クレジット:
UTokyo Online Education 統計データ解析 I 2017 小池祐太

東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

統計データ解析 (I) 第 10 回

小池祐太

2017 年 12 月 6 日

1. 基礎的な記述統計量とデータの集約
   - モーメントに基づく統計量
     - 相関と共分散
   - 順序に基づく統計量
     - メディアン・分位点
     - ばらつきの指標
   - 頻度に基づく統計量

2. 推定
   - 点推定
   - 最尤法
   - 区間推定
   - 正規母集団の場合の区間推定
   - 漸近正規性による区間推定

基礎的な記述統計量とデータの集約

- **記述統計量**とはデータを簡潔に要約して表すための統計値のことで, 要約統計量, 基本統計量とも言われる
- ヒストグラム (あるいは密度関数) や箱ひげ図などのグラフと併用して, その集団全体の特徴を表す重要な指標となる
- ここでは, 比較的良く用いられる統計量を, その背景となるモーメント, 順序, 分布という考え方に基づいて分類する

モーメントに基づく統計量: 相関と共分散

- 複数のデータが与えられた場合, それらのデータの間の関係性を知りたい場合が頻繁に生じる
  そのような目的のための最も基本的な記述統計量に**相関**がある
  - 2 種類のデータ間の比例関係の大きさを計測
- 2 種類のデータ $x_1, x_2, \ldots, x_N$ および $y_1, y_2, \ldots, y_N$ に対して, それらの相関は

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{1}$$

  で定義される ($\bar{x}$ および $\bar{y}$ はそれぞれ $x_1, x_2, \ldots, x_N$ および $y_1, y_2, \ldots, y_N$ の平均)

モーメントに基づく統計量: 相関と共分散

- 相関は -1 以上 1 以下の値をとり, 1 に近いほど正の比例関係が強く, -1 に近いほど負の比例関係が強いことになる
- なお, (1) の分子の統計量を $N-1$ で割ったものは**共分散**と呼ばれる
- 相関および共分散はそれぞれ関数 `cor()` および関数 `cov()` で計算できる
  - 2 種類のデータ `x` および `y` が与えられたとき, それらの相関は `cor(x,y)` で計算できる
  - `x` がデータフレームのとき, `cor(x)` は (i, j) 成分が `x` の i 列と j 列の間の相関であるような行列 (**相関行列**) を計算
  - 共分散についても同様
- 実行例 `cor.r`

順序に基づく統計量: メディアン・分位点

- データの順序にもとづく記述統計量もよく利用される
- 例えば, $X_1, \ldots, X_n$ の**最大値**は関数 `max()` で, **最小値**は関数 `min()` でそれぞれ計算できる
- データを

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

のように昇順に並べ替えた際に中央の位置にくる値を**メディアン**もしくは**中央値**と呼ぶ

  - n が奇数の場合, メディアンは $X_{((n+1)/2)}$ であり, n が偶数の場合は $(X_{(n/2)} + X_{(n/2+1)})/2$ である
  - メディアンは関数 `median()` で計算できる
  - メディアンは平均と同様データを代表する値だと考えられるが, 平均と比較して, 計算結果がデータに含まれる異常な値 (**外れ値**と呼ばれる) の影響を受けにくい

順序に基づく統計量: メディアン・分位点

- メディアンの一般化として, $\alpha \in [0, 1]$ に対して, その点以下のデータの個数が全体の (約)100α%になるような点を 100α**%分位点**と呼ぶ
  - 特に 25%分位点および 75%分位点をそれぞれ**第 1 四分位点**, **第 3 四分位点**と呼ぶ
  - **第 2 四分位点**は 50%分位点となるが, これはメディアンのことである
- ベクトル `x` の 100`alpha`%分位点は `quantile(x, alpha)` で計算できる
  - 分位点は一意的には定まらず, いくつかの計算方式がある: `help(quantile)` を参照すること
- 実行例 `order.r`

