データ作成に反映される人間

椙山女学園理事長
森棟公夫
平成25年1月7日

概要

本稿は拙著『教養　統計学』（新世社2012）などから統計学に現れる人間の特色を取り出して、説明する。統計分析には人間が係わるが、データ作成には、人間が有する偏りが意識的あるいは無意識に反映される。そのようなデータでは、分析結果は正確にはなりえない。逆に、分析対象をランダムに選ぶならば、データさえ大きければ、大数の法則が成立し、真の関係が見つかる可能性が高まる。しかし、分析対象をランダムに選ぶためには、乱数とか無作為抽出といった人間の知恵によってもたらされた技術が新たに必要とされる。

1　序文

統計分析には人間が係わるが、データ作成においては、人間が有する偏りが意識的あるいは無意識的に反映されてしまう。この様な偏りを排除するには、無作為標本という人間が考え出した技術が必要となる。

データは統計分析の根幹である。データを取るためには実験や調査が不可欠だが、実験や調査の対象はランダム（random）に選ばれていないといけない。ランダムに選ばれた調査対象から得た調査結果の集合を無作為標本という。標本という用語を避け、数値の集まりをデータということもある。データが無作為に作成されていなければ、分析対象の真の性質を見い出すことができない。データが無作為に作成されていれば、分析を行う時に、大数の法則が成立する。大数の法則とは、たとえば分析対象の中心の値を計算しようとする場合なら、データが大きければデータの平均は真の平均に近い可能性が高くなるという性質である。

データは時に恣意的に作成されるが、多くの場合、恣意性は分析者の意図によって生じる。分析者が自分の恣意性に気づいていない場合もある。さらに、調査対象が無作為に選ばれても、測定が公正に行われないこともある。本稿ではまず、データ作成において生じるこのような偏りを例を用いて説明する。次に、データが無作為な場合に成立する大数の法則の意味を、実験によって説明する。

2　偏ったデータ

偏ったデータとは、無作為（バラバラ）にとられていないデータだけを意味するのではなく、とり方に癖があったり、調査を行う者が望む結果が生まれるように仕組んだデータを含む。東北大震災の後、九州電力によるやらせメール事件が世間を賑わした。玄海原子力発電所（佐賀県玄海町）2、3号機の再稼働を巡りメールによる世論調査を行っていたところ、九州電力は原発再稼働賛成のメールを送るよう関連会社に依

頼し、そのおかげで再稼働賛成が過半数を超えたということだ。典型的な仕組まれた標本だが、やらせが発覚し世論は再稼働に厳しい反応を示した。仕組まれた標本では、賛否の正しい比率は決して分からない。ここで、統計学でよく知られている偏ったデータの例を見よう。

2.1 大統領選挙の誤った予測

最も有名な例は、1936年に行われたアメリカ大統領選挙の予測である。当時、アメリカ最大の調査会社(Literary Digest、LD社と略す)は、共和党のアルフレッド・ランドン候補が圧倒的な勝利を収めると予測した。対立候補は民主党のフランクリン・ルーズベルトであった。ルーズベルトは、後に大不況を解決するために行ったニューディール政策などで有名になった大統領である。

選挙予測では、LD社は、自社の定期出版物の購読者、電話を所有している人、車を所有している人などを調査の対象として、当時の有権者の(1/5)の千万人に郵便葉書によるアンケート調査を実施した。回答が得られたのは(1/5)だった。結果はランドン129万票、ルーズベルト97万票となり、ランドンの圧倒的な勝利を予測したが、この予測は誤りだった。正しい予測をしたのは小規模なランダム標本を利用した対立会社で、この会社は大きな名声を得た。この対立する会社が今日も名を聞くギャラップ(Gallup)社である。調査費用も割安だった。

LD社の調査は、1936年に電話を持っている、車を持っているという条件から理解できるように、裕福な階層に偏っていた。自社出版物の購読者というのも政治的な偏りを示している。しかしこの例では、LD社は意図的に偏った市民を選んでアンケートを行ったのではない。当時ではランダムに選べば正しい結果を導き得るといった知識が普及していなかったと考えることができる。

