健康統計学 第8回

今回は、確率分布（テキスト53〜63ページ）について学習します。確率分布を学習するために必要な、確率変数、確率密度関数についても説明します。

テキスト

- 『やさしい保健統計学 改訂第5版』縣 俊彦著 (南江堂)

今回の内容

1. 確率分布と確率密度関数
2. 確率変数の期待値と分散
3. 二項分布と関係する分布
4. 正規分布と関係する分布
    a. 補足：標準化を用いた確率の計算

確率分布と確率密度関数

確率変数と確率分布

確率変数 (random variable)

- 試行の結果、ある値をとる確率が決まる変数を、「**確率変数**」という
- サイコロを1回投げる場合を考える
  - サイコロの出た目の数 {1, 2, 3, 4, 5, 6} を X (確率変数) とする
    - 確率変数は大文字で書く
  - $X=1$ (つまり1の目がでる) の事象の確率は、次のように表すことができる

$$\mathrm{P}(X=1)=\frac{1}{6}$$

  - 同じように、1以外の目が出る確率は、次のように表せる

$$\mathrm{P}(X=2)=\cdots=\mathrm{P}(X=6)=\frac{1}{6}$$

  - なお、$X=1$ という事象は、$\{X=1\}$ とも表せる

確率分布 (probability distribution)

- サイコロを1回投げたときにでた目の数を確率変数 X を使うと、その確率は次のようになる

$$\mathrm{P}(X=1)=\cdots=\mathrm{P}(X=6)=\frac{1}{6}$$

- 確率変数 X のとる値と、それに対応する確率を表にまとめたもの、つまり確率変数 X に対応する確率の分布を、「**確率分布**」という

X	1	2	3	4	5	6	計
確率	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

- 一般に、確率変数 X が、次のような n 個の値をとるとき、

$$x_1,\ x_2,\ \cdots,\ x_n$$

- その確率が次のようになるのであれば、

$$\mathrm{P}(X=x_k)=p_k(k=1,2,\cdots,n)$$

- 次のことが成り立つ

$$\begin{cases} p_1\geq 0,\ p_2\geq 0,\ \cdots,\ p_n\geq 0 \\ p_1+p_2+\cdots+p_n=1 \end{cases}$$

X	x_1	x_2	$\cdots$	x_n	計
確率	p_1	p_2	$\cdots$	p_n	1

確率分布の例

- サイコロを1回投げたときにでた目の数が奇数か偶数かを考える
  - 奇数がでたときの確率変数を $Y=0$ 、偶数 (=奇数でない) がでたときの確率変数を $Y=1$ とする
  - 確率変数 Y の確率分布は、次のようになる

Y	0	1	計
確率	$\frac{3}{6}$	$\frac{3}{6}$	1

- コインを1回投げたときに表が出るか裏が出るかを考える
  - 表が出る回数を、確率変数 X で表すと、その確率分布は次のようになる

X	0	1	計
確率	$\frac{1}{2}$	$\frac{1}{2}$	1

確率密度関数

確率変数の種類

- これまで考えてきたのは、確率変数が離散的な（飛び飛びの値を取る）場合である
  - このような確率変数を、離散型確率変数（または離散変数）という
- 確率変数が連続的な値を場合もある（身長、体重、年齢など）
  - このような確率変数を、連続型確率変数（または連続変数）という

確率密度関数と累積分布関数

- 「確率変数 X のとる値が x 以下である」という事象とその確率を

$$\{X \leq x\}, \quad \mathrm{P}(X \leq x) = F(x)$$

と表し、関数 $F(x)$ を「**累積分布関数（または分布関数）**」という
  - つまり累積分布関数は、確率変数 x が最小値から指定された値までをとる間の確率（最小値に対応する確率から指定された値までに対応する確率をすべて足したもの）を与える
- 連続型の確率変数の場合は、その分布関数も連続的になる（グラフは曲線になる）
- 連続した確率変数 x がある区間 $a \leq x \leq b$ の値をとる確率は、関数の曲線と x 軸の囲む図形の面積になる
- このような、面積が確率を与えるような関数を「**確率密度関数**」という

$$\mathrm{P}(a \leq x \leq b) = \int_a^b f(x)dx$$

確率分布の分類

確率変数が離散的か連続的かで、次のように確率分布を分類することがでます。テキストで取り上げられている、おもな分布を挙げておきます。

- 離散型確率変数（コイン、サイコロ、トランプなど）
  - 二項分布（60ページ）
    - ポアソン分布（61ページ）
    - 幾何分布（63ページ）
  - 累積一様分布（59ページ）
- 連続型確率変数（大きさ、重さ、長さなど）
  - 正規分布（53ページ）
    - 標準正規分布（55ページ）