順序に基づく統計量: メディアン・分位点

- 確率分布に対しても分位点が定義され, 推定や検定において重要な役割を果たす
- $0 < \alpha < 1$ に対して, 連続分布の 100α**%分位点**は, その分布に従う確率変数を X としたとき, 不等式

$$P(X \leq x) \geq \alpha$$

を満たす実数 x のうち最小のものとして定義される
- そのような実数は常に存在し, それを q_α とすると,

$$P(X \leq q_\alpha) = \alpha$$

が成り立つことが知られている

順序に基づく統計量: メディアン・分位点

- $X_1, X_2, \ldots, X_n$ が独立同分布な確率変数の列のとき, $X_1, X_2, \ldots, X_n$ の $100\alpha\%$分位点は $n \to \infty$ のとき X_i たちの従う分布の $100\alpha\%$分位点の一致推定量となることが知られている
- 確率分布の分位点は, その分布の省略形が `xxx` の場合, 関数 `qxxx()` で計算できる
  - 例えば, 平均 `mu`, 標準偏差 `sigma` の正規分布の `100alpha`%分位点は,

    ```
    qnorm(alpha, mean = mu, sd = sigma)
    ```

    で計算できる
- 実行例 `quantile.r`

順序に基づく統計量: ばらつきの指標

- 順序に基づいてデータのばらつきを測るための記述統計量もいくつか存在する
- そのようなものとして, 最大値と最小値の差である**範囲**がある
- 範囲は外れ値の影響を大きく受けるので, 第 3 四分位点と第 1 四分位点の差である**四分位範囲**もよく使われる
- また, データ $X_1, X_2, \ldots, X_n$ のメディアンを m としたとき, $|X_1 - m|, |X_2 - m|, \ldots, |X_n - m|$ のメディアンを**メディアン絶対偏差**と呼ぶ
- 実行例 `range.r`

頻度に基づく統計量

- データの中で最も頻度が高く現れる値を, **モード**もしくは**最頻値**と呼ぶ
- モードはデータが有限個の値を取る場合に特に有効であるが, データが連続で無限に多くの値を取ることができる場合には注意が必要である
- 連続なデータの場合でも有限個の観測データに対してモードは定義できるが, 偶々観測値として現れた値なので, その意味はよく考えなくてはならない
  - 必要に応じて, 例えば区分的に集計するなどの工夫をすることもある
- 実行例 `mode.r`

推定

- これまでの講義で学習したように, 確率統計学では観測データをある確率変数たちの実現値と考えてモデル化する
- このとき, 確率変数たちの従う分布のもつなんらかの特性量 (平均や分散など) を評価したり, 分布そのものを決定することが統計解析の目標の 1 つとなるが, この作業を一般に**推定**と呼ぶ
- 以下では, 統計学で広く利用されている代表的な推定方法を説明する

推定

- 以下, 観測データが独立同分布な確率変数列 $X_1, X_2, \ldots, X_n$ がモデル化されている状況を考える
- この場合, X_i たちが従う共通の分布 $\mathcal{L}$ に関する推定を行うことが目標となる
- $\mathcal{L}$ としてすべての分布を考察対象とすると, 対象とする範囲が広くなりすぎて, サンプル数 n が十分大きくない限り応用上意味のある結論を導き出すことが困難
- そのため, 以下では確率分布 $\mathcal{L}$ を特徴づけるなんらかのパラメーター θ を考察対象とする
  - 例: $\mathcal{L}$ の平均・分散・歪度, ・尖度など

点推定

- $\mathcal{L}$ に含まれるあるパラメーター θ を $X_1, \ldots, X_n$ のある関数

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

  で推定することを, **点推定**と呼ぶ
- このとき, $\hat{\theta}$ は θ の**推定量**と呼ばれる
- 前回・今回の講義でみたように, 多くの記述統計量は, 適当なモデル化の下でなんらかのパラメーターの推定量とみなせる
  - 例: $\mathcal{L}$ の平均 μ を標本平均 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ によって推定することが点推定であり, $\bar{X}$ は μ の推定量となる

点推定

- 一般に, 1 つのパラメーターに対して推定量は無数に存在するため, 使うべき推定量を決定するために, 推定量の良さを評価する基準が必要
- 前章で議論したように, そのような基準として代表的なものに, **不偏性**と**一致性**がある
  - $\hat{\theta}$ が θ の不偏推定量であるとは, $\hat{\theta}$ の平均が θ, すなわち

$$E[\hat{\theta}] = \theta$$

    が成り立つこと
  - $\hat{\theta}$ が θ の (強) 一致推定量であるとは, $n \to \infty$ のとき $\hat{\theta}$ が θ に収束する確率が 1 であること
  - 例: 標本平均, 不偏分散はそれぞれ $\mathcal{L}$ の平均, 分散の不偏性かつ一致性をもつ推定量