2.2 化粧品のアンケート調査に騙されるな

仕組まれた標本は商品の宣伝で多く使われる。「お客様のアンケート調査の結果、95%の方々にご満足いただいている」といった類の広告である。

誇大広告の批判を避けるために会社は何らかのアンケート調査を行っており、それを根拠として宣伝をする。しかし、アンケートで顧客が選べる選択は、「満足」、「おおよそ満足」、「全く不満足」の三つくらいしか与えられない。そうすると、アンケートに答える顧客は「満足」、「おおよそ満足」の人と、「全く不満足」の人だけになる。わざわざ参加が自由なアンケートに「全く不満足」と答える人は喧嘩でもしたい人だけだから、回答者のほとんどが「満足」、「おおよそ満足」の人達となり、その結果、9割以上の顧客は満足しているという広告が生まれる。仕組まれた標本の好例である。

数字のマジックもあるので、この例は詳しく見てみよう。例えば、「満足」、「おおよそ満足」と答える人が60%、「全く不満足」の人が3%とする。三択以外の37%は無回答となり、除かれる。そうすると、「満足」或いは「おおよそ満足」という回答の比率は

$$\frac{0.60}{0.60+0.03} \fallingdotseq 0.95$$

となる。ここでのトリックは、「全く不満足」が3%で非常に少ないということである。割合で計算すると理解が難しいかも知れないが、500人中300人が「満足」あるいは「おおよそ満足」、

15人が「全く不満足」と答えたとしても同じである。

もし、「全く不満足」が1%しかないと、「満足」あるいは「おおよそ満足」が20%あれば、

$$\frac{0.20}{0.20+0.01+0.01} \fallingdotseq 0.95$$

という計算になるから、アンケート結果は95%の消費者の支持を得ていると書くことができる。

ここでの問題は、少々不満足、不満足といった感想を持つ消費者がアンケートから外されるように標本が仕組まれているということである。こういった感想を持つ消費者が例えば10%いたとしても、商品に「満足」あるいは「おおよそ満足」する人は

$$\frac{0.20}{0.20+0.10+0.01} \fallingdotseq 0.65$$

だから、65%に低下する。

2.3 有能な不動産屋のPR

2011年7月6日に放映されたNHKの番組「ためしてガッテン」で、大阪大学の近所の或る不動産屋さんは、大阪大学の受験生に入試の日からアパートを斡旋しているという話があった。「発表前に契約をした人の合格率は9割を超える」という宣伝文句を使っており、もし入試に落ちたら無料でキャンセルでき、斡旋費用は不必要だという。入試の合格率は3〜4割だから、このような商いをすると、不動産屋さんは空きアパートばかり抱えてしまって商売が大変ではないかと想像される。ところが、発表前に事前契約をする受験生に関しては、合格率が実際に9割を超えるそうだ。一般の合格率が3割であるのに比べてアパートを発表前に契約をした受験生の合格率が9割を超えるのだから、不思議である。また、誇大広告はなく、空き室が増え不動産屋さんの商売に支障がでることもない。

不思議な現象に見えるが、これが偏った標本の例になっている。なぜなら、発表の前にアパートの契約をする受験生は、自主的な判断だが、合格する自信がある受験生になっているからである。つまり、受験生が事前に契約をすれば合格率が9割になるのではなく、合格しそうな成績を取った受験生が事前契約をしている。この宣伝は、仕組まれた標本ではなく、受験生の心理を掴んだ不動産屋さんの巧みな商いである。試験に自信が無い人もどんどん契約するようになるとランダム標本になるが、そうなると不動産屋さんはキャンセル料を要求するようになるだろう。（この例は、大阪大学工学部狩野裕教授の資料を参考にした。）

2.4. 身長測定の不備

最後に、調査対象が無作為に選ばれていても、測定によって不備が生じる例をみる。測定結果を被調査者が自己申告する場合では、このような不備は容易に生じる。例えば、もし所得が自己申告であり、それに対して税金を支払うのであれば、所得の申告額は実際の所得より低くなるのは当然であろう。このような不備があるデータは多く存在すると思われるが、統計学で有名な例と、最近、筆者が見つけた日本の例をここで紹介する。

