2000/4/17
第　1　講

比較統計システム論　講義録（2000 年度）

統計学の基本的考え方

母集団(population)の性格を知るために十分なサンプルを確保し、そこから得られた統計的性格が母集団のそれと一致しているということを推量する。

①母集団全体を調査する場合
全数調査、悉皆調査、Census、国勢調査が良い例である。

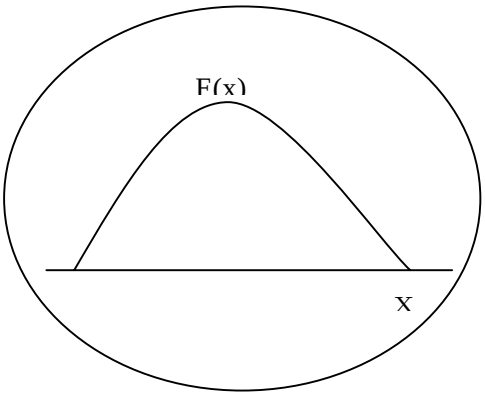
②母集団から一部を選び出して、母集団の性質を推測する場合
統計的推測（statistical inference）を用いる。

母集団から分析のため選び出された要素またはその属性値を標本（sample）、選び出す作業を標本抽出（sampling）または抽出と呼ぶ。

（注）　われわれが直接調査するのは標本であるが、知ろうとするのは母集団の分布（統計的性質）についてである。
統計的手法を用いた研究では、研究の対象としている母集団は何で、用いているサンプルはその全体のどのぐらいのウェイトを占め、また母集団をバイアスなく代表しているかどうかを検討する必要がある（少なくとも標本抽出の手法についてはチェックしなければならない）。

（例）　わが国の家計の貯蓄行動はライフサイクル仮説に基づいているという命題を検討する場合の母集団はわが国の家計全体であり、実証する場合には全数調査はないのでサンプルを抽出する必要がある。しかしそのサンプルは母集団の分布を反映したものでなければならない。

母集団と母集団分布



母集団（無限母集団）

標本抽出
（サンプリング）

標本（n＝10）

x_1、x_2、x_3、x_4、x_5

x_6、x_7、x_8、x_9、x_{10}

random sampling
この母集団分布に従う確率変数

標本x_iは同一の母集団分布 f(x) に従うn個の独立名確率変数である。
nを標本の大きさ（sample size）という。

①母集団分布の確率分布が事前に分かっている場合（パラメトリック）

交通事故→ポアソン分布

身長、IQ→正規分布

母集団分布を決定する定数、パラメータを母数（parameter）と呼ぶ。

②母集団分布の具体的な形が事前に分からない場合（ノン・パラメトリック）

この場合、母集団分布に関して様々な統計量を用いて分析する必要がある。

母平均　中位値（median）、最頻値（mode）、

母分布　歪度（skewness）、尖度（kurtosis）

標本の抽出

母集団に属する要素すべてについて観測値を得ることが出来ない場合は、この中から標本を抽出して母集団分布について推定する。

抽出した要素を再び母集団に戻し、その後の抽出の対象とする場合：復元抽出
（sampling with replacement）

戻さない場合：非復元抽出（sampling without replacement）

通常は非復元抽出が行われており、Nがnに比べて十分大きい場合には、いずれの方法によってもほとんど差がない。

非復元抽出によってN個の母集団からn個の標本の可能な選び方の総数は組み合わせ数

$$ {}_N C_\rho = \frac{N!}{n!(N-n)!} $$

で与えられる。

実際に利用できる標本は、この組み合わせの中からたまたま無作為に選ばれた一つであると考えられる（単純ランダム・サンプリング＝単純無作為抽出）

単純ランダム・サンプリングでは、母集団の各要素が標本に含まれる確率（抽出率）を等しく n／N とするものである。

（注） 現実の統計調査では、国勢調査などの全数調査をもとに地域毎の人口分布に応じて各地域から抽出すべき人口（家計）をまず決定し、次にその地域の中で無作為に人口（家計）を抽出するという二段抽出法が取られることが多い。

