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 9 回

小池祐太

2017 年 11 月 29 日

1. 確率分布
   - 連続分布
     - ガンマ分布
     - t 分布
     - F 分布

2. 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
     - 平均・分散・標準偏差
     - 歪度と尖度
     - 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

確率分布

- X を確率変数とする
  - 値がランダムに決定される変数で, すべての実数 $a \leq b$ に対して, その値が区間 $[a, b]$ に含まれる確率があらかじめ定められているような変数
- 各区間 $[a, b]$ $(a \leq b)$ と, X が区間 $[a, b]$ に含まれる確率

$$P(a \leq X \leq b)$$

との対応を示したものを, X の**確率分布**または単に**分布**といい, X はこの分布に**従う**という

連続分布

- 任意の正の実数値を取りうるデータのように, 取りうる値が連続的なデータのモデル化には連続分布を用いる
- 一般に, 確率変数 X が**連続型**であるとは, 非負の値をとる実数直線上の関数 f があって, $a \leq b$ なるすべての実数 a, b に対して

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

  が成り立つことをいい, 対応する確率分布を**連続分布**と呼ぶ
- また, 関数 f をこの確率分布の**確率密度関数**, あるいは単に**密度**と呼ぶ

連続分布

- 確率変数 X をシミュレーションした際のヒストグラムのビン $[a, b]$ における高さは

$$\frac{1}{b-a} P(a \leq X \leq b)$$

で与えられる (関数 `hist()` でオプション `freq` を `FALSE` に指定した場合)

- 従って, 確率密度関数 f は, ビン $[a, b]$ の幅を限りなく小さくした場合のヒストグラムの形状の極限として現れるグラフに対応する
- 和のかわりに積分を用いることで, 離散分布の場合と同様に, 連続分布に対しても平均, モーメント, 分散, 標準偏差の概念が定義される

ガンマ分布

- ν, α を正の実数とする
- 確率密度関数が

$$f(x) = \frac{1}{\Gamma(\nu)}\alpha^{\nu}x^{\nu-1}e^{-\alpha x} \quad (x > 0), \qquad f(x) = 0 \quad (x \leq 0)$$

で与えられる連続分布をパラメータ ν, α の**ガンマ分布**と呼び，記号 $\Gamma(\nu, \alpha)$ や $G(\alpha, \nu)$ で表す

  - $\Gamma(\nu)$ はガンマ関数

$$\Gamma(\nu) = \int_0^{\infty} x^{\nu-1}e^{-x}dx$$

を表す

ガンマ分布

- ν, α はそれぞれ**形状パラメーター**, **レート**と呼ばれることがある
- 平均は ν/α, 分散は ν/α^2 で与えられる
- ガンマ分布に従う乱数の生成は関数 `rgamma()` で行う
- 実行例 `rgamma2.r`

指数分布

- ガンマ分布はいくつかの応用上重要な確率分布を特殊な場合として含む
- 正の実数 λ に対して, $\Gamma(1, \lambda)$ をパラメータ λ の**指数分布**と呼び, 記号 $\mathrm{Exp}(\lambda)$ で表す
- λ は**レート**と呼ばれることがある
- 指数分布の平均, 分散はそれぞれ λ^{-1}, λ^{-2} で与えられる
- 指数分布に従う乱数の生成は関数 `rexp()` で行う
- 実行例 `rexp.r`

χ^2 分布

- 正の実数 k に対して, $\Gamma(k/2, 1/2)$ を自由度 k の χ^2 **分布**と呼び, 記号 $\chi^2(k)$ で表す [1]
- 自由度 k の χ^2 分布の平均, 分散はそれぞれ k, $2k$ で与えられる
- χ^2 分布に従う乱数の生成は関数 `rchisq()` で行う
- 実行例 `rchisq.r`

[1] χ^2 は「カイ二乗」と読む

χ^2 分布

- 標準正規分布に従う k 個の独立な確率変数の二乗和は自由度 k の χ^2 分布に従うことが知られている
- 実行例 `rgamma-chi2.r`

t 分布

- ν を正の実数とする
- 確率密度関数が

$$f(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$

  で与えられる連続分布を, 自由度 ν の (Student の) t 分布と呼び, 記号 $t(\nu)$ で表す [2]
- 平均は $\nu > 1$ のときに限り存在し, 0 で与えられる
- 分散は $\nu > 2$ のときに限り存在し, $\nu/(\nu-2)$ で与えられる
- t 分布に従う乱数の生成は関数 `rt()` で行う
- 実行例 `rt2.r`

