統計(医療統計)
第8回　13章 区間推定

授業担当：徳永伸一

東京医科歯科大学教養部　数学講座

第13章 推定

Ⅰ．母集団と標本

Ⅱ．点推定

- 不偏性，不偏推定量

＊＊＊前期試験範囲ここまで＊＊＊

Ⅲ．区間推定

Ⅳ．母平均の区間推定

1. 母分散が既知のとき
2. 母分散が未知のとき

Ⅴ．母分散の区間推定

Ⅵ．母比率の区間推定

Ⅲ. 区間推定

区間推定とは

- 「母数の値をズバリ推定するより，母数が（高い確率で）存在する区間を推定する」という考え方.
- 「推定値が母数にどのくらい近いか（誤差がどのぐらいあるか）」も含めて推定する.

具体的には

- 「母数θが区間Ⅰに含まれる確率が○%」といった形の推定を行なう.
  - 「θの推定値θ'の誤差が確率○%でd以下」と考えても同じ。
  - $|\theta-\theta'| \leqq d \Leftrightarrow \theta \in [\theta'-d, \theta'+d]$（＝Ⅰ）
- ↑における
  - 区間・・・（○%）信頼区間
  - 確率・・・信頼度（または信頼係数）・・・95%，99%など

Ⅳ. 母平均の区間推定

母平均μの区間推定(母分散既知の場合)

問題設定:

- 標本サイズn, 標本平均 $\bar{X}$ の実測値が与えられており, 母分散σ^2は既知とする.
- 母集団分布・・・正規分布なら好都合だが, 任意の分布でもnが大きければ, 標本平均の分布は中心極限定理により正規分布で近似可能.

以上の条件のもとで,

「μの 100γ%信頼区間」

を求める(γは信頼度＝信頼係数で, 具体的には 0.95, 0.99, 0.90など).

→続いて信頼区間の求め方, でもその前に・・・

[あらためて前期の復習] Ⅵ.標本分布（2）

標本平均の期待値と分散･標準偏差

$X_1, X_2, \cdots, X_n$を平均μ，分散σ^2である母集団から無作為抽出した標本とするとき，

$X_1, X_2, \cdots, X_n$はそれぞれ，期待値μ，分散σ^2の互いに独立な確率変数と見なせる．

よって標本平均$\overline{X}$について

$$E(\overline{X}) = \mu \times n \times (1/n) = \mu$$

$$V(\overline{X}) = \sigma^2 \times n \times (1/n)^2 = \sigma^2/n$$

（期待値･分散の加法性↑）　（↑積に関するE,Vの性質より）

$$\sigma(\overline{X}) = \sqrt{V(\overline{X})} = \sqrt{(\sigma^2/n)} = \sigma/\sqrt{n}$$

[前期の復習]標本平均の分布・まとめ（対比して再確認）

定理（正規分布の性質より）

$X_1, X_2, \cdots, X_n$を正規分布$N(\mu, \sigma^2)$に従う母集団から無作為抽出した標本とすると

$$\text{標本平均 } \bar{X} \sim N\left(\mu, \ \sigma^2/n\right)$$

定理（中心極限定理の系）

$X_1, X_2, \cdots, X_n$を平均μ，分散σ^2である任意の母集団から無作為抽出した標本とするとき，

標本サイズnが十分大きければ，近似的に

$$\text{標本平均 } \bar{X} \sim N\left(\mu, \ \sigma^2/n\right)$$

となる.