点推定

- シンプルな状況では, 1 つのパラメーターに対して複数の不偏推定量が存在する場合も起こりうる
- 例えば, $\mathcal{L}$ の平均 μ の不偏推定量としては, 標本平均 $\bar{X}$ 以外にも, 例えば X_1 が考えられる
- より “自然な” 推定量の例を挙げると, $\mathcal{L}$ が直線 $x = \mu$ に関して線対称な密度をもつ連続分布であったならば, $X_1, \ldots, X_n$ のメディアンも μ の不偏推定量
- このような場合, 不偏推定量の中から使うべきものを決定するための基準を設ける必要がある

点推定

- 自然な基準として, 推定値のばらつき (分散) が最も小さいものを選ぶという基準が考えられる
- すなわち, パラメーター θ の不偏推定量 $\hat{\theta}$ で, θ の任意の不偏推定量 $\hat{\theta}'$ に対して

$$\mathrm{Var}[\hat{\theta}] \leq \mathrm{Var}[\hat{\theta}']$$

  を満たすようなものを選ぶということ
- このような推定量 $\hat{\theta}$ を**一様最小分散不偏推定量**と呼ぶ
- 一様最小分散不偏推定量を見出す方法の 1 つとして, 次の結果が利用できる:

点推定

- $\mathcal{L}$ は 1 つの (1 次元) パラメーター θ を含む連続分布であるとし, その確率密度関数 $f_\theta(x)$ は θ に関して偏微分可能であるとする
- このとき, 緩やかな仮定の下で, θ の任意の不偏推定量 $\hat{\theta}$ に対して以下の不等式が成り立つ:

$$\mathrm{Var}[\hat{\theta}] \geq \frac{1}{nI(\theta)}. \tag{2}$$

ただし,

$$I(\theta) = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial \theta} \log f_\theta(x) \right)^2 f_\theta(x) dx.$$

点推定

- 不等式 (2) は **Cramér-Rao の不等式**と呼ばれ, その下界 $1/(nI(\theta))$ は **Cramér-Rao 下界**と呼ばれる
- また, $I(\theta)$ は **Fisher 情報量**と呼ばれる
- Cramér-Rao の不等式より, もし θ の不偏推定量 $\hat{\theta}$ で分散が Cramér-Rao 下界 $1/(nI(\theta))$ に一致するものが存在すれば, それは一様最小分散不偏推定量となる

点推定

- 例として, $\mathcal{L}$ が平均 μ, 分散 σ^2 の正規分布でモデル化されている状況を考える
- このとき, 平均パラメーター μ に関する Fisher 情報量は

$$I(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sigma^4} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sigma^2}$$

  となるから, Cramér-Rao 下界は σ^2/n となる
- 従って, 標本平均 $\bar{X}$ の分散は Cramér-Rao 下界に一致するので, $\bar{X}$ は μ の一様最小分散不偏推定量である
- 実行例 `umvue.r`

最尤法

- 興味あるパラメーターが, 平均や分散といった記述統計量と自然に関連づけられるパラメーターではない場合, 推定量の構成が自明ではないことがある
- このような場合でも, 確率分布 $\mathcal{L}$ に対するモデルがいくつかのパラメーター $\theta_1, \theta_2, \ldots, \theta_p$ を除いて特定されている状況であれば, 一般的に適用可能な $\theta_1, \theta_2, \ldots, \theta_p$ の推定量の構成方法がいくつか知られている
- ここでは代表的な方法として最尤法を説明する
  - その他の有名な方法として**モーメント法**がある

最尤法

- まず $\mathcal{L}$ が離散分布の場合を考え, その確率関数を $f_{\boldsymbol{\theta}}(x)$ と書くことにする
  - 確率関数のパラメーター $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_p)$ への依存を明示するために添え字 $\boldsymbol{\theta}$ をつけている
- このとき, パラメーター $\boldsymbol{\theta}$ を一つ定めれば, 観測値として $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ が得られる理論上の確率を

$$\prod_{i=1}^{n} f_{\boldsymbol{\theta}}(x_i) = f_{\boldsymbol{\theta}}(x_1) \cdot f_{\boldsymbol{\theta}}(x_2) \cdots f_{\boldsymbol{\theta}}(x_n)$$

