

生物・社会調査のための統計解析入門：調査・研究の現場から(その1)

—統計解析の基礎知識(I)—

Statistical Analysis for Biological and Social Research (1)

—The basic knowledge of the statistical analysis (I) —

執行 盛之[†]

(SHIGYO Moriyuki)

講座の趣旨

農業土木学会誌編集委員会講座小委員会

近年、社会のボーダーレス化が進行する中、技術、学術分野とも、その境界領域はますます重要性を帯びてきています。とりわけ、農業土木分野においては、平成13年、土地改良法に環境との調和への配慮が原則化され、周辺領域と融合した新しい農業土木技術が求められているといっても過言ではない局面を迎えております。

農業・農村のもつ景観や生物等、生産以外の多面的機能が見直され、たとえば、農林水産省、環境省連携の「田んぼの生き物調査」の実施にみられるように、境界領域における活動が盛んとなってきています。このような活動をさらに充実させていくためにも、周辺領域である生態学や社会学との融合、また、領域間の意思疎通を図るインタープリターは欠くことができません。

現在、さまざまに行われている生物、社会調査結果を材料にその統計解析を進めていくことは、周辺領域との間のインタープリターとして有効であると考えられます。これが、本講座で「統計」を取上げる一つ目の理由です。

本講座では、農業土木技術者や農業土木を学ぶ学生を読者層と考えました。なぜなら、すでに各種調査結果を抱えている、あるいは、調査中であるが、その後の解析において統計に不慣れなため未解析の情報も多いと考えられるからです。これが、本講座で「統計」を取上げる二つ目の理由です。

農業土木分野では「統計」はまだまだ馴染みの薄い学問です。とくに、その理論的背景は、物理的に慣れ親しんだ農業土木技術者等にとってとっつきにくい一面を有しています。しかし、さまざまな統計解析ソフトウェアの普及が進んだ現在、もはや理論的背景の理解なしには統計手法を使えないという時代ではありません。むしろ、そのようなソフトウェアを用いて解析する中で、必要に応じて理論的背景の理解を進めた方が、「統計」に

馴染みの薄い読者にとっては有益であろうと考えます。

そこで、本講座では、生物・社会調査結果および統計解析ソフトウェアといった材料がすでに手元あるいは身近にあることを前提に、解析方法の選択、データセットの作成、解析の実施、解析結果の読取り方等に主眼をおいて解説していきたいと考えています。理論的背景は、必要に応じた極力簡単な解説にとどめ、参考文献等で示していきたいと考えています。また、読者が解析方法を選択する際、その助けとなるよう、各タイトルは極力解析目的を示すものとしました。

今回の講座では、生物・社会調査およびその解析を実施されている方々に、ツールとしての統計解析をご執筆いただけることとなりました。この場をお借りして、執筆をご快諾いただきました皆様にお礼申し上げます。

本講座を通じて、各種調査結果からより多くの情報を的確に得ることのできる技術者が増えることを期待しております。

講座構成

1. 統計解析の基礎知識(I)
(独) 中央農業総合研究センター 執行盛之
2. 統計解析の基礎知識(II)
(独) 中央農業総合研究センター 執行盛之
3. 2群を比較する
秋田県立大学 神宮 寛
4. 3群以上を比較する
(独) 農業工学研究所 小出水規行
5. ノンパラメトリック検定
(独) 農業工学研究所 若杉晃介
6. 因果関係を探る 長野県環境保全研究所 北野 聡
7. 判別する (独) 農業環境技術研究所 楠本良延
8. 分類する (独) 農業工学研究所 嶺田拓也
(独) 農業環境技術研究所 山中武彦
9. 総合化する (独) 農業工学研究所 合崎英男
(変更することもあります。ご承知おきください。)

[†](独)中央農業総合研究センター

I. 統計的手法への誘い

統計解析や実験計画法の勉強を始めるのに、まず書店へ行き、統計学の本を購入して、おもむろに机に向かい、振り鉢巻の勉強を始めることを、私は皆さんに勧めない。緊張した面持ちで統計的方法とは？ 正規分布とは？ と進めていくとすると、正規分布に基づいた検定と推定の入口辺りまでが、赤いサインペンでアンダーラインを引くことのできる限界で、分布によらないノンパラメトリック検定に取りかかるころには途方にくれている筈である。私共は、目の前の調査データをどう解析するかが当面の目的であって、数学と統計学そのものを究めようとしているわけではない。また調査データの解析に、統計の知識が絶対に必要というわけではなく、むしろ知らなくて済む場合だって多い。したがって、統計解析や実験計画法の解説書は、余程親しみやすい学問にしておかねば、誰も活用してくれなくなるであろう。

