データ解析：現代における共通のたとえ話

北海道大学　情報基盤センター
先端データ科学研究室　水田 正弘

１．はじめに

「人類が登場したのは、400 万年前」といっても、私たちが直感的に理解するのは困難である。実生活において、実感できる時間の長さは高々、数十年である。それを、「地球が誕生から今までを 1 年とするならば、人類が発生したのは、12 月 31 日の午後 5 時である。さらに新人(現代人)が生まれたのは午後 11 時 37 分である。」と表現すると、地球の歴史における人類の占める位置が分かったような気分になる。また、日本の道路の総延長は 1266770.4 ㎞であるが、これも把握しにくい値である。「日本の道路の総延長は、地球から月まで行って、帰ってきて、さらに行くより少し長い」とか「時速 200km の新幹線で 264 日かかる」というと、私たちが持っている地球と月との距離のイメージや新幹線に乗った経験から、ものすごい長さであることを感じることができる。さらに、一年間で消費されたビールの総量は、東京ドーム 4.8 杯分というと、その量を表現できる。これらは、データに対する「たとえ話」である。

複雑な事象や大規模な事象を直接的に理解・把握することは容易ではない。それを、巧妙な「たとえ話」を使うと、本質的なことを直感的に伝えることが出来る。データ解析においては、解析対象を、数値などで記述し、さらに少数の値や判断に集約する。集約した結果はデータ解析の利用者にとって理解しやすいものでなくてはならない。データ解析の研究者は、有効で理解しやすい手法を開発しなくてはいけない。また、ユーザは、集約結果を理解する能力を得なくてはならない。すなわち、データの解析法は現代社会における「共通のたとえ話」である。

２．絵で表してみよう

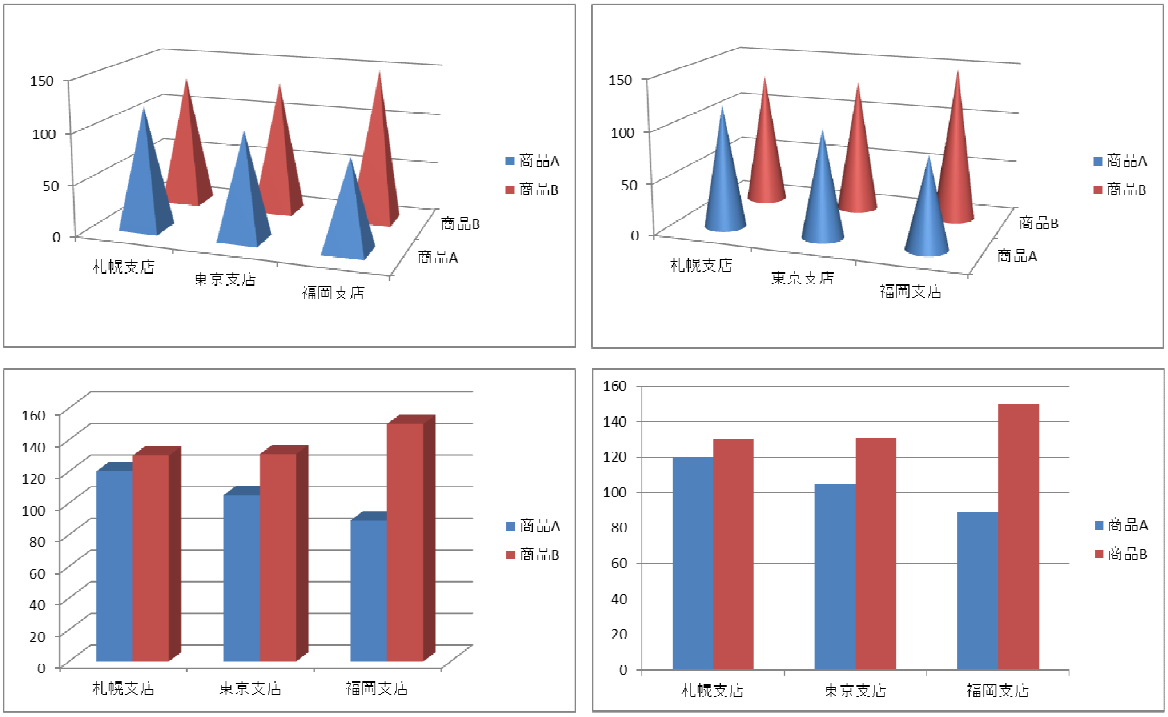
私たちは、図形を認識する高い能力を持っている。幼少のころ、兄弟や友達とケーキを分け合うとき、より大きな方を選ぼうとしてきた。成長するにつれ、どの程度の大きさの差なら我慢できるかを認識してくる。そのため、複雑なデータも適切な統計グラフで表すと、理解しやすくなることが多い。これは、典型的な「たとえ話」である。

しかし、統計グラフで表したものが「共通のたとえ話」になるためには、統計グラフの作成者と、そのグラフを見る人の間に共通認識がなくてはならない。人間の図形認識能力は高い部分もあるが、間違って認識することも少なくない。単に長さを比較する能力は長けているが、面積を認識する能力は弱い。さらに、3 次元物体を平面(紙)で描いたものは正確な把握が困難である。統計グラフでは、人間の誤解を招かない図形を利用するのが重要である。世間で広く使われている Excel によるグラフを例にとる。非常に単純なデータを

