するものかどうかを統計学で直接判定することはできない(第 1 章第 2 節を参照のこと)。

既に述べた、散布図や相関表などの図・表によってデータを整理して、相関関係を視覚に訴えることができる。この方法を補完するものとして相関係数がある。相関係数は、2 変数間の相関関係のうち、特に直線的な関係の強さを測る指標である。この節では、この相関係数についてまとめておこう。

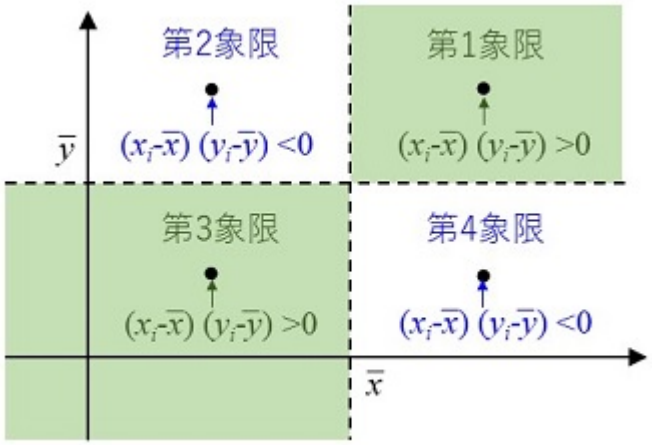
(2)相関係数

通常、単に「相関係数」といった場合には、ピアソンの積率相関係数(Pearson's product-moment correlation coefficient あるいは Pearson's r)を指すことになる。2 変数データ $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ のピアソンの積率相関係数は次のように定義される。

$$r_{xy} = \{\Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\} / [\{\Sigma_{i=1}^{n} (x_i - \bar{x})^2\}^{1/2} \{\Sigma_{i=1}^{n} (y_i - \bar{y})^2\}^{1/2}]$$

この相関係数は何を意味しているのだろう。定義式の分母は平方和の平方根の積なので、その符号は正に決まっている。それでは定義式の分子はどうだろうか。相関係数の定義式の分子に注目して、相関係数の意味を考えてみよう。そこで、平均 $(\bar{x}, \bar{y})$ を原点とする直交座標を考えると、図 4.4 のようになる。

図 4.4　相関係数の分子と平均



つまり、相関係数の定義式の分子の Σ の中身 $(x_i - \bar{x})(y_i - \bar{y})$ は、第 1 象限や第 3 象限にある点で計算すると正、第 2 象限や第 4 象限にある点で計算すると負の値をとることになる。したがって、

1. データが主に第 1 象限と第 3 象限(影の部分)に散布していれば(正の相関のケース)、相関係数は正の値をとる: $\Sigma(x_i - \bar{x})(y_i - \bar{y}) > 0$
2. データが主に第 2 象限と第 4 象限(白地の部分)に散布していれば(負の相関のケース)、相関係数は負の値をとる: $\Sigma(x_i - \bar{x})(y_i - \bar{y}) < 0$
3. データが各象限にまんべんなく散布していれば(相関がないケース)、相関係数は 0 に近い値をとる: $\Sigma(x_i - \bar{x})(y_i - \bar{y}) \fallingdotseq 0$

ということになる。さらに付言すれば、直線 $x = \bar{x}$ について左右対称、または直線 $y = \bar{y}$ について上下対称の分布をしていれば、相関係数は $r_{xy} = 0$ ということもわかる。さらに図 4.4 の四つの象限が平均値線で区切られていることからも想像がつくと思うが、平均が少数の外れ値、異常値に大きく影響を受けるように、相関係数もまた少数の外れ値、異常値に驚くほ

ど大きな影響を受けることになる(→演習問題 4.4)。したがって、相関係数が 0 であっても、変数間には明白な関係が存在することも多いし、逆に、少数の外れ値のために、相関係数が異常に大きくなることも多い。本当にデータを生かして変数間の関係を調べるためには、やはり散布図が重要なのである。

このようなことから、相関係数は、散布図など相関関係を視覚に訴える方法を補完するものであると考えていた方が間違いがない。相関係数は 2 変数間の直線的な関係を要約するものにすぎないからである。

ところで、相関係数の定義式の分子を n で割ったもの

$$s_{xy}=\{\Sigma_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\}/n$$

は変数 x と y の共分散(covariance)と呼ばれる。変数 x と x の共分散は、定義より、

$$s_{xx}=s_x^2$$

つまり、変数 x の分散となる。この共分散を用いれば、相関係数は

$$r_{xy}=s_{xy}/(s_x s_y)$$

と表すこともできる。この形はこの章の中でもたびたび用いられるので、記憶にとどめておいてほしい。

相関係数 r_{xy} の定義では、分母 $s_x s_y\neq 0$ を暗黙のうちに仮定していることになる。しかし、これは実際上はなんら制約にならない。なぜなら、例えば、$s_x=0, s_y\neq 0$ であれば、変数 x の標準偏差、分散は 0 ということになり、x は一定値の定数だったことになるからである。したがって、このデータは実質的に 2 変数データではなかったことになり、もはや相関関係を考える必要はない。

(3)相関係数の性質

次に、相関係数のもっている性質を列挙しておこう。これらの性質は、計算ミスを見つける際や、SAS のような統計パッケージを使う際に、常識として知っておかねばならない性質ばかりを集めたものである。特に、性質 1、性質 2 は、相関係数が相関関係の客観的指標になりうることを示すものであるが、知っていれば統計パッケージ使用の際の色々な手間を省くのに役立つ。

性質 1　相関係数 r_{xy} は x と y とを入れ換えても変化しない。すなわち $r_{xy}=r_{yx}$

性質 2　相関係数 r_{xy} の値は、両変数の 1 次変換 $x'=ax+b, y'=cy+d, ac>0$ によっても変わらない。すなわち $r_{x'y'}=r_{xy}$

<証明>

$$r_{x'y'}=[\Sigma\{ax_i+b-(\overline{ax+b})\}\{cy_i+d-(\overline{cy+d})\}]$$

$$/[[\Sigma\{ax_i+b-(\overline{ax+b})\}^2]^{1/2}[\{\Sigma cy_i+d-(\overline{cy+d})\}^2]^{1/2}]$$

$$=\{ac\Sigma(x_i-\bar{x})(y_i-\bar{y})\}/[a\{\Sigma(x_i-\bar{x})^2\}^{1/2}c\{\Sigma(y_i-\bar{y})^2\}^{1/2}]=r_{xy}\quad\square$$

性質 3　$-1\leqq r_{xy}\leqq 1$

<証明>

最初に t に関する次のような 2 次式を考える。

$$h(t)=\Sigma_i[a_i t-b_i]^2=t^2\Sigma a_i^2-2t\Sigma a_i b_i+\Sigma b_i^2$$

$h(t)\geqq 0$ だから、この 2 次式の判別式が正になることはなく、

$$\text{判別式}=[\Sigma a_i b_i]^2-\Sigma a_i^2\Sigma b_i^2\leqq 0$$

これは、シュルツの不等式(Schwarz's inequality)と呼ばれる。そこで $a_i=x_i-\bar{x}$, $b_i=y_i-\bar{y}$ とおくと

$$r_{xy}^2=\{\Sigma_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})\}^2 / \{\Sigma_{i=1}^n (x_i-\bar{x})^2 \Sigma_{i=1}^n (y_i-\bar{y})^2\} \leqq 1 \quad \square$$

性質 4　$|r_{xy}|=1$ の必要十分条件は (x_i, y_i), $i=1, 2, ..., n$ がすべて (x, y) を通る同一直線上にあることである。

<証明>

(十分性の証明)

(x_i, y_i), $i=1, 2, ..., n$ がすべて同一直線上 $y=ax+b$ 上にあれば、$y_i=ax_i+b$, $i=1, 2, ..., n$ となり、$\Sigma y_i=a\Sigma x_i+nb$ ゆえに、$\bar{y}=a\bar{x}+b$。つまり、$(\bar{x}, \bar{y})$ もこの直線 $y=ax+b$ 上にある。したがって、

$$(x_i-\bar{x})t-(y_i-y)=0,\ i=1, 2, ..., n \qquad (4.1)$$

いま次の 2 次式

$$h(t)=\Sigma[(x_i-\bar{x})t-(y_i-\bar{y})]^2 = t^2\Sigma(x_i-\bar{x})^2-2t\Sigma(x_i-\bar{x})(y_i-\bar{y})+\Sigma(y_i-\bar{y})^2$$

を考えると、常に $h(t)\geqq 0$ なので、t は(4.1)式から方程式 $h(t)=0$ の重根でなければならない。よって、

$$\text{判別式}=\Sigma[(x_i-\bar{x})(y_i-\bar{y})]^2-\Sigma(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2 = 0$$

これから $r_{xy}^2=1$

(必要性の証明: 背理法で証明する)

(4.1)式を満たす t が存在しないと仮定すると、

$$h(t)=\Sigma[(x_i-\bar{x})t-(y_i-\bar{y})]^2>0$$

となり、判別式は負でなければならないために $r_{xy}^2<1$ となる。これは矛盾。 よって $r_{xy}^2=1$ となるのは、(4.1)式を満たす t が存在するときに限られる。したがって、(x_i, y_i), $i=1, 2, ..., n$ はすべて同一直線上にある。□

(4)相関係数の検定

第 3 章で扱った平均値の差の検定の際の t 検定、F 検定と同様に、相関係数についても検定することができる。母集団での相関係数を ρ とおくと、仮説

$$H_0: \rho=0$$

を検定するためには、

$$t=r_{xy}(n-2)^{1/2}/(1-r_{xy}^2)^{1/2}$$

が、自由度 $n-2$ の t 分布 $t(n-2)$ にしたがうことがわかっているので、これを用いればよい。

(5)スピアマンの順位相関係数

表 4.1 の(　)内に示した資本金順位、従業員数順位のような順序尺度データに対して相関係数を適用したらどうなるだろうか。尺度の問題(第 1 章第 3 節(2)を参照のこと)を考えると乱暴としか言いようがないが、いま、表 4.1 の(　)内の順位を抜き書きして得られるような表 4.3 の形の 2 変数の順位データ $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ に相関係数を適用することを考えよう。

表 4.3　順位データ

観測対象	x	y
1	x_1	y_1
2	x_2	y_2
:	:	:
i	x_i	y_i
:	:	:
n	x_n	y_n

ここで、x_i, y_i はともに 1 から n までの整数なので、

$$\Sigma x_i = \Sigma y_i = 1+2+\cdots+n = n(n+1)/2$$

$$\Sigma x_i^2 = \Sigma y_i^2 = 1^2+2^2+\cdots+n^2 = n(n+1)(2n+1)/6$$

となる。このことを用いると、x と y の平方和は、

$$ns_x^2 = ns_y^2 = \Sigma x_i^2 - n\bar{x}^2 = n(n+1)(2n+1)/6 - n(n+1)^2/4 = n(n^2-1)/12$$

また

$$ns_{xy} = \Sigma x_i y_i - n\bar{x}\cdot\bar{y} = [-\Sigma(x_i-y_i)^2 + \Sigma x_i^2 + \Sigma y_i^2]/2 - n\bar{x}^2$$

$$= -\Sigma(x_i-y_i)^2/2 + \Sigma x_i^2 - n\bar{x}^2$$

$$= -\Sigma(x_i-y_i)^2/2 + n(n^2-1)/12$$

これから

$$r_{xy} = s_{xy}/s_x s_y = 1 - 6\Sigma(x_i-y_i)^2/\{n(n^2-1)\} \qquad (4.2)$$

となる。これは、スピアマンの順位相関係数(Spearman's rank correlation coefficient あるいは Spearman's ρ)と呼ばれる。ここでは r_S と書くことにしよう。ただし、(4.2)式は順位データが 1 から n までの整数であることを利用して簡略化した計算式であって、相関係数の定義自体はピアソンの積率相関係数と同じである。

この計算式(4.2)式の分子にある $x_i - y_i$ に着目すると、

1. 変数間の順位が完全に一致するならば、$x_i - y_i = 0,\ i = 1, 2, ..., n$ であるから、$r_S = 1$ となる。
2. 変数間の順位が完全に逆ならば、$y_i = n+1-x_i,\ i = 1, 2, ..., n$ であるから、$\Sigma(x_i-y_i)^2 = \Sigma[2x_i-(n+1)]^2 = n(n^2-1)/3$ したがって、$r_S = -1$ となる。

この両極端のケースからも分かるように、$0 < r_S \leqq 1$ のときは、変数間の順位が一致している傾向があり、$-1 \leqq r_S < 0$ のときは、変数間の順位が一致していない傾向がある。

順位相関係数は順位という粗い情報しか用いない。実は、通常の相関係数は少数個の大きな値で決まってしまい、実際には無意味な数字になることがある。このような場合、情報を少し落して、順位に変換し、おとなしくさせて、順位相関係数を求めればよいこともある。

ただし、r_S は順位を表す数値をあたかも間隔尺度であるかのように取り扱って、ピアソンの積率相関係数を求めたものなので、より正確には、順位を順序尺度として扱う順位相関係数を使うべきである。このことを考えているのが、次のケンドールの順位相関係数である。

(6)ケンドールの順位相関係数

ケンドールの順位相関係数(Kendall's rank correlation coefficient) τ (タウ)は、純粋に観測値の大小関係のみを反映した相関係数である。したがって、どちらがどれくらい大きいとか小さいとかを一切考慮しない。先ほどと同様に表 4.3 のような順位データが与えられているとしよう。このような順位データの相関係数であるケンドールの τ は、次の二つのケースに対応して、τ_a と τ_b の 2 種類の τ が考え出されている。

(a)同順位がない場合

個体二つずつの対を考えて(全部で ${}_nC_2 = n(n-1)/2$ 個の観測値の対)、この各対を次のように分類する。

- A･･･**正順** : x の大小関係の方向と y の大小関係の方向が一致しているもの。
- B･･･**逆順** : x の大小関係の方向と y の大小関係の方向が反対になっているもの。

A、B に該当する対の個数を ΣA、ΣB と表す。全部で $n(n-1)/2$ 個の対に占める、正順の対の個数 ΣA と逆順の対の個数 ΣB との差の割合を順位相関係数と考えるのである。

$$\tau_a = (\Sigma A - \Sigma B)/\{n(n\text{-}1)/2\} \qquad (4.3)$$

ここで、$n(n\text{-}1)/2 = \Sigma A + \Sigma B$ であるから、当然のことながら、$-1 \leqq \tau_a \leqq 1$ となる。τ_a が 1 や-1 をとるのは、次の場合である。

1. 全部の対が正順: x の大小関係の方向と y の大小関係の方向がすべて一致しているとき、$\Sigma A - \Sigma B = n(n-1)/2 - 0$ となり、$\tau_a = 1$。
2. 全部の対が逆順: x の大小関係の方向と y の大小関係の方向がすべて反対になっているとき、$\Sigma A - \Sigma B = 0 - n(n-1)/2$ となり、$\tau_a = -1$。

(b)同順位がある場合

既に定義した正順 A、逆順 B に加えて、次の 2 種類の同順位を考える必要がある。

- C･･･**同順位** : x の値が等しい(＝同順位)場合。
- D･･･**同順位** : y の値が等しい(＝同順位)場合。

こうして定めた A、B、C、D に該当する対の個数を ΣA、ΣB、ΣC、ΣD と表すと、

$$\tau_b = (\Sigma A - \Sigma B)/[\{n(n-1)/2 - \Sigma C\}^{1/2}\{n(n-1)/2 - \Sigma D\}^{1/2}] \qquad (4.4)$$

先ほどの(a)同様に、$-1 \leqq \tau_b \leqq 1$ となる。

同順位がない場合($\Sigma C = \Sigma D = 0$)には、(4.4)式にこのことを代入すると、(4.3)と同じ式になる。すなわち $\tau_b = \tau_a$ となる。したがって、同順位のあるなしにかかわらず、τ_b を計算すればよいことになる。

(7)SAS による相関係数の計算と検定

これまで述べてきた、ピアソンの積率相関係数、スピアマンの順位相関係数、ケンドールの順位相関係数を SAS で求めることは、非常に簡単である。まずピアソンの積率相関係数から説明すると、これには CORR プロシジャが使われる。

PC 版 SAS

```
LIBNAME libname '￥パス名';
PROC CORR DATA=libname.永久 SAS データ・セット名本体;
[VAR  変数名の並び;]
[WITH  変数名の並び;]
[OPTIONS NOCENTER;]
RUN;
```

例)

```
LIBNAME SAVE '￥MYDIR';
PROC CORR DATA=SAVE.SHITETSU;
        OPTIONS NOCENTER;
RUN;
```

CMS 版 SAS

```
PROC CORR DATA=永久 SAS データ・セット名;
[VAR  変数名の並び;]
[WITH  変数名の並び;]
[OPTIONS NOCENTER;]
RUN;
```

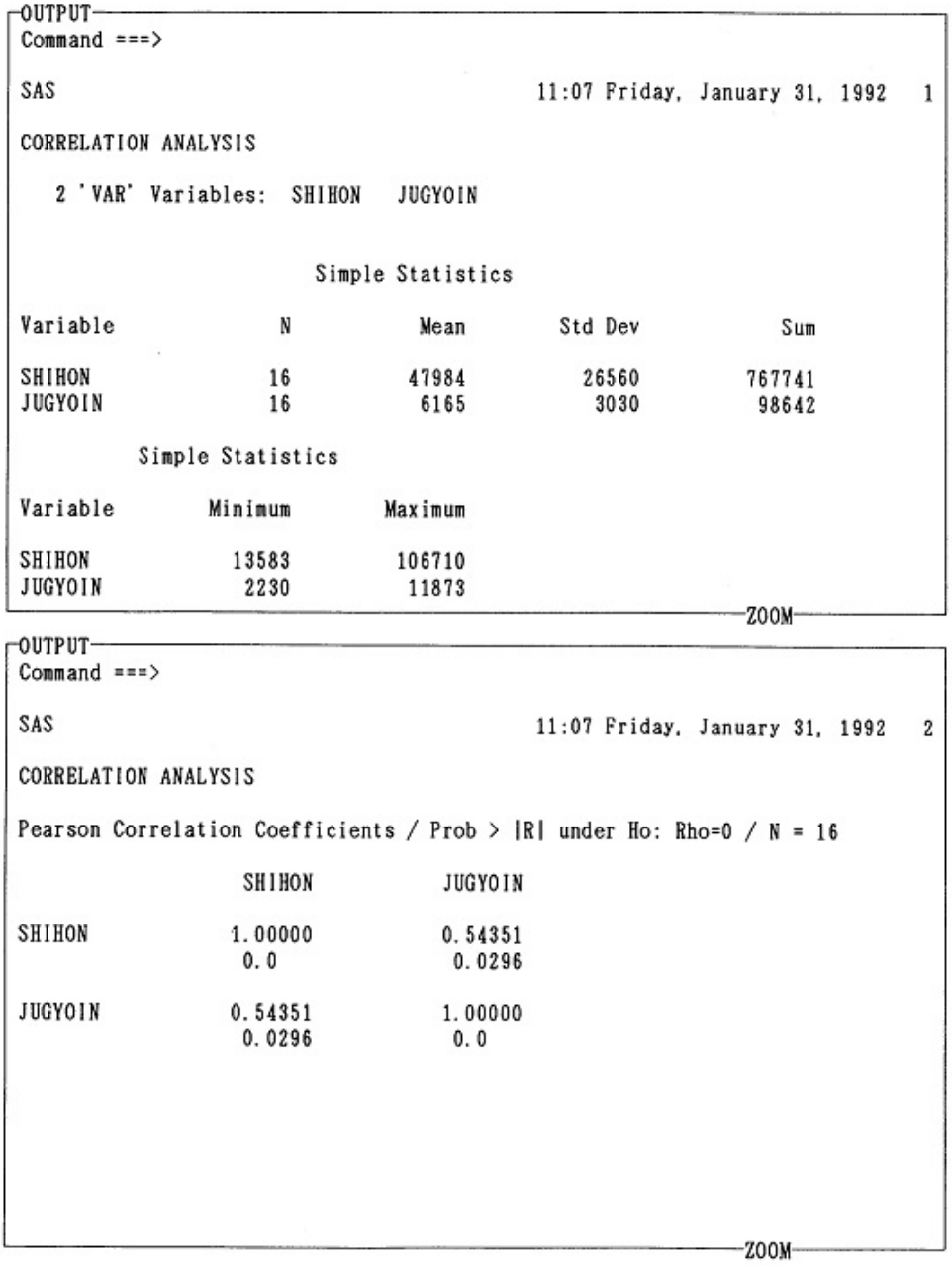
例)

```
PROC CORR DATA=SAVE.SHITETSU;
        OPTIONS NOCENTER;
RUN;
```

PC 版 SAS のプログラムの 1 行目を削除すると形式的には CMS 版 SAS のプログラムとなる。このプログラム例では、さきほど生成した永久 SAS データ・セット(PC 版 SAS では SHITETSU.SSD、CMS 版 SAS では SHITETSU SAVE)を使って、全数値変数間の相関係数を計算させている。このプログラム例を実行させると、PC 版 SAS では 5 秒以内に図 4.5 の画面が出力される。第 1 画面では、各変数のオブザーベーション数(N)、平均(Mean)、標準偏差(Std Dev)、総和(Sum)、最小値(Minimum)、最大値(Maximum)といった単純統計が表示され、第 2 画面では、相関係数行列が表示される。

第 2 画面で、相関係数行列の各セルの上段に表示されているのが相関係数、下段に表示されているのが有意確率である。この例の場合、SHIHON と JUGYOIN との間のピアソンの積率相関係数は 0.54351 で、有意確率は 2.96%ということになる。通常のように、有意水準を 5%と設定している場合には、「5%水準で有意」ということになる。ただし、この場合、ここで取り扱っている私鉄大手 16 社のデータは何か別の母集団からの無作為標本のデータというわけではない。むしろこれ自体が母集団といってよいだろう。したがって、統計学的には、相関係数は意味があるが、それを検定すること自体は無意味である。もっとも、このような場合でも、有意確率や有意水準を併記しておくのが、現在のところ「流儀」になっている。「それは本質的におかしい！」と目くじらを立てずに、この場合には、相関係数 0.54351 が「相関がある」といってもよいほどの水準の値なのかどうかの一つの判断材料になっているという程度に理解していただきたい。客観的な判断材料になりうるものに対しては貪欲であり続けるのが統計処理の基本的姿勢なのである。ただし、統計学的には、あくまでも無作為標本データでない限り、有意確率に意味はないということは頭の片隅にとどめておいてほしい。

図 4.5　SAS による相関係数行列の計算



```
OUTPUT
Command ===>

SAS                                        11:07 Friday, January 31, 1992   1

CORRELATION ANALYSIS

   2 'VAR' Variables:  SHIHON   JUGYOIN


                        Simple Statistics

Variable             N          Mean       Std Dev           Sum

SHIHON              16         47984         26560        767741
JUGYOIN             16          6165          3030         98642

          Simple Statistics

Variable        Minimum        Maximum

SHIHON            13583         106710
JUGYOIN            2230          11873
                                                              ZOOM
OUTPUT
Command ===>

SAS                                        11:07 Friday, January 31, 1992   2

CORRELATION ANALYSIS

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 16

                   SHIHON           JUGYOIN

SHIHON            1.00000          0.54351
                   0.0              0.0296

JUGYOIN           0.54351          1.00000
                   0.0296           0.0
                                                              ZOOM
```

また単純統計は、データの特性をチェックするという点では重要な情報であるが、既に、第 3 章で行ったような、1 変数データとしての吟味が終了している場合には、改めてチェックする必要はない。そのときは、PROC 文で単純統計はいらないということで、NOSIMPLE と次のように下線部を追加指定してやると、単純統計は出力されず、相関係数行列だけが出力される。

PROC CORR NOSIMPLE DATA=SAS データ・セット名;

ただしはじめの頃は、その都度、単純統計をまめにチェックする習慣をつけておいた方が無難である。

相関係数行列は、変数の数が増えてくると、有意水準を

\+ $p<0.1$ (10%)

* $p<0.05$ (5%)

** $p<0.01$ (1%)

*** $p<0.001$ (0.1%)

のように+印や*印で表示した方が見やすい。もちろん、有意水準をどのように設定するのかは分析する側の意図に依拠しているのであるが、ここに挙げたような有意水準の区切り、10%、5%、1%、0.1%は、筆者も含めて多くの人に用いられているものである。特別な主義、主張のない限り、これらを採用した方が無難であろう。また、通常は、5%水準ではじめて有意と言われることが多く、10%水準では「有意」とは扱われないので、「参考程度に」という意味もこめて、*印を使わずに+印で表示する。

したがって、レポートや論文で相関係数行列を示す場合には、表 4.4 のように表示すればよい。相関係数は対角線上はすべて 1 になるし、また対角線に対して対称になっているので、右上半分もしくは左下半分だけを示せば済むことになる。紙幅の都合でそのように省略することもあるが、見やすいということをいえば、紙幅の許す限り、省略せずに表示してもよいだろう。

表 4.4　相関係数行列

	資本金	従業員数
資本金	1.000***	0.544*
従業員数	0.544*	1.000***

ところで、SAS プログラムの「変数名の並び」の中の変数は数値変数に限られる。またプログラム例では省略してしまったが、もちろん VAR 文も WITH 文も省略せずに使用し、相関係数を計算する変数名を明確に指定するのが基本である。WITH 文や VAR 文を省略した場合に、それぞれどういった処理がなされるのか整理しておこう。

(a)WITH 文の省略

この場合には、省略された WITH 文には VAR 文と同じ変数名の並びが指定されていると解釈される。つまり、次の二つのプログラムは同じ機能を果たす。

```
(A)
PROC CORR DATA=SAVE.SHITETSU;
        VAR SHIHON JUGYOIN;
RUN;
```

```
(B)
PROC CORR DATA=SAVE.SHITETSU;
        VAR SHIHON JUGYOIN;
        WITH SHIHON JUGYOIN;
RUN;
```

(b)WITH 文と VAR 文の省略

この場合には、指定された SAS データ・セットの全ての数値変数が処理の対象となる。つまり、次の二つのプログラムは同じ機能を果たすことになる。

```
(C)
PROC CORR DATA=SAVE.SHITETSU;
RUN;
```

```
(D)
PROC CORR DATA=SAVE.SHITETSU;
        VAR _NUMERIC_;
RUN;
```

結局、このプログラム例で使用している永久 SAS データ・セットでは、数値変数は SHIHON と JUGYOIN の二つしかないので、4 つのプログラム例(A)(B)(C)(D)は同じ処理を行うことになる。

スピアマンの順位相関係数、ケンドールの順位相関係数を求めたい場合には、ピアソンの積率相関係数を求めるプログラムの CORR のオプション部分をそれぞれ

PROC CORR SPEARMAN DATA=SAS データ・セット名;

あるいは、

PROC CORR KENDALL DATA=SAS データ・セット名;

と下線部を書き加えるだけで同様にして求められる。

表 4.1 のデータを使うと、資本金と従業員数との間の相関係数は表 4.5 のようになる。この表からもわかるように、SPEARMAN や KENDALL を指定した時には、順位データではなく、金額・人数データを与えた場合でも、その順位についての相関係数が求められる。したがって、金額・人数データでも順位データでも相関係数 r_S、τ_b は変わらない。また、順位データを与えて求めたピアソンの積率相関係数 r が、定義通りスピアマンの順位相関係数 r_S と一致することも確認できる。

表 4.5　表 4.1 のデータについて SAS で求めた各種相関係数と有意確率

	金額・人数データ		順位データ	
	相関係数	有意確率	相関係数	有意確率
ピアソンの積率相関係数 r	0.54351	0.0296	0.46471	0.0697
スピアマンの順位相関係数 r_S	0.46471	0.0697	0.46471	0.0697
ケンドールの順位相関係数 τ_b	0.30000	0.1051	0.30000	0.1051

4. 単回帰と相関係数

(1)二つの問題

相関(correlation)が、x と y を対等に見て、相互の関係を扱っていたのに対して、回帰(regression)分析では、例えば x から y が決定される一方向の関係を扱う(もちろん逆方向に、y から x が決定されるという関係を扱うこともできる)。2 変数データに直線(1 次式)をあてはめるには、次の二つの問題を考える必要がある。

問題 1　$y=a+bx$ をサイズ n の 2 変数データ $(x_1, y_1), ..., (x_n, y_n)$ にうまくあてはまるように a, b を決める。

ここで、$y=a+bx$ という式は変数 x によって変数 y を説明しようとするものなので、変数 x は説明変数または独立変数(independent variable)と呼ばれ、変数 y は被説明変数または従属変数(dependent variable)と呼ばれる。この問題を解く方法には、最小 2 乗法と呼ばれる方法がある。最小 2 乗法によって得られる直線を「y の x への回帰直線(regression line)」または「y の x への回帰式(regression equation)」と呼ぶ。このように、最小 2 乗法によって直線を求めることを回帰分析と呼び、この問題 1 のように説明変数が一つの場合には、単回帰と呼ぶ。また、より一般的に説明変数が p 個の場合には、重回帰(multiple regression analysis)と呼ぶ。単回帰は重回帰の特殊なケース($p=1$)であるが、通常は、重回帰といえば、特に、$p \geqq 2$ の場合を指すことが多い。

問題 2　問題 1 で求められた直線(1 次式)のデータへのあてはまり具合を測る。

これは、正確には重相関係数や決定係数を求める問題になるが、単回帰では、相関係数の絶対値は重相関関係数と一致し、相関係数の 2 乗は決定係数とも一致するので、相関係数によってあてはまり具合を測ることができる。

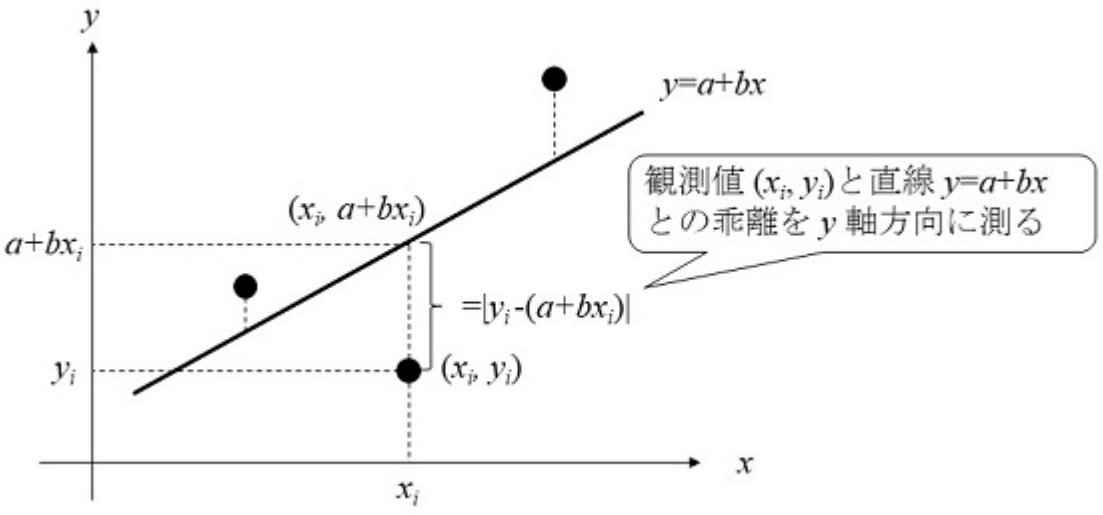
(2)最小 2 乗法

それでは、問題 1: $y=a+bx$ が 2 変数データ $(x_1, y_1), ..., (x_n, y_n)$ にうまくあてはまるように a, b を決めることを考えてみよう。最小 2 乗法(method of least squares)とは、

$$\Sigma_{i=1}^{n}[y_i-(a+bx_i)]^2 \qquad (4.5)$$

を最小にするように a, b を決める方法である。式の中の[　]内は残差 e_i と呼ばれるもので、(4.5)式は残差 2 乗和ということになる。最小 2 乗法はこの残差 2 乗和を最小にするように a, b を決める方法なのである。このことを図示すると、図 4.6 のようになる。

