標本調査

データを整理し、傾向をつかんだり推測したりするとき、いちばん確実なのは**全数調査**（悉皆（しっかい）調査とも言う）を行うことである。しかし、いつでも全数のデータが手に入るわけではなく、また、全数のデータが必ずしも必要なわけでもない。調査対象の全体を**母集団**と呼ぶが、母集団から一部の**標本**を抽出して全体を推測する**標本調査**が現実的な選択であることは多い。

ただ、標本調査では注意すべきことがある。標本として抽出するデータに偏りがあっては困るのである。そのためには**乱数**を用いて母集団から標本を選ぶようにするとよい。これは**無作為抽出**と呼ばれ、一般に標本調査は無作為抽出された**無作為標本**を用いるものである。

細かいことを言えば抽出の仕方は 2 通りある。ひとつは抽出の度に要素をもとに戻して抽出を続ける方法で、もうひとつは抽出した要素をもとに戻さず抽出を続ける方法である。どちらでも同じように見えるが、要素をもとに戻す抽出から得られたデータは独立であるが、要素をもとに戻さない抽出から得られたデータは独立ではない。なぜなら、要素をもとに戻さないことで、残されたデータのある要素が選ばれる確率が影響を受けるからである。しかし、母集団の要素数が少ない場合はそうかもしれないが、母集団の要素数が極めて大きければその心配はない。実際、母集団の要素数が極めて大きい場合は、抽出されたデータを独立とみなしてよいことが知られている。

標本平均、標本標準偏差

大きさ N の母集団の分布から得られる**母平均**を m とする。また、母集団から実際に得た、大きさ n の無作為標本 $(x_1,\ x_2,\ \ldots,\ x_n)$ の**標本平均**は $\overline{X} = \dfrac{1}{n}(x_1 + x_2 + \cdots + x_n)$ である。当然、母集団の平均 m と標本平均 $\overline{X}$ の値は若干異なるであろう。しかし、大きさ N の母集団から n 個の標本をとる取り方は ${}_NC_n$ 通りあるので、とった標本の平均を見積もることにして

$$\boxed{E(\overline{X}) = m}$$

でよいだろう。つまり、${}_NC_n$ 通りの中の i 番目の無作為標本の平均 $\overline{X_i} = \dfrac{1}{n}(x_{i_1} + x_{i_2} + \ldots + x_{i_n})$ の期待値は、そのまま母集団の母平均 m に等しい。ただし、ひとつひとつの標本平均 $\overline{X_i}$ は、m に対してばらついている。しかし、無作為に標本をとっているので、ばらつきの確率分布は母集団の確率分布にしたがうと考えてよい。すなわち、無作為標本の確率変数 x_i の分散 $V(x_i)$ は母集団の母分散 σ^2 に等しいのだが、標本平均 $\overline{X}$ の分散 $V(\overline{X})$ とは違うものを指していることに注意さ

れたい。大きさ n の標本平均の分散 $V(\overline{X})$ とは $\overline{X} = \dfrac{1}{n}(x_1 + x_2 + \cdots + x_n)$ の分散のことであるが、前節の確率変数の性質により、

$$\begin{aligned} V(\overline{X}) &= V\left(\frac{1}{n}(x_1 + x_2 + \cdots + x_n)\right) \\ &= \frac{1}{n^2}\{V(x_1) + V(x_2) + \cdots + V(x_n)\} \end{aligned}$$

ということである。そして、$V(x_i) = \sigma^2$ であったので、

$$V(\overline{X}) = \frac{1}{n^2}\left\{n \cdot \sigma^2\right\} = \frac{\sigma^2}{n}$$

が成り立つ。よって、

$$\boxed{V(\overline{X}) = \frac{\sigma^2}{n}}$$

が言える。これは**中心極限定理**と呼ばれ、標本の大きさ n が大きいほど正規分布 $N\left(m,\ \dfrac{\sigma^2}{n}\right)$ に近づく。具体的な n の大きさとしては、母集団は 100 程度以上、標本数は 30 程度以上で実用的な分布になるようである。

ちなみに、母集団は必ずしも正規分布でなくとも、標本の分布は正規分布に近づくことが知られている。ちょっと意外な感じもするが、以前のデータ G のヒストグラムを見ながら考えると理解しやすいかもしれない。標本は母集団から独立にとっていくわけだから、データが集中しているところの値が選ばれやすいはずだ。データ G は 300–400（万円）が最頻値であるから、ここの値がいちばん選ばれやすい。続いて 200–300（万円）、400–500（万円）、. . . の順に選ばれやすいだろう。そうすると標本の分布は、300–400（万円）を中心に左右対象になるように思えるだろう。

