社会統計学(清水剛)の　しけぷり

presented by　幻日

□データ□

データの収集には確固たる目的や問題意識が前提となる。統計的手法ばかりがもてはやされる中、目的や問題意識が曖昧なままデータを分析しても誤った結果を生み出しかねない。統計的手法を選ぶということは、そのデータの背後にあるモデルを推定することであるからだ。データを取るということは、社会現象に能動的に働きかけることで、主観的に社会現象をある断面から見ることにすぎない、つまりデータとは取られた時点で客観的事実ではなくなるのだ。また得られたデータを数値化し量的に分析する手法が客観的かと言われれば必ずしもそうとも言えず、質的分析の方が現実性を持つかというと、歪んだ観察はいくらでも生じうるから分析段階においても主観性は排し得ない。

しかし、本来分析とは主観性をもって行われるものなのだから、これらの主観性は当然のものと言える。真に問題となるのは主観的な分析をあたかも客観的なもの、科学的なものとしてとらえ、疑わない無批判的な読者であるのだ。

社会学研究には、データを得る際の主観性、質問を作成する際の主観性、データを読み取る際の主観性、データを分析する際の主観性など、おおよそ様々な主観性が存在する。
以下に記すのは、標本調査や社会調査、データ分析に関する単語を解説することでこれらの主観性の一端に触れてもらい、無批判的な読者から脱するためのものである。

□尺度水準□

スティーヴンズは「名義尺度」、「順序尺度」、「間隔尺度」、及び「比率尺度」を提唱し、後者三つを定量的なものと、前者一つを定性的なものとした。

●名義尺度

この水準では数字を単なる名前として対象に割り振る。2 つの対象に同じ数字がついていればそれらは同じカテゴリに属する。変数値間の比較は等しいか異なるかでしか行えない。順序もないし加減などの演算もできない。

例としては電話番号、背番号など。代表値の指標として使えるのは最頻値のみである。統計的バラツキは変動比や情報エントロピーで評価できるが、標準偏差などの概念はありえない。名義尺度でのみ測定されるデータはカテゴリデータとも呼ばれる。

なおカテゴリデータを、ある性質が「あるかないか」という表現に直し、さらにこれを「1 か 0 か」で表現したものをダミー変数という。ダミー変数またはそれから算出されるスコア（点数）を、順序尺度以上の水準に準じて扱う方法もよく用いられる。

●順序尺度

この水準では対象に割り振られた数字は測定する性質の順序を表す。数字は等しいかどうかに加え、順序による比較ができる。しかし加減などの演算には意味がない。

心理学や社会科学の測定のほとんどは順序尺度で行われる。例えば社会的態度や階級は順序水準で測定されるものである。また客の嗜好のデータもこれで表現できる。順序尺度の代表値は最頻値や中央値で表されるが、中央値の方が多くの情報を与える。順序尺度で測定されるデータは順序（または順位）データと呼ばれる。

●間隔尺度

対象に割り振られる数字は順序水準の性質を全て満たし、さらに差が等しいということは間隔が等しいということを意味する。つまり測定値のペアの間の差を比較しても意味がある。加減の演算にも意味があるが、尺度上のゼロ点は任意で負の値も使える。

例にはカレンダーの日付がある。値の間の比には意味がなく、直接の乗除の演算は行えない。とはいえ差の比には意味がある。代表値は最頻値、中央値あるいは算術平均で表され、算術平均が最も多くの情報を与える。間隔尺度で測定されるデータは間隔データと呼ばれる。摂氏または華氏で測る温度も間隔尺度である。

●比尺度

対象に割り振られた数字は間隔尺度の性質を全て満たし、さらにその中のペアの比にも、乗除の演算にも意味がある。比率水準のゼロ点は絶対的である。

比率尺度で測定される変数の代表値は最頻値、中央値、算術平均あるいは幾何平均で表されるが、間隔尺度と同じく算術平均が最も多くの情報を与える。比率尺度で測定されるデータは比率データと呼ばれる。比率尺度で表される社会的変数には年齢、ある場所での居住期間、収入などといったものがある。

□調査手法□

母集団を設定し、社会の全体像を把握するために大量のデータをとる調査を統計的社会調査という。この方法は、以下のように面接調査、留置調査、郵送調査、電話調査、電子調査などに分類される。