図 4.6　観測値への直線のあてはめ



つまり、残差の絶対値 $|y_i-(a+bx_i)|$ で観測値 (x_i, y_i) と直線 $y=a+bx$ との乖離を y 軸方向に測り、この y 軸方向の乖離の 2 乗和によって、この直線の 2 変数データ $(x_1, y_1), ..., (x_n, y_n)$ へのあてはまり具合を計ってやろうというのである。

残差 2 乗和を最小にするような a, b を求めてみると、つまり最小 2 乗法によると、a, b は次のような簡単な式で求められる。

$$b=\Sigma_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})/\Sigma_{i=1}^{n}(x_i-\bar{x})^2=s_{xy}/s_x^2=r_{xy}s_y/s_x \qquad (4.6)$$

$$a=\bar{y}-b\bar{x}$$

このようにして求められた b を回帰係数(regression coefficient)、a を回帰定数あるいは y 切片(y-intercept)と呼ぶ。回帰係数 b の式から明らかなように、変数 x と y の分散が変わらなければ、回帰係数 b は両変数の相関係数の大きさに比例することになる。相関係数が 0 のときは回帰係数 b も 0 になって、回帰直線は x 軸に平行になる。

<証明>

$$Q(a, b)=\Sigma[y_i-(a+bx_i)]^2$$

とおき、微分の公式

$$[\{f(x)\}^n]'=n\{f(x)\}^{n-1}f'(x)$$

を使って、$Q(a, b)$ の a, b に関する偏導関数を求め、0 とおくと、

$$\partial Q(a, b)/\partial a=\Sigma 2(y_i-a-bx_i)(-1)=0$$

$$\partial Q(a, b)/\partial b=\Sigma 2(y_i-a-bx_i)(-x_i)=0$$

これを整理すると正規方程式(normal equation)と呼ばれる次の連立方程式が得られる。

$$an+b\Sigma x_i=\Sigma y_i$$

$$a\Sigma x_i+b\Sigma x_i^2=\Sigma x_i y_i$$

この連立方程式を解くには、第 1 式の両辺を n で割って得られる $a=\bar{y}-b\bar{x}$ を第 2 式に代入することによって、

$$b=(\Sigma x_i y_i-n\bar{x}\bar{y})/(\Sigma x_i^2-n\bar{x}^2)=\{\Sigma(x_i-\bar{x})(y_i-\bar{y})\}/\Sigma(x_i-\bar{x})^2 \quad \square$$

(3)2 本の回帰直線と相関

同様にして、x の y への回帰直線 $x=a'+b'y$ の回帰係数、回帰定数も次のように求められる。

$$b'=s_{xy}/s_y^2=r_{xy}s_x/s_y$$

$$a'=\bar{x}-b'\bar{y}$$

このようにして、2 変数データ $(x_1, y_1), ..., (x_n, y_n)$ については、どちらを説明変数にし、どちらを被説明変数にするかで、2 本の回帰直線を引くことができる。この 2 本の回帰直線については、次の関係が成り立つ。

性質 1　2 本の回帰直線は点 (x, y) で交わる。

<証明>

y の x への回帰直線 $y=a+bx$ に $a=\bar{y}-b\bar{x}$ を代入すると $y-\bar{y}=b(x-\bar{x})$ と書けることから、この回帰直線は $(\bar{x}, \bar{y})$ を通る。同様に、x の y への回帰直線は $x-\bar{x}=b'(y-\bar{y})$ と書け ることから、この回帰直線も $(\bar{x}, \bar{y})$ を通る。□

性質 2　変数 x と変数 y の相関係数 r_{xy} の絶対値は b と b' の絶対値の幾何平均となる。すなわち、$b'b=r_{xy}^2$

<証明>

$$b'b=(r_{xy}s_x/s_y)(r_{xy}s_y/s_x)=r_{xy}^2 \quad \square$$

この性質 2 から次の性質がすぐに導かれる。

性質 3　$r_{xy}=\pm1$ のとき $1/b'=b$ で 2 本の回帰直線は重なる。

これは性質 2 によらなくても明らかであろう。なぜなら、$r_{xy}=\pm1$ のとき、すべての点は同一直線上にある(相関係数の性質 4)わけだから、当然、その直線が回帰直線と重ならなければおかしいからである。

さらに、b と b' の式から

性質 4　相関係数 r_{xy} と回帰係数 b, b' の符号は一致する。

<証明>

$b=r_{xy}s_y/s_x$, $b'=r_{xy}s_x/s_y$ から明らか。□

例えば、無相関で $r_{xy}=0$ となっているときは、$b=b'=0$ で 2 本の回帰直線はそれぞれ x 軸、y 軸に平行な直線となり、直交することになる。

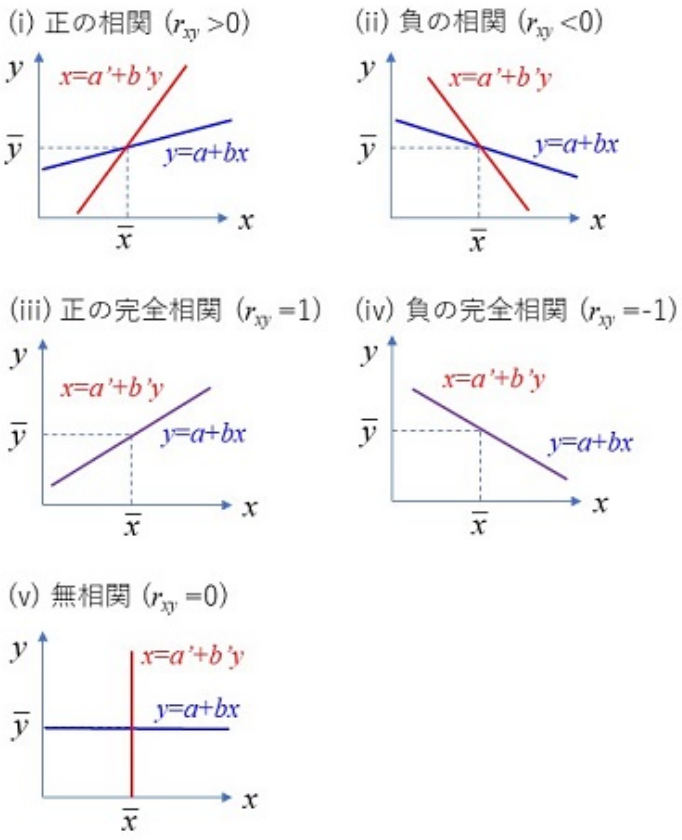
性質 5　$b'\neq0$ のとき $|1/b'|\geqq|b|$

<証明>

$0\leqq b'b=r_{xy}^2\leqq1$

である。したがって、$b>0$ のとき $b'>0$ であることから $1/b'\geqq b>0$。$b<0$ のとき $b'<0$ であることから $1/b'\leqq b<0$。□

図 4.7　2 本の回帰直線と相関



以上の性質をまとめると、相関の大きさと 2 本の回帰直線の関係は図 4.7 のように図示される。図 4.7 からもわかるように、y の x への回帰直線 $y=a+bx$ の方が常に傾きがゆるやかなものになる(性質 5)。しかも、相関が弱いほど、2 本の回帰直線の傾きの開きは大きなものになる。回帰直線の傾きは回帰係数 b で決まるので、回帰係数 b の絶対値が小さくて、回帰直線 $y=a+bx$ の傾きが小さいときには、実際にはこのように相関が弱く、ただ単に回帰直線のあてはまりの程度が低く、回帰直線の説明力が弱いだけという可能性のあることに注意する必要がある。実は、回帰直線のあてはまりの程度、説明力を見るには、相関係数の大きさを見ればよいのである。

回帰直線の説明力を見るためには、被説明変数の全平方和 $\Sigma(y_i-\bar{y})^2$ のうち、回帰直線 $y=a+bx$ で説明のつく平方和部分の割合がどの程度になっているのかを調べればよいわけだが、それには、回帰直線によって説明できなかった残差 2 乗和の全平方和に占める割合を考え、それを 1 から引いたものに着目してやればよい。これを決定係数(coefficient of determination)と呼び、R^2 で表す。すなわち、

$$R^2=1-[\Sigma\{y_i-(a+bx_i)\}^2]/\Sigma(y_i-\bar{y})^2$$

単回帰の場合には、説明変数と被説明変数の相関係数の 2 乗がこの決定係数に等しくなることがわかっている。すなわち、

$$r_{xy}{}^2=R^2$$

また回帰直線による予測値 $a+bx_i$ と実際の観測値 y_i との間の相関係数 R は重相関係数(multiple correlation coefficient)と呼ばれ、回帰直線のあてはまりの程度を表していると考えられる。重相関係数は $0\leqq R\leqq 1$ であり、また重相関係数の 2 乗は決定係数に一致することがわかっている。すなわち、

$$(R)^2=R^2$$

このことから、単回帰の場合には、相関係数の絶対値と重相関係数が等しいことになる。したがって、相関係数が大きいほど、回帰直線のあてはまり程度も良く、説明力も高いということになるのである。

決定係数は相関係数の 2 乗になっているので、たとえば、相関係数が $r=0.7$ のとき、決定係数は $R^2=0.49$ となる。つまり、被説明変数の全平方和の 0.49 つまりほぼ半分の 49%が回帰直線によって説明されていることになり、相関係数の数字に比べて説明力が意外と低いことがわかる。相関係数が $r=0.9$ にもなれば、決定係数は $R^2=0.81$ になるが、逆に相関係数が小さく $r=0.3$ であれば、決定係数は $R^2=0.09$、つまり被説明変数の全平方和のわずか 10%弱しか説明できていないことになる。

このように、回帰直線の説明力を示す決定係数が、相関係数の 2 乗となっていることからもわかるように、回帰分析は相関分析と深い関係があり、相関係数を使うときと同様の注意が必要となってくる。すなわち、

1. 散布図の形が、平均値線に対して対称に近いものであれば、相関係数はほとんど 0 になり、したがって、回帰直線を求めても、ほとんど説明力がないはずである。ところが、こうした場合でも、平均値線を境にしてデータを二分して回帰分析を行うと、説明力の高い回帰直線が得られることがあるので注意がいる。(→演習問題 4.3)
2. 平均も相関係数も少数の外れ値、異常値に大きく影響を受けたように、回帰直線もまた少数の外れ値、異常値に大きく影響を受けるので注意がいる。(→演習問題 4.4)
3. したがって、相関係数が散布図など相関関係を視覚に訴える方法を補完するものであると考えた方がよいというのと同様に、回帰分析の際も、必ず散布図を描いてみる必要がある。

(4)回帰係数・回帰定数の検定

無作為標本のデータ $(X_1, Y_1), ..., (X_n, Y_n)$ にもとづいて回帰分析を行ったとき、求められた回帰定数 a、回帰係数 b は確率変数となる。母集団での回帰直線が $y=\alpha+\beta x$ で、そのときの残差が x の値によらない等分散の正規分布で近似される場合(x と y の同時分布が2次元正規分布で近似される場合にこうなることがわかっている)、回帰係数についての仮説

$H_0: \beta=\beta_0$

を検定するためには、

$t=(b-\beta_0)/\text{s.e.}(b)$ (4.7)

が自由度 $n-2$ の t 分布にしたがうので、そのことを用いる。同様に、回帰定数についての仮説

$H_0: \alpha=\alpha_0$

を検定するためには

$t=(a-\alpha_0)/\text{s.e.}(a)$

が自由度 n-2 の t 分布にしたがうので、そのことを用いればよい。ここで、s.e.(b)、s.e.(a)は標準誤差(standard error)と呼ばれ、次のようになる。

$\text{s.e.}(b)=s/\{\Sigma(X_i-\bar{X})^2\}^{1/2}$

$\text{s.e.}(a)=s/\{1/n+\bar{X}^2/\Sigma(X_i-\bar{X})^2\}^{1/2}$

ただし、

$s^2=\Sigma[Y_i-(a+bX_i)]^2/(n-2)=(1-r^2)\Sigma(Y_i-\bar{Y})^2/(n-2)$

ここで、仮説の中の β_0、α_0 は分析を行う人が決める定数であるが、特に回帰係数 β_0 について一般によく用いられるのは $\beta_0=0$ である。つまり、「説明変数 x は y に影響を与えない」という仮説である。この仮説のときの t の値を t 値という。すでに(4.6)式で回帰係数 b が

$b=r_{xy}s_y/s_x$

となることを述べたが、この式からも想像がつくように、この式と s.e.(b)を(4.7)式に代入すると、単回帰の回帰係数についての仮説 $H_0: \beta=0$ の検定と相関係数についての仮説 $H_0: \rho=0$ の検定は同等ということがわかる。

(5)SAS による回帰分析

回帰分析を手計算で行うことは、データ数が増えてくると、一般に大変な作業である。そこで、いよいよ SAS のような統計パッケージが有用となってくる。回帰分析を行う SAS プログラムもいたって簡単である。REG プロシジャを使えばよい。

PC 版 SAS

```
LIBNAME libname '¥パス名';
PROC REG DATA=libname.永久 SAS データ・セット名本体;
MODEL  被説明変数=説明変数;
RUN;
```

例)

```
LIBNAME SAVE '¥MYDIR';
PROC REG DATA=SAVE.SHITETSU;
MODEL SHIHON=JUGYOIN;
RUN;
```

CMS 版 SAS

```
PROC REG DATA=永久 SAS データ・セット名;
MODEL  被説明変数=説明変数;
RUN;
```

例)

```
PROC REG DATA=SAVE.SHITETSU;
MODEL SHIHON=JUGYOIN;
RUN;
```

PC 版 SAS のプログラムの 1 行目を削除すると形式的には CMS 版 SAS のプログラムとなる。このプログラム例を実行すると、PC 版 SAS では 10 秒ほどで、図 4.8 のように、2 画面が表示される。第 1 画面は分散分析(ANALYSIS OF VARIANCE)表を表示し、第 2 画面は回帰係数・回帰定数の推定値(PARAMETER ESTIMATES)及びその検定結果が出力される。さらに必要であれば、PROC 文のオプションとして SIMPLE を追加指定すれば、相関係数のときと同様の単純統計を出力させることもできる。

図 4.8　SAS による回帰分析(SHIHON の JUGYOIN への回帰直線)

```
OUTPUT
Command ===>

                                    SAS    13:11 Friday, January 31, 1992   1

Model: MODEL1
Dependent Variable: SHIHON

                          Analysis of Variance

                            Sum of          Mean
      Source          DF      Squares        Square      F Value       Prob>F

      Model            1 3125851981.5 3125851981.5        5.869       0.0296
      Error           14 7455827395.0 532559099.64
      C Total         15  10581679376

          Root MSE   23077.24203     R-square       0.2954
          Dep Mean   47983.81250     Adj R-sq       0.2451
          C.V.         48.09381
```

ZOOM

```
OUTPUT
Command ===>

                                    SAS    13:11 Friday, January 31, 1992   2

                        Parameter Estimates

                     Parameter      Standard    T for H0:
    Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

    INTERCEP   1         18612  13426.372740         1.386        0.1874
    JUGYOIN    1      4.764214    1.96648659         2.423        0.0296
```

ZOOM

分散分析については、ここでは説明しないが、決定係数 R^2 (R-square)が 0.2954 と既に第 3 節図 4.5 で求めていた相関係数 0.54351 の 2 乗に等しいこと、また回帰分析のモデルの有意確率(Prob>F) 0.0296 が、相関係数の有意確率と等しくなっていることを確認してほしい。この有意確率は、説明変数 JUGYOIN の回帰係数の有意確率(Prob > |T|)とも等しい。既

に述べたように、単回帰の回帰係数についての仮説 H_0: β＝0 の検定と相関係数についての仮説 H_0: ρ＝0 の検定は同等なのである。

回帰係数、回帰定数(INTERCEP)の推定、そして、*t* 値(T)から、次のような回帰分析の結果が得られたことになる。

SHIHON＝18612＋4.7642 JUGYOIN

(1.386)　(2.423*)

ここで、(　)内は *t* 値を示している。*t* 値に付けられた*印は、5%水準で有意であることを示している。また、もとになった表 4.1 の資本金データの有効桁数などを考えると、回帰係数、回帰定数は、SAS の出力結果を全部書かずに、5 桁も表示すればいいだろう。

ちなみに、プログラム例の MODEL 文で、説明変数と被説明変数を入れ替えて

MODEL JUGYOIN=SHIHON;

として実行させると、図 4.9 のような結果が出力される。

図 4.9　SAS による回帰分析(JUGYOIN の SHIHON への回帰直線)

```
OUTPUT
Command ===>

                                SAS    13:11 Friday, January 31, 1992   3

Model: MODEL1
Dependent Variable: JUGYOIN

                          Analysis of Variance

                            Sum of          Mean
      Source          DF    Squares        Square      F Value       Prob>F

      Model            1 40681751.404 40681751.404        5.869       0.0296
      Error           14 97034702.346 6931050.1675
      C Total         15 137716453.75

          Root MSE    2632.68877     R-square       0.2954
          Dep Mean    6165.12500     Adj R-sq       0.2451
          C.V.          42.70293
                                                                    ZOOM
OUTPUT
Command ===>

                                SAS    13:11 Friday, January 31, 1992   4

                          Parameter Estimates

                     Parameter      Standard    T for H0:
    Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

    INTERCEP   1   3189.917412  1393.3068057         2.289        0.0381
    SHIHON     1      0.062004    0.02559306         2.423        0.0296
                                                                    ZOOM
```

説明変数と被説明変数を入れ替えても、分散分析の *F* 値(F Value)、有意確率(Prob>F)、決定係数(R-square)は先ほどの値と全く同じになることを確かめてほしい。また説明変数が入れ替わったにも関わらず、説明変数の *t* 値(T)、有意確率(Prob > |T|)が変わらず、同じであることも確認してほしい。もっとも、当然、回帰直線は次のように別のものになっている。

JUGYOIN＝3189.9＋0.062004 SHIHON
(2.289*)　(2.423*)

先ほど求めた回帰係数 4.7642 とここで求めた回帰係数 0.062004 の幾何平均は、

$(4.7642\times0.062004)^{1/2}=0.54351$

と相関係数に等しくなることも確認してほしい。

この章の最初に描いておいた図 4.1 の散布図に、2 本の回帰直線を書き込んでみると、図 4.10 のようになる。

図 4.10　2 本の回帰直線

演習問題

4.1 相関・回帰分析　表 4.1 の私鉄大手 16 社のデータを、資本金(SHIHON)、従業員数(JUGYOIN)そのままではなく、その対数をとった上で、相関、回帰分析を行なってみよ。SAS のプログラムでは、変数 X の対数は、底が e の自然対数、底が 10 の常用対数、底が 2 の対数については、それぞれ

```
Y=LOG(X);
Y=LOG10(X);
Y=LOG2(X);
```

といったように(　)つきの関数で求めることができる。次の PC 版 SAS のプログラム例を参考にするとよい。この 1 行目を削除すると CMS 版 SAS のプログラム例になる。

```
LIBNAME SAVE '¥MYDIR';
DATA SAVE.SHITETSU;
 SET SAVE.SHITETSU;
SLOG=LOG(SHIHON);
JLOG=LOG(JUGYOIN);
PROC PLOT;
     PLOT SLOG*JLOG=KCODE;
RUN;
PROC REG;
MODEL SLOG=JLOG;
RUN;
```

4.2 相関・回帰分析　表 4.6 のデータは、第 6 章の D 電気㈱の 1991 年 3 月末現在の 100%子会社全 14 社の資本金、従業員数である。相関・回帰分析を行ってみよ。

表 4.6　D 電気㈱の 100%子会社の資本金、従業員数(1991 年 3 月末現在)

会社	資本金(百万円)	従業員数(人)
A	2,000	779
B	2,200	218
C	300	1,108
D	300	1,214
E	40	579
F	80	404
G	50	830
H	50	1,199
I	450	32
J	50	590
K	200	1,284
L	50	156
M	200	84
N	450	16

4.3 相関・回帰分析　東日本旅客鉄道(株) (いわゆる「JR 東日本」)は、1990 年 3 月末当時、資本金 200,000 百万円、従業員数約 80,000 人といわれた。表 4.1 の私鉄大手 16 社に JR 東日本も加えて相関・回帰分析を行ってみよ。

第 5 章　クロス表

章目次

1. はじめに
2. 2×2 クロス表
3. $s \times t$ クロス表と V 係数
4. 2×2 クロス表への相関係数の適用
5. SAS によるクロス表の作成と相関係数
6. エラボレイション

1. はじめに

2 変数データの表は、一般には、質的データの場合にはクロス表(cross table)あるいは関連表・分割表(contingency table)、量的データの場合には相関表(correlation table)と呼ばれる。質的データの表と量的データの表の両者を総称して、二重分類表(double classification table)と呼ぶこともある。いずれにせよ、これらの表は、2 変数を同時にとりあげた同時度数分布の表である。

実際に表を作成するときには、まず二つの変数の値を質的データではカテゴリー、量的データでは階級に分割して、さらに両者が交差(クロス)する升目(セル)に度数を書き込むのである。その一般的な形式は表 5.1 のようになる。

表 5.1　2 変数データの表の一般形式

(表側)　　(表頭)

x	y					
	y_1	・・・・・・	y_j	・・・・・・	y_k	計
x_1	f_{11}	・・・・・・	f_{1j}	・・・・・・	f_{1k}	$f_{1\cdot}$
：	：		：		：	：
x_i	f_{i1}	・・・・・・	f_{ij}	・・・・・・	f_{ik}	$f_{i\cdot}$
：	：		：		：	：
x_I	f_{I1}	・・・・・・	f_{Ij}	・・・・・・	f_{Ik}	$f_{I\cdot}$
計	$f_{\cdot 1}$	・・・・・・	$f_{\cdot j}$	・・・・・・	$f_{\cdot k}$	n

この形式で作成された表を、質的データ×質的データの場合はクロス表と呼び、量的データ×量的データの場合には相関表と呼ぶのである。したがって、質的データのクロス表のときは、x_i, y_i はカテゴリーを表すことになるし、量的データの相関表のときには、x_i, y_i は階級を表すことになるのである。一般的には、表側の変数(ここでは x)に、より基本的なものをもってくることが多い。

ここで、表 5.1 の中の各セルの f_{ij} が同時度数をあらわしている。他方、「計」として表の縁すなわち周辺に表示されているものは、周辺度数と呼ばれる。つまり、

$f_{i\cdot}=\Sigma_j f_{ij}$　……x の周辺度数
$f_{\cdot j}=\Sigma_i f_{ij}$　……y の周辺度数
ということになる。

量的データについて、前章で扱った散布図は、一般的には非常に有力な方法であるのだが、社会科学分野、特に調査データの場合には、あまり深く考えもせずに、単純に散布図を描いても、整理の実用性、効果という点ではあまり期待できない。なぜなら、本書で扱っている「組織活性化のための従業員意識調査」のようなケースでは、変数が離散変数であることが多いし(もっとも、既に第 3 章第 1 節で述べたように、観測・測定データは厳密にはすべて離散的である)、しかも、とりうる値の数に比べてデータのサイズが大きいので(つまり、調査対象者数が多いので)、散布図を描いてみても、結果的に散布図では同じ位置にいくつもの点が重なり合うことになり、結局、点では分布がわからず、度数で示さざるをえなくなってしまう。特に前章で扱った PLOT プロシジャを使って SAS に描かせると、画面の大きさの制約もあって、描かせた散布図が、実質的には相関表と同じことにならざるを得ないような場合が多いのである。したがって、「とりあえず」散布図を描かせてみることの効用を否定しはしないが、正式には相関表をきちんと作成すべきであろう。

第 3 章で扱ったように、量的データとはいっても、各変数は適切に階級が設定されるべきであり、仮にそれが既に行なわれているならば、安易に散布図を描いてみる(SAS を使えば手間はかからないが)よりも、相関表を作成した方がはるかに実用的で、きれいに整理ができる。実は、第 3 章で扱ったような 1 変数でのデータの吟味は、こうした後の複雑な処理を間違いや誤解を回避して、すんなりと進めるために、極めて重要である。2 変数の相関表を作成する段になって、このようなことに悩むとしたら、それは、こうした階級設定の前工程がしっかりと行われていないことの証拠なのである。

この章では、2 変数データの表の基本である質的データのクロス表について説明しよう。量的データの相関表や、質的データと量的データを組み合わせた表も、このクロス表の応用となる。

2. 2×2 クロス表

2 変数の質的データのクロス表で、一番基本的でかつ重要なものが、各変数が各々二つのカテゴリーをもっている場合のクロス表で、2×2 クロス表と呼ばれる。四分表(four-fold table)と呼ばれることもある。例として「組織活性化のための従業員意識調査」で調べた質問を使って、クロス表を作ってみよう。いま男女の性別と二つの Yes-No 形式の質問

V.2.　終身この会社で仕事をしていきたいと思う。　1. Yes　2. No
V.12.　人生にゆとりがあり、楽しい。　1. Yes　2. No

との間でそれぞれクロス表を作ってみると、表 5.2 の(A)(B)が得られる。これは筆者がよく使う形式のクロス表である。

表 5.2　クロス表の例

(A)

	質問 V2. 終身この会社で仕事をしていきたいと思う。		
	1. Yes	**2. No**	**計**
1. 男	508　(66.67)	254　(33.33)	762　(100)
2. 女	38　(35.19)	70　(64.81)	108　(100)
計	546	324	870

Cramer's V＝0.215　　χ^2＝40.112***

(B)

	質問 V12. 人生にゆとりがあり、楽しい。		
	1. Yes	**2. No**	**計**
1. 男	282　(37.01)	480　(62.99)	762　(100)
2. 女	68　(62.39)	41　(37.61)	109　(100)
計	350	521	871

Cramer's V＝-0.171　　χ^2＝25.550***

表 5.2 のクロス表の中の(　)内には行方向の百分率を示している。クロス表の下についている Cramer's V とか χ^2については、これからこの章の中で説明していくことになるが、いずれも相関の程度を表しているものである。***はその有意確率が 0.1%よりも小さいことを示している。Cramer's V とか χ^2についてまだ知らなくても、表 5.2(A)からは男子従業員の方が、終身この会社で仕事をしていきたいと思う傾向があることわかるし、表 5.2(B)からは女子従業員の方が、人生にゆとりがあり、楽しいと思っているという興味深い傾向が読んで取れるだろう。（→演習問題 5.1）

表 5.3　解釈のはっきりしているクロス表

(A)無相関の場合

x	***y***		
	1. Yes	**2. No**	**計**
1. Yes	25	25	50
2. No	25	25	50
計	50	50	100

(B)正の完全相関の場合

x	***y***		
	1. Yes	**2. No**	**計**
1. Yes	50	0	50
2. No	0	50	50
計	50	50	100

一般的には、あらゆるクロス表は最終的には 2×2 クロス表に帰着させるといわれるほどで、2×2 クロス表をきちんと理解し、きちんと解釈できることが、クロス表利用の基本となる。もっとも 2×2 クロス表をきちんと理解し、きちんと解釈することは決して難しくない。もし表 5.3(A)のようなクロス表が得られたら、変数 x と y との間に何の関連もないということは、はっきりしている。もし表 5.3(B)のようなクロス表が得られたら、x に Yes ならば y も Yes、x に No ならば y も No と答えるという完全な関連があることが一目瞭然である。それでは、次の例題はどうなるだろうか。

例題) 表 5.4 のクロス表では周辺度数だけが与えられている。もし、x と y との間に相関が全くなく無相関であったら、度数はどうなっているはずだろうか。表 5.4(A)のクロス表に度数を書き入れよ。また x と y との間に正の完全相関が見られる例も表 5.4(B)のクロス表に度数を書き入れよ。

表 5.4　例題の解答用クロス表

(A)無相関の場合

x	y		
	1. Yes	**2. No**	**計**
1. Yes			60
2. No			40
計	60	40	100

(B)正の完全相関の場合

x	y		
	1. Yes	**2. No**	**計**
1. Yes			60
2. No			40
計	60	40	100

ここでは、この例題のように周辺度数が具体的に与えられている例ではなく、一般の場合について解説しよう。例題の解答は、以下の解説を読めばすぐにわかるはずである。それぞれ自分で考えられたい。

(1)無相関(独立)の場合

クロス表で、x と y との間に相関が全くない場合、「無相関」と呼んだり、あるいは「独立」と呼んだりする。実は確率論的な意味では、無相関よりも独立の方が強い性質なわけであるが、独立ならば無相関となることがわかっている。

それでは、独立とはどのような状態のクロス表を意味するのであろうか。もし x と y の両者の間に関係がないとしたら、表 5.5 で示されているような関係があるはずである。

表 5.5　独立の場合のクロス表

x	y		
	y_1	y_0	計
x_1	f_{11}	f_{10}	$f_{1\cdot}$
x_0	f_{01}	f_{00}	$f_{0\cdot}$
計	$f_{\cdot 1}$	$f_{\cdot 0}$	n

$f_{11}/f_{1\cdot} = f_{01}/f_{0\cdot} = f_{\cdot 1}/n$

$f_{11}/f_{\cdot 1} = f_{10}/f_{\cdot 0} = f_{1\cdot}/n$

たとえば、いまカテゴリー y_1 の占める比率に着目してみよう。x と y の両者の間に関係がなければ、カテゴリー x_1 の行の度数 $f_{1\cdot}$ のうちカテゴリー y_1の占める比率も、カテゴリー x_0 の度数 $f_{0\cdot}$ のうちカテゴリー y_1の占める比率も等しく、$f_{\cdot 1}/n$ のはずである。ならば

$$f_{11} = f_{1\cdot} \cdot f_{\cdot 1}/n \qquad f_{01} = f_{0\cdot} \cdot f_{\cdot 1}/n \qquad (5.1)$$

となっているはずである。同様に、カテゴリー y_0の占める比率に着目すれば、

$f_{10}=f_{1\cdot}\cdot f_{\cdot 0}/n$　　$f_{00}=f_{0\cdot}\cdot f_{\cdot 0}/n$　　(5.2)

クロス表で、この(5.1)式または(5.2)式が成立するとき、x と y は独立(independent)であると定義する。この独立の定義は、確率の独立の概念と一致している。なぜなら、いま確率の意味で独立とするならば、たとえば(5.1)式の第 1 式に対応して、

$Pr(X=x_1, Y=y_1) = Pr(X=x_1 | Y=y_1)\, Pr(Y=y_1) = Pr(X=x_1)\, Pr(Y=y_1)$

となるからである。ここで、左辺は f_{11}/n、右辺は $(f_{1\cdot}/n)\cdot(f_{\cdot 1}/n)$ に対応しているので、結局(5.1)式の第 1 式が導出されることになる。

(2)完全相関(完全関連)の場合

質的データの場合には、量的データの相関(correlation)の概念に対応するものとして、関連(association)の概念がある。もっとも、両者を総称して「相関」ともいうので、クロス表の場合には関連と呼ばなければならないというものではなく、相関と呼んでもかまわない。本書では一貫して相関と呼ぶことにしよう。

