統計分析シリーズ（Ⅲ）

茨城大学教授　所　一夫

Ⅲ　平均と比率の推定（大標本の場合）

1. 正規分布

前号で標本平均の分布は，標本の大きさnが大であれば中心極限定理によりほとんど正規分布と見てよい事を示したが，本号ではこの大標本の場合を考えて，正規分布について概説しよう。

正規分布というのはそれに属する単位数は無限であり（無限母集団），数学的に理想化されたモデルで，現実の問題ではこれに近いと思われる母集団が考えられるものである。そして正規分布に近いと見られる母集団ではその母平均をM，標準偏差をS，各単位の持つ数値をxとすれば，$x=M$のとき　$x\pm d$（dは小さい正数と考えてよい）の範囲内にはいる母集団の単位の数が最大で，xがMを離れるにしたがってこの数は小さく，全体の分布は$x=M$を中心として左右対称になっている°

くわしくは平均をM，標準偏差をSとした正規分布では（これを$N(M, S^2)$と書く）xがMから，M+aSの範囲内にある単位数（M−aSの範囲内にある単位数も同じ）の全体に対する割合を$f(a)$とするとき，aの種々の値に対する$f(a)$の値が表示されている。この一部分を次に示す。

a	0	0.5	1.0	1.5	2.0	2.5
f(a)	0.000	0.1915	0.3413	0.4332	0.4772	0.4938

3.0	3.5	'00
0.4987	0.4998	0.5

この表によれば正規分布ではM−2SとM+2Sの間には全体の　0.4772×2S＝0.9544すなわち約95%がはいって居り，M−3SとM+3Sの間には全体の0.4987×2=0.9974すなわち約99.7%がはいっていることがわかる。なおくわしい正規分布表があれば，一般にbとcの間に全体の何%がはいって居るかは，上表のaに対応する値が（c−M）／Sなどであることから

$$f\left(\frac{c-M}{S}\right)-f\left(\frac{b-M}{S}\right)$$

として上表から算出できる。（ただし$f(-a=-f(a)$とする。）

2.母平均の推定

（i）　母平均の区間推定

大きさNの母集団から大きさnのランダムサンプルを抽出しその標本平均$\bar{x}$を算出した場合に　(a)　$\bar{x}$は母平均Mにはならない，(b)　$\bar{x}$の平均はMである，(c)　$\bar{x}$の標準偏差$s(\bar{x})$は$\sqrt{(N-n)/(N-1)}\,S/\sqrt{n}$，(d) nが大きい場合には$\bar{x}$の分布は正規分布に近い事を前回述べた。

そうすると母平均M，母集団の標準偏差がSなる母集団から大きさnのランダムサンプルを抽出した場合に，$\bar{x}$は種々の値をとるが，正規分布の性質から，これらの値が　$M\pm 2s(\bar{x})$　の範囲内にはいっている場合は約95%で，$M\pm 3s(\bar{x})$の範囲内にはいっている場合は約99·7%と考えられる。このことは$\bar{x}$が$M\pm 2s(\bar{x})$の範囲にはいる確率が95%，$M\pm 3s(\bar{x})$の範囲内にはいる確率が99.7%　であることを示している。

すなわち$\bar{x}$はMに等しくはないが確率　95%　で$M\pm 2s(\bar{x})$，確率　99.7%　で$M\pm 3s(\bar{x})$の範囲内にあると判断ができる事を示して居り，この事実を他の面から見ると，知ろうとする数値Mは$\bar{x}$に等しいとはいえないが，Mが　$\bar{x}\pm 2s(\bar{x})$の範囲内にある確率が95%で，$\bar{x}\pm 3s(\bar{x})$の範囲内にある確率が99.7%と判断できることを示している。

この場合の範囲をその区間推定に対する信頼区間と言いその確率を信頼係数といっている。

以上より標本から算出した平均を$\bar{x}$，要求された信頼係数を95%，母集団の大きさをN，その標準偏差をSとするときMの信頼区間は

$$\left(\bar{x}-2\sqrt{\frac{N-n}{N-1}}\frac{S}{\sqrt{n}},\ \bar{x}+2\sqrt{\frac{N-n}{N-J}}\frac{S}{\sqrt{n}}\right)$$

であり信頼係数を 99.7% まで要求された場合には上の−2，+2の代りに−3，+3とすればよい。

上の話ではまだ未解決の問題が残っている。それは母集団の標準偏差Sが未知の場合が多いからである° この解決策として実際に用いられている方法は

(a)　過去の資料，実験，または同様な資料について調査した結果から得られた標準偏差S′をもってSの代用とする。

(b)　nが大である場合を考えているので，この場合には抽出した標本のn個の値から標本標準偏差sを算出してこれを母集団標準偏差Sの代用とすることである。

例　世帯数5,000の町で，大きさ100のランダムサンプルを抽出して家計調査を行なった結果，1ケ月の平均支出は95,000円，標準偏差は10,000円であった。この結果より全町での平均支出はいくらと推定されるか，信頼係数95%でその信頼区間を求めよ。

解　$s(\bar{x})\fallingdotseq\sqrt{\frac{N-n}{N-1}}\frac{s}{\sqrt{n}}=\sqrt{\frac{5,000-100}{5,000-1}}\frac{10,000}{\sqrt{100}}=990$