で求めることができる
  - 独立性より

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$$

となることに注意

最尤法

- 実際に得られている観測データは $X_1, X_2, \ldots, X_n$ であるから, パラメーター θ に対してそのような観測データが得られる理論上の確率は,

$$L(\theta) := \prod_{i=1}^{n} f_\theta(X_i)$$

  で与えられる
- $L(\theta)$ は観測データ $X_1, X_2, \ldots, X_n$ が現れるのにパラメーター θ の値がどの程度尤もらしいか測る尺度と解釈でき, θ の**尤度**と呼ばれる
- $L(\theta)$ を θ の関数とみなしたものを**尤度関数**と呼ぶ

最尤法

- **最尤法**は観測データに対して「最も尤もらしい」パラメーターを $\boldsymbol{\theta}$ の推定量として採用する方法
- すなわち, 尤度関数 $L(\boldsymbol{\theta})$ を最大化するパラメーター値 $\hat{\boldsymbol{\theta}}$ を $\boldsymbol{\theta}$ の推定量とする:

$$L(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}).$$

  - Θ は尤度関数の定義域
- 上の式は以下のようにも書かれる:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}).$$

- $\hat{\boldsymbol{\theta}}$ は**最尤推定量**と呼ばれる

最尤法

- なお, 尤度関数は積の形をしていて扱いにくいので, 和の形に直すために対数を取ることが多い:

$$\ell(\boldsymbol{\theta}) := \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(X_i).$$

  - 対数関数は狭義増加であるから, $\ell(\boldsymbol{\theta})$ の最大化と $L(\boldsymbol{\theta})$ の最大化は同義
- $\ell(\boldsymbol{\theta})$ は**対数尤度関数**と呼ばれる

最尤法

- 例として, $\mathcal{L}$ がパラメーター $\lambda > 0$ の Poisson 分布としてモデル化されている場合を考える
- このとき, 未知パラメーターは λ であり, 対数尤度関数は

$$\ell(\lambda) = \sum_{i=1}^{n} \log \frac{\lambda^{X_i}}{X_i!} e^{-\lambda} = \sum_{i=1}^{n} (X_i \log \lambda - \log X_i!) - n\lambda$$

で与えられる

最尤法

- いま, 少なくとも 1 つの i について $X_i > 0$ であると仮定する
- このとき, $\ell(\lambda)$ を微分すると,

$$\ell'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} X_i - n, \quad \ell''(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} X_i < 0$$

  を得る
- 従って方程式 $\ell'(\lambda) = 0$ の解が $\ell(\lambda)$ を最大化するから, $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$ が λ の最尤推定量である

最尤法

- $\mathcal{L}$ が連続分布の場合は, 確率関数の代わりに確率密度関数を用いて尤度を計算する
- 例として, $\mathcal{L}$ がパラメーター $\lambda > 0$ の指数分布としてモデル化されている場合を考える
- このとき, 未知パラメーターは λ であり, 対数尤度関数は

$$\ell(\lambda) = \sum_{i=1}^{n} \log \lambda e^{-\lambda x_i} = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

で与えられる

最尤法

- $\ell(\lambda)$ を微分すると,

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i, \quad \ell''(\lambda) = -\frac{n}{\lambda^2} < 0$$

  を得る
- 従って方程式 $\ell'(\lambda) = 0$ の解が $\ell(\lambda)$ を最大化するから,

$$\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

  が λ の最尤推定量である

最尤法

- 別の例として, $\mathcal{L}$ がパラメーター $\nu, \alpha > 0$ のガンマ分布としてモデル化されている場合を考える
- このとき, 未知パラメーターは ν, α であり, 対数尤度関数は

$$\ell(\nu, \alpha) = \sum_{i=1}^{n} \log \frac{\alpha^{\nu}}{\Gamma(\nu)} X_i^{\nu-1} e^{-\alpha X_i}$$

$$= n\nu \log \alpha - n \log \Gamma(\nu) + \sum_{i=1}^{n} \{(\nu - 1) \log X_i - \alpha X_i\}$$

