ビジネスにおける初期的データ解析の重要性

東　渕　則　之

目次

1．はじめに

統計分析を経営活動に適用しようとしても，自然科学の場合のようにはうまく行かず，利用の度合いも低調であるのが実態である(東渕，1992)。その原因は，ビジネス分野の統計分析環境において，統計分析と経営の間にギャップ・懸隔が存在しているからである。その懸隔は3つに分けられ，①統計手法における問題，②データの問題，③利用者側の問題からなっている(東渕，1993)[1]。本稿では，このような問題を解決するために，ビジネス分野において有効な統計手法はどうあるべきかについて，一つの提案を行う。

2．ビジネス分野で統計分析が満たすべき条件

統計分析が使われる環境は自然科学分野の場合とビジネス分野の場合ではかなり異なる。自然科学分野では，統計手法の適用前提条件が比較的満たされていたり，データの品質もある程度の水準をクリアしていることが多い。それに対して，ビジネス分野では，適用前提条件が満たされにくく，データの質もよくないことが多い[2]。また，統計分析結果の使われ方にも差がある。自然科学分野

ではデータやその分析結果は，第一に尊重されるべきものである。それに対して，ビジネス分野では，統計分析情報は，定性情報・事実情報，経験，直観などと同列に位置するものであり，必ずしも第一に重視されるものではない。これらからわかるように，自然科学分野とビジネス分野では統計分析に求められるものはおのずと異なってくる。以下では，ビジネス分野で求められる統計分析の性質について考える(東渕，1995)。

①ある程度の精度・正確さ

統計分析では，絶対的な正確さはめったに達成されるものではない。とくにビジネス分野の統計分析では絶対的な正確さは必要とされない。その理由として，一つには，ビジネス分野の統計分析情報においては，もとのデータは対象の数値化できる側面だけを表しているにすぎず，分析結果はあくまで定性情報・事実情報，経験，直観と統合され意思決定に変換されるものであり，仮に手にしたデータについてのみ100%正確な分析結果が得られたとしても，それだけで完璧な決定が行えるというものではないからである。もう一つには，分析に用いられるデータの質がビジネス分野では必ずしもよくないため，分析の段階で正確さにこだわりすぎても意味がないからである。したがって，いたずらに精度・正確さを求め，高度な手法をこだわりすぎるのは意味がない。

②迅速さ

ビジネス分野では，一刻も早い意思決定が求められている。統計分析結果が１ヶ月後にならないと出ないようではほとんど使い物にならない。ある程度の精度が確保されていれば十分であり，今の今，すぐに分析結果が欲しいという場合が大半である。[3] このようにビジネス分野の統計分析では迅速さが極めて重視される。迅速な分析が可能な手法が望まれる。

③頑健さ

ビジネス分野のデータの質は一般によくないため，分析結果も，データにはある程度の誤差があることを見込んで解釈する必要がある。つまり，データの誤差や外れ値があっても，分析結果がそれほど外れないような分析手法，ある

いはそれが人間にわかるような手法が望まれる。このことはデータの質からだけでなく，統計手法の前提条件の面からも言える。パラメトリック手法を中心として統計分析ではデータ数や変数の独立性等の前提条件が課せられており，それをクリアしたとき適用が可能となる。しかし，現実には，必ずしもこれらが確認されないまま，あるいは確認する術がないまま分析が行われることが少なくない。したがって，分析手法には前提条件が緩いか，あるいは前提条件が満たされていない場合であってもそれほど外れた結果にならないような頑健な分析手法が望まれる。

④分析結果の理解しやすさ

統計分析情報が意思決定者に採用され，意思決定にその情報が生かされるか否かは重要である。意思決定者は自分の意思決定に責任を負う以上，自分が理解していない手法や信じていない方法で出された分析結果を積極的に使おうと言う気にはなれないであろう。ビジネス分野では，分析を他者が行おうと自分で行おうと統計分析結果を使って意思決定するのはビジネスマンである。ビジネスマンの統計分析の知識は一般に高くはない。したがって，ビジネス分野で用いられる統計分析手法は，その結果の解釈に苦労するような高度・複雑なものは不適当であると言える。

⑤分析の容易さ

前項でも述べたように，統計分析を行うのも分析結果を解釈し利用するのもビジネスマンであり，ビジネスマンの統計分析の知識は一般には高くないと思われる。よって，組織内で統計分析が有効に活用され普及するには，そこで利用される分析手法は適用が容易なものであることが望まれる。

