データサイエンスのための統計学

代表的な確率分布

by 尚晋　(名古屋大学経済学研究科助教)
on 2025 年 6 月 03 日

目次

* おもな離散型の確率分布
* おもな連続型の確率分布

目次

* おもな離散型の確率分布
* おもな連続型の確率分布

» 二項分布

ベルヌーイ試行　結果が 2 通り（成功 p ／失敗 $q(=1-p)$）しかない試行をベルヌーイ試行という。

二項分布　同じ条件且つ独立にベルヌーイ試行を n 回行って x 回成功する確率は

$$f(x) = {}_nC_x p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n \tag{1}$$

で与えられる。ただし、${}_nC_x = \frac{n!}{x!(n-x)!}$ である。確率関数が上の $f(x)$ で与えられる確率分布を2 項分布 (binomial distribution) といい，X が 2 項分布に従う事を $X \sim Bi(n, p)$ と表す。

- 二項分布の期待値と分散は：

$$E(X) = np, \quad V(X) = np(1-p) \tag{2}$$

» 二項分布の性質

* $f(x) \geq 0$ は明らか。また、2 項定理により、$\Sigma f(x) = 1$ を満たす。
  (二項定理：$\sum_{x=0}^{n} {}_nC_x p^x q^{n-x} = (p+q)^n = 1$)
* X の平均は定義に基づいて計算すると、

$$E(X) = \sum_{x=0}^{n} x f(x) = \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!(n-x)!} p \cdot p^{x-1} q^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$x-1 = y, n-1 = m$とおくと, $m-y = n-x$ であるから

$$E(X) = np \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^y q^{m-y} = np \cdot 1 = np$$

最後の行は、$f(y)$ は $Y \sim Bi(m,p)$ となる確率変数 Y の確率関数である事を用いた。

* 同じで X の分散は定義に基づいて計算できる

» 二項分布

例

ある池の魚には，それを獲るとき確率 0.2 でその尾部に赤い色の標識が付いている．いま魚を 5 匹獲ったとき，その中で印の付いた魚がそれぞれ $x = 0, 1, 2, 3, 4, 5$ 回出る確率は？
これは試行回数 $n = 5$、成功確率 $p = 0.2$ の二項分布 $X \sim Bi(5, 0.2)$：

$$P(X = 0) = f(0) = \frac{5!}{0!(5-0)!}(0.2)^0(0.8)^5 = 0.328$$

$$P(X = 1) = f(1) = \frac{5!}{1!(5-1)!}(0.2)^1(0.8)^4 = 0.410$$

$$P(X = 2) = 0.205, P(X = 3) = 0.051, P(X = 4) = 0.006, P(X = 5) = 0.0003$$

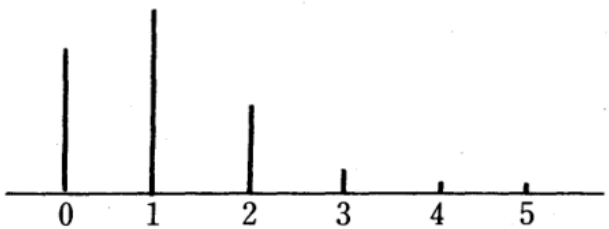


図 6.3　魚の匹数の分布
$n=5$, $p=1/5$ の二項分布であるが，5 匹中 1 匹の確率が最大という点に意味がある．

$E(X) = np$ であることが確認できる

» ポアソン分布

ポアソン分布　二項分布 $Bi(n, p)$ について、n が十分大きく、かつ p が非常に小さい場合、成功回数 X の分布は

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots \quad (\lambda > 0) \tag{3}$$

に収束。これを**ポアソン分布 (Poisson distribution)**と呼び、
$X \sim Po(\lambda)$ と表す。λ（ラムダ）はポアソン分布固有のパラメータ。
$e = 2.718\ldots$ は自然対数の底、定数。

- ポアソン分布の期待値と分散は：

$$E(X) = \lambda, \quad V(X) = \lambda \tag{4}$$

常に「期待値 = 分散 = λ」という、珍しい性質。

» ポアソン分布

例

ゴール数の例：サッカー J リーグ一部 2012 年第 1 節 ～ 第 3 節、のべ 54 チームのゴール数のデータを入手したとする。データ（サンプル数 54）の平均 $\bar{X} = 1.116$、分散 $s^2 = 1.001$、標準偏差 $s = 1.001$。
ポアソン分布の未知パラメータ λ を $\lambda = 1.116$ と置き、ゴール数実現値 $x = 0, 1, 2, \ldots$ の確率を計算する。この計算した結果とサンプルデータの相対度数と比較しみれば：