確率変数の期待値と分散

期待値（平均値）

期待値とは

- 確率変数 X の確率分布が次のようなとき、

X	x_1	x_2	$\cdots$	x_n	計
確率	p_1	p_2	$\cdots$	p_n	1

- 確率変数 X の平均値、または期待値は、次のように表せる

$$\mu = E(X) = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n$$

- 期待値とは、1回の試行の結果として期待される値の大きさを表す

期待値の計算例（1）

- サイコロを1回投げたときにでた目の数を確率変数 X を使うと、その確率分布は次のようになる

X	1	2	3	4	5	6	計
確率	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

- 確率変数 X の期待値（平均値）は、 $n=6$ なので、

$$x_1 = 1,\ x_2 = 2,\ \cdots,\ x_6 = 6$$

$$p_1 = \frac{1}{6},\ p_2 = \frac{1}{6},\ \cdots,\ p_6 = \frac{1}{6}$$

- したがって、次のようになる

$$E(X) = 1\times\frac{1}{6} + 2\times\frac{1}{6} + \cdots + 6\times\frac{1}{6}$$

$$= \frac{21}{6} = 3.5$$

- つまり、サイコロを何回も投げたときに、でた目の平均をとると 3.5 になることを示している

期待値の計算例（2）

- サイコロを5回連続で投げたときに1の目が出る回数を確率変数 X とすると、その確率分布は次のようになる

X	0	1	2	3	4	5
確率	0.4019	0.4019	0.1608	0.0322	0.0032	0.0001

- 確率変数 X の期待値（平均値）は、 $n=6$ なので、

$$x_1 = 0,\ x_2 = 1,\ \cdots,\ x_6 = 5$$

$$p_1 = 0.4019,\ p_2 = 0.4019,\ \cdots,\ p_6 = 0.0001$$

- したがって、次のようになる

$$E(X) = 0\times0.4019 + 1\times0.4019 + 2\times0.1608 + 3\times0.0322 + 4\times0.0032 + 5\times0.0001$$

$$\simeq 0.83$$

- つまり、サイコロを5回連続投げて1の目が出るのは1回あるかないか程度であることを示している

期待値の計算例（3）

- 宝くじの期待値を求めることもできる。宝くじの場合は「当せん金×当せん確率」の合計が期待値となる。
- 例えば、平成21年年末ジャンボ宝くじは、1ユニット（1000万枚）あたり、次のような当せん本数になっている。なお、当せん確率は「当せん本数÷1000万×100」から求めている。

等級	当せん金	当せん本数	当せん確率
1等	200,000,000円	1本	0.00001%
1等前後賞	50,000,000円	2本	0.00002%
1等組違い賞	100,000円	99本	0.00099%
2等	100,000,000円	2本	0.00002%
3等	5,000,000円	10本	0.0001%
4等	100,000円	600本	0.006%
5等	10,000円	10,000本	0.1%
6等	3,000円	100,000本	1%
7等	300円	1,000,000本	10%
元気に2010年賞	1,000,000円	100本	0.001%

- 宝くじがいくら当たるかの期待値を調べるには、「当せん金×当せん確率」の合計を求めるので、

$$\begin{aligned} E(X) &= 200{,}000{,}000\times 0.00001\%+50{,}000{,}000\times 0.00002\%+\cdots+300\times 10\%+1{,}000{,}000\times 0.001\% \\ &= 141.99 \end{aligned}$$

- つまり、宝くじ1枚（300円）を買うと、1枚につき141.99円の還元が期待できる、ということを示している。

期待値と算術平均との関係

- n 個のデータ $x_1, x_2, \cdots, x_n$ の平均値は、次のように表せる

$$\begin{aligned} \overline{x} &= \frac{x_1+x_2+\cdots+x_n}{n} \\ &= x_1\times\frac{1}{n}+x_2\times\frac{1}{n}+\cdots+x_n\times\frac{1}{n} \end{aligned}$$

- ここで確率について、$p_1=\frac{1}{n}, p_2=\frac{1}{n}, \cdots, p_n=\frac{1}{n}$ とおく、つまり各々の確率が等しいと考えると、

$$\begin{aligned} \overline{x} &= x_1p_1+x_2p_2+\cdots+x_np_n \\ &= E(X) \end{aligned}$$

- すなわち、各々の確率が等しくても等しくなくても、平均値（期待値）を求めることができる

分散と標準偏差

確率変数の分散と標準偏差

- 確率変数 X の確率分布が次のようなとき、

X	x_1	x_2	$\cdots$	x_n	計
確率	p_1	p_2	$\cdots$	p_n	1

- 確率変数 X の分散は次のように表す

$$\sigma^2 = V(X) = (x_1-\mu)^2 p_1 + (x_2-\mu)^2 p_2 + \cdots + (x_n-\mu)^2 p_n$$
$$= \sum_{i=1}^{n} (x_i-\mu) p_i$$