⑥実施コストの安さ

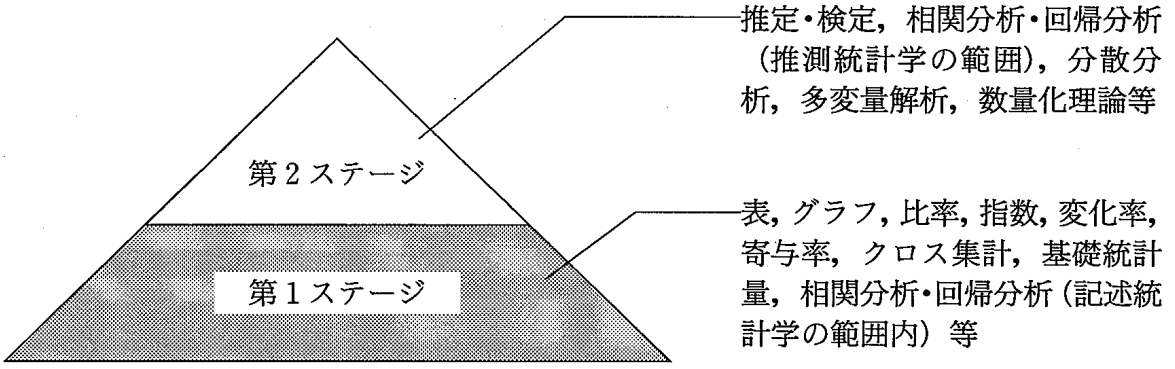
実施コストには，人的，時間的，金銭的なコスト，及び分析に伴う困難などが含まれる[4]。統計分析は，それによって得られるベネフィットがそれを実施するコストを上回る場合に利用されるので，実施コストが安いことが，統計分析の利用を促進することになる。よって，ビジネス分野の統計分析としてはこれ

らのコストが安い手法が望まれる。

3．統計分析の2つのステージとビジネス適合性

統計分析の手法は第1ステージ，第2ステージの2つに分類できる。第1ステージの手法は，加減乗除で実施できるような比較的簡易な分析手法であり，具体的には，表，グラフ，比率，指数，変化率，寄与率，クロス集計，各種基礎統計量，記述統計学分野の相関分析・回帰分析などである。これらの手法は，生データをわかりやすく整理・記述してくれる（それを見てデータの持つ意味，すなわち情報を抽出するのは人間の役目である)。①比較的表面的な情報しか抽出できないものの，データが持っている情報のかなりの部分を抽出することができ，ビジネスの場面では精度や情報等の面で十分である場合が多い。②分析のやり方が簡単であるため，誰でも手軽に迅速に実施できる。③データの加工・処理が単純であるため，途中がブラックボックス化せず，データの質の悪さの影響が深く潜在し結果を大きく歪めているにもかかわらずそれに気がつかないという危険も少ない。適用に当たっての前提条件も比較的緩やかであり，分析結果が大外れすることも少ない（グラフ等を描いてチェックすれば大きな火傷はしない)。④同じく，データの加工が単純であるため，分析結果の意味するところも理解しやすい。結果がわかりやすいということは，定性情報・事実情報，経験，直観と脳裏で統合するとき都合がよい。⑤分析も容易である。⑥表計算

図表1．第1ステージと第2ステージの統計分析



ソフトがあれば実施できコストも安い。このように第1ステージの分析手法は，第2節で述べたビジネス分野で統計分析に求められる性質である「ある程度の精度・正確さ」「迅速さ」「頑健さ」「分析結果の理解のしやすさ」「分析の容易さ」「実施コストの安さ」を満たしていることがわかる。言いかえると，人間の身の丈に合った分析手法，「等身大の統計分析手法」であると言える。