ポアソン確率（%）	32.76	36.56	20.40	7.59	2.12	0.47
データ相対度数（%）	29.63	35.19	25.93	7.41	1.85	0.00
x（ゴール数）	0	1	2	3	4	5

ポアソン分布での結果は実際のサンプルデータの分布とよく似ていることが分かる。

» 離散型一様分布

離散型一様分布 さいころを振ったときに出る目 X の確率分布が一様分布の例で,

$$f(x) = 1/N, \quad x = 1, 2, \ldots N \quad (N \text{は正整数}) \tag{5}$$

を, $1, 2, \ldots, N$ 上の離散型一様分布 (uniform distribution of discrete type)という。
一例として, さいころでは $N = 6$ である. この期待値は $E(X) = \frac{N+1}{2}$, 分散は $V(X) = \frac{N^2-1}{12}$。

目次

* おもな離散型の確率分布
* おもな連続型の確率分布

» 指数分布

指数分布 確率密度関数は

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & x < 0 \end{cases} \tag{6}$$

で与えられる確率分布を指数分布 **(exponential distribution)** という。ただし、$\lambda > 0$ である。$Exp(\lambda)$ と表す。

- 指数分布の期待値と分散は：

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \frac{1}{\lambda^2} \tag{7}$$

- 指数関数の累積分布関数：指数関数の確率密度関数式の定積分で、時点 t までにイベントが起こる確率 $P(X \leq t)$ を求めると

$$F(t) = P(X \leq t) = \int_0^t f(x)dx = 1 - e^{-\lambda t} \tag{8}$$

一方、t 以降にイベントが起こる確率 $S(t) = P(X > t) = 1 - F(t) = e^{-\lambda t}$ を、生存関数と呼ぶ。

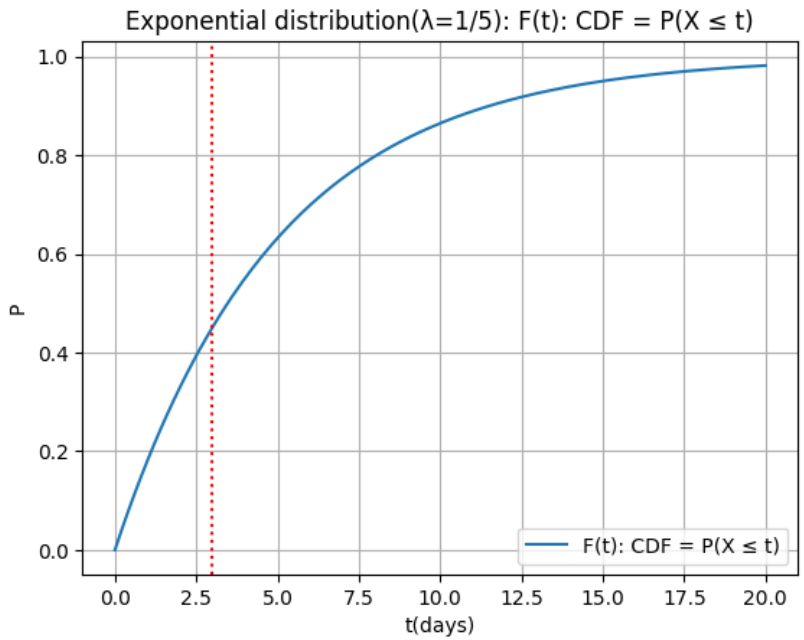
» 指数分布　例: 地震の待ち時間（累積分布関数）

ある地域では, 地震の平均発生間隔が **5** 日であると観測されている (つまり, 次の地震が起こるまでの平均待ち時間が **5** 日). このとき, 次の地震が **3** 日以内に起こる確率はいくらか?

- 平均待ち時間 $\Rightarrow \frac{1}{\lambda} = 5$ より $\lambda = \frac{1}{5}$. 指数分布 $X \sim Exp(1/5)$.
- 求めたいのは: $P(X \leq 3) = F(3) = 1 - e^{-\lambda \cdot 3}$

$$P(X \leq 3) = 1 - e^{-3/5} = 1 - e^{-0.6} \approx 1 - 0.5488 = 0.4512$$

結論: 次の地震が **3** 日以内に起こる確率は約 **45.1%** である.