2.4.1 徴兵検査

ケトレー（1844）は、男性に関する身長分布が左右対称なベル型（正規分布）になることに気づいた。そして、この性質から、フランス男性の徴兵逃れを見つけ出した。徴兵検査において

身長を調べ、測定結果から身長のヒストグラム（棒グラフ）を作れば、滑らかなベル型（正規分布）になるはずである。しかし、もし特定の身長区間の密度（棒）がその前後の区間より一際目立つ程高くなっていれば、測定に不備があると判断することができる。このような不自然に高い棒をコブ（凸）と呼ぶが、ケトレーは徴兵検査の対象となった男性について、そのようなコブを発見した。つまり、徴兵を逃れるために、徴兵検査の対象となった男性の多くが、徴兵基準に満たないよう身長を低く申告していたのである。（『統計学とは何か』（ラオ著、藤越、柳井、田栗訳、ちくま学芸文庫、71頁、2011年）この例に関する実際のデータは手に入っていないし、また身長が自己申告であったのか或いは測定された値であったのかも分からない。しかし、もし測定された値だったとすれば、検査に来た若者が膝を僅かに折り曲げるといった背を低く見せる工夫をしていたと考えられる。

2.4.2 身体測定

図1は、「学校保健統計調査」（文部科学省）平成13年から22年までの10年分から作った高3男女に関する身長のヒストグラムである。調査対象者は男女ともほぼ20万人に達し、区間の幅を1cmとしても滑らかなヒストグラムを求める事ができる。この10年間は、各学年の平均身長の変化がほとんどない事が分かっているのだが、1cmごとの度数分布表を作ると、ケトレーが見つけたような不備が女子については159～160cm区間ではっきりと見つかる。男子でも、169～170、179～180cm区間で不備が見つかる。これらの区間にコブ（凸）があり、その直前の区間が凹になっていることが、ヒストグラムから確認できよう。図中の数値は、1cm区間の中点である。

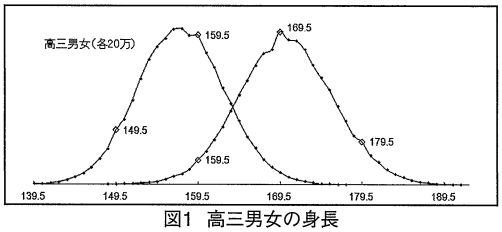

図1 高三男女の身長

女子では159～160cm区間、男子では160～161cm区間にはっきりとしたコブが見えるが、同じく、女子では149～150cm、男子では159～160cmと179～180cm区間にも小さなコブが見える。これらのコブを無視して滑らかな曲線を描けば、ベル型（正規分布）になる。この現象は、高3だけでなく他の学年でも見つかる。図2と図3の原資料と作成方法は図1と同じである。図2は高2男女だが、分布は高3とあまり変わらない。特に、コブがある区間がほぼ同じであることに注意して欲しい。

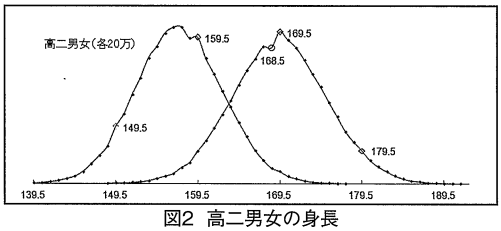

図2 高二男女の身長

図3は高1男女の身長分布だが、高1でもだいたい同じ区間にコブ（凸）がある。特に男子の169～170cm区間ははっきりとしている。

これらのコブは各学年の特徴に過ぎないとすると、次の様な矛盾が生じる。データは2001年から2010年までの10年であるが、2001年から2009年の高1は、2002年から2010年の高2になっており、高2データの9割は高1データと同じ生徒の身長である。同様なことは、高2と高3生徒

