データ分析の面白さとこれから

2016 年 1 月 26 日

久　米　　均

インターネットの普及により、統計的母集団を想定した従来の少数例による推測統計学的方法に加えて、ビッグデータによる記述統計による新しい手法が生まれつつある。これまでのデータ分析の発展を振り返りながら今後の方向を考えてみたい。

１．統計的方法の適用について

ばらつきを持つ製品群の複雑な品質問題を解析するためには統計的方法は不可欠な手法であるが，統計的方法ですべてのことがうまくいくわけではない．改善のためのデータ分析には 3 種類の問題がある。

ａ．　第一はわざわざデータをとらなくても，不具合の現象やその工程を見ればすぐにその原因がわかり，対策がとれるものである（対策が困難なものもあるが)．たとえば，設備に突起があり，製品がそれに当たってキズがつく場合や，パイプに孔があきそこから油が漏れているというような場合である．このような場合は，現場でデータをわざわざとるまでもなく，すぐに対策をとることができるし，とらなければならない．

ｂ．　第二は多くの品質問題がそうであるが，いくつかの要因が複雑に入り組んでおり，統計的にばらついているため，製品や工程を見ただけでは不具合の原因を見出すことが困難で，データによる現象の分析，原因の解析が必要な場合である．統計的方法はこの場面でもっとも有力である．

製造作業で発生する品質不良の原因を解析するために統計的方法が有効であるのは，不良の発生状況を統計的に把握することにより，その原因が解明できることが多いからである．１個の欠陥製品を顕微鏡で精密に観察しても解明できない原因が，欠陥の発生状況を多数の不良品について調べることによって解明できるのである．多くの不良品についてその傾向を統計的に調べることによってその原因が明らかになり，これが不良の解消につながるのである．

わが国ではほぼ 5 年に 1 回国勢調査が行われるが，これは行政の基礎資料とするために，人口動態およびこれに関する諸種の状況を把握するために行う．

これによりミクロ的観察では見えないものが見出される．たとえば，出生率の変化，農村から都市への人口移動の状況などは個別的調査では把握しにくい事柄であるが，定期的に統計的調査を行ってデータを集め，それを分析することにより，明らかにすることができる．

個別では見えないものが，全体的状況の中で見える．ミクロでは見えないものが，マクロで見えるのである．マクロ的観察には多くの記録やデータが用いられる．統計的方法はこの情報を処理するための道具である．

ｃ． 第三はデータやこれまでの経験だけでは，処理することが困難な場合で，洞察力，新しい発想，創造的活動，試行錯誤のトライアルを必要とする場合である．たとえば，実験にどのような因子を取り上げるか，市場調査のデータをどう解釈するかなどの新しい要素の多い問題ではデータを集めて眺めるだけではうまくいかない．大きな変化がある時には，それまでのデータによる予測は外れるのが一般である．この場合，データから読み取らなくてはならない事柄は，傾向ではなく兆候である．それが何の兆候であるか，その底にあるものは何かが重要なのである．この段階では従来の統計的手法が活用できるほど状況がよくわからないのが普通である．しかし，「よくわからないから，もっとデータを集めよう」と言ってはならない．「分析を行えるほどにはまだわかっていない．だから出かけて行って，観察し，質問して聞いてくることにしよう」と言わなければならない．アフリカのある地域に靴を売るために商社が二人のセールスマンを派遣した。しばらくして、第一のセールスマンから、「この地域では靴は売れません。みんな裸足で歩いています」という報告が来た。その後で第二のセールスマンから「この地域は有望です。みんな裸足で歩いています」という報告が来た。これらの報告ではアクションが取れない。もう一歩突っ込んで‘何故靴を履かないのか’その理由を調査することが必要である。

２．1960 年代のデータ分析の問題

1960 年代の始めに日科技連にはいろいろな研究会があった。水野先生、石川先生、朝香先生、増山先生などの大先生がそれぞれ研究会を持たれ、統計的方法部会、石川部会、抜取検査部会、M^2部会などで統計的方法の勉強、研究が行われていた。石川部会でよく議論になった問題の一つに、データの構造模型に関する問題があった。