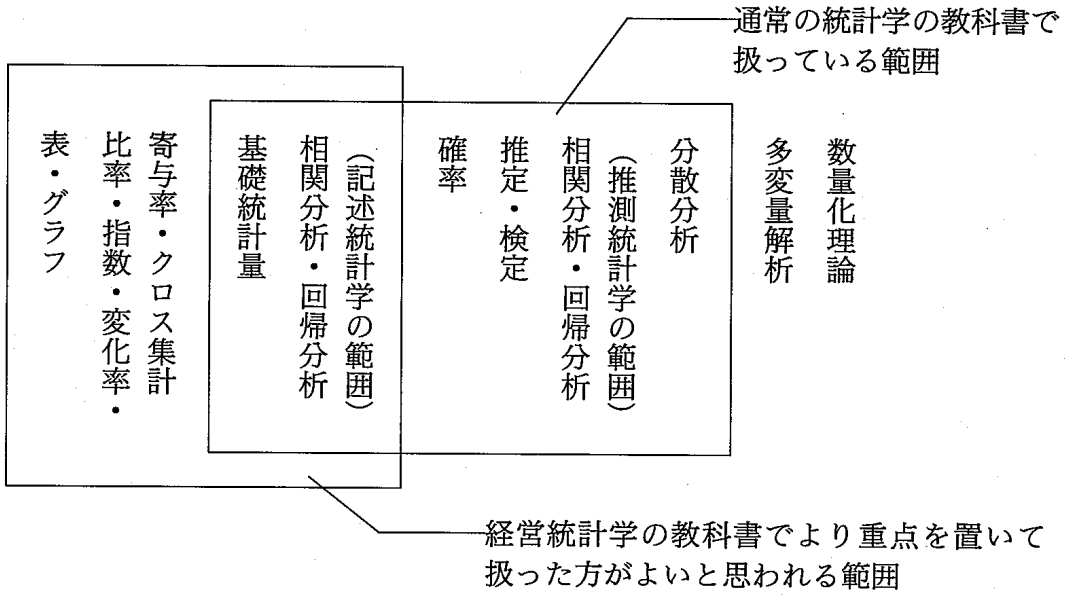
それに対して，第2ステージの手法には，推定と検定，推測統計学分野の相関分析，回帰分析，分散分析，多変量解析，数量化理論などより高度な解析手法が相当する。第2ステージの手法は，データからより深く複雑な情報を抽出することを目的としている。しかし，その適用に当たっては一般に専門的な分析能力と解釈能力が必要であり，また，データ数やその分布条件など適用前提が比較的厳しいものが多く，第2節の議論をもとに考えるとビジネス実践には必ずしも向かない（もちろん，条件が満たされていれば，ビジネス場面でも有効に活用できる。）

以上から，ビジネス分野では，第1ステージの分析手法の方が第2ステージの手法より適していると言える。第1ステージを，まずデータを前にしたとき行う分析という意味で「初期的データ解析」と呼ぶことにするが，先ほど述べたようにこれで十分な場合が多い。それに対して第2ステージの手法は「発展的データ解析」と呼ぶことができる。発展的データ解析も，もちろんデータ面や手法面，利用者側の能力等の面での諸条件がクリアされれば，ビジネス場面でも有効に活用されうるし，これらを適用しないと見えてこない情報もたくさんあることは言うまでもない。

4．初期的データ解析のいくつかの例

初期的データ解析の手法のうち，各種基礎統計量，記述統計学の範囲での相関分析・回帰分析は通常の統計学の教科書で必ず扱われているが，グラフ，比率，指数，変化率，寄与率，クロス集計については，あまり扱われていない。これらの手法は，数学的な理論展開が必要ないため数学的な展開を旨とする通

図表2．経営統計学で扱う統計手法



常の統計学の教科書で扱われていないのかもしれない。しかし，実際にビジネスデータを分析する第1歩としては強力な武器になりうる。その意味で，経営統計学の教科書の有り方の一つとして，扱う手法を，次の図表のようにシフトさせることも考えてよいであろう[5)]。

もちろん，これらについて特別に注意すべき点がなく常識で分析が進められるのであれば，教科書で扱うこともないかもしれない。しかし，これらの初歩的に見える加工・分析でも，事前に知っておくと有益なこと，あるいは知らないと思わぬ落とし穴に入ってしまうことがよくある。それらの例は，Neter & Wasserman (1961)，古寺(1980)，上田(1981)，Ehrenberg (1982)，林(1993)に詳しい。詳しくはこれらの文献を参照されたい。ここでは，データが持つ情報を正確に抽出する上で初期的データ解析の注意事項が案外知られていないことを示すため，「データの視覚化」と「データの簡単な加工」に分け，例をいくつか簡単に紹介する。ちなみに，筆者もデータ分析に熱心なビジネスマンに話をする機会を時々持つが，これらのことは案外知られていない。なお，これらの大半は上田(1981)からの引用である。

