5

推定

５．１ 統計的推定の考え方

統計的推定

⇒ 母集団のパラメータを母集団からの大きさ n の標本 $X_1, X_2, \cdots, X_n$ を用いて理論的に推定すること.

統計的推論にはいくつかの種類があるが、今回は点推定と区間推定の２つについて述べる。

- 点推定 …… 推定を１点で行う
- 区間推定 $[a, b]$ … 確率 α を事前に決定し，母集団のパラメータの値がその区間 $[a, b]$ に入る確率が α になるように推定する

それぞれの特徴

点推定は統計量をつかって１点で推定を行う．そのため標本の関数として簡単に計算できるが，推定された値が持つ誤差等の精度がわからない．区間推定は事前に確率の計算が必要で，多少計算が面倒であるが，推定の精度が分かりやすい．

５．２ 点推定

点推定の推定量として持っていることが望ましい性質

１．不偏性（不偏推定量）

２．一致性(一致推定量)

３．最小分散性（有効推定量）

４．十分性（十分推定量）

の４つがある。

５．２．１ 不偏推定量

不偏推定量とは …

推定量の期待値がパラメータと一致するときその推定量を不偏推定量と呼ぶ.

式で表すと母集団のパラメータθについての推定量$g(X_1, X_2, \cdots, X_n)$に対して

$$E[g(X_1, X_2, \cdots, X_n)] = \theta$$

が成立するとき$g(X_1, X_2, \cdots, X_n)$はθの不偏推定量である.

例）母平均μの不偏推定量は標本平均$\bar{X}$である

母平均μを持つ母集団に対して，標本平均$\bar{X}$の期待値は

$$
\begin{aligned}
E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu
\end{aligned}
$$

であり，標本平均$\bar{X}$の期待値が母平均μと一致していることがわかる．

5．2．2　一致推定量

n 個の標本 $X_1, X_2, \cdots, X_n$ から母集団のパラメータ θ を推定する推定量 $g(X_1, X_2, \cdots, X_n)$ が，$n \to \infty$ のときに

$$P(|g(X_1, X_2, \cdots, X_n) - \theta| > \varepsilon) \to 0$$

となる推定量のことを一致推定量とよぶ.

例）正規分布の標本平均$\bar{X}$は一致推定量である。

まず$\bar{X}$の分散は$\frac{\sigma^2}{n}$よりチェビシェフの不等式を使うと

$$P(|\bar{X}-\theta|>\varepsilon)<\frac{\sigma^2}{n\varepsilon^2}$$

が成り立つ。

ここで$n\to\infty$のとき与式は0になるので

$$P(|\bar{X}-\theta|>\varepsilon)\to 0\ (n\to\infty)$$

が成り立ち、一致推定量であることがわかる。

5．2．3　尤度と最尤推定

n個の標本$X_1, X_2, \cdots, X_n$の実現値（標本値）$x_1, x_2, \cdots, x_n$が得られた場合の尤もらしさを表す値として，尤度関数Lを次のように定義する．

$$L(\theta|x_1, \cdots, x_n)$$
$$= f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta)$$

例）指数分布$f(x|\lambda) = \lambda e^{-\lambda x}$のとき

$$L(\theta|x_1, \cdots, x_n)$$
$$= \lambda e^{-\lambda x_1} \times \lambda e^{-\lambda x_2} \times \cdots \times \lambda e^{-\lambda x_n}$$
$$= \lambda^n e^{-\lambda(x_1 + x_2 + \cdots + x_i)}$$

また，離散型確率変数の場合は確率密度関数 $f(x|\theta)$ のかわりに確率 $P(X=x|\theta)$ を用いて次のように定義する．

$$L(\theta|x_1,\cdots,x_n) = P(X=x_1|\theta)\times P(X=x_2|\theta)\times\cdots\times P(X=x_n|\theta)$$

例) 幾何分布 $P(X=k|p)=p(1-p)^k$ のとき

$$\begin{aligned} &L(\theta|x_1,\cdots,x_n) \\ &= p(1-p)^{k_1}\times p(1-p)^{k_2}\times p(1-p)^{k_n} \\ &= p^n(1-p)^{k_1+k_2+\cdots+k_n} \end{aligned}$$

いずれにしても標本x_iが現れた出現確率の積とみなすことができる（厳密にいえば連続型の場合，確率密度の積である）.

よって，この尤度関数の値が大きいほど分布の当てはまりが良いことを示している．そこでn個の標本が得られた場合の尤度関数の値を最大にするような，パラメータθの推定量を最尤推定量と呼ぶ.