実は、クロス表の場合でも、通常の相関と同様の考え方をする(第 4 節で後述するように、同じ相関係数を用いることもある)。ピアソンの積率相関係数の意味の説明の際に用いた第 4 章第 3 節の図 4.4 と対応させるとよくわかるが、

1. $f_{11}>f_{1\cdot}f_{\cdot 1}/n$ または $f_{00}>f_{0\cdot}f_{\cdot 0}/n$ のときは正の相関
2. $f_{11}<f_{1\cdot}f_{\cdot 1}/n$ または $f_{00}<f_{0\cdot}f_{\cdot 0}/n$ のときは負の相関

となる。その両極端である相関の最も強い正の完全相関、負の完全相関の場合には、クロス表は表 5.6 のようになる。

表 5.6　完全相関の場合のクロス表

(A)正の完全相関

x	y		
	y_1	y_0	計
x_1	f_{11}	0	$f_{1\cdot}$
x_0	0	f_{00}	$f_{0\cdot}$
計	$f_{\cdot 1}$	$f_{\cdot 0}$	n

(B)負の完全相関

x	y		
	y_1	y_0	計
x_1	0	f_{10}	$f_{1\cdot}$
x_0	f_{01}	0	$f_{0\cdot}$
計	$f_{\cdot 1}$	$f_{\cdot 0}$	n

(3)χ^2 検定と φ 係数

全く相関のない独立な場合とその逆に完全相関の場合のクロス表はわかったが、通常はそのような極端なケースは希であろう。その両者の間の中間的なクロス表の場合には、相関の程度、あるいは逆に独立ではない程度を表すにはどうしたらよいだろうか。それには、独立であるときの状態から、どの程度乖離しているのかを表現すればよい。

仮に、x と y とが独立とすると、その際に期待される度数つまり期待度数は、(5.1)式と(5.2)式から

$f_{11}=f_{1\cdot}f_{\cdot 1}/n$　　$f_{10}=f_{1\cdot}f_{\cdot 0}/n$

$f_{01}=f_{0\cdot}f_{\cdot 1}/n$　　$f_{00}=f_{0\cdot}f_{\cdot 0}/n$

のはずである。この独立と仮定した場合の期待度数と実際の観測度数との差が大きければ、x と y とが独立ではない、つまり相関している度合が高くなる。そこで、O を観測度数(observed frequency)、E を期待度数(expected frequency)とするとき、次の量を考えるので

ある。

$\chi^2=\Sigma_{セル}\ (O-E)^2/E=\Sigma_i\Sigma_j\ (f_{ij}-f_{i\cdot}f_{\cdot j}/n)^2/(f_{i\cdot}f_{\cdot j}/n)$

これは Pearson の χ^2検定量と呼ばれる。ここで χ^2は「カイ 2 乗」(chi-square)と読む。母集団で 2 変数 x、y が独立であるときには、この母集団からの無作為抽出標本におけるこの χ^2 の分布が近似的に自由度 1 の χ^2分布にしたがうことが、Pearson によって示されている。

したがって、仮説

H_0: 変数 x と y とは独立である。

数式で書けば

H_0: すべての i, j に対して $f_{ij}=f_{i\cdot}f_{\cdot j}/n$

を検定するためには、χ^2を用いればよいことになる。これを独立性の χ^2検定(χ^2 test for independence)という。

しかし、χ^2をそのまま相関係数として使う場合には問題がある。それは、χ^2のとりうる値の範囲が、通常用いられている各種の相関係数のとりうる値の範囲 -1〜1 と一致しないのである。例えばクロス表の各セルの度数にある定数 k をかけたときには、χ^2 の値 $\chi_{(k)}^2$ は同じく k 倍になってしまうのである。実際、

$$\chi_{(k)}^2=\Sigma_i\Sigma_j\ (kf_{ij}-kf_{i\cdot}kf_{\cdot j}/kn)^2/(kf_{i\cdot}kf_{\cdot j}/kn)=k\chi^2 \qquad (5.3)$$

となる。したがって、この Pearson の χ^2検定量は標本の大きさ n に比例していくらでも大きくなり、0〜∞の値をとることになってしまう。そこで、χ^2を n で割ってやることで、

$\varphi^2=\chi^2/n$

を定義してやればよい。これは、相対度数によりクロス表の χ^2を求めたのと同じことになる。つまり、 $\Sigma\Sigma f_{ij}=n$、$\Sigma\Sigma f_{i\cdot}f_{\cdot j}=n^2$ に注意すれば、

$$\begin{aligned}\varphi^2&=\chi_{(1/n)}^2=\chi^2/n\\&=\Sigma\Sigma\{f_{ij}^2-2f_{ij}f_{i\cdot}f_{\cdot j}/n+(f_{i\cdot}f_{\cdot j}/n)^2\}/(f_{i\cdot}f_{\cdot j})\\&=\Sigma\Sigma f_{ij}^2/(f_{i\cdot}f_{\cdot j})-\Sigma\Sigma 2f_{ij}/n+\Sigma\Sigma f_{i\cdot}f_{\cdot j}/n^2\\&=\Sigma\Sigma f_{ij}^2/(f_{i\cdot}f_{\cdot j})-1 \qquad (5.4)\end{aligned}$$

導出の過程からもわかるように、この(5.4)式は、2×2 以外のクロス表にも使える。さらに、2×2 クロス表に関しては、この(5.4)式の分母が $f_{1\cdot}f_{0\cdot}f_{\cdot 1}f_{\cdot 0}$ となるように通分した上で、その分子を f_{11}、f_{10}、f_{01}、f_{00} だけからなる式にして展開すると、次の形の式も得られる。

$$\varphi^2=(f_{11}f_{00}-f_{10}f_{01})^2/(f_{1\cdot}f_{0\cdot}f_{\cdot 1}f_{\cdot 0}) \qquad (5.5)$$

どちらの式を用いて計算するにせよ、この φ^2の平方根をとって、

φ 係数: $\varphi=(\chi^2/n)^{1/2}$

を定義する。

φ 係数はこの定義からも明らかなように、

$\varphi\geqq 0$

である。実際、(5.5)式の分子は $f_{11}f_{00}=f_{10}f_{01}$ のとき、すなわち、x と y とが独立のとき、0 をとるので、このとき $\varphi=0$ となる。

他方、周辺度数 $f_{1\cdot}$、$f_{0\cdot}$、$f_{\cdot 1}$、$f_{\cdot 0}$ を固定したとき、(5.5)式の分子は、

1. $f_{10}=f_{01}=0$ のとき(正の完全相関の場合)、最大値 $f_{11}f_{00}$ をとる。このとき $\varphi=1$
2. $f_{11}=f_{00}=0$ のとき(負の完全相関の場合)、最小値 $-f_{10}f_{01}$ をとる。このとき $\varphi=1$

となって、正負の完全相関のとき値 1 をとる。実は後の第 4 節で述べるが、このφ係数はピアソンの積率相関係数 r の絶対値と一致する。

ところで、(5.3)式から、各セルの総数 n が大きくなってくれば、χ^2検定は有意になりやすくなってくることがわかる。経験的には、n が 400〜500 を超えるような大きさになると、多少なりとも相関のありそうな(つまり、$\varphi=0$ ではない)クロス表はほとんど有意にな

ってしまう。こうした「クロス表の χ^2 検定は n が大きいとき有意になりやすい」という性質は、分析の際に十分知っておく必要がある。しかし、このことをもってして、n の大きなクロス表の χ^2 検定は信用できないという人がいるが、これは正しくない。こうした見解は、検定、推定の考え方を理解しないままに、「χ^2 検定で有意ならば相関が高い」と短絡的に考えている人が陥りやすい誤解である。第 1 章で既に述べたように、標本の大きさが十分に大きければ、推定の精度が十分に向上しているのである。つまり、χ^2 検定の結果が教えてくれているように、例えば、$\varphi=0.15$ なる標本での値を「母集団では $\varphi=0$ だったのではないか」などと疑ってみる必要も、検定してみる必要も、もはやないのである。n の大きなクロス表の χ^2 検定は信用できないのではなく、そもそも必要ないのだ。n の大きなクロス表だからこそ、φ の値はかなりの精度で信用ができるし、そのために、検定すること自体の意味も、必要もなくなっているのである。

3. $s\times t$ クロス表と V 係数

表 5.7 に示す $s\times t$ クロス表（$s\leqq t$）のような場合には、相関はどのように考えればよいだろうか。

表 5.7　$s\times t$ クロス表

x	y				
	y_1	y_2	・・・・・・	y_t	計
x_1	f_{11}	f_{12}	・・・・・・	f_{1t}	$f_{1\cdot}$
x_2	f_{21}	f_{22}	・・・・・・	f_{2t}	$f_{2\cdot}$
：	：	：		：	：
x_s	f_{s1}	f_{s2}	・・・・・・	f_{st}	$f_{s\cdot}$
計	$f_{\cdot 1}$	$f_{\cdot 2}$	・・・・・・	$f_{\cdot t}$	n

変数が名義尺度(第 1 章第 3 節参照のこと)によるものであれば、カテゴリー間に何か固有の順序が存在しているわけではない。それでもカテゴリーが二つの場合には、φ 係数のように相関係数の符号を考えないのであれば問題はない。ところが、カテゴリーが 3 つ以上になった場合には、カテゴリーの並べ方は固定していないことになる。カテゴリーが 3 つのときでもその並べ方は、3 個のものから 3 個をとる順列の個数である ${}_3P_3=3\cdot 2\cdot 1=6$ 通りもあることになる(第 1 章第 5 節参照のこと)。

このように、一般の $s\times t$ クロス表では、二つの変数のカテゴリーは、それぞれ必ずしも固定した順序で並んでいるわけではないので、単純に相関を考えるのは難しい。したがって、次のどちらかの方法を考えることになる。

1. 全体、あるいは部分を 2×2 クロス表の形に直せば、既に述べたような相関係数が使える。ただし、どのような区切りで 2×2 クロス表の形に直すのかという点で、分析者の恣意性が入り込む可能性がある。
2. なんらかの相関係数を求める。ただし、カテゴリーに固定した順序はないので、係数の値の符号の正負を考えることは無意味である。求めたとしても係数は非負に限られる。

実用的には、1 の方法が直感的に容易に理解が可能であり、もっとも説得力がある。したがって、可能ならば 1 の方法をとることが望ましいが、それができない場合もあるので、ここでは、2 の方法について考えることにしよう。

2×2 クロス表と同様に $s \times t$ クロス表でも、全く相関のない場合とは、二つの変数が独立な場合である。したがって独立性の検定を考えることができる。つまり、2×2 クロス表と同様に $s \times t$ クロス表でも Pearson の χ^2検定量を計算することができるのである。O を観測度数(observed frequency)、E を期待度数(expected frequency)とすると、Pearson の χ^2検定量は、

$$\chi^2 = \Sigma_{セル} (O-E)^2/E = \Sigma_i \Sigma_j (f_{ij} - f_{i\cdot} f_{\cdot j} /n)^2 /(f_{i\cdot} f_{\cdot j} /n)$$

Pearson はこの χ^2の分布が近似的に自由度 $(s-1)(t-1)$ の χ^2分布になることを示した。2×2 クロス表の自由度が 1 になることは、 $s \times t$ クロス表の自由度 $(s-1)(t-1)$ の特殊な場合だということは容易にわかるだろう。

しかし、χ^2は 2×2 クロス表のときと同様に、標本の大きさ n に比例して、いくらでも大きくなってしまうという性質があるので、χ^2を相関係数として使うことには、2×2 クロス表のときと同様の問題がある。それでは、2×2 クロス表のときと同様に、

$$\varphi = (\chi^2/n)^{1/2}$$

で定義した φ 係数が使えるだろうか。無関連のときは、χ^2が 0 になるので、$\varphi = 0$ となるので問題はない。完全相関のときはどうだろうか。

そこでまず、 $s \times t$ クロス表での完全相関の例を考えてみよう。$s = t$ のときの代表的な例は、対角線上のセルだけが度数をもち、他のセルが全て度数 0 となる表 5.8 のようなクロス表となる。この表 5.8 が完全相関を示していることは、前章の量的データの散布図からも納得できるだろう。相関表の場合には、正の完全相関というと、この表 5.8 の場合に限られるわけだが、さらに、カテゴリー間に何か固有の順序が存在しているわけではないときには、この表 5.8 の変数 x のカテゴリーの順序を入れ替えたもの、変数 y のカテゴリーの順序を入れ替えたものも完全相関となっている。すなわち、各行、各列において、度数 0 ではないセルが 1 個だけのとき、完全相関というのである。

表 5.8　完全関連のクロス表の一例($s = t$ のとき)

x	y					
	y_1	y_2	y_3	・・・・・・	y_s	計
x_1	f_{11}	0	0	・・・・・・	0	f_{11}
x_2	0	f_{22}	0	・・・・・・	0	f_{22}
x_3	0	0	f_{33}	・・・・・・	0	f_{33}
：	：	：	：		：	：
x_s	0	0	0	・・・・・・	f_{ss}	f_{ss}
計	f_{11}	f_{22}	f_{33}	・・・・・・	f_{ss}	n

しかし、$s < t$ のときは、こう単純にはいかない。完全相関とは、y の値が決まると、x の値も決まるということであると考えて、

1. 各列において度数 0 でないセルは 1 個
2. 各行において度数 0 でないセルは 1 個以上 $t-s+1$ 個以下

となっているときに､完全相関と呼ぶことにしよう｡例えば表 5.9 のようなクロス表になる。

表 5.9　完全相関のクロス表の一例($s<t$ のとき)

x	y				
	y_1	y_2	y_3	y_4	計
x_1	f_{11}	0	0	0	$f_{1\cdot}$
x_2	0	f_{22}	f_{23}	0	$f_{2\cdot}$
x_3	0	0	0	f_{34}	$f_{3\cdot}$
計	$f_{\cdot 1}$	$f_{\cdot 2}$	$f_{\cdot 3}$	$f_{\cdot 4}$	n

このとき、φ 係数を計算すると、(5.4)式が使えて、

$$\varphi^2=f_{11}^2/(f_{1\cdot}f_{\cdot 1})+f_{22}^2/(f_{2\cdot}f_{\cdot 2})+f_{23}^2/(f_{2\cdot}f_{\cdot 3})+f_{34}^2/(f_{3\cdot}f_{\cdot 4})-1$$
$$=1+f_{22}/f_{2\cdot}+f_{23}/f_{2\cdot}+1-1=2$$

となる。一般的に、$s\times t$ クロス表($s<t$)の φ^2の値は、完全相関のときに 1 ではなく $s-1$ となるのである。そこで完全相関のときに 1 の値をとるように、次のような係数を考える。

$$V=\{\varphi^2/(s-1)\}^{1/2}\qquad s\leqq t$$

これは、クラマーの V 係数(Cramer's V)と呼ばれるもので、これまでのことからもわかるように、

$$0\leqq V\leqq 1$$

という、相関係数としてふさわしい性質をもっている。$2\times t$ クロス表のとき

$$V=\varphi$$

という性質があり、もちろん、2×2 クロス表のときも、両係数は一致するので、2×2 クロス表のときも含めて、一般にクロス表の相関係数としては、統一的にクラマーの V 係数を用いるようにしておくとよい。(表 5.2(B)で、$V=-0.17$ と負の値を示している理由については、第 5 節(2)を参照のこと。)

4. 2×2 クロス表への相関係数の適用

(1)ピアソンの積率相関係数

実は 2×2 クロス表に対してもピアソンの積率相関係数を適用することができる。考え方としては、2 値質的データとして扱おうというのである。そこでいま表 5.10 のように、一方のカテゴリーに 0、他方のカテゴリーに 1 という数値を与え、ピアソンの積率相関係数を適用してみよう。

表 5.10　2×2 クロス表

x	y		
	1	**0**	計
1	f_{11}	f_{10}	$f_{1\cdot}$
0	f_{01}	f_{00}	$f_{0\cdot}$
計	$f_{\cdot 1}$	$f_{\cdot 0}$	n

$x = f_{1\cdot}/n$

$y = f_{\cdot 1}/n$

$s_x^2 = (1 - f_{1\cdot}/n)\, f_{1\cdot}/n = (f_{0\cdot}/n)(f_{1\cdot}/n)$

$s_y^2 = (1 - f_{\cdot 1}/n)\, f_{\cdot 1}/n = (f_{\cdot 0}/n)(f_{\cdot 1}/n)$

$s_{xy} = \Sigma\, x_i\, y_i\, /n - x\cdot y = f_{11}/n - x\cdot y = f_{11}/n - (f_{1\cdot}/n)(f_{\cdot 1}/n)$

であるから

$$r_{xy} = s_{xy}/s_x s_y$$
$$= \{ f_{11}/n - (f_{1\cdot}/n)(f_{\cdot 1}/n)\} / \{(f_{0\cdot}/n)(f_{1\cdot}/n)(f_{\cdot 0}/n)(f_{\cdot 1}/n)\}^{1/2}$$
$$= (n f_{11} - f_{1\cdot} f_{\cdot 1}) / (f_{0\cdot}\, f_{1\cdot}\, f_{\cdot 0}\, f_{\cdot 1})^{1/2}$$
$$= \{(f_{11} + f_{10} + f_{01} + f_{00})\, f_{11} - (f_{11} + f_{10})(f_{01} + f_{11})\} / (f_{0\cdot}\, f_{1\cdot}\, f_{\cdot 0}\, f_{\cdot 1})^{1/2}$$
$$= (f_{11}\, f_{00} - f_{10}\, f_{01}) / (f_{0\cdot}\, f_{1\cdot}\, f_{\cdot 0}\, f_{\cdot 1})^{1/2} \qquad (5.6)$$

この 2 乗は(5.5)式の φ^2 と一致するので、

$r_{xy}^2 = \varphi^2 = V^2$

2×2 クロス表すなわち四分表に対して適用したピアソンの積率相関係数は、四分点相関係数(four-fold point correlation coefficient)、略して、点相関係数(point correlation coefficient)、あるいは四分積率相関係数または四分表に対するピアソンの相関係数といわれることもある。

ピアソンの積率相関係数 r_{xy} の絶対値は両変数の正の 1 次変換によって変わらない(相関係数の性質 2)ので、二つのカテゴリーが数値データとして与えられていさえすれば、このような $r_{xy}^2 = \varphi^2$ という関係は保証される。

(2)ケンドールの順位相関係数

2×2 クロス表の場合には、さらに、ケンドールの順位相関係数 τ_b(Kendall's τ_b)との間に、$\tau_b = r_{xy}$ という関係もある。いまケンドールの順位相関係数の考え方をクロス表にそのまま適用してみよう。

$\Sigma A = f_{11}\, f_{00}$

$\Sigma B = f_{01}\, f_{10}$

さらに、ΣC について考えてみよう。同じセルに属する二つの個体(観測値)のつくる対は、x についても y についても同順位なので、各セル (i, j) に属する f_{ij} 個の個体のすべての組み合わせ

${}_{fij}C_2 = f_{ij}\,(f_{ij} - 1)/2$

通りの対はすべて同順位となる。またセル $(1, 1)$ と $(1, 0)$ に属する個体間の対 f_{11}・f_{10} 通りも x について同順位であるし、セル $(0, 1)$ と $(0, 0)$ に属する個体間の対 f_{01}・f_{00} 通りも x について同順位である。こうしたことから、

$\Sigma C = f_{11}(f_{11} - 1)/2 + f_{10}(f_{10} - 1)/2 + f_{01}(f_{01} - 1)/2 + f_{00}(f_{00} - 1)/2 + f_{11}\, f_{10} + f_{01}\, f_{00}$

$2\Sigma C = f_{11}^2 + 2 f_{11}\, f_{10} + f_{10}^2 + f_{01}^2 + 2 f_{01}\, f_{00} + f_{00}^2 - f_{11} - f_{10} - f_{01} - f_{00}$
$= (f_{11} + f_{10})^2 + (f_{01} + f_{00})^2 - n$
$= f_{1\cdot}^2 + f_{0\cdot}^2 - n$

よって、

$n(n-1) - 2\Sigma C = n(n-1) - (f_{1\cdot}^2 + f_{0\cdot}^2 - n) = n^2 - f_{1\cdot}^2 - f_{0\cdot}^2 = (f_{1\cdot} + f_{0\cdot})^2 - f_{1\cdot}^2 - f_{0\cdot}^2 = 2 f_{1\cdot}\, f_{0\cdot}$

同様にして、ΣD についても、

$\Sigma D = f_{11}(f_{11} - 1)/2 + f_{10}(f_{10} - 1)/2 + f_{01}(f_{01} - 1)/2 + f_{00}(f_{00} - 1)/2 + f_{11}\, f_{01} + f_{10}\, f_{00}$

$n(n-1) - 2\Sigma D = 2 f_{\cdot 1}\, f_{\cdot 0}$

したがって、τ_b の定義と(5.6)式から、

$\tau_b = (f_{11}\, f_{00} - f_{10}\, f_{01}) / (f_{0\cdot}\, f_{1\cdot}\, f_{\cdot 0}\, f_{\cdot 1})^{1/2} = r_{xy}$

このように、2×2 クロス表では、ピアソンの積率相関係数やケンドールの順位相関係数を適用しても、その絶対値は φ 係数そして　V 係数に一致する。つまり、2×2 クロス表については、相関の大きさを測ることについては、これらの相関係数の間で一致が見られている。

(3)さまざまな相関の 2×2 クロス表

例題)　表 5.11 の 9 枚の 2×2 クロス表の相関係数として V 係数を求めよ。($V=\varphi$、$V^2=\varphi^2=r^2=\tau_b{}^2$、$r=\tau_b$ に注意)

表 5.11　様々な相関係数の 2×2 クロス表

(1) $V=$

x	y		
	y_1	y_0	計
x_1	55	45	100
x_0	45	55	100
計	100	100	200

(2) $V=$

x	y		
	y_1	y_0	計
x_1	60	40	100
x_0	40	60	100
計	100	100	200

(3) $V=$

x	y		
	y_1	y_0	計
x_1	65	35	100
x_0	35	65	100
計	100	100	200

(4) $V=$

x	y		
	y_1	y_0	計
x_1	70	30	100
x_0	30	70	100
計	100	100	200

(5) $V=$

x	y		
	y_1	y_0	計
x_1	75	25	100
x_0	25	75	100
計	100	100	200

(6) $V=$

x	y		
	y_1	y_0	計
x_1	80	20	100
x_0	20	80	100
計	100	100	200

(7) $V=$

x	y		
	y_1	y_0	計
x_1	85	15	100
x_0	15	85	100
計	100	100	200

(8) $V=$

x	y		
	y_1	y_0	計
x_1	90	10	100
x_0	10	90	100
計	100	100	200

(9) $V=$

x	y		
	y_1	y_0	計
x_1	95	5	100
x_0	5	95	100
計	100	100	200

《解答》　まず、$V=0$ (独立)のときの期待度数を求めてみよう。9 枚のクロス表はすべて同じ周辺度数なので、期待度数も同じになり、表 5.12(A)に示すような期待度数になる。つまり、どのセルにも同じ度数 50 が入ることになるので、これを a と置くことにしよう。すると、実は 9 枚のクロス表にある観測度数は、表 5.12(B)のように表現することができる。

表 5.12　期待度数と観測度数

(A)V＝0 (独立)のときの期待度数(a＝50)

x	y		
	y_1	y_0	計
x_1	a	a	$2a$
x_0	a	a	$2a$
計	$2a$	$2a$	$4a$

(B)クロス表(1)～(9)にある観測度数(a＝50)

x	y		
	y_1	y_0	計
x_1	$a+b$	$a-b$	$2a$
x_0	$a-b$	$a+b$	$2a$
計	$2a$	$2a$	$4a$

9 枚のクロス表のこの性質を使って、9 枚のクロス表の相関係数をまとめて計算してしまおう。

$$\chi^2 = b^2/a + b^2/a + b^2/a + b^2/a = 4b^2/a$$

$$V^2 = \varphi^2 = \{4b^2/a\}/(4a) = b^2/a^2$$

したがって、

$$V = b/a$$

このことから、クロス表 (i) の相関係数は $V = i/10$。つまり、クロス表(1)の相関係数は V＝0.1、クロス表(2)の相関係数は V＝0.2,..., クロス表(9)の相関係数は V＝0.9 ということになる。例えば表 5.11(2)は直感的に相関があるように見えるが(実際、$\chi^2 = 4\times100/50 = 8$ となり、0.5%水準で有意である)、相関係数は V＝0.2 しかない。このように、量的データの散布図などと比較して、クロス表では相関係数は一般に低めであり、高い相関係数は出にくいということを知っておくとよい。

5. SAS によるクロス表の作成と相関係数

(1)クロス表の作成

これまで述べてきたようなクロス表を作成し、独立性の χ^2検定とＶ係数などの相関係数を求めることは、SAS を使えば非常に簡単にできる。それには、第 2 章第 6 節で単純集計をする際に用いた FREQ プロシジャの TABLES 文中の変数の指定の仕方を変えるだけでよい。

PC 版 SAS
LIBNAME libname '￥パス名'; PROC FREQ DATA=libname.永久 SAS データ・セット名本体; TABLES 変数リスト*変数リスト/CHISQ; [OPTIONS NOCENTER;] RUN;

例)

```
LIBNAME SAVE '￥MYDIR';
PROC FREQ DATA=SAVE.JPC;
        TABLES I1*V2/CHISQ;
        OPTIONS NOCENTER;
RUN;
```

CMS 版 SAS
PROC FREQ DATA=永久 SAS データセット名; TABLES 変数リスト*変数リスト/CHISQ; [OPTIONS NOCENTER;] RUN;

例)

```
PROC FREQ DATA=SAVE.JPC;
        TABLES I1*V2/CHISQ;
        OPTIONS NOCENTER;
RUN;
```

やはり PC 版 SAS のプログラムの 1 行目を削除すると、形式的には CMS 版 SAS のプログラムと同じになる。

独立性の χ^2検定と *V* 係数などの相関係数を求めるために、TABLES 文のオプションとして、ここでは /CHISQ を指定してあるが、後で触れるようにさらに色々なオプションを追加指定することができる。

一般に、TABLES 文で「変数 1*変数 2」と指定すると、表 5.13 の形式のクロス表が作られる。

表 5.13　TABLES 文で「変数 2*変数 2」と指定したとき作成されるクロス表の一般形式

変数 1　　変数 2
Frequency
Percent
Row Pct

Col Pct	値 1	値 2	値 3	……	TOTAL
値 1					
値 2					
値 3					
⋮					
TOTAL					

さらに、TABLES 文で「変数リスト*変数リスト」を指定すると、「*」をはさんでできる変数の組合せすべてについてクロス表が作られる。例えば、

TABLES (A B)*C;　　は　TABLES A*C B*C;　　と指定したのと同じこと

TABLES (A B)*(C D); は　TABLES A*C B*C A*D B*D; と指定したのと同じことになるのである。

プログラム例を実行すると、表 5.14 のようなクロス表が作成されるはずである。しかし、実際には、このクロス表をディスプレイ画面で見ることはできない。これは OUTPUT ウィンドウに出力されたものをファイルに一旦保存してから、筆者が編集してつなぎあわせたものである。

表 5.14　プログラム例で作成されるクロス表

TABLE OF I1 BY V2

I1(SEX)　　V2

Frequency Percent Row Pct Col Pct	1. YES	2. NO	Total
1. MALE	508 58.39 66.67 93.04	254 29.20 33.33 78.40	762 87.59
2. FEMALE	38 4.37 35.19 6.96	70 8.05 64.81 21.60	108 12.41
Total	546 62.76	324 37.24	870 100.00

Frequency Missing＝37

注) ただし、ディスプレイ画面では、このクロス表は 1 行ずつに分割されて OUTPUT ウィンドウに表示される。

一応、この形式のクロス表について解説しておくと、このクロス表では各セルに表示されている数値は、クロス表の左上に表示されているとおり、

Frequency ‥‥‥度数
Percent　‥‥‥相対度数
Row Pct　‥‥‥行相対度数
Col Pct　‥‥‥列相対度数

という順に上から並んでいる。盛りだくさんの内容ではあるが、実はそのことが災いして、1 画面にたった 1 行しか入らないのである。1 画面 1 行に分割されて表示されるために、これでは視覚的にまったくクロス表の役割を果たさない。しかも、たとえ 1 画面で表示されたとしても、実際に表 5.14 を見てもわかるように、これでは各セルの中身が繁雑な印象を受ける。実は、これは表 5.2(A)のクロス表と同じものなのだが、とても表 5.2(A)のようには視覚的に訴えない。

(2)オプション

そこで、TABLES 文で、必要ではないものを表示させない次のようなオプションを使うことになる。

NOFREQ ‥‥‥度数を表示しない
NOPERCENT ‥‥‥相対度数を表示しない
NOROW ‥‥‥行相対度数を表示しない
NOCOL ‥‥‥列相対度数を表示しない

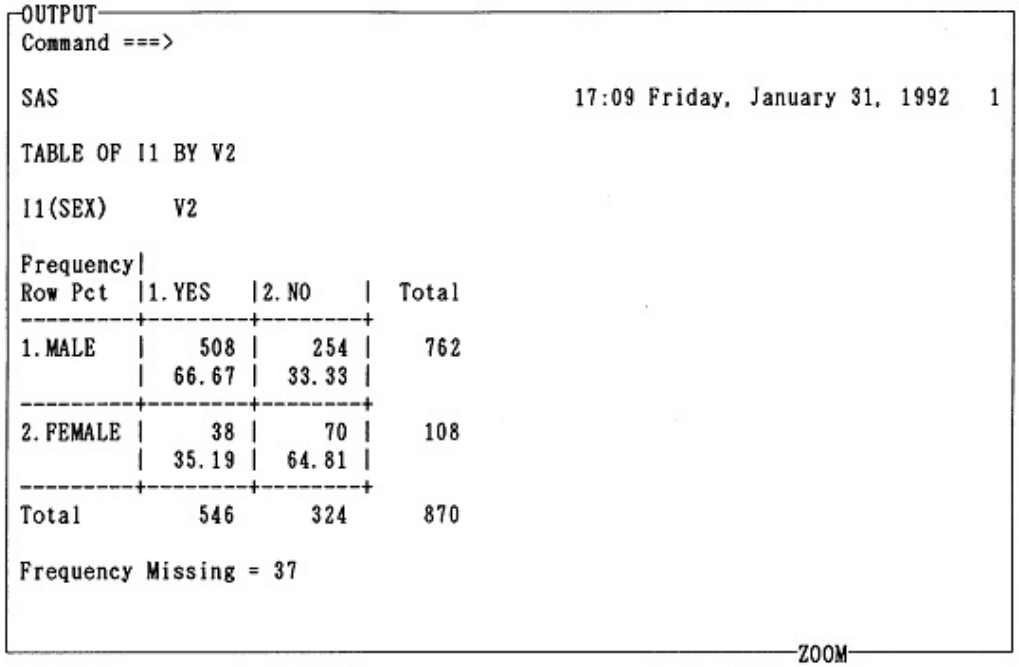
例えば、PC 版 SAS のプログラム例でいうと、筆者がよく指定するオプションは次の下線部のようになる。

例)