●面接調査●

面接調査は、調査者が調査対象者に直接会って質問を発し、回答を得る方法である。調査者が対象者に実際に会って行う為、データ一件あたりの費用が高くなる反面、身代わり回答や無回答が少なく比較的信頼性の高いデータを得る事ができる。

●留置調査●

留置調査とは調査票を一定期間対象者に渡しておき、後日に訪問して調査票を回収する方法である。調査票を郵送し、回収は調査員が訪問する場合は郵送留置調査と呼ばれる。

低コストだが、身代わり回答や無回答が多く、データの質は面接法と比べ、やや落ちる。ただし、家計調査や生活時間調査などにおいて、家計簿や日記などを見て、回答者が考えながら答えるのに時間がかかる場合は有効。

●郵送調査●

調査票を郵送し、郵送で返送してもらう方法。郵便代金だけで実施可能だが、通常、回収率は 3 割前後であり、学術調査としては不適切とされることが多い。ただし、質問数が少なく、依頼状を工夫し、返送先が大学で信用があり、何度か繰り返し調査票を送付した場合は、7 割前後の回収率となる場合もあった。

●電話調査●

電話をかけて質問を行い、結果を聴取する方法。通常、かけるべき電話番号はランダムに作成される。ランダムに作り出した番号に電話をかける手法は、RDD 法と呼ばれ、選挙の結果予測調査などによく用いられる。

電話調査の問題点としては以下の点が指摘されている。

- 日本では電話の普及率が高いので問題になりにくいが、普及率の低い地域では調査対象が母集団の標本たりえない場合がある。ダイヤル対象は固定電話であることが多いが、携帯電話の普及により固定電話の普及率は低下傾向にある。この傾向は都会の一人暮らし世帯や若年者に顕著であると言われている。日中では女性が電話に出る確率が高く、年代や性別等に偏りが発生することがあるため、標本の無作為性に疑問が抱かれる。
- 個人ではなく世帯対象になる。
- 相手の信用が分からないため協力してくれる人は多くはない。厳密に回収率を出せば10%以下となることもあり、回答の偏りがある。
- RDD 法では、ランダムに番号を作るもののあらかじめ掛けてはいけない番号を決めるのが普通である。これは警察などの公共機関に繋がるのを防ぐためである。そのため、番号の作り方に制限をかけることで、恣意的な運用が可能になる。

●電子調査●

インターネットで調査フォームを公開して、回答を募る方法。調査・集計が手軽で安価であるなどの理由で利用が増えている。回答者は事前に調査会社に登録している人の中から無作為抽出で選ばれることが多い。

この手法の問題点として以下の点が指摘されている。

- 男女区分や職業等の属性を、母集団比率を反映するように調査会社が調整していることが普通だが、現実には、回答者はインターネットを積極的に活用する層に限定され、回答者に偏りがある。実際には、年齢と性別を調整するだけのことが多いが、登録者の詳細な属性まではわからないばかりでなく、属性の正確性も保証されない。属性絞り込みを行わない方式では、コンピューター関連企業に勤める 40 歳代以下の男性が多くなるなど、回答の代表性に関する問題はさらに大きくなる。
- インターネット上では、個々人の同一性を識別できないため、同一人物が複数回回答することがありえる。調査会社では、認証などの手段で重複回答を避けるようにしているが、チェックは完全ではないし、調査謝礼目的で、一人で数十回も回答する人もいる。

□調査書の設計□

実際に調査をするにあたって質問の仕方、言葉の選び方によって得られる回答が異なってくる場合がある。特に言葉の定義は、それが日常的な意味とかけ離れていた場合、読者に誤った印象を与えることがある。

調査書を設計するにあたって注意が必要な点として以下が挙げられる。

- 「社員」や「理事」など、回答者によって多様な解釈が可能な言葉を用いることは回答バイアス(後述)を生むことにつながるので避けること。
- 面接調査や電話調査以外では、質問の意図を正確に伝えることが肝要なので、意味が分かりにくい言葉を使用しないこと。
- 「煙草は健康に悪いから、禁煙する方がいいと思うか？」のような、ダブル・バーレル、つまり二つの質問が混ざった質問を避けること。
- 回答者の回答意欲を減じて無回答バイアス(後述)を生むことにつながるので質問数をやたらと増やさないこと。
- 質問順序によっては図らずも回答が誘導されることがあること。
- 答えにくい質問や、質問者や社会への同調が少なからず現れそうな質問については、その点に留意して得られた結果を分析すること。
- 回答の選択肢では、回答者の意見を正確に反映しきれないということ。

