記述統計量

基礎的な記述統計量とデータの集約

村田 昇

2020.06.12

基礎的な記述統計量

記述統計量

- **記述統計量**：(または要約統計量・基本統計量)
  - データを簡潔に要約して表すための統計値
  - その集団全体の特徴を表す重要な指標
  - ヒストグラム・密度関数・箱ひげ図などのグラフと併用
- 比較的良く用いられる統計量を以下の観点で分類
  - モーメント
  - 順序
  - 頻度

記述統計量の推定

- 記述統計量は背後の確率分布 (集団全体) で決まる量
- 一般に確率分布は未知
- 手に入るのは (少数の) サンプル (観測データ)
  (観測データを $X_1, X_2, \ldots, X_n$ で表す)
- **推定** = 観測データから知りたい量を計算する方法
- 真の値と観測データによる推定には **差** がある

独立同分布性

- 統計解析における重要な仮定とその帰結:
  - 確率変数 $X_1, X_2, \ldots, X_n$ が **同分布**
    共通の平均 μ および分散 σ^2 を考えることができる
    (適切な次数のモーメントの存在を仮定)
  - 確率変数 $X_1, X_2, \ldots, X_n$ が **独立同分布**
    標本平均はサンプル数 $n \to \infty$ のとき確率 1 で真の平均に収束
    (大数の強法則)
- データは **偏っていない** ことを仮定している

推定量の一致性

- サンプル数が大きい場合に「まともな」推定量となる根拠の 1 つ
  - **(強) 一致性** (consistency):
    推定量がサンプル数 $n \to \infty$ のとき確率 1 で真の値に収束する性質
  - **(強) 一致推定量**：一致性をもつ推定量

推定量の不偏性

- サンプル数が小さい場合の推定量の良さに関する性質の 1 つ
  - **不偏性** (unbiasedness):
    推定量 $\hat{\theta}$ が不偏であるとは, $\hat{\theta}$ の平均が真の値 θ となる性質

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

  - **不偏推定量**：不偏性をもつ推定量

“モーメント”に基づく記述統計量

平均

- **平均** (mean):

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

- **標本平均** (sample mean):

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

  データの代表値を表す記述統計量

分散・標準偏差

- **分散** (variance):

  $$\mathrm{Var}(X) = \sigma^2 = \mathbb{E}[(X-\mu)^2]$$

- **標本分散** (sample variance):

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

  データのばらつき具合を表す記述統計量

- **標本標準偏差** (sample standard deviation):
  標本分散の平方根

標本平均・分散の不偏性

- 標本平均は μ の **不偏推定量である**:

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

- 標本分散は σ^2 の **不偏推定量ではない**:

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

  (標本分散は平均的には真の分散を **過小推定** する)

不偏分散

- 不偏性を担保した分散の推定量
- バイアス補正: 標本分散に $n/(n-1)$ を乗じたもの

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

は σ^2 の不偏推定量となる

標本標準偏差

- 標本標準偏差: 通常, 不偏分散の平方根 s を指す
- 一般に s は標準偏差 σ の **不偏推定量ではない**

平均・分散・標準偏差の計算

- 基本書式

```
mean(x, trim = 0, na.rm = FALSE, ...) # 標本平均
var(x, na.rm = FALSE, ... ) # 不偏分散
sd(x, na.rm = FALSE) # 標本標準偏差
```

- 関数の引数
  - `x`: ベクトル，データフレームなど
  - `na.rm`: 欠損値を取り除くか否か

標準化

- 複数データの分析のために単位や基準を揃える
- データ $X_1, X_2, \ldots, X_n$ の標準化:

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

(s の代わりに S で割って定義する文献もある)

- 定義から $Z_1, Z_2, \ldots, Z_n$ の標本平均は 0, 不偏分散は 1 に規格化される
- Z_i: **標準得点** あるいは **Z スコア**

偏差値

- 別の基準での標準化
  - 教育学や心理学では, 平均 50, 標準偏差 10 が好まれる
- 標本平均 50, 標準偏差 10 に線形変換:

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

- T_i: **偏差値得点** あるいは **T スコア**

標準化の計算

- 基本書式

```
scale(x, center = TRUE, scale = TRUE) # 標準化
10 * scale(x) + 50 # 偏差値得点に変換
```