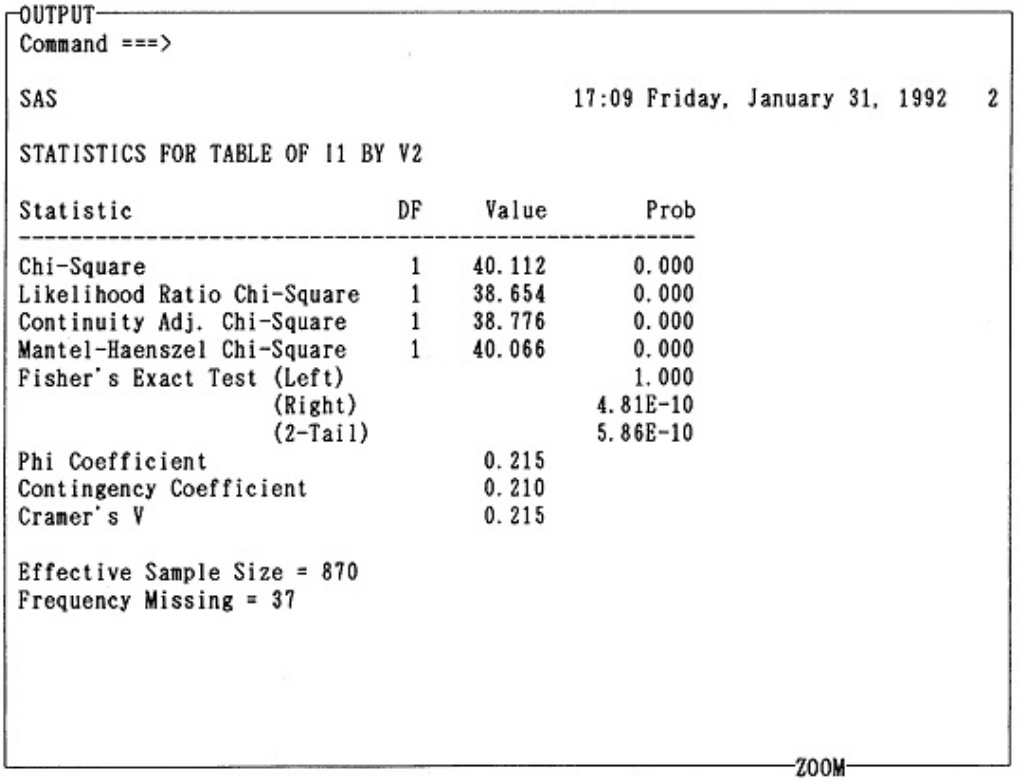
```
LIBNAME SAVE '￥MYDIR';
PROC FREQ DATA=SAVE.JPC;
        TABLES I1*V2
        /CHISQ NOPERCENT NOCOL;
        OPTIONS NOCENTER;
RUN;
```

この PC 版 SAS プログラムの 1 行目を削除すると、CMS 版 SAS プログラムとなる。このようにオプションを指定すると、度数と行相対度数だけがセルの中に表示されることになり、PC 版 SAS では 20 秒ほどで、図 5.1 のような簡素化されたクロス表が画面に出力されることになる。もっとも、簡素化されたといっても、図 5.1 の第 1 画面のクロス表は表 5.2(A)のクロス表と同じものである。情報量の多いことが良いクロス表の条件ではない。本当に必要な情報がコンパクトにまとめられていることが重要なのである。

図 5.1　SAS によるクロス表



```
OUTPUT
Command ===>

SAS                                          17:09 Friday, January 31, 1992   1

TABLE OF I1 BY V2

I1(SEX)     V2

Frequency|
Row Pct  |1.YES   |2.NO    |  Total
---------+--------+--------+
1.MALE   |    508 |    254 |    762
         |  66.67 |  33.33 |
---------+--------+--------+
2.FEMALE |     38 |     70 |    108
         |  35.19 |  64.81 |
---------+--------+--------+
Total         546      324      870

Frequency Missing = 37

                                                                        ZOOM
```



```
OUTPUT
Command ===>

SAS                                          17:09 Friday, January 31, 1992   2

STATISTICS FOR TABLE OF I1 BY V2

Statistic                     DF      Value        Prob
------------------------------------------------------
Chi-Square                     1     40.112       0.000
Likelihood Ratio Chi-Square    1     38.654       0.000
Continuity Adj. Chi-Square     1     38.776       0.000
Mantel-Haenszel Chi-Square     1     40.066       0.000
Fisher's Exact Test (Left)                        1.000
                    (Right)                    4.81E-10
                    (2-Tail)                   5.86E-10
Phi Coefficient                      0.215
Contingency Coefficient              0.210
Cramer's V                           0.215

Effective Sample Size = 870
Frequency Missing = 37

                                                                        ZOOM
```

この場合、2×2 クロス表なので、クラマーの V 係数(Cramer's V)と φ 係数(Phi Coefficient)はともに 0.215 で等しい。χ^2 (Chi-Square)は自由度(DF) 1 で、その値(Value)は 40.112 となる。独立性の χ^2検定の有意確率(Prob)は 0.000、つまり、有意水準を 0.1%に設定してある場合でも、有意であり、独立性の仮説は棄却されたことになる。

またオプションとして CHISQ を指定してあるときは、それに追加的に、次のようなオプションを指定して、クロス表のセル中に表示させることができる。

EXPECTED ・・・・・・「期待度数」を表示する

DEVIATION ・・・・・・「度数－期待度数」を表示する

CELLCHI2 ・・・・・・「(度数－期待度数)^2/期待度数」を表示する

計算されたこれらの数値を用いることで、各セルの度数が期待度数とどのような関係にあるか、特に CELLCHI2 では、χ^2値に対する各セルの貢献度を見ることができる。指定の仕方としては、例えば、PC 版 SAS のプログラム例でいうと、次の下線部のようにオプションを指定する。

例)

```
LIBNAME SAVE '￥MYDIR';
PROC FREQ DATA=SAVE.JPC;
        TABLES I1*V2
        /CHISQ NOPERCENT NOCOL DEVIATION;
        OPTIONS NOCENTER;
RUN;
```

この PC 版 SAS プログラムの 1 行目を削除すると、CMS 版 SAS プログラムとなる。

ここで注意が必要なのは、SAS では 2×2 クロス表の場合に限って、「φ 係数」「*V* 係数」でも負の相関があった場合には、負の値をとるようになっているということである。その意味では、このピアソンの積率相関係数を出力するようになっているともいえる。しかし、2×2 以外のクロス表では、本来の定義を用いて、正の値しか出力されない。本来の φ 係数、*V* 係数の定義や背景となる考え方からして、負の値をとることは間違っているが、ユーザーとしては、そのことを知った上で、2×2 クロス集計表の場合、出力される「φ 係数」「*V* 係数」の値をそのまま使うか、あるいは絶対値を使うかを決めれば良い。この章では、SAS の実行結果との対応をつけるために、出力された数値を負の値でもそのまま使っている。

(3)多数のクロス表への対処の仕方

ところで、変数の数が増えてくると、クロス表を 1 枚 1 枚見ていくのは大変な作業になってくる。「組織活性化のための従業員意識調査」では、Yes-No 形式の 75 変数を調べているが(詳しくは第 6 章の資料参照)、過去に行ってきた調査でもだいたい 100 変数前後の調査を行ってきている。仮に 100 変数のデータであったとすると、総当り方式では、100×100＝10,000 枚。このうち、同じ変数同士のクロス表 100 枚は見る必要がないし、同じ変数の組み合せのクロス表が 2 枚ずつ含まれているので、これらはどちらか一方だけ見ればよいことにしても、なんと (10,000-100)/2＝4,950 枚、実に 5,000 枚も見なくてはならない計算になる。メインフレームであれば、この程度の集計作業をさせてもあっという間に結果が出力されてくる。問題なのは、それを読む人間の能力の方に限界があるということである。まさに限定された合理性である。やってみればすぐにわかることだが、直感に訴える 2×2 クロス表でも 100 枚読むのは大変な作業である。一般の *s*×*t* クロス表では 1 枚 1 枚のクロス表の解読に時間がかかり、絶望的な作業になる。ラインプリンターに出力した場合、通常、用紙 1 ページに 1 枚のクロス表が印刷されるので、約 5,000 枚ということは、実にラインプリンター用紙の段ボール箱 2～3 箱分に相当することになる。私の知る範囲で、これだけの量のクロス表の山を読みこなした人間は存在しない。だいたいラインプリ

ンター用紙をめくるだけで、腕が筋肉痛になってしまう。内容を理解するなどもはや不可能である。しかし、総当りで変数間の関係をつかんでおくことは、分析に際して重要なことなのである。

それでは、筆者はどのようにしてきたのであろうか。それは

1. できる限り、質問を Yes-No 形式に統一し、2×2 クロス表で済むようにする。
2. 2×2 クロス表であれば、*V* 係数の絶対値とピアソンの積率相関係数は一致する。しかも、ピアソンの積率相関係数は正負がわかるので、クロス表をいちいち見なくても、相関係数によって大まかな相関関係がわかる。
3. クロス表は総当りで作成することはしないで、代わりに相関係数行列を作成する。特に必要になった場合に限って、必要なクロス表を 1 枚 1 枚作成、出力させる。

ちなみに、筆者がよく作成する 15 変数×15 変数の相関係数行列表 1 枚で 225 枚のクロス表の要約が可能なので、さきほどの 100 変数のケースでも、20 枚ちょっともあれば十分である。

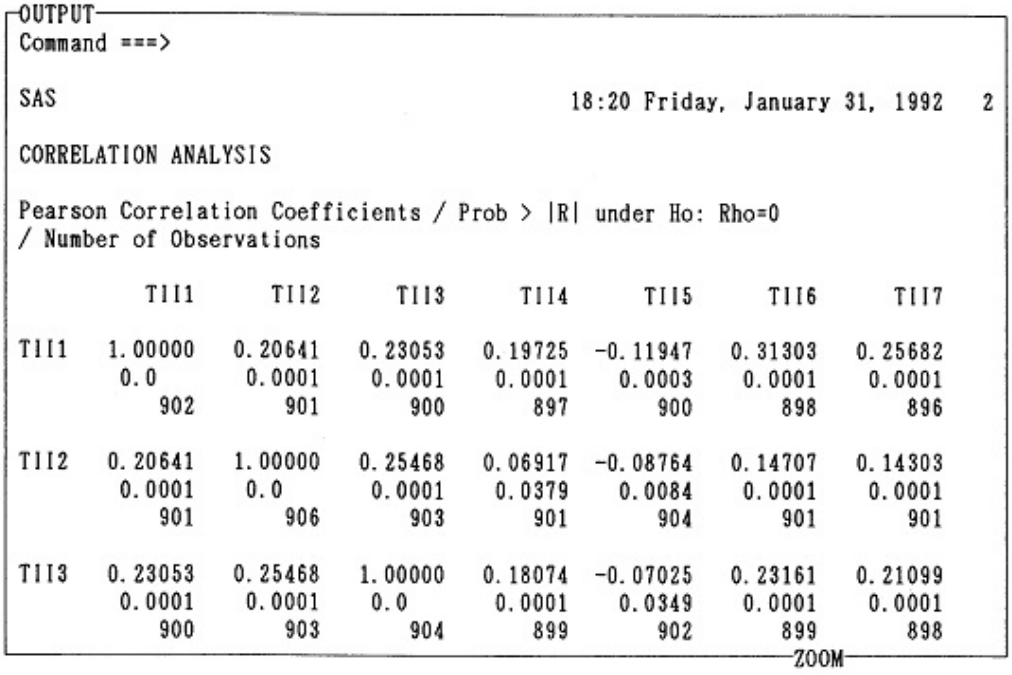
CORR プロシジャを使えば、PC 版 SAS のプログラム例でいうと、

例)

```
LIBNAME SAVE '￥MYDIR';
PROC CORR NOSIMPLE DATA=SAVE.JPC;
        VAR TII1 -TII7;
        WITH TII1 -TII3;
        OPTIONS NOCENTER;
RUN;
```

のように指定すると、30 秒ほどでちょうど 1 画面分に相当する 3 変数×7 変数の相関係数行列が図 5.2 のように画面に表示される。各セルで 1 行目は相関係数、2 行目はその相関係数の有意確率(帰無仮説は母相関係数 ρ＝0)、3 行目はその相関係数を計算する際に欠損値となったものを除いたオブザーベーション数を示している。15 変数×15 変数の相関係数行列を出力するには約 2 分を要する。この PC 版 SAS プログラム例の 1 行目を削除すると、CMS 版 SAS プログラムとなる。

図 5.2　SAS によるクロス表の相関係数行列



```
OUTPUT
Command ===>

SAS                                          18:20 Friday, January 31, 1992   2

CORRELATION ANALYSIS

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0
/ Number of Observations

          II1       II2       II3       II4       II5       II6       II7

II1   1.00000   0.20641   0.23053   0.19725  -0.11947   0.31303   0.25682
        0.0       0.0001    0.0001    0.0001    0.0003    0.0001    0.0001
          902       901       900       897       900       898       896

II2   0.20641   1.00000   0.25468   0.06917  -0.08764   0.14707   0.14303
        0.0001    0.0       0.0001    0.0379    0.0084    0.0001    0.0001
          901       906       903       901       904       901       901

II3   0.23053   0.25468   1.00000   0.18074  -0.07025   0.23161   0.21099
        0.0001    0.0001    0.0       0.0001    0.0349    0.0001    0.0001
          900       903       904       899       902       899       898
                                                                    ZOOM
```

この図 5.2 のような SAS によって出力されたクロス表の相関係数行列は、そのままでは使わない。この OUTPUT ウィンドウを一旦ファイルに保存してから、そのファイルを編集して、表としての形式を整えて、表 5.15 のような表を作成するのである。ここで、有意確率は数字のままでは、視覚的にピンとこないので、記号化しておくとともに、各変数には第 6 章資料 D の入力フォーマットにある変数ラベルを使って、簡単な内容説明をしておく。

表 5.15　クロス表の相関係数によるまとめの例

	II1. 常に仕事改善心掛	**II2.** 先例に拘らず仕事	**II3.** 境界に拘らず仕事	**II4.** 自分他社でも通用	**II5.** 上司に素直に従う	**II6.** 経営方針考え仕事	**II7.** 上司から権限委譲
II1. 常に仕事改善心掛	1.00000 902***	0.20641 901***	0.23053 900***	0.19725 897***	-0.11947 900***	0.31303 898***	0.25682 896***
II2. 先例に拘らず仕事	0.20641 901***	1.00000 906***	0.25468 903***	0.06917 901*	-0.08764 904**	0.14707 901***	0.14303 901***
II3. 境界に拘らず仕事	0.23053 900***	0.25468 903***	1.00000 904***	0.18074 899***	-0.07025 902*	0.23161 899***	0.21099 898***

+ p <0.1;　* p <0.05;　** p <0.01;　*** p <0.001

6. エラボレイション

(1)クロス表のより深い分析

まず例から始めよう。いま「組織活性化のための従業員意識調査」で調べた次の二つのYes-No 形式の質問に対する回答のクロス表を作ってみよう。

IV7. 指示が出されても、やり過ごしているうちに、立ち消えになることがある。

1. Yes　2. No

VI7. 福利厚生面は充実している。

1. Yes　2. No

クロス表は表 5.16 のようになる。この 2 問の質問、常識的に考えてほとんど関係のない質問のように思えるのだが、驚いたことに、この両者の間には有意な負の相関、しかもクロス表としてはかなり大きめの相関係数で-0.218 の相関がある。実は、やり過ごしについての質問 IV7 と他の 74 の質問項目との相関をすべてとってみたが、この VI7 の福利厚生との相関係数の大きさは、74 問中 2 番目の大きさだったのである。

表 5.16　指示のやり過ごしと福利厚生のクロス表

IV7 指示やり過ごし可	**VI7　福利厚生面は充実**		
	1. Yes	**2. No**	**計**
1. Yes	267 (45.56)	319 (54.44)	586
2. No	205 (68.56)	94 (31.44)	299
計	472	413	885

Cramer's V＝-0.218　χ^2＝42.075***

こうした他変数との比較と表 5.16 とをふまえれば、このクロス表の示唆する傾向は明らかである。つまり、指示のやり過ごしができないところでは福利厚生面が充実しているし、逆に、指示のやり過ごしができるところでは福利厚生面が充実していないというのである。総数が 885 人もいると、χ^2検定は有意になりやすいのは確かだが、表 5.16 のクロス表を見ても、その傾向ははっきりしている。要するに、福利厚生面がしっかりしているところは、指示もしっかりしているということなのだが……。

こうした、どうも常識的に納得できない、根拠のはっきりしない調査結果がでた場合、これぞ事実発見と手放しで喜ぶのはまだ早い。だいたいは何か他に、よりもっともらしい理由があるもので、そうした常識的な可能性をしらみつぶしにしてみるまでは結論は出せない。

そこで試しに、調査対象企業を公益事業関係とそれ以外の流通業などに分けた上で、別々にクロス表を作ってみると、表 5.17 のようになった。まだどちらのクロス表も有意な負の相関関係があるが、相関係数であるクラマーの V は随分と小さくなった。つまり、指示のやり過ごしと福利厚生面での充実との相関は随分と小さくなっている。しかも、表 5.17 のクロス表をよく見ると、非公益事業では、指示のやり過ごしが多く、福利厚生面で

はまだ充実していないということがわかる。それに対して、公益事業では、福利厚生面は充実しているし、指示のやり過ごしも非公益事業に比べれば少なくなっているということがわかる。

表 5.17　指示のやり過ごしと福利厚生のクロス表

(A)公益事業

IV7 指示やり過ごし可	VI7　福利厚生面は充実		
	1. Yes	2. No	計
1. Yes	223 (66.57)	112 (33.43)	335
2. No	187 (76.95)	56 (23.05)	243
計	410	168	578

Cramer's V＝-0.113　χ^2＝7.371**

(B)非公益事業

IV7 指示やり過ごし可	VI7　福利厚生面は充実		
	1. Yes	2. No	計
1. Yes	44 (17.53)	207 (82.47)	251
2. No	18 (32.14)	38 (67.86)	56
計	62	245	307

Cramer's V＝-0.141　χ^2＝6.066*

表 5.16 に見られる高い相関は、このように性質の異なる 2 群、公益事業と非公益事業とを合わせて集計したために出現したものと考えられそうである。常識的に考えてみると、巨大な設備や装置を管理運用している公益事業で、指示のやり過ごしが少ないのは当然であるし、独占や地域独占を前提としている公益事業が福利厚生面を充実させる余裕があるというのもまた当然かもしれない。これとてまだ推測の域を出ないわけで、フォロー・アップ・ヒアリングや追試を必要としてはいるが、指示のやり過ごしと福利厚生という直接的には本来無関係なはずの変数が、実はともに企業の業種、業態、市場環境などとは密接に結び付いていそうだということは、この表 5.17 でも十分に示されているといえるだろう。

(2)3 重クロス表

いまの分析は、やり過ごしと福利厚生という 2 変数のクロス表に、さらに第 3 の変数として一種の「業種」を導入して行なったものである。このように、これまでに扱ってきたような 2 変数のクロス表に、さらに第 3 の変数を導入して、変数間の関係を明らかにしていく方法を一般に、エラボレイション(elaboration)と呼ぶ。より具体的にいえば、確認したいと思っている独立変数 x と従属変数 y との間の関連を示すクロス表に、第 3 の変数 t を導入して、3 重クロス表を作成するのである。このときの第 3 変数 t のカテゴリー(t_1, t_2, ...)ごとに作られた x と y のクロス表の相関を分割相関(split correlation)、または、条件相関(conditional correlation)、層別相関(stratified correlation)と呼ぶが、この分割相関を調べることで、変数間の関係が明らかにされるのである。

既に(1)の表 5.16、表 5.17 で示したようなケース、極端には x と y との間には相関が見られるのに、分割相関が全くない表 5.18 のようなケースは何を意味しているのだろうか。これには次の二つの可能性が考えられる。

表 5.18　極端にした架空例

(A)公益事業　＋　(B)非公益事業　＝　(C)全体

IV7	VI7		
	Yes	No	計
Yes	9	1	10
No	81	9	90
計	90	10	100

Cramer’s V＝0

χ^2＝0

IV7	VI7		
	Yes	No	計
Yes	9	81	90
No	1	9	10
計	10	90	100

Cramer's V＝0

χ^2＝0

IV7	VI7		
	Yes	No	計
Yes	18	82	100
No	82	18	100
計	10	100	200

Cramer's V＝-0.64

χ^2＝81.92***

(a)エクスプラネーション(explanation)

これは、第 3 の変数 t が x と y に対する先行変数(antecedent variable)になっている場合である。図式化すれば、

$x \leftarrow t \rightarrow y$

ということになり、疑似相関(spurious correlation)を説明することになる。疑似相関とは、 x と y には因果関係が存在せず、実際、x を人為的に変化させても y に変化は生じない。さきほどの(1)の例もこの疑似相関に相当する。つまり、「福利厚生を充実させると、指示のやり過ごしが減少するように見えるけれども、実は、これは大部分が、『業種』という先行変数があるための見かけ上の疑似相関であると説明される」のである。したがって、この調査結果から、指示のやり過ごしを減らすためには、福利厚生を充実させればよいのだと、短絡的に考える人がいれば(そんな人はいないと思うが)、その人は見かけ上の疑似相関に惑わされて、真の因果関係を見過ごしたということになる。

しかし、このような x と y との間には相関が見られるのに、分割相関がない場合にはもう一つ別の次のような可能性も考えられるので注意がいる。

(b)インタープリテーション(interpretation)

これは、第 3 の変数 t が、x と y との間の媒介変数(intervening variable)になっている場合で、図式化すると次のようになる。

$x \rightarrow t \rightarrow y$

間接的な因果関係をより詳しく解釈したものである。しかしこの場合には、直接的には因果関係がないとはいえ、間接的には因果関係があるので、見かけだけの疑似相関とは異なる。

ただし、(a)のタイプになるか、(b)のタイプになるか、つまり因果関係の矢印の方向がどうなるかは常識や学問的蓄積等を動員して、思考実験によって行う以外にはない。つまり、それは統計処理上の問題ではなく、論理の問題である。

最後に、(a)(b)とは逆のケースとして、次の(c)があるのであげておこう。

(c)スペシフィケーション(specification)

もとの x と y との相関が t_1 と t_2 とによって程度が異なっていることが発見される場合である。極端な場合には、全体の単純相関は 0 であるが、それが実は、相反する方向の二つの分割相関の合成の表面的結果であったことを示している。つまり、疑似無相関

(spurious non-correlation)である。疑似無相関の場合には、全体では無相関であっても、実際には適当にグループ化、層別化(stratification)を行うことで、各グループ内での相関を明確に示すことができる。

(3)SAS による 3 重クロス表の作成

3 重クロス表を作成し、*V* 係数のような相関係数を求めることは、SAS の FREQ プロシジャを使えば非常に簡単にできる。

PC 版 SAS

```
LIBNAME libname '￥パス名';
PROC FREQ DATA=libname.永久 SAS データ・セット名本体;
TABLES  変数 1*変数 2*変数 3
[/オプション];
[OPTIONS NOCENTER;]
RUN;
```

例)

```
LIBNAME SAVE '￥MYDIR';
PROC FREQ DATA=SAVE.JPC;
        TABLES KCODE*IV7*VI7
        /CHISQ NOPERCENT NOCOL;
OPTIONS NOCENTER;
RUN;
```

CMS 版 SAS

```
PROC FREQ DATA=libname.永久 SAS データ・セット名;
TABLES  変数 1*変数 2*変数 3
[/オプション];
[OPTIONS NOCENTER;]
RUN;
```

例)

```
PROC FREQ DATA=SAVE.JPC;
        TABLES KCODE*IV7*VI7
        /CHISQ NOPERCENT NOCOL;
OPTIONS NOCENTER;
RUN;
```

CMS 版 SAS のプログラム例は、形式的に PC 版 SAS のプログラム例の 1 行目を削除したものになる。このプログラムによって、変数 1 の各カテゴリーごとに、変数 2*変数 3 のクロス表が作られることになる。ただし、全体のクロス表はこれでは作られないので、別に作る必要がある。(→演習問題 5.1)

演習問題

5.1 エラボレイション　第 2 節で表 5.2 のクロス表を作成するときに使用した質問 V2 と質問 V12 との間でクロス表を作成するとともに、第 3 の変数として、性別に関する質問 I1 を用いた 3 重クロス表を作成すると次のようになった。この SAS の出力(一部省略)をどのように理解できるかを述べよ。

TABLE OF V2 BY V12

V2 V12

Frequency	1.YES	2.NO	Total
1.YES	239	325	564
2.NO	117	216	333
Total	356	541	897

Statistic	DF	Value	Prob
Chi-Square	1	4.586	0.032
Cramer's V		0.072	

TABLE 1 OF V2 BY V12
CONTROLLING FOR [1=1.MALE

V2 V12

Frequency	1.YES	2.NO	Total
1.YES	207	300	507
2.NO	75	178	253
Total	282	478	760

Statistic	DF	Value	Prob
Chi-Square	1	9.046	0.003
Cramer's V		0.109	

TABLE 2 OF V2 BY V12
CONTROLLING FOR [1=2.FEMALE

V2 V12

Frequency	1.YES	2.NO	Total
1.YES	29	9	38
2.NO	38	32	70
Total	67	41	108

Statistic	DF	Value	Prob
Chi-Square	1	5.075	0.024
Cramer's V		0.217	

5.2　疑似無相関　全体の相関は 0 であるが、それが実は、相反する方向の二つの分割相関の合成の表面的結果であったというような疑似無相関のもっともらしい架空例を考え、表 5.16、表 5.17、表 5.18 を参考にしてクロス表を作成せよ。

第 6 章　調査の手順と実際:

「組織活性化のための従業員意識調査」マニュアル

章目次

1. はじめに
2. 基本設計
3. 調査票の設計
4. 質問票調査
演習問題
資料 A. 基礎調査票と調査の基本設計
資料 B. 質問調査票
資料 C. 現地調査の手引
資料 D. 入力フォーマット
資料 E. 配布・回収状況一覧表
資料 F. 単純集計

1. はじめに

本書では、統計調査の実際の姿をできるだけ具体的かつ明確にイメージしてもらうために、統計処理、分析に用いるデータ例をはじめ、SAS プログラムの例なども、一貫して「組織活性化のための従業員意識調査」を素材としている。この調査手順と方法を、実際に用いられた詳細な資料とともに取り上げよう。もっとも、この章は単に具体例を提示するというだけではなく、もともと統計調査マニュアルを意図して書かれているので、読者が近い将来、自分の手で統計調査を企画、実施する際のマニュアルとして生かされることを希望している。

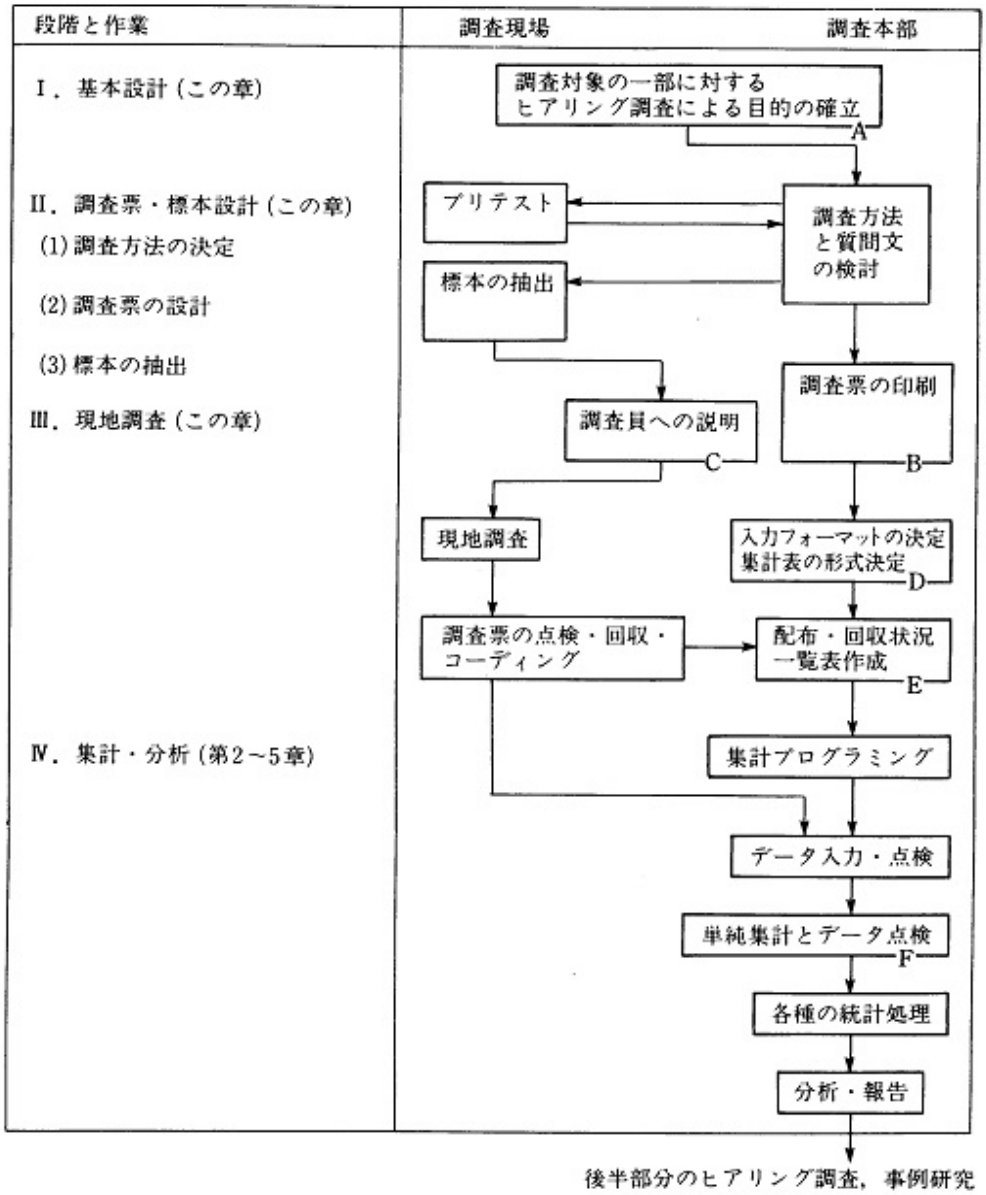
「組織活性化のための従業員意識調査」自体は 1986 年以来毎年 6 月～翌年 1 月の 8 カ月をかけて、(財)日本生産性本部経営アカデミー「人間能力と組織開発」コースを舞台にして、筆者によって繰り返し繰り返し企画・実施されてきたものである。1991 年までに、のべで約 50 社、約 5,000 人を調べてきたことになる。

この調査では、企業間での横断的な研究グループを作り、このメンバー企業の間で同時に同一の意識調査を行い、企業間・職場間の比較集計を行う。調査結果の企業間比較によって自社の抱える問題点を探り、他社の実例や知恵も借りながら、自社等の事例研究を行うというプロセスを繰り返すことで、研究グループを中心として組織の活性化に関する問題発見を図るプログラムとなっている。この「組織活性化のための従業員意識調査」では、従業員の意識調査を行なうことを契機として問題発見プロセスが進んでいくのである。

そのうち本書で扱うのは、調査研究プロセスの前半に相当する質問票調査、いわゆるアンケート調査の企画、設計から集計、分析に至るまでであるが、集計、分析については、既に第 2～5 章で扱ってきたので、この章では通常の社会調査のプロセスの区分で言うと、基本設計、調査票・標本設計、現地調査を扱うことにする。

