教育ノート

データを見直そう
―より良い統計解析を行うために―

高際　睦

はじめに

筆者は大学時代から統計学の研究をして来たが，統計学の中でも特に，データサイエンスとかデータ科学と呼ばれる分野の研究に携わってきた。名前からもわかるように，データサイエンスという学問は，標本抽出や実験計画などのデータの取得から，モデルの構築，データ解析，モデルの検証に至るまで，データの流れの上にあるすべてのことを科学的に検証するもので，主に解析手法が研究の中心であった従来の統計学に較べ，よりデータを重要視する分野である。データを重要視するのは，統計学を使った研究，調査結果をより厳密なものにするためには，解析手法に関する研究だけを行っても限界があり，さらに精確な結果を求めるためには，どうしても結果のもととなるデータにも注目しなければならないからである。しかも，例えば，データ取得に関する新しい方法を研究することで，得られるデータの精度，信頼性が向上するだけでなく，データの情報量が増えることにより，そのデータに適したモデルの構築，解析も可能になるなどデータの流れの上にあるすべてのことへの効果も期待できるからである。

このデータを重要視するという考えは，最近，多くの分野で取り入れられている。当然，歯科医学の研究においても，データの重要性は変わらないので，多少なりともデータに関心を示すべきである。しかし，本学の統計相談などを通して，多くの研究者の統計解析の手伝いをする機会があったが，ほとんどの研究者は統計解析の手法や解析結果にしか興味がなく，データに関心を持つ人は皆無に近かった。確かに，研究目的に適した解析手法を選ぶことは大切である。しかし，適切な手法を選択するためにも，さらに重要なことである，より良い研究結果を得るためにも，少なくとも研究データに関するきちんとした理解は不可欠である。本来ならば，解析と同等，もしくは，それ以上にデータにも注意を向けてもらいたいものである。

では，実際問題として，データのどの辺りに注意すれば良いかということになると，それを一概に説明することは難しい。例えば，医療系と社会科学系のデータでは着目するポイントが大きく異なるであろう。どの分野においても，データを取得するための手続きや計画に関する部分，つまり，標本抽出や実験計画などと呼ばれる分野が重要であることには間違いないが，これらに関しては多くの文献があるので詳細はそれらにまかせたい。一旦，データを取得すれば，後は解析を行うだけだと思われがちであるが，実は，解析の前後のデータの取り扱いが非常に重要であり，それらについては，残念ながらあまり文献等で触れられることはない。そこで，本稿では，筆者の今までの本学における統計相談などの経

キーワード：データの種類，外れ値，データの表し方，
データの誤差
東京歯科大学数学研究室
（2011年９月11日受付）
（2011年10月３日受理）
別刷請求先：〒261-8602　千葉市美浜区真砂1－2－2
東京歯科大学数学研究室　高際　睦

Mutsumi TAKAGIWA : Taking a New Look at the Data −Achieving a better statistical analysis−（Laboratory of Mathematics, Tokyo Dental College）

験から，データ収集やそのハンドリングなどデータに関することで，多くの人に是非知っておいてもらいたいこと，知っておいて損のないことをいくつか紹介したい。すでによく知っていることであれば，それについての話は飛ばしてもらっても構わない。多くの人に理解してもらえるよう，ほとんど数式を使わず，また，あまり専門的になり過ぎないように説明したつもりである。肩肘張らず，気軽に読んでもらいたい。

1．データの尺度

データは数値で表されることが多い。しかし，数値データだからと言って，データ間の演算が必ずしも自由に行なえるわけではない。統計解析を行なう場合，まずは，扱っているデータの種類，性質などを良く理解したうえで，解析を始めるべきである。