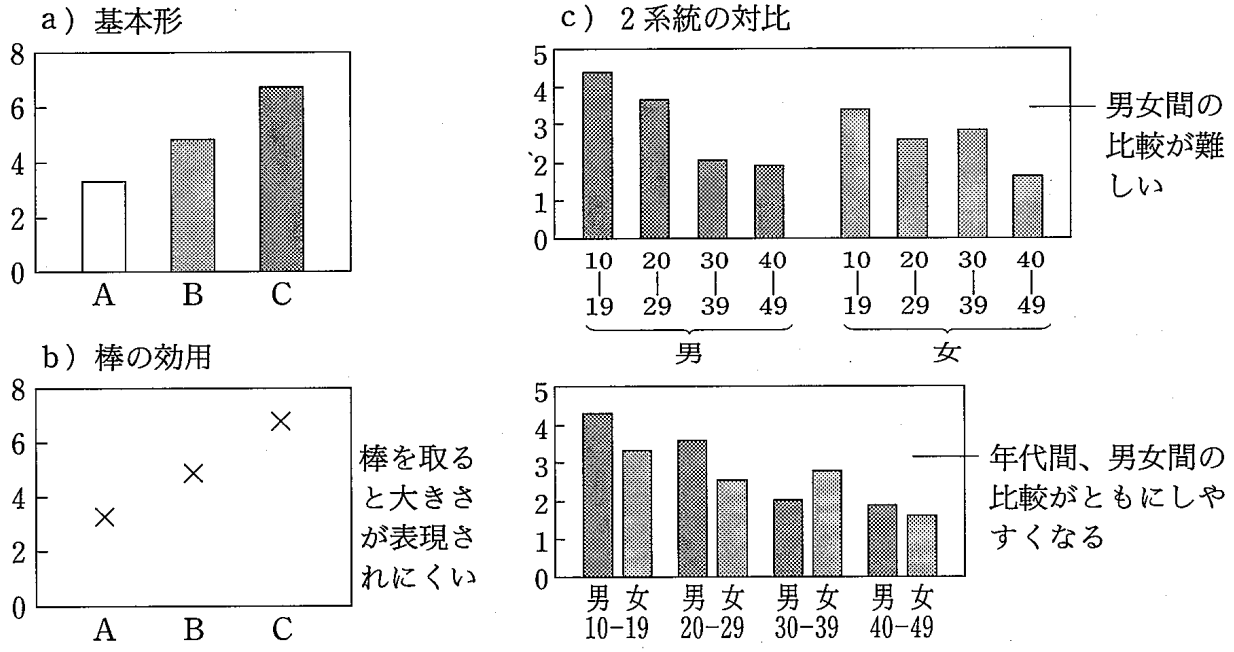
(1) データの視覚化

生データから情報を抽出する際，まず行うのはデータの視覚化である。グラフに表現することによって，データの概要が把握でき，分析方針を立てることができる。場合によってはデータのミスを発見することもできる。また，グラフは分析した結果を表現するためにも有効な手段である。

① 棒グラフ

数値を対比するための基本グラフである。棒を取り去ると大きさが印象づけられない。2系統の棒を組み合わせるときは，区分数の少ない方を隣接させると2系統の対比が共にしやすくなる。

図表3．棒グラフ

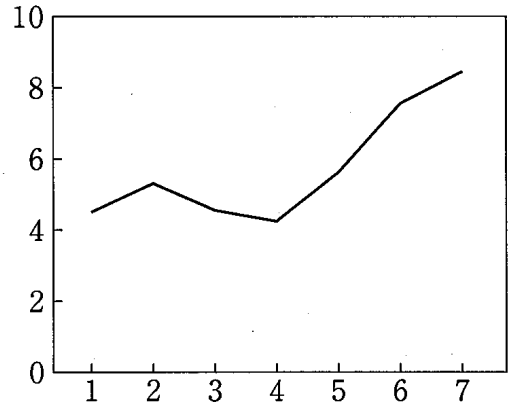


② 線グラフ

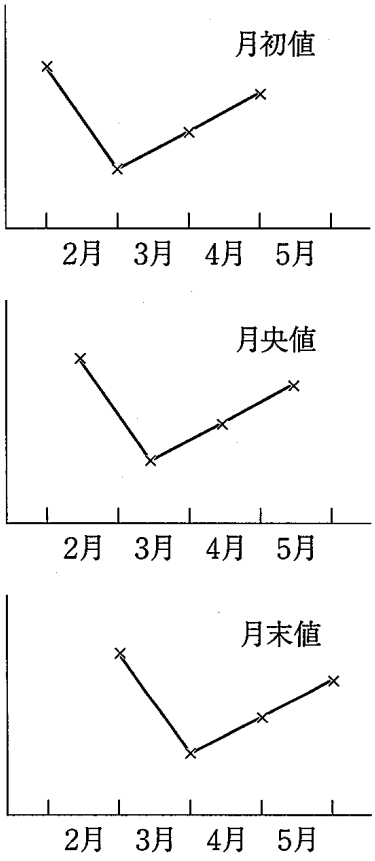
時系列データを表現する基本グラフである。線に沿ってデータの推移を読み取ることができる。2種類の線を重ね合わせる場合には，それぞれの軸の0点を一致させる必要がある。また，点を打つ位置は，ストックの場合，月初，月央，月末などの区別が必要である。