最尤推定量の求め方

$L(\theta)$ が θ に関して微分可能ならば、その導関数を0に等しいとおいた式から極値を求め、その中で $L(\theta)$ を最大にするものが θ の最尤推定量である。実際に $L(\theta)$ を最大にするには $L(\theta)$ の対数をとった対数尤度関数

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i : \theta)$$

を最大にすればよいので、方程式

$$\frac{\partial}{\partial \theta} \log L(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i : \theta) = 0$$

を解いて得られる解の中から $L(\theta)$ を最大にするものを求めればよい.

例）正規分布を仮定した場合，標本平均$\bar{x}$とnで割った標本分散 $\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ が最尤推定量である．

◎ 注意点

最尤推定量と不偏推定量が常に同時に成立しているわけではないことに注意が必要である．実際正規分布を仮定した場合，母平均μの推定量$\bar{x}$は不偏推定量かつ最尤推定量であるが，母分散σ^2の不偏推定量は$n-1$で割った標本分散s^2であり，最尤推定量はnで割った標本分散である．実際，分散の最尤推定量の期待値は実際の値より少し小さい値となる．

５．２．４　その他の性質

最小分散推定量

　母数θの不偏推定量のうち，分散が小さいものほどよい推定量であり，最小となるものを最小分散推定量とよぶ．しかし，多くの不偏推定量の分散を求め，　それを比較するのは大変である．母集団分布がある正則条件を満たしていれば，不偏推定量$\hat{\theta}$についてクラーメルラオの不等式

$$Var(\hat{\theta}) \geq \frac{1}{n\int_{-\infty}^{\infty}\left\{\frac{\partial f(x:\theta)}{\partial\theta}\right\}^2 f(x:\theta)dx}$$

が成り立つ．これから$\hat{\theta}$の分散は上式の右辺より小さくすることができないことがわかる．不等式の等号が成り立つ$\hat{\theta}$をθの有効推定量という．

十分性

統計量$T(X_1, X_2, \cdots, X_n)$が母数θに関する情報をすべて含んでいるときTをθの十分統計量であるという．Tがθの十分統計量であるという必要十分条件は尤度関数Lが

$$L(\theta|x_1, x_2, \cdots, x_n)$$
$$= g(\theta|t) \cdot h(x_1, x_2, \cdots, x_n)$$

と分解できることである(分解定理)．ここに，gとhは非負関数であり，hはθによらないものとする.推定量$\hat{\theta}$が十分統計量であるとき十分推定量であるという.

例) ポアソン分布の場合

$X_1, \cdots, X_n$ をパラメータ λ のポアソン分布に従う独立な確率変数とする．このとき，標本平均 $\bar{X}$ は λ に対する十分統計量である．

＜証明＞ 尤度関数 $L(\lambda|x_1, \cdots, x_n)$ は

$$L(\lambda|x_1, \cdots, x_n) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdot \cdots \cdot \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$

となる．この式は

$$= e^{-n\lambda}\lambda^{(x_1+x_2+\cdots+x_n)}\frac{1}{x_1!x_2!\ldots x_n!}$$

$$= \left(e^{-n\lambda}\lambda^{n\bar{x}}\right)\left(\frac{1}{x_1!x_2!\ldots x_n!}\right)$$

のように，分解できるので十分推定量である．

５．３ 区間推定

５．３．１ 平均の区間推定（母分散が既知のとき）

標本平均$\bar{X}$に基づいて母平均μの区間推定を行うために，$\bar{X}$の分布を考える．母集団が平均μ，分散σ^2の正規分布$N(\mu, \sigma^2)$に従うとき，無作為抽出されたn個の標本から計算された標本平均$\bar{X}$は$N(\mu, \frac{\sigma^2}{n})$に従うことが理論的にわかっている．このことから標本平均$\bar{X}$を

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

のように変換すると，Zは標準正規分布$N(0, 1)$に従うことがわかる．

そこで

$$P(-z_\alpha < Z < z_\alpha) = 1 - \alpha$$

になるようにz_αを求める.

このとき（　）内は

$$-z_\alpha < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_\alpha$$

$$\Rightarrow \quad \bar{X} - z_\alpha \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

と変形できるので，$[\bar{X} - z\alpha \cdot \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}]$の間に母平均$\mu$が入っている確率が$1 - \alpha$であることを示している.

このことを利用して

母集団から大きさnの標本$x_1, \cdots, x_n$が得られたときに，母平均μが確率$1-\alpha$で入っている区間を

$$\left[\bar{x} - u(\alpha) \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + u(\alpha) \cdot \frac{\sigma}{\sqrt{n}}\right]$$

とする.

このような区間を母平均μの$100(1-\alpha)\%$信頼区間とよぶ.