* * *

標本平均については**大数の法則**が関与している。大数の法則とは、母平均 m の母集団から、大きさ n の標本を無作為抽出したとき、標本平均 $\overline{X}$ は n が大きいほど m に近づく、というものである。n は最大で母集団の要素数まで増やせて、そのとき標本は母集団と一致する。この様子を実感するには、Microsoft Excel はちょうどよい。

◇	A	B	C	D	E	F
1	(※ **A1**)	(※ **B1**)				
2		0.1				
3	(※ **A3**)	(※ **B3**)				
4	↓下へコピーする	↓下へコピーする				
5	↓	↓				
6	↓	↓				

※ セルの式

(**A1**)　=**AVERAGE(A3:A100)**
(**B1**)　=**AVERAGE(B3:B100)**
(**A3**)　=**100*RAND()**
(**B3**)　=**IF(RAND()<B$2,A3,"")**

A 列には母集団として A3 セルから A100 セル程度まで乱数を表示させる。ここでは、0 以上 100 未満の実数をとるようにした。B 列には標本を A 列から取ってくるのだが、選ぶかどうかは B2 セルの値で決まる。B2 セルの値が大きいほど A 列の値が選ばれやすい。

こうすることで、標本数 n を大きくするにしたがい、標本平均 $\overline{X}$ が A1 セルの母平均に近づく様子が見られるだろう。標本数を大きくするために B2 セルの値を 0.5, 0.6, 0.7, ... と増やしていけば、なんとなく様子がつかめるはずである。▌

標本平均の標準化

さて、無作為標本の平均と分散を求めたわけだが、実用的な指標としては分散より標準偏差を用いたい。標本平均の標準偏差 $\sigma(\overline{X})$ は $\sqrt{V(\overline{X})}$ のことであるから

$$\boxed{\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}}$$

である。この分布は正規分布 $N\left(m\ , \frac{\sigma^2}{n}\right)$ にしたがうといってよい。

分布が正規分布にしたがうのであれば、それは標準正規分布 $N(0,\ 1^2)$ へと標準化することができる。この場合は、$Z = \dfrac{\overline{X} - m}{\sigma/\sqrt{n}}$ とすれば、n が大きいときにほぼ $N(0,\ 1^2)$ とみなすことができる。

母平均の推定

母集団から得られた標本が標準正規分布とみなせるなら、標本をもとに母集団の推定を行えるだろう。前節で、標準正規分布において、

$$\begin{aligned} P(-2 \leqq X \leqq 2) &\approx 0.9545 \\ P(-3 \leqq X \leqq 3) &\approx 0.9973 \end{aligned}$$

であるという話をした。これより、平均 0 から標準偏差 1 以内には、データの 95.45%が含まれることが分かる。標準偏差 3 以内なら 99.73%だ。

逆の見方をすれば、データの 95.45%が含まれるデータの範囲は、平均 0 ± 2 に収まるのである。このようなことは、あるデータが集団からかけ離れているかどうかを見極める役に立つ。あるデータの値が、95.45%の確率で一定の集団に含まれるかどうかを知りたければ、データの値が平均と比べて標準偏差の 2 倍以内にあるかどうかを調べればよい。この範囲を**母平均に対する信頼区間**というが、一般に 95.45%の確率という中途半端な値は用いない。

よく使われるのが、母平均 m に対する**信頼度 95%の信頼区間**、または、**信頼度 99%の信頼区間**である。信頼度 95.45%で標準偏差 2 以内の全てのデータが該当するので、信頼度 95%であればそれよりほんの少し内側のデータが該当するだろう。実際、それは標準偏差がおよそ 1.96 以内の範囲である。また、信頼度 99.73%で標準偏差 3 以内の全てのデータが該当するので、信頼度 99%であればそれよりだいぶ内側のデータが該当するだろう。実際、それは標準偏差がおよそ 2.58 以内の範囲である。

では、どのようにして 1.96 や 2.58 という値を特定できたのだろう。それは

$$P(-c \leqq X \leqq c) = 2\int_0^c \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = 1.96 \text{ や } 2.58$$

を解いて求めるのである。と、簡単に言ってみたものの、実は e^{x^2} は初等関数による積分はできないのである。そこで数値積分などの特別な方法で、正規分布表が作成されている。

