第 8 章 母集団とその推測

浅川 伸一

2006 年 06 月 08 日

1 中心極限定理 (例題)

コイントスを 400 回行った結果 215 回表が出た。このコインの表の出る確率は $p = 1/2$ と言って良いか。

問題はコイン投げなので二項分布を使う必要があるが、中心極限定理により正規分布で近似してもよい。表の出る確率を 1/2 とすると 400 回コインをトスしたときの二項分布の期待値は $np = 400 \times 1/2 = 200$ である。分散は $npq = np(1-p) = 400 \times 1/2 \times 1/2 = 100$ となる。標準偏差は $\sqrt{100} = 10$ となるので、表の出た回数 215 を標準化得点 (z スコア) に変換すると $(215 - 200)/10 = 1.5$ である。この時の確率を

$$\begin{aligned} \int_{-\infty}^{1.5} N(x; 0, 1^2)dx &= \int_{-\infty}^{1.5} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx && (1) \\ &= \text{normdist(1.5,0,1,1)} && (2) \\ &= 0.93 && (3) \end{aligned}$$

となるので、400 回コイントスして 215 回以上表の出る確率は $1 - 0.93 = 0.07$ となる。後はこの数値 0.07 を十分に大きいとみなすか否かの問題である。一般に 0.05 か 0.01 を基準に考える。

中心極限定理は強い定理であるが証明するのは難しい。一石賢 (2004)「道具としての統計解析」118 ページから 130 ページに証明が載っている。この証明では、積率関数、キュムラント母関数、およびマクローリン展開が使われている。しかし、正規分布を仮定したときの例であり、全ての分布があてはまる証明にはなっていない。

2 無限母集団 (補足)

例えば、生後 36 ヶ月児の語彙数を調査したとすると、想定される母集団は、すべての 36 ヶ月児である。これは将来産まれてくるかも知れない全ての子どもを考えれば、母集団は無限母集団である。

サイコロは、その気になりさえすれば何回でも振ることができるので、無限母集団である。

実験心理学 (視知覚、認知、学習など) では、ある被験者、被験体のある実験条件における反応全体と考える。この場合も実験を気の済むまで繰り返すことができるので、いくらでもデータを集めることができる。ゆえに、無限母集団と仮定される。

演習：今、自分のもっとも興味のあるテーマでデータを集めてレポートを書くとしたら、そのときの母集団は何か。それは有限母集団か、それとも無限母集団か。

3　母集団と統計量 (補足)

母集団の分布を表す母数と標本統計量とを明確に区別しなければならない。このアイデアはフィッシャー (Sir Ronald Aylmer Fisher) に由来する。標本平均、標本分散、標本標準偏差は具体的に計算可能である。一方、母平均、母分散、母集団は概念上の数であって、その数を推測することしかできない。

そこで母集団分布の母数を推測するためにデータが集められる。つまり心理実験やアンケートや面接などによって、データを採取し、得られたデータを分析することによって、母集団の μ や σ^2 を推定することになる。データが無作為抽出されたのなら、標本平均 $\bar{x}$ や標本分散 s^2 は、一定の値を取らず確率的に変動する。この意味において $\bar{x}$ や s^2 は確率変数と見なすことができる。これらの値が確率変数であるのならば、それぞれの分布を表すための期待値（平均）や分散が存在するはずである。そして母数を推定するために、標本平均の期待値 (平均)、標本平均の分散、標本分散の期待値を考えなければならない。

4　無作為抽出の重要性

統計的データ解析の目標は母集団を記述するパラメータを推定することである。

ここで重要なのは、データが母集団からランダムに抽出されていることが保証されなければならないことである。このランダムなデータ採取 (random sampling) が守られていなければ、どんなに高度で洗練された統計的手法を用いたとしても正しい推測をすることができず、結果として信頼できないデータ解析になってしまう。

ところが、無作為抽出が重要であるにもかかわらず、卒業論文レベルでは、無視されることがある。

「各データが無作為に抽出される」と同じ意味合いで、「各データが独立である」と呼ばれることもある。「独立」という用語は、一番目のデータがどの

ような値を取ろうとも、二番目のデータは同一母集団からの無作為に抽出されたものでなければならない、ということである。

例えば、サイコロをふって最初に 1 が出たとしても 4 が出たとしても、サイコロに記憶装置がない以上、次にどの目が出るかには影響しない。サイコロにおける 1 試行目と 2 試行目とは独立である。

今、隣りに座っている学生の身長は、貴方の身長とはおそらく無関係、独立である。ところが貴方の両親の身長は貴方の身長が分かれば正確ではなくとも類推することが可能である。この意味において貴方の身長と貴方の両親の身長とは独立ではない。