で与えられる

最尤法

- $\ell(\nu, \alpha)$ を最大化するような ν, α は解析的には求まらないため, 実際の計算では数値的に求めることになる (以下の実行例を参照)
- 数値計算の際に対数尤度関数の勾配 (偏導関数からなるベクトル) があると便利なので計算しておく:

$$\frac{\partial \ell}{\partial \nu}(\nu, \alpha) = n \log \alpha - n\psi(\nu) + \sum_{i=1}^{n} \log X_i,$$

$$\frac{\partial \ell}{\partial \alpha}(\nu, \alpha) = \frac{n\nu}{\alpha} - \sum_{i=1}^{n} X_i$$

ここに, $\psi(\nu) = \frac{d}{d\nu} \log \Gamma(\nu)$ であり, ディガンマ関数と呼ばれる (R では関数 `digamma()` で計算できる)

最尤法

- 広い範囲の確率分布に対して, 最尤推定量は一致性を持つことが知られている
- 実行例: `mle.r`

区間推定

- 未知パラメーター θ を推定量 $\hat{\theta}$ で点推定した場合, 通常推定値は真のパラメーター値とは異なるため, 推定誤差が必ず存在する
- そのため, 推定結果の定量的な評価には, 推定誤差の評価が重要となる
- 統計学では, ある値 $\alpha \in (0, 1)$ を固定したとき,

$$P(l \leq \hat{\theta} - \theta \leq u) \geq 1 - \alpha$$

が成り立つような l, u を観測データから推定することで推定誤差の評価を試みる

  - 上の式の意味するところは, 「誤差 $\hat{\theta} - \theta$ が区間 $[l, u]$ の外側にある確率が α 以下」ということ

区間推定

- 上の式を変形すると,

$$P(\hat{\theta} - u \leq \theta \leq \hat{\theta} - l) \geq 1 - \alpha$$

となる

- 従って, ここで行っているのは, パラメーター θ が含まれているような確率が $1 - \alpha$ 以上となるような区間 $[\hat{\theta} - u, \hat{\theta} - l]$ を推定することだと言い換えられる
- このように, 未知パラメーターが含まれている確率があらかじめ決められたある値以上となるような区間を推定することを**区間推定**と呼ぶ

区間推定

- より一般には, 未知パラメーター θ とある値 $\alpha \in (0,1)$ に対して,

$$P(L \leq \theta \leq U) \geq 1 - \alpha$$

が成り立つような確率変数 L, U を観測データから求めることになる
  - 区間 $[L, U]$ を $100(1-\alpha)\%$**信頼区間**
  - L を $100(1-\alpha)\%$**下側信頼限界**
  - U を $100(1-\alpha)\%$**上側信頼限界**
  - $1-\alpha$ を**信頼係数**

とそれぞれ呼ぶ

- 慣習として $\alpha = 0.01, 0.05, 0.1$ とすることが多い

区間推定

- 信頼区間は幅が狭いほど真のパラメーターが取りうる値の範囲を限定することになるため, 推定精度が良いといえる
- 一方で, 信頼区間 $[L, U]$ の幅が狭いほど確率 $P(L \leq \theta \leq U)$ は小さくなるため, 最も推定精度の良い $100(1-\alpha)\%$信頼区間 $[L, U]$ は,

$$P(L \leq \theta \leq U) = 1 - \alpha$$

  を満たす
- そのため, 実行可能である限り, $100(1-\alpha)\%$信頼区間 $[L, U]$ の構成では上の式を満たすように L, U を決定する

正規母集団の場合の区間推定: 平均

- 最も基本的な場合として, X_i たちの従う分布が平均 μ, 分散 σ^2 の正規分布の場合に, μ および分散 σ^2 の区間推定をする方法を説明する
- はじめに, 分散 σ^2 がすでにわかっている場合に平均 μ の区間推定をする方法を説明する
- これには次の結果を用いる:

命題 1

$Z_1, Z_2, \ldots, Z_k$ を独立な確率変数列とし, 各 $i = 1, 2, \ldots, k$ に対して Z_i は平均 μ_i, 分散 σ_i^2 の正規分布に従うとする. このとき, $a_0, a_1, \ldots, a_k$ を $(k+1)$ 個の 0 でない実数とすると, $a_0 + \sum_{i=1}^k a_i Z_i$ は平均 $a_0 + \sum_{i=1}^k a_i \mu_i$, 分散 $\sum_{i=1}^k a_i^2 \sigma_i^2$ の正規分布に従う.