それでは、「組織活性化のための従業員意識調査」の手順を、順を追って説明していこう。実際の調査手順は図 6.1 のようなフローチャートにまとめることができる。このフローチャートにそって、ある年(19XX 年)に行われた「組織活性化のための従業員意識調査」で、実際に使用された資料をこの章末に掲載してある。A～F の資料が調査のどの段階で使用されたものかは、図 6.1 のフローチャートの中に示してあるので、参考にしてほしい。ただし、調査内容には非公開を前提にしているものも含まれており、会社名等、公開して各方面にご迷惑のかかるおそれのあるものについては、一部、名前を伏せたり、特定不可能なように内容を変えたりしている。

図 6.1　「組織活性化のための従業員意識調査」の前半部分(質問票調査)の調査手順と使用資料



2. 基本設計

(1)目的の確立

まず調査を企画する際には、調査の目的を明確にする必要がある。具体的に、経験、一般常識、あるいは理論に基づいて仮説を立てると、調査目的は次第に明確化、具体化してくる。

この際に注意が必要なのは、多くの仮説をそれぞれ少数の調査項目で確かめるよりも、少数の仮説を多数の調査項目で確かめる方が調査の効率も高く、また分析にも深みが出るということである。したがって、調査目的となる仮説は、できるだけ絞った方がよい。多くの仮説が互いにきちんとした関係も与えられずに並立しているような場合は、仮説を立てるという立場からすると努力不足である。モデルは単純で、明解でなければ説得力をもたないし、モデルとはいえないのである。

(2)ヒアリング(聴き取り)調査

調査の目的を確立し、統計調査の基本的な方針を立てるためには、まず入念なヒアリング調査から始めるのがよい。事前にかなりはっきりした目的や知識をもっているような場合でも、実際の調査対象予定者に対してヒアリング調査を行うことは、調査票の設計等を考えると極めて有益である。ヒアリング調査は面接調査の一種であるが、第 1 章第 4 節で述べた調査票を用いた面接調査法が、調査票を用いることで、(a)質問の仕方、(b)回答の仕方、を厳格に定め、統一していて、指示的(directive)面接調査と呼ばれるのに対して、不定型で調査者の自由に任されているために、非指示的(non-directive)面接調査と呼ばれる。

ただし、実際的には、非指示的なヒアリング調査に指示的な面接調査を組み合わせて行った方が効率的である。「組織活性化のための従業員意識調査」では、章末資料 A にあるように、各組織が抱える問題点や各メンバーの問題意識を探る非指示的なヒアリング調査を行っているが、その前に、それと抱合せの形で、「基礎調査票」を使って、指示的な面接調査も必ず行っている。これは、一つには、ヒアリング調査の際の背景的知識を獲得するということであるが、この目的とともに、企業によって異なる「言葉」や「概念」を明確に認識し、ヒアリング調査の際に、各社ごとの「方言」をできるだけ「標準語」に翻訳しながら会話することが出来るようになるために必要な作業なのである。こうした一連の作業とヒアリング調査を通じて、はじめて人々が各企業、各組織の中で、方言を使い、独特な論理を使って動いているということが、実感として認識されるのである。また、こうしたプロセスを経なければ、複数の企業、部門で共通して使用が可能な質問調査票を作ることができない。一般の社会調査、世論調査と比較しても、経営組織調査が難しい最大のポイントは、企業活動の核心的部分に近付けば近付くほど、企業によって「方言」が多く使われていて、しかもそのことに内部の人間が気づいていないという点である。

そうしたこともあって、この段階で、調査対象予定の者の一部に対して、入念なヒアリング調査を行うことが望ましい。質問票調査によって検証してみたい仮説がある場合には、ヒアリング調査によって、仮説の妥当性を(仮説が、調査対象にとって意味のあるものかどうかも含めて)ある程度確かめておく必要がある。

調査によって出てくる統計数字は、こうしたヒアリング調査による心証の精度を上げるものに過ぎない、という程度に考えていた方がよい。

(3)ヒアリング調査の進め方

それでは、「組織活性化のための従業員意識調査」で、ヒアリング調査が具体的にどのように行われるのかをまとめておこう。ヒアリング調査は章末資料 A の基礎調査票を併用

しながら行われる。「組織活性化のための従業員意識調査」では、調査の開始に当って、例年 6 月に第 1 回の合宿が行われる。章末資料 A にもあるように、第 1 回合宿の第 1 目標は、今後のグループ研究の討議の際の基礎となるキーワード、コンセプトの認識をできるだけ統一しておくこと、もしくは認識の違いを明確にしておくことにある。もちろん親睦が大事なわけだが、ただなかよくなるのではなく、各々の企業の中では、その企業の方言を使って会話が行われていること、そして、「うち常識＝よその常識」という暗黙の前提あるいは仮説が間違いであることをある程度気付かせることが大切である。実際には調査のプロセスが進んで、質問票調査の統計数字が出てくるまでは、なかなか納得の出来ないものらしいが、この最初の合宿は、そのきっかけとなるように行われる必要がある。

合宿は、例年 3 日間にわたって、グループのメンバー全員をホテルに缶詰にして、公式には 12 時間前後、実際には非公式にも夜、アルコールが入ってから深夜に至るまで行われるので、それプラス 8 時間程度はヒアリングが行われる。このヒアリングでは、まず 1 人 1 時間程度を目安にしてもらって、会社の概要紹介と会社の特長と問題点を報告してもらう。報告は一方的に発表するのではなく、質問がある人は誰でも、いつでも、その場で質問が出来るようにしておく。この段階では、第三者として入っている研究者が、率先して質問していく必要があるので注意がいる。何回かこうした調査を繰り返してみると自然にわかってくることだが、業種や企業の個性によっても、使われる状況や意味の異なる用語、概念が必ずあるので、そうした点をこまめに質問することで、次第に議論の糸口が見えてくるものである。そうした経験がたとえなくても、知ったふりをしないで、素朴に質問を積み上げていくことが肝要である。

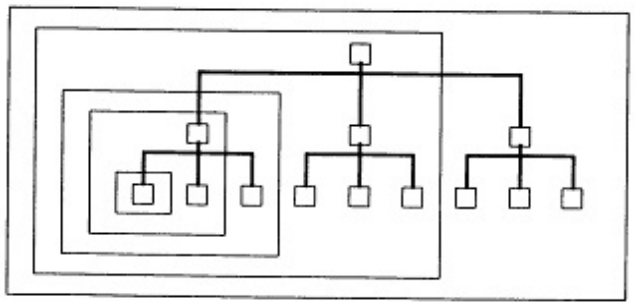
会社の概要紹介は所定の用紙に則った形で要点だけということになっているが、あらかじめ、章末資料 A にもある所定の基礎調査票に記入してきた上で概要紹介を行うので、情報量としては十分なものとなる。質問が許されているので、通常公表されているデータを読むのとは異なり、その企業の実像や雰囲気を知ることが出来る。基礎調査票にある項目は例年の合宿の際に、こうしたグループ研究で話題になった事柄、あるいは、当初の基礎調査票には記入欄がなかったために、その場で黒板などを使って説明が行われたようなことを年々追加しながら形成されてきたものである。言い換えれば、最初に明らかにしておかなかったために、議論の混乱の原因となったものを公約数的にまとめたものということもできる。

こうした基礎知識を基にして、会社の特徴と問題点を会社の活性化に対する自分の問題意識を中心にして報告してもらう。これは非指示的なヒアリング調査であるので、所定の用紙は用いないが、身近なエピソードも交えて、活性化に対する自分の考え、問題意識をまとめてもらい、一応口頭だけではなく、紙に書いてきてもらうことにしている。これは「書く」という作業が、自分の考えをまとめるために有用だからである。

章末資料 A の基礎調査票の組織図は、質問票調査の調査対象となる組織単位を選ぶためにも用いられるもので、調査開始時期に当るこの第 1 回合宿の段階で、早めに各自が想定している調査対象組織単位を表示してもらうことになる。これは、ヒアリング調査は一般論を話されても調査にはならないので、具体的に、企業のどの部分を念頭に話しているのかを明らかにしておく必要があるからである。さらにいえば、実は今日の大企業では企業規模が非常に大きいために、1 人の人が説明できる「企業」は実際の企業のごく一部だけにすぎない。後で行われる質問票調査は、こうした「説明可能」な部分だけに対象を限定して行われる。そうしなければ、統計数字のもっている意味、つまり実態を明確に確定することは出来ないからである。したがって、自分の所属部署を明示の上、調査してみたいと思っている組織単位をいくつか設定、図示してみてもらい、その際、できれば組織単位ごとの大まかな人数も記入してもらい、「説明可能」な部分を確認してもらうのである。

具体的には、人員規模 50 人以上 100 人未満のホワイトカラーの集団を組織単位として、一つまたは複数選んでもらうことになる。ここで、「組織単位」とは組織図上で同一の上司を持つ職場もしくは職場の集合で、例えば、図 6.2 のような組織図では細線で囲まれているような各レベルの部分に相当するのだということをあらかじめきちんと説明しておく。

図 6.2　「組織単位」の設定



このように、単に「職場」とはせずに、組織単位を設定するのは、分析の段階で、組織単位間の比較分析も行うわけだが、職場間比較では、各職場の人員規模に開きがありすぎるために(場合によっては 10 倍もの差がある)、統計的には比較が難しいためである。ただし、性質の全く異なる職場を一つの組織単位に押し込めるようなことはしない。その場合には、たとえ人員規模が小さくても、組織単位は別立てにする。

(4)質問票調査についての了解

この第 1 回合宿の段階で、調査スケジュールの見通しを明らかにしておくことは重要である。前年度を踏襲する形で行うと、どういう調査スケジュールになるのかを示し、そのスケジュールで何か問題はないか、あるいは、改善すべき点はないかをこの段階で既に確認しておく必要がある。おおざっぱにいえば、「組織活性化のための従業員意識調査」では例年第 1 回合宿の後、7 月に 2 回のグループ研究を行い、質問案の検討・討議を行った後で、筆者がそれらを取りまとめる形で質問調査票に仕上げ、8 月末から 9 月はじめにかけて、1 週間程度をかけて調査票配布・回収が行われる。集計は 1 週間もあればできるので、比較的早く調査結果をフィードバックすることが出来る。フィードバックされる集計表は、具体的には次のようなものである。

1. 配布・回収状況一覧表(章末資料 E)
2. 単純集計(章末資料 F)
3. 個人属性によるクロス表(たとえば第 5 章第 2 節の表 5.2)
4. 企業間比較のためのクロス表(非公開)
5. 組織単位間比較のためのクロス表(非公開)
6. Yes-No 形式の質問間の相関係数行列(たとえば第 5 章第 5 節の表 5.15)

3. 調査票の設計

調査票の設計、特に質問文の作成は、もっとも注意深く行われるべき作業の一つであり、また、もっとも面白い作業の一つでもある。「組織活性化のための従業員意識調査」の場合には、このプロセスは、各企業の担当者から質問文の候補を募る形で始められる。つまり、自分の組織を調査するにあたって、これだけは調べてほしいこと、これだけは知

りたいということを、各自が自分で質問文として考えてくるように「課題」として与えるのである。

この作業は少なくとも 2 度繰り返される。1 度目は各自ができるだけ「自分の頭で」質問文を考えてくるという作業で、できるだけ具体的な、実際の企業人が直感的に理解できるような質問文を考えてもらう。そうして作られた質問文は、グループ全体で討議される。その際の討議のポイントは

1. 各質問の文言だけでなく、質問の真意と目的。
2. 各質問文が、そのままで他の企業においても使用可能かどうか、つまり通じるかどうか。
3. 各質問に対して、その考案者が自分の組織でどのような回答を期待しているのか。

ということである。経験的には、この段階で持ち寄られる各質問文は、無意識のうちに、それぞれの企業の「方言」で書かれていることが多い。そこで基本的には、どの企業でも使えるように「標準語」に書き直す作業が行われるのである。

2 度目は、1 度目に他のメンバーが考えてきた質問項目も含めたグループ全体の全質問案の中から、「他人の質問を盗んで」もいいから、これだけは質問案の中に入れておいてほしいという質問文をもう一度持ち寄り、1 度目と同じ作業を繰り返すのである。

こうした作業での表立った目的は、「標準語」で書かれた共通の質問票を作成するということであるが、実は、もう一つ重要な目的がある。それは、各メンバーの問題意識を掘り起こすということである。既にヒアリング調査の際に、各自の問題意識についても聞いているはずではあるが、その段階では、明解かつ直接的に問題意識が表明されるということを期待するのは実はまだ無理である。しかし、各自に、これだけは調べたいという質問文を考えさせ、また選択させることによって、潜在的であいまいだった問題意識が明解な形で表現されてくることになる。しかも、こうした質問文はそれぞれが一つの仮説を表明していると考えられる。各メンバーには意識されてはいなくても、少なくとも「自分の組織では、この質問に対してこんな回答が返ってくるだろう」あるいは「自分の組織のこの質問に対しての回答結果は、上司、同僚に見せるときっと問題だと言われるにちがいない」という仮説を各自が暗黙のうちにでも、もっているものだからである。だからこそ、ぜひその質問を聞いてみたいと思ったはずなのである。

したがって、こうした作業は、問題発見のプロセスとしては非常に重要なものとなる。仮説を立てるという立場からしても、この作業は研究者のような第三者が独立かつ一方的に行うべきではなく、あくまでも、グループ参加者からの自発的な提案を促し、それを生かす形で進めるべきなのである。

そこで、ここでは質問調査票を設計する際の一般的な注意事項、および一般的なテクニック、ノウハウについてまとめておくことにしよう。

(1)実態方式と常態方式

統計調査での質問票の質問文の作成にあたっては、聞き方に、実態方式と常態方式の二つの方式があり、両者の特性をふまえて質問文を作る必要がある。

1. 常態(usual status)方式とは、普段の、あるいは平常の状態を尋ねる方式である。
2. 実態(actual status)方式とは、調査時点に近い一定期間内の実態を求める方式である。

ただし、2 の実態方式における調査時期、期間は、調査結果に直接影響するので、慎重に選択する必要がある。一般には、期間を短く設定するほど正確だが、その分、時間的変動を受

けやすくなる。例えば、飲酒の量を聞く際には、休日前、年末・年始や年度末・年度始めには一般的に飲酒の機会が増えるので、その時期に、通常の平均的な飲酒量を聞くことは適切ではない。したがって、実態方式を採用する際には、調査目的に照らして、調査時期、期間を適切に設定することが必要になる。また、この実態方式をとる場合でも、例えば「この 1 年間で」というように期間を長くするほど実質的には「常態」に近くなるので、この点でも注意が必要である。

(2)コーディングと質問形式

質問の内容が決ったら、次に質問に対する回答の仕方を決めることになる。そこでまず、質問票調査の回答を統計的に処理するために必要なコーディングの作業についてまとめておこう。コーディング(coding)とは、次の作業の総称である。

1. 調査対象者の回答をいくつかのカテゴリーに分類し、各カテゴリーに一定の記号(code)を定めること。そして、
2. 個々の回答を所定のコードで記号化することである。

このうち、2 だけの場合を狭義のコーディングと呼ぶ。質問調査票の統計処理をするからには、少なくとも狭義のコーディングは欠かせないことになる。

コーディングには、プリコーディングとアフターコーディングの 2 種類の方法がある。プリコーディング(pre-coding)とは多項選択形式のことである。この形式では、質問に対し、あらかじめ回答の選択肢が用意されていて(すなわち、あらかじめ回答のカテゴリー化、記号化が行われていて)、回答者がそれらの選択肢の中から回答を選択すると事実上、コーディングが終了することになるのである。

アフターコーディング(after-coding)とは自由回答形式のことである。あらかじめ回答の選択肢を用意しないで、質問に対する回答を回答者に思い付くままに自由に回答してもらい、回答終了後に、回答のカテゴリー化(categorization)によってコーディングの作業を行うことになる。

コーディングには 2 種類の方法があるといっても、統計調査では可能な限りプリコーディングを行い、多項選択形式を採用すべきである。なぜなら、

1. 自由回答形式は調査後の処理が繁雑である。
2. 自由回答形式は回答の微妙なニュアンスを知ることができるかもしれないが、集計の作業の前にはアフターコーディングがあるために、回答の微妙なニュアンスはその段階で切り捨てられることになる。
3. 自由回答形式では、質問票の設計者の意図、期待とは全く異なった次元で回答されることがある。
4. 自由回答形式では、回答の差異が、意見・意識の差異ではなく、回答を文章にまとめる際の文章力の差異であることが多い。

これらのことは一度自分でやってみると身に染みる。自由回答形式が意外に情報量に乏しく、その割には手間ばかりがかかるというのは実感である。

この自由回答形式の欠点と関係があるのだが、多項選択方式の場合でも、選択肢の中に無造作に「その他」を含めるべきではない。「その他」は欠損値と同じ意味しかもたないことが多いのである。「その他」を入れずに済むように、徹底的に事前のヒアリング調査を行うべきである。「その他」を選択肢として含めることは、事前のヒアリング調査の不備不足を質問調査票の中で告白しているようなものである。

それでも何らかの理由でヒアリング調査が十分に行えず、質問文の回答の選択肢をあらかじめ予想できないような場合には、プリテストを行ってみるとよい。プリテスト(pre-test)とは、本調査に先立ち、少数の調査対象に対して、質問文の検討や回答の分布に見当をつけるために行う調査である。質問文の回答の選択肢をあらかじめ予想できないような場合は、プリテストでは自由回答形式で回答してもらい、出現する回答を調べて分類し、本調査の選択肢を作成する。主に、アフターコーディングつまり自由回答形式は、プリテストで本調査におけるプリコーディングのカテゴリーを設定するために用いられると考えるべきであろう。

プリコーディングで既に他項選択形式の選択肢を用意してしまっているような場合でも、プリテストは、1 度は行っておくべきである。ただし、こうした場合のプリテストは、本番同様の方式で規模を小さくして行うというものでなくてもよい。面接調査法や面前記入法の形をとりながらも、選択肢「案」についての率直な意見を聴取するという気持ちで、質問文、選択肢の洗練を行ってもよいのである。

ところで、自由回答形式は、調査とは別の目的で用いられることがある。つまり、実際の企業などでは、回答者のもっている不満のガス抜きの目的ために、自由回答形式を多用したアンケートを使用することがみられる。しかし、調査の本来の目的は事実としての統計データを得ることであり、このようなガス抜きの目的も含めて調査することは、本来の目的を損なう可能性があり、行うべきではない。

(3)選択肢回答の諸形式

プリコーディングの多項選択形式の場合でも、選択肢の中から該当する回答を選ばせる際の選ばせ方には、次のような種類がある。

1. 一つだけ選ばせる(single answer: SA)
2. いくつでも選ばせる(multiple answer: MA)
3. ある限定数以内で選ばせる(limited answer: LA)

というような種類がある。このうち、1 と 2 の統計処理は比較的自由度が大きいが、3 の形式については、単純集計のみが可能なので、統計処理という点ではあまりうまみのない形式である。

後述するようなクロス表を作成する際の容易さ、クロス表自体の簡明さと説得力を考えると、できるだけ 1 の択一方式(SA)が望ましい。さらに、質問数が 100 項目にもなるような場合には、総当りで二つの質問項目間の相関をみていくと、第 5 章第 5 節(3)でも見たように、ざっと (100×100-100)÷2＝4,950 枚 のクロス表と向き合うことになる。これほどの枚数でもメインフレームの側ではあっという間に集計してくれるが、問題はこれを見て理解する人間の側の能力の方に限界があるということである。私を含めて私の知る限り、この無謀な試みにうまく成功した人間はいない。したがって、択一方式(SA)の中でも、できる限り、Yes-No 形式のように、二者択一型の質問にすることが望ましい。というのは、第 5 章第 5 節でも見たように、2×2 クロス表ばかりであれば、各クロス表を一つの相関係数で代表させ、相関係数だけで比較することが可能になるからである。これならば、総当りでも相関関係を調べることが出来る。

複数回答方式(MA)でも、質問の中の各選択肢を形式的に、一つの Yes-No 形式の質問として扱ってクロス表を作ることは可能である。

回答選択肢のカテゴリーの設定上の注意としては、

1. あらゆる回答がいずれかのカテゴリーに分類できることにすることである。つまり、カテゴリーの全体が回答の全範囲をカバーすること。
2. 択一方式では、回答はどれか一つのカテゴリーに分類できるようにすること。つまり、各カテゴリーは互いに排反で、カテゴリー間の区別は合理的かつ明解であること。

過去に優れた調査がある場合には、できれば類似の分類カテゴリーを採用しておくと、その調査結果との比較が可能になり、調査の回答の偏り等を考える際に参考になる。

最後に技術的なことであるが、コードはどうしても避けられないような事情がない限り数字を用いること。統計パッケージ(SAS 等)とコンピュータの都合ではあるが、後々の処理が楽になり、例えば SAS ならば、PROC ステップでの制約も気にせずに済む。そして、データ・エントリーの際も、テン・キーだけで入力ができるので、エントリー・ミスの誘発を防ぐとともに、所要時間も短くて済み、外注した場合でもデータ・エントリー費用の節約につながるのである。

(4)基本特性

質問調査票には回答者の属性(性別、年齢、職種、職位など)を聞く部分を必ず設けておく。こうした基本特性(classification data)は、分析の基準となるもので、他の質問についても、基本特性との関係を分析しておくことが望ましい。この基本特性を聞く欄は調査票の冒頭に置かれることが多かったので、基本特性項目のことを、しばしば、フェース・シート(face-sheet)項目と呼ぶが、これは和製英語。現在では、末尾に置くことも多い。継続的に調査を行う場合には、早い機会にこうした属性の分類を確立、確定しておくことが、継続的なデータの集積、比較の際に基礎として必要である。

基本特性は、そのつもりになればどんどん詳しい内容の質問を作ることができるが、詳しくなくてはいけないとか、詳しければ詳しいほど良いということはない。むしろ、あまり詳しい内容の個人情報を聞くと、回答者の側に警戒心を呼び起こすことになり、回収率、有効回答率の低下を招くことになる。常識的に考えても、基本特性の項目が多くなれば、個人の特定が容易になり、匿名性が維持できなくなる。これまでの経験からすると、未婚・既婚の別や学歴を聞くことは分析上のメリットはほとんどなく、かえってこうしたデメリットの方が大きい。

例えば、未婚・既婚の別については、この属性がこれまでに分析に役立ったことはないし、この質問自体に不快感をもつ人がかなりいる。それに、未婚・既婚の他に、実際には離婚、死別、その後の再婚のケースなどもあり、あまりに選択肢が細分化しすぎる。「配偶者の有無」を聞く方法もあるが、これとて内縁の妻や夫の存在をどう扱うべきか、調査の意図とも関係してくる。結局、本書で考えているような調査では、こうした質問をすること自体に、どんな意味があるのかわからない。

また学歴の場合、特定企業の特定部門は、ほとんど同質的な学歴をもっている集団になっているので、企業と部門を特定すれば、学歴はほぼ特定できてしまうのが普通である。このことがあるために、学歴によって回答に差異が見られるようなケースでも、詳しく検討すると、実際には企業間差異、部門間差異の反映にすぎない場合がほとんどである。学歴による分析は、企業間比較、部門間比較の際には、実質的な意味をほとんどもたないといっていいだろう。

(5)質問番号の付け方について

このことについては、既に第 2 章第 4 節(3)でも触れたが、章末資料 B の質問調査票のように、これだけ質問の項目数が多くなると、変数名が互いに重複しないように配慮しながら、各質問項目に対応した変数名を考えることはたいへんな作業となる。また類似の変数名も多くなるために、ごく限られた字数(SAS では英数字で 8 文字以内)の変数名だけを見て、すぐに元の質問を特定することは至難の技となる。そこで、変数名には質問の番号をそのまま用い、変数の意味内容については変数ラベルを用いることで集計表の上で表示することの方が実用的である。変数ラベルについては、字数の制限も大幅に緩くなっているし、漢字も使えることが多いので、少ない字数で情報量の大きな変数ラベルを作ることもできる。つまり、変数名で変数の意味内容を表現するよりはずっと確実で実用的である。

ただし、変数名に質問番号をそのまま用いるといっても、第 2 章第 4 節でも述べたように、変数名の最初の 1 字は数字ではいけないというソフト上の制約があるので、質問の大分類はローマ数字にしておくと便利である。ローマ数字は I、V、X、L などのアルファベットで構成されるので、変数名の頭にもってきても、そのまま変数名として用いて差し支えない。例えば、「質問 II の 1」「質問 V の 5」などはそれぞれ「II1」「V5」と質問番号をそのまま変数名にできるのである。

4. 質問票調査

(1)調査対象となるべき組織単位の設定と抽出

「組織活性化のための従業員意識調査」では、既にヒアリング調査の段階で設定しておいた組織単位を対象として、その正社員の全数調査をすることを原則としている。全数調査にする理由は、次のようなものである。

1. 各企業は規模で大きな違いがあるので、正確には、各企業の組織単位を母集団として、複数母集団間の比較調査を行う。この調査で想定しているような、各企業 100 人程度の大きさの母集団については、第 1 章第 7 節でも述べたように、標本誤差の目標精度を 5% 程度に抑えるために必要な標本の大きさは母集団の大きさの 80%にもなり、ほぼ全数調査にならざるを得ない。
2. これまでに企業内での標本調査を行った経験からすると、10 人中 1 人か 2 人にしか質問票が当らない中で、「無作為抽出によって標本抽出を行った」と説明しても、調査対象に選ばれた側で、自分が選ばれた理由について勘ぐりたくなるのが人情であり、このことは、企業のおかれた状況によっては非標本誤差の増大をともなう調査の質の低下(回収率の低下、回答の偏り、回答者の偏り)を招く大きな要因となってしまう。

以上のような調査対象の設定意図を説明、確認した上で、調査の仕方・手順については、章末資料 C「現地調査の手引」という説明文書をもとにして事前に口頭で説明し、確認しておく。この「現地調査の手引」は、各社における調査手順を説明したもので、各社における調査は、次の 2 段階に分けて行うことになる。

1. 第 1 段階: 調査対象となるべき組織単位の設定の確認。
2. 第 2 段階: 第 1 段階で抽出した組織単位について、それを構成する正社員全員を対象とした質問票調査。

調査の進め方としては、まず第 1 段階として、各社単位に「配布・回収状況調査票」に記入してもらった上で、回収し、この段階で調査対象を最終的に確定する。なお、この調査対象の設定は慎重に進める必要がある。調査の質を決定するものだからである。したがって、念には念を入れて、企業側に疑問点等があれば、いつでも、気軽に問い合わせが出来るように配慮しておくことが重要である。企業側にとっては小さな疑問点でも、第 2 段階の質問調査票の配布前に解決しておかないと、取り返しのつかないことになる可能性がおうおうにしてあるものである。例えば、多くの場合問題はないのだが、組織単位が複数の「職種区分」（配布・回収状況調査票を参照）にまたがらないように設定されていることを確認することも必要である。つまり、「1. 事務・スタッフ」「2. 技術・製造」「3. 研究・開発」「4. 販売・営業」のどれかに収まることを確認するのである。

そして、このときに、経験的には調査票の回収部数が全体で 400～500 部もあれば、標本誤差を抑えることが出来、統計処理上、かなり見栄えのする結果が得られるということ、したがって、この調査での問題は回収部数よりも、むしろ回収率の方であり、非標本誤差を抑えるために、回収率の目標は 90%程度としておきたいということを明確に打ち出し、組織単位を設定する際には、できるだけ高い回収率を維持したままで、必要な回収部数が得られるような配慮を強調しておくことを忘れないように。回収率は質問票調査の質と信頼性を決定するのである。

以上のことを注意した上で、調査対象となる組織単位をいくつか設定してもらい、「配布・回収状況調査票」のコピーを回収し、確認して、この段階でまめにきちんとした管理をしておくことが重要である。

(2)質問票調査と並行作業

第 2 段階は、第 1 段階で既に設定してある組織単位について、それを構成する正社員全員を対象とした質問票調査である。配布から回収までの期間が 1 週間というのは短いと感じられるかもしれないが、経験的には、回収のピークは配布直後と提出期限前後の 2 回で、この間の期間は、それがかなり長期に設定してあっても、ほとんど回収がないものである。また実際には、提出期限後も、回収打ち切りまで 1 週間程度の余裕をもたせているが、この期間に回収できる部数も経験的にはそれほど多いものではない。1 週間でできるだけかき集めるのが一番効率的で楽な方法だということを理解・納得してもらうしかない。

また質問調査票のコード・ボックスの記入の仕方は十分に説明しておく必要がある。質問調査票を配布する際には、あとでわからなくならないように、実際に配布した部数を配布・回収状況調査票の「配布部数」の「実際」の欄に記入してもらう。質問調査票を回収した質問調査票は組織単位別に分類し、組織単位ごとに質問調査票の 2 ページ目のコード・ボックスの後 2 桁に、一連番号を記入してもらう。そして、実際に回収した質問調査票の部数を配布・回収状況調査票の「回収部数」欄に記入してもらい、回収率を正確に把握するのである。

実際に回収を打ち切るのは、公式の締め切りのほぼ 1 週間後である。やむをえない場合やその後回収できた分はできるだけ生かすが、実際に集計作業が始まってしまえば、だらだらと追加するわけにはいかないので、回収打ち切りは明言しておくこと。

また章末資料 B にもあるように、質問調査票の表紙には、集計分析者を明らかにし、調査票が個票のまま利用されることのないことを保証した上で、その集計分析者の連絡先を明確にしてあるので、その対応も怠らないこと。これは、例年、せいぜい 1 件問い合わせがあるかどうかではあるが、集計分析者を確認できるという安心感が、調査の質を向上させるのである。

(3)入力フォーマット

調査票の設計がまずいと、コーディングをしなくてはならないようなはめに陥ることになるのであるが、「組織活性化のための従業員意識調査」では、調査票の設計の段階でプリコーディングを入念に行い、データもすべて数字データにしてあるので、入力は調査票を見ながら直接行うことが出来る。しかし、章末資料 D にあるように、入力フォーマットだけは、疑問の余地のないような正確に定義しておくべきである。入力フォーマットは、調査票の回収が終る前に(実際には回収期間は 1 週間しかないので、質問調査票の完成直後に)、作成しておくことが望ましい。また、SAS で集計する場合には、欠損値はピリオド "." で入力するように確認しておく。データ・エントリーを業者に外注する場合には、データは磁気テープ、いわゆる MT に入れられて納入されることが多いので、磁気テープに書き込む形式については、業者とも相談して、あらかじめ決めておく必要がある。

(4)配布・回収状況一覧表

調査票の回収が終わると、データ・エントリーと並行して、章末資料 E のような「配布・回収状況一覧表」を作成する。ここで重要なのは組織単位ごとの回収数を正確に確定しておくことである。この数字は、単純集計を行って、データとプログラムのチェックを行う際に参照すべき基礎的数字なので、必ず集計作業の始まる前に作成、確定しておくこと。

以上が実際の調査手順である。ここまでくれば、あとは実際の集計作業に入るだけであり、本書の第 2 章で扱っている単純集計から順に分析を進めていけばよいのである。

演習問題

6.1　基礎調査票　『会社四季報』『会社情報』など公にされている最新版の資料を使って、自分の興味のある会社 1 社について、章末資料 A の基礎調査票の会社概要調査票に記入し、できるだけ完成させてみよう。それから何がわかるか。また、そのような公にされている資料ではわからなかった情報の価値について考えてみよ。

6.2　プリテスト　被験者になったつもりで、章末資料 B の質問調査票に実際に回答してみよ。その際、次の点に注意しながら回答せよ。

1. 全問を回答するのに要した時間を計測せよ。
2. 答えられなかった質問、答えにくかった質問について、どのように質問文を変更、改善すれば答えられるようになるか、具体的に提案せよ。