図表4．線グラフ

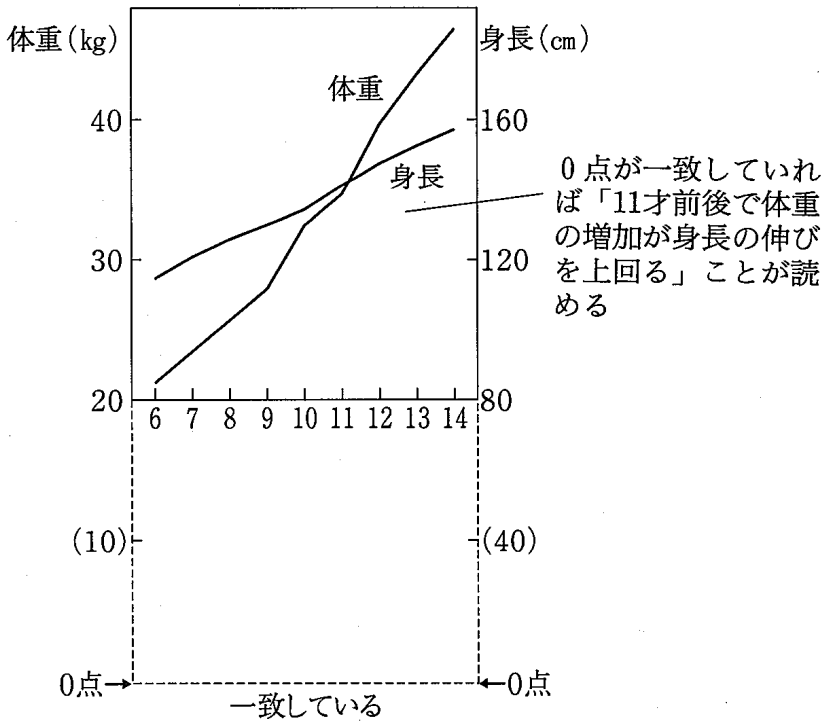
a）基本形



b）点を打つ位置



c）2系統の重ねあわせの場合は
0点を一致させる

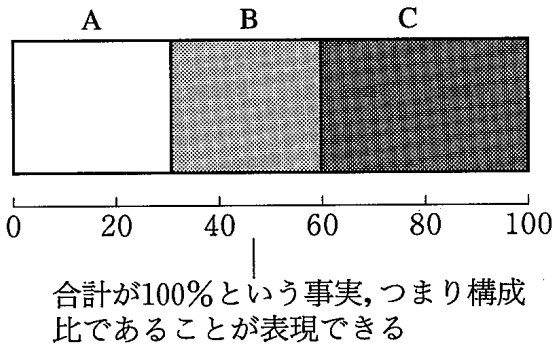


③　帯グラフ

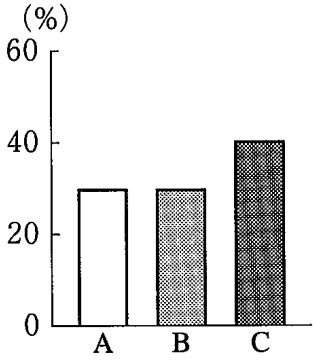
内訳・構成比を表現するための基本グラフである。棒グラフで構成比を表現するのは好ましくない。構成比を2つ以上の対象について比較する場合には最適である。

