連　載

研究で君が光り輝くために

第２回　データと向き合おう

筑波大学教授　野村　港二

実験や観察では，測定値や記録あるいは画像などのデータに基づいて考察するのが基本です。データは客観的な事実なのですが，注意深く取り扱わないと，気づかぬうちに誤った結論を導いてしまうこともあります。そこで，今回はデータとどう向き合うか考えてみたいと思います。今回も，SSH 校で指導している石井葉子先生にご意見を頂きながら，話を進めます。

そもそもデータとは何だろう

第１回でも少し触れましたが，その語源を知ることがデータと向き合う第一歩だと私は考えています。Oxford English Dictionary では datum, data の単数形，は「A thing given or granted; something known or assumed as fact, and made the basis of reasoning or calculation... [L. given]」と説明されています。データはラテン語で given を意味する語，すなわち既に存在していたものなのです。私も若い頃には，データは実験で作り出すものと思っていました。しかし，そうではないのです。データは，もともとそこにあった自然の営みの一端に我々が気づき，言葉や数字で表現したものです。

もともと存在はしていたものの一端，すなわち母集団から抽出した標本がデータです。ですからデータと向き合うときには，我々が意図的にデータを選んでいないか，データに余分な脚色を付け加えて表現していないか，データから母集団の状態が推測できるかなど気をつけるべき事柄が，実はたくさんあります。

データの取り方，予備実験，再現性

自然科学は，同じ条件でデータを得れば，同じ結論が導かれるという前提で進んできました。たとえば，ある質量の金属球を１mの高さから落とすといつも同じことが起きるといったことから，法則が導かれたのでしょう。もちろん，測定には誤差が生じますし，我々には制御できない僅かな条件の相違もありますから，毎回の数値がぴったり一致することはないでしょう。しかし，同じ条件で実験すれば，誰が行っても基本的に同じ結果が得られる事が研究の基本です。同じ条件なら同じ結果が得られることを再現性と呼び，特に実験ではとても大切です。

再現性を得るためには，条件が定まっている必要があります。実験系の論文に，材料と方法という項目があるのは，そのためです。条件を定めるために，プロの研究者は予備実験を繰り返し，試行錯誤を重ねます。逆にオーバーな言い方ですが，実験を始めたら定まった方法を繰り返すだけですから，頭脳労働は停止させます。高校では予備実験を行う時間も無いでしょうから，見切り発車で研究を進める必要はあると思います。そうであっても，再現性を保証するために，どんな材料で，どんな方法で研究を進めたかの詳細な実験記録をとることが必要です。

これは実験に限った話ではありません。ある大学院生は，膨大な文献から自分が抽出すべき情報がどのようなものかを，あらかじめ具体的に定義をしています。文献調査では，思いついたままに情報を集める研究者もいるのですが，何を集めるかの定義をすれば，実は調査は誰かに任せても，同じ質と量の情報が集まります。いわゆる文系の研究でも，方法を定義することで，再現性を得ることが可能です。

サイエンスの決まりとしての比較

実験を伴う研究では，何かの操作を加えるグループと，それを加えないグループを作り，２つのグループ間に生じる差異を測定して，操作の効果を測るのが一般的です。異なる地点，時間での変化などを比較するということもあります。大学などでは，１つだけの条件を動かす，たとえば温度

や湿度は同じにして磁場だけを変えるなどの操作が可能です。しかし，高校生にとっては，温度や湿度などは成り行きまかせということもあるでしょう。それでも，比較するグループ間では，動かすべき条件以外を，できるだけ同じにすること，また，違いが感じられるときには，状況をきちんと記録することが大切です。

平均値は同じでも

測定して得られたデータの特徴を表すために，代表値である平均値，場合によっては最頻値か中央値を用いることは数学Ⅰで習います。表１は，キツネ組とタヌキ組の化け方テストの結果です。どちらも 40 匹，平均は 66.5 点ですから，同じような生徒がいるように思えます。

ところが，両者のグラフは様子が違います（図１）。キツネ組は平均点近くの者が多く，タヌキ組は成績が散らばっています。両者の点数分布を，平均値と標準偏差で表すと，キツネ組は 66.5±6.6 点，タヌキ組は 66.5±16.4 点となり，数値として違いを示すことができます。標準偏差は，データのばらつき具合を数値化したもので，正規分布をとるデータの場合，平均値を X，標準偏差を SD とすると，X±2SD の範囲に約 95 % のデータがはいります。