[2] Student は t 分布を導入した統計学者 Gosset のペンネームである

t 分布

- Z を標準正規分布に従う確率変数, Y を自由度 k の χ^2 分布に従う確率変数とし, Z, Y は独立であるとする. このとき, 確率変数

$$\frac{Z}{\sqrt{Y/k}}$$

  は自由度 k の t 分布に従うことが知られている
- 実行例 `normal-t.r`

F 分布

- ν_1, ν_2 を正の実数とする
- 確率密度関数が

$$f(x) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B(\nu_1/2, \nu_2/2)} \frac{x^{\nu_1/2-1}}{(1+\nu_1 x/\nu_2)^{(\nu_1+\nu_2)/2}} \qquad (x > 0),$$
$$= 0 \qquad (x \leq 0)$$

  で与えられる連続分布を, 自由度 ν_1, ν_2 の F 分布と呼び, 記号 $F(\nu_1, \nu_2)$ で表す
- 平均は $\nu_2 > 2$ のときに限り存在し, $\nu_2/(\nu_2 - 2)$ で与えられる
- 分散は $\nu_2 > 4$ のときに限り存在し, $\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ で与えられる
- F 分布に従う乱数の生成は関数 `rf()` で行う
- 実行例 `rf2.r`

F 分布

- Y_1 を自由度 k_1 の χ^2 分布に従う確率変数, Y_2 を自由度 k_2 の χ^2 分布に従う確率変数とし, Y_1, Y_2 は独立であるとする
- このとき, 確率変数

$$\frac{Y_1/k_1}{Y_2/k_2}$$

  は自由度 k_1, k_2 の F 分布に従うことが知られている
- 実行例 `normal-f.r`

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: 平均・分散・標準偏差

- $X_1, X_2, \ldots, X_n$ を観測データとする
- **(標本) 平均**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

  - データの代表値を表す記述統計量

- **(標本) 分散**

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n}$$

  - データのばらつき具合を表す記述統計量
  - 標本分散の平方根である **(標本) 標準偏差**も広く利用される

モーメントに基づく統計量: 平均・分散・標準偏差

- これまでの講義で説明してきたように, 確率統計学では, 一つ一つのデータ X_i をある確率変数の実現値とみなすことで, データの背後にある現象に対する統計解析を行う
- 確率変数 $X_1, X_2, \ldots, X_n$ が同分布であれば, 以前定義したように, X_i たちに共通の平均 μ および分散 σ^2 を考えることができる (もちろん適切な次数のモーメントの存在を仮定した下で)
- さらに, $X_1, X_2, \ldots, X_n$ が独立であれば, 大数の強法則より, 標本平均・標本分散・標本標準偏差はそれぞれ $n \to \infty$ のとき確率 1 で平均 μ・分散 σ^2・標準偏差 σ に収束する

モーメントに基づく統計量: 平均・分散・標準偏差

- これは, 標本平均・標本分散・標本標準偏差をそれぞれサンプル対象の集団の「真の」平均・分散・標準偏差の推定量と考えた場合に, これらの推定量がサンプル数 n が十分大のときに「まともな」推定量であるという根拠の 1 つを与える
- このような性質を推定量の **(強) 一致性**と呼び, 一致性をもつ推定量を **(強) 一致推定量**と呼ぶ

モーメントに基づく統計量: 平均・分散・標準偏差

- 一致性はサンプル数が十分大きい場合に推定量がまともであることの１つの根拠を与えるが, サンプル数が小さい場合の推定量の性質については何も語っていない
- そのような場合の推定量の良さに関する性質の１つとして**不偏性**がある
- 一般にパラメーター θ の推定量 $\hat{\theta}$ が不偏であるとは, $\hat{\theta}$ の平均が θ, すなわち

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

が成り立つことをいう

モーメントに基づく統計量: 平均・分散・標準偏差

- 標本平均は μ の不偏推定量である:

$$E[\bar{X}] = \mu$$

  が成り立つ
- 一方で, 標本分散は σ^2 の不偏推定量ではない. 実際,

$$E[S^2] = \frac{n-1}{n}\sigma^2$$

  が成り立つ
  - この式は, 標本分散は平均的には真の分散を過小推定する傾向にあることを意味する

モーメントに基づく統計量: 平均・分散・標準偏差

- このバイアスを補正するには, 標本分散に $n/(n-1)$ をかけてやれば良い. すなわち,

$$s^2 := \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

  は σ^2 の不偏推定量となる
- わざわざ不偏性を持たない S^2 を σ^2 の推定量として使う理由は通常ないので, 標本分散という場合には s^2 のことを指す場合もあるが, バイアス補正をしていることを強調するために**不偏分散**と呼ぶ場合もある
- R には不偏分散を計算するための関数として `var()` が用意されている

モーメントに基づく統計量: 平均・分散・標準偏差

- 同様に, 標本標準偏差という場合は, 通常, 不偏分散の平方根 s を指し, R では関数 `sd( )` で計算できる (ただし, 一般に s は標準偏差 σ の不偏推定量ではない)
- 実行例 `unbiased.r`

