統計（医療統計）

前期・第６回　推測統計の初歩

授業担当:徳永伸一

東京医科歯科大学教養部　数学講座

もういちど Overview

- 確率（9章：6ページ） ･･･第1回授業
- 記述統計（10章： ４ページ） ･･･第2回授業

確率モデル（11章:18ページ）･･･第3 ～5回

- 推測統計（13章:7ページ, 14章:15ページ）
  - 推定（点推定、区間推定）
  - 仮説検定

（注:前期は推測統計のさわりまで）

第11章　確率変数と確率分布

Ⅰ．確率変数と確率分布の定義
Ⅱ．確率変数の特性値
- 期待値（平均），分散など
- 期待値と分散の性質
- 確率変数の標準化

Ⅲ．確率変数の独立性
Ⅳ．代表的な確率分布
- 2項分布，正規分布など　・・・前々回ここまで

Ⅴ．中心極限定理と正規近似・・・前回ここから
Ⅵ．標本分布

[復習] Ⅴ.中心極限定理と正規近似（1）

<u>中心極限定理1</u>

[仮定]$X_1, X_2, \cdots, X_n$が
（任意の！）同じ分布に従う独立な確率変数
ならば，
[結論]$n \to \infty$のとき，
和$X_1 + X_2 + \cdots + X_n$の分布は

正規分布に収束する！

[復習] Ⅴ.中心極限定理と正規近似（2）

中心極限定理2（言い換え）

「互いに独立な確率変数$X_1, X_2, \cdots, X_n$の分布が同一で，

$$E(X_k)=\mu,\ V(X_k)=\sigma^2 \quad (k=1,2,\cdots,n)$$

であるとき，nが十分大きければ，和 ΣX_k の分布は $N(n\mu, n\sigma^2)$ にで近似できる」

【注意】仮定すべき条件は独立性と同一分布性のみ．元の分布は任意．

[復習] Ⅴ.中心極限定理と正規近似（3）

二項分布の正規近似

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．

よって標準化変数

$$Z = (X-E(X)) / \sqrt{V(X)}$$

$$= (X-np) / \sqrt{(np(1-p))}$$

は近似的に$N(0,1)$に従う．

∵ $B(n,p)$に従う確率変数は，$B(1,p)$に従う 独立なn個の確率変数の和と見なせるから．

［復習］ Ⅴ.中心極限定理と正規近似（4）

半整数補正

- nが大きければかなり良い近似であると思われるが，nが小さいときはどのくらい誤差が出るのだろうか？
  - p.97問題10のケースで厳密値と正規近似の値を比較せよ．

- nが小さいときに少しでも誤差を減らす方法はないか？

⇒X～B(n,p)とする．整数 a, b に対し P(a≦X≦b) を正規近似で求める際，P(a−0.5≦X≦b+0.5) と補正して計算した方が誤差が減る．この補正を「不連続補正」ないし「半整数補正」といい，特にnが小さいときに効果的．
  - 区間を広げる方向に0.5ずらす（教科書p.97図11-7で確認 ）．
  - 再びp.97問題10で誤差の減少を確認．

S. TOKUNAGA

第11章 確率変数と確率分布

Ⅰ．確率変数と確率分布の定義

Ⅱ．確率変数の特性値
- 1-期待値と分散・標準偏差の定義
- 2-確率変数の期待値と分散の性質
- 3-確率変数の標準化

Ⅲ．確率変数の独立

Ⅳ．代表的な確率分布
- 2項分布，正規分布など

Ⅴ．中心極限定理と正規近似　←ここまできた

Ⅵ．標本分布　←次ここ

S. TOKUNAGA

［復習］ Ⅵ.標本分布（1）

1−母集団分布と標本分布

KEYWORDS：母集団⇔標本，無作為抽出，母集団分布，統計量，標本分布

★母集団から無作為抽出した個々のデータの値を確率変数をみなして，確率分布の理論を適用することができる！

2−標本平均の分布

- 個々の標本データの値 $X_1, X_2, \cdots, X_n$ はもちろん確率変数と見なすことができる.
- 標本平均 $\bar{X}$ も1つの確率変数とみなすことができる！

（一定の大きさの標本を繰り返し抽出し，その度に標本平均の値を計算すれば，「標本平均の分布」を観察することができる）.

よって・・・

［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 $\bar{X}$ について

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

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

（期待値・分散の加法性↑）　　（↑積に関するE,Vの性質より）

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

［復習］ Ⅵ.標本分布（3）

正規母集団を仮定すると･･･

<u>定理（正規分布の性質より）</u>

$X_1, X_2, \cdots, X_n$をN(μ , σ^2)に従う母集団から無作為抽出した標本とすると

- 和 $\Sigma X_k \sim N(n\mu, n\sigma^2)$
- 標本平均 $\bar{X} \sim N(\mu, \sigma^2/n)$

さらに $\bar{X}$ の標準化変数Zについて：

- $Z = (\bar{X} - \mu) / \sqrt{(\sigma^2/n)} \sim N(0,1)$

［復習］ Ⅵ.標本分布（4）

さらに！

$$n \times \bar{X} = X_1 + X_2 + \cdots + X_n$$

であるから，

nが十分大きければ，母集団分布が正規分布でなくても中心極限定理によって標本平均の分布を正規分布で近似できる！

注意：

- 同一分布性：同一の母集団から抽出したから
- 独立性：無作為抽出により保証される
- 正規分布に従う確率変数はnで割っても正規分布.

