統計分析シリーズ(Ⅳ)

茨城大学教授　所　一夫

Ⅳ　小標本の場合の母平均の推定

1.　正規型母集団からの標本

前号では標本の大きさが大で中心極限定理が用いられる場合，すなわち標本調査といわれる技法についてその考え方と方法を述べたが，実は推計学の大きな特徴は標本の大きさが小さい場合（小標本）であっても，調査した数値を通して「どの範囲までの事がどの程度の確率で確認できるか。」という問題が考えられることである。

その例としてここでは母集団が正規分布N（M, S^2）と見られる場合（これを正規型母集団という。）について，調査して得られた標本値から，指定された信頼係数のもとに母平均の信頼区間がどのように求められるかを考えて見よう。

正規型母集団といえば，何か特別の場合にしか適用できない理論のようにも見えるが，分布を示す図表の山が一つで，そこから左右が対称となるような分布では正規型と見てよい場合が多く，またもとの分布では正規分布と見られない場合でも，適当に変数を変換すれば変換した変数に関しては正規分布と見られる場合が多い。したがって以下の理論は実際問題では広く用いられるものである。

またこの場合には必ずしも乱数表によりランダムサンプルを抽出するということでなく，その母集団から無作為に抽出されたと考えられる標本の調査値を標本値(x_1, x_2, …, x_n）として考える場合が多い。

このような場合について前と同様に標本平均$\bar{x}$を作ると，この$\bar{x}$は種々の値をとるが，この$\bar{x}$の分布は標本調査の場合と同様に

(a)　$\bar{x}$の平均は母平均Mに等しく，

(b)　$\bar{x}$の標準偏差s($\bar{x}$)はもとの母集団の標準偏差Sに対して，s($\bar{x}$)＝S／$\sqrt{n}$となる。これは標本調査の場合で母集団の大きさNが無限大になった場合である。

(c)　また$\bar{x}$の分布は正規分布になっていることなどが数理統計学により明らかにされている。

これらの結果をもとにして標本調査の場合と同様に母平均Mの区間推定が考えられる。

2.　母集団標準偏差Sが既知の場合

この場合は標本調査の場合と比べて，標本標準偏差がs($\bar{x}$)＝S／$\sqrt{n}$となっただけのちがいで他は全く同様である。すなわち　$\bar{x}$＝（$x_1+x_2+\cdots+x_n$）／n　とすれば，信頼係数 95% ではMの信頼区間は

$$\left(\bar{x}-2\frac{S}{\sqrt{n}},\quad \bar{x}+2\frac{S}{\sqrt{n}}\right)$$

であり，信頼係数99.7%では次のようになる。

$$\left(\bar{x}-3\frac{S}{\sqrt{n}},\quad \bar{x}+3\frac{S}{\sqrt{n}}\right)$$

例　ある年度の入学試験の採点について某高校から発表された結果によると，某中学からの志願者は9名あり，それらの受験生の得点は

290, 280, 275, 260, 255, 243, 237, 220, 217,

で，全志願者の得点の標準偏差Sは 22.5 であった。この結果から，その中学の生徒たちの（母集団における）平均は何点ぐらいと見られるか。信頼係数 95％ でその信頼区間を求めよ。

解　問題の意味はこの9人によって代表される母集団の母平均を区間推定することであるが，この母集団を正規型母集団と考えて問題を解こうとするものである。試験の点数の分布は一般には正規分布とは限らない。しかし高校入試の点数は例年正規型に近いと見られるので，ここに例題として取りあげたものである。

示された結果によれば

$$\bar{x}=(290+280+\cdots+217)/9=253$$

S＝22.5より　s($\bar{x}$)＝22.5／$\sqrt{9}$＝7.5

であるから，信頼係数95%では信頼区間は

(253－2×7.5, 253＋2×7.5). すなわち (238, 268)

とわかる。

すなわち「もし判断が誤ることがあったとしてもその確率は5％以下である。」という条件のもとでは母平均は238点と268点の間と考えてよいという結論が出たわけである。

上例では標本の大きさが9であるからnが小さく，標本調査の考えで処理することはできない。しかし「入試の点数の分布が正規型と見られる。」という従来からの知識を判断の資料として加えることによって，この事実に対する確率判断がなされる事を示したものである。

この例のほかにパックされた食品の実際の目方とか，工場から大量生産方式によって製造された製品の計量などについて，その計量に対する保証を与える方法として上の区間推定法が用いられる事が多い。

3．母集団の標準偏差が未知の場合

前例の入試の点数の場合で，高校側から母集団標準偏差Sの値が知らされていない場合にはどのようにしてこの中学の生徒の母平均を区間推定したらよいか，ということがここの問題である。