□抽出方法□

標本調査とは、母集団をすべて調査対象とする全数調査に対して、母集団から標本を抽出して調査し、それから母集団の性質を統計学的に推定する方法である。その抽出方法によっては標本調査の信頼性が揺らいでしまう。ギャラップ社とリテラリィー・ダイジェスト社の明暗を分けたのもこれによると言われている。

●単純無作為抽出●

これは全要素を平等に扱い、分割はしない方法である。無作為抽出は、全ての要素の組合せの起こる確率がわかっている場合に有効で、標本が母集団を適切に代表しないリスクを、統計学理論により抽出に伴う誤差から計算し、適切な標本サイズを選ぶことができる。無作為抽出のうちで最も基本的な単純無作為抽出では、各要素を同じ確率で選び出す。これは理想的なものの実用的ではないので、抽出法に関して他の方法が用いられることが多い。

●系統的抽出●

最初のサンプルを任意に選び、そのサンプルから一定の間隔をあけたサンプルを順々に選んでいく方法。例としては、電話帳から 10 番目毎に抽出する方法がある。簡単ではあるが、データの非対称性と偏りから、結果の偏りが出やすい。電話帳自体が無作為化されていない限り非確率的抽出になるのだ。

●多段階抽出●

無作為抽出を段階的に繰り返す方法。例としては、まず都道府県を無作為抽出し、県の中で市町村を無作為抽出する等。

●層化抽出●

母集団を互いに重ならず、また一つの要素が複数の層に属さないように層別した後に、層ごとに抽出を行う方法。各層の標本サイズは層の標準偏差、あるいは母集団における層の占める割合に比例したものにする。各層は、平均が互いに十分異なり、分散が全体の分散よりは小さいように選ぶ。層化抽出法はしばしば標本誤差を減らし、サンプルの質を高めるが、母集団の層を適切に把握する必要があり、適切に層別できない場合は信頼に足る結果が得られない。

●有意抽出●

代表的と考えられる調査対象を抽出する方法。この方法では主観的判断に頼ることになるので、集団全体を適切に代表する標本を選ぶことは母集団の性質を十分理解しない限り難しい。

□偏り□

標本調査や社会調査によって得られたデータには常にいくつかの偏りがある。

●選出バイアス●

母集団の一部の要素が他よりも標本として選ばれやすい場合に、標本に偏りがあるという。この種の偏りは単純に標本数を大きくしただけでは取り除けず、抽出方法そのものを見直す必要がある。

●無回答バイアス●

無回答であることと回答があることが必ずしもランダムであるとは言えない事。この場合、質問にある一定の意味合いが込められていたり、答えにくい質問であったりすることがある。

●回答バイアス●

回答に与えられた選択肢または、得られた回答そのものが、回答者の意思を正確に反映しているわけではないという事。社会や質問者に対する同調による影響、解答順による影響等がある。

□代表値□

データ分析において、我々が出来ることはデータの要約、仮説の検定、推定などと限られている。データの要約の際、標本を適切に代表する値を考えることで、データの分布によって使い分けが必要だが、演算や操作、伝達においてある程度の客観性を持たせられる。

●平均●

もっともよく使われる代表値であり、その中でも特に算術平均を指す場合が多い。観測値に度数をかけた観測値の和を観測値の総数で割ることによって得られる。算術平均はその定義からデータの重心ともいわれる。データが単峰型であり、外れ値が無く、分布に歪みが無い場合に有用とされる。

●中央値●

観測データに大きな外れ値がある場合に有効な代表値。その値より大きな観測値と小さな観測値の数が等しくなるような値。

●最頻値●

データ分布の峰に対応する値で、単峰型でないデータ分布においては有効でない。

□相関□

二次元データにおける 2 変数間の関係を見るために、各観測対象を平面上にプロットして散布図を作成する方法や、分割表を作る方法がある。散布図や分割表から得られる直感的な 2 変数間の関係を裏付ける値がいくつか存在する。