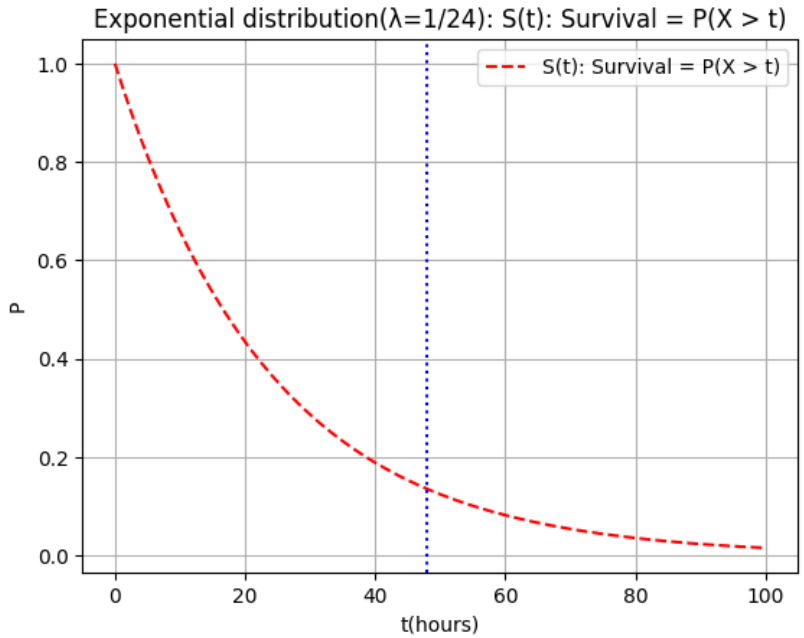
» 指数分布　例: 遭難者の生存確率 (生存関数)

ある地域の原始森林では、過去の救助記録から「遭難者が生存して発見されるまでの平均時間は 24 時間」とされている. このとき、48 時間以上発見されない場合に生存している確率はいくらか？

- 平均が 24 時間 $\Rightarrow \frac{1}{\lambda} = 24 \Rightarrow \lambda = \frac{1}{24}$
- 求めたいのは：$P(X > 48) = S(48) = e^{-\lambda \cdot 48}$ (生存関数)

$$S(48) = P(X > 48) = e^{-\lambda \cdot 48} = e^{-\frac{1}{24} \cdot 48} = e^{-2} \approx 0.1353$$

結論：**48** 時間以上発見されない場合に、まだ生存している確率は約 **13.5%** である。



» 正規分布

正規分布　正規分布 (normal distribution) の密度関数：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \quad (= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}) \tag{9}$$

で与えられる。確率変数 X がこの確率密度関数関数を持つ時，$X \sim N(\mu, \sigma^2)$ と表す。正規分布をガウス分布 (Gaussian distribution) ということもある。

- 正規分布の期待値と分散は：

$$E(X) = \mu, \quad V(X) = \sigma^2 \tag{10}$$

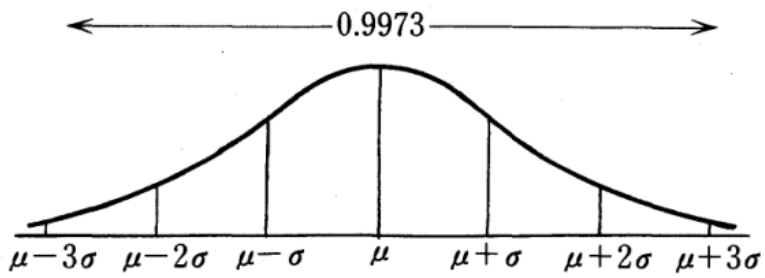


図 6.8　3 シグマ範囲
3 シグマ範囲の外へはずれる確率は千に三つ，いわゆる「千三つ」である．この言葉は，「きわめて成り立ちにくい」「稀にしか真実でない」の意に使われる（広辞苑）．