「統計」は大変有効な手法であり、理解してしまえば、研究者・技術者にとって、これほど頼もしい相棒は見当たらない。嬉々として勉強できるような統計の解説書はないだろうか？ 多くの人に利用してもらい、みんなが身近に感じるようになって初めて統計学は実学となり得る。もっともっと平易に書ける筈であり、書かなければならない。以降では、これまでに私が試みてきた経緯を紹介しながら、初心者を対象とした統計的手法への誘いとしたい。

最近解析ソフトが進化し、データの統計処理がパソコンで自在にできるようになった。おかげで統計学は、科学技術の分野においても確固たる地位を占めることができた。時宜を得たソフトの開発と解析結果のビジュアル化は、私から見ても大変魅力的である。こういう便利な時代だから、これを享受して存分に活用した方がよい。一方、解析する調査データには、欠測値があったり、繰返し数が揃っていないあったり、0・1データであったり、%データであったり、分類データであったりとさまざまである。このような現実のデータに対応するために、統計学自体もその幅を少しづつ広め、解析手法もますます多様になってきている。しかし、研究者・技術者とは、本来各自の専門分野を究めつつ、新たな知見を加えていくのが責務だから、すべての人に統計解析の専門家になれるというのも無理な話である。したがって、データの統計解析が一般化し、いろいろな手法が開発されるようになればなるほど、研究者・技術者は解析手法の理論を正確に理解しないまま、誤って適用している場合も見受けられる。そういうとき、解析の方法を穏やか

に指し示し、その方向に沿って具体的な解析手順を教えてくださいの指南書があれば、統計学は身近に感じられるようになる。

本講座では、①検定と推定、②多変量解析、③数量化理論などの解析手法が紹介される。これらは農業土木分野の裾野を広げるために、生物学や社会学の知見が急速に導入されており、これに対応するものとして選定された。統計学が全体として発展させている学問体系は広範である。概略すれば、①得られたデータをどのように解析するかという統計的検定法、②線形計画法などの数理統計学、③一対比較などの官能検査法、④直交表を代表とする実験計画法などの分野がある。本講座では、①統計的検定を主とする手法の解説を行う。

この講座では、研究者・技術者の目の前には、得られたデータが据えられており、「これをどう統計解析すればいいのか？」ということが主に説明される。紹介内容の一つである数量化理論は、わが国の林知己夫氏（前統計数理研究所長）によって独自に開発された。これは、刑期を終えて社会に復帰された方々が再び罪を犯さないようにするには、どのような社会的な配慮を必要とするかの提言を行う過程で案出された手法で、林式数量化理論と称することもある。

以降の講座では、Fisher や Tukey を始めとして外国の統計学者の名前を冠した多くの手法が紹介されよう。しかし、わが国の先人の貢献も素晴らしいものがあり、それらがいずれも社会の問題を解決し、また工業製品の品質向上に指針を与える実際的な統計的手法であることに、私達は誇りをもってよい。

II. 頼りになる統計学のホームページ

今やインターネットが自由に利用できる時代である。検索エンジン google で「統計 多重検定」を検索してみると、4千件余がヒットする。「チューキー」を加えると11件まで絞込むことができる。しかし、インターネットを利用して統計的手法を学ぼうと漠然と検索しても単に道に迷うだけである。そこで、大変頼りになる統計学のホームページを紹介する。

群馬大学社会情報学部青木繁伸教授が開設されている「<http://aoki2.sigunma-u.ac.jp/>」である。直接の面識はないが、優しい人柄が思われる楽しい内容はHP前半に配置されている。統計学と疫学を専門とされており、HP後半には授業内容と関連する魅力的な統計学のコーナーがある。「統計学自習ノート→検定・推定」を選択すると、関連する54の手法について例題と解析結果が説明してある。もちろん、搭載してある解析プログラム

も利用できる。また、「→多変量解析」の項では、回帰分析から数量化理論までが十分に説明してある。

そして、このHPのもう一つの魅力は、統計解析に悩むユーザーの質問に直接答える「統計学関連・・・何でも掲示板」コーナーが併設されていることである。質問の内容を適切に整理して送信すれば、解析の指針となる回答を期待できる。このコーナーで学んだ人も、回答者として質問に答える一種の互助会のような形になっているので、農学関連分野からの質問も散見されるし、回答者として農林水産省法人研究者も活躍している。