図表 5．帯グラフ

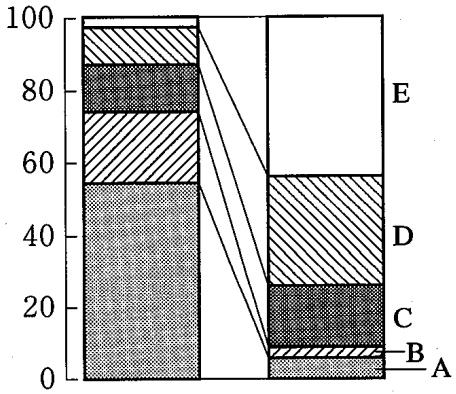
a）基本形



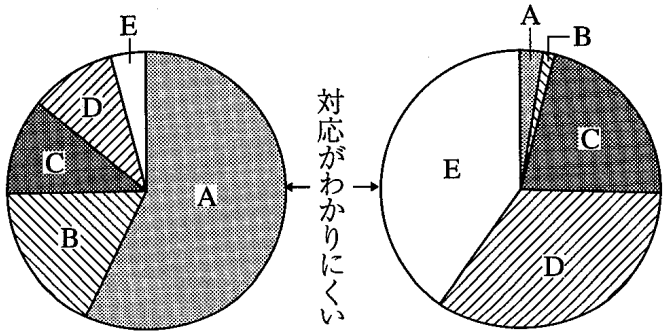
b）棒グラフでは構成比という事実が表現できない



c）帯グラフは対比もしやすい



d）円グラフは対比に向かない

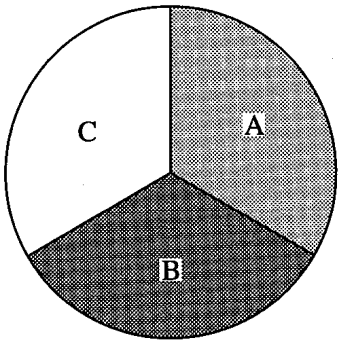


④　円グラフ

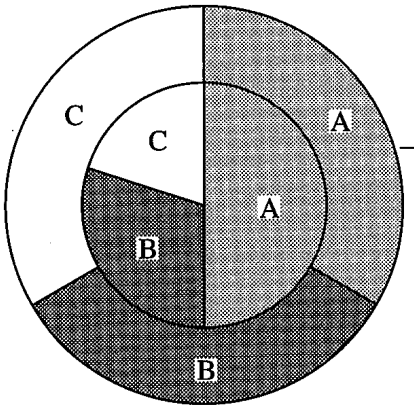
帯グラフと同じく，内訳・構成比を表現するための基本グラフである。ただし，構成比を 2 つ以上の対象について比較する場合には不適である。

図表 6．円グラフ

a）基本形



b）2 つの同心円で対比させるケース

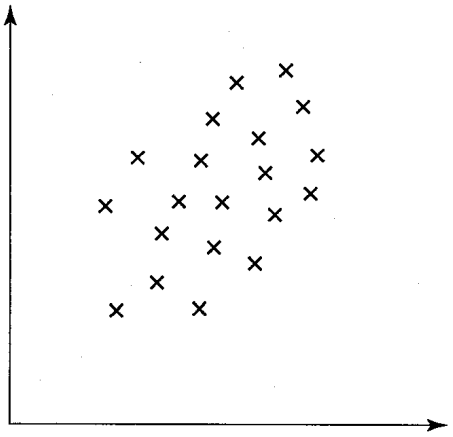


⑤　散布図

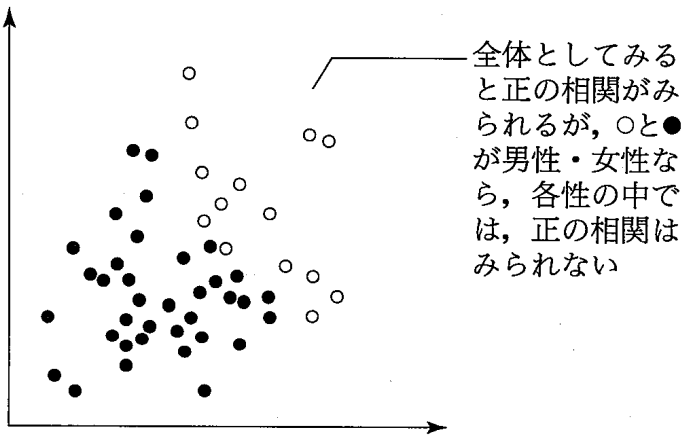
2つの変数をx軸y軸にとり，データをプロットし，変数間の関連性を表現する。複数のグループが混在している場合には，それぞれのグループ内で関連性を読まなければならない。

図表7．散布図

a）基本形



b）異質なグループからなる相関図は
みせかけの相関しか表現しない



(2)　データの簡単な加工

第1ステージの分析手法の中から，ここでは比率，指数，クロス集計を取り上げ，問題形式で間違いやすい点を指摘する。

①　比　率

比率はサイズ効果を除いて比較するための加工手法である。サイズ効果を除くには，分母として対比しようとする集団のサイズ，分子として対比しようとする指標を当てはめて分数を作ればよいが，何を分母にとるか，そう簡単ではない。