資料 A. 基礎調査票と調査の基本設計

1. 19XX 年度第 1 回合宿について

第 1 回合宿の第 1 目標は、今後のグループ研究の討議の際の基礎となるキーワード、コンセプトの認識をできるだけ統一しておくこと、もしくは認識の違いを明確にしておくこと(＝親睦)にあります。

(1)日程

6 月 14 日 14:00～18:00 グループ討議
9:00～12:00 グループ討議
15 日 13:00～17:30 グループ討議
19:00～ 翌日の中間発表の準備作業
16 日 9:00～11:30 中間発表(問題意識のまとめ、できれば研究テーマ)

(2)グループ討議では 1 人 1 時間程度で次の内容について発表していただきます。

a. 会社の概要紹介(所定の用紙に則った形で要点だけ)
- 所定の基礎調査票に記入 してきてください。組織図については、原図またはコピーを貼り付けてもかまいません。基礎調査票にある項目は例年、合宿の際に、グループ研究で話題になった事柄、あるいは最初に明らかにしておかなかったために、議論の混乱の原因となったものを公約数的にまとめたものです。
- わからない部分は有価証券報告書等まで参照するつもりで、徹底的に調べあげてください。
- 組織図は、質問票調査の調査対象となる組織単位を選ぶためのものです。各自が想定している調査対象にしたい組織単位を次ページを参照しながら表示してください。
- 所定の用紙の他に 補足資料等を付け加えても構いません。会社のパンフレット等を添付した方が発表がスムーズにいくようです。

b. 会社の特長と問題点(会社の活性化に対する自分の問題意識を中心にして)
- 所定の用紙はありませんが、A4 横書きで箇条書形式にしてまとめてください。その際には、全社的なもの、抽象的なもの、というよりは、想定している調査対象にしたい組織単位を十分に意識の上、できるだけ具体的にまとめてください。
- まとめる際のポイントには、次のようなものが考えられます。(単なるヒント)
  - 会社の特徴についてのトップの発言
  - 世間から見た会社イメージと内部から見た実像
  - 活性化の必要性を感じる、とき、ところ
  - 自分にとっての活性化、会社にとっての活性化

  以上のようなポイントに留意しながら、身近なエピソードも交えて、活性化に対する自分の考え、問題意識をまとめてきてください。

(3)その他

- 例年、合宿では時間が大幅に不足します。
- 発表をスピーディに行うために、(2)の内容について、レジュメ(発表要旨)を作っていただきます。当日コピーを配布してください。
- 当日、合宿先でのコピーは混雑する上に、料金も高いので、事前にコピーしておいてください。コピー部数は、(グループ・メンバー数＋1)部をお願いします。

限られた時間内に参加者相互の理解を深めるために、是非ご協力下さい。

2. 質問票調査について

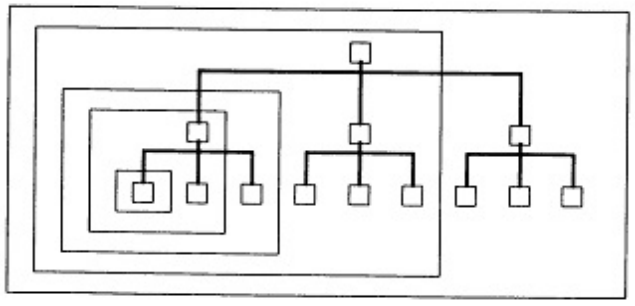
(1)調査スケジュールの見通し

前年度を参考にすると

7 月	2 日	質問案の検討・討議
	16 日	質問案の検討・討議
8 月	27 日	質問調査票完成
	28 日	調査票配布
9 月	3 日	調査票点検・整理(調査票記載の回収期限は 2 日)
9 月	24 日	集計表報告
10 月	8 日	集計表をもとにした企業間比較
10 月	18〜20 日	合宿: 集計表をもとにした職場間比較

(2)調査対象

本社の中の人員規模 50 人以上 100 人未満のホワイトカラーの組織単位を一つまたは複数選び、その組織単位の正社員の全数調査を原則とします。ここで、「組織単位」とは組織図上で同一の上司を持つ職場もしくは職場の集合です。例えば、次の図で、細線で囲まれているような各レベルの部分です。



《全数調査にする理由》

1. 各企業は規模で大きな違いがあるので、正確には、各企業の組織単位を母集団として、複数母集団間の比較調査を行う。今回想定しているような、各企業 50〜100 人程度の大きさの母集団については、標本誤差の目標精度を５％程度に抑えるために必要な標本の大きさは母集団の大きさの 80〜90%にもなり、ほぼ全数調査にならざるを得ない。
2. これまでに企業内での標本調査を行った経験からすると、10 人に 1 人か 2 人にしか質問票が当らない中で、「無作為抽出によって標本抽出を行った」と説明しても、調査対象に選ばれた側で、自分が選ばれた理由について勘ぐりたくなるのが人情で

あり、このことは、企業のおかれた状況によっては調査の質の低下(回収率の低下、回答の偏り、回答者の偏り)を招く大きな要因と実質的になってしまうため。

(3)集計表

1. 配布・回収状況一覧表
2. 単純集計
3. 個人属性によるクロス表
4. 企業間比較のためのクロス表
5. 組織単位間比較のためのクロス表
6. Yes-No 形式の質問間の相関係数行列

基礎調査票①　会社概要調査票(1/2)

氏名　　19XX 年　　月　　日現在（金額の単位は百万円）

会社名	本店所在地
英語名	資本金

沿革
　設立年月日　西暦　　年　　月　　日

主要製品・サービス	売上高比率	主要株主(持株比率順)	持株比率
1.		1.	
2.		2.	
3.		3.	
4.		4.	
5.		5.	

	19XX 年度	19XX 年度	財務状況についてのコメント
(自己)資本			
負債			
売上高			
売上総利益			
営業利益			
営業外収益			
営業外費用			
経常利益			

社是・社訓

19　　年制定

基礎調査票①　会社概要調査票(2/2)

関係会社総数_____社 …… そのうち　子会社(持株比率 50%超)数　_____社 子会社を除いた関連会社数　_____社					
重要な関係会社名	資本金	持株比率	事業内容	設立年	所在地
1.					
2.					
3.					
4.					
5.					

	正社員	(うち出向中)	準社員(パート等)	備考
従業員数	人	(　　人)	人	
平均年齢	歳	(　　歳)	歳	
女子比率	%	(　　%)	%	

	正式役職名	資格・等級	最短到達年齢	平均到達年齢
部長クラス				
課長クラス				
係長クラス				
一般				

基礎調査票②　会社制度・施策調査票

次の制度・施策が既に実行されていれば、その開始時期とその具体的名称を、予定されていれば、その予定開始時期とその予定名称とを記入して下さい。(（　）内は、該当するものを一つ選んで下さい。)

１．異部門間でのジョブ・ローテーション（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

２．引っ越しを必要とするような距離での転勤（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

３．年功序列ではない能力主義等を看板にした人事評価制度（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

４．自己申告制度（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

５．必要な技能・スキルを身につけるような社内教育制度（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

６．専門職制度（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

７．財形制度・持ち家制度（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

８．保養施設（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

９．持ち株制度（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

10. 文化・体育・レクリエーション施策（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

11. 海外留学制度（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

12. 小集団活動・改善提案制度（既に実行・予定・未定）
　　開始時期　　年　月
　　具体的名称

13. 労働組合
　　開始時期　　年　月
　　具体的名称

基礎調査票③　組織図

（大まかなもので結構ですが、御自分の所属部署を明示の上、調査してみたいと思っている組織単位を２の(2)を参考にして、いくつか設定、図示してみてください。その際、できれば組織単位ごとの大まかな人数も記入してください。）

資料 B. 質問調査票

グループの皆さんへ

先日のグループ研究での討議を基に、質問調査票の原案を作成しました。ご検討ください。

1. 質問調査票の内容については、細かいことであっても、修正を要する箇所について、8 月 22 日(木曜日)までに 、電話で連絡してください。［連絡先］TEL 000-000-0000 高橋伸夫 宛
2. 調査の仕方・手順については、「C. 現地調査の手引」という説明文書を今回、添付いたしておりますので、その中の第 1 段階については、各自実施の上、「C. 現地調査の手引」の中にある「配布・回収状況調査票」の太枠内に記入し、そのコピーを 8 月 27 日(火曜日)に 、日本生産性本部経営アカデミーまでお持ちください。

完成版は 8 月 27 日(火曜日)に、私が日本生産性本部経営アカデミーまで持参し、お渡しするつもりです。その際に、口頭で調査の仕方についてもご説明するつもりでおりますが、事前に、今回添付している「C. 現地調査の手引」という説明文書を御熟読いただければ、調査の仕方はご理解いただけるものと思います。ご不明の点は、その際にまとめてご質問いただいても結構ですが、疑問点等がありましたら、早い機会にお電話いただければ幸いです。

なお今後の日程は「C. 現地調査の手引」の中に「現地調査・集計スケジュール」がありますのでご参照ください。

19XX 年 8 月

東京大学 教養学部
高橋伸夫

組織活性化のための従業員意識調査

本調査票は統計的に集計し、個票のまま外部へは公表致しません

調査ご協力のお願い

拝啓、時下益々御清栄のこととお慶び申し上げます。

さて、近年、従業員の会社に対する意識に変化が見られると言われ、私ども日本生産性本部経営アカデミーの参加者の間でも、このたび、従業員の意識の多様化と組織の活性化の関係について学術的に調べるために、新時代における人間能力と組織開発についての調査を企画、実施することにいたしました。結果が得られれば、学術的にも貴重な資料になると考えられます。

この調査の結果は純粋な統計数値としてのみ集計され、個々の調査票のままで外部へ公表されることはありませんので、ご迷惑をおかけすることは決してございません。業務にご多忙のこととは存じますが、是非ご協力下さるようにお願い申し上げます。

敬具

企画実施　(財)日本生産性本部 経営アカデミー
『人間能力と組織開発』コース C グループ
代表幹事 A 電鉄株式会社　　□□□□
事務局　(財)日本生産性本部 □□□□
00-0000-0000

集計分析　東京大学教養学部 助教授　高橋伸夫

記入にあたってのお願い事項

○全質問に対する回答をお願いいたします。特に質問 I は統計分析のために重要な質問ですので、必ずお答え下さい。また、どうしても回答したくない質問にぶつかったときにはお手数でも下記の連絡先まで連絡の上、ご相談下さい。この調査に関してのご質問等につきましても、お気軽にお尋ねくだされば幸いです。

［連絡先］東京大学教養学部 助教授 高橋伸夫　電話 00-0000-0000
（貴社名、ご氏名等を名乗っていただく必要は全くありませんので御安心ください。）

○この調査票は、19XX 年 9 月 2 日（月曜日）まで に貴社の担当者

氏名		所属・役職	

までご提出ください。

組織活性化のための従業員意識調査

（提出期限　19XX 年 9 月 2 日 月曜日）

Ⅰ．あなたの 9 月 1 日現在の年齢等をお教え下さい。選択肢の中から該当するものを一つ〇で囲んで下さい。

性別	1. 男 2. 女	職種	1. 事務・スタッフ 2. 技術・製造 3. 研究・開発 4. 販売・営業
年齢	1. 20～24 歳 2. 25～29 歳 3. 30～34 歳 4. 35～39 歳 5. 40～44 歳 6. 45～49 歳 7. 50～54 歳 8. 55～60 歳	職位	1. 部長クラス 2. 課長クラス 3. 係長・主任クラス 4. 一般

Ⅱ．あなたの仕事への意欲・やりがいに関する次の各質問について、Yes (該当する)、No (違う) のどちらか近いと思われる方一つを選んで〇をつけていって下さい。

1. 自分の仕事については、人並の仕事のやり方では満足せずに、常に問題意識をもって取り組み、改善するように心がけている。　1.Yes　2. No
2. 従来のやり方・先例にこだわらずに仕事をしている。　1.Yes　2. No
3. 必要な仕事はセクションにとらわれずに積極的に行っている。　1.Yes　2. No
4. 自分の実力は他の会社でも充分通用すると思う。　1.Yes　2. No
5. 上司がこうだと言えば、自分に反対意見があっても素直に従う。　1.Yes　2. No
6. トップの経営方針と自分の仕事との関係を考えながら仕事をしている。　1.Yes　2. No
7. 上司からの権限委譲がなされている。　1.Yes　2. No
8. 自分の意見が尊重されていると思う。　1.Yes　2. No
9. 21 世紀の自分の会社のあるべき姿を認識している。　1.Yes　2. No
10. 良いと思ったことは、周囲を説得する自信がある。　1.Yes　2. No
11. 難しい仕事や慣れない仕事にはあまり手を出したくない。　1.Yes　2. No

12. 今の職場では、業績を残すよりも、大きな問題やミスを起こさないようにしたい。 1.Yes 2. No
13. できれば人よりも早く昇進したいと思っている。 1.Yes 2. No
14. ポスト不足の実感から、出世を半ばあきらめている。 1.Yes 2. No
15. 現在の職務に満足感を感じる。 1.Yes 2. No

Ⅲ. <u>あなたの会社・職場の雰囲気</u>に関する次の各質問について、Yes (該当する)、No (違う) のどちらか近いと思われる方一つを選んで〇をつけていって下さい。

1. 仕事上の個人の業績、貢献の高い人は、昇進、昇格あるいは昇給などを確実に果たしている。 1.Yes 2. No
2. 失敗をしながらでも業績を挙げていくよりは、失敗をしないで過ごした方が評価されると思う。 1.Yes 2. No
3. 新しい仕事にチャレンジしていこうという雰囲気がある。 1.Yes 2. No
4. 個性を発揮するよりも、組織風土に染まることを求められる。 1.Yes 2. No
5. 目標達成に向けて競争的雰囲気がある。 1.Yes 2. No
6. 斬新な発想や創意工夫を生かそうという雰囲気がある。 1.Yes 2. No
7. 職場の雰囲気を「ぬるま湯」だと感じることがある。 1.Yes 2. No
8. 出る杭は打たれる風土である。 1.Yes 2. No
9. 金太郎飴的な人間の集まりである。 1.Yes 2. No
10. 安全性が最優先される雰囲気がある。 1.Yes 2. No
11. 有給休暇は自由にとれる。 1.Yes 2. No
12. 社内で、県人会、同窓会などのインフォーマルな組織の活動が活発である。 1.Yes 2. No
13. 評価は、業績、貢献度でというより、上司の好き嫌いで判断される傾向がある。 1.Yes 2. No
14. 20 代の若手社員の能力や個性が職場で生かされていると思う。 1.Yes 2. No
15. 外部からの役員就任にもこだわらない雰囲気がある。 1.Yes 2. No

Ⅳ．あなたの会社の意思決定スタイルとコミュニケーション に関する次の各質問について、　Yes (該当する)、No (違う) のどちらか近いと思われる方一つを選んで○をつけていって下さい。

1. 意見が異なる場合、対立を避け、できるだけ歩み寄ろうとする。 1.Yes 2. No
2. 本音の議論は、就業時間中というより、社外で行うことが多い。 1.Yes 2. No
3. 上下関係を意識せずに、自由闊達に議論のできる雰囲気がある。 1.Yes 2. No
4. 互いに納得のいくまで議論ができる雰囲気がある。 1.Yes 2. No
5. 議論を議論のままで終らせずに、委員会の組織化や根回し等、実行に向け乗り出すことが多い。 1.Yes 2. No
6. 仲間内でも、会社全体の長期的経営方針を議論することがある。 1.Yes 2. No
7. 指示が出されても、やり過ごしているうちに、立ち消えになることがある。 1.Yes 2. No
8. 問題点はわかっていても、誰も問題提起せず、見過ごされることがある。 1.Yes 2. No
9. 文書・資料等がいつでも引き出せるように整理がなされている。 1.Yes 2. No
10. 最近は資料作りが多すぎると思う。 1.Yes 2. No
11. 課と課の間の情報交換がスムーズに行われている。 1.Yes 2. No
12. 社員間の横のコミュニケーションが充分なされていると感じる。 1.Yes 2. No
13. 現場からの情報が、充分に本社・本部に伝わっていると思う。 1.Yes 2. No
14. 本社・本部から、現場に必要な情報が適切に伝達されていると思う。 1.Yes 2. No
15. 複数の系統から指示を受けることがある。 1.Yes 2. No

Ⅴ．あなたと会社との関わりに関する次の各質問について、Yes (該当する)、No (違う) のどちらか近いと思われる方一つを選んで○をつけていって下さい。

1. 会社に対して忠誠心をもっている。 1.Yes 2. No
2. 終身この会社で仕事をしていきたいと思う。 1.Yes 2. No
3. チャンスがあれば転職したいと思う。 1.Yes 2. No
4. 自分自身の 10 年後の具体的な将来設計が、なかなか立てられない。 1.Yes 2. No
5. 会社の進んでいる方向は、自分の望む会社の将来像と結びついている。 1.Yes 2. No
6. 自分は会社の経営方針を充分理解していると思う。 1.Yes 2. No
7. 役員層に人間としての魅力を感じる人がいる。 1.Yes 2. No

8. 経営理念を自分の具体的行動計画にまでブレークダウンしている。 1.Yes 2. No
9. 職場の目標がはっきりわかっている。 1.Yes 2. No
10. 今の職場で自分のやるべきこと、自分の役割を自覚している。 1.Yes 2. No
11. 自分なりの行動理念をもって行動している。 1.Yes 2. No
12. 人生にゆとりがあり、楽しい。 1.Yes 2. No
13. 会社は社会奉仕活動に理解があると思う。 1.Yes 2. No
14. 自分としては、職場よりも家庭を重視している。 1.Yes 2. No
15. 自分の仕事上の成果に対する上司の評価は、適切で公平だと思う。 1.Yes 2. No

Ⅵ. あなたの会社・職場に対する評価に関する次の各質問について、Yes (該当する)、No (違う) のどちらか近いと思われる方一つを選んで○をつけていって下さい。

1. タテ、ヨコの縛りがゆるく、弾力的に仕事ができる。 1.Yes 2. No
2. 人材育成のための教育研修制度(off JT)がしっかりしていると思う。 1.Yes 2. No
3. 人材育成は職場まかせになっていると思う。 1.Yes 2. No
4. いわゆる敗者復活の人事が行われたことがある。 1.Yes 2. No
5. 他の企業に就職した同期と比較して、わが社の出世スピードは遅い方だと思う。 1.Yes 2. No
6. 関連企業等への出向を積極的に行う方がよい。 1.Yes 2. No
7. 福利厚生面は充実している。 1.Yes 2. No
8. 役員層と社員の間にもっと対話が必要だと思う。 1.Yes 2. No
9. 会社の方針が明確に示されていると思う。 1.Yes 2. No
10. 上司から仕事上の目標をはっきり示されている。 1.Yes 2. No
11. 基準・規程・マニュアルがあるにもかかわらず、有効には利用されていない。 1.Yes 2. No
12. 改善に向け、小集団活動 (例えば QC 活動) が効果的に行なわれている。 1.Yes 2. No
13. 長期的展望に立った仕事というより、短期的な数字合わせになりがちである。 1.Yes 2. No
14. 21 世紀に向けた企業イメージを積極的に社外にアピールしていくべきだと思う。 1.Yes 2. No
15. 会社に来ること自体が仕事のようになっている人がいる。 1.Yes 2. No

ご協力ありがとうございました。

資料 C. 現地調査の手引

1. はじめに

この「現地調査の手引」は、各社における調査手順を説明したものです。各社における調査は、次の 2 段階に分けて行います。

1. 第 1 段階: 調査対象となるべき組織単位の設定の確認。
2. 第 2 段階: 第 1 段階で抽出した組織単位について、それを構成する正社員全員を対象とした質問票調査。

調査の進め方としては、まず第 1 段階として、各社単位に「配布・回収状況調査票」に回答の上、8 月 27 日(火曜日)に 、日本生産性本部経営アカデミーまでお持ちください。

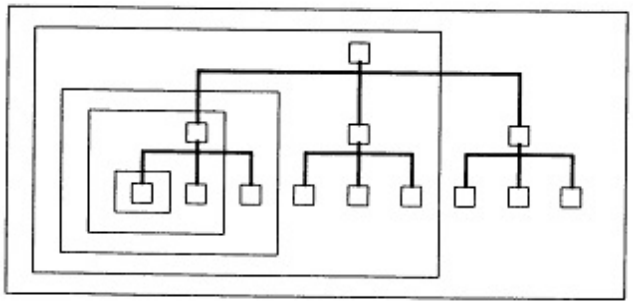
それでは、早速、「2. 第 1 段階調査の手順」にある説明を御熟読の上、添付してある「配布・回収状況調査票」の太枠内に記入しながら、第 1 段階調査を進めてください。

なお疑問点等がありましたら、いつでも、お気軽にお問い合わせください。疑問点は第 2 段階の質問調査票の配布前に解決しておかないと、取り返しのつかないことになる可能性がありますので、どんな小さな疑問点でも、お気軽にお問い合わせください。［連絡先］TEL 000-000-0000 高橋伸夫

2. 第 1 段階調査の手順

既に、どこを調査対象としたいのかについては、おおまかにご説明いただいておりますが、質問票調査を目前にして、調査対象を明確に定義し、その人数的な大きさも確定しておく必要がありますので、これから述べるような考え方に沿って、組織単位を実際に設定してください。

第 2 段階調査の質問票調査では正社員の人員規模 30～50 人程度のホワイト・カラーの「組織単位」をいくつか選び、その「組織単位」については、その正社員の全数調査を原則とします。ここで、「組織単位」とは組織図上で同一の上司を持つ職場もしくは職場の集合です。例えば、次のような組織図では、細線で囲まれているような各レベルの部分が組織単位の候補として考えられます。



この候補となるレベルの中から、どのレベルに組織単位を設定するかは、

1. 組織単位が複数の「職種区分」(配布・回収状況調査票を参照してください)にまたがらないように注意しながら、
2. 正社員の人員規模が 30～50 人程度になるように、

という二つの基準を一応の目安にして決めてください。なお、基準 1 を満たすようにすると、どうしても 30 人を下回ってしまう場合には、基準 1 を優先し、基準 2 の方はかなり緩めて

適用していただいて結構です。ただし、そのような少ない人数の組織単位は、通常の大きさの組織単位との比較が難しいので、注意してください。また、組織単位には、別会社の形をとっているものを選んでいただいてもかまいません。

経験的には調査票の回収部数が全体で 400～500 部もあれば、統計処理上、かなり見栄えのする結果が得られます。したがって、問題は回収部数よりも、むしろ回収率の方になります。これまでの経験から、回収率の目標は 90%以上としておきたいものです。この目標回収率は明確な基準には致しませんが、組織単位を設定する際には、できるだけ高い回収率を維持したままで、必要な回収部数が得られるように配慮してください。回収率は質問票調査の質と信頼性を決定します。

以上のことに留意された上で、調査対象となる組織単位をこの基準 1、2 でいくつか設定した上で、添付の「配布・回収状況調査票」の太枠内に記入して、8 月 27 日(火曜日)に 、そのコピーを日本生産性本部経営アカデミーまでお持ちください。

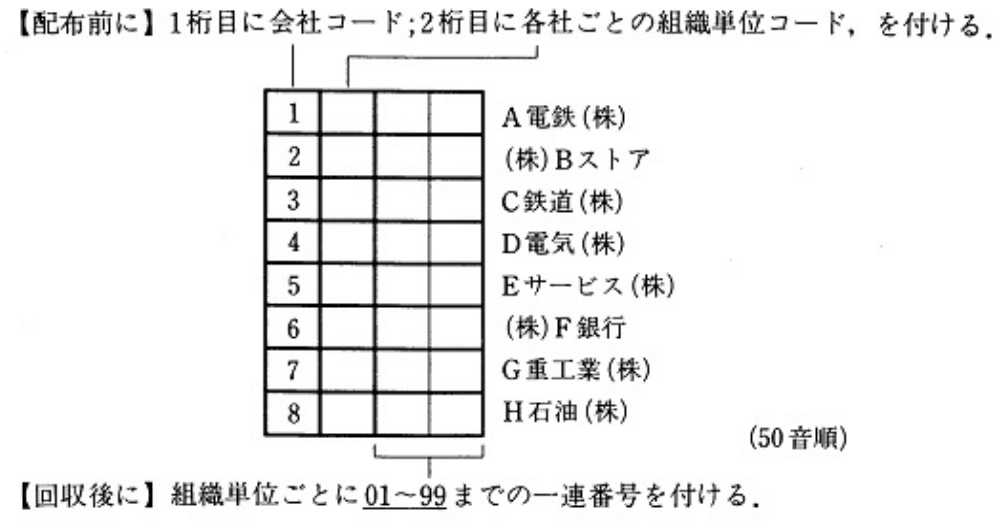
3. 第 2 段階調査の手順

第 2 段階は、第 1 段階で既に設定してある組織単位について、それを構成する正社員全員を対象とした質問票調査です。8 月 28 日(水曜日)に一斉に配布していただき、質問調査票に記載されている提出期限、9 月 2 日(月曜日)までに回収してください。配布から回収までの期間が 1 週間というのは短いと感じられるかもしれませんが、経験的には、回収のピークは配布直後と提出期限前後の 2 回で、この間の期間はほとんど回収がないもののようです。また、実際には提出期限後も、回収打ち切りまで 1 週間程度の余裕をもたせる予定ですが、この期間に回収できる部数も経験的にはそれほど多いものではありません。ここはひとつ覚悟を決めて、1 週間でできるだけかき集めるのが一番効率的で楽な方法だと思います。第 2 段階の手順は以下の通りです。

(1)質問調査票の用意をしてください。

1. 質問調査票の 1 ページ目にある「担当者」欄に記入し、2 ページ目にあるコード・ボックスの 1 桁目に会社コードを記入した上で(下の注の説明を参照のこと)、質問調査票を必要部数だけコピーしておきます。
2. 配布する前に、配布・回収状況調査票で既に 1, 2, 3, ... と順に割り付けてある組織単位コードを各質問調査票の 2 ページ目にあるコード・ボックスの 2 桁目に記入しておいてください。

注) 質問調査票の 2 ページ目にあるコード・ボックスの記入の仕方



(2)8 月 28 日(水曜日)に一斉に質問調査票を配布してください。

その際、実際に配布した部数を配布・回収状況調査票の「配布部数」の「実際」の欄に記入してください。

(3)質問調査票を回収し、次の作業を行った後で、9 月 3 日(火曜日)の日本生産性本部経営アカデミー(グループ研究)にお持ちください。

1. 回収した質問調査票の 1 ページ目は不要ですので、取り除いてください。
2. 回収した質問調査票は組織単位別に分類し、組織単位ごとに質問調査票の 2 ページ目のコード・ボックスの後 2 桁に、01, 02. 03, ...と順に一連番号を記入してください。
3. 実際に回収した質問調査票の部数を配布・回収状況調査票の「回収部数」欄に記入してください。

(4)やむをえない場合やその後回収できた分は、回収でき次第、(3)と同様の作業を行った上で、日本生産性本部経営アカデミーまでお持ちいただくか、送ってください。9 月 9 日(月曜日)に回収を打ち切ります。

その際、配布・回収状況調査票は、回収した質問調査票の発送時に、その都度、ファックスで下記の東京大学宛に送ってください。私に対する調査票の「発送通知」も兼ねております。

それ以後の到着分は、集計作業が始まってしまいますので、集計に用いることができません。間に合わせるために、部数が少なければ、ファックスで経営アカデミーまで送っていただいてもかまいません。

回収した質問調査票の送付先：

〒150 東京都渋谷区渋谷 3-1-1 (財)日本生産性本部経営アカデミー □□□□ 氏
FAX 00-0000-0000

配布・回収状況調査票の送付先：

〒153 東京都目黒区駒場 3-8-1 東京大学教養学部社会科学科 高橋伸夫
FAX 00-0000-0000

資料 D. 入力フォーマット

- データはすべて数字データです。
- データは各質問調査票がパンチ・カード 2 枚に相当するように、2 行にわたって、指定するカラムに入力してください。(質問調査票 1 部が 2 レコードになる。)
- 欠損値はピリオド(.)で入力してください。
- データは次の仕様でＭＴに入れてください。
    1. ノンラベル
    2. レコード長 80 バイト(固定長)
    3. ブロック化係数 40(ブロック・サイズ 3200 バイト)
    4. EBCDIC コード(ただし、データはすべて数字データ)

入力フォーマット

【入力フォーマット１／２】

ｶﾗﾑ	変数名	質問No.	変数ラベル	変数値ラベル
1	KC KCODE		会社コード	1.A ER 2.B STORE 3.C RAIL 4.D EP 5.E SER 6.F BANK
2	SC		組織単位コード	(SCODE=KC*10+SC) 7.G HI 8.H OIL
3	IC		個人コード	【十の位】
4				【一の位】
5	I1	Ⅰ.	性別 SEX	1.MALE 2.FEMALE 7.50-54 8.55-60
6	I2		年齢 AGE	1.20-24 2.25-29 3.30-34 4.35-39 5.40-44 6.45-49
7	I3		職種 OCCUPATION	1.STAFF 2.E&P 3.R&D 4.SALES
8	I4		職位 RANK	1.GEN 2.MGR 3.HEAD 4.OTHERS
9	II1	Ⅱ.	１．常に仕事改善心掛	1.YES 2.NO
10	II2		２．先例に拘らず仕事	1.YES 2.NO
11	II3		３．境界に拘らず仕事	1.YES 2.NO
12	II4		４．自分他社でも通用	1.YES 2.NO
13	II5		５．上司に素直に従う	1.YES 2.NO
14	II6		６．経営方針考え仕事	1.YES 2.NO
15	II7		７．上司から権限委譲	1.YES 2.NO
16	II8		８．自分の意見が尊重	1.YES 2.NO
17	II9		９．自社の将来像認識	1.YES 2.NO
18	II10		10．良い事説得に自信	1.YES 2.NO
19	II11		11．難しい仕事尻込み	1.YES 2.NO
20	II12		12．業績よりミス回避	1.YES 2.NO
21	II13		13．人よりも早く昇進	1.YES 2.NO
22	II14		14．出世を半ば諦める	1.YES 2.NO
23	II15		15．現在職務に満足感	1.YES 2.NO
24	III1	Ⅲ.	１．高業績の人は昇進	1.YES 2.NO
25	III2		２．減点主義的な人事	1.YES 2.NO
26	III3		３．新しい仕事に挑戦	1.YES 2.NO
27	III4		４．風土に染められる	1.YES 2.NO
28	III5		５．目標に向け競争的	1.YES 2.NO
29	III6		６．創意生かす雰囲気	1.YES 2.NO
30	III7		７．職場にぬるま湯感	1.YES 2.NO
31	III8		８．出る杭は打たれる	1.YES 2.NO
32	III9		９．金太郎飴的な集団	1.YES 2.NO
33	III10		10．安全性が最優先に	1.YES 2.NO
34	III11		11．有給休暇は自由に	1.YES 2.NO
35	III12		12．非公式組織が活発	1.YES 2.NO
36	III13		13．上司の好嫌で評価	1.YES 2.NO
37	III14		14．若手を生かす職場	1.YES 2.NO
38	III15		15．外部役員に拘らず	1.YES 2.NO