5 期待値と分散 (再)

確率変数 x の期待値と分散は以下のような定義であった。

$$
\begin{aligned}
E(x) &= \sum x_i\, p(x) \text{(離散型の場合)} && (4)\\
&= \int x\, f(x)\, dx \text{(連続型の場合)} && (5)\\
V(x) &= \sum x_i^2 p(x) - \left(E(x)\right)^2 \text{ (離散型の場合)} && (6)\\
&= \int x^2\, f(x)\, dx - \left(\int x\, f(x)\, dx\right)^2 \text{ (離散型の場合)} && (7)
\end{aligned}
$$

言葉で書くと、確率変数 x の値に確率密度関数をかけて、全定義区間に渡って合算 (離散型の分布の場合は合計、連続型のそれにおいては積分) したものが期待値 (平均) の定義であった。分散の場合は 2 乗の期待値 (平均) から期待値 (平均) の 2 乗を引いたものである。

さらに確率変数 x と y とを考え、a および b を任意の定数とすると

$$
\begin{aligned}
E(a) &= a && (8)\\
E(x+a) &= E(x) + a && (9)\\
E(ax) &= aE(x) && (10)\\
E(x+y) &= E(x) + E(y) && (11)\\
E(x-y) &= E(x) - E(y) && (12)\\
E(ax+by) &= aE(x) + bE(y) && (13)\\
E(xy) &= E(x)\, E(y) && (14)
\end{aligned}
$$

が成り立つ。ただし式 (14) が成り立つのは、x と y とが「独立な」場合に限

る。独立の意味合いは前述した。分散に関しては、

$$
\begin{aligned}
V(a) &= 0 && (15)\\
V(x+a) &= V(x) && (16)\\
V(ax) &= a^2 V(x) && (17)\\
V(x+y) &= V(x)+V(y) && (18)\\
V(x-y) &= V(x)+V(y) && (19)
\end{aligned}
$$

という関係がある。

演習：それぞれの式の意味を言葉で記述せよ。

式 (17) について離散型の分布を考えれば、

$$
\begin{aligned}
V(ax) &= \sum_{i=1}^{n}(ax_i)^2 p(x_i) - (E(ax))^2 && (20)\\
&= a^2 \sum_{i=1}^{n} x_i^2 p(x_i) - a^2\ (E(x))^2 && (21)\\
&= a^2 \left[\sum_{i=1}^{n} x_i^2 - (E(x))^2\right] && (22)\\
&= a^2\ V(x) && (23)
\end{aligned}
$$

だからである。式 (18) について連続型の分布を考えれば、

$$
\begin{aligned}
V(x+y) &= V(x+y) = E\left((x+y)^2\right) - \{E(x+y)\}^2 && (24)\\
&= \int\int (x+y)^2 f(x)g(y)dx\ dy - \{E(x)+E(y)\}^2 && (25)\\
&= \int\int \left(x^2+2xy+y^2\right) f(x)g(y)dx\ dy\\
&\quad - \left\{E(x)^2 + 2E(x)E(y) + E(y)^2\right\} && (26)\\
&= \int\int x^2\ f(x)\ g(y)\ dx\ dy\\
&\quad +2\int\int xy\ f(x)g(y)dx\ dy\\
&\quad +\int\int y^2\ f(x)\ g(y)dx\ dy\\
&\quad - \left\{E(x)^2 + 2E(x)E(y) + E(y)^2\right\} && (27)\\
&= \int x^2\ f(x)\ dx + 2E(x)E(y) - \int y^2\ g(y)\ dy\\
&\quad - (E(x))^2 - 2E(x)E(y) - (E(y))^2 && (28)\\
&= V(x)+V(y) && (29)
\end{aligned}
$$

となる。正負の符号に注意すれば $V(x-y)=V(x)+V(y)$ であることがわかる。

5.1 母数と統計量の関係

ある母集団から無作為に抽出されたデータ x に対して

$$E(x) = \mu \tag{30}$$
$$V(x) = \sigma^2 \tag{31}$$

が成り立つ。ここで母数と標本統計量の関係

$$E(\bar{x}) = \mu \tag{33}$$
$$V(\bar{x}) = \frac{\sigma^2}{n} \tag{34}$$
$$E(s^2) = \frac{\sigma^2}{n-1} \tag{35}$$

が重要である。言葉で表現すると標本平均は期待値は母平均と一致する。あるいは、標本平均は母平均の良い推定量であると表現される。標本平均の分散は母分散をデータ数で除したものである。