想定する。

	商品 A	商品 B
札幌支店	120	130
東京支店	105	131
福岡支店	89	150

このデータに対して、Excel を使って各支店ごとの商品 A、商品 B の売り上げを棒グラフで表示する。非常に簡単にグラフを作成できる。



これらの４つのグラフは、形式的には同じ情報を表現している。しかし、左上および右上のグラフでは、４角錐や円錐を使っていることと、3 次元の物体を 2 次元で表現しているため、量の把握が困難である。また、体積は錐の高さに比例しているとはいえ、その質感に惑わされる。左下のグラフは、多少は改善されている。しかし、実際の数字をグラフから読み取ることは容易ではない。例えば、福岡支店の商品 B の値が 150 であることを左下のグラフから読み取るのは、ほとんど絶望的である。右下のグラフは、デザイン的な面白さはないが、素直にデータを表現している。このようなグラフであれば、他の人も理解しやすい「共通のたとえ話」となっている。

不適切な統計グラフは、人々の誤解を招く。ここに掲載しないが、当日、いくつかの好ましくない統計グラフを報告する。

３．サイコロでたとえ話を作る

確率計算は、数学問題の中でも混乱しやすい課題である。しかし、統計的に表現された

ことを正しく理解するために、確率を使いこなすことは重要である。例えば、中学校における２つのクラスで数学のテストを１回実施して、クラスの点数を比較することを考える。統計学の教科書で t 検定を見つけて機械的に実行するのは容易である。しかし、結果として得られた「２つのクラスの平均点は、5%で有意な差がある」の意味を、「２つのクラスの平均点に差があるというのは 95%正しい」などと解釈してはいけない。正確には、「２つのクラスの平均点に差がないと仮定した場合、このような平均点の差が起きる確率は5%である。5%は小さな確率なので仮定(帰無仮説)、すなわち平均点に差がない、ということが間違っていると解釈するのが妥当である。」という論理である。当然、5%より 1%の方が起こりにくいので、「1%で有意」のほうが強く帰無仮説を否定することになる。数学でよく使われる背理法と類似した論法である。

統計学の教科書でサイコロやコインを使った説明を見ることがある。特殊な状況を除くと、実生活でサイコロの目やコインの裏表が問題になることはほとんどない。しかし、サイコロやコインという理解しやすい「たとえ話」を使って確率や統計的仮説検定を説明している。それが自然な「共通のたとえ話」になるためには、残念ながら多少の努力が必要である。

４．たとえ話を作る切り口

解析対象をどのように記述するか、言い換えると、どの切り口から解析するかは重要な課題である。それに対応して、記述方法を工夫しなくてはならない。たった１つの数値、例えば試験の平均点で記述するのと、各教科の点数で記述するのではデータとして異なる。先の例では、２つのクラスの平均点だけに着目して差の検定をした。しかし、テストにおいて大切なのは、クラスの平均点ではなく、各生徒の達成度である。全国学力・学習状況調査では、各県の平均点のみが注目されているが、県ごとの得点の分布には興味深い情報が隠されている。当日、具体例を報告する。

また、データをどのような集まりとして扱うかについては、シンボリックデータ解析として研究・開発されている。データを、いくつかの数値、離散的な値、テキスト、区間、分布、さらにはそれらの組み合わせとして記述することが考えられている。それぞれの記述方法に対応した解析手法の研究が推進されている。

５．おわりに

PISA(OECD 生徒の学習到達度調査)における我が国の数的リテラシーの順位は、2000 年には世界 1 位であったが、2003 年は 6 位、2006 年は 10 位、2009 年は 9 位となった。これをもって日本の数学のレベルが低下していると結論付けるのは適切ではないが、学校教育における重要な検討課題である。PISA の問題を見ると、統計学に関係した出題が多数を占めている。また、これらの問題は、単なる数学の計算問題ではなく、実際にデータを解釈する力を評価している。PISA の結果も一つの原因となり、平成 20 年度および平成 21 年度

に公示された小学校・中学校の新学習指導要領および高等学校の新学習指導要領では、統計学関係の内容が追加されている。

小学校では、２年生に対する簡単な表やグラフからはじまって、５年生では百分率や円グラフと帯グラフ、６年生では資料の考察が扱われる。中学校では、資料の活用として多くの概念、特に標本調査などが示されている。さらに、高校では、正規分布、二項分布、四分位偏差、母平均の統計的な推測などがある。以上は、算数・数学に関する内容であるが、社会や理科にも統計と関係する内容が明示されている。大部分は、これまでの学習指導要領には記載されていない内容である。このレベルの内容が義務教育および高校の授業によって普及するならば、データ解析は現代における「共通のたとえ話」として有効に利用されていくと思う。

参考文献

Diday, Edwin and Noirhomme-Fraiture, Monique (2007), Symbolic Data Analysis and the SODAS Software, Wiley.

水田正弘・山本義郎・南 弘征・田澤 司(2005), S-PLUSによるデータマイニング入門, 森北出版.

国土交通省道路局Web Page: http://www.mlit.go.jp/road/soudan/soudan_10b_01.html

ビール酒造組合Web Page: http://www.brewers.or.jp/data/t11-kuni-syohi.html

ウキペディア 東京ドーム(単位)

統計学習の指導のために（先生向け）http://www.stat.go.jp/teacher/c3index.htm

文部科学省、PISA（OECD生徒の学習到達度調査）
http://www.mext.go.jp/b_menu/toukei/data/pisa/index.htm