さらに、1999年から現在までに寄せられたすべての質問と回答がデータベース化されており、画面上でこの全文検索が手軽にできることもIT化の恩恵である。先ほどの「チューキー」で検索すると、5件が選択される。ちなみに、このHPへのアクセス数はすでに150万件を超え、インターネットの特徴を生かして統計学を普及させようとする新しい試みは成功裏に進められている。言葉を変えれば、社会が分かりやすい統計学を求めていることの反映でもある。

III. 統計解析手法を学ぶポイント

統計学の書籍よりも魅力的なHPを紹介した。しかし、統計を学ぶのに、このHPで説明されている一つ一つの解析手法を順に修得していけばよいのか？と言えば、そうではないと思う。統計学を専門とする分野も少しづつ研究の幅を広げているわけだから、解析手法の懐も深い。一つの手法、次の手法と学び始めても、広範な内容をもつ統計学の路頭に迷いかねない。冒頭に書いたとおり、私共農業に関わる研究者・技術者は、自らの専門分野を究めつつ新しい知見をつけ加えていくのが社会的責務である。得られたデータを解析することを目的に統計手法を利用しようとしているのであるから、すべての解析手法に通暁する必要もないだろう。

そこで、「こういうデータの場合には、次のような手順で解析手法を選択して下さい」という選択経路を示した整理図があれば、初心者であっても誤りなく適切な手法にたどり着くことができる。過去にも多くの人がこのような整理図の作成を試みているが、なかなか難しい。しかし、本講座は、「生物・社会調査のための統計解析