したがって･･･

[復習] Ⅵ.標本分布（5）

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大きさnの標本とするとき，

標本平均 $\bar{X}$ の分布は，nが十分大きければ，正規分布$N(\mu, \sigma^2/n)$で近似できる．

さらに $\bar{X}$ の標準化変数$Z = (\bar{X} - \mu)/\sqrt{(\sigma^2/n)}$は標準正規分布$N(0,1)$で近似できる．

【注意】母集団分布が任意でよいことにあらためて注目．これにより，（十分大きい標本さえ得られれば）未知の分布を持つ母集団の母平均を推定・検定する際，正規分布が利用できる！

S. TOKUNAGA

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

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

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

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

となる．

★「3−その他の重要な標本分布」は後回し（後期に）．

S. TOKUNAGA

第11章 確率変数と確率分布

Ⅰ. 確率変数と確率分布の定義
Ⅱ. 確率変数の特性値
1-期待値と分散・標準偏差の定義
2-確率変数の期待値と分散の性質
3-確率変数の標準化
Ⅲ. 確率変数の独立
Ⅳ. 代表的な確率分布
- 2項分布，正規分布など

Ⅴ. 中心極限定理と正規近似
Ⅵ. 標本分布 ←ここまで終了

いよいよ推測統計 へ

第13章 推定

Ⅰ. 母集団と標本
Ⅱ. 点推定
- 不偏性，不偏推定量

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

Ⅰ．母集団と標本

KEYWORDS：

- 母集団⇔標本sample， 無作為標本
- 母平均， 母分散
- 層別， 層別抽出stratified sampling
- 乱数

Ⅱ．点推定（1）

KEYWORDS：

- 母数，母集団パラメータ
  - 母平均，母分散
- （標本）統計量，統計値（統計量の実現値）
  - 標本平均，標本分散，標本比率
- 推定量（母数の推定に用いる統計量），
  推定値（推定量の実現値）

点推定と区間推定

- **点推定**：「無作為に選んだ標本から数値を得て，それから推定量の推定値を計算し，その推定値がイコール母数であるとする推定方法」
- **区間推定**：「推定値から，”ある確率である数値の区間の中にある”とする推定方法」

Ⅱ．点推定（2）

推定量が「不偏unbiasedである（偏りがない）」とは：

- 「対応する母数より大きい（or小さい）値が得られやすい」といった傾向がない.
- その推定量を繰り返し実測し、得られた値（推定値）の平均値は，繰り返しの回数を増やすほど対応する母数に近づく.

厳密には：

- ［定義］母数θの推定量θ'に対し，

$$E(\theta') = \theta$$

のときθ'をθの不偏推定量（不偏性を持つ推定量）であるという.

Ⅱ．点推定（3）

$X_1, X_2, \cdots, X_n$に対し

不偏分散 $U^2 := \{ \Sigma (X_k - \bar{X})^2 \} / (n-1)$

は，母分散σ^2の不偏推定量.

- すなわち，$E(U^2) = \sigma^2$ である（証明は割愛）.

ということは

- 分散$S^2 := \{ \Sigma (X_k - \bar{X})^2 \} / n$ については，

$$E(S^2) = E(((n-1)/n)U^2) = ((n-1)/n)\sigma^2$$

- つまり S^2の実測値は，母分散より小さめの値をとる傾向にある（すなわち不偏でない）.

不偏性に関する補足

- 標本平均 $\bar{X} := \Sigma X_k/n$ も母平均 μ の不偏推定量.

∵ $E(\bar{X}) = E(\Sigma X_k/n) = \Sigma E(X_k)/n = \mu$

- たとえばn=3のとき $X' = (X_1 + X_2 + 2X_3)/4$ とおいても X' は μ の不偏推定量(確認せよ).
- 不偏分散の平方根 $U = \sqrt{U^2}$ は, σ の不偏推定量ではない!

∵ $E(U) = E(\sqrt{U^2}) \neq \sqrt{E(U^2)} = \sigma$

第13章 推定

Ⅰ. 母集団と標本

Ⅱ. 点推定

- 不偏性, 不偏推定量 ←ここまで終わった

Ⅲ. 区間推定 ←次ここ

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に含まれる確率が〇％」といった形の推定を行なう.
  - 「θの推定値θ'の誤差が確率〇％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･･･（〇%）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･･･

［あらためて前回の復習］　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 $\bar{X}$ について

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

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

（期待値・分散の加法性↑）　（↑積に関するE,Vの性質より）

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

［前回の復習］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

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

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

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

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$\sigma=12$(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

<u>例題1に関する補足</u>

- 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．

29　S. TOKUNAGA

Ⅳ．母平均の区間推定

「μ の（100× γ）%信頼区間」:

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

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

30　S. TOKUNAGA

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（μ，σ^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（後期）

第13章　推定

Ⅰ．母集団と標本

Ⅱ．点推定

- 不偏性，不偏推定量

Ⅲ．区間推定

Ⅳ．母平均の区間推定

1. 母分散が既知のとき　←ここまで終わった
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　←ここを先取り

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」.

S. TOKUNAGA

33

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(？)

S. TOKUNAGA

34

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

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

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：

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

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

<u>推定</u>を行う際は，

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

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：

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

★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

Ⅲ. 区間推定

Ⅳ. 母平均の区間推定

1. 母分散が既知のとき ←ここまで終わった
2. 母分散が未知のとき ←後期ここから

Ⅴ. 母分散の区間推定

Ⅵ. 母比率の区間推定 ←ここを先取り