標本平均の分散とは、無限母集団から大きさ n の標本を繰り返し取り出し、その都度平均を求めるという手続きを考える。このとき標本 (データ) は無作為に取り出されるので、取り出されたデータによって定まる（計算される）標本平均の値も、その都度異なる値をとる。このように、無作為抽出したデータの標本平均を求める作業をを繰り返すと、標本平均の分散は母分散をデータ数で割ったものに一致するということである。

加えて式 (34) の意味するところは、データ数 n を増やせば増やすほど分散は小さくなり母平均の μ の近くに集まる。$n \to \infty$ の極限では標本平均の分散は 0 となり母平均と一致する。

式 (34) を利用して、母平均の信頼区間の推定が可能となる。ただし、標準偏差については、$\sqrt{n}$ に比例して小さくなることに注意。

$$V(\bar{x}) = V\left(\frac{1}{n}\sum_i^n x_i\right) \tag{36}$$
$$= V\left(\frac{1}{n}\{x_1 + x_2 + \cdots + x_n\}\right) \tag{37}$$
$$= V\left(\frac{1}{n}x_1\right) + V\left(\frac{1}{n}x_2\right) + \cdots + V\left(\frac{1}{n}x_n\right) \tag{38}$$
$$= \underbrace{\frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2}_{\text{同じものが } n \text{ 個}} \tag{39}$$
$$= \frac{1}{n^2}n\sigma^2 \tag{40}$$
$$= \frac{\sigma^2}{n} \tag{41}$$

これはパラドックスではないか？

式 (33) と式 (34) とを用いて母集団のパラメータを推定することができる。母集団分布がどのような分布であっても (ただしデータ数 n が多ければ)、標本平均は $E(\bar{x}) = \mu$ と母平均の推定量であり、標本平均の分散は $\frac{\sigma^2}{n}$ なる正規分布 $N\left(\mu, \frac{\sigma^2}{n}\right)$ に従う (中心極限定理)。これは、母集団の分布を問わないことに注意。

区間推定について簡単に触れる。母平均が含まれる区間を危険率 5% で推定することを考える。このとき、正規分布は左右対称で下側 2.5% 以上から上側 2.5% 以下までを含む領域を考えれば良い。$\mathrm{normdist}(1.96, 0, 1, 1) = 0.975$ なので区間 $-1.96\frac{\sigma}{\sqrt{n}} + \bar{x} \leq x + 1.96\frac{\sigma}{\sqrt{n}}$ の確率は 0.95 となる。これを信頼率 95% の信頼区間と言う。区間推定については、たいていの統計学の教科書に載っているが、母分散 σ^2 が既知でなければ使いものにならない。

そこで以下では母分散の推定値を考えることにする。

6 不偏分散と自由度

標本分散 s^2 は $1/n\sum_{i=1}^{n}(x_i - \bar{x})^2$ であった。ある標本分散が与えられたときの母分散 σ^2 を推定するために $E(s^2)$、すなわち標本分散の期待値を考える。$E(s^2)$ を求めるために s^2 を次のように変形する。

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 && (42)\\
ns^2 &= \sum_{i=1}^{n}(x_i - \bar{x})^2 && (43)\\
&= \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + (\bar{x})^2\right) && (44)\\
&= \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2 && (45)\\
&= \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 && (46)
\end{aligned}
$$

従って、標本分散 (の n 倍) の期待値は式 (46) の右辺第 1 項と第 2 項とを合わせたものである。右辺第 1 項は確率変数の 2 乗の期待値 $E(x_i^2)$。を合計することを意味する。一方、右辺第 2 項は標本平均 $\bar{x}$ の 2 乗の n 倍の期待値 $E(n\bar{x}^2)$ を求めることになる。

分散の定義が 2 乗の期待値（平均）から期待値の 2 乗であったことを思い出せば、一つのデータの 2 乗の期待値は $E(x_i^2)$ は

$$
\begin{aligned}
E(x_i^2) &= V(x_i) + [E(x_i)]^2 && (47)\\
&= \sigma^2 + \mu^2 && (48)
\end{aligned}
$$

である。式 (11) から確率変数の和の期待値は、それぞれの期待値の和に等しい。ゆえに、

$$
\begin{aligned}
E\left(\sum_{i=1}^{n} x_i^2\right) &= \sum_{i=1}^{n} E(x_i^2) && (49)\\
&= n\left(\sigma^2+\mu^2\right) && (50)
\end{aligned}
$$