モーメントに基づく統計量: 平均・分散・標準偏差

- 複数のデータを同時に分析する場合, 単位や基準を揃えた方が扱いやすい
- このような目的でよく使われる方法に, データの**標準化**がある
- データ $X_1, X_2, \ldots, X_n$ の標準化は

$$Z_i = \frac{X_i - \bar{X}}{s} \qquad (i = 1, 2, \ldots, n)$$

で定義される [3]

  - 定義から明らかなように, $Z_1, Z_2, \ldots, Z_n$ の標本平均は 0, 不偏分散は 1 である (むしろ, そうなるようにデータを一次変換したものが標準化である)
  - 標準化は**標準得点**あるいは **Z スコア**とも呼ばれる

[3] s の代わりに S で割って定義する文献もある

モーメントに基づく統計量: 平均・分散・標準偏差

- 教育学や心理学では, データを標本平均が 50, 標準偏差 (不偏分散の平方根) が 10 となるように一次変換したもの

$$T_i = 10Z_i + 50 \qquad (i = 1, \ldots, n)$$

  を使う場合が多い
- これを**偏差値得点**あるいは **T 得点**と呼ぶ
- 実行例 `scale.r`

モーメントに基づく統計量: 歪度と尖度

- 中心極限定理が示すように, 正規分布は確率分布のうち最も基本的なものと考えられる
- 正規分布は平均と分散を決めると完全に決定されるから, 正規分布に従うデータを考える際には標本平均と標本分散 (不偏分散) を考えれば十分
- しかし, 現実には正規分布では捉えきれない特徴をもつデータに遭遇することがしばしばある
- そのようなデータを考える場合の最初のアプローチとして, 正規分布からのずれを調べるということがしばしば行われる
- そのための統計量として代表的なものに歪度と尖度がある

モーメントに基づく統計量: 歪度と尖度

- X を平均 μ, 分散 σ^2 をもつ確率変数とする
- X が 3 次のモーメントをもつとき,

$$\frac{E[(X-\mu)^3]}{\sigma^3}$$

を**歪度**と呼ぶ

  - 歪度は分布の非対称性を表す統計量で, 正の場合分布の右の裾の方が重く, 負の場合分布の左の裾の方が重いと考えられる
  - 左右に対称的な分布の歪度は 0 であり, 従って正規分布の歪度は 0 である
  - 正の歪度をもつ分布としては, 例えばガンマ分布 $\Gamma(\nu, \alpha)$ があり, その歪度は $2/\sqrt{\nu}$ で与えられる

モーメントに基づく統計量: 歪度と尖度

- X が 4 次のモーメントをもつとき,

$$\frac{E[(X-\mu)^4]}{\sigma^4}$$

を**尖度**と呼ぶ
  - 尖度は平均の周囲の分布の尖り具合を表す統計量だと考えられる
- 正規分布の場合 3 であるため, 正規分布との比較のため上の定義から 3 を引いた量

$$\frac{E[(X-\mu)^4]}{\sigma^4} - 3$$

のことを尖度と呼ぶ文献も多いが, 後者を前者と区別するために**超過尖度**と呼ぶ場合もある

モーメントに基づく統計量: 歪度と尖度

- 超過尖度が正の分布は正規分布よりも平均の周囲の分布が尖っており, 負の分布は丸みを帯びていると考えられる
  - 前者の場合, 平均まわりの密度が分布の裾の方にまわっていることが多いため, 正規分布より裾が重いと解釈されることが多い
- 正の超過尖度をもつ分布としては, 例えば自由度 $\nu > 4$ をもつ t 分布 $t(\nu)$ があり, その超過尖度は $6/(\nu - 4)$ で与えられる ($\nu \leq 4$ のときは $t(\nu)$ は 4 次モーメントをもたない)
- また, ガンマ分布 $\Gamma(\nu, \alpha)$ は超過尖度 $6/\nu$ をもつ

モーメントに基づく統計量: 歪度と尖度

- 観測データ $X_1, X_2, \ldots, X_n$ から歪度と尖度を推定するには, それらの標本バージョンを考えればよい
- すなわち, 歪度の推定量としては, **標本歪度**

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$$

を考えればよく, 尖度の推定量としては, **標本尖度**

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4}$$

を考えればよい

モーメントに基づく統計量: 歪度と尖度

- 標本歪度・標本尖度を計算するための関数はデフォルトでは R には実装されていないため, 自作するかパッケージを利用する
- 例えば, パッケージ `e1071` には標本歪度を計算するための関数 `skewness()` および標本尖度を計算するための関数 `kurtosis()` が実装されている
  - 後者は上の定義の標本尖度から 3 を引いたもの (すなわち標本超過尖度) を計算することに注意
- 標本歪度・標本尖度の値は標本平均・分散に比べてばらつきが大きくなる傾向があるため, サンプル数が少ない場合の計算結果の解釈には注意を要する
- 実行例 `skewkurt.r`

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`