第３回　データを分析する（1）

宮井 信行（和歌山県立医科大学大学院 教授）

前回（通信 No.134）の関西福祉科学大学の大川尚子先生に引き続き，今回と次回の 2 回にわたって,「データを分析する」のテーマで，データ分析の基礎となる統計学の考え方や基本的なデータ処理の方法について解説させていただきます。

調査や測定を行って収集したデータをどのように整理・分析すれば研究仮説に対して科学的で客観的な解を得ることができるのか，あるいは，先行研究の論文に書かれている結果をどのように読み解いたらいいのか，統計解析は難しいと思い悩むことも多いのではないでしょうか。最近は，統計解析ソフトが進歩して多変量解析などの複雑な解析手法も多く用いられるようになってきたため，このように感じることもしばしばです。しかし，どのようなデータ分析を行うにしても，統計学の理論や考え方を正しく理解しておくことが大切です。今回は，データの性質や尺度，母集団と標本の関係，データの記述的解析などついて述べたいと思います。

1. データの特性と尺度

1）カテゴリーデータと数量データ

データとは，調査や測定を行って得られた数値や，対象となる集団の観測結果のことをいいます。観察する対象の特性には，形や色といった「質的（qualitative）」なものと，大きさや数量といった「量的（quantitative）」なものがあります。質的な特性とは対象を単に区別するだけのものですが，量的な特性は回数や個数を数えたり，計器を用いて測定することで得られるもので，大小関係による距離の情報を持ちます。また，質的な特性のデータは「カテゴリーデータ（categorical data）」，量的な特性のデータは「数量データ（numerical data）」とよばれ，カテゴリーデータは整数値を，数量データは連続した数値をとります。

2）データの尺度

データの特性に対してある数値を対応させる基準のことを「尺度」といい，「名義尺度（nominal scale）」,「順序尺度（ordinal scale）」,「比例尺度（ratio scale）」,「間隔尺度（interval scale）」があります（表 1）。名義尺度とは,「性別（男・女）」,「職業（製造業・農林業・運輸業）」などのようにデータを数値としてではなくグループとして扱う尺度です。例えば, 男に「1」, 女に「0」などのように，同一カテゴリーに属するものに数値を与えたとしても，その数値は単に標識として与えられるものですので互いに区別する働きしか持たず，順序を変えても構いません。このように質的な特性に付与された数値が名義尺度のデータです。一方，順序尺度とは,「病気の経過（改善・不変・悪化）」や「成績（優・良・可・不可）」などのように，カテゴリーデータのなかでも大小関係の情報を含むデータです。

比例尺度や間隔尺度は，カテゴリーではなく連続した数値をとり，大小関係とともに距離の情報も含みます。このうち，比例尺度は，本質的な零点を持ち,「身長」,「血圧」,「コレステロール」など測定して得られるデータの多くがこれに相当します。間隔尺度は，零点が任意に決められているもので,「摂氏の温度」,「自尊感情尺度のスコア」などがこれにあたります。

表 1．データの特性と尺度

■ カテゴリーデータ	
名義尺度（nominal scale） （質的な特性に与えた数値）	職業，死因，性別，人種， 出身地，資格，趣味
順序尺度（ordinal scale） （大小関係の情報を含む）	著効/有効/不変/悪化 痛みが強い/弱い/ない
■ 数量データ	
比例尺度（ratio scale） （本質的な零点を持つ）	年齢，身長，体重 血圧，総コレステロール
間隔尺度（interval scale） （本質的な零点を持たない）	温度（摂氏・華氏），西暦， 尺度のスコア

3）データの情報量

一般に，個々のデータの持つ情報の詳しさには，数量データ＞カテゴリーデータ，また，比例尺度＞間隔尺度＞順序尺度＞名義尺度の関係が成り立ちます。情報量の多い数量データは情報量の少ないカテゴリーデータに変換することができますが（カテゴリー化），カテゴリーデータを数量データに変換することはできません。例えば，体温が実測されているとき，37.5℃以上を発熱と定義して，「発熱あり」，「発熱なし」に区分することや，発熱ありをさらに層別化し，「発熱なし」，「軽度発熱」，「高度発熱」のように区分することは可能です。しかし，「発熱なし」と「発熱あり」の情報のみが得られている場合は体温の実測値を導くことは不可能です。したがって，研究で必要な情報を収集するときには，できるだけ情報量の多い尺度で測定しておくことが望ましいといえます。なお，順序尺度のデータに数値を割り当てることで便宜的に間隔尺度として扱って数量的な分析を行うこともあります（数量化）。