正規母集団の場合の区間推定: 平均

- 上の命題を,

$$k = n,\ \mu_i = \mu,\ \sigma_i^2 = \sigma^2,\ a_0 = 0,\ a_i = 1/n \quad (i = 1, \ldots, n)$$

として適用すると, 標本平均 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ は平均 μ, 分散 σ^2/n の正規分布に従うことがわかる

- 再び命題 1 を $k = 1$, $\mu_1 = \mu$, $\sigma_1^2 = \sigma^2/n$, $a_0 = -\sqrt{n}\mu/\sigma$, $a_1 = \sqrt{n}/\sigma$ として適用すると, 確率変数

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

が標準正規分布に従うことがわかる

正規母集団の場合の区間推定: 平均

- 従って, $\alpha \in (0,1)$ を定めたとき, $z_{1-\alpha/2}$ を標準正規分布の $100(1-\alpha/2)$%分位点とすれば,

$$P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

  が成り立つことがわかる (詳細は配布資料参照)
- カッコ内を μ について解くと,

$$P\left(\bar{X} - z_{1-\alpha/2}\cdot\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\cdot\sigma/\sqrt{n}\right) = 1-\alpha$$

正規母集団の場合の区間推定: 平均

- 従って, σ が既知であれば,

$$[\bar{X} - z_{1-\alpha/2} \cdot \sigma/\sqrt{n}, \bar{X} + z_{1-\alpha/2} \cdot \sigma/\sqrt{n}]$$

  が平均 μ の $100(1-\alpha)$%信頼区間を与える
- 実行例 `ci-mean.r`

正規母集団の場合の区間推定: 平均

- 分散 σ^2 が既知であることは稀
- σ^2 が未知の場合, 不偏分散 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ で代用するのが自然な考えである
- s^2 の分布については次の結果が知られている ($n \geq 2$ とする):

命題 2

$X_1, X_2, \ldots, X_n$ は独立同分布な確率変数列で, 平均 μ, 分散 σ^2 の正規分布に従うとする. このとき, $\bar{X}$ と s^2 は独立であり, 確率変数 $(n-1)s^2/\sigma^2$ は自由度 $n-1$ の χ^2 分布に従う.

正規母集団の場合の区間推定: 平均

- 上の命題と $\sqrt{n}(\bar{X}-\mu)/\sigma$ が標準正規分布に従うことから, 確率変数

$$\frac{\sqrt{n}(\bar{X}-\mu)}{s}\left(=\frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\sqrt{\frac{(n-1)s^2}{\sigma^2}\Big/(n-1)}}\right)$$

  は自由度 $n-1$ の t 分布に従うことがわかる (6.3.5 節参照)
- 従って, $\alpha \in (0,1)$ を定めたとき, $t_{1-\alpha/2}(n-1)$ を自由度 $n-1$ の t 分布の $100(1-\alpha/2)$%分位点とすれば,

$$P\left(-t_{1-\alpha/2}(n-1) \leq \frac{\sqrt{n}(\bar{X}-\mu)}{s} \leq t_{1-\alpha/2}(n-1)\right) = 1-\alpha$$

  が成り立つことがわかる

正規母集団の場合の区間推定: 平均

- カッコ内を μ について解くことで,

$$[\bar{X} - t_{1-\alpha/2}(n-1) \cdot s/\sqrt{n}, \bar{X} + t_{1-\alpha/2}(n-1) \cdot s/\sqrt{n}]$$