- 関数の引数
  - x: ベクトル，データフレームなど
  - center: 中心化 (平均 0) するか否か
  - scale: 正規化 (分散 1) するか否か

演習

練習問題

- 東京都の気候データ (tokyo_weather.csv) の中の気温, 日射量, 風速の項目について以下の問に答えよ．

```
myData <- read.csv("data/tokyo_weather.csv", fileEncoding="utf8")
```

  - 全てのデータを用いて各項目の平均・分散・標準偏差を求めよ．(データ数 365)
  - 毎月 5 日のデータのみを用いて各項目の平均・分散・標準偏差を求めよ．(データ数 12)
  - 5 の付く日 (各月の 5,15,25) のデータを用いて各項目の平均・分散・標準偏差を求めよ．(データ数 36)
  - ランダムに選んだ 36 日分のデータで各項目の平均・分散・標準偏差を求めたとき，推定量のばらつきを確認せよ．

歪度と尖度

歪度と尖度

- 正規分布からのずれを調べるための統計量
- 正規分布の特徴:
  - 確率分布のうち最も基本的なもの **(中心極限定理)**
  - 平均と分散を決めると完全に決定される
- 正規分布に従うデータでは標本平均と標本分散 (不偏分散) を考えれば十分
- 現実には正規分布では捉えきれない特徴をもつデータも多い

歪度

- 分布の非対称性を表す統計量
- **歪度** (skewness): 平均 μ , 分散 σ^2 で 3 次モーメントをもつ確率変数 X に対して以下で定義

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

- 左右に対称的な分布の歪度は 0 (正規分布の歪度は 0)
  - 歪度が正の場合: 分布の右の裾の方が重い
  - 歪度が負の場合: 分布の左の裾の方が重い
- 正の歪度をもつ分布の例:
  - ガンマ分布 $\Gamma(\nu, \alpha)$ の歪度は $2/\sqrt{\nu}$

尖度

- 平均の周囲の分布の尖り具合を表す統計量
- **尖度** (kurtosis): 4 次のモーメントをもつ確率変数 X

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

- **超過尖度** (excess kurtosis): 正規分布との比較のため尖度から正規分布の尖度 3 を引いた量

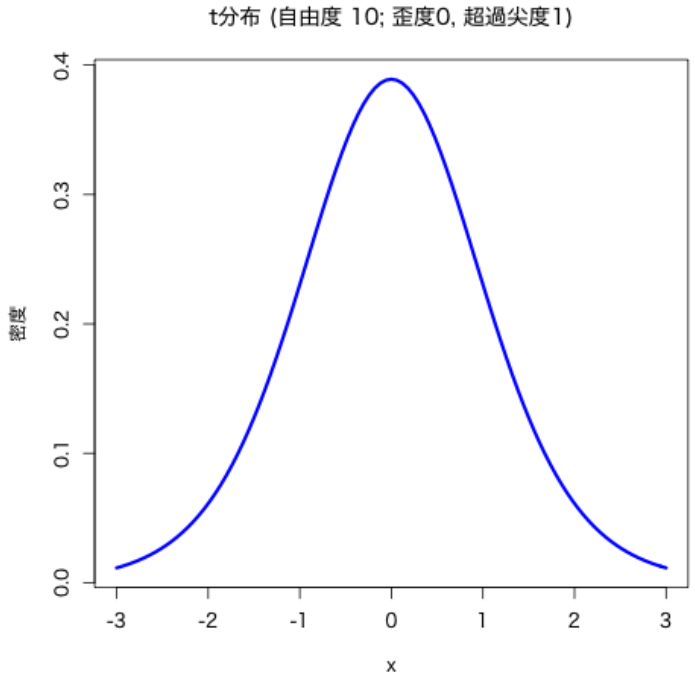
$$\text{excess kurtosis} = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4} - 3$$