５．３．２ 平均の区間推定（母分散が未知のとき）

基本的には母分散が既知の場合と同じ考え方であるが，母分散σ^2の代わりに標本分散S^2を用いた

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

を考える．

このとき，tは自由度$n-1$のt分布に従うことがわかっている．

そこで

$$P(-t_\alpha(m) < T < t_\alpha(m)) = 1 - \alpha$$

になるように$t_\alpha(m)$を求めると，確率の括弧の中身は母分散既知と同様に

$$-t_\alpha(m) < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < t_\alpha(m)$$

$$\Rightarrow \quad \bar{X} - t_\alpha(m) \times \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_\alpha(m) \times \frac{s}{\sqrt{n}}$$

変形できる.

つまり

母分散が未知の場合の母平均μの$100(1-\alpha)\%$信頼区間は

$$\left[\bar{x} - t_\alpha(m) \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_\alpha(m) \cdot \frac{s}{\sqrt{n}}\right]$$

となる.

５．３．３　２標本の平均の差の区間推定

二つの母集団がそれぞれ母分散が等しい正規分布 $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2)$ に従うとき, 母平均の差 $\mu_1 - \mu_2$ の区間推定を行うことを考える.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,,\ \bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

とすると

正規分布の再生性より $\bar{X} - \bar{Y}$ は正規分布 $N(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$ に従う

- 分散が既知のとき

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

と変換すればZが標準正規分布に従うので,
前回同様に,

$$P(-z_\alpha < Z < z_\alpha) = 1 - \alpha$$

になるようにz_αを求め,
次のような信頼区間を得ることができる.

$$(\bar{x} - \bar{y}) - z_\alpha \cdot \sigma\sqrt{\frac{1}{n} + \frac{1}{m}} < \mu < (\bar{x} - \bar{y}) + z_\alpha \cdot \sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

○ 分散が未知のとき

母分散σ^2の代わりに標本分散S^2を使うことを考える

$$S^2 = \frac{1}{(n-1)+(m-1)}\left\{(n-1)S_x^2 + (m-1)S_y^2\right\}$$

ここで

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\,,\ S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2$$

を得る．これからZの変換のかわりに

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

と変換するとTは自由度$f = (n-1)+(m-1)$のt分布に従っている．

よって

$$P(-t_\alpha(f) < T < t_\alpha(f)) = 1 - \alpha$$

になるように$t_\alpha(f)$を求め,

$$(\bar{x} - \bar{y}) - t_\alpha(f) \cdot s\sqrt{\frac{1}{n} + \frac{1}{m}} < \mu < (\bar{x} - \bar{y}) + t_\alpha(f) \cdot s\sqrt{\frac{1}{n} + \frac{1}{m}}$$

のような信頼区間を得ることができる

５．３．４ 母比率の区間推定

$X_1, X_2, \cdots, X_n$を,

$$P(X=1)=p\ ,\quad P(X=0)=q=1-p$$

のようなベルヌーイ試行からの標本とする.

このとき、$\hat{p}=\frac{1}{n}\sum_{i=1}^{n} X_i$ は pの不偏推定量となる.

また、$n\hat{p}=X_1+\cdots+X_n$は２項分布$B(n,p)$に従うが、nが十分大きいとき正規分布$N(np,npq)$で近似することができる.

したがって

母比率pの区間推定は平均の区間推定と同様に標準正規分布から

$$Z = \frac{n\hat{p} - np}{\sqrt{np(1-p)}}$$

$$P(-z_\alpha < Z < z_\alpha)) = 1 - \alpha$$

となるz_αを求め, 母数pを推定量$\hat{p}$で置き換えた

$$\hat{p} - z_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \mu < \hat{p} + z_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

が母比率pの$100(1-\alpha)$%信頼区間となる.

５．４　生存時間関数の推定

生存時間データとは

生存時間データは，最初に観察を始めてから，死亡や故障などの事象(イベント)が発生するまでの時間（時点）として与えられる．また、ある開始時点から，それに続く事象が発生するまでの時間は，生存時間とよばれる．

例えば …

- 出生から死亡までの時間
- 脳卒中や心筋梗塞などの疾患が発生するまで時間
- 疾患を治療してから再発にいたるまでの時間

生存時間データを扱う場合の問題

1. 観察開始時間が同等でないことがある.
2. 途中で観察不能が生じることがある.
3. ある一定期間で観察を中止し，結論を出す必要がある.

観察の途中で転居や，試験参加の拒否などで観察不能になるデータは，打ち切り（センサリング）とよばれる.