（注） 抽出法（sampling theory）については、W. G. Cochran, (1977)Sampling Thechniques, 3rd ed. (Weily)が古典であるが、最も実務的な入門書に清水誠(2000)『推計統計初めの一歩』（講談社ブルーバックス）がある。

比率の推定

母集団において、ある事象の起こる確率を母比率と呼び、これが ϕ ならば、n回の観察におけるこの事象の出現回数xは2項分布に従う。母集団は、観察の仕方によって容易に2つに分けることができ、男性と女性、良品と不良品、就業者と失業者などに分けて、それぞれ男性の割合、良品の割合、就業者の割合などを ϕ とみなせばよい。

ϕ を変化させて実際のデータが起こる確率が最も高くなるように ϕ の値を定める方法を最尤法（maximum likelihood method）という。

標本において、ある事象の起こる確率（標本確率）をpとすると、母集団が2項分布に従うなら $E(x)=n\phi$, $v(x)=n\phi(1-\phi)$ であることから、

$$E(p) = \phi, \qquad V(p) = \frac{1}{n}\frac{N-n}{N-1}\phi(1-\phi),$$

$$\sigma(p) = \sqrt{\frac{1}{n}\frac{N-n}{N-1}\phi(1-\phi)}$$

母集団が大きければ、

$$V(p) = \frac{\phi(1-\phi)}{n}, \qquad \sigma(p) = \sqrt{\frac{\phi(1-\phi)}{n}}$$

となる。
標本比率の標準誤差は標本数の平方根に反比例する。

母数と統計量

母集団分布の代表的な母数は母平均 population mean（μ）と母分散 population variance（σ^2）である。しかし全体を調べることは難しいので、大きさnの標本 x_1、…x_nを抽出し、その標本平均（sample mean）は

$$\bar{x} = (x_1 + x_2 + \ldots x_n)$$

標本分数（sample variance）は

$$S^2 = \frac{1}{n-1}\left\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots (x_n - \bar{x})^2\right\}$$

で与えられる。

ここで注意すべきは、標本分数はn－1で割ったものであることである（n－1は不偏分散の自由度）。

標本分散 S^2は、期待値が $E(s^2) = \sigma^2$で母分数に一致し、母分数を過大あるいは過小に推定することがない。この S^2を母分数の不偏推定量、不偏分散という。
S^2のかわりに、

$$Z^2 = \frac{1}{n}\left\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots (x_n - \bar{x})^2\right\}$$

を定義すると、Z^2も標本分散であるが、偏りのある標本分散であることが知られている。すなわち、

$$E(Z^2) = \frac{n-1}{n}\sigma^2$$

となり、n＝10 ならば 10%程度のσ^2の過小評価が起こるので、S^2と Z^2の違いに注意すべきである。

＜不偏分散の不偏性＞　確率変数 x_1, x_2, … x_nは独立、同一の母集団分布に従うとする。
$E(x_i)=\mu$、$V(x_i)=\sigma^2$ (i＝1，2，…n)とする。自由度n－1で割って定義した分散

$$S^2=\frac{\sum_{i=1}^{n}(x^i-\bar{x})^2}{n-1}\quad,\quad \bar{x}=\frac{\sum_{i=1}^{n}x^i}{n}$$