» 正規分布の著しい特徴

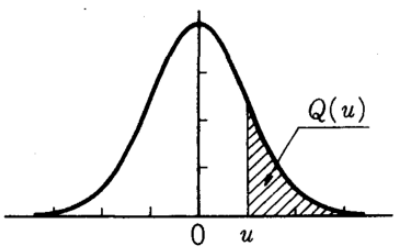
* X が正規分布 $N(\mu, \sigma^2)$ に従っているとき，その線形変換 $Y = aX + b$ は $N(a\mu + b, a^2\sigma^2)$ に従う。
* 標準化変数 $Z = (X - \mu)/\sigma$ は正規分布 $N(0, 1)$ に従う。これを標準正規分布 (standard normal distribution) という。いかなる正規分布の確率計算も標準正規分布に帰着する。
* 標準正規分布については，累積分布関数

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{11}$$

が，巻末に正規分布表 (上側確率) で確認できる。なお，
$\Phi(-z) = 1 - \Phi(z)$ の関係があるので，$z > 0$(上側確率) の表だけが与えられる。

付表 1　正規分布表(上側確率)

$Q(u)=1-\Phi(u)=\int_u^\infty \phi(u)du$



u	.00	.01	.02	.03
.0	.50000	.49601	.49202	.48803
.1	.46017	.45620	.45224	.44828
.2	.42074	.41683	.41294	.40905
.3	.38209	.37828	.37448	.37070
.4	.34458	.34090	.33724	.33360

» 正規分布の著しい特徴 (続き)

区間の確率が

$$P(-k \leq Z \leq k) = P(Z \leq k) - P(Z < -k) = \Phi(k) - \Phi(-k) \tag{12}$$

で計算できる。従って、主な区間の確率が $k = 1, 2, \ldots$ として下記で与えられる：

$$\begin{aligned}
P(-1 \leq Z \leq 1) &= \Phi(1) - \Phi(-1) = 0.6827 && \text{(1/3 の確率で } [-1, 1] \text{ の区間外に落ちる)} \\
P(-2 \leq Z \leq 2) &= \Phi(2) - \Phi(-2) = 0.9545 && \text{(1/20 の確率で } [-2, 2] \text{ の区間外に落ちる)} \\
P(-3 \leq Z \leq 3) &= \Phi(3) - \Phi(-3) = 0.9973 && \text{(3/1000 の確率で } [-3, 3] \text{ の区間外に落ちる)} \\
P(-4 \leq Z \leq 4) &= \Phi(4) - \Phi(-4) = 0.9999 && \text{(1/10000 の確率で } [-4, 4] \text{ の区間外に落ちる)}
\end{aligned} \tag{13}$$

なお, $-3 \leq Z \leq 3$ は, もとの X でいえば, $\mu - 3\sigma \leq X \leq \mu + 3\sigma$ に相当する. 常識的にいえばこれで事実上すべて (全体の確率 $= 1$) である.「事実上のすべて」の意味で, 区間 $[\mu - 3\sigma, \mu + 3\sigma]$ を, 3 シグマ範囲という (図 6.8)。

» 正規分布

例

例 1:　$X \sim N(1, 9)$ について $P(X \leq 2)$ を求める.

$$\frac{X-1}{3} \sim N(0,1) \text{より}, P(X \leq 2) = P(\frac{X-1}{3} \leq \frac{2-1}{3})$$
$$= \Phi(\frac{1}{3}) \approx 1 - 0.3707 = 0.6293$$

例 2:　$T \sim N(50, 10^2)$ について $P(50 \leq X \leq 51)$ を求める.

$$\frac{T-50}{10} \sim N(0,1) \text{より}, P(50 \leq T \leq 51) = P(0 \leq \frac{T-50}{10} \leq 0.1)$$
$$= \Phi(0.1) - \Phi(0)$$
$$= (1 - 0.46017) - 0.5 = 0.03983$$

例 3:　$Z \sim N(0, 1)$ について $P(|Z| \leq 1.96)$ を求める.

$$P(|Z| \leq 1.96) = P(-1.96 \leq Z \leq 1.96)$$
$$= P(-1.96 \leq Z < 0) + P(0 \leq Z \leq 1.96)$$
$$= P(0 < Z < 1.96) + P(0 < Z < 1.96)$$
$$= \Phi(1.96) - \Phi(0) + \Phi(1.96) - \Phi(0)$$
$$= [(1 - 0.024998) - 0.5] \times 2 \approx 0.95$$