授業評価などのアンケート結果を使って，どちらが良い評価を得ているか比較したい場合がある。例えば，表1のデータにおいて，AとBのどちらが良い評価であるかを考えてみよう。良く行なわれる方法としては，“大変悪い”，“悪い”，……，“大変良い”をそれぞれ，1，2，……，5と数量化し，A，Bそれぞれの平均を求め，その値で比較する方法である。表1のデータの場合，A，Bの平均は，それぞれ，3.2，3.1であるので，Aの方が良いということになるが，この結論についてどう思われるであろうか。この結果はあくまでも1つの目安でしかない。なぜならば，もし，“大変悪い”という評価をつけることは本当に悪いに違いないということで，“大変悪い”の数値だけを－5とすれば，Bの平均は3.1のままであるのに対し，Aの平均は2.9になり，Bの方が良いという結論になる。数量化によって，どちらの結果も起こりうるということは，この方法で得られた結論が絶対的なものでないということである。そもそも，この数量化した数値が何かと言うと，これは，“大変悪い”，“悪い”，……などの各カテゴリーを表すための記号でしかなく，本来の数とはまったく意味合いが異なる。したがって，数量化した数を足すとか，その平均を求めるということからして，何の意味もないのである。その意味では，このような操作は数量化というより符号化と言った方が適切であるかもしれない。あらかじめ与えられたカテゴリーの中から1個，または，複数個のカテゴリーを測定値としたデータのことを質的データ，もしくは，カテゴリカルデータと言う。カテゴリカルデータを解析するとき，各カテゴリーを適当に数量化して行うのが一般的であるが，それはコンピュータで処理するためなどの便宜上のことであって，あくまでもその数値はもとのカテゴリーを表すだけのものでしかない。特に，今の例のようなカテゴリー間に順序がある順序カテゴリカルデータと呼ばれるデータの場合は，順序があるので各カテゴリーを数値で表すことが自然なことと思われがちであるが，それは大きな誤解である。それでは，表1のアンケートにおけるAとBの比較はどのようにすれば良いかということになるが，実は，この種の問題は，特に，統計的に有意な差であるかを判断したいときはそれほど簡単ではない。

もう一つ別の例を考えてみよう。ある治療の前後で痛み具合に有意な差があるかを調べたいとする。この研究を行うためには，何人かの被験者に対し，何らかの方法で治療前，治療後の痛み具合を測定し，そのデータに対して適切な検定を行うのが一般的であろう。痛み具合を測定する方法としては，図1のような視覚アナログ尺度(VAS)が良く使われる。VASの直線上に，現在の痛みの度合いに応じた場所にマークしてもらい，そのマークの位置を測ることで，痛みを数値データとして得ることができる。では，このVASを用いた痛みのデータが得られたとして，どの検定を行えば良いであろうか。同じ被験者の治療前と後の痛みであるので，データは対になっている。したがって，通常は，対応のある

表1　アンケート結果のデータ（架空のデータ）

	大変悪い	悪い	普通	良い	大変良い
A	1	3	9	5	2
B	0	1	16	3	0

図1　視覚アナログ尺度(VAS)

t 検定，もしくは，対応のある符号付き順位和検定を用いれば良いはずである。検定の詳細については述べないが，どちらの検定でも，まずは，同一被験者の治療後の痛みから治療前の痛みの差(便宜上，この差のことをスコアと呼ぶことにする)を計算し，対応のある t 検定の場合には，すべての被験者のスコアの平均に基づいて，符号付き順位和検定の場合はこのスコアの絶対値に順位を付けることによって検定を行う。ここで少し，このスコアについて考えてみよう。VAS によって得られた痛みの数値は，あくまでも被験者の主観的な値であり，客観的な基準で測られたものではない。同一被験者のデータであれば，痛みの基準はほぼ同じであるので，その値を比較することや，スコアを求めることには問題ないだろう。ところが，異なる被験者間では，痛みの基準が異なるので，スコアに順位を付けることや，スコアに関する平均を求めるなどの演算にはほとんど意味がない。したがって，このようなデータに対する検定として，対応のある t 検定や符号付き順位和検定は適切ではないのである。では，どうすれば良いかと言うと，スコアの値そのものではなく，その符号＋(治療後に痛みが増した)，－(治療後に痛みが減じた)，0(治療前後で痛みに変化なし)を使った符号検定を行うべきである。実は，表１のアンケートデータにおいても，有意な差があるか比較したい場合には，符号検定を用いるのが正しい方法である。ただし，アンケートデータの場合は，0(タイ，差がない)データが多いのでよい結果が得られるとは限らない。