したがって信頼係数　95%　の信頼区間は

(95,000−2×990　95,000+2×990)

すなわち（93,020,　96,980）である。

実際問題ではnを大にすればs($\bar{x}$)は小さくなり，上

例では次の(ii)で示すよでに，n=400 とすればs=10,000ならば$2s(\bar{x})$=990で信頼区間は95,000円±1,000円以内となり，これを表示する場合に1,000円単位で±990円を略して95,00円とする場合が多い。

(ii) 標本の大きさnの決定

上例を場合，$\bar{x}\pm 2s(\bar{x})$ によつて示された$2s(\bar{x})$ の値を定められた値以下にするためにはnをどの程度にすればよいかを調べることが調査者にとって問題である。

これは結局$s(\bar{x})=\sqrt{(N-n)/(N-1)}\ S/\sqrt{n}$を与えられたb以下におさえる問題で

$$\sqrt{\frac{N-n}{N-1}}\frac{S}{\sqrt{n}}<\frac{S}{\sqrt{n}}\leqq d \quad より \quad \frac{S}{d}\leqq\sqrt{n},\quad \frac{S^2}{d^2}\leqq n$$

とすればよい。ここにdは与えられた値でおり，Sは未知であるが，前に行なった調査または同様な他の調査より推定するか，または予備調査を行なって少し過大に見積っておいて，このbとS（またはその代用値）より上式によってnを求めればよい。

上例で$\pm 2s(\bar{x})$ を±1,000とするようにnを求めて見ると

$$s=10{,}000,\quad 2d=1{,}000 \quad より \quad d=500$$
$$s^2/d^2=(10{,}000)^2/(500)^2=400\leqq n$$

すなわち S=10,000 と考えられるならば n=400 でよいことがわかる。

3. 比率の推定

(i) 母集団比率の信頼区間

前号で示したように，母集団比率はその属性を持った単位は1とし他の単位を0とした場合の平均Mである。したがってN個の中でその属性を持ったものをR個とすると比率Pは P=M=R/N であり，この場合の標準偏差Sは

$$S=\sqrt{\frac{(1-P)^2R+(0-P)^2(N-R)}{N}}=$$
$$\sqrt{(1-P)^2P+P^2(1-P)}=\sqrt{P(1-P)}$$ である。

いま大きさnのランダムサンプルについての標本比率をpとすればp=$\bar{x}$であり標本比率pについての標準偏差S(p)はS（$\bar{x}$）であるから

$$s(p)=s(\bar{x})=\sqrt{\frac{N-n}{N-1}}\frac{S}{\sqrt{N}}=\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{P(1-P)}{n}}$$

したがって標本より求めた比率がpであるとき母集団における比率は信頼係数を 95% とすれば

$$\left(p-2\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{P(1-P)}{n}},\ p+2\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{P(1-P)}{n}}\right)$$

である。しかしここでもPは未知であるので，nが大である場合には標本から求めたpによってPり代用をさせることが多い。すなわち上の信頼区間を

$$\left(p-2\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}},\ p+2\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}\right)$$

として実用に供するのである。

またPはどのような値であっても常に $\sqrt{P(1-P)}\leqq 0.5$, $\sqrt{(N-n/(N-1)}<1$ であることより少し範囲を広くとって上の信頼区間を

$$\left(p-2\frac{0.5}{\sqrt{n}},\ p+2\frac{0.5}{\sqrt{n}}\right)$$すなわち$$\left(p-\frac{1}{\sqrt{n}},\ p+\frac{1}{\sqrt{n}}\right)$$

とすれば真のPがこの区間内に存在する確率は 95% 以上となる。（信頼係数は95%より大とるな。）

例 農林省統計情報部がまとめた「46年度農家の形態別にみた農家経済」では全国約517万戸の農家より11,100戸を抽出して諸調査をした結果，稲作経営の場合の所得の農業使存度は28%と発表された。この結果より信頼係数95%で全国稲作農家についての農業依存度の信頼区間を求めて見よう。

解 Nは大きいので$\sqrt{(N-n)/(N-1)}\fallingdotseq 1$と見る，n=11,100 p=0.28 より

$$2\sqrt{\frac{p(1-p)}{n}}=2\sqrt{\frac{0.28\times 0.72}{11{,}100}}\fallingdotseq 2\times 0.00426\fallingdotseq 0.0085<0.01$$

したがって信頼係数 95% の信頼区間は（0.28−0.0085, 0,28+0.0085）すなわち（0.272, 0.289）となる。この結果より依存度を28%と発表された意味もわかる。

(ii) 標本の大きさ，の決定

上述より信頼係数95%で考える場合には$\pm 1/\sqrt{n}$を与えられたdより小さくしたい場合にnをどの程度にすればよいかの計算が直ちにできる。すなわち $1/\sqrt{n}\leqq d$ として$1/n\leqq d^2$, $1/d^2\leqq n$ とすればよい。すなわち±の部分を 1% 以下におさえようとすれば上式から $n\geqq 10{,}000$とすれば十分であることがわかる。

統計ニュース

—— 9 月 の 行 事 ——

9月 3日～ 5日　地方統計職員業務研修（専門研修）
4日～ 6日　刊行物ブロック会議
13日～14日　労働力特調新設集団住宅ブロック会議
14日　統計グラフコンクール
19日～20日　漁業センサス本調査ブロック会議
30日　工作機械調査日