★以上を踏まえて区間推定の具体的な方法へ

Ⅳ. 母平均の区間推定

母平均μの区間推定(母分散既知)の解法

- $\bar{X} \sim N(\mu, \sigma^2/n)$と近似(中心極限定理による).
  - 注意:正規母集団を仮定すれば厳密.
- $Z=(\bar{X}-\mu)/(\sigma/\sqrt{n})$ と標準化すると $Z \sim N(0, 1)$
- $P(-z(\alpha/2) \leqq Z \leqq z(\alpha/2))=\gamma$
  - ただし$\gamma=1-\alpha$. $z(\alpha)$ は $P(Z \geqq z(\alpha))=\alpha$ を満たす値.
  - たとえば$\gamma=0.95$ のとき$\alpha/2=0.025$, $z(\alpha/2)=z(0.025)=1.96$
  - $z(\alpha/2)$は「上側$100(\alpha/2)$%点」と呼ばれる.
- ↑を同値変形すると

$$P(\bar{X}-z(\alpha/2)\,\sigma/\sqrt{n} \leqq \mu \leqq \bar{X}+z(\alpha/2)\,\sigma/\sqrt{n})=\gamma$$

よって「μの(100×γ)%信頼区間」は

$$(\bar{X}-z(\alpha/2)\,\sigma/\sqrt{n},\ \bar{X}+z(\alpha/2)\,\sigma/\sqrt{n})$$

Ⅳ. 母平均の区間推定

教科書p.107例題1

コレステロールの平均値μの区間推定

- 母集団・・・成人男子
  （正規分布はp.106で仮定）
- 標本サイズn＝36（人）
- 標本平均$\bar{X}$＝63（mg/dl）・・・μの点推定値
- 母標準偏差σ＝12（ mg/dl ）

以上の条件のもとで,

「μの95%信頼区間を求めよ」という問題.

（信頼度95%）

Ⅳ. 母平均の区間推定

例題1に関する補足

- 正規母集団の仮定がどこにも明記されていなければ，下記のいずれかの方針で考える.
  - 正規母集団に十分近い分布であると考えて，正規母集団の仮定を導入.
  - n＝36を十分大きいと見なし，標本平均の分布を正規分布で近似（中心極限定理より）.

- 公式（$\bar{X} - z(\alpha/2)\sigma/\sqrt{n}$　$\bar{X} + z(\alpha/2)\sigma/\sqrt{n}$）
  に値を代入すれば一応答えは出ます.

- 母標準偏差σの値が既知というのは，かなり虫のいい仮定.
  - 現実的にはあまりないケースと思われるが，基本的な原理と手法を理解するためにあえて導入している.

Ⅳ. 母平均の区間推定

「μの(100×γ)%信頼区間」:

$$\left(\bar{X} - z(\alpha/2)\,\sigma/\sqrt{n} \;,\; \bar{X} + z(\alpha/2)\,\sigma/\sqrt{n} \right)$$

について:

- 標本平均(の実測値)を中心とする区間である.
- $\sigma/\sqrt{n}$ は**標準誤差**と呼ばれる.

[その他の重要な考察]

- γを大きくすると・・・
  - $\alpha = 1-\gamma$ は小さくなる → $z(\alpha/2)$ は大きくなる.
    →信頼区間の幅が大きくなる.
    (外れる確率を減らすのだから、幅を大きく取る必要があるのは当然)
- nを大きくすると → 信頼区間の幅が小さくなる.
  - 情報量が増えるのだから, 誤差が減るのは当然

Ⅳ. 母平均の区間推定

「母分散既知の場合」のポイントをまとめると

- 母分散σ^2を用いて標本平均の分布が表せる.
- 母集団分布が正規分布なら標本平均の分布も正規分布.
- 正規母集団を仮定せずとも，標本サイズnが十分大きければ標本平均の分布は正規分布で近似でき，いずれにしても正規分布の問題に帰着できる.
- だがその（標本平均が従う）正規分布$N(\mu, \sigma^2/n)$は，母分散σ^2を用いて表されているのだから，σ^2の値がわからないと推測できない.
- 現実にはσ^2は未知のケースが多い！

→「母分散未知」のケースへ.

Ⅳ. 母平均の区間推定（母分散未知）

母平均μの区間推定（母分散未知の場合）

母分散σ^2が未知であっても

標本の不偏分散　$U^2 := \{ \Sigma (x_i - \bar{x})^2 \} / (n-1)$

は，常にわかる（標本データから計算できる）.