●相関係数●

共分散を各変数の標準偏差で割ったもので、その定義から絶対値は 1 以下である。相関係数が正である時、正の相関関係があるといい、各変数は増減をともにする。相関係数が負である時はこの逆で、負の相関関係があるといい、片方が増加する時、もう一方は減少する。相関係数の絶対値が 1 であるときを完全相関であると言い、観測値はある直線上に完全に乗る。ただし相関係数の大きさは必ずしも、散布図から得られる直感的な理解と一致するわけではなく、散布図のスケールによってはこの逆が起きる場合もある。外れ値には滅法弱い。

この係数の問題点として以下の点が指摘されている。

- 相関係数が高いことから、2 変数間の強い相関関係が説明されるが、このことは必ずしも二つのデータ間の因果関係を説明するものではない。これは被説明変数と説明変数が定かではないからである。また、相関係数はあくまで直線的な関係しか説明できないのに対して、一般に因果関係ははるかに複雑な関係を内包している場合が多いのもその理由である。
- 2 変数に相関する外部変数がある場合、つまり交絡がある場合も相関係数は高い値を出してしまうことによって直接的に関係のない 2 変数にも見かけ上の相関が現れてしまうことがある。データを取る段階から直接的な関係が認められような 2 変数を取る必要があるのだ。
- 全体では相関係数が小さく相関が無いと考えられても、適切に層別することによって、各層で相関が現れる場合もある。

いずれの場合も調査者が標本を十分に理解し、変数や標本を適切に選べば起こることではないが、恣意的な相関関係を述べる温床となっている。

●ピアソンのX^2●

分割表の各セルの値の、2 変数の関連性が最も小さくなる値や、帰無仮説から導かれる理論的な値とのずれから、2 変数の関連性の度合いを示す値。具体的には、理論値との差の 2 乗を理論値で割ったものをセルごとに足し合わせる事で得られ、2 変数の関連性が高いほど大きくなる。ただしX^2はサンプル数が 2 倍になると、連動して 2 倍になるため、次のΦ係数を使用する事もある。

●Φ係数●

分割表における 2 変数の関連性の度合いを示す基本的な統計量。X^2を観測数で割った値の平方根として与えられる。Φ係数は 0 から 1 の間の値を取り、2 変数間の関連性が高いほど大きくなる。この係数は 2×2 分割表でのみ利用ができるのでこれ以外の場合では分割表の行と列の少ない方の本数を k として、X^2を観測数と(k-1)で割った値の平方根として得られるクラメールのV係数を使う。

□分析□

一次元データであるか二次元データであるか、あるいは多次元データであるかによって、データを分析する手法は様々だが、いずれにしても分析の目的に合ったモデルを返してくれるものを選ばなくてはならない。

●回帰分析●

データのうち一つを被説明変数として、被説明変数をいくつかの説明変数と誤差項の線形結合である回帰方程式で表そうとする手法。非線形な場合もあるが、往々にして線形モデルに帰着できる。説明変数が一つであるものを単回帰、複数あるものを重回帰とよぶが、いずれも回帰方程式の係数(回帰係数)は、回帰方程式から得られる当てはめ値と観測値の差、つまり残差の 2 乗和を最小にするように選ばれる。

説明変数が適切に選ばれているかの尺度として決定係数がある。これは回帰方程式が観測値を説明できている割合を示す値で、1 から残差の 2 乗和を観測値の全変動で割った値を引くことで求められる。

●判別分析●

いくつかの母集団が混ざっているグループの中から得られた観察対象が、どの母集団に属しているかを判別するための手法。どの母集団に属しているかがわかっている観察対象から変数を定めて、データを取り、その変数から成る線形判別関数を作る。線形判別関数の係数は各群を最も効率的に分類できるように、判別関数に各データを当てはめて得られた値である判別得点が、各群の中では分散が小さく、群と群の間では大きくなるように選ぶ。そうして得られた判別関数によって、判別を要する観察対象の判別得点を求めてどの母集団に属しているかを分析する。

●主成分分析●

ある観測対象について多数のデータが得られた場合、データを少数の指標にまとめるための手法。元の変数群を線形変換によって新しい変数群にして、新しい変数に関してデータを取り直すことで主成分得点を定めて、分析を進める。線形変換によって得られた変