次に標本平均の 2 乗の期待値 $E\left(\bar{x}^2\right)$ を求める。

$$
\begin{aligned}
E\left(\bar{x}^2\right) &= E\left(\frac{x_1+x_2+\cdots+x_n}{n}\right)^2 && (51)\\
&= \frac{1}{n^2}E\left[\left(x_1+x_2+\cdots+x_n\right)^2\right] && (52)\\
&= \frac{1}{n^2}E\left[\left(\sum_{i=1}^{n} x_i^2+\sum_{i=1}^{n}\sum_{j\neq i}^{n} x_i x_j\right)\right] && (53)\\
&= \frac{1}{n^2}\left[E\left(x_1^2\right)+E\left(x_2^2\right)+\cdots+E\left(x_n^2\right)\right]\\
&\quad +\frac{1}{n^2}\left[\sum\sum E\left(x_i\right)E\left(x_j\right)\right] && (54)\\
&= \frac{n\left(\sigma^2+\mu^2\right)}{n^2}+\frac{1}{n^2}\left(n(n-1)\right)\mu^2 && (55)\\
&= \frac{\sigma^2+\mu^2}{n}+\frac{(n-1)}{n}\mu^2 && (56)\\
&= \frac{1}{n}\left(\sigma^2+\mu^2+n\mu^2-\mu^2\right) && (57)\\
&= \frac{\sigma^2}{n}+\mu^2 && (58)
\end{aligned}
$$

これらより標本分散の期待値は、

$$
\begin{aligned}
E(s^2) &= E\left\{\frac{1}{n}\left(\sum_{i=0}^{n} x_i^2-n\bar{x}^2\right)\right\} && (59)\\
&= \frac{1}{n}E\left(\sum_{i=1}^{n} x_i^2\right)-\frac{n}{n}E\left(\bar{x}^2\right) && (60)\\
&= \frac{1}{n}n\left(\sigma^2+\mu^2\right)-\frac{\sigma^2}{n}-\mu^2 && (61)\\
&= \sigma^2+\mu^2-\frac{\sigma^2}{n}-\mu^2 && (62)\\
&= \sigma^2-\frac{\sigma^2}{n} && (63)\\
&= \frac{n-1}{n}\sigma^2 && (64)
\end{aligned}
$$

となる。s^2 の期待値が $\frac{n-1}{n}\sigma^2$ であるから、

$$s^2 = \frac{n-1}{n}\sigma^2 \tag{65}$$

$$ns^2 = (n-1)\sigma^2 \tag{66}$$

$$\frac{ns^2}{n-1} = \sigma^2 \tag{67}$$

である。この関係が標本分散と母分散との関係である。ns^2 は偏差平方和と呼び $\sum (x_i - \bar{x})^2$ であった。偏差平方和をデータ数 n で割って標本分散を求めるかわりに、$n-1$ で割ると母分散の不偏推定量になる。

データ数 n が大きくなれば大きくなるほど、偏差平方和を n で割っても $n-1$ で割っても、さほど差が無くなるほとんど一致してしまう。$n \to \infty$ の極限では標本分散は母分散に一致する。

偏差平方和を $n-1$ で割った量を不偏推定量、不偏分散と呼ぶ。不偏 unbiased とは聞き慣れない言葉であるが、統計量の持っている性質を表している。不偏性以外にも一致性、有効性、十分性、最尤性、最良線型性などの概念があるが省略する。

偏差平方和をデータ数で割らずに、(データ数 -1) で割ることを、偏差平方和を自由度で割ると呼ぶ。あるいは不偏分散を求めるために偏差平方和を自由度で割る、という。

自由度 degree of freedom とは統計学の専門用語であって、理解しにくい概念である。自由度の直感的な意味合いは、以下のようになる。標本平均 $\bar{x}$ は、無作為に抽出された n 個のデータに基づいて計算される確率変数である。ところが、無作為に n 個のデータを選んだという情報と、その情報のもとで計算された確率変数である標本平均 $\bar{x}$ との間には、一定の関係が生じてしまう。換言すれば、標本平均 $\bar{x}$ とは n 個のデータに基づくものであるが、標本平均を求めてしまった後では $n-1$ 個のデータしか自由ではない。標本平均と $n-1$ 個のデータを知ってしまえば、最後の n 番目のデータは一意に定まる。$n-1$ 個のデータしか自由にならない、独立ではない、という意味である。

これは、確率論からみると都合が悪い。なぜなら n の互いに独立な確率変数を考えねばならぬときに、1 個のデータだけは自由ではない、独立ではない、ので無作為抽出とは言えなくなるからである。