カテゴリーを数量化したときの数値に対する演算に違和感を覚える人は少なくないと思うが，VAS で得られたデータのように，一見数値データと思われるものでも，その数値の演算には制約があるかもしれないことはぜひ覚えておいてもらいたい。最初にも書いたが，扱っているデータの種類，性質がわかっていないと，どの演算が行なえるかもわからず，ひいては，正しい統計手法の選択ができなくなる。

２．外れ値

通常，統計解析はすべてのデータから総合的に判断してその結果を導く。そのとき，個々のデータに軽重はなく，すべてのデータは同等に扱われる。しかし，解析手法の中には，わずか数個のデータの影響を強く受け，それらのデータによってほとんど解析結果が決まってしまうものもある。

例えば，図２に示された２変量データの相関係数はどの程度であると予測できるだろうか。良く知られているように，相関係数(r)は２つの変数間の直線関係の方向と強さを表す数値的尺度である。簡単に説明すると，相関係数は－１≦ｒ≦１の値をとり，変数間に直線関係があるとき，その直線が正の傾きを持つときには r＞0，傾きが負であるときは r＜0になる。また，直線関係が強くなる，つまり，その直線の近くに存在するデータの割り合いが多くなるほど，相関係数ｒは±１に近づく。図２の散布図を見れば，ほとんどのデータが正の傾きを持つ直線の近くにあるので，相関係数は１に近い値であると思われるかもしれない。しかし，実際に相関係数を求めてみると，－0.154である。なぜ，このような結果になるのか。グラフの右下にある１個のデータ(* でプロットしてある)を除いて相関係数を計算しなおすと，相関係数は0.905となるので，この右下の１個のデータにより相関係数が予想外の値になったことになる(なぜかは各自考えてもらいたい。相関係数の定義式を考えればわかるであろう)。この右下の測定値のようにデータ全体から極端に離れている測定値のことを外れ値，または，異常値と言う。この例からもわかるように，１個，もしくは，わずか数個の外れ値によって，予期したも

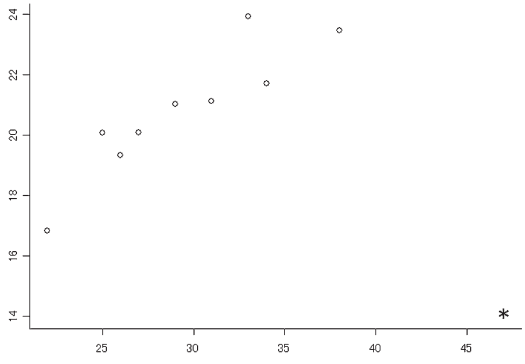

図２　２変量データの散布図。右下のデータ(＊)は外れ値の可能性が高い

のとはまったく異なる解析結果になることがある。

上記の例のように，わずか数個の外れ値によって解析結果が決まってしまうということは，他のデータと較べて，外れ値をより価値があるものと評価しており，あまり望ましいことではない。このようなことを防ぐためには，解析を行う前に，データのグラフを作り，外れ値があるかなどのデータの特徴を注意深く眺める必要がある。データに外れ値がある場合は，なぜそのようなデータがあるかを検証する。外れ値の原因の多くは，測定や入力のミスである。また，高齢者のデータに若年者のデータが交じっていたりするなど，性質の異なるデータが混在している場合もある。このように外れ値である原因がはっきりと特定できる場合にはそのデータを取り除いても構わない。ただし，原因がわからない，もしくは，はっきりしない場合，特に，生命に係わるような場合には，無闇に外れ値を取り除くことは慎まなくてはならない。どちらかと言えば，特異なデータとして，より慎重に取り扱うべきである。もし，外れ値を取り除くことができないデータに対し，それでも解析を行う場合には，外れ値の影響をあまり受けない頑健（ロバスト）な統計手法を使うことが望ましい。例えば，データの中心を表す尺度を求めたいとき，平均が最もよく使われる尺度であるが，平均は外れ値の影響を強く受けるので，外れ値がある場合には，その影響をあまり受けない中央値を用いる方が良い（なぜ，中央値が平均に較べ，外れ値の影響をあまり受けないのかも各自考えてもらいたい。これも，２つの尺度の定義を考えれば，明らかだろう）。

