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 8 回

小池祐太

2017 年 11 月 22 日

1. 確率分布
   - 離散分布
     - 二項分布
     - Poisson 分布
     - 幾何分布
   - 連続分布
     - 一様分布
     - 正規分布
     - ガンマ分布
     - t 分布
     - F 分布

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

離散分布

- 取りうる値が有限個, もしくは可算無限個 (例えば整数値のみとる場合) であるような確率変数は**離散型**であるといい, 対応する確率分布を**離散分布**と呼ぶ
- 離散分布は, その分布に従う確率変数 X が取りうる値 x のそれぞれに対して, $X = x$ となる確率 $P(X = x)$ を対応させる関数 $f(x) = P(X = x)$ を考えることで完全に決定される
- この関数 f を**確率質量関数**, あるいは単に**確率関数**と呼ぶ
- 平均や分散が, 取りうる値が有限個の確率変数の場合の一般化として定義される (前回のスライド参照)

二項分布

- n を正の整数, p を 0 以上 1 以下の実数とする
- 取りうる値が $0, 1, \ldots, n$ であり, 確率関数が

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n$$

  で与えられる離散分布を, 試行回数 n, 成功確率 p の**二項分布**と呼ぶ
- 平均は np, 分散は $np(1-p)$ で与えられる

実際, X を試行回数 n, 成功確率 p の二項分布に従う確率変数とすると,

$$
\begin{aligned}
E[X] &= \sum_{x=0}^{n} x f(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n-1} \frac{n!}{x!(n-x-1)!} p^{x+1} (1-p)^{n-x-1} \\
&= np \sum_{x=0}^{n-1} \binom{n-1}{x} p^x (1-p)^{n-1-x} \\
&= np
\end{aligned}
$$

であり, また,

$$
\begin{aligned}
E[X^2] &= \sum_{x=0}^{n} x^2 f(x) = \sum_{x=0}^{n} x(x-1)\binom{n}{x} p^x (1-p)^{n-x} + E[X] \\
&= \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!} p^x (1-p)^{n-x} + np \\
&= \sum_{x=0}^{n-2} \frac{n!}{x!(n-x-2)!} p^{x+2} (1-p)^{n-x-2} + np \\
&= n(n-1)p^2 \sum_{x=0}^{n-2} \binom{n-2}{x} p^x (1-p)^{n-2-x} + np \\
&= n(n-1)p^2 + np = (np)^2 + np(1-p)
\end{aligned}
$$

となるから,

$$
\mathrm{Var}[X] = E[X^2] - (E[X])^2 = (np)^2 + np(1-p) - (np)^2 = np(1-p)
$$

二項分布

- 特に, 試行回数 1 の二項分布を **Bernoulli 分布**と呼ぶ
- 例えば, 表が出る確率が p のコインを n 回投げたときに表が出る回数は試行回数 n, 成功確率 p の二項分布に従う
- 前回も述べたように, 二項分布に従う乱数の発生には関数 `rbinom()` を用いる
- なお, 原則として, ある確率分布に従う乱数を生成するための R の関数の命名規則は,「`r` + その乱数が従う分布の名前の省略形」となっている (離散一様分布など一部例外がある)
- また, 離散分布の場合, その確率関数を計算するための関数が, 同じ省略形の文頭に `d` をつけることで得られる
  - 例えば, 二項分布の確率関数は関数 `dbinom()` で計算できる
- 実行例 `rbinom2.r`

二項分布

演習 1

$X_1, \ldots, X_n$ を成功確率 p の Bernoulli 分布に従う独立同分布な確率変数列とすると, $\sum_{i=1}^{n} X_i$ の分布は試行回数 n, 成功確率 p の二項分布に従うことが知られている. このことを上の実行例を参考にしながらグラフを描画して確認せよ.

Poisson 分布

- λ を正の実数とする
- 取りうる値が 0 以上の整数であり, 確率関数が

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, \ldots$$

で与えられる離散分布をパラメーター λ の **Poisson 分布**と呼び, 記号 $P_o(\lambda)$ で表す
  - λ は**強度**と呼ばれることがある
- 平均, 分散はともに λ で与えられる

実際, X をパラメーター λ の Poisson 分布に従う確率変数とすると,

$$E[X] = \sum_{x=0}^{\infty} xf(x) = \sum_{x=0}^{\infty} x\frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x}{(x-1)!}$$
$$= \lambda e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = \lambda$$

であり, また,

$$E[X^2] = \sum_{x=0}^{\infty} x^2 f(x) = \sum_{x=0}^{\infty} x(x-1)\frac{\lambda^x}{x!}e^{-\lambda} + E[X]$$
$$= e^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^x}{(x-2)!} + \lambda = \lambda^2 e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} + \lambda = \lambda^2 + \lambda$$

となるから,

$$\mathrm{Var}[X] = E[X^2] - (E[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Poisson 分布

- 放射性物質から一定時間に放射される粒子の数や, 一定期間に起こる交通事故の数などは Poisson 分布に従うことが知られている
- また, 前回観察したように, 発生確率が低い事象が十分長い期間のあいだに起こる回数の分布は Poisson 分布で近似できる (少数の法則)
- Poisson 分布に従う乱数の発生には関数 `rpois()` を用いる
- 実行例 `rpois2.r`

Poisson 分布

演習 2

X, Y を独立な 2 つの確率変数とし, それぞれパラメーター λ_1, λ_2 の Poisson 分布に従うとする. このとき, 和 $X + Y$ の分布はパラメーター $\lambda_1 + \lambda_2$ の Poisson 分布に従うことが知られている. このことを上の実行例を参考にしながらグラフを描画して確認せよ.