標準偏差と標準誤差

平均値と標準偏差で，ある集団でのデータの分布を表すことができました。ところで，キツネ組とタヌキ組は，世界中のすべてのキツネやタヌキ，すなわちキツネやタヌキの母集団の中の 40 匹です。キツネやタヌキの母集団の成績の平均値を，40 匹ずつの組の平均値から推察したいときもあります。実際の研究では，繰り返し何回か測定した値から，母集団の平均値を推定したい，あるいは複数の母集団の平均値を比較で考察するのが普通です。そのようなときには，標準偏差 SD をデータ数 n の平方根$\sqrt{n}$で割った標準誤差 SE =SD/$\sqrt{n}$を使います。標準誤差はデータのばらつきではなく，母集団の平均値の区間推定量です。標準誤差を使うと，X±2SE の範囲に母集団の平均値が存在する確率が約 95 % となり，母集団の平均値を推測する事ができます。研究では，データの信頼性が重要なので，論文では，測定値などを平均値と標準誤差で示すことがほとんどです。

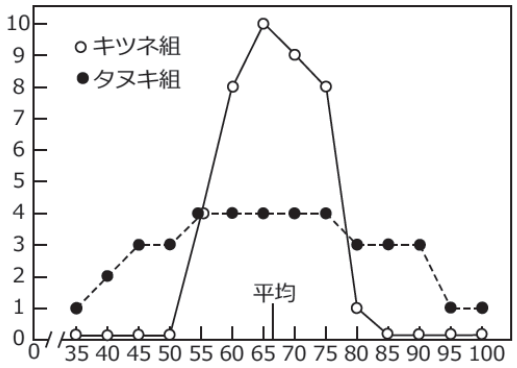

図１　キツネ組とタヌキ組の化け方テスト

安易なグラフの危険性

図２のグラフ３つは，ある魔物の体重の分布です。グラフごとに魔物の体重について考察してください。図２A からは，分布が連続的ではないので，魔物は段階的に大きくなるらしいという結論が得られます。図２B のヒストグラムからは，魔物はまだ重くなるかもしれないこと，図２C からは，体重は少し右寄りに中心を持って分布していることが読み取れます。実は，この３つのグラ

キツネ組

点数	35	40	45	50	55	60	65	70	75	80	85	90	95	100	
匹	0	0	0	0	4	8	10	9	8	1	0	0	0	0	合計 40 匹 / 平均 66.5 点

タヌキ組

点数	35	40	45	50	55	60	65	70	75	80	85	90	95	100	
匹	1	2	3	3	4	4	4	4	4	3	3	3	1	1	合計 40 匹 / 平均 66.5 点

表１　キツネ組とタヌキ組の化け方テスト

体重	9	10	11	12	13	14	15	16	17	18	19	20	21
個体数	0	1	1	3	2	4	3	4	11	2	12	7	0

表２　魔物の体重

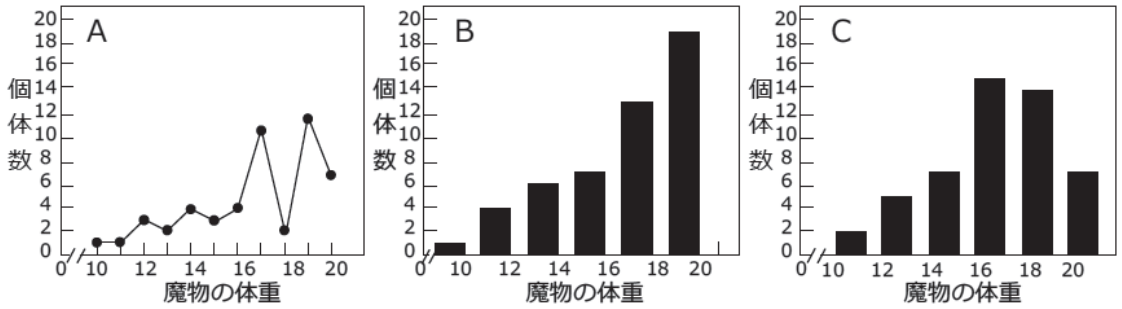

図２　魔物の体重

フは，表２に示したデータをもとに作成したものです。図２Ａは，データをそのまま折れ線グラフにしたもの，ＢとＣは，階級の取り方を「奇数・偶数」「偶数・奇数」と，１つずらして作成したものです。階級をどう取るかという，作為のない操作で，同じデータから，異なる結論を導く複数のグラフが作れます。

このように，悪気は無くても安易なデータ処理は，誤った結論を導きます。信じられないかもしれませんが，もっとも身近な落とし穴の一つが表計算ソフトです。表計算ソフトはグラフの横軸と縦軸を不適切な比率で取ってしまい，差がはっきりしないグラフや，逆に不必要に差があるように見えるグラフを作りがちです。データが得られたら，まずフリーハンドで良いので散布図や単純なグラフを描いてデータの大まかな傾向を知り，この段階で仮説と照らし合わせることが大切です。