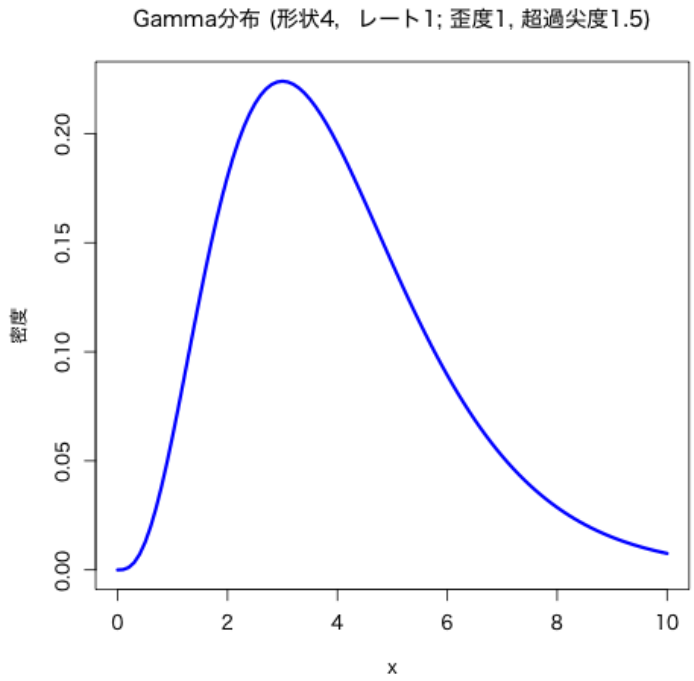
(こちらを単に尖度と呼ぶ文献もあるので注意)

超過尖度

- 正規分布と比較して
  - 超過尖度が正の場合: 平均の周囲の分布の形状が尖っている
  - 超過尖度が負の場合: 分布の形状は丸みを帯びている

    正の場合，正規分布に比べて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の例:
  - 自由度 $\nu > 4$ をもつ t 分布 $t(\nu)$ の超過尖度は $6/(\nu-4)$
    ($\nu \leq 4$ のときは $t(\nu)$ は 4 次モーメントをもたない)
  - ガンマ分布 $\Gamma(\nu, \alpha)$ の超過尖度は $6/\nu$





標本歪度と標本尖度

- 観測データ $X_1, X_2, \ldots, X_n$ からの歪度と尖度の推定
- **標本歪度** (sample skewness):

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

- **標本尖度** (sample kurtosis):

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

- 分子の計算は $1/n$ としているが，別の場合もあるので注意

歪度と尖度の計算

- 歪度・尖度を計算する関数は R の標準機能にはないので `package::e1071` を利用 (自作してもよい)

  標本歪度・標本尖度の値は標本平均・分散に比べて**ばらつきが大きい** ので，サンプル数が少ない場合の計算結果の解釈には注意が必要

- 基本書式

```
skewness(x, na.rm = FALSE, type = 3) # 標本歪度
kurtosis(x, na.rm = FALSE, type = 3) # 標本超過尖度 (尖度ではない)
```

- 関数の引数
  - `x`: ベクトル，データフレームなど
  - `na.rm`: 欠損値を取り除くか否か
  - `type`: 計算法の指定 (通常は既定値でよい)

演習

練習問題

- 東京都の気候データ (`tokyo_weather.csv`) の中の気温, 日射量, 風速の項目について以下の問に答えよ.
  - 全てのデータを用いて各項目の歪度と超過尖度を求めよ. (データ数 365)
  - 5 のつく日のデータのみを用いて各項目の歪度と超過尖度を求めよ. (データ数 36)
  - それぞれの値から正規分布から逸脱していると思われる項目はいずれか考察せよ.
  - 各データのヒストグラムを描き，データから計算される平均と分散を持つ正規分布と比較せよ.

相関と共分散

共分散

- 複数のデータ間の関係を知るための記述統計量
- **共分散** (covariance):

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

- **標本共分散** (sample covariance):
  $X_1, X_2, \ldots, X_n$ および $Y_1, Y_2, \ldots, Y_n$ に対して

$$\mathrm{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

相関

- 複数のデータ間の正規化した記述統計量
- **相関** (correlation):

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

- **標本相関** (sample correlation):
  $X_1, X_2, \ldots, X_n$ および $Y_1, Y_2, \ldots, Y_n$ に対して

$$\rho = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- 直感的には 2 種類のデータ間の比例関係の大きさ
- 相関の値は -1 以上 1 以下
  - 1 に近いほど正の比例関係が強い
  - -1 に近いほど負の比例関係が強い

相関と共分散の計算

- 基本書式

```
cov(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman")) # 共分散
cor(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman")) # 相関
```