ブロック因子

変量模型、母数模型をどのように取り扱うかの議論が行われたが、今から考

えると、実験計画法で取り扱うブロック因子について多少方向違いの議論がなされていたように思われる。

ブロック因子とは実験単位の全体の集合より均一となるように実験単位をまとめたもの。ブロックという用語は元来農場を、風当り、地下水への近さ、耕地層の厚さといった共通条件で分割する農事試験に由来する。ブロックとして用いられる因子の他の例は、**原材料のバッチ、作業者、日**などがある。(JIS Z 8101)

実験計画法では処理の効果とブロック効果を分離し、処理の効果の差を精度よく推定する方法として、乱塊法、釣合い型不完備ブロック計画（BIB)、部分釣合い型不完備ブロック計画（PBIB）などがあり。BIB、PBIB では複雑な理論に基づいて解析が行われる。そのためこれらの方法は高度な方法とされ、日科技連のベーシックコースでは取り上げられず、その上級コースとされていた実験計画法コースで教えられていた。しかし、JIS の記述はおかしいのではないかという議論があった。例えば実験日による効果、すなわち、実験日によるデータの違いをブロック効果とすることは、品質管理の観点からは日によって品質に差が出るということである。原材料のバッチ、作業者についても同様で、原材料のバッチが変わると品質に違いがでてくるということでは具合が悪い。原材料バッチの品質規格に問題があるわけで、この原因を調べて除去しなければならない。ブロック効果がなければこれらは一元配置法に帰着する。品質管理では多くの場合ブロック効果は変量でなく母数としなければならない。これを強く主張されたのは田口玄一さんであった。

交互作用

交互作用とは応答変数に対する一つの説明変数の影響が、他のいくつかの説明変数に依存している程度（JIS Z 8101）である。

要因効果の線型加法性からのずれ、ということもできる。

応答変数を z 、説明変数を x, y とし、f, g を任意の関数とするとき

$$z = af(x) + bg(y)$$

と書けるときは x と y の間に交互作用は存在しない。しかし、

$$z = cxy$$

の関係では、x と y の間に交互作用がある。それでも、両辺の対数をとり、$Z = \log z$, $X = \log x$, $Y = \log y$ とすれば

$$Z = C + X + Y$$

となり、X と Y の間には交互作用は存在しない。しかし、

$$z = af(x) + bg(y) + cxy$$

の場合は、変数変換により非線形性を取り除くことはできないので、この場

合は真性交互作用、前者の場合は仮性交互作用ということができる。

実験計画法で要因効果の線型加法性からのずれを全部ひっくるめて交互作用としてしまっていることは、交互作用に対する技術的理解を妨げ、その研究を行う機会を取り除いてしまっているように思われる。以下は交互作用を変数の型で分類する一つの試みである。実際の実験で、交互作用が現れたとき、それがどのような型に属するものであるかを分類してみることは交互作用に関する理解を深め、実験計画法の発展に資するのではないかと考えている。

外延量と内包量

質量･長さ･体積などの同じ種類で加え合わせることのできる量は外延量と呼ばれる。例えば５Kg のものと 10Kg のものを合わせると 15Kg になり、加法性がある。一方 20°の水と 30°の水を合わせてかきまぜても，50°にはならないで，20°と 30°とのあいだのある値をとる。このように加法的でない量が内包量である。密度・濃度・温度・純度・不良率などがそうである。

応答変数が内包量である場合は交互作用が出る。

離散因子、分類因子

反応装置、作業者、材料メーカーなどの因子は水準と水準の間が存在せず定量的に差を示すことはできない。これらの水準は多くの因子の集合の質的、量的違いによって構成されている。材料メーカーが異なる場合同じ条件で作業を行っても、同じ品質にならないということはしばしば経験する所であるが、これは同じと思っている材料が異なるからである。規格通りの材料だから同じ材料というわけにはいかない。他の規格で定められていないもろもろの特性に差があるので、同じ品質の製品が得られないのである。このような場合には交互作用が現れる。離散因子の場合、異なる水準では何が違うのかはっきり分からないことが少なくなく、他の因子との関係も変わってくるので、この効果が大きい場合は交互作用として現れる。

３．少数例からビッグデータへ

ビッグデータとは、市販されているデータベース管理ツールや従来のデータ処理アプリケーションで処理することが困難なほど巨大で複雑な データ集で、以下のようないろいろな状況で処理されるデータが総称してビッグデータと呼ばれている。