  が平均 μ の $100(1-\alpha)$%信頼区間を与えることがわかる
- 実行例 `ci-mean-unknown.r`

正規母集団の場合の区間推定: 分散

- 分散 σ^2 の区間推定には命題 2 を利用する
- すなわち, $(n-1)s^2/\sigma^2$ が自由度 $n-1$ の χ^2 分布に従うので, $\chi^2_{\alpha/2}(n-1)$, $\chi^2_{1-\alpha/2}(n-1)$ をそれぞれ自由度 $n-1$ の χ^2 分布の $100\alpha/2\%$, $100(1-\alpha/2)\%$分位点とすれば,

$$P(\chi^2_{\alpha/2}(n-1) \leq (n-1)s^2/\sigma^2 \leq \chi^2_{1-\alpha/2}(n-1)) = 1-\alpha \qquad (3)$$

が成り立つ (詳細は配布資料参照)

正規母集団の場合の区間推定: 分散

- (3) の左辺のカッコ内を σ^2 について解くと,

$$P\left((n-1)s^2/\chi^2_{1-\alpha/2}(n-1) \leq \sigma^2 \leq (n-1)s^2/\chi^2_{\alpha/2}(n-1)\right) = 1-\alpha$$

  が得られる
- 従って,

$$\left[(n-1)s^2/\chi^2_{1-\alpha/2}(n-1), (n-1)s^2/\chi^2_{\alpha/2}(n-1)\right]$$

  が σ^2 の $100(1-\alpha)$%信頼区間を与える
- 実行例 `ci-variance.r`

漸近正規性による区間推定

- 前節のように信頼区間を正確に計算できることは稀
- しかし, 未知パラメーター θ のある推定量 $\hat{\theta}$ について, 推定誤差 $\hat{\theta} - \theta$ の分布がある正規分布で近似できる状況はしばしばある
- このような推定量の性質を**漸近正規性**と呼ぶ
  - 中心極限定理によって, 多くのモーメントに基づく記述統計量は漸近正規性をもつ
  - 最尤推定量は広い範囲の確率分布に対して漸近正規性をもつことが知られている
- 漸近正規性をもつ推定量がある場合, 推定誤差がある区間に含まれる確率を近似的に求めることができるから, 近似的に正しい信頼区間を構成することが可能となる

漸近正規性による区間推定

- これまでの講義で確認したように, 確率分布 $\mathcal{L}$ が 2 次のモーメントを持てば, 中心極限定理より, $\mathcal{L}$ の平均 μ の推定量である標本平均 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ は漸近正規性をもつ
- より正確に述べると, $\mathcal{L}$ の標準偏差を σ とすれば, 任意の $a \leq b$ に対して,

$$P\left(a \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq b\right) \to \int_a^b \phi(x)dx \quad (n \to \infty) \tag{4}$$

が成り立つ (ϕ は標準正規分布の確率密度関数)

漸近正規性による区間推定

- 従って, $\alpha \in (0,1)$ を定めたとき, $z_{1-\alpha/2}$ を標準正規分布の $100(1-\alpha/2)$%分位点とすれば,

$$P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq z_{1-\alpha/2}\right) \to 1-\alpha \quad (n \to \infty)$$

  が成り立つ
- カッコ内を μ について解くと,

$$P\left(\bar{X} - z_{1-\alpha/2} \cdot \sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \cdot \sigma/\sqrt{n}\right) \to 1-\alpha \quad (n \to \infty)$$

  が得られる

漸近正規性による区間推定

- 従って, σ が既知であれば,

$$[\bar{X} - z_{1-\alpha/2} \cdot \sigma/\sqrt{n}, \bar{X} + z_{1-\alpha/2} \cdot \sigma/\sqrt{n}]$$

  はサンプル数 n が十分大きい場合に近似的に正しい平均 μ の $100(1-\alpha)\%$信頼区間を与える
- 通常は σ は未知であるが, 近似 (4) は σ をその一致推定量 $\hat{\sigma}$ で置き換えてもそのまま成立することが知られている

漸近正規性による区間推定

- 従って, 上の式で σ を $\hat{\sigma}$ で置き換えたもの

$$[\bar{X} - z_{1-\alpha/2} \cdot \hat{\sigma}/\sqrt{n}, \bar{X} + z_{1-\alpha/2} \cdot \hat{\sigma}/\sqrt{n}]$$

  も, サンプル数 n が十分大きい場合に近似的に正しい平均 μ の $100(1-\alpha)\%$信頼区間を与える
- $\hat{\sigma}$ としては例えば不偏分散の平方根

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

  を使うことができる
- 実行例 `ci-mean-asymp.r`