【入力フォーマット２／２】

ｶﾗﾑ	変数名	質問№.	変数ラベル	変数値ラベル
1	【ブランク】			
2	【ブランク】			
3	【ブランク】			
4	【ブランク】			
5	IV1	Ⅳ.	１．対立避け歩み寄り	1.YES 2.NO
6	IV2		２．本音の議論は社外	1.YES 2.NO
7	IV3		３．自由闊達に議論可	1.YES 2.NO
8	IV4		４．納得いくまで議論	1.YES 2.NO
9	IV5		５．議論は実行に直結	1.YES 2.NO
10	IV6		６．仲間で方針を議論	1.YES 2.NO
11	IV7		７．指示やり過ごし可	1.YES 2.NO
12	IV8		８．問題見過ごしあり	1.YES 2.NO
13	IV9		９．資料の整理は充分	1.YES 2.NO
14	IV10		10．資料作り多すぎる	1.YES 2.NO
15	IV11		11．課間の情報交換良	1.YES 2.NO
16	IV12		12．社員間の伝達充分	1.YES 2.NO
17	IV13		13．現場から情報伝達	1.YES 2.NO
18	IV14		14．本社から情報伝達	1.YES 2.NO
19	IV15		15．複数系統から指示	1.YES 2.NO
20	V1	Ⅴ．	１．会社に対し忠誠心	1.YES 2.NO
21	V2		２．終身この社で仕事	1.YES 2.NO
22	V3		３．機会があれば転職	1.YES 2.NO
23	V4		４．自分の将来設計難	1.YES 2.NO
24	V5		５．望む方向に社進む	1.YES 2.NO
25	V6		６．経営方針充分理解	1.YES 2.NO
26	V7		７．役員に人間的魅力	1.YES 2.NO
27	V8		８．経営理念を具体化	1.YES 2.NO
28	V9		９．職場の目標を理解	1.YES 2.NO
29	V10		10．自分の役割を自覚	1.YES 2.NO
30	V11		11．自分なり行動理念	1.YES 2.NO
31	V12		12．人生にゆとりあり	1.YES 2.NO
32	V13		13．会社社会奉仕理解	1.YES 2.NO
33	V14		14．職場より家庭重視	1.YES 2.NO
34	V15		15．上司の評価は適切	1.YES 2.NO
35	VI1	Ⅵ.	１．緩く弾力的に仕事	1.YES 2.NO
36	VI2		２．研修制度しっかり	1.YES 2.NO
37	VI3		３．人材育成職場委せ	1.YES 2.NO
38	VI4		４．敗者復活人事あり	1.YES 2.NO
39	VI5		５．わが社の出世遅い	1.YES 2.NO
40	VI6		６．出向積極的が良い	1.YES 2.NO
41	VI7		７．福利厚生面は充実	1.YES 2.NO
42	VI8		８．役員と対話が必要	1.YES 2.NO
43	VI9		９．会社の方針は明確	1.YES 2.NO
44	VI10		10．上司から目標明示	1.YES 2.NO
45	VI11		11．基準有効利用まだ	1.YES 2.NO
46	VI12		12．小集団活動効果的	1.YES 2.NO
47	VI13		13．短期的数字合わせ	1.YES 2.NO
48	VI14		14．企業像主張すべき	1.YES 2.NO
49	VI15		15．いること仕事の人	1.YES 2.NO

資料 E. 配布・回収状況一覧表

会社 コード KCODE	会社名	組織単位 コード SCODE	組織単位名	職　種	配布数	回収数	回収率
１．	Ａ電鉄㈱	11	運輸部	事務・ｽﾀｯﾌ	57	53	
	［A ER］	12	自動車事業部	事務・ｽﾀｯﾌ	67	52	
		13	賃貸事業部	事務・ｽﾀｯﾌ	41	35	
					165	140	84.8
２．	㈱Ｂストア	21	人事部	事務・ｽﾀｯﾌ	60	56	
	［B STORE］	22	能力開発部	事務・ｽﾀｯﾌ	30	20	
		23	食品事業部	販売・営業	30	26	
					120	102	85.0
３．	Ｃ鉄道㈱	31	総務部	事務・ｽﾀｯﾌ	33	33	
	［C RAIL］	32	情報管理室	事務・ｽﾀｯﾌ	42	42	
		33	車両部	技術・製造	43	38	
					118	113	95.8
４．	Ｄ電気㈱	41	人事部研修室	事務・ｽﾀｯﾌ	33	33	
	［D EP］	42	工務部施設業務課	技術・製造	42	42	
		43	総務部文書課	事務・ｽﾀｯﾌ	51	46	
					126	121	96.0
５．	Ｅサービス㈱	51	総括部門	事務・ｽﾀｯﾌ	30	22	
	［E SERVICE］	52	営業部門	事務・ｽﾀｯﾌ	42	28	
		53	トラヒック情報部門	事務・ｽﾀｯﾌ	12	12	
		54	第一顧客サービス部門	事務・ｽﾀｯﾌ	16	15	
		55	第二顧客サービス部門	事務・ｽﾀｯﾌ	10	9	
		55	料金部門	事務・ｽﾀｯﾌ	29	21	
		56	設備品質管理部門	事務・ｽﾀｯﾌ	8	8	
		57	所内保全サービス部門	事務・ｽﾀｯﾌ	6	5	
		58	所外保全サービス部門	事務・ｽﾀｯﾌ	7	7	
					160	127	79.4
６．	㈱Ｆ銀行	61	秘書室	事務・ｽﾀｯﾌ	12	11	
	［F BANK］	62	事務部	事務・ｽﾀｯﾌ	25	25	
		63	業務調整部	事務・ｽﾀｯﾌ	17	16	
		64	開発企画部	事務・ｽﾀｯﾌ	23	21	
		65	営業部	事務・ｽﾀｯﾌ	13	13	
					90	86	95.6
７．	Ｇ重工業㈱	71	人事部、教育部	事務・ｽﾀｯﾌ	45	42	
	［G HI］	72	国内営業本部	事務・ｽﾀｯﾌ	44	35	
		73	広報室	事務・ｽﾀｯﾌ	32	30	
					121	107	88.4
８．	Ｈ石油㈱	81	財務部	事務・ｽﾀｯﾌ	23	23	
	［H OIL］	82	人事部	事務・ｽﾀｯﾌ	49	47	
		83	総務部	事務・ｽﾀｯﾌ	45	41	
					117	111	94.9
					1017	907	89.2

資料 F. 単純集計

組織活性化のための従業員意識調査

（提出期限　19XX 年 9 月 2 日 月曜日）

Ⅰ．あなたの 9 月 1 日現在の年齢等をお教え下さい。選択肢の中から該当するものを一つ〇で囲んで下さい。

性別	１．男	765(87.2)	職種	１．事務・スタッフ	648(74.8)
	２．女	112(12.8)		２．技術・製造	62(7.2)
年齢	１．２０～２４歳	60(6.7)		３．研究・開発	14(1.6)
	２．２５～２９歳	169(18.8)		４．販売・営業	142(16.4)
	３．３０～３４歳	175(19.5)	職位	１．部長クラス	72(8.3)
	４．３５～３９歳	153(17.1)		２．課長クラス	180(20.7)
	５．４０～４４歳	159(17.7)		３．係長・主任クラス	281(32.3)
	６．４５～４９歳	99(11.0)		４．一般	336(38.7)
	７．５０～５４歳	62(6.9)			
	８．５５～６０歳	20(2.2)			

Ⅱ．あなたの仕事への意欲・やりがいに関する次の各質問について、Yes (該当する)、No (違う) のどちらか近いと思われる方一つを選んで〇をつけていって下さい。

１．自分の仕事については、人並の仕事のやり方では満足せずに、常に問題意識をもって取り組み、改善するように心がけている。

1. Yes　738(81.8)　2. No　164(18.2)

２．従来のやり方・先例にこだわらずに仕事をしている。

1. Yes　581(64.1)　2. No　325(35.9)

３．必要な仕事はセクションにとらわれずに積極的に行っている。

1. Yes　603(66.7)　2. No　301(33.3)

４．自分の実力は他の会社でも充分通用すると思う。

1. Yes　378(41.9)　2. No　524(58.1)

５．上司がこうだと言えば、自分に反対意見があっても素直に従う。

1. Yes　315(34.8)　2. No　590(65.2)

６．トップの経営方針と自分の仕事との関係を考えながら仕事をしている。

1. Yes　658(72.9)　2. No　244(27.1)

７．上司からの権限委譲がなされている。

1. Yes　543(60.3)　2. No　358(39.7)

８．自分の意見が尊重されていると思う。

1. Yes　663(73.4)　2. No　240(26.6)

９．21 世紀の自分の会社のあるべき姿を認識している。

1. Yes　454(50.4)　2. No　447(49.6)

10. 良いと思ったことは、周囲を説得する自信がある。

1. Yes　593(65.5)　2. No　312(34.5)

11. 難しい仕事や慣れない仕事にはあまり手を出したくない。

1. Yes　298(32.9)　2. No　608(67.1)

12. 今の職場では、業績を残すよりも、大きな問題やミスを起こさないようにしたい。

1. Yes　348(38.5)　2. No　556(61.5)

13. できれば人よりも早く昇進したいと思っている。

1. Yes　436(48.3)　2. No　467(51.7)

14. ポスト不足の実感から、出世を半ばあきらめている。

1. Yes　298(33.4)　2. No　595(66.6)

15. 現在の職務に満足感を感じる。

1. Yes　457(50.6)　2. No　447(49.4)

Ⅲ．<u>あなたの会社・職場の雰囲気</u>に関する次の各質問について、Yes (該当する)、No (違う) のどちらか近いと思われる方一つを選んで〇をつけていって下さい。

１．仕事上の個人の業績、貢献の高い人は、昇進、昇格あるいは昇給などを確実に果たしている。

1. Yes　361(40.3)　2. No　534(59.7)

２．失敗をしながらでも業績を挙げていくよりは、失敗をしないで過ごした方が評価されると思う。

1. Yes　367(40.9)　2. No　531(59.1)

３．新しい仕事にチャレンジしていこうという雰囲気がある。

1. Yes　461(51.2)　2. No　439(48.8)

４．個性を発揮するよりも、組織風土に染まることを求められる。

1. Yes　496(55.1)　2. No　404(44.9)

５．目標達成に向けて競争的雰囲気がある。

1. Yes　250(27.7)　2. No　651(72.3)

６．斬新な発想や創意工夫を生かそうという雰囲気がある。

1. Yes　432(47.9)　2. No　470(52.1)

７．職場の雰囲気を「ぬるま湯」だと感じることがある。

1. Yes　619(68.5)　2. No　284(31.5)

８．出る杭は打たれる風土である。

1. Yes　376(42.1)　2. No　518(57.9)

９．金太郎飴的な人間の集まりである。

1. Yes　402(44.9)　2. No　493(55.1)

10. 安全性が最優先される雰囲気がある。

1. Yes　655(72.9)　2. No　244(27.1)

11. 有給休暇は自由にとれる。

1. Yes　582(64.4)　2. No　322(35.6)

12. 社内で、県人会、同窓会などのインフォーマルな組織の活動が活発である。

1. Yes　207(23.1)　2. No　691(76.9)

13. 評価は、業績、貢献度でというより、上司の好き嫌いで判断される傾向がある。

1. Yes　362(40.4)　2. No　535(59.6)

14. 20 代の若手社員の能力や個性が職場で生かされていると思う。

1. Yes　361(40.1)　2. No　539(59.9)

15. 外部からの役員就任にもこだわらない雰囲気がある。

1. Yes　514(58.0)　2. No　372(42.0)

Ⅳ. あなたの会社の意思決定スタイルとコミュニケーションに関する次の各質問について、Yes (該当する)、No (違う) のどちらか近いと思われる方一つを選んで〇をつけていって下さい。

１．意見が異なる場合、対立を避け、できるだけ歩み寄ろうとする。

1. Yes　641(71.3)　2. No　258(28.7)

２．本音の議論は、就業時間中というより、社外で行うことが多い。

1. Yes　440(48.9)　2. No　460(51.1)

３．上下関係を意識せずに、自由闊達に議論のできる雰囲気がある。

1. Yes　444(49.2)　2. No　458(50.8)

４．互いに納得のいくまで議論ができる雰囲気がある。

1. Yes　405(45.0)　2. No　496(55.0)

５．議論を議論のままで終らせずに、委員会の組織化や根回し等、実行に向け乗り出すことが多い。

1. Yes　380(42.5)　2. No　514(57.5)

６．仲間内でも、会社全体の長期的経営方針を議論することがある。

1. Yes　425(47.3)　2. No　474(52.7)

７．指示が出されても、やり過ごしているうちに、立ち消えになることがある。

1. Yes　597(66.3)　2. No　303(33.7)

８．問題点はわかっていても、誰も問題提起せず、見過ごされることがある。

1. Yes　642(71.3)　2. No　258(28.7)

９．文書・資料等がいつでも引き出せるように整理がなされている。

1. Yes　443(49.0)　2. No　461(51.0)

10. 最近は資料作りが多すぎると思う。

1. Yes　680(75.7)　2. No　218(24.3)

11. 課と課の間の情報交換がスムーズに行われている。

1. Yes　215(23.9)　2. No　686(76.1)

12. 社員間の横のコミュニケーションが充分なされていると感じる。

1. Yes　259(28.7)　2. No　643(71.3)

13. 現場からの情報が、充分に本社・本部に伝わっていると思う。

1. Yes　138(15.4)　2. No　758(84.6)

14. 本社・本部から、現場に必要な情報が適切に伝達されていると思う。

1. Yes　227(25.3)　2. No　670(74.7)

15. 複数の系統から指示を受けることがある。

1. Yes　594(66.1)　2. No　304(33.9)

Ⅴ．あなたと会社との関わりに関する次の各質問について、Yes (該当する)、No (違う) のどちらか近いと思われる方一つを選んで〇をつけていって下さい。

１．会社に対して忠誠心をもっている。

1. Yes　674(74.8)　2. No　227(25.2)

２．終身この会社で仕事をしていきたいと思う。

1. Yes　565(62.8)　2. No　334(37.2)

３．チャンスがあれば転職したいと思う。

1. Yes　360(40.1)　2. No　538(59.9)

４．自分自身の 10 年後の具体的な将来設計が、なかなか立てられない。

1. Yes　658(73.2)　2. No　241(26.8)

５．会社の進んでいる方向は、自分の望む会社の将来像と結びついている。

1. Yes　416(47.0)　2. No　470(53.0)

６．自分は会社の経営方針を充分理解していると思う。

1. Yes　490(54.5)　2. No　409(45.5)

７．役員層に人間としての魅力を感じる人がいる。

1. Yes　548(61.3)　2. No　346(38.7)

８．経営理念を自分の具体的行動計画にまでブレークダウンしている。

1. Yes　338(37.7)　2. No　558(62.3)

９．職場の目標がはっきりわかっている。

1. Yes　640(71.0)　2. No　262(29.0)

10. 今の職場で自分のやるべきこと、自分の役割を自覚している。

1. Yes　822(90.7)　2. No　84(9.3)

11. 自分なりの行動理念をもって行動している。

1. Yes　786(87.3)　2. No　114(12.7)

12. 人生にゆとりがあり、楽しい。

1. Yes　357(39.7)　2. No　543(60.3)

13. 会社は社会奉仕活動に理解があると思う。

1. Yes　422(47.2)　2. No　473(52.8)

14. 自分としては、職場よりも家庭を重視している。

1. Yes　486(54.0)　2. No　414(46.0)

15. 自分の仕事上の成果に対する上司の評価は、適切で公平だと思う。

1. Yes　646(72.1)　2. No　250(27.9)

Ⅵ. あなたの会社・職場に対する評価に関する次の各質問について、Yes (該当する)、No (違う) のどちらか近いと思われる方一つを選んで〇をつけていって下さい。

１．タテ、ヨコの縛りがゆるく、弾力的に仕事ができる。

1. Yes　474(52.7)　2. No　426(47.3)

２．人材育成のための教育研修制度(off JT)がしっかりしていると思う。

1. Yes　449(49.8)　2. No　452(50.2)

３．人材育成は職場まかせになっていると思う。

1. Yes　515(57.4)　2. No　382(42.6)

４．いわゆる敗者復活の人事が行われたことがある。

1. Yes　351(40.3)　2. No　519(59.7)

５．他の企業に就職した同期と比較して、わが社の出世スピードは遅い方だと思う。

1. Yes　449(51.1)　2. No　430(48.9)

６．関連企業等への出向を積極的に行う方がよい。

1. Yes　662(74.8)　2. No　223(25.2)

７．福利厚生面は充実している。

1. Yes　474(53.3)　2. No　415(46.7)

８．役員層と社員の間にもっと対話が必要だと思う。

1. Yes　762(85.6)　2. No　128(14.4)

９．会社の方針が明確に示されていると思う。

1. Yes　513(58.0)　2. No　372(42.0)

10. 上司から仕事上の目標をはっきり示されている。

1. Yes　565(63.7)　2. No　322(36.3)

11. 基準・規程・マニュアルがあるにもかかわらず、有効には利用されていない。

1. Yes　571(64.4)　2. No　316(35.6)

12. 改善に向け、小集団活動（例えば QC 活動）が効果的に行なわれている。

1. Yes　220(24.8)　2. No　667(75.2)

13. 長期的展望に立った仕事というより、短期的な数字合わせになりがちである。

1. Yes　730(82.3)　2. No　157(17.7)

14. 21 世紀に向けた企業イメージを積極的に社外にアピールしていくべきだと思う。

1. Yes　841(94.6)　2. No　48(5.4)

15. 会社に来ること自体が仕事のようになっている人がいる。

1. Yes　688(77.8)　2. No　196(22.2)

ご協力ありがとうございました。

付章　CMS 入門

章目次

1. はじめに
2. CMS のファイル
3. 入出力装置とのやりとり
4. 端末からホスト・コンピュータへの接続
5. CMS ファイルの管理
6. XEDIT による CMS ファイルの作成と編集
7. 磁気テープ装置
演習問題

1. はじめに

CMS(Conversational Monitor System)は IBM のメインフレーム用に開発された会話型専用 OS である。OS (Operating System)とは、第 2 章第 2 節でも述べたように、簡単に言えば、外部記憶装置や入出力装置といったハードウェアを操作するソフトウェアのことである。

実際には 1 台のメインフレームをホスト・コンピュータにして、それにたくさんの端末をぶらさげて相手をさせている。つまり、1 台のホスト・コンピュータを複数の利用者によって共用しているので、中央演算処理装置、主記憶装置をはじめ外部記憶装置や入出力装置といった資源の配分、管理を利用者間で調整しなくてはならないはずである。ところが、CMS では、端末で利用者がログオンすると、その利用者専用の仮想計算機(Virtual Machine)を起動することになる。物理的なコンピュータ・システムを意識せずに、各自の利用環境を自由に設定できるし、仮想計算機には仮想入出力装置が接続されていることになる。各利用者は他の利用者やメインフレームの全体的な管理を気にすることなく、あたかも自分専用のパーソナル・コンピュータを使用しているかのように、端末を利用することができるのである。この CMS の使用方法、利用の決め事さえ覚えれば、自分がどのようなホスト・コンピュータのハードウェア、機種を使用しているのかを知らずとも、メインフレームのホスト・コンピュータを使用することができる。

コマンド入力の表記方法については、第 2 章第 2 節(4)で説明したものと基本的に同じであるが、異なるのは、原則的に、コマンド入力の際には、リターン・キーの代わりに、(実行)キーを押すということである。リターン・キーにあたる(改行)キーではないので注意してほしい。(改行)キーを押す場面も出てきて複雑なので、この章では、その都度キーの種類を表示することにしよう。端末の種類によっては、(実行)キーは(ENTER)キー、(改行)キーは(矢印)キーのこともあるので、その場合には読み替えること。

2. CMS のファイル

CMS におけるファイル(file)またはデータ・セット(data set)とは、物理的にはホスト・コンピュータの固定ディスクに格納されたプログラムやデータの集合のことである。ファイルは論理的には利用者ごとに振り分けられた仮想の「ミニ・ディスク」に保管されている。

パンチ・カードを使って入力を行っていた時代には、プログラムやデータはパンチ・カードにパンチされ、パンチによって開けられた穴を光学的にカード・リーダーと呼ばれる読取機械で読み取ってコンピュータに入力していた。パンチ・カード1枚に80字の数字、アルファベットもしくはカタカナをパンチすることができた。ここでいっている1字分を書き込むべき場所はパンチ・カードの場合にはカラム(column)と呼ばれたが、この1字もしくは1カラムは、情報の量としては1バイト(byte)に相当している。つまりパンチ・カード1枚は80カラムからなり、80バイトの情報を入力することができたことになる。

現在では、パンチ・カードはほとんど使用されず、端末から入出力が行われるが、端末の画面の上の1行がこのパンチ・カード1枚に相当し、この1行に書かれたデータをレコードと呼ぶ。ファイルは、正確にはこのレコードの集合である。パンチ・カードの時代には、コンピュータに実行させたいプログラムやデータをパンチ・カードにパンチし、そのカードの束(これをカード・デックと呼んだ)を枚数が少なければ輪ゴムでまとめ、枚数が多ければファイリング・ケースに入れて保管したり、持ち歩いたりしていた。カード・リーダーにカードを読み取らせるときにも、この束単位でカード・リーダーにかけていた。この束こそが「ファイル」だったのである。ちなみに、一般的な端末の画面の左端から右端まで1行いっぱいに文字を入力すると半角文字で80字(つまり80バイト)入るようになっているのは、このパンチ・カードの名残である。

(a)ファイルの属性

物理的に用紙の大きさの規格が統一されていたパンチ・カードとは違って、端末からの入力では改行するまでは1行であり、別に各行を80字に固定したり、(80字ではなくとも)各行の長さを揃えたりしなければならない理由はない。したがって、色々な形式のファイルが存在しうることになるわけだが、CMSではファイルの属性として、次の4つを表示してくれることになっている(第4節で後述するFILELISTコマンド参照)。

1. レコード形式(FORMAT): 1行の長さが固定長(F)か可変長(V)かということ。
2. レコード長(LRECL): レコード形式が固定長の時は1行の長さ、可変長の時は1行の長さで最大のものをバイト数で表示したもの。
3. レコード数(RECORDS): ファイルのレコード数、すなわち行数を表示したもの。
4. ブロック・サイズ(BLOCKS): ファイルの大きさをブロック数で表示したもの。

このうち4については、もう少し説明を要するだろう。計算機ではデータの保存効率を高めるために、いくつかのレコードをまとめて、ブロック化することが行われる。その際の1ブロックの大きさはCMSでは4KB。CMSではこのブロック・サイズに基づいたブロック数を表示している。

標準的なファイルの形式は、

1. レコード形式＝固定長
2. レコード長＝80バイト

で、この形式のファイルを「カード・イメージ」のファイルと呼ぶこともある。特別な事情や理由がない限り、ファイル形式はこの標準型にしておいた方が無難である。カード・イメージのファイルでトラブルに巻き込まれることはほとんどないが、それ以外の形式のファイルではソフトウェアによってはトラブルに巻き込まれることがある。ちなみに、CMS版SASでは、固定長のファイルしか受け付けないことになっていて、可変長のファイルではSASの

システムは作動しない。

(b)行番号

各レコードは行番号を付けることも付けないこともできる。ここでいっている行番号は、後で編集のことに触れる際に出てくる編集画面上、左端に表示される「行番号」のことではなく、各レコード固有の行番号である。これもカード時代の名残で、かつてはカードの束を落したりして順番がメチャメチャになったときでも、この行番号のおかげで救われたものである。したがって、現在では付けても付けなくても、ほとんど意味はない。ただ問題になるのは、行番号をつける際には、各レコードの最後部 8 バイト(73 カラム目～80 カラム目)を使って行番号が書き込まれるということである。このことは注意を要する。もし行番号なしのファイルを行番号付きのファイルに変えたときには、レコード最後部の 8 バイト分のデータが欠落してしまうからである。データのファイルは後でコーディング・ミスやパンチ・ミスの訂正を行うことが多いので、その際によく訳もわからないままに、オプションを指定してしまったりして、行番号を誤って書き込んでしまうおそれがないとはいえない。したがって、例えば、固定長、80 バイトのカード・イメージのファイルを作成する際には、特別の事情がない限りは、もしもの用心のために各行の 1～72 カラムのみを使い、最後部の 8 カラムは使用しないようにする。SAS プログラムや入力データは行番号付きでも行番号なしでもよいことになっているので、行番号なしで使っても、この方が無難である。

(c)ファイル識別子

ファイル識別子(fileid とも書き表される)はデータ・セット名(data set name: dsname)とも呼ばれるが、SAS ではファイルをデータ・セットと呼ぶので、SAS の中ではデータ・セット名と呼ぶのが通例である。ファイル識別子の例としては、

NEWFILE SAS A1

のようになり、ファイル識別子は空白(ブランク)をはさんだ

1. ファイル・ネーム(fn)
2. ファイル・タイプ(ft)
3. ファイル・モード(fm)

の 3 部分から構成される。この例では "NEWFILE" はファイル・ネーム、"SAS" はファイル・タイプ、"A1" はファイル・モードを表している。具体的な表記方法としては、1、2 は 8 文字以内で、英数字 A～Z、0～9 および$、#、@、+、-(ハイフン)、:(コロン)、_(下線)からなる文字列である。3 は 1 桁の英文字と 1 桁の数字の 2 文字で表されるが、ただし、通常、利用者側が指定するときには英文字部分だけでよい。各ユーザー名に属するミニ・ディスクのファイル・モード(fm)は "A"（既定値）となっている。

わかりやすくいうと、1. ファイル・ネーム(fn)と 2. ファイル・タイプ(ft)の組でファイルの名前を表している。いわば人の姓名と同じで、2. ファイル・タイプ(ft)が姓、名字、1. ファイル・ネーム(fn)が名に相当している。SAS のプログラムは「SAS 家の一員」ということで「SAS 姓」を名乗り、ファイル・タイプとして SAS を用いることになっている。SAS はそれ以外のファイル・タイプをもったファイルを SAS プログラムのファイルとして「認知」せず、受け付けないので注意がいる。それに対して、3. ファイル・モード(fm)は住所に相当し、どこの地区(ミニ・ディスク)にファイルが「住んでいるのか」を示している。特

に断らなければ、ミニ・ディスク A、つまり自分も住んでいる「A 地区」に住んでいると考える約束になっている。

3. 入出力装置とのやりとり

CMS では、入出力装置は既定値として端末のみが指定されている。逆にいうと、端末だけで通常の入出力は事足りているわけである。しかし、データが非常に大きいときや、他のコンピュータからデータをもらってくるとき、あるいはハード・コピーとしてきちんと紙の上に印刷して残したいときでも、端末だけで我慢しろというわけではない。端末以外の入出力装置、例えば、読取装置、穿孔装置、磁気テープ装置、印刷装置といった装置との入出力は、基本的には次のように考えて行うことになっている。

1. 磁気ディスク上の CMS ファイルを論理的に入出力装置(仮想入出力装置)と見立てた上で、既定値で端末となっている入出力装置の設定をその CMS ファイルに変更し、
2. 仮想計算機の入出力はその CMS ファイルに対して行い、
3. 入出力の実装置の操作はその CMS ファイルに対する管理の形で行う。

このうち、1 で磁気ディスク上の CMS ファイルを論理的に入出力装置(仮想入出力装置)と見立てた上で、既定値で端末となっている入出力装置の設定をその CMS ファイルに変更するには、データ定義名(data definition name: ddname)を用いて定義を行う。この一連の操作についてまとめておこう。

fileid で示された磁気ディスク上の CMS ファイルを入出力装置と見立てる際には、

FIledef ddname DISK fileid(実行)

例えば、次のように入力すればよい。

FI 18 DISK OUT91 DATA(実行)

ここで、例からもわかるように ddname は「仮想入出力装置名」とはなっているが、実際には入出力装置番号でよく、筆者は例にあるような "18" や "IN" などもよく使う。ただし番号でよいとはいっても、11〜15 は使えないことがある。同様に、SAS を使用する際には、次の ddname は使用できないので、注意がいる。

LIBRARY, SASDUMP, SASLIB, SORTWKxx, SORTLIB, SYSIN, $SYSLIB, SYSOUT, WORK

ここで指定した ddname は、2 で仮想計算機の入出力を CMS ファイルに対して行う際に、仮想入出力装置名として当該装置番号が指定されることになる。

こうした仮想入出力装置名の指定を解除し、既定値の端末に戻すには

FIledef ddname CLEAR(実行)

例えば、次のように入力すればよい。

FI 18 CLEAR(実行)

このように各装置名ごとに指定を解除するのでは面倒な場合には、全ての設定を解除し、既定値の端末に設定を戻すことも可能である。このときには、ワイルド・カード"*"を指定して、

FIledef * CLEAR(実行)

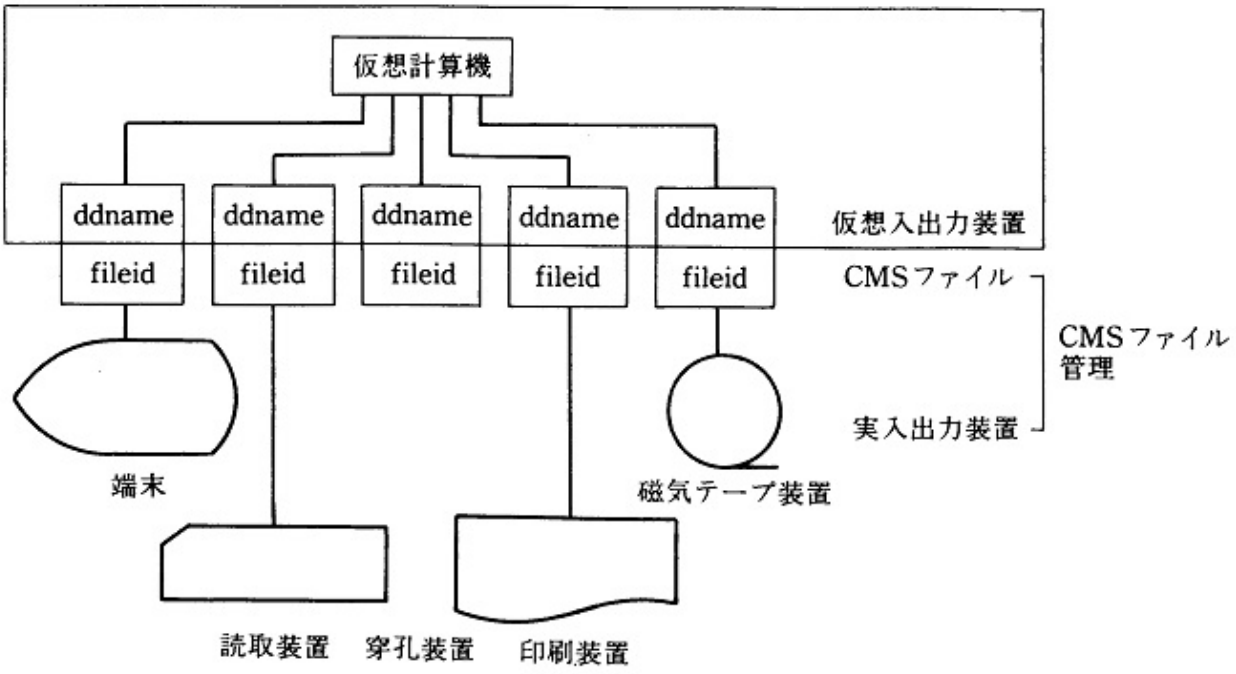
と入力すればよい。ワイルド・カードつまり星印 "*" は「該当する項目のすべて」を意味している。

ここでは、端末からその都度 CMS にコマンドを入力することを考えているが、実際のケースでは、SAS のプログラム中で「X CMS コマンド」の形が使えるので、SAS プログラム内で次のように指定した方が混乱しないですむ。