パターン認識、異常検出

銀行詐欺、構造欠陥、医療診断、パスポート写真照合、文書中の誤り検出など。

クラウドソーシング

テレビ放送での動画番組の素材提供を一般の視聴者に求め、多くの応募の中から適当なものを選出して放送する。パッケージソフトのβ版をマニアに無料で提供し、プログラムエラーの情報提供を受ける。

データ融合と統合

地上観測データ（点観測、連続的）と航空機、衛星からのデータ（面的観測、スナップショット的)、数値モデルからのデータを組み合わせて統合的に利用する。

アンサンブル学習

いくつかの学習法、訓練法を組み合わせて統合的に活用する。

相関データ抽出

データベースに蓄積された大量のデータから、頻繁に同時に生起する事象同士を相関の強い事象の関係、すなわち相関ルールとして抽出する技術。

例：本 A を買う人は、後に本 B を買うことが多い。→　本 A の購入者に本 B を勧めるダイレクトメールを送る。

クラス分類

与えられたデータに対応するカテゴリーを予測する

例：薬品の化合物のデータから，その化合物に薬効がある・ないといったカテゴリを予測する。

４．データマイニング

特定の目的のためだけに集められたものではない大量のデータから、目的に沿ったモデルを掘り起こす技術。データマイニングではそれまで「未知」だった知識を発見できるかどうかで評価される。知識の有用性は最初に設定した目標に対して決定される。データそのものは確たる目的なしに集められたものかもしれないが、それを用いる分析（マイニング）には、適正な目的の設定が必要である。

以下はデータマイニングの一般的ステップである。

データマイニングのステップ

① データの同定・収集・選択

一般にデータマイニングの対象となるデータは特定の目的を念頭に置いて収集されたものではなく、かつ一つのデータベースに集中しているとは

限らない。データハウスと呼ばれるデータ集合からマイニングに必要な目標データセットを選択する。

② 前処理・変換
目標データからノイズや異常値を除去、欠損データの補完、単位の変換、などデータを成型する。
データの構造の統一、医療情報、化合物・薬品情報など構造を持つデータ、赤外線吸収など曲線データなどもあり、特殊な変換が必要になる。

③ パターンの発見
興味あるパターン（知識）の候補を抽出する。
統計的手法、機械学習、概念・クラス記述、分類規則、相関規則、決定木、などデータマイニングの手法が用いられる。

④ 結果の解釈・評価・活用
抽出したパターンを解釈・評価して知識を得る。このステップまで十分な知識が得られなければ、一つ前のステップに戻る。どこまで戻らなければならないかは事前には分からない。新たなデータを収集追加しなければならないこともある。

・知識の有用性は最初に設定した目標に対応して決定される。データそのものは確たる目的なしに集められたものかもしれないが、それを用いるマイニングには、適正な目的の設定が必要である。
・これまでの多くの研究はパターンの発見のアルゴリズムに集中しているが、一番重要なのはデータの獲得、選択、前処理である。データマイニングでは、データは何らかの形ですでに蓄積されているとの前提であるが、マイニングに使える形に持っていくこの部分が全プロセスに占める割合は７０～８０％であるといわれている。

データマイニング事始め

現場データを使った不良解析は本質的にデータマイニングである。

例　強化ガラスの不良解析

1959 年、筆者が石川馨先生のご指導のもとに行った不良の解析は、今から考えるとデータマイニングのはしりであった。当時は計算機が未だなかったが、算盤の上手な現場の女性がデータの整理を手伝ってくれた。

当初利用できそうなデータとしては

製造部　生産月日、担当直、　品種、強化炉 No.　操業条件、加工枚数
検査部　寸法不良、疵不良　その他の不良
工務部　加熱炉、風冷装置、その他の修理、部品交換記録

などがあったが、製造と検査のデータは別々に取られ、その対応が記録されていなかったので、あらためて、操業条件と不良の発生状況の対応が付くようにデータシートを設計し、全部で 10 基あった強化炉ごとに生産品種、加工炉、操業条件などの対応のあるデータの獲得に努めた。約半年にわたって不良項目ごとに作業者、加工条件の関係を分析し、その操業方法を細かく分析することにより、不良の原因が少しずつ明らかになり、不良は少しずつ減少した。