解析を行う前だけでなく，解析を行った後でも，結果が予想と異なるときには，グラフなどを用いてデータを良く見直すべきである。そのときも，外れ値があるかなどのデータの特徴を探り，なぜそのような結果になったかを良く考えてもらいたい。ただし，そのためには，用いた統計手法がどのようなものか，典型的なデータに対してどのような結果になるかなど，その手法の最低限の知識はあらかじめ知っておかなければならない。

３．データの表し方

統計解析の中には，データ全体の様子，つまり，データの分布によって，適切な手法を選択するものもある。したがって，論文等に解析結果を記すときには，なぜその手法を用いたかを明らかにするためにも，結果とともにデータの分布も示した方が良い。分布は，数値，グラフなど何を使って表しても構わないが，その分布に適した表し方をしなければならない。

統計解析に関する説明で，以下のような記述を見かける：

「2つの群ＡとＢの標本数はともに20で，群Ａ，Ｂの平均±標準偏差はそれぞれ10. 03±3. 02，11. 94±2. 55であった。この２つの群のデータに対して，Mann-Whitney のＵ検定を行ったところ，統計学的に有意な差が見られた（図３）。」

この説明，結果についてどう思うだろうか。統計学に多少詳しい人であれば，図３のＡ，Ｂ両群の平均と標準偏差から，本当に有意な差があるのか疑わしいと感じるかもしれない。なぜ，そのような疑いが生じるかと言えば，データの表し方が適切でないからである。２つの群の中心に関する検定を行うときに，データの母集団分布が正規分布であると仮定できるときにはｔ検定を，そうでないときにはMann-Whitney のＵ検定（Wilcoxon の順位和検定とも言う）を使うことは良く知られている。上の統計解析において，Ｕ検定が用いられているということは，扱っているデータの分布に正規性が仮定できないことを示している。実は，正規性のないデータの分布を平均と標準偏差を使って表すことはあまり意味がないし，この例のように，解析結果に誤解を

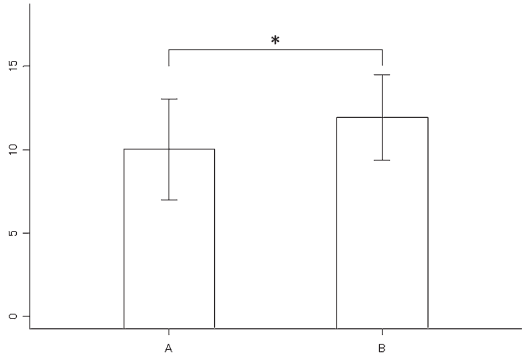

図３　Ａ，Ｂそれぞれの群のデータを棒グラフで表したもの

与えてしまう可能性もある。

平均，標準偏差が，それぞれ分布の中心，広がりを表す尺度であることは，今さら説明する必要もないであろう。では，この２つの値で，データの分布の特徴を表すことができるのはどのような場合であろうか。外れ値のところでも説明したように，データに外れ値がある場合は，平均はその値の影響を強く受けるので，中央値を用いた方が良い。同様な理由で，分布が偏っている場合も平均よりは中央値を用いるべきである。また，標準偏差は分布の広がりを１つの数値で表すので，データの分布が中心に関して対称でないとあまり意味がない。なぜならば，非対称な広がりをしているときには，平均の右側(上側)および，左側(下側)それぞれの広がりを表す量が知りたいからである。このようなことを考えれば，平均と標準偏差でデータが表せるのは，中心に関して対称な分布のときに限られる。さらに，分布の形が釣鐘型(つまり，正規分布に近い)をしていれば，平均±２×標準偏差の区間にデータの約95%が存在しているなど，平均と標準偏差である程度データの様子が予測できる。確かに，どんなデータの分布に対しても平均±２×標準偏差の区間に少なくともデータの75%以上が存在することなどを保証するチェビシェフの定理というものもあるが，これは保守的過ぎてあまり実用的な定理ではない。