さて、ここで話を少し戻すが、信頼度 95%となる確率は $P(|Z| \leqq 1.96) = 0.95$ であった。Z は標本平均 $\overline{X}$ をもとに、$Z = \dfrac{\overline{X} - m}{\sigma/\sqrt{n}}$ で変換してものだったから、$|Z| \leqq 1.96$ であることは $|\overline{X} - m| \leqq 1.96 \times \dfrac{\sigma}{\sqrt{n}}$、すなわち

$$\overline{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} \leqq m \leqq \overline{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

である。これは、母平均 m がこの範囲にある確率が 95%であることを意味し、比較的高い確率で母平均を推定しているのである。

だだ、母集団の標準偏差 σ が分からない場合、要するにそれが分からないから標本をとるのだが、標本の大きさ n を十分大きくとることができれば、母標準偏差 σ を標本標準偏差 s に置き換えることは可能だ。実際、そのようにして、母平均 m に対して信頼度 95%の信頼区間は

$$\left[\overline{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \overline{x} + 1.96 \times \frac{s}{\sqrt{n}}\right]$$

で見積もってよいことが知られている。

*　　*　　*

正規分布表に頼らずに、信頼度 t%の信頼区間に相当する標準偏差 c の値を求めることはできるだろうか。現実的には、WolframAlpha に頼るのがよかろう。たとえば信頼度 95%を与える標準偏差 c の値は WolframAlpha のサイトで

```
= solve{1/sqrt(2 pi) integral_0^c e^(-x^2/2)dx=0.475, c}
```

と入力すればよい。すぐさま、`c ≈ 1.95996398454005...` と返してくれる。至れり尽くせりである。▌

母比率の推定

母平均の推定をしたところだが、これだけでは大した意味はない。もう一歩進めて、**母比率**が推定できるとよい。母比率とは、母集団の中である性質 A を持つ要素の割合 p を指す。母集団から直接母比率を求めることができればよいのだが、大抵は標本から推定することになる。すなわち、標本の中に性質 A を持つ要素の割合を求め、そこから母集団の中の性質 A の割合を推定するのである。では、母比率はどのようにして推定するのだろう。

まず、性質 A を持てば 1、持たなければ 0 である確率変数 X を考える。その場合の母平均は $E(X) = 1 \times p + 0 \times (1-p) = p$ であるから、$E(X)$ がそのまま母比率 p になっている。

母集団から抽出した大きさ n の標本からは、性質 A を持つ標本を正しく選別できるので、性質 A の標本比率 p_0 と標本平均 $\overline{x}$ は確実に求められる。であれば、$\overline{x} = 1 \times p_0$、$\overline{x^2} = 1^2 \times p_0$ より、標本標準偏差 s は、

$$s = \sqrt{\overline{x^2} - (\overline{x})^2} = \sqrt{p_0 - p_0^2} = \sqrt{p_0(1-p_0)}$$

となる。このことから以下の結論が導ける。

母集団から十分な大きさ n の標本を抽出すると、母比率 p に対する...

信頼度 95%の信頼区間は、$\left[p_0 - 1.96\sqrt{\dfrac{p_0(1-p_0)}{n}},\ p_0 + 1.96\sqrt{\dfrac{p_0(1-p_0)}{n}}\right]$

信頼度 99%の信頼区間は、$\left[p_0 - 2.58\sqrt{\dfrac{p_0(1-p_0)}{n}},\ p_0 + 2.58\sqrt{\dfrac{p_0(1-p_0)}{n}}\right]$

具体的に何に応用できるのだろう。たとえば有権者 10 万人から十分な人数を任意抽出して、候補者 A に投票するかどうか尋ねたとしよう。このとき、1,000 人の抽出に対して 650 人が投票する、350 人が投票しない、と答えた。この場合、（標本中では）投票する意思を持つ有権者の比率は $p_0 = 0.65$ である。したがって、信頼度 95%の信頼区間は

$$\left[0.65 - 1.96\sqrt{\frac{0.65(1-0.65)}{1000}},\ 0.65 + 1.96\sqrt{\frac{0.65(1-0.65)}{1000}}\right] \approx [0.62,\ 0.68]$$

である。このことから、有権者の 62%から 68%ぐらいの人が候補者 A に投票すると推定される。選挙当日の投票率が仮に 55%であれば、少なくとも

$$100{,}000\text{（人）} \times 55\% \times 62\% = 34{,}100\text{（票）}$$

の票が獲得できると見込まれる。