についても言える。高1と高3では、8割が同じ生徒である。

そうすると、高1男子の169cm～170cmのコブ（凸）は、翌年はどうなるのであろうか。平均身長は高1から高3では167.7cmから170.2cmへ2.5cm伸びるので、1年に1cmほどは伸びており、同じ区間にコブができるのは不自然である。高1の169cm～170cmのコブが、高3では172～174cmに移っていれば理解もできるが、区間が同じ169cm～170cmであるのは測定が不備であるいう以外の説明ができない。男子では170cmを目前にして、多くの生徒が身長を高く見せるために踵を上げているとするのが自然であろう。他の区間についても同様である。

女子についても高3ほどではないが、高1に同様のコブが見られる。平均身長は高1の157.1cmから高3の157.4cmへ3mm伸びるだけだが、159～160区間のコブが目立ってくる。これも、背を高く見せるために踵を上げる生徒の割合が多いと理解するのが自然であろう。

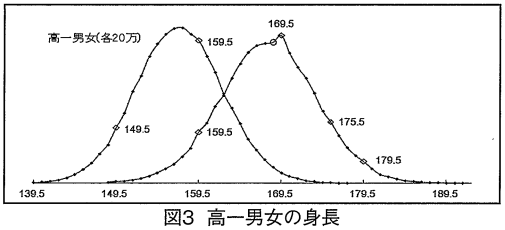

図3 高一男女の身長

3 無作為な標本と大数の法則

無作為に選ばれたデータを使うと、データから計算される平均はすばらしい性質を示す。硬貨を繰り返し投げる例を用いて説明するが、以下では、この様な例を実験と呼ぼう。硬貨投げの実験では、表が出る比率は平均になっている。表が出れば1点、裏が出れば0点とすれば、合計点数は表が出る回数だから、平均点数は表が出る比率になる。そして、千回ほど硬貨を投げ表が出る比率を求めれば、大数の法則により、平均は真の比率に近い値になる。標本が大きければ（データのサイズが大きければ）、真の比率が大体分かるというのが大数の法則である。

3.1 大数の法則

新たに千回投げ直すと、異なる比率が求まる。千回投げるのも大変だが、千回投げを3回繰り返すと3回とも答えが違う。したがって、千回投げの実験を繰り返しても、真の比率は分からない。

大数の法則とは、投げる回数を増やしていくと、平均が、真の値の例えばプラス・マイナス0.1以内に入る割合が高くなるという性質である。逆に言えば、平均が、真の値から0.1以上外れる割合が低くなる性質と言うこともできる。数学の法則だから、何回投げればどのくらいの比率で0.1以内になるか、あるいは0.1以上外れるかは分からない。真の値も分からない。しかし、千回投げを3度繰り返すと、真の値は分からないにしろ、真の値の予測は千回投げ1度より確実になる。

3.1.1 男女比率

男女の比率は等しいという常識がある。しかし、小さな小学校などでは男女比率は等しくならない。或る小学校の実際のデータだが、男女の順で1年生から6年生までが、(11,13)、(14,14)、(15,15)、(5,15)、(18,12)、(13,16)となっていた。全体では(76,86)である。

小さな小学校の新入生が50人だったとする。男女は25人ずつ分かれるのが当然かもしれないが、男子20、女子30といった風に10％くらいのずれが起きても不思議はない。しかし、学校

全体が500人だとすると、全体で10%のずれが起き、男子が200人、女子が300人となる事はほとんど起きない。もちろんある市の子供全体で見てみると、男女比が2分の1から10%も外れることは絶対に起こらない。調査の対象が多くなるほど男女比は(1/2)に近づいていくが、真の男女比は決して求まらない。

3.2 実験

実際に硬貨を投げてみるとどうなるだろうか。表を1、裏を0として、表が出る比率(平均)を求める実験をした。

3.2.1 実験1(25回投げ)

表は1、裏は0とし、硬貨を25回投げて平均を求めてみる。1回目のランダムな標本は{1,1,1,0,1,0,0,0,0,1, 1,0,0,0,0,0,0,0,1,1, 1,0,1,0,0}となった。標本(サンプル)の大きさは25である。25回のうち10回は表だから、平均は0.4だった。平均は、表の回数を25で割った値だから、比率になっている。標本が1組とれれば平均が1個計算できる。