幾何分布

- $0 < p \leq 1$ とする
- 取りうる値が 0 以上の整数であり, 確率関数が

$$f(x) = p(1-p)^x, \qquad x = 0, 1, \ldots$$

  で与えられる離散分布を成功確率 p の**幾何分布**と呼ぶ
- 平均は $(1-p)/p$, 分散は $(1-p)/p^2$ で与えられる

実際, X を成功確率 p の幾何分布に従う確率変数とすると,

$$
\begin{aligned}
E[X] &= \sum_{x=0}^{\infty} x f(x) = \sum_{x=0}^{\infty} x p (1-p)^x = p(1-p) \sum_{x=1}^{\infty} x (1-p)^{x-1} \\
&= -p(1-p) \frac{d}{dp} \sum_{x=1}^{\infty} (1-p)^x = -p(1-p) \frac{d}{dp} \frac{1-p}{p} \\
&= -p(1-p) \cdot \left(-\frac{1}{p^2} \right) = \frac{1-p}{p}
\end{aligned}
$$

であり，また，

$$
\begin{aligned}
E[X^2] &= \sum_{x=0}^{\infty} x^2 f(x) = \sum_{x=0}^{\infty} x(x-1)p(1-p)^x + E[X] \\
&= p(1-p)^2 \sum_{x=2}^{\infty} x(x-1)p(1-p)^{x-2} + \frac{1-p}{p} \\
&= p(1-p)^2 \frac{d^2}{dp^2} \sum_{x=2}^{\infty} (1-p)^x + \frac{1-p}{p} \\
&= p(1-p)^2 \frac{d^2}{dp^2} \frac{(1-p)^2}{p} + \frac{1-p}{p} \\
&= p(1-p)^2 \frac{2}{p^3} + \frac{1-p}{p} = 2\left(\frac{1-p}{p}\right)^2 + \frac{1-p}{p}
\end{aligned}
$$

となるから，

$$
\mathrm{Var}[X] = E[X^2] - (E[X])^2 = \left(\frac{1-p}{p}\right)^2 + \frac{1-p}{p} = \frac{1-p}{p^2}
$$

幾何分布

- 表が出る確率が p のコインを投げ続けて, 初めて表が出るまでに出た裏の回数は, 成功確率 p の幾何分布に従う
- 幾何分布に従う乱数の発生には関数 `rgeom()` を用いる
- 実行例 `rgeom.r`

連続分布

- 実際のデータでは, 取りうる値が任意の実数またはある範囲の実数である場合, もしくは取りうる値のパターンが数多いため近似的にすべての実数値またはある範囲の実数値をとりうると考えられる場合が頻繁にある
  - 具体例: 株価, 気温, 風速など
- このようなデータのモデル化には, しばしば連続分布に従う確率変数が用いられる

連続分布

- 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

連続分布

- 離散分布の場合と同様に, 連続分布に対しても平均, 分散, 標準偏差の概念が定義される
- X を連続型の確率変数, f を X の分布の確率密度関数とする
- 積分 $\int_{-\infty}^{\infty} xf(x)dx$ が絶対収束するとき, X の**平均**を

$$E[X] := \int_{-\infty}^{\infty} xf(x)dx$$

で定義する

  - 平均は**期待値**とも呼ばれる
  - 積分 $\int_{-\infty}^{\infty} xf(x)dx$ が絶対収束しないとき, X は平均をもたない

連続分布

- より一般に, X の関数 $\varphi(X)$ に対して, 積分 $\int_{-\infty}^{\infty} \varphi(x)f(x)dx$ が絶対収束するとき, $\varphi(X)$ の期待値を

$$E[\varphi(X)] := \int_{-\infty}^{\infty} \varphi(x)f(x)dx$$

で定義する

- 特に, 正の整数 p に対して

$$E[X^p] = \int_{-\infty}^{\infty} x^p f(x)dx$$

であり, これを p 次の**モーメント**あるいは**積率**と呼ぶ
  - 積分 $\int_{-\infty}^{\infty} x^p f(x)dx$ が絶対収束しないとき, X は p 次のモーメントをもたない
  - 離散型の確率変数の場合と同様に, 一般に, ある正整数 p に対して X が p 次のモーメントをもてば, $q \leq p$ なるすべての正整数 q に対して X は q 次のモーメントをもつことが知られている