数群のうち、主成分得点の分散が大きいものほど情報を保持していると言え、通常は最も大きいものか、二番に大きいものまでを使うことで、多次元データを一次元データあるいは二次元データに要約しなおす。

●因子分析●

ある観測値が合成データであると仮定して、それを構成する個々の変数を得て、その変数から説明しようとする手法。主成分分析と混同されがちだが、主成分分析が主成分という変数をつくり、主成分得点を求めるのにたいして、因子分析では観測値である因子得点から、それを構成する因子を得ようとする点で方向性がまるで逆である。

●クラスター分析●

観測対象に関するデータから、対象同士の類似度を考え、その程度に基づいてグループ分けをする手法。まず類似度を定義した上で、あるデータと最も類似度の高いものとを一つのグループとし、類似度に従って近いものを結び付けていき、一つのグループを徐々に大きくしていく方法を取る。

□練習問題□

Ⅰ　以下の質問文に問題点があれば全て述べて、自分なりに改善せよ。

・　「日本の景気は悪いので、日本の企業は内部留保を削減するべきだと思うか？」

Yes-No

・　「安倍首相の経済政策をどの程度評価するか？」

5-4-3-2-1

「支持政党を一つ答えてください」

・　「あなたはアドルフ・ヒトラーを尊敬しますか？」

Yes-No

Ⅱ　以下の調査手法に問題点があれば全て述べて、自分なりに改善せよ。

・　幻日君は電話調査を行って、日本の景気について調べることにした。まず、都道府県ごとに 100 人サンプルを取ることに決めて、各都道府県の電話帳をもとに最初の番号から順に一定の間隔をあけて電話をかけた。実際に調査をしたのは平日の午後 3 時から 4 時の間で、最初に電話に出た人に質問した。

・　幻日君は親子丼、かつ丼、牛丼の中でどの丼が一番人気なのかを調べるため、駅前で聞き込み調査を行った。東京駅、渋谷駅、巣鴨駅で聞き込みをすると決めて、聞き込みは朝、昼、晩と 3 つの時間帯に実施して、時間帯ごとに 100 人の回答を得た時点で聞き込みを終了した。回答者には各丼を好きな順に並べてもらい、順位が高かった

丼から 3,2,1 点と点を与えて、各丼の点数を集計し、最も点数が高いものを最も人気が高い丼とした。

Ⅲ　以下の分析に問題点があれば全て述べよ。

・　幻日君は会社に勤める男性の給料と血圧を調べることによって、血圧の高さと給料の高さには正の相関関係があることを発見して、血圧を説明変数とする回帰直線とともに、夫の血圧を上げれば、夫の給料が上がるという報告書をまとめた。

・　幻日君は国連に加入する全ての国に関して国力を調べるために、各国の一人当たりの GNP や、合計特殊出生率、平均寿命、軍事支出、義務教育の年数などを調べて主成分分析を行った。主成分得点から、 1 位の国と 10 位の国では国力が 2 倍違うことを発見して、報告書をまとめた。

Ⅳ　以下の語句を説明せよ。

・ Φ係数
・ 選出バイアス
・ 層化抽出法

解答例は特に設けない。

□注意□

　気を付けてもらいたいのは以上の解説を一度も授業に参加していない者がしていることである。ここまで読んだ人々には申し訳ないが、ここにある内容が授業内容と一致しているかは完全に不明である。そのため運用はほどほどにするのがよいだろう。

□参考文献□

東京大学教養学部統計学教室(2013)　統計学入門
縄田和満(2014)　Excel 統計解析ボックスによるデータ解析
http://www.stat.go.jp/teacher/c2hyohon.htm
http://ja.wikipedia.org/wiki/%E5%B0%BA%E5%BA%A6%E6%B0%B4%E6%BA%96
http://ja.wikipedia.org/wiki/%E7%A4%BE%E4%BC%9A%E8%AA%BF%E6%9F%BB
http://ja.wikipedia.org/wiki/%E6%A8%99%E6%9C%AC%E8%AA%BF%E6%9F%BB
http://ja.wikipedia.org/wiki/%E5%88%86%E5%89%B2%E8%A1%A8