では，データの分布に正規性がない場合には，どのように分布を表せばよいであろうか。このような場合は，３数要約，もしくは，５数要約を用いるのが一般的である。データを小さい順に並べ替えたとき，下から25%，50%，75%の位置にある測定値のことをそれぞれ，下側四分位点(Q_1)，中央値(M)，上側四分位点(Q_3) と言うが，３数要約は，この３つの値をこの順で並べたものを，５数要約は，さらに，データの最小値(Min)，最大値(Max)を加え，Min，Q_1，M，Q_3，Max の順番に並べたものを言う。定義からもわかる通り，３数要約，５数要約ともに，隣り合う数値の間にデータが約25%ずつ存在する。先ほども述べたが，分布が非対称である場合には，中心(この場合は，中央値)より右側，左側の広がり具合を知りたいが，それは，それぞれ Q_3-M と $M-Q_1$で求められる。また，５数要約であれば，Min，Max の値から外れ値があるかないかの検証もできる。

分布をグラフで表す場合も，事情はまったく同じである。図３の棒グラフは平均，標準偏差だけを描いたものであるので，正規分布に従うデータにしか役に立たない。それ以外のデータの場合は，５数要約をグラフ化した箱形図(もしくは，箱ひげ図)を使うのが良い。箱形図は，下側四分位点と上側四分位点で長方形(箱)を描き，箱の両端から(外れ値でない)最小値，最大値まで線(ひげ)を引く。また，箱の中の線は中央値を表す。先ほどの例のデータを箱形図で表したものが図４である。A群，B群の中心(中央値)の位置を較べれば，有意な差が見られることの妥当性に納得がいくであろう。

多くの論文などで，データの分布を平均，標準偏差だけで表しているが，これは，データにある程度の正規性が示されている場合，もしくは，今までの調査などで，あらかじめデータの分布に正規性が仮定できる場合だけに有効である。分布に正規性がないときには，３数要約，または，５数要約を，グラフで表すときには箱形図を使うべきである。箱形図は，データに正規性がある場合に使っていけないわけではないので，データに正規性があることをはっきりと示すためにも，むしろ積極的に活用するべきである。

４．データの誤差

データに誤差はつきものである。統計解析の目的はその誤差を含んだデータから，なるべく正しい結論を得ることである。当然，データに含まれる誤差

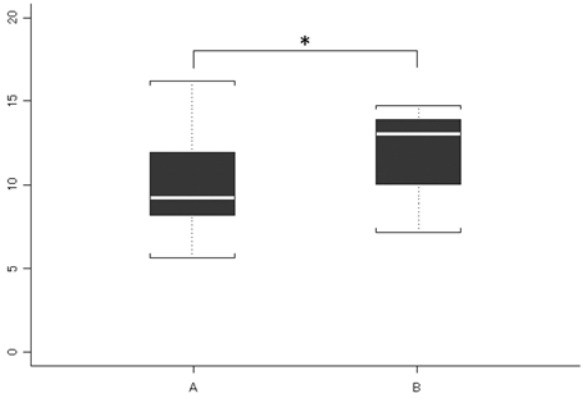

図４　A，Bそれぞれの群のデータを箱形図で表したもの

が大きければ，そのデータから導かれた結論の信頼性は低いものにしかならない。より精確な結論を得るためには，できるだけ誤差の小さなデータを使えば良いわけであるが，では，どのようにすればそのようなデータを得ることができるのであろうか。

