統計的データ解析

随　　筆

白　旗　慎　吾*

Statistical Data Analysis

Key Words：Data Analysis, Statistics

1．はじめに

最近統計学が世の注目を集めています．西内啓氏の「統計学が最強の学問である」がベストセラーになり，日本統計学会が主導して2011年に始まった統計検定の受験者数は1222名，2012年度は2692名と倍増しています．

特にビッグ・データに関する話題が，特にビジネス界で注目されています．ビッグ・データは，コンピュータ，デジタル通信機器の普及発展により大規模なデータ収集が可能となり，ツイッターやSNS，コンビニやスーパーの大きなチェーンの顧客データ，官庁統計など，爆発的に増大しています．大規模なデータベースには有意義な情報が含まれており，多くのソフトウエアも作られています．うまく活用すれば企業の利益や国の活性化に役に立つのは間違いないでしょう．ただし，私には規模の大小以外には，少し前の流行語であるデータ・マイニングとの区別がつきません．データ・マイニングは情報系の人達が推進したように思いますが，推進している人達の層が広がっただけではないでしょうか．もちろん層が広がるということはそれが有用であることの証明でもあります．そこで用いられている手法の多くは，伝統的に統計で用いられていた手法と名前が異なるだけで統計的手法であり，データ解析に他なりません．

* Shingo SHIRAHATA
1947年6月生
大阪大学大学院基礎工学研究科　数理系専攻　修士課程修了（1972年）
現在、大阪大学　名誉教授
理学博士（九州大学）　統計科学
TEL：0797-81-3312
FAX：0797-81-3312
E-mail：sirahata@sigmath.es.osaka-u.ac.jp

では統計的データ解析とは何でしょうか．統計の源流はドイツ語のStatistikにあるようにstateの状態の記述，フランスの確率論，イギリスの政治算術やロンドンの死亡統計にある，と言われています．しかしデータを集めることは古代から行われています．味方と敵の人口・兵力や生産力の把握なしに戦争も政治も不可能です．そのために古代ローマで国勢調査のような調査がすでに行われていました．しかしこれらはまだデータを集めているとは言えても，解析しているとは言えないでしょう．

私はもともと数学の出身で，実際のデータ解析に従事した経験は乏しいのですが，大雑把に言えば，C.R. Raoも言うように「不確実性（誤差，個体差，個人差等）を含んだデータの不確実性を評価すること」と考えます．しかしながら統計データはすべての科学で得られ，不確実性のタイプも多岐に渡ります．実際のデータを気持ちよく解析できることは実は滅多にありません．以下で乏しい経験を披露します．

2．統計学の知識のあまりない人からの相談

かつて教養部で統計学の授業をしていた関係で，しばしば，データがおかしいから見て欲しい，と飛び込みがあります．某学部の若い人から，血中のある成分AとBの濃度はほとんど関係がないはずなのに強い相関がある，どうしてか分からないので教えて欲しい，と電話があり，承知するとすぐに車で飛んできました．一見してすぐA，Bは共に年齢とともに上昇する傾向があることが分かり，偏相関を教えて5分で解決しました．あの学部では統計学は必修のはずなのですが．ただし大学に入学したての1年生の統計学では偏相関まで講義する余裕はありません．統計学の知識があまりない人の持ってくるデータは，すぐ返事ができるか，もしくは何がや

りたいのかさっぱり分からず，考えようのない場合が多いようです．

3．某電力会社の電力ケーブル取り替え問題

大都市では電柱，電線の地下化が進められています．前から決まっている都市計画にしたがって地下化が進められているのですが，初期の埋設からかなり時間が経過しており，想定された寿命期限が来つつあります．電力会社は新規の地下化と既存の古いケーブルの置換のバランスに神経を使っていました．古いケーブルの送電容量は22KV，33KV規格であり，一応設定された寿命が来るごとに66KV，77KVに置換していきます（今はもっと大容量かも知れません）．ただし寿命が高精度に推定できれば新規の地下化を優先できます．寿命に最も関係するのは水です．実験室でのデータではどの程度の水分にどのくらい暴露されれば寿命が尽きるか，がはっきり分かります．しかし実際に敷設されているケーブルがどのくらいの水に暴露されたかは不明であり，かつデータを得るために道路を掘り返すことはできません．たまたま道路工事など他の事情で撤去されたケーブルに高電圧の電流を流してダウンするまでの時間を測定する，切断してケーブル内部をチェックする，などで得られたデータを用いて寿命を推定するのですが，寿命も直接は測定できません．いわゆる野外データの解析は泥沼を行くがごとしです．そこで，電力会社では回収された電線に，寿命に関連するはずのいくつかの実験を行ってデータを得ています．担当の研究員は統計理論とその実践にもかなり詳しい人達です．告白すると，統計に詳しい人からの相談に，的確に応えることができることは珍しいのです．かなり多数の多変量データが得られているので説明変数の選択やその変換によって回帰分析を行った結果，データに適合する有意な回帰関係が多数得られました．実際のデータでは，データに十分適合し，かつ簡明な構造のモデルは一般に複数得られ，かつモデルの良さを測る尺度でもその多くにはほとんど差がありません．その中から現場の人達の感覚に合うモデルを採用しましたが，寿命に関する信頼区間を求めても実用的な結果は得られず，参考資料を提供しただけのようでした．