(問)「人口当たりの高校数を市町村別に人口で割って対比することは妥当だろうか？」

「人口当たりの児童公園数を県別に人口で割って対比することは妥当だろうか？」

(答)「まず，高校は市町村の枠を越えて通学しあっているので，市町村より広い範囲で比較するべきであり，分母はもっと広くとる必要がある。児童公園は，子供が気軽に利用できる範囲にないと意味がないので，数百メートルの範囲で比較すべきであり，分母はもっと狭くとらないと利便性が表現できない。」

(問)「大病院の過不足を対比する指標として，分母に人口をとるのは妥当だろうか？」

(答)「必ずしも妥当でない。例えば，10万人あたりに一つあると言っても，大都市では多すぎるし，田舎では広い地域になり少なすぎる。病院から○○キロメートル以内の範囲の人口，という形で人口に面積の要素を組み込んで分母とする必要がある。」

② 指　数

指数は，基準時点の値に対する比率のこと。評価時点の値を基準時点の値で割って100を掛けて求める。基準時点における絶対値の大小は捨て去られる。よって，指数では絶対値の大小は読めない。

(問)「2つの消費財A，Bの価格指数（昭和60年＝100）は次のようになっている。これについて，Bの価格はAのほぼ2倍で推移している，という説明は正しいだろうか？」

図表8．2つの消費財A，Bの価格指数

	平成1年	3年	5年	7年	9年
A	120	140	150	150	160
B	250	280	300	310	320

(答)「正しくはない。基準時点でのA，Bの価格が不明だから。」

(問)「次の表は，2つの地点での鶏卵の値段の推移を指数で示したものであるとする。これを見たa，bの2人が次のようにコメントした。a…A市と

比べてB市の方が高い。b…A市とB市の価格差が開いた。彼らのコメントはそれぞれ正しいだろうか？」

図表9．鶏卵の値段の推移

地域＼時点	0	1	2	3
A　市	100	102	105	106
B　市	100	108	110	114

(答)「基準年の鶏卵の値段がわからないから，B市の指数がA市より高くとも，値段が高いとは言えない。よって，aは間違い。bは，次のような場合があるから間違い。例えば，A市200円→212円，B市180円→205円。この場合，B市の指数がA市より大きくなるが，価格差はむしろ小さくなる。」

③　クロス集計

あるカテゴリーにサンプル数を別の軸のカテゴリー毎に集計することをクロス集計と言うが，これにより2つ(以上)の軸の関連性を抽出できる。クロス集計の解釈に際しては，混同要因の存在に注意すべきである。

(問)「男性ドライバーと女性ドライバーを比べ，どちらが交通事故を起こしやすいか調べるために，運転免許者台帳によって選んだサンプル250人について，過去1年間に事故を起こしたことのある人数を調べたところ，男性は200人中50人，女性は50人中10人であったとする。この結果から男性の方が女性より交通事故を起こしやすいと結論してよいだろうか？」

図表10．男性ドライバーと女性ドライバーの事故件数

	事故あり	事故なし	事故率	計
男　性	50人	150人	25%	200人
女　性	10人	40人	20%	50人

(答)「よくない。サンプルは運転免許者台帳から選んだ，とあり，自分では車を持っていない人やめったに運転しない人も含まれている可能性がある。よって，男女差の他に，走行距離の長短に起因する差があり，これらが合わさったものがこの表の結果である。事実，この調査で，走行距離の長短で250人を分類し，長いグループ，短いグループで，それぞれ男女の事故率を求めるとほぼ同じであった。」

5. むすび

本稿では，ビジネス分野で統計分析に求められる性質として「ある程度の精度・正確さ」「迅速さ」「頑健さ」「分析結果の理解のしやすさ」「分析の容易さ」「実施コストの安さ」が重要であることを述べ，それを満たす手法として初期的データ解析が適合することを述べた。一般にビジネス実践ではこれらの手法さえ十分に活用されておらず，まして，第4節で例示したようなポイントは認識すらされていないのが実態である。情報技術の発展により，企業ではデータウェアハウスのような大規模なデータ倉庫が実現しつつある。そのデータ資源を生かして使えるか否かは，企業の業績を大きく左右する。近年，その分析手段としてOLAPやデータマイニングが注目されるようになってきているが，その基礎的部分を支えるのが初期的データ解析であるとも言える。その意味で，ビジネスマンに初期的データ解析を中心とする統計分析の知識が普及することが望まれる。