このようにランダムな標本をとり、平均を計算するという手続きを繰り返す。ランダムな標本が1組とれれば平均が1個計算できるだけだから大変だが、25回投げ実験を500回繰り返して、平均を500組求めた。平均の値を500個書き出しても意味が無いので、結果を相対度数分布(区間の比率)として表1、5行目にまとめた。表の2行目と3行目は区間と区間の中点である。

表1 平均の相対度数分布(区間の比率)

区間と中点(区間は下限超～上限以下.15を.15などと略した。)						
.15～.25	.25～.35	.35～.45	.45～.55	.55～.65	.65～.75	.75～.85
0.2	0.3	0.4	0.5	0.6	0.7	0.8
実験1(25回投げ)						
0.018	0.048	0.31	0.29	0.276	0.052	0.006
実験2(1000回投げ)						
0	0	0.002	0.996	0.002	0	0

25回の硬貨投げから求まる平均は、0.15から0.85までの比率を取ることが分かる。区間中点が0.5である中央の区間には29%、その両側を入れた区間(0.35～0.65)には、ほぼ88%が含まれる。しかし、その左右の区間の比率も0ではない。したがって、硬貨を25回投げただけでは表が出る真の比率を予想することは難しい。そこで、硬貨を投げる回数を25回から千回に増やしてみた。

3.2.2 実験2(千回投げ)

硬貨を千回投げて平均を求めた。標本の大きさは千、手順は実験1と同じである。表が出る比率は平均だから、ランダムな標本は千個の値を含むが、スペースを取るので実験1のようにランダムな標本の例は示さない。大変だが、この実験も500回繰り返し、相対度数分布表を表1の7行目にまとめた(実際のところ硬貨をこれだけの回数投げるのは無理である。Excelを使って計算をしている)。中央の区間(0.45～0.55)から外れるのは500回のうち2回だけである。500回

のうち498回は(0.45～0.55)区間に入っている。真の比率は(0.45～0.55)区間に入るようだ。千回投げ実験を1回しかしなくても、かなり真の値に近い結果になる。

3.2.3 実験のまとめ

真の比率を探すには、硬貨を投げる回数が重要である。統計用語では、標本（データ）の大きさ（サイズ）と言う。25回では平均の値はばらついて、真の比率が(0.45～0.55)区間に入っているとは予想できない。千回投げれば平均は真の比率にかなり近くなるようだ。真の比率は中央の区間(0.45～0.55)に入ると予想できる。区間を(0.45～0.55)に固定し、平均がこの区間から外れる割合を求めると、硬貨投げの回数が増えるほど外れる割合は減少する。千回投げでは(2/(500))である（1万回投げの実験では0であった）。大数の法則とは、データの大きさが大きいほど、中央の区間から外れる割合が減少していくという性質である。データの大きさを非常に大きくすれば、中央の区間を100分の1の狭い区間にすることもできる。大数の法則が成立するには、硬貨投げのように人が影響を及ぼすことができない実験が不可欠である。硬貨投げの結果は、無作為なデータになっている。

4 結論

調査を行う者の好みや癖などによってデータが偏ってしまうと、データの大きさ（サイズ）にかかわらず真実を見いだすことができない。逆に、データが無作為であれば、大数の法則により、データの平均が中央の区間から外れる割合はどんどん減っていく。データさえ大きければ、真の関係を高い確率で予測することができる。

ここでは無作為という条件が不可欠であり、人間の恣意性がデータの作成に含まれないならば大数の法則が成立する。つまり、無作為データの作成において人間性が無いことが、統計学に不可欠である。しかし、無作為なデータを作成することは容易ではない。単に調査対象をデタラメに選ぶだけではデータが無作為になる保証はなく、乱数とか無作為抽出といった人間の知恵を利用する事によって、初めて無作為なデータを作成することができる（本稿では、乱数の作り方などは説明しない）。結局、人間の恣意性を除くためには、新たに人間の知恵を用いることが必要となる。