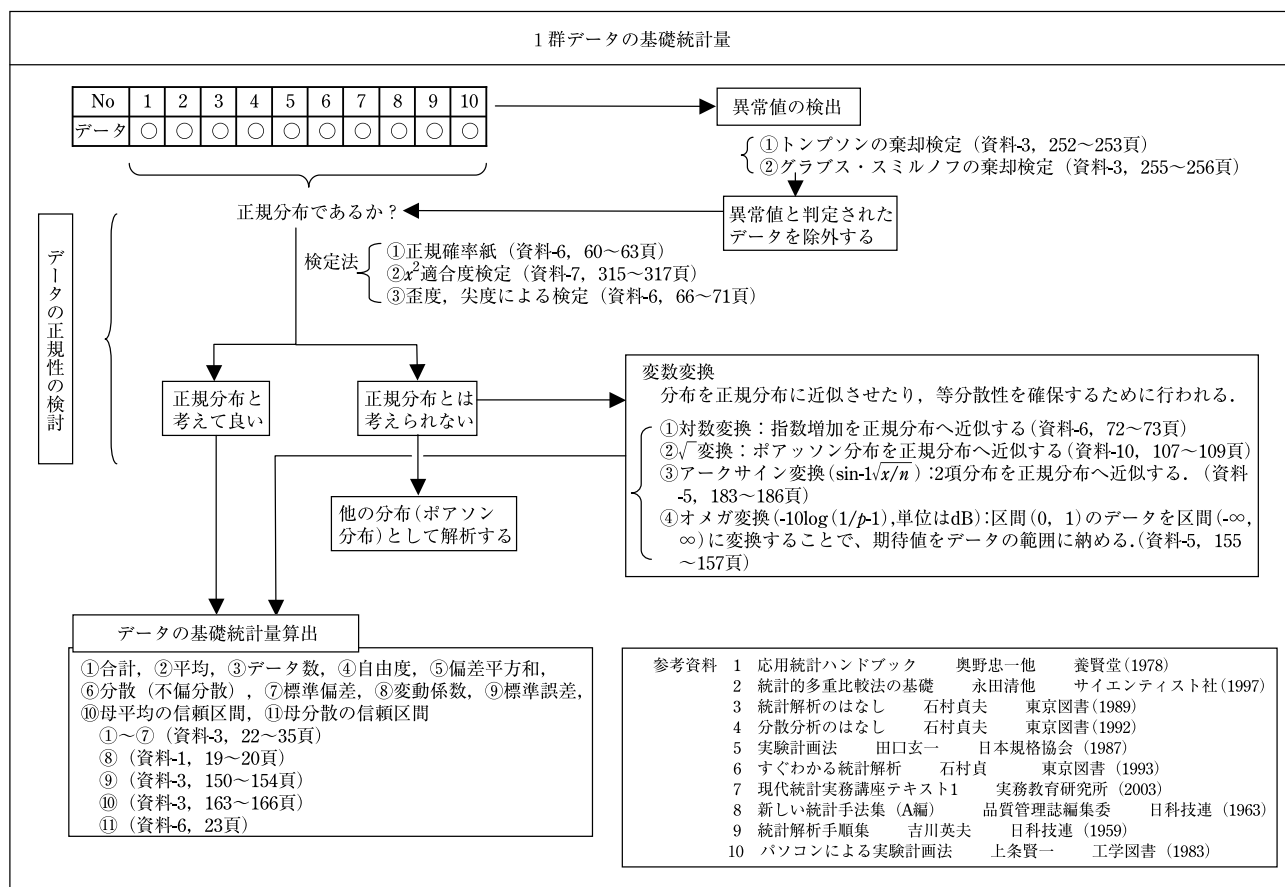


図-1 1 群データの基礎統計量

入門」という内容であるから、これまでの私の経験を基に、(独)農業環境技術研究所・生態システム研究グループと昆虫研究グループの協力を得ながら、データ構造を異にする次のような3つの場合の選択経路整理図を作成した。

①1群データの基礎統計量では、一般に正規分布への適合度検定・変数変換・異常値の検出を経て、データの基礎統計量が算出されることになるので、その選択経路を示した。初心者の理解を深めるためにおのおのの手法には参考資料とそのページ数も付した。また、先ほどのHPの記載内容も大変有効である(図-1)。

②2群データの解析と検定では、データに対応が有るか無いかを見極めて解析手法を選択するので、この点が理解できる整理図とした(次講の図-4)。

③3群以上データでは、どの群とどの群の間に有意差があるかを検定する多重比較までが解析手順に含まれる。ここでは、データが等分散の場合の多重比較法と、等分散でない場合のノンパラメトリック多重比較法を選択する経路を整理した(次講の図-5)。

統計学の最新の研究成果までを網羅した整理図とは言えないが、統計的解析手法を学ぶ道標として十分な内容を持たせている。

IV. 栃木県鬼怒川流域のメダカは？

II. では、隅々にまで配慮の行き届いた統計学のHPを紹介した。続けて III. では選択経路図を示し、併せて参考書籍のページ数も付した。これで、一通りの解析手法を概観できるようになり、統計的な解析手法を利用できる環境になった。だが、だからと言って「使いこなせる！」という段階には程遠い距離にある。自分の専門分野である農業工学や生物学の研究と技術開発の場面で、統計手法を活用するには、今少しの努力と精進を覚悟した方がよい。

統計学の研究者はさまざまな手法を開発してくれているが、すべてのデータに万能な解析手法は無い。データに制限がある中で、数学的な厳密性を保ちながらより良く解析できる考え方と手順を提供してくれているだけである。たとえば、2群データの場合、対応の有るデータと対応の無いデータで解析手順が異なるのは、対応が有る・無いというおのおのの条件下で最も有意差を検出しやすい手法を案出するためである。また、等分散が保証されないデータの場合でも、平均値の差が比較できるようにノン・パラメトリック検定の諸法が確立された。多重比較法として一般的な Tukey の方法では、データが同数であることが前提となるが、データ数が異なる場合

もあろう。その場合に備えて、Tukey の方法の発展形として Tukey-Kramer の方法が開発された。したがって、ユーザーが農業工学や生物学という自らの専門分野の中で、提供された統計手法の特徴を活かした利活用を図らなければならないということになる。

さて、ある研究者が栃木県鬼怒川流域に設定した現地試験に赴き、終日メダカ生息数の調査を行った。立地の異なる4条件下で調査を続けているが、日暮れも近づいたので心が急いてくる。A・B・Cの条件下では10データを得たが、これに時間を取られ過ぎて最後に行ったD条件の試験区では未だ8データしか得られていない。D条件でも10データが得られれば、多重比較には Tukey の方法を適用できる。しかし、8データでは、4条件でデータ数が揃わないので Tukey-Kramer の方法となる。もう少し粘って2データを取るべきか？、帰途を急ぐべきか？ このとき Tukey と Tukey-Kramer の検出能力の違いを理解していれば、調査を粘って10データを揃える方が得策と判断がつく。統計解析を適用するには、データを取る段階から解析を念頭においた調査を行う姿勢が求められる。これが統計手法を活かす方策である。単なる手持ちデータを解析することで、目的とする解析結果が得られることは少ない。