注

1）東渕(1993)では，経営と統計学との間の懸隔例として以下のようなものをあげている。

①統計手法における問題…・前提条件が経営の場面では満たされにくい
・統計処理の手間や費用が小さくない

②データの問題……・経営データは一般に質が悪く量も少ない
・関心のある事象が適切にデータに表現されているとは限らない

・データ化までのタイムラグが小さくない
③利用者側の問題……・意思決定者に数量的判断志向性がない場合も多い
・意思決定者が必ずしも統計学に精通していない
・統計分析による情報には適・不適がある

2）企業の統計分析でもっとも多く用いられるのは日常の各種業務記録である。各種業務記録の質は必ずしもよくない。企業の業務の遂行状況を忠実に映していない場合が少なくないからである。例えば，今月のノルマ達成のために納品先に無理やり商品を押し込んでしまったり，逆に，来月のノルマ達成のため今月末の売上げを一部来月に繰り延べしたりすることはしばしばある。また，部署によって販売の成立が契約成立時点か，納品完了時点かで統一できなかったりもする。こうなると業務記録からだけでは，例えば本来の月別需要量を正確に把握することはできない。また，経費を記録する場合，勘定科目への分類が部署や担当者によって異なることが少なくない。その他，記録上の単純な書き間違いや入力間違いなども少なくない。これらの例からもわかるように経営現場で発生するデータにはかなりの恣意性や誤差が含まれていることを認識しておかねばならない。

3）精度はある程度確保できていれば十分であり，なるべく速やかに分析結果を入手することの方が重要である (Whitehead & Whitehead, 1984 ; McArther, 1980)。どんな美しい結果であろうと１ヶ月も遅れれば使われない。むしろ不完全であろうと，意思決定を助けるように時間内にすばやく反応する方が大変価値がある (Porter, 1993)。簡単な数字で予測し，統計的妥当性をとやかく言わずに簡潔な市場調査を行い，少数の代替案と大まかな見積もりのみを用い，速やかな費用便益計算を行うことが大切である (Mintzberg, 1973)。

4）分析の困難さとは，「技術的に高度で難しい」というものから「単純に面倒くさい」というものまで含まれる。

5）経営統計学の教科書の問題については，東渕(1993)を参照されたい。

参　考　文　献

1) Chatfield, C., "The Initial Examination of Data," Journal of The Royal Statisitical Society, Series A, 1985, vol. 148, Part 3, pp. 214-253.
2) Cox, D. R. & Snell, E. J., Applied Statistics, Chapman and Hall, 1981.
3) Ehrenberg, A. S. C., A Primer in Data Reduction, John Wiley & Sons, 1982.
4) McArther, D. S., "Decision Scientists, Decision Makers, and The Gap," Interfaces, vol. 10, no. 1, 1980, pp. 110-113.
5) Mintzberg, H., The Nature of Managerial Work, Harper Collins Publishers Inc., 1973. (奥村哲史，須貝栄訳，『マネージャーの仕事』，白桃書房，1983.)
6) Neter, J. & Wasserman, W., Fundamental Statistics for Business and Economics, Allyn & Bacon, 1961. (保田順三郎監訳『経営と経済学のための基礎統計学＜上＞データ

の分析と提示』、ダイヤモンド社，1964.）
7) Porter, M. A., "The Role of The Statistician in Industry," The Statistician, vol. 42, 1993, pp. 217-227.
8) Whitehead, P. & Whitehead, G. Statistics for Business, Pitman, 1984.
9) 上田尚一，『統計データの見方・使い方』，朝倉書店，1981.
10) 古寺雅美，『統計学以前の統計入門』，東京法令出版，1980.
11) 東渕則之，「経営統計学の方向性に関する序論」，松山大学論集，第5巻第3号，1993.8，pp. 449-469.
12) 東渕則之，「米国流経営統計学の意義と限界—BS教科書を題材に—」，松山大学論集，第5巻第4号，1993.10，pp. 461-485.
13) 東渕則之，「意思決定者における統計分析情報の取り込み条件に関する一考察」，松山大学論集，第7巻第4号，1995.10，pp. 79-96.
14) 林知己夫，『行動計量学序説』，朝倉書店，1993.

（本稿は，平成8年度松山大学特別研究助成による研究成果の一部である。）