に対して、$E(s^2)=\sigma^2$となる場合に、標本分数は不偏であるという。
$y_i=x_i-\mu$とする。

$$E(y_i)=0\qquad E(\bar{y})=0$$

となる。そこで、

$$\sum(x_i-\bar{x})^2=\sum(y_i-\bar{y})^2=\sum y_i^{\,2}-n\bar{y}^2$$

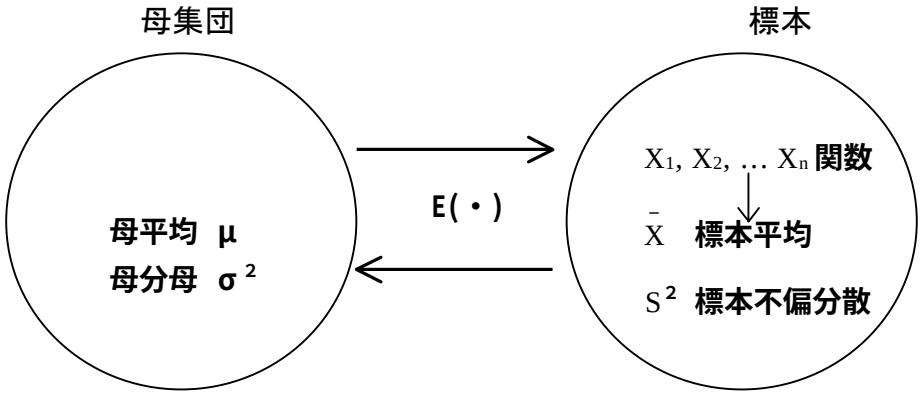
に対して

$$E(y_i^{\,2})=V(y_i)=V(x_i)=\sigma^2\qquad E(\bar{y}^2)=V(\bar{y})=\frac{V(y_i)}{n}=\frac{\sigma^2}{n}$$

である。ここで

$$E(s^2)=-\frac{1}{n-1}E\left(\sum\left(x_i-\bar{x}\right)^2\right)=\frac{1}{n-1}\left\{E(\sum y_i^{\,2})-nE(\bar{y}^2)\right\}$$

$$=\frac{1}{n-1}\left\{n\sigma^2-\frac{\sigma^2}{n}\bullet n\right\}=\sigma^2$$



統計量の標本分布

標本分布が母集団分布に依存するパラメトリックな場合で、かつその分布が再生性を持っていれば、母集団分布は求められる。

再生性とは、独立な二つ以上の確率変数が同一の分布族に属する場合、その和もそれに属することで、二項分布、ポアソン分布、正規分布等がこの性質を満たす。

有限母集団と有限母集団修正

いままでは、一般に無限母集団を仮定してきたが、この課程は母集団の大きさNがあまり大きくない場合や、n／Nが大きい場合には適当ではない。このような場合はNが有限であることを考慮した修正を行う必要がある。

有限母集団では

(a) $\bar{x}$の期待値 $E(\bar{x})$は母平均μと等しい。

(b) $\bar{x}$の分散$v(\bar{x})$は母分散をσ^2とすると

$$v(\bar{x}) = \frac{N-n}{N-1} \bullet \frac{\sigma^2}{n}$$

ここで

$$C_n = \frac{N-n}{N-1}$$

は無限母集団の分散を修正する係数であり、有限母集団修正と呼ばれる。有限母集団修正 C_Nは無限母集団の極限 $N \to \infty$で $C_N \to 1$となり消える。

総量の推定

母集団の総量を標本から推定するにはどうすればよいだろうか。
一般に総量Tは平均値の推定値（μ）に母集団の個数（N）を掛けた数になる。

$$T = \mu N$$

標本の総量は同様にして

$$\overline{T} = n\overline{x}$$

となるので、標本総量の期待値は

$$E\left(\overline{T}\right) = nE\left(\overline{x}\right) = n\mu$$

となる。
標本総量の分散は

$$V\left(\overline{T}\right) = V(n\overline{x}) = n^2 V(\overline{x}) = n^2\left(\frac{1}{n}\frac{N-n}{N-1}\sigma^2\right) = n\frac{N-n}{N-1}\sigma^2$$

標本総量の標準誤差は

$$\sigma(\overline{T}) = n\sigma(\overline{x}) = \sqrt{n\frac{N-n}{N-1}}\sigma$$

つまり、n人の分散や標準誤差は各個人の分散や標準誤差をn倍したものより有限母集団修正の分だけ小さくなる。母集団が大きければ、

$$V\left(\overline{T}\right) = n\sigma^2,$$

$$\sigma\left(\overline{T}\right) = \sqrt{n}\sigma = n\frac{\sigma}{\sqrt{n}}$$

となり、平均値のn倍となる。