そこで母分散未知の場合には：

- 方針1：U^2をσ^2の近似値（推定値）として利用.
  - U^2はσ^2の（不偏）推定量ですからね.
  - 「近似」を認めてしまえば，またしても正規分布の問題に帰着．推定の方法はほとんど同じ.

    （Uの値をσのところに代入するだけ．同じ公式が使える）
  - 標本サイズが大きければ，良い（誤差の少ない）近似値であることが期待できる.
  - だが小標本の場合は誤差が無視できないはず.

    →方針2へ

Ⅳ. 母平均の区間推定（母分散未知）

＊＊＊ここからまったく新しい内容＊＊＊

- 方針2：σ^2を用いずに表せる（U^2を含む）統計量を導入する．

→第11章Ⅴ

3-その他の重要な標本分布[2] t分布

以下の定理を利用：

定理（教科書p.100）

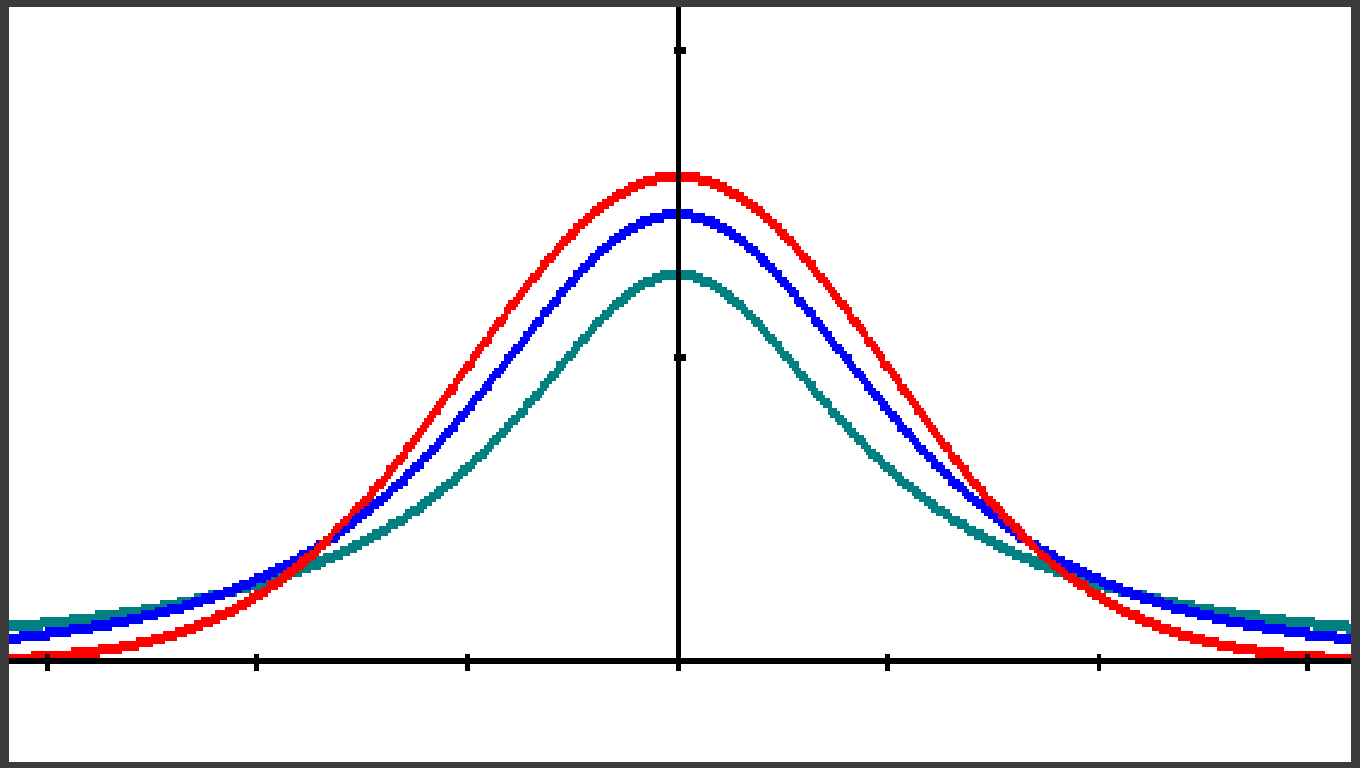
正規母集団$N(\mu, \sigma^2)$からの，大きさ n の無作為標本 $\{X_1, X_2, \cdots, X_n\}$ について，標本平均X’，不偏分散を U^2とする．このとき，確率変数