4．眼に関するコホートデータ

某医科大学の眼科教室に協力してデータ解析を行いました．データはある幼稚園から高校まである学校の生徒（4歳から18歳）約4000名の1984年からの眼に関する測定データです．データは視力だけではなく，左右のどちらが利き目なのか，眼球を楕円体と見なしての角膜前面の曲率半径，など非常に詳しく測定しています．曲率半径は乱視の程度の指標です．視力の定義を初めて知りました．利き目は悩ましく，左右どちらが利き目か判定できなかったり年ごとに変わったりすることもあり，きわめて不安定です．

データを一見して驚いたのは空白の多さでした．事情を聞くと，子供が実験動物にされる，と誤解した保護者が検査を拒否した，ということでした．また，まず平均，分散などの基礎統計量を計算すると，とんでもない外れ値が散見されました．ほとんどは小数点が抜け落ちて数値が100倍に記録されているらしいことが分かりました．絶対に誤記入がある，というのでチェックを依頼すると10日，時には1ヶ月近くかかります．測定はともかくファイル化は業者まかせのようでした．皆さん，特にお医者さんは忙しいんですね．

データを採取している目的は，基礎データを収集することにあり，解析は一度も行われていないようでした．世の中には眠っている大事なデータが結構ある，と感じました．平均さえ計算していませんでした．かつ，小学校から中学校，中学校から高等学校にかけてかなりの数の脱落と新規の入学があり，純粋なコホートデータでもありません．15年分のデータが欠測なしに揃っているのはごく少数です．

解析結果は，欠測の多い経時測定データと見なして，いろいろな検定が可能なように一般的な分散分析モデルを構成することによって，多くの項目が年齢とともに下降（視力），もしくは上昇（曲率）する傾向を示し，またばらつきが学年進行に伴い大きくなることを示しました．また，男女間には差が見られないが，幼稚園と小学校低学年，高学年，中学校，高校の各相の間で大きな差があることが分かりました．

5．マウスの発する超音波データ

これは国立遺伝学研究所と統計数理研究所の人達

と現在も共同で解析しているデータです．マウスの雄は雌に対して超音波で話しかけており，そのマウスの種類による超音波の特徴をパターン認識することによって遺伝情報を得る目的でデータの収集・解析を行っています．

データは1秒に44100回観測される分解能で1分間観測された長大な一次元データであり，観測値256個ごとに（ただし半数を重複させる）高速フーリエ変換して時点，周波数，周波数成分の3次元データに変換します．そのうち超音波領域が解析対象です．高速フーリエ変換した後のデータでは0.425マイクロ秒につき125個の周波数ごとの強度データとなります．ただし，バックグラウンド・ノイズが無視できない大きさで存在するので，ノイズを軽減するために移動平均をくり返した後に，周波数強度最大の部分を取り出します．しきい値を設定し，しきい値以下の部分は無視します．図1，2は3次元に変換したデータ，しきい値で小さい部分を無視した結果です．これだけではまだ超音波音声とは見なしがたい断片やジャンプが残ります．そこでさらに曲線と見なせる基準を作成し，曲線を切り出し，それらの曲線をジャンプが許されるタイプのB－スプライン関数により平滑化する，という方法をとります．図3はその例です．これらの関数群をスプライン関数の係数を用いてデンドログラムを作成してクラスター分析により分類しています．マウスの種類によっては，時には2つの周波数で話しかけているとしか考えられないなど，曲線群の分類の様子はかなり複雑であり，遺伝的な類似性や相違との関係を調べています．