データは最終的には 1,000 件近くになったが、用いた統計的手法は主にデータの層別だけで、推測統計学のこれといった手法を活用することはなかった。

５．データ解析法の変遷

データを収集したり要約したりすることに関する統計的方法の部門を通常**記述統計**と呼び、抜取検査などデータの源泉に関する推論を行う部門は**推測統計**と呼ばれる。推測統計的方法の利用は 20 世紀に入って著しく増加した。生物学、社会科学、自然科学、工学分野でそれが顕著であった。わが国では日科技連や規格協会の品質管理セミナーで推測統計学を基盤とする統計的方法の教育が行われた。統計的方法はその発展とともに複雑、多様となったが、最も重要な方法の多くはきわめて単純で、層別・分類および時系列プロットなど記述統計の方法である。計算機による高速データ処理、インターネット技術の普及発展に伴い大量のデータを比較的簡単に処理することが可能になり、ビッグデータと呼ばれる新しいデータ分析の分野が生まれつつあるが、統計技術に関しては今のところ大部分が記述統計の技術である。データ解析の方法は

記述統計 → 推測統計 → 記述統計

と元に回帰している。

６．力（ちから）と技（わざ）

ソ連式品質管理
現代の大砲と那須与一
平塚のタクシー　小田原からの戻り車
台湾の仕立屋

（この項は日本規格協会発行の「標準化と品質管理」4 月号の随筆欄に掲載予定）

７．作り方の研究から使い方の研究へ（石川　馨）

品質管理において、これまで多くのデータ解析の方法が開発され、応用されてきたが、その大部分はモノづくりに関する領域であった。ものをどう作るか（how）の研究は重要であることは言うまでもないが、品質管理においては何を作るか（what）に関する研究が今後一層積極的に行われることが重要である。

ナイロン

工業材料は使い方をよく研究してやらないと失敗します。日本のナイロンメーカが、ナイロンを作ったが売れない。何故売れないかというと作る研究ばかりやって使う研究をやっていなかったからです。アメリカのデュポンは、作ったものをどう使うかという研究を大変良くやっています。そこで、最初に魚網の研究から始めました。魚網にはどんな糸が良いのかをいろいろ調べました。魚網に良い糸と、ストッキングにする糸とではその要求される性質は同じではありません。調査の結果、こういう糸が良いのではないかということで作ったところ今度は売れました。

男性用化粧品

男性用化粧品の場合は、買っているのは奥さんかお母さん、お姉さんです。本人は買いに来ない。今はそんなことはないと思いますが、その当時男の子が化粧品店に入るなんて恥ずかしいことでした。そこで商品を簡単に積めるよう容器をデザインし、店頭に置きました。男の子でも持って入るのは比較的楽ですから。このようなプル商品は女性の化粧品と違って広告が極めて大切です。また、男性用化粧品と女性用化粧品は使う場所が違うのです。女性用は化粧台の前ですから暖かいところなのですが、男性用化粧品はだいたい洗面所に置いてあります。ところが洗面所というのは非常に過酷な条件で、夏は暑くなり、冬は零下何度に下がってしまうところがあります。だからそのような条件でも変質しないことが必要です。又、最近の洗面台はみな陶器ですから、ガラスの容器だと落としたときに割れてしまう。それで材料は全部プラスチックにしました。市場調査をやってパッキンがどのくらい残っているかを調べたところ、ほとんどパッキンを落としてなくしている。そこでパッキングレスにしてグッと締めれば止まるような瓶にしました。そういったデータをいろいろ集めて企画委員会に持ち込みました。その結果、これは売れるだろうということになり、値段は、２０歳の男の子が買うので、一瓶３００円ということで直ちに生産に入り、うまく売れました。

引用文献

１．久米　均　「品質経営入門」, 2005,　日科技連出版社
２．元田　浩　他　「データマイニングの基礎」, 2006, オーム社
３．ホーエル, P.G. 「初等統計学 改訂版」, 1970, 培風館

４．久米　均編　「石川　馨　品質管理とは」, 2015, 品質月間テキスト

■公開資料ページ
弊社ウェブページで各種資料をご覧いただけます http://www.i-juse.co.jp/statistics/jirei/

■お問い合わせ先
(株)日科技研 数理事業部 パッケージサポート係 http:/www.i-juse.co.jp/statistics/support/contact.html