$$T := (X' - \mu)/(U/\sqrt{n})$$

は自由度n-1のt分布（と呼ばれる分布）に従う．

t分布の密度関数のグラフ

（明星大学 船津好明先生のウェブサイトより転載）



青：自由度3，暗緑：自由度1（赤：標準正規分布）

Ⅳ. 母平均の区間推定（母分散未知）

t分布について

- 「スチューデントのt分布」とも呼ばれる.
  - 「スチューデントstudent」は発案者のペンネーム.
  - 定理中の$T=(X'-\mu)/(U/\sqrt{n})$はスチューデント比とも呼ばれる.
- 厳密に定義するのは少々やっかいなので省略. ともかく理論的に「わかっている」分布として（巻末の数表などを用いて）利用する.
- 密度関数の概形は正規分布と似ている. 対称性のある釣鐘型.
- 自由度と呼ばれるパラメータを持つ.
  - サイズnの標本から求めたスチューデント比Tにおいては，自由度ν（ギリシャ文字の「ニュー」）$=n-1$.
  - 自由度とは大雑把に言えば「独立なパラメータの個数」.
  - ν大きいほど標準正規分布N(0,1)に近づき，$\nu=\infty$で完全に一致.
  - 自由度νのt分布の上側100α%点を$t_\nu(\alpha)$または$t(\nu,\alpha)$で表す.
  - $t_\nu(\alpha)$の値は教科書p.137の表から.

Ⅳ. 母平均の区間推定（母分散未知）

注意と補足

- 「正規母集団」の仮定はt分布を利用する上で本質的に重要.
- 標本平均X'の標準化変数:

  $Z=(X'-\mu)/(\sigma/\sqrt{n})$ との類似に注意.
  - つまりnが大きければ，T≒Zであるということ.
    （ただし，今考えているのはnが小さいケースでした）

Ⅳ. 母平均の区間推定（母分散未知）

t分布を利用したμの推定

- 母分散σ^2が未知の正規母集団$N(\mu, \sigma^2)$からの，大きさnの無作為標本$\{x_1, x_2, \cdots, x_n\}$について，標本平均$x'$，不偏分散を$U^2$とする．

このとき

$$t = (x' - \mu) / (U/\sqrt{n}) \sim (\text{自由度}n-1\text{のt分布})$$

だから

- 母平均μの信頼度$\gamma = 1-\alpha$の信頼区間は

$$\left(x' - t_{n-1}(\alpha/2)\, U/\sqrt{n},\ x' + t_{n-1}(\alpha/2)\, U/\sqrt{n} \right)$$

【注意】やってることは母分散既知の場合と大差ない．自由度という新たなパラメータが出てきただけ．

[RECALL]母分散既知の場合のμの$100(1-\alpha)$%信頼区間：

$$\left(\bar{X} - z(\alpha/2)\, \sigma/\sqrt{n},\ \bar{X} + z(\alpha/2)\, \sigma/\sqrt{n} \right)$$

Ⅳ. 母平均の区間推定（母分散未知）

例題2（教科書p.108）

- 母集団：成人女性HDLコレステロール値
  （正規母集団はp.106で仮定）
- 標本サイズn＝10，標本平均x'＝70
- 標本の不偏分散U^2＝14×14×（10/9）・・・★
  （「標準偏差」の定義に注意！）
- 信頼度＝0.95

［解答］

自由度＝n－1＝10－1＝9

教科書p.137の表より $t_9(0.025)$＝2.262を読み取る.