- 関数の引数
  - `x,y`: ベクトル，データフレームなど (データフレームの時は列間の共分散行列，相関行列を計算)
  - `use`: 欠損値などの扱いに関する指定
  - `method`: 計算法の指定 (通常は既定値 `pearson` でよい)

演習

練習問題

- 東京都の気候データ (`tokyo_weather.csv`) の中の気温, 降水量, 日射量, 降雪量, 風速, 気圧, 湿度 (数値データ) の項目について以下の問に答えよ．
  - それぞれの項目間の共分散，および相関を求めよ．
  - 相関の高い項目の組 (絶対値が大きい)，および相関の低い項目の組 (0 に近い) を求めよ．
  - その項目同士の散布図を描け．

“順序” に基づく統計量

中央値

- データの順序にもとづく記述統計量
- **中央値** もしくは **メディアン** (median):

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

  データを昇順に並べ替えたとき中央にくる値
  - n が奇数の場合: $X_{((n+1)/2)}$
  - n が偶数の場合: $(X_{(n/2)} + X_{(n/2+1)})/2$
- 中央値は平均と同様にデータを代表する値
- データ中の **外れ値** (異常な値) の影響を受けにくい

分位点

- 100α % **分位点** (percentile/quantile):
  メディアンの一般化
- $\alpha \in [0, 1]$ 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
  - **第 1 四分位点**: 25%分位点
  - **第 2 四分位点**: 50%分位点 (中央値と等価)
  - **第 3 四分位点**: 75%分位点

中央値・分位点の計算

- 基本書式

```
median(x, na.rm = FALSE, ...) # 中央値
quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE,
 names = TRUE, type = 7, ...) # 分位点
summary(x) # 最大，最小，四分位点，平均を計算する
```

- 関数の引数
  - `x`: ベクトル

- `na.rm`: 欠損値を取り除くか否か
- `probs`: 計算する分位点の値
- `names`: 出力に関する指定，多数の分位点を計算する場合は `FALSE` とした方がよい
- `type`: 計算法の指定 (`help(quantile)` を参照)

連続分布の分位点

- 分位点は推定や検定において重要な役割を果たす
- 連続分布の 100α **%分位点** :

  $0 < \alpha < 1$ に対して，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.
  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.

- 分位点の性質

  $X_1, X_2, \ldots, X_n$ が独立同分布な確率変数の列のとき, $X_1, X_2, \ldots, X_n$ の 100α %分位点は, $n \to \infty$ のとき $X_1, X_2, \ldots, X_n$ の従う分布の 100α %分位点の **一致推定量** となる.

分布の分位点の計算

- 基本書式

```
# 正規分布の例
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
# xxx 分布の場合は以下の形式
qxxx(p, "分布の特性を決める option の指定")
```

- 関数の引数
  - `p`: 分位点 (100p%)
  - `mean`, `sd`: 正規分布の特性を決める option
  - `lower.tail`: `TRUE` なら $P(X \leq x)$ を計算，`FALSE` なら逆
  - `log.p`: 出力を対数とするか否か (値が小さい場合に利用)

ばらつきの指標

- 分位点を利用したデータのばらつきの指標
- **範囲** (range):
  最大値と最小値の差 (外れ値の影響を大きく受ける)
- **四分位範囲** (interquantile range):
  第 3 四分位点と第 1 四分位点の差
- **中央絶対偏差** (median absolute deviation):
  $X_1, X_2, \ldots, X_n$ の中央値を m としたとき, $|X_1 - m|, |X_2 - m|, \ldots, |X_n - m|$ の中央値

“頻度” に基づく統計量

頻度に基づく統計量

- **最頻値** もしくは **モード** (mode): データの中で最も頻度が高く現れる値
- データが有限個の値を取る場合に特に有効
- 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
  - 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
  - ただし，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

演習

練習問題

- 東京都の気候データ (`tokyo_weather.csv`) の中の気温 (数値データ) と最多風向 (ラベルデータ) を用いて以下の問に答えよ.
  - 全てのデータを用いて気温の四分位点を求めよ．(データ数 365)
  - 5 の付く日 (各月の 5,15,25) の気温の四分位点を求めよ．(データ数 36)
  - ランダムに選んだ 36 日分のデータで気温の四分位点がどのくらいばらつくか確認せよ.
  - 風向の最頻値を求めなさい.