例）４人の患者にある治療法を行い，治療開始からある時点までの死亡を観察した次のデータを考える．

患者の生存時間

患者番号	生存時間(年)
1	1
2	2
3	4*
4	5

生存時間の欄に書かれている数字，たとえば2は，２年生存して，その後に死亡したことを意味しているが，*がついているものは，観察不能が生ずるまでの時間を表していて，打ち切りを意味している．

T：生存時間（非負値確率変数）

とすると、その分布関数を

$$F(t) = P(T \leq t) = \int_{-\infty}^{t} f(u)du$$

とする．ここに，$f(t)$ は T の確率密度関数である．

生存時間解析においては，ある個人がある時点 t を越えて生存する確率に関心がある．

これは生存時間の分布において，T が時点 t より大きい確率 $S(t)$ であって

$$S(t) = P(T > t) = 1 - F(t)$$

と表わされる．

$S(t)$ は生存時間関数とよばれ，そのグラフは生存曲線とよばれる．生存期間 T に関する情報は，生存時間関数 $S(t) = P(T > t)$ で示される．

カプラー・マイヤー法

前の例で考えると、ゼロ時点に死亡した患者はいないので$\hat{S}(0) = 1$と推定する．そして，この値は次の死亡が生じるまで一定であると考える．次の死亡は$t = 1$で生じ，その時点における生存確率は確率の乗法定理を利用して求める．

1. 時点$t = 1$の直前まで生存したという条件のもとで，時点$t = 1$における死亡率を

$$q_1 = \frac{(\text{時点}t = 1\text{における死亡数})}{(\text{時点}t = 1\text{の直前まで生存していた患者数})} = \frac{1}{4}$$

と推定する．

2. ある患者が$t=1$の直前まで生存する事象をAとし、その患者が時点$t=1$において生存している事象をBとする．このとき，患者が$t=1$より長生存する確率は

$$S(1)=P(T>1)=P(A\cap B)=P(A)P(B|A)$$

と表すことができ、 この関係を利用して，$S(1)$を

$$\begin{aligned}\hat{S}(1)&=\hat{S}(0)\times(1-q_1)\\&=1\times\left(1-\frac{1}{4}\right)=0.75\end{aligned}$$

と推定できる．

この値は次の死亡が生じる$t=2$の直前まで同じとする．

以下同様に

時点$t=2$の直前まで生存したという条件のもとで,
時点$t=2$における死亡率を

$$q_2 = \frac{(\text{時点}t=2\text{における死亡数})}{(\text{時点}t=2\text{の直前まで生存していた患者数})} = \frac{1}{3}$$

と推定し、さらに,

$$\hat{S}(2) = \hat{S}(1) \cdot \left(1 - \frac{1}{3}\right) = 0.5$$

と推定できる.

次に死亡が生じるのは，$t=4$ で打ち切りがあるがこれは死亡であるとは見なさないで，$t=5$ と考える．

時点 $t=5$ の直前まで生存したという条件のもとで，時点 $t=5$ における死亡率は

$$q_5 = \frac{(\text{時点}\,t=5\,\text{における死亡数})}{\text{時点}\,t=5\,\text{の直前まで生存していた患者数}} = \frac{1}{1}$$

と推定する．

ここでの計算においては，時点 $t=4$ で打ち切られた患者は，$t=5$ の直前まで生存していた患者には含めないで計算されることに注意する．

$$\hat{S}(5) = \hat{S}(2) \cdot \left(1 - \frac{1}{1}\right) = 0$$

と推定できる.

$t = 5$ 以降には死亡がないので, $5 \leq t$ において $\hat{S}(t) = 0$ と推定する.

死亡があった時点で計算される死亡率の推定値 q_1, q_2, q_5 に対して, $1 - q_1, 1 - q_2, 1 - q_5$ は生存率成分とよばれる.

- カプラン・マイヤー推定法は次のように一般化される.

死亡があった時点を

$$t_1 < t_2 < \ldots < t_k$$

とする.

d_j : 時点 t_j における死亡数

n_j : 時点 t_j の直前まで（t_j を含まない）の生存者数

このとき

カプラン・マイヤー推定量$S(t)$は次のように定義される.

t_1より小さいtに対しては$S(t)=1$であり,

$t_1 \leq t < t_2$のとき;

$$S(t) = \left(1 - \frac{d_1}{n_1}\right),$$

$t_2 \leq t < t_3$のとき;

$$S(t) = \left(1 - \frac{d_1}{n_1}\right) \cdot \left(1 - \frac{d_2}{n_2}\right),$$

$$\vdots$$

$t_{k-1} \leq t < t_k$のとき;

$$S(t) = \left(1 - \frac{d_1}{n_1}\right) \cdot \left(1 - \frac{d_2}{n_2}\right) \cdots \left(1 - \frac{d_{k-1}}{n_{k-1}}\right)$$

とし, $t_k \leq t$のとき$\hat{S}(t)=0$とする.