95%信頼区間は

$$\left(x' - t_{n-1}(0.025)\, U/\sqrt{n},\ x' + t_{n-1}(0.025)\, U/\sqrt{n} \right)$$

に該当する値を代入して

＝（　　，　　）

Ⅴ．母分散の区間推定

都合により割愛します

→Ⅵ．母比率の区間推定　へ

Ⅵ. 母比率の区間推定

「母集団比率に関する推測」とは

- ベルヌーイ母集団に関する推測.
  - 2項分布の応用.
  - ある程度大きな標本を扱うケースがほとんどなので，たいていは「2項分布の正規近似」を利用する.
- 「世論調査の類」．支持率調査など，身近に興味深い例が多い.
  - 「統計的な理解を深めるよいチャンス」.

Ⅵ. 母比率の区間推定

2項分布の正規近似RECALL

中心極限定理により，nが十分大きいとき，

B(n,p)はN(np, np(1−p))で近似できる.

∵ B(n,p)に従う確率変数は，B(1,p)(という同一の分布)に従う独立なn個の確率変数の和と見なせるから.

- 従って，標本比率P＝X/nの分布も，
  正規分布 N(p, p(1−p)/n) で近似できる.
- さらにPの標準化変数：

$$Z=(P-p)/\sqrt{(p(1-p)/n)}$$

  は近似的に標準正規分布に従う.
- 分布が決まれば，あとはこれまでと同じ考え方で進めればよいはず(？)

Ⅵ. 母比率の区間推定

とりあえずやってみる.

（以下母比率をp，標本比率P＝X/n の実現値をP_0とする）

$$Z=(P-p)/\sqrt{(\ p(1-p)/n)}$$

が近似的に標準正規分布に従うので

$$P\left(-z(\alpha/2) \leqq \ (P_0-p)/\sqrt{(\ p(1-p)/n)}\ \leqq z(\alpha/2)\right) = 1-\alpha$$

左辺のカッコ内を同値変形すると

$$P_0-z(\alpha/2)\sqrt{(\ p(1-p)/n)} \leqq \ p \ \leqq P_0+z(\alpha/2)\sqrt{(\ p(1-p)/n)}$$

すなわち、区間:

$$\left(P_0-z(\alpha/2)\sqrt{(\ p(1-p)/n)},\ P_0+z(\alpha/2)\sqrt{(\ p(1-p)/n)}\right) \quad \cdots(*)$$

に母比率pが含まれる確率が$1-\alpha$.

ところが(＊)は未知数であるpそのものを含んでいるので、これをそのまま信頼区間とすることはできない！

そこで・・・

Ⅵ. 母比率の区間推定

推定を行う際は,

$$\left(P_0 - z(\alpha/2)\sqrt{p(1-p)/n},\ P_0 + z(\alpha/2)\sqrt{p(1-p)/n}\right)$$

においてpを近似値（推定値）P_0で置き換える.

すなわち信頼度$\gamma = 1-\alpha$の信頼区間は:

$$\left(P_0 - z(\alpha/2)\sqrt{P_0(1-P_0)/n},\ P_0 + z(\alpha/2)\sqrt{P_0(1-P_0)/n}\right)$$

★ただし，誤差の最大値を見積もりたいときはp=0.5 を採用.

- p(1−p)はp=0.5のとき最大値0.25を取ることに注意（2次関数のグラフを思い出せ！）.
- 「誤差の最大値を見積もりたいとき」の例:

→誤差が一定値以下となるような標本サイズを決定する問題など.
（標本サイズを決定する時点ではP_0の値は得られていない！）

★（14章でやる）「母比率の検定」との違いに注意！

第13章 推定

Ⅰ．母集団と標本

Ⅱ．点推定

- 不偏性，不偏推定量

＊＊＊今回ここから＊＊＊

Ⅲ．区間推定

Ⅳ．母平均の区間推定

1. 母分散が既知のとき
2. 母分散が未知のとき

Ⅴ．母分散の区間推定・・・省略

Ⅵ．母比率の区間推定

＊＊＊今回ここまで＊＊＊