連続分布

- X が 2 次のモーメントをもつとき, X の**分散**を

$$\mathrm{Var}[X] := E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx$$

  で定義する
- 分散の平方根 $\sqrt{\mathrm{Var}[X]}$ を**標準偏差**と呼ぶ
- 離散型の場合と同様に, 次の恒等式が成り立つ:

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

連続分布

- 連続分布の平均, モーメント, 分散, 標準偏差は, その分布に従う確率変数の平均, モーメント, 分散, 標準偏差で定義する
- 定義より明らかなように, 連続型の確率変数の平均, モーメント, 分散, 標準偏差もその分布のみに依存して定まるため, この定義は確率変数の選び方によらない
- 離散分布の場合と同様に, むしろ確率変数の平均, モーメント, 分散, 標準偏差はその確率変数が従う分布のものとみなす方が本質的である

連続分布

- 離散型の確率変数の場合と同様に, 大数の法則, 中心極限定理および重複対数の法則は, 確率変数たちが 2 次のモーメントをもつ限り, 連続型の確率変数の列についても成り立つ
  - 離散型の確率変数列の場合と同様に, 2 次のモーメントをもたない場合, 中心極限定理と重複対数の法則は成立しない (そもそも分散が定義できない)
  - 大数の強法則は平均が存在すれば成立する
- 以下に代表的な連続分布を列挙する

一様分布

- $a < b$ とする
- 確率密度関数が

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \text{ のとき}, \\ 0 & \text{上記以外のとき} \end{cases}$$

で与えられる連続分布を区間 (a, b) 上の**一様分布**と呼び, 記号 $U(a, b)$ で表す

- 平均は $(a + b)/2$, 分散は $(b - a)^2/12$ で与えられる

実際, X を区間 (a, b) 上の一様分布に従う確率変数とすると,

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{b-a}\int_{a}^{b} xdx = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$$

であり, また,

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \frac{1}{b-a}\int_{a}^{b} x^2 dx = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}$$

となるから,

$$\begin{aligned}\mathrm{Var}[X] = E[X^2] - (E[X])^2 &= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\ &= \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}\end{aligned}$$

一様分布

- 前述のように, 一様分布に従う乱数の発生には関数 `runif()` を用いる
- 連続分布の場合, 分布の省略形の文頭に `d` をつけることで, 確率密度関数を計算するための関数が得られる
- 例えば, 一様分布の確率密度関数は関数 `dunif()` で計算できる
- 実行例 `runif2.r`

正規分布

- μ を実数, σ を正の実数とする
- 確率密度関数が

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

で与えられる連続分布を平均 μ, 分散 σ^2 の**正規分布**または **Gauss 分布**と呼び, 記号 $N(\mu, \sigma^2)$ で表す

- 言葉通り, 平均は μ, 分散は σ^2 で与えられる
- 特に, 平均 0, 分散 1 の正規分布を**標準正規分布**と呼ぶ

正規分布

- 物理実験等の観測誤差の分布はしばしば正規分布でモデル化される
- また, 前回観察したように, 真の平均を標本平均で推定した際の推定誤差の確率分布は, サンプル数が大きくなるに従って正規分布に近づいていく (中心極限定理)
- 実行例 `rnorm2.r`

二項分布

演習 3

U_1, U_2 を 2 つの独立な確率変数とし, ともに $(0,1)$ 上の一様分布に従うとする. このとき,

$$\begin{cases} X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2), \\ X_1 = \sqrt{-2\log(U_1)}\sin(2\pi U_2) \end{cases}$$

とおくと, X_1, X_2 は独立かつともに標準正規分布に従うことが知られている (この変換を Box-Müller 変換と呼ぶ). このことを上の実行例を参考にしながらグラフを描画して確認せよ.

正規分布

- Y を試行回数 n, 成功確率 p の二項分布に従う確率変数とすると, n が十分大きいとき, $(Y - np)/\sqrt{np(1-p)}$ の分布は標準正規分布で近似できる
- これは **de Moivre-Laplace の定理**として知られているが, 中心極限定理の特別な場合である
- 実際, $X_1, \ldots, X_n$ を成功確率 p の Bernoulli 分布に従う独立同分布な確率変数列とすると, $\sum_{i=1}^n X_i$ は試行回数 n, 成功確率 p の二項分布に従う
- Bernoulli 分布は平均 p, 分散 $p(1-p)$ であったから,

$$\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} = \frac{\sqrt{n}(\frac{1}{n}\sum_{i=1}^n X_i - p)}{\sqrt{p(1-p)}}$$

  の分布は中心極限定理によって標準正規分布で近似できる
- 実行例 `rbinom-normal2.r`

正規分布

- ただし, 上において, p が非常に小さい場合, 特に np がそれほど大きくならない程度に p が小さい場合は, $(Y - np)/\sqrt{np(1-p)}$ の分布の正規近似よりも, Y の分布のパラメーター np の Poisson 分布による近似の方が精度がよい (後者は前章で述べた少数の法則の特殊な場合である)
- 実行例 `rbinom-poisson.r`

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
- 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
- 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
- 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
- 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
- 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`