誤差に関する話の前に，実験データについて簡単に説明したい。実験データには大きく分けて２種類のものがある。１つは，研究の目的がまだ漠然とした状態で，とりあえず実験を行い，その実験結果から興味ある研究対象を探ろうとするためのデータであり，もう１つは，研究目的がはっきり決まっており，それを立証するための十分に計画された実験から得られたデータである。一般に，前者を探索的データ，後者を検証的データと言う。この２つのデータは，実験の目的が異なるものなので，データ収集に関する考え方も異なる。探索的データを収集するための実験を行う時点では，どの因子(変数)が重要であるかわからないので，できるだけ多くの因子を考慮した実験を行わなければならない。取り入れる因子の数が多くなるので，標本の数も可能な限り増やす必要がある。探索的データを収集するときには，データの量に重点が置かれることが多い。一方，検証的データを収集する場合には，もちろん，標本数も重要である(通常，研究のゴールが決まれば，必要な標本数も決まる)ことに間違いはないが，それよりも，様々な因子をきちんとコントロールするなどの適切な計画のもとで実験を行うことにより，より信頼性の高いデータを収集することが望まれる。つまり，データの質がより重要視される。

このように説明すると，検証的データは，きちんと計画された実験から得られたデータなので，その物理的な誤差が小さくなると思われがちであるが，そうではない。探索的データを収集する場合でも，実験をきちんと行えば，データの誤差は検証的データのものとほとんど変わらない(当たり前である)。では，何が違うかと言えば，検証的データの場合，解析時おける誤差の影響を小さく抑えられる点にある。例えば，男性，女性の比較実験を各６人の被験者で行うとき，実験を

１日目：男性，男性，男性，男性，男性，男性
２日目：女性，女性，女性，女性，女性，女性

と行うことが適切でないことに異論はないと思う。それは，もし，このような実験で男女間に有意な差が認められたとしても，それが，本当に男女間の差なのか，それとも，１日目と２日目の天候などの環境による違い，もしくは，測定器具の精度，測定者などの違いも影響しているのかを判断することができないためである。そこで，実験環境などの因子の影響を小さくするために，通常は，各日，男性，女性３人ずつの実験を行う。ただし，

１日目：男性，男性，男性，女性，女性，女性
２日目：男性，男性，男性，女性，女性，女性

などと行った場合には，まだ，実験順序による影響が残る可能性がある。実験の回数を重ねることにより，段々と手慣れてきて，後に行った実験ほど測定の誤差が小さくなるかもしれないし，逆に，実験の手順が雑になって，後になるほど誤差が増える可能性もある。このような日間の系統的な誤差をなくすためには，各日３人ずつにした上で，さらに，それぞれの日で実験順序をランダム化した

１日目：男性，男性，女性，男性，女性，女性
２日目：女性，男性，女性，男性，男性，女性

と行えば良い。こうすることにより，実験環境における誤差，実験順序による系統的な誤差は男女とも同じ程度と考えられ，もし，男女間に差があるとすれば，それは，まさしく，性別間の差であることになる。実は，このように日などのブロックの中で順序をランダム化する方法はブロック無作為化と呼ばれる，様々な分野で良く使われている実験計画の一つである。

実験の誤差を本当に小さくするためには，測定機器の精度を挙げるなどのハードウェアの進歩がないとそう容易なことではない。それに較べ，解析における誤差の影響を少なくすることは，因子の水準を適切に割付けることや，実験順序をランダム化することにより，簡単に行うことができる。これは，検証的データを収集する場合に限ったことでなく，探索的データを取得する場合にもあてはまる。探索的データを集めるからと言って，何も考えずに実験を行うのではなく，後の解析のことを考慮し，どのような実験が適切であるかをあらかじめ考えてから，実験を行ってもらいたい。

おわりに

美味しい料理を食べるためには，一流のシェフを雇うことも必要かもしれないが，まずは，その料理にあった最良な食材を見つけることである。良い食材さえ手に入れば，あとはレシピ通りに作ったとしても，それなりのご馳走にありつける。良い解析結果を得るのも同じでことある。最も重要なことは，その目的に適したデータを手間暇掛けて集めることである。後は，データを様々な角度から眺め，データに適した解析を行えば，間違いなく，望んだ解析結果が得られるはずである。

本稿のタイトル「データを見直そう」というのは，統計解析におけるデータの価値を認識してもらいたいということと，そのための一つの方法は，データを眺め，さらに良く見直すことであるという二つの意味を掛けたものである。本稿を読んで，データの重要性，データの見方について少しでも理解してもらえたら，筆者としても喜ばしい限りである。