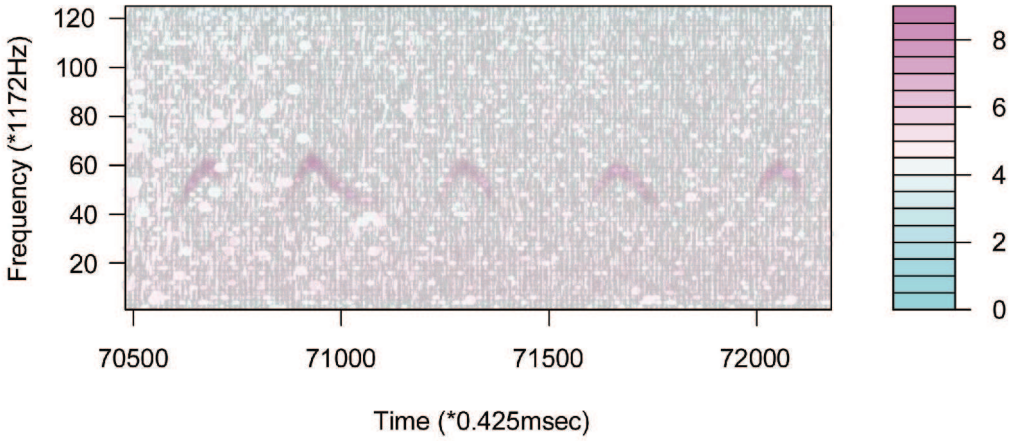

図1．FFTして得られたマウス (BALB/cAnN) 超音波のデータの一部

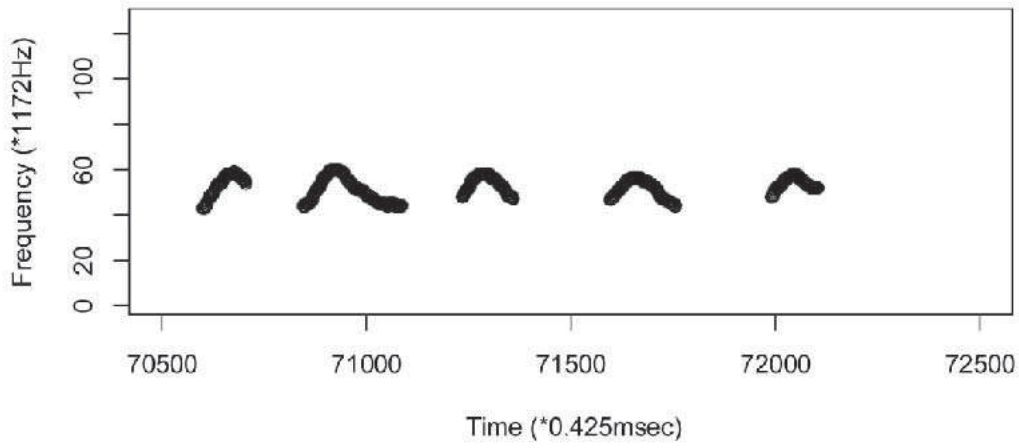

図2．移動平均を取り、ピーク周波数を推定し、得られた超音波発声の波形

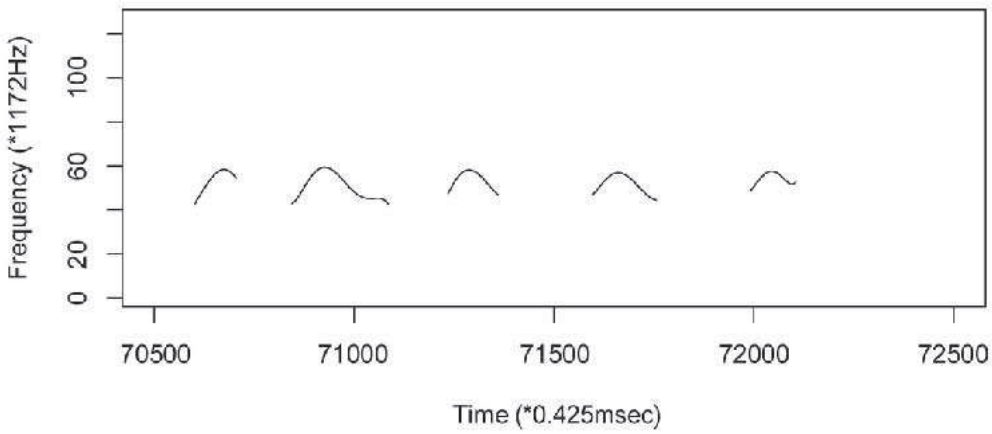

図3．関数として得られた音声の波形

参考文献

(1) 西内啓 (2013)．統計学が最強の学問である．ダイヤモンド社．

(2) Rao, C. R. (1997)．Statistics and Truth. Putting Chance to Work, 2nd ed., World Scientific Publishing Co. Pte. Ltd.（藤越康祝，柳井晴夫，田栗正章 (2011) 訳　統計学とは何か．ちくま学芸文庫．）