相関関係と因果関係

図３の散布図は，キツネたちが，一週間に食べる油揚げの枚数と，化け方の上手さの関係です。油揚げをたくさん食べるキツネほど，化けるのが上手という相関関係があります。では，油揚げを食べると，化けるのが上手になるのでしょうか。それは分かりません。両者に相関関係はあっても，因果関係は分かりません。もともと化けるのが得意なキツネが，人里に下りてきて，油揚げを手に入れているのかもしれません。世の中には，たまたま２つの事象の間に相関が見られることが少なくありません。私たちは自分に都合の良いように物事を解釈しがちなので，つい，相関関係と因果関係を混同して結論を下そうとするので注意が必要です。さらに，作為的にデータを選べば，全く関係のない事柄の間で，相関があるようなグラフすら作れます。

ところで，食べる油揚げの枚数と，化け上手に因果関係があるかないかを知るには，どのようにしたら良いでしょう。方法は一つではありませんが，たとえば，油揚げを食べる前より，食後の方が上手に化けられたなら，油揚げに効果があるのかもしれません。このように相関が見られた２つの事象に一定の時間差が見られるか否かを調べるとヒントが得られることもあります。実験が可能なら，油揚げを食べさせたグループと食べさせないグループで比較する方法もありますね。

画像データ

写真は，具体的な情報を見せることができる有力な手段です。ところが，写真として見せられるのは，個別の事例一つです。たとえば，タヌキの写真を見せても，それは，その一匹のタヌキの写真です。それが平均的なタヌキなのか，集団中のタヌキの個体差がどれくらいなのか分かりません。言い換えれば統計的な有意性は，写真では示すことができません。

一つの解決策は，できるだけ，美しい写真を選

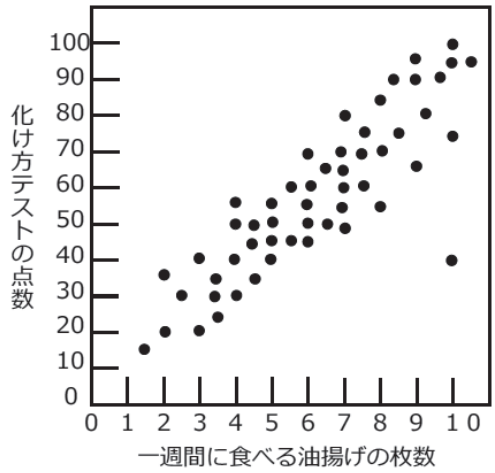

図3　油揚げと化け方テスト

ぶことです。美しいタヌキを選ぶのではありません。まずは，集団を代表できるような大きさや毛並のタヌキを選ぶこと，タヌキの全身像が必要なのであれば，鼻の先から尻尾の先まで，きちんとピントが合った写真であること。余計なもの，たとえば後ろにキツネが映り込んだりしていないこと。極端な構図ではないことなどがポイントです。必要な要素すべてにピントが合って写っており，余計なものが入り込んでいない写真を，一発で撮るのは意外と難しいものです。当然，たくさん撮って，マスターピースを見せることになります。実は，その過程で私たちはタヌキをよく観察することになり，集団を代表するのにふさわしいタヌキの写真が選ばれることになります。

画像の加工

デジタルカメラでは，撮影者も知らないうちに画像が加工されることもあります。そして，発表する写真は，明るさやコントラストを調整したくなりますし，見やすい画像を用意するのは研究者の良心でもあります。調整をどこまでやって良いかは，分野や状況によっても異なると思います。研究倫理の専門家でも議論されている問題です。

試しに顔写真にヒゲをいたずら書きするように，画像処理ソフトを使って写真を改変してみたら，どうでしょう。元と違うメッセージを与えてしまう写真ができてしまいますね。これが，そこまでやってはいけない，ということです。

なお，求められたら，多少の調整をする前のオリジナルのデータをいつでも示せるようにしておくと安心です。

引用文献もデータの一つ

引用文献もデータの一つと考えられます。ある小説から怪獣が暴れている部分を引用した時，怪獣自体は空想のものですが，その小説にその文字列が存在するというのは，まぎれもない事実，すなわちデータです。研究をまとめるときに，引用した文献が特定できるようにリストを整えるのは，先人を敬うためではありません。そうではなくて，歴史上そのように考えた方や表現が存在したことを明らかにして，自らの結論を支持してもらい，反対意見を引用することで徹底的な議論を展開するためです。引用文献は，自分が研究をまとめるより前に存在していたデータと考えてください。

文献を特定するためには，著者名，書籍や論文の題目，出版社あるいは学術雑誌名，雑誌の場合号数やページ，発行年などを明記します。ウエブサイトを閲覧して引用した場合は，URL と最終閲覧日時を明記するのが一般的です。はじめは，よく読む雑誌の引用文献欄を参考にして書けば良いと思います。

まとめ

データと向き合うためには，データを得るための条件，再現性，統計的な処理，グラフの作成，画像の信頼性など，気をつけなければならないことが，たくさんあります。きちんと取られたデータは客観的な事実ですが，だからと言って，何も考えずに示してしまうのは危険です。データは，ちょっと引いた目線で冷静に見てください。