X FIledef …… ;

このようにデータ定義名を定義することにより、図 7.1 のように仮想入出力装置と CMS ファイルの対応をつけることができる。

図 7.1　仮想計算機と仮想入出力装置



この図の中の仮想読取装置、仮想穿孔装置、仮想印刷装置はまとめて仮想ユニット・レコード装置といい、実装置との間での入出力は CMS ファイルの管理の形で行われる。ただし、現在では実際のコンピュータ・システムが読取装置、穿孔装置の実装置をもたないことが多く、この場合には、それぞれが入力用 CMS ファイル、出力用 CMS ファイルの意味しかもたない。また磁気テープ装置は接続台数が少なく、しかも磁気テープ装置には、各利用者が使用する磁気テープが常時かけられているわけではないので、磁気テープ装置を使用する際には、磁気テープ装置を仮想計算機の入出力機器として接続する手続きが特別に必要になる(磁気テープ装置の使用については、第 7 節を参照のこと)。

4. 端末からホスト・コンピュータへの接続

(1)ログオンとログオフ

以上で、もっとも基本的な部分の解説は終ったので、あとは実際にホスト・コンピュータをいじりながら話を進めていこう。それでは、とにかく端末機の前に座って、ホスト・コンピュータと接続してみよう。このことをログオンするという。自分のユーザー名とパスワードさえもっていれば、ログオンの仕方はいたって簡単である。次のような手順でやってみると、すぐにホスト・コンピュータとの会話が可能になる。

1. 端末の電源を入れる。
2. 専用端末は 3 に進む。パーソナル・コンピュータを端末として使用する場合には、ここで「プログラム選択」のメニュー画面が表示されるので、パーソナル・コンピュータを端末として機能させるためのプログラムを起動する。多くの場合は「区画」の 1 を選ぶようになっている。つまり、

   1(実行)
3. ロゴ(例えば、「CT」)が表示されるのを確認し、その下の指定された場所に、次のようにユーザー名とパスワードを入力する。

   USERID ===>ユーザー名
   PASSWORD ===>パスワード(実行)

ただし、3 については、ユーザー名の後では、(実行)キーは押さずに、パスワードを入力した後で、(実行)キーを押す。ユーザー名を入力すると、他のキーに触れなくても自動的に改行し、パスワード行にカーソルが動いているはずなので、画面をよく見ること。もし、ユーザー名を入力しても、パスワード行にカーソルが移動しない場合には、(改行)キーを押すことになるが、端末で、入力を促す "===>" の表示の後ろに入力する場合には、その直後の 1 カラムには入力できないことになっているので、例えば、

===> TAKAHASI

というように、必ず 1 字分、カーソル・キー(スペース・キーではない)であけて入力すること。以上の手続きが無事終了すると、

LOGON ユーザー名
LOGON AT hh:mm:ss JST day mm/dd/yy
VM/SP REL 5 11/06/90 01:15

と表示される。これでログオンが済んで、TSS 処理が開始され、端末はホスト・コンピュータであるメインフレームに接続されたオンラインの状態にある。この状態になると、もう端末のキーボードを使って、ホスト・コンピュータにコマンドを送ることができる。プログラムやデータを入力して、ホスト・コンピュータに仕事(job)をさせることができるのである。ホスト・コンピュータへの用が済んだら、端末をホスト・コンピュータから切り放しておく。このことをログオフするといい、次のようにする。

1. LOGoff(実行) と入力する。
2. 端末の電源を切る。

このうち 1 の段階で、端末機はホスト・コンピュータとは切り放されて、パソコン端末の場合には独立したパーソナル・コンピュータとして機能することになるので、たとえトラブルがあっても、ホスト・コンピュータには迷惑をかけずに済む。この状態になれば、2 として端末の電源を切れば、使用は終了する。

(2)システムの状態表示

まずメインフレームの初心者に対する一般的なアドバイスから始めよう。端末のディスプレイ画面に赤色の表示が出たら、ホスト・コンピュータからの「警告」である。こうした警告を解除するのはいたって簡単である。(取消)キーまたは(RESET)キーを押せばよい。しかし、これでは何ら問題の解決にはなっていない。

警告が出たのなら、解除することをあせっても仕方がない。まず気を落ちつけて、なぜ警告が出されたのかを考えてみる必要がある。原因は必ず利用者の側にある。多くの場合

は「タイプ・ミス」が原因である。しかし、このタイプ・ミスは広い意味でのタイプ・ミスで、端末を操作している当人が、「真面目に間違えて操作」している場合、つまり、「確信をもってミス」している場合でも、それがコンピュータにとって「意味不明」ならば、コンピュータの側ではタイプ・ミスとして形式的に判断するのである。したがって、なぜ警告が出たのかを確認する必要がある。ほとんどは本書のようなテキストやマニュアルの読み間違い、誤読が原因であるので、もう一度入念に読み直して、確認してほしい。

それともう一つ。ホスト・コンピュータは利用者と会話したがっている。もちろんコンピュータが声を出して話すわけではないので、コンピュータ側からの呼びかけは、メッセージとして端末のディスプレイ画面を通して行われる。ところが、初心者はほとんど画面を読まない。これでは話が成立しない。ホスト・コンピュータからのメッセージは、画面の隅につつましく出されることが多いので、コンピュータと会話するためには、画面の隅々まで気を配ってやる必要がある。相手が機械だからといって、利用者は一方的にコンピュータに命令することばかりを考えずに、コンピュータの意見やアドバイス、メッセージをきちんと受け止めてやる寛容さが必要である。もし皆さんがコンピュータの立場だったら、人の質問や意見、忠告にも耳を貸さず、ただ一方的にガミガミと一貫性のない命令をし、思い通りにならないと暴力まで振るうようなヤツ(たまに見かけるが絶対にやめてほしい)の言うことは聞かないはずである。コンピュータもそうした場合には言うことを聞かない。たまに口論をすることはあっても、コンピュータと良好な「友人」関係を築き上げてほしいものである。

そこで、パソコンとはちょっと勝手の違うホスト・コンピュータからのメッセージの読み方の初歩、基本中の基本を表 7.1 にまとめておくので、最低限の会話を試みてほしい。ログオンしてホスト・コンピュータと接続すると、画面右下隅に次のようなステータス(＝システムの状態)が表示される。それぞれが利用者に対するメッセージとなっているので、覚えておくと計算機の「考えていること」がわかる。この表のうち最初のケース、すなわち左下隅に表示かなく、右下隅に「RUNNING」だけが表示されている場合は、通常のステータスで、正常に処理を行っているという表示である。他のステータス表示になっている場合には、それなりに対処する必要があるので、表を参考にしてほしい。

表 7.1　画面右下隅に表示されるステータスと対処方法

ステータス		対処方法と《意味》	
RUNNING	左下隅に表示なし	《入力待ち: 通常の表示》	
	左下隅に「X ○」「X SYSTEM」	《ホスト作業中》	その終了待ち
	左下隅に赤い「XX …… 」	《誤操作なので取消せ》	(取消)キーで解除
VM READ		(実行)キーで「RUNNING」表示	《通常の表示》
		(実行)キーで「CMS」表示	《コマンドの入力待ち》
		(実行)キーで「?」表示	《実行プログラムのデータ入力》
CP READ		ログオン前: Logon(実行)	
		CMS 使用中: B(実行)	それで変化なければ I CMS(実行)
NOT ACCEPTED		《ホスト作業中》	数秒待ってコマンド再入力
MORE...		《1 分後に次画面表示》	停止するには(実行)キー 即刻表示には(CLEAR)キー
HOLDING		《現画面表示を保持》	その解除には(CLEAR)キー

注) 端末によっては、(実行)キーは(ENTER)キー、(取消)キーは(RESET)キーのこともある。

5. CMS ファイルの管理

CMS ファイルの管理は、原則的に FILELIST コマンドによって呼び出されたファイルの一覧表の画面を使って行われる。

(1)FILELIST コマンドの使用開始と終了

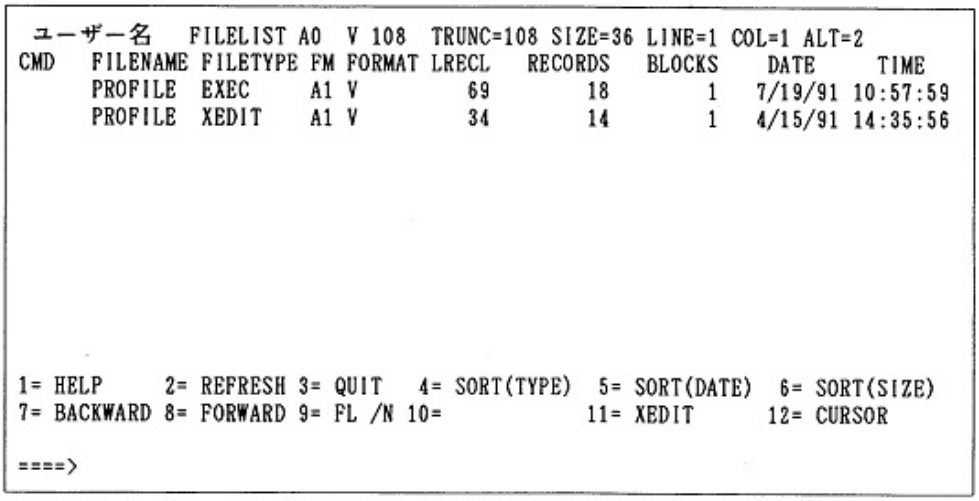
次のどちらの方法でも、FILELIST コマンドの使用を開始し、ファイルの一覧表を表示することができる。

FILEList(実行)　‥‥‥ 自分の保有するファイル(ただし fm=A)の全部を表示。

FILEList * SAS(実行)　‥‥‥ 「ft=SAS のファイルのみを表示」(ワイルド・カード「*」を使った例)

例えば、前者の方法で、FILELIST コマンドによって、自分のミニ・ディスク A の全てのファイルを表示させると、図 7.2 のようにファイルの一覧表が画面上に表示される。

図 7.2　FILELIST コマンドによるファイルの一覧表(初めてログオンしたときの状態)



この画面はまだファイルを自分で作成したことがない使い始めの最初の状態を示したものである。この画面からもわかるように、この状態で既に、各利用者のファイルとして PROFILE EXEC と PROFILE XEDIT の二つのファイルが必ず入っている。この二つのファイルは、ログオンと同時に自動的に実行される初期設定ファイルなので、いじらないこと。もしこのファイルが壊れると、特に PROFILE EXEC ファイルの方が壊れると、基本的な環境設定が出来ず、通常の使用は不可能になるので注意が必要である。

この図のような FILELIST 画面の状態で、キーボード上の PF キーを使って CMS ファイルの管理を行うことができる。FILELIST 画面の PF キーの設定された機能は画面の下部に表示された通りで、これらの機能は比較的よく使用されるために PF キーに設定されている。

PF1	＝HELP	コマンド、機能についての説明を表示する。
PF2	＝REFRESH	余分な表示を除いて、FILELIST 画面をリフレッシュする。
PF3	＝QUIT	FILELIST コマンドを終了し、システム・モードに戻る。
PF4	＝SORT(TYPE)	ファイル・タイプ(FILETYPE)で分類して並べ直したファイルの一覧表を表示する。
PF5	＝SORT(DATE)	作成年月日(DATE)にしたがって並べ直したファイルの一覧表を表示する。
PF6	＝SORT(SIZE)	ファイルの大きさ順に並べ直したファイルの一覧表を表示する。
PF7	＝BACKWARD	前の画面に戻る。
PF8	＝FORWARD	次の画面に進む。
PF9	＝FL /N	カーソル行のファイルと同じファイル・ネーム(FILENAME)のファイルの一覧表を表示する。
PF10	＝	
PF11	＝XEDIT	カーソル行のファイルを XEDIT で編集する。
PF12	＝CURSOR	カーソルをコマンド行に移動させる。

このうち一番重要なのは、FILELIST コマンドを終了させる PF キーで、(PF3)キー(＝QUIT)がそれである。このキーを押すと、

READY; T=n.nn/n.nn hh:mm:ss

と表示され、CPU 使用時間とシステム・モード(CMS コマンドが使える状態)に戻った時刻とが示される。その他の機能と使用方法については(2)で触れる。

(2)FILELIST 画面を使ったファイル管理

(a)ファイルの消去

1. 消去したいファイルの CMD 欄に次のコマンドを重ね書きしていく。

   DISCARD(実行)

2. ファイルが消去されると、画面上の消去されたファイル(ファイル・ネーム fn)の行に

   fn ** HAS BEEN DISCARDED

   と表示される。
3. (PF2)キー(＝REFRESH)を押すと、消去されたファイルを除いたファイルの一覧表が表示される。

(b)ファイルのコピー

1. コピーしたいコピー元のファイルの CMD 欄に、コピーする先のファイル名を次のように重ね書きして指定する。

   COPYfile / newfn newft newfm(実行)

   ここで、"/" は現在行のファイル名を示している。またコピー先のファイル識別子のファイル・モード newfm は省略できないので注意がいる。通常は A1 を newfm として指定する。
2. 画面上ではコピー元のファイルの CMD 欄に "*" が付く。
3. (PF2)キー(＝REFRESH)を押すと、コピー先のファイルも含めたファイルの一覧表が表示される。

(c)ファイルのプリンタでの印刷

1. プリンタで印刷したいファイルの CMD 欄に、次のコマンドを重ね書きする。

   PRint(実行)

(d)他の利用者へのファイルの送信

1. 送信したいファイルの CMD 欄に、送信先のユーザー名を次のように重ね書きして指定する。

   SENDFile / ユーザー名 [AT ノード名](実行)

2. 画面上に

   FILE fn ft fm SENT TO ユーザー名 AT ノード名 ON dd/mm/yy hh:mm:ss

   と表示されると送信成功である。

同じホスト・コンピュータを使っている他の利用者に送信する場合は、1 の[ ]内のノード名は省略してもかまわない。ノード名は例えば "JPNUTKOM" とすると、東京大学教養学部 2 号館 6 階の機器室にある IBM メインフレームを意味することになる。実はこれは bitnet と呼ばれるネットワーク内のノード名で、世界中どこでも、その bitnet に加入しているホスト・コンピュータのノード名を指定すれば、そのホスト・コンピュータの利用者に対して、ファイルを送ることが出来る。逆に、名刺にユーザー名とノード名を例えば

TAKAHASI AT JPNUTKOM
と刷り込んでおけば、世界中の bitnet 利用者からファイルを受信することが出来る。

(3)他の利用者から転送されてきたファイルの受信

他の利用者から転送されてきたファイルを受信するためには、FILELIST コマンドと同レベルの RDRLIST コマンドを使って行う。FILELIST コマンドを使用している場合には、FILELIST 画面を一旦終了し、システム・モードに戻した上で、RDRLIST コマンドを使用する。

1. CMS のシステム・モードで RDRLIST コマンドを入力する。
   RdrList(実行)
2. 送信されてきているファイルの一覧が表示される。
3. 受信したいファイルの CMD 欄にカーソルを合わせる。
4. 同じファイル識別子をもつファイルが存在しない場合、CMD 欄に
   RECEIVE(実行)
   と入力する。同じファイル識別子をもつファイルが存在する場合、既存のファイルに上書きするならば、CMD 欄に
   RECEIVE / (REPLACE(実行)
   と入力する。ただし、このときは上書きされることで、既存のファイルの内容は失われてしまう。それを避けるためには、新規のファイル newfn newft として登録する必要がある。このときは CMD 欄に
   RECEIVE / newfn newft(実行)
   と入力すればよい。

(4)ホストと端末の間でのファイル転送

ホスト・コンピュータの CMS ファイルを端末となっているパーソナル・コンピュータの MS-DOS ファイルにコピーすることが出来る。その逆も可能である。この作業は、ホスト・コンピュータと端末となっているパーソナル・コンピュータとの間でのファイル転送と呼ばれている。

ホスト・コンピュータにログオンした後は、二つのキーを同時に
(前面)＋(終了)
と押せば、CMS のシステム・モードと MS-DOS のシステム・モードとがスイッチの on/off のように切り替わる。ホスト・コンピュータと端末になっているパーソナル・コンピュータの間でファイル転送をするには、ホスト・コンピュータにログオンした後で、MS-DOS のシステム・モードに切り替え、端末側のソフトウェアを使って行う。いま端末側のディスク・ドライブ A のフロッピィ・ディスクのある MS-DOS ファイルを CMS ファイル fn ft に転送(正確にはコピー)するとき、及びその逆のケースでは、MS-DOS のシステム・モードで、次のように入力すればよい。

《端末→ホスト》
C:> SEND A:MS-DOS ファイル名 fn ft [(JISCII CRLF(改行)

《端末←ホスト》
C:> RECEIVE A:MS-DOS ファイル名 fn ft [(JISCII CRLF(改行)

ここで、"JISCII" は、ASCII コードを EBCDIC コードに変換し、PC 漢字コードを IBM 漢字コードに変換するという意味、"CRLF" は、ファイルの中の復改(CRLF)コードをそのまま転送するという意味のパラメータである。

6. XEDIT による CMS ファイルの作成と編集

SAS のプログラムやデータは、第 2 章では SAS の PROGRAM EDITOR ウィンドウで作成することにしている。しかし、CMS のエディターである XEDIT によって作成することもできる。本書では XEDIT を使わなくても済むようにしているが、プログラムやデータを移植、転送する人は、この節の(3)の部分は必要になるので、最低限そこだけは読んでおくこと。

(1)CMS ファイルの作成

新しい CMS ファイルを作る時は、次のような手順にしたがって行う。

1. システム・モードで XEDIT コマンドを用いて、新しくこれから作成しようとしているファイル識別子 fn ft を指示して、ファイルを開く。fm は特に指示しなければ、A1 が既定値となる。

   Xedit fn ft [fm](実行)

   これによって、XEDIT の編集画面がディスプレイ画面に表示される。
2. コマンド行にカーソルがあるので、次のように入力して入力モードに設定する。

   ====> Input(実行)
3. 順にデータを画面に入力していく。画面の最終行まで入力が終ったときは、(実行)キーを押すと、新たな入力領域が確保される。
4. 入力作業を終了したら(実行)(実行)と 2 回キーを押す。これで入力モードから編集モードに戻るので、編集については次の(2)の②を参照のこと。再び入力モードに戻りたいときには、②に戻って操作を繰り返す。
5. コマンド行にカーソルが移動していることを確認して、次のように入力して作業結果を保存する。

   ====> FILE(実行)

   ただし、それまでに行った作業結果を破棄することもできて、そのときには、次のように入力する。

   ====> QQuit(実行)

(2)CMS ファイルの編集

既存の CMS ファイルを編集する時には、

1. FILELIST コマンドでファイルの一覧表を表示させる。
2. CMS ファイル fn ft にカーソルを合わせ、(PF11)キー(＝XEDIT)を押す。すると、画面は図 7.3 のようにファイル fn ft の編集画面に変わるので、編集作業を行う。この編集画面の例は空白行 10 行からなるファイルを表示したものである。XEDIT はフル・スクリーン・エディターで、しかも第 2 章第 3 節(4)にある SAS の編集コマンドと同じ編集コマンドが使える。ただし、SAS の編集コマンドの行コマンドのうち、複写、移動のときに用いた A(After)、B(Before)はそれぞれ F(Follow)、P(Perced)になる。
3. 編集作業を終了したら、コマンド行にカーソルを移動させ、次のように入力して作業結果を保存する。

   ====> FILE(実行)

   これにより、ファイルをディスクに書き込んで編集を終了する。編集を終了させな

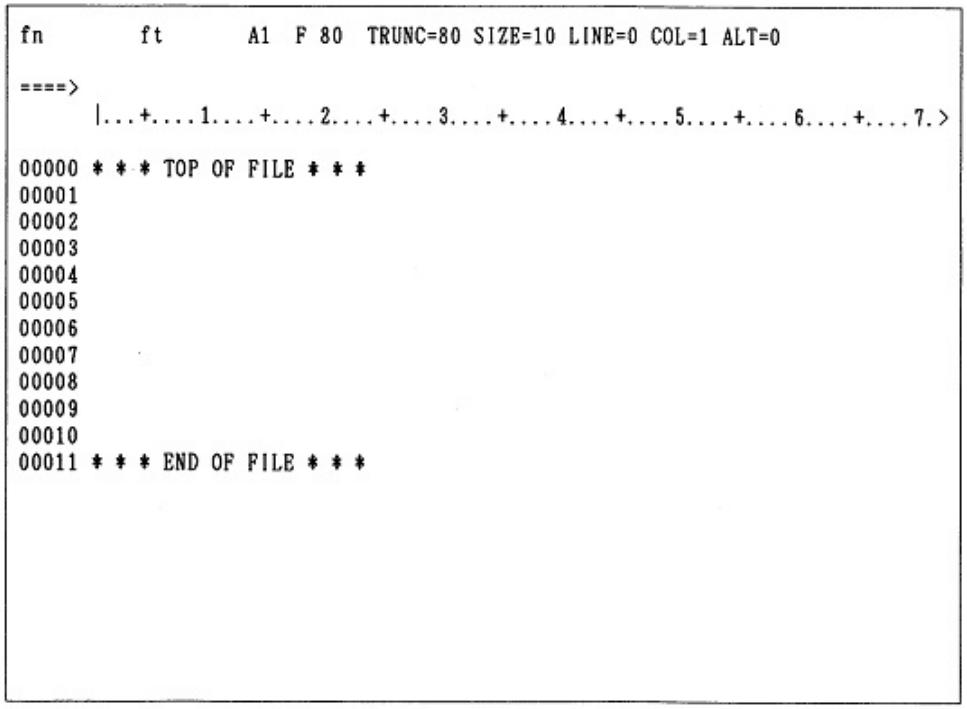
いで、とりあえず現時点でのファイルをディスクに書き込むには、

====> SAVE(実行)

と入力すればよい。入力モードと同様に、それまでに行った作業結果を破棄することもできて、そのときには、やはり次のように入力すればよい。

====> QQuit(実行)

図 7.3　XEDIT の編集画面



(3)応用例：可変長のファイルを固定長のファイルにコピーする

SAS ではプログラム・ファイルは固定長でないと受け付けないが、他からプログラムを移植、転送した場合などには、ときどき可変長のファイルとなってしまっていることがある。こうした場合には、可変長のファイルを固定長のファイルに変えなくてはいけないわけだが、そのときは、新規のファイルを作成しておき(自動的に固定長のファイルとなっている)、そのファイルに可変長のファイルを読み込むことで、内容としては全く同一だが、形式だけが固定長に変わったファイルを作成することが出来る。つまり、正確にいえば、可変長のファイルを固定長のファイルにコピーすることが出来るのである。具体的な手順は、

1. まず、システム・モードで XEDIT コマンドを用いて新規のファイル fn1 ft1 を開く。

   Xedit fn1 ft1 [fm1](実行)
2. この新規ファイルの編集画面のコマンド行にカーソルがあるので、次のように入力して、対象となる可変長のデータ・ファイル fn2 ft2 を読み込む。

   ====> GET fn2 ft2 [fm2](実行)
3. ファイル fn2 ft2 の内容が読み込まれたことを画面で確認して、コマンド行に次のように入力して、そのまま保存する。

   ====> FILE(実行)

7. 磁気テープ装置

ダウンサイジングの進行する中でも、メインフレームを使用する最大の理由は、大量のデータを取り扱う必要があるということであろう。そこで、この章の終りに、大量のデータのデータ・エントリーを外注した場合を考え、磁気テープ装置の取り扱いについて説明しよう。

(1)外注したデータ・エントリーの磁気テープでの納入

磁気テープ(MT: Magnetic Tape)に次のような形式で書かれているファイルを自分のユーザー名のディスク上の CMS ファイルとして読み込む(正確にはコピーする)ケースを扱う。

1. ノンラベル
2. レコード長 80 バイト(固定長)
3. ブロック化係数 40 (ブロック・サイズ 3200 バイト)
4. 6250/1600 BPI (Byte Per Inch)

これは 4 を除いて第 6 章資料 D の入力フォーマットと同じケースである。ここで、2 はこのファイルがカード・イメージであることを意味している。3 のブロック化係数 40 とは、40 レコードを 1 ブロック化しているという意味で、ブロック・サイズは 80×40＝3200 バイトとなる。4 は磁気テープの 1 インチ当り何バイト書き込むかということで、書き込まれる密度を意味している。磁気テープ装置の種類によっては 1600 BPI しかできないものがあるので、事前に確認しておく必要がある。もっとも、いまや 6250BPI も 1600BPI もどちらも可能なものが普通で、この場合には、読み込む際に特にどちらかを指定する必要はない。実際の読み込み手順は次のようになる。

1. 磁気テープ装置(ここでは、装置番号＝580; 記号識別子＝TAP1; 仮想装置アドレス＝181 とする)の蓋を開けてテープを装着し、テープの端を吸い込み口に入れ、蓋を閉める。(この際、MT の黄色のリングを外すと、その MT は書き込み禁止となる。)
   その上で、
   (ロード／巻き戻し)ボタンを２回押す
2. コンソールに次のように入力する。
   ATTach 580 ユーザー名 181(実行)
3. 自分でログオンしている端末で
   FI IN TAP1 (RECFM FB LRECL 80 BLOCK 3200(実行)
   FI OUT DISK fn ft fm (RECFM FB LRECL 80 BLOCK 3200(実行)
   MOVE IN OUT(実行)
4. FILELIST コマンドを使って、指定の fn ft fm が読み込まれていることを確認した後、端末に
   DETach 181(実行)
5. テープを取り外す。

ただし 1 と 2 の操作は、磁気テープ装置の種類などハードウェアによって異なるので、センター等に確認すること。

(2)磁気テープ上の CMS ファイルの操作

この読み込み手順のうち、1 と 2 の準備さえ整えば、3 の磁気テープ上のファイル操作の部分には、代わりに次のような TAPE コマンドを使って、磁気テープ装置上の CMS ファイルの操作を行うこともできる。

TAPE REW	・・・・・・	テープをロード開始点まで巻き戻す。
TAPE RUN	・・・・・・	テープを巻き戻しアンロードする。
TAPE SCAN fn ft	・・・・・・	走査してファイル識別子をリストし、ファイル fn ft の直前に停止。
TAPE SKIP fn ft	・・・・・・	走査してファイル識別子をリストし、ファイル fn ft の直後に停止。
TAPE LOAD fn ft	・・・・・・	テープ・ファイル fn ft をディスクに読み込む。
TAPE DUMP fn ft (TAP1 DEN 6250	・・・・・・	ディスク・ファイル fn ft をテープにダンプする(書き込む密度は 6250BPI)。

オプションとして、SCAN, SKIP, LOAD, DUMP されるファイルのリストを印刷したいときには "(PRint"、 端末に表示したいときには "(Term" を指定することが出来る。

こうした TAPE コマンドの典型的な使用例は、大学のホスト・コンピュータを使用する際に年度の変わり目(つまりユーザー名の変更・更新時期)に、自分のユーザー名のディスクに保管されている CMS ファイルを磁気テープに退避・保存する際の使用例で、3 の部分を

TAPE REW(実行)
TAPE DUMP * * A (PR TAP1 DEN 6250(実行)
TAPE RUN(実行)

とすると、自分のユーザー名に保管されている全ての CMS ファイルを 6250BPI で磁気テープにコピーし、コピーしたファイルの識別子のリストをプリンタに印刷させることが出来る。この磁気テープを使って、ユーザー名が更新された後で、やはり 3 の部分で

TAPE LOAD * *(実行)

とすると、磁気テープ上の全ファイルをディスクに読み込むことが出来る。

演習問題

7.1　ファイルの作成・編集　CMS ファイルとして、LETTER DATA A1 を作成せよ。ファイルは 1 画面に収まる行数であれば、内容、表現等は自由である。例えば、英文やローマ字の手紙、感想文でもよいし、英数字を使って描いた図形、絵でもよい。とにかく、XEDIT を使って各自でオリジナリティー溢れる内容のファイルを作成してみること。作成したファイルは、端末キーボードの(ページ印刷)キーを押して、画面をそっくりまるごとプリンタにハード・コピーをとること。

《ヒント：大文字・小文字を混ぜて使用しているのに、大文字だけで表示されてしまうような場合、ファイルに大文字・小文字を混在させたければ、XEDIT 編集画面のコマンド行に

====> SET CASE Mixed(実行)

と入力するとよい。》

7.2　ファイルのコピーと転送　ファイル PROFILE EXEC は可変長のファイルであることを FILELIST コマンドで確認せよ。これを PRO DATA A1 という固定長のファイルにコピーして保管せよ。その後、このファイル PRO DATA A1 をユーザー名=_____ に転送せよ。

参考文献

本書を入口としてさらに学習を進める読者のために、読者が比較的入手しやすくかつ読みやすいように、和書のテキスト(翻訳を含む)を新しい順に列挙しておこう。

入門的統計学テキスト

- 東京大学教養学部統計学教室 編『統計学入門』東京大学出版会, 1991.
- 宮川公男『基本統計学[新版]』有斐閣, 1991.
- 中村隆英・新家健精・美添泰人・豊田敬『統計入門』東京大学出版会, 1984.
- 林周二『統計学講義第２版』丸善, 1973.

やや詳しい統計学テキスト

- 竹内清・佃良彦 編『経営統計学』有斐閣, 1990.
- R.E.ヘンケル『統計的検定－統計学の基礎－』朝倉書店, 1982.
- S.チャタジー・B.プライス『回帰分析の実際』新曜社, 1981.
- H.B.アッシャー『因果分析法』朝倉書店, 1980.
- P.G.ホーエル『入門数理統計学』培風館, 1978.
- A.M.ムード・F.A.グレイビル・D.C.ボウズ『統計学入門（上)』マグロウヒル好学社, 1978.
- I.ガットマン・S.S.ウィルクス『工科系のための 統計概論』培風館, 1968.
- 肥田野直・瀬谷正敏・大川信明『心理教育 統計学』培風館, 1961.

調査法テキスト

- 宝月誠・中道實・田中滋・中野正大『社会調査』有斐閣, 1989.
- 浅井晃『調査の技術』日科技連出版社, 1987.
- 杉山明子『社会調査の基本』朝倉書店, 1984.
- 原純輔・海野道郎『社会調査演習』東京大学出版会, 1984.
- 直井優 編『社会調査の基本』サイエンス社, 1983.
- 安田三郎・海野道郎『改訂２版 社会統計学』丸善, 1977.
- 続有恒・村上英治 編『質問紙調査』東京大学出版会, 1975.
- 安田三郎『社会調査の計画と解析』東京大学出版会, 1970.

SAS のテキスト

- 奈良久・川添良幸・金井浩『SAS への招待』共立出版, 1989.
- 高橋行雄・大橋靖雄・芳賀敏郎『SAS による実験データの解析』東京大学出版会, 1989.
- 市川伸一・大橋靖雄『SAS によるデータ解析入門』東京大学出版会, 1987.

マニュアル類

- 『SAS ランゲージ・リファレンスガイド』Release 6.03 Edition, 1990.
- 『SAS プロシジャ リファレンスガイド』Release 6.03 Edition, 1990.
- 『SAS/STAT ユーザーズ ガイド』Release 6.03 Edition, 1990.
- 『PC 版 SAS システム導入及び運用と保守の手引き バージョン 6.04』1990.
- 『SAS Technical Report: J-115 メインフレーム版 SAS システム バージョン 6 使用の手引き』1990.
- 『SAS Technical Report: J-107 PC SAS 操作の手引き』1988.
- 『東北大学情報処理教育センター利用の手引き[第４版]』1989.