この意味で、統計解析は場数を踏むことで理解が深まるとも言える。しかし、1回目の調査データを統計解析したら処理間に有意差はなかったものの、4試験区を概観するとメダカの数には確かに違うようである。そこで2回目の調査を行って解析したが、それでも差が認められない。さて、このような場合には、どうすれば良いのだろうか。いろいろな対応策が考えられる。①異常値があって、データの前提条件を満たしていないため、解析が円滑に進まないのではないかと？ ②別の解析手法を適用すべきデータかも？ ③メダカ生息数の他にも、別の調査項目を加えたらどうか？ ④調査期間を短くして、1週間単位の変化をみたらどうか？ ⑤条件が異なってもメダカの生息数に影響がないのなら、この調査は中止して別の方向に試験を発展させたらどうか？

試験の途中では、研究の発展や次善の策を何時も楽しく考えているので胸躍らせて調査を進めることができる。「どうしようかな？」と戸惑っている時こそ、統計手法は研究者に指針を与える良き相棒でなければならない。このように考えると、おのおのの統計解析手法の適用データには前提があって、限界のあることに立ち戻らざるを得ない。すなわち、自らの専門分野を十分に意識した上で、解析手法の特徴を理解を深めておく必要がある。したがって、統計解析手法を使いこなす、自らのも