2. データの統計解析

1)母集団と標本

研究は，漠然とした疑問をできるだけ明確にしたうえで，研究テーマを設定することから始まります。また，テーマが定まった時点で研究の概念上の対象集団「母集団（population）」が自ずと決まることになります。ある集団の特性や傾向を知るためには，母集団の構成員全員を対象に調査する（全数調査）のが最も正確ですが，規模が大きくなるほど費用や労力の点で調査が実行不可能となります。多くの場合は，母集団から実施可能な規模の対象集団「標本（sample）」を選び出して調査（標本調査）が行われます。

標本調査を行った後，平均や割合を求めたり，度数分布や散布図を描いたりします。しかし，これは標本の姿を知ろうとしているのではなく，標本が得られた背後の母集団の姿を描こうとしているのです。標本から母集団を推測しようとするときは，標本が母集団の特性を忠実に反映していることが必要になります。このような母集団と標本の関係はデータ分析の基本となりますので正しく理解しておくことが大切です（図 1）。

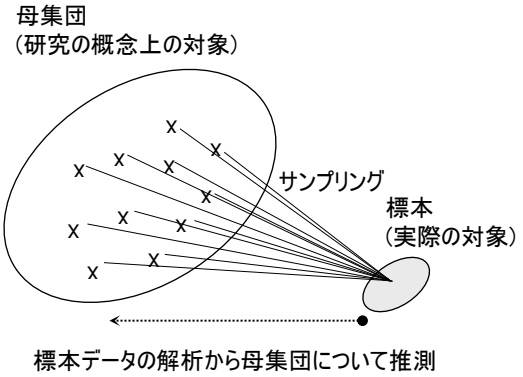

図 1. 母集団と標本の関係

母集団を代表する標本を得るためには，調査者の意思や主観を排除し，無作為に母集団から標本を抽出する（無作為抽出）ことが望ましいといえます。しかし，現実的には無作為抽出の手続きをとることは困難であることが多く，例えば，調査者と縁を持つ人や呼びかけに応じた人など，対象者を便宜的，恣意的に選んで調査が行われます。この場合，対象者が母集団を反映しない可能性があり，その程度が大きいほど，標本の結果と母集団の真値との間の「誤差（ズレ）」が拡がることになります。データ分析に際しては，研究の対象者が母集団をどの程度反映しているかを検討し，母集団からの誤差の大きさによっては結果を母集団に当てはめる（一般化・普遍化）ことに限界があることを意識しておくことが大切です。

2)記述的解析と統計的推論

データを分析するときは，標本において得られた数値を平均や割合などのかたちで要約して示すことが多く，このようなデータ処理を「記述統計（descriptive statistics）」といいます。また，標本の平均や割合から，数学的な確率論を用いて母集団の平均や割合を推測したり（推定），標本でみられた平均や割合の差または変数間の関連が母集団にも当てはまるかどうか

を判定する（検定）ことが行われ，これらは「推計統計（inferential statistics）」とよばれます。つまり，研究においては，さまざまな情報を整理・要約して客観的に表現したり，標本の結果からもとの母集団の真値を推論する解析が行われます。

3. 記述的解析

1)データの要約

データはそれぞれを個別に眺めただけでは全体を捉えにくいため，個々のデータの持つ情報は捨て，全体としての傾向や特性を要約して記述します。通常は，ある変数の数値の度数分布をみて傾向を掴んだり，分布特性を端的に表す記述統計量（基本統計量，要約統計量）を求めることが行われます。

2)分布の特徴をみる

変数のデータは特定の範囲の数値をとります。ある値には人数が多く，他の値では少ないなどの数値のとり方の状態のことを「分布（distribution）」といいます。カテゴリーデータの場合は，性質や特性ごとの人数（度数）で表し，比較的簡単に分布を知ることができます。例えば，血液型であれば，「A・B・AB・O」の 4 つに分類され，どの血液型が多いか少ないかは容易に判断できます。

一方，数量データの場合は，連続的に無数の値をとるため，カテゴリーデータのように個人を数値で分類することはできません。そこで，データを大きさの順に並びかえ，ある階級（範囲）を設定して，それぞれの階級における度数を求め，度数分布表として整理します。さらに，度数分布表をもとに度数分布図を作ると分布の特徴が一層捉えやすくなります。度数分布図としては，ヒストグラム（柱状図）や度数折れ線，累積度数折れ線などが用いられます。