問題を整理して見ると（x_1, x_2, …, x_n）を$N(M, S^2)$なる正規型母集団（M, Sは未知）からの無作為標本とするとき，（$\bar{x}-d<M<\bar{x}+d$）となる確率を与えられた信頼係数に等しくするようなdを求めることである。

この場合には標本平均を$\bar{x}$，標本から求めた標準偏差（これを標本標準偏差という）をs'，すなわち

$$\bar{x}=(x_1+x_2+\cdots, +x_n/n$$

$$s^2=\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots(x_n-\bar{x})^2}{n}$$

$$=\frac{x_1^2+x_2^2+\cdots+x_n^2}{n}-\bar{x}^2$$

として　$t=\frac{(\bar{x}-M)\sqrt{n-1}}{s}$

なる式を作ると，このtの値は抽出された標本の値によって種々の値をとるが，その分布はt＝0を対称の軸とした（正規分布と似た）分布をしている。

この分布を自由度n－1のt分布といっている。そしてこの分布に対して自由度(n－1)の種々の値に対して，

（$|t|>t_0$）となる確率が0.05とか0.01などになるようなt_0の値が表示されている。この表をt分布の表という。（この表も統計数値表または多くの数理統計学の本に付いている。）

次に確率が0.05の場合のt_0の値を示すt分布表の一部を示す。

確率0.05のt分布の表

自由度	t_0の値	自由度	t_0の値	自由度	t_0の値
5	2.571	14	2.145	23	2.069
6	2.447	15	2.131	24	2.064
7	2.365	16	2.120	25	2.060
8	2.306	17	2.110	26	2.053
9	2.262	18	2.101	27	2.052
10	2.228	19	2.093	28	2.048
11	2.201	20	2.086	29	2.045
12	2.179	21	2.080	30	2.042
13	2.160	22	2.074	∞	1.960

この表で自由度n－1，（$|t|>t_0$）となる確率が0.05となるようなt_0を示した表の値を$t_{n-1,0.05}$のように書くが，この表記法によると表より$t_{5,0.05}=2.571$，$t_{10,0.05}=2.228$である。そして自由度が5のt分布では（$|t|>t_{5,0.05}=2.571$）となる確率が0.05となっているのである。これらの関係より

$$t=(|\bar{x}|-M\sqrt{n-1}/s)>t_{n-1,0.05}$$

となる確率は0.05であるから

$$\frac{|\bar{x}-M|\sqrt{n-1}}{s}\leqq t_{n-1,\ 0.05}$$

となる確率は0,95である。したがって

（$|\bar{x}-M|\leqq t_{n-1,\ 0.05}\ s/\sqrt{n-1}$）となる確率が95%である，ゆえに（$\bar{x}-t_{n-1,0.05}\ s/\sqrt{n-1}$）$\leqq M\leqq$（$\bar{x}+t_{n-1,0.05}\ s/\sqrt{n-1}$）

なる確率が95%であるから求めようとする信頼区間は

$$\left(\bar{x}-t_{n-1,0.05}\frac{s}{\sqrt{n-1}},\ \bar{x}+t_{n-1,0.05}\frac{s}{\sqrt{n-1}}\right)$$

である。

例　前例の9人の点数から（S未知），この9人が抽出されたと考えられる母集団の母平均を信頼係数95%で区間推定せよ。

解 $\bar{x}=\frac{290+280+275+260+255+243+237+220+217}{9}$

$=253$

$$\frac{s^2}{n-1}=\frac{(290-253)^2+(280-253)^2+\cdots+(217-253)^2}{9\ (9-1)}$$

$$=\frac{5376}{72}$$

$\therefore\ s/\sqrt{n-1}=\sqrt{5376/72}\fallingdotseq 8.64$

またt分布の表より　$t_{8,0.05}=2.306$

$\therefore\ t_{8,0.05}\ s/\sqrt{n-1}=2.306\times 8.64\fallingdotseq 20$

ゆえに上の計算より求める信頼区間は

（253－20, 253＋20）すなわち大略（233, 273）

である。

この結果を前例の結果（253－15, 253＋15）と比べてみると信頼区間の幅がだいぶ大きくなっているが，これは前の場合には，後の場合のほかに S＝22.5 という情報がはいっていたために，より精密な判断がなされたわけである。

以上小標本でも判断の下せる母平均の推定法を紹介したが，この場合についても信頼係数と信頼区間の幅を与えてこれを満足させる標本の大きさnを求める問題がある。しかしこの問題は前号の方法を参考にして適当に解決されるものと思うので今回は詳述しない。