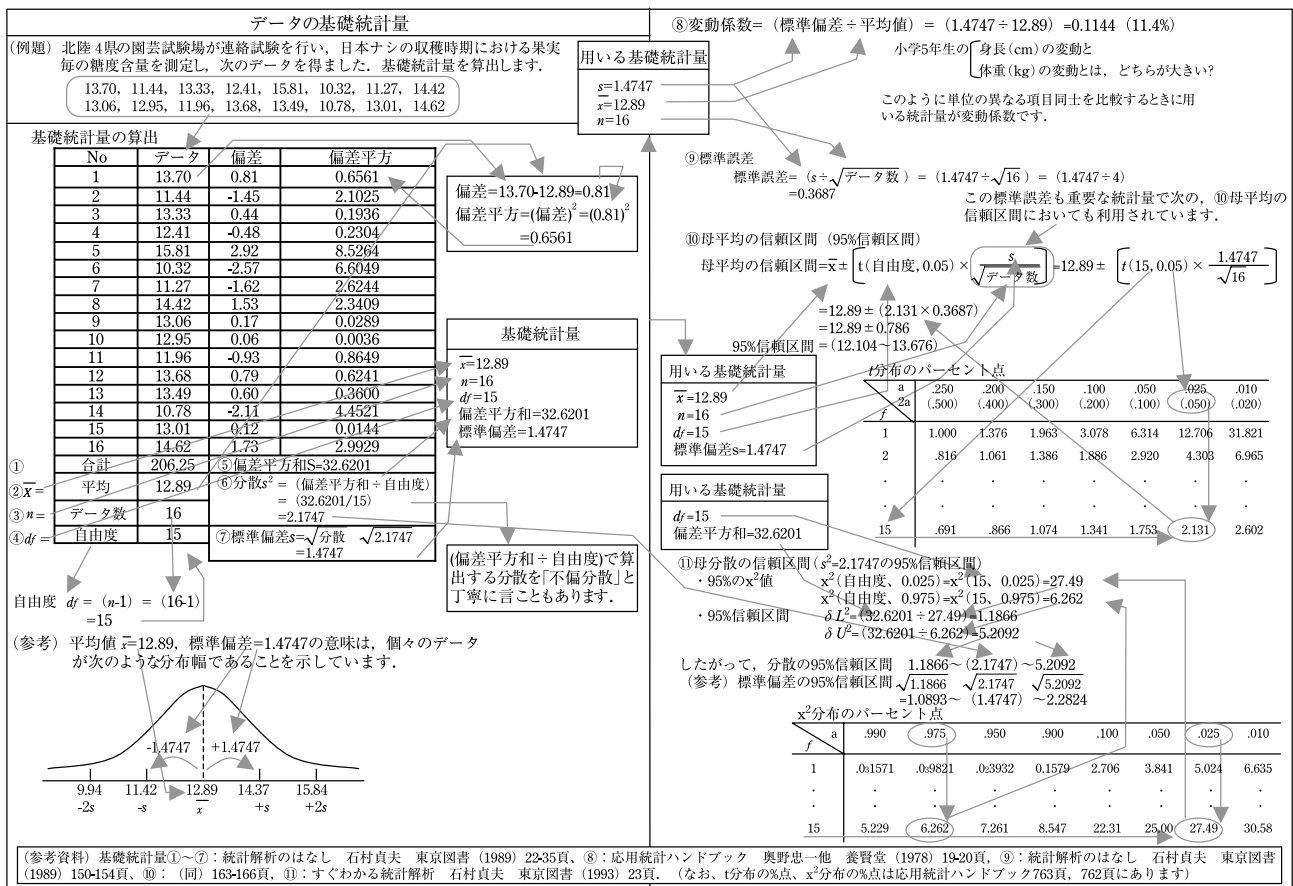


図-2 統計解析メモの例示「データの基礎統計量」

のとするには今少しの努力と精進を重ねよう。

V. 私から勉強方法の提案があります

私にも、統計学やデータ解析上の疑問があると必ず尋ねる友人が居る。雑談のなかで次のようなことを聞いた。友人が、アメリカで開催された統計学に関する国際学会に参加して、ある研究所を訪ねた。この研究所には実験計画法や統計解析の専門部署があって、所員は試験設計の段階で相談に来て、またデータ解析の相談にもやってくる。研究所が所掌する分野で十分な経験を積んだ統計学の専門家を所内に常駐させ、所内研究に対応させると同時に、彼らは所を越えた統計学のネットワークで更なる研鑽を積んでいるという。しかし、わが国でこのような恵まれた環境がすぐに実現するとは思われない。鬼怒川流域のメダカ試験のように、試験途中で立往生をしなくて済むようになるには、研究者・技術者が自ら学んでいくしかないようである。

ここで、私が提案するのは、学問に近道は無い、あるいは急がば廻れという古い諺どおりの地味な勉強方法である。それは、自分なりに統計解析のメモを作り、しか

も電卓を用いても解析できるように算出手順も明確に示した整理を地道に積み上げていくことである。この方法を皆さんに勧める。実例として、図-1の1群データの基礎統計量の中で「データの基礎統計量」に係わる私の実際のメモを、図-2に紹介する。

この統計メモの作成に当たり、私は次のような配慮を施している。

①用紙は、A3版の見開き(A4版2枚)としている。なるべく広い用紙を使えば、解析手法の全体像が一目で理解できるからである。

②電卓でも解析できるよう、順次引用していく数字には矢印で引用先を示している。こうしておけば、何時でも解いても悩むことなく参照できる。

③3色のボールペンを用いて作成している。基本的な計算の流れは黒ボールペンで書く。数字の引用先は赤矢印で明確に示している。コメントや参考になる知識は、余白に青で囲み書きする。