3)記述統計量を求める

①カテゴリーデータ

名義尺度や順序尺度として得られているデータは，「割合（proportion）または百分率（percent）」で示されることが多く，その他に「比（ratio）」が使わることもあります。割合と比はどちらも分数で表される点において似ていますが，割合は分子がすべて分母に含まれるのに対して，比には分子と分母の間に全体と部分の関係がありません。例えば，BMI を求めた合計人数で，そのうち肥満であった人数を割ったものは肥満「割合」で，肥満者について，男の人数を女の人数で割ったものは男女「比」（または性比）となります。

②数量データ

比例尺度や間隔尺度のデータについて，数値の分布特性を表現する指標には，分布の中心がどのあたりにあるかを示す「代表値」と，測定値がどの程度ばらついているのかを示す「散布度」があり，この代表値と散布度の 2 つの情報を示すことで分布の特徴を捉えることが可能になります。

A. 分布の代表値（中心位置の指標）

代表値とは，分布の中心位置を示す数値のことで，「平均値（mean）」，「中央値（median）」，「最頻値（mode）」があります。平均値は，ある変数についての個々の観測値を合計して標本数で除した数値のことです。平均値には計算方法によって算術平均，幾何平均，調和平均，移動平均などがありますが，単に平均というときは算術平均を指します。

中央値は，ある変数の個々の観測値を小さいものから大きさの順に並べたとき，その中央に位置する数値のことです。データ数が奇数のときは，「(n＋1)／2 番目の数値」，データ数が偶数のときは，「n／2 番目の数値」と「(n＋1)／2 番目の数値」を平均して求めます。この中央値は分布に歪みや外れ値がある場合に用いられます。

最頻値は，ある変数の個々の観測値のうちで最も出現頻度が高い数値で，度数分布表やヒストグラムをもとに計算されます。

B. 分布の散布度（バラツキの指標）

散布度とは，データが代表値の付近に集中しているか，散らばっているかを示す数値のことで，「分散（variance）」，「標準偏差（standard deviation）」，「範囲（range）」，「四分位範囲（interquartile range）」などがあります。ある変数の個々の観測値と平均との差（偏差）を 2 乗し，これを全データについて合計した後に，「標本数－1」で割ったのが分散です。また，分散は偏差を 2 乗して求めているので，分散の平方根をとってもとの単位に戻したものが標準偏差です。分散や標準偏差は平均値と対応させて使用され，数値が大きいほどデータのばらつきの程度が大きいことを意味します。

ある変数の個々の観測値を大きさの順に並べたときの最小値と最大値の差が範囲です。また，全体を 1／4（25%）に区切る値を四分位数とよび，小さい方から第 1 四分位数，第 2 四分位数（＝中央値），第 3 四分位数となります。四分位範囲は，第 1 四分位数と第 3 四分位数の差で，データの半数（50%）がこの範囲に入ります。範囲や四分位範囲は分布の歪みの影響を受けにくく，中央値に対応させて用いられます。

③分布の形状と記述統計量

比例尺度や間隔尺度などの数量データの数値は，左右対称の釣り鐘型の分布（正規分布）をとることが多いですが，分布の裾が右（高値側）に長かったり，左（低値側）に長かったりといった左右非対称の歪んだ分布を示す場合もあります。平均値は，分布の形状が正規分布に近いときには分布の中心付近に位置しますが，歪んだ分布では中心から離れてしまうため，代表値として適切とはいえません（図 2）。

データを要約するときは，あらかじめヒストグラムや棒グラフを描いて分布の形状を確認し，正規分布と見なせる場合には平均値や標準偏差で，正規分布から逸脱した歪んだ分布の場合には中央値と範囲または四分位範囲で示すことになります。

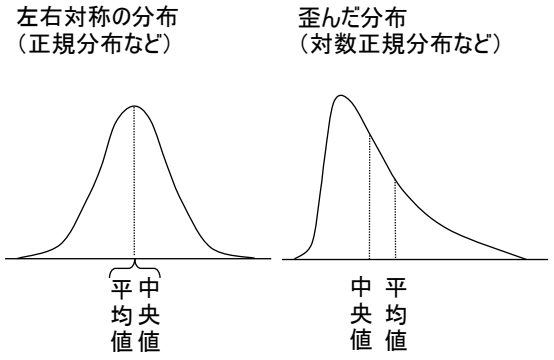

図 2. データの分布と代表値