④関係する統計表(部分)もコピーして同じページに入れておく。こうすれば理解の流れも良くなる。

⑤引用文献の記載欄も付けて、ページ数まで明記して

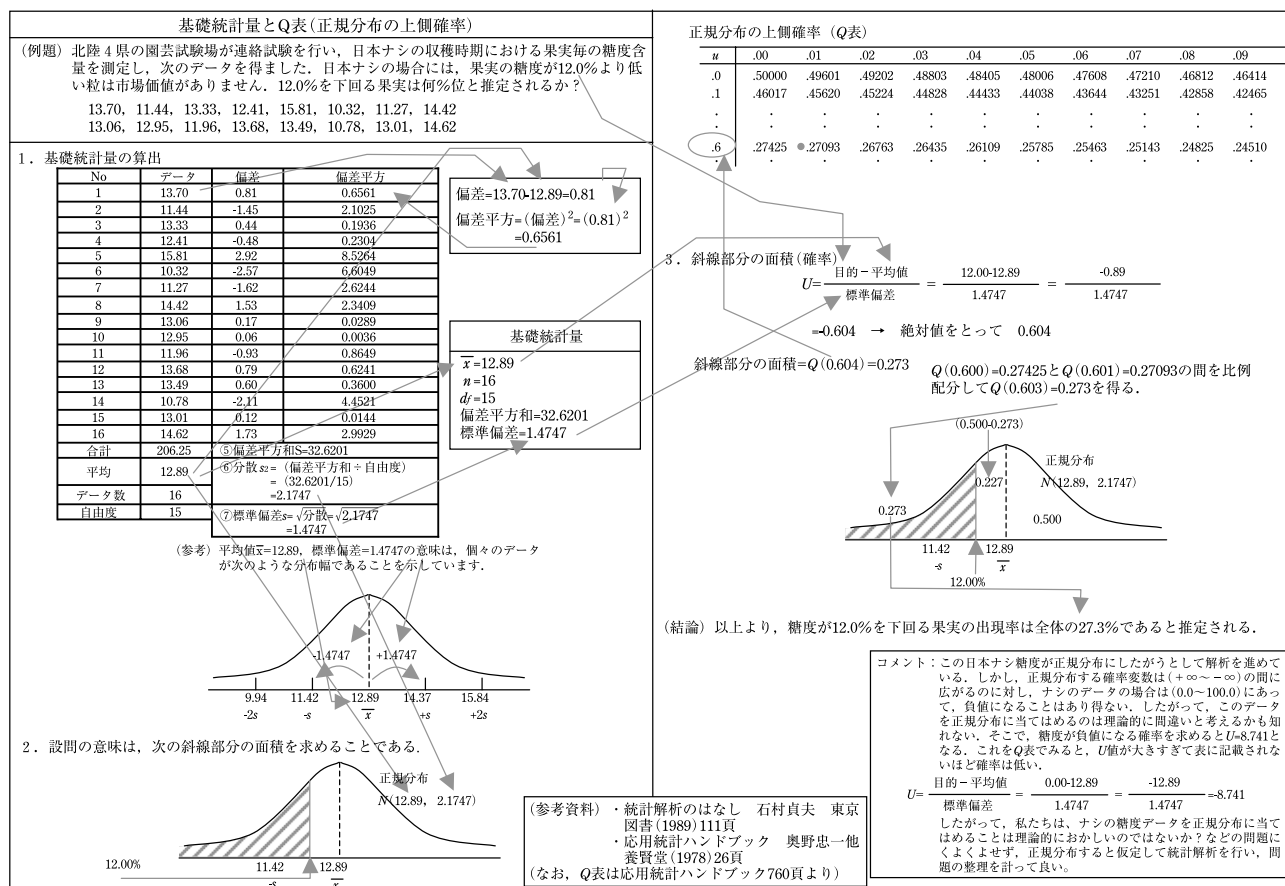


図-3 「基礎統計量とQ表」

おく。

このようなメモの作成により、「基礎統計量の算出」手順は確実に自分の手法となり、何時でも引き出せる状態に整理できる。一人でコツコツと勉強を重ねてもよいが、長続きしないようなら若手研究者が統計ゼミに集まり、分担しながらメモ作りを楽しむ方法もある。

VI. 統計解析のメモは発展する

実は、先に示した日本ナシのデータ(図-2)には、次のような解析の目的があった。果実の糖度が12.0%より低い粒は甘みに欠けるということで、市場価値はなく、選果場の糖度測定機によって弾かれて廃棄となる。得られた糖度データから、収穫した果実のどの程度が糖度12.0%を下回って廃棄になるか、を推測したい。この場合の私のメモを紹介する。

さて、このような解析手順を詳細に整理したメモを作ったことで、何を明らかにできただろうか? まず、日本ナシのデータから糖度が12.0%を下回る果実の出

現率は27.3%であると推測でき、データが示唆することを具体的に理解できた。したがって、「ナシの糖度は、12.89±1.4747(平均値±標準偏差)」と概要書に記載しておけば、別の技術者でも果実の出現率を容易に算出できる。

次に、どのような統計表においても必ず冒頭に載せてある「Q表:正規分布の上側確率」(図-3)の重要性もわかった。Q表とは、こういう場合に利用される基本的な値を示したものである。正規分布は左右対称の広がりをもち、上側確率さえ示しておけば下側の確率も同じ値をとるので、両側確率の計算も単純になる。メモ中にQ表を加えることで、Q表の意味合いが明確になり、全体の理解を容易にする。メモ前半の基礎統計量算出の部分は図-2と同じであるが、一話完結型のメモとしておくことで後々引用しやすい。書籍には、コメントで示したような関連知識も多いので、理解したときに丹念に記載しておけば、メモの完成度は格段に高まる。

[2004.6.2.受稿]