- μ は期待値 $E(X)$ 簡単に表したもの
- 分散の正の平方根を、確率変数 X の標準偏差といい、次のように表す

$$\sigma = \sqrt{V(X)}$$

確率変数の分散と標準偏差の特徴

- 分散や標準偏差が小さいほど、確率変数の値は平均に集中し、ばらつきが小さい
- 分散や標準偏差が大きいほど、確率変数の値は平均から離れ、ばらつきが大きい
- 分散は変数の単位の2乗を表す（例えば変数の単位がcmなら、分散の単位cm^2）ため、元の単位と同じ標準偏差を用いて平均からのばらつきを表す

確率変数の分散と標準偏差の計算

- サイコロを1回投げたときにでた目の数を確率変数 X を使うと、その確率分布は次のようになる

X	1	2	3	4	5	6	計
確率	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

- したがって、分散は次のようにして求められる

$$\sigma^2 = V(X) = (1-3.5)^2 \times \frac{1}{6} + (2-3.5)^2 \times \frac{1}{6} + \cdots + (6-3.5)^2 \times \frac{1}{6}$$
$$= \frac{35}{12} \simeq 2.92$$

- また、標準偏差は次のようになる

$$\sigma = \sqrt{V(X)}$$
$$= \sqrt{\frac{35}{12}} \simeq 1.71$$

- つまり、サイコロを何回も投げたとき、そのでた目の平均が 3.5 ± 1.71（1.79〜5.21）の範囲になる確率が高いことを示している

二項分布と関連する分布

二項分布 (binomial distribution)

二項分布を考える準備

- ある独立した試行を何回か連続して繰り返す場合を考える
- 例：1個のサイコロを5回連続して投げたときに、1回は1の目が出る確率について
  - 1の目がでたかどうかを確率変数 X とする
  - でた場合の事象を $\{X=1\}$ 、でない場合の事象を $\{X=0\}$ とすると、それらの確率は次のようになる

$$\mathrm{P}(X=1)=\frac{1}{6},\ \mathrm{P}(X=0)=\frac{5}{6}$$

  - また、5回サイコロを投げたとき1の目が出る組み合わせは、5通りになる

$${}_5C_1=\frac{5!}{1!\times 4!}=5$$

  - このうち1回だけ1の目が出る確率は、次のようになる

$$\left(\frac{1}{6}\right)^1\times\left(\frac{5}{6}\right)^4$$

  - したがって、サイコロを5回連続して投げたときに1回は1の目が出る確率は、次のとおりになる

$${}_5C_1\times\left(\frac{1}{6}\right)^1\times\left(\frac{5}{6}\right)^4\simeq 0.4019$$

- あらためて、サイコロを5回連続で投げて1の目がでる回数を、確率変数 X とすると、確率は次のようにして求められる

$$\mathrm{P}(X=0)={}_5C_0\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5\simeq 0.4019$$

$$\mathrm{P}(X=1)={}_5C_1\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4\simeq 0.4019$$

$$\mathrm{P}(X=2)={}_5C_2\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3\simeq 0.1608$$

$$\mathrm{P}(X=3)={}_5C_3\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2\simeq 0.0322$$

$$\mathrm{P}(X=4)={}_5C_4\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^1\simeq 0.0032$$

$$\mathrm{P}(X=5)={}_5C_5\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^0\simeq 0.0001$$

確率分布は次のようになる

X	0	1	2	3	4	5
確率	0.4019	0.4019	0.1608	0.0322	0.0032	0.0001

二項分布

- ある独立な試行で、事象 A が起こる確率を p 、起こらない確率を $q(=1-p)$ とする
- この試行を独立に n 回繰り返したときに、事象 A が起こる回数を確率変数 X としたとき、 $X=x$ （つまり x 回起こる）となる確率は次のようになる

$$
\begin{aligned}
\mathrm{P}(X=x) &= {}_nC_x p^x q^{n-x} \\
&= \frac{n!}{x!(n-x)!} p^x q^{n-x} \\
&= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}
\end{aligned}
$$

- このような確率分布を、「**二項分布**」といい、$B(n,p)$ と表す
  - 1回の試行で生じる事象が2種類しかない場合（不良品の発生など）の分布などに使われる

二項分布の特徴

二項分布 $B(n,p)$ には、次のような特徴がある。

- 確率が $p=0.5$ のとき、平均を中心とした左右対称な分布になる
  - p が小さいと、左側に寄った非対称な分布になる
  - p が大きいと、右側に寄った非対称な分布になる
- 回数 n が非常に大きくなると、左右対称な分布になる（ラプラスの定理）
- 平均は np 、分散は $npq=np(1-p)$ になる
- $np \geq 5$ で正規分布に近似できる

ポアソン分布 (poisson distribution)

ポアソン分布とは

- 二項分布の平均 np が一定の値 λ をとる場合を考えると、二項分布の式は次のように変形できる

$$
\mathrm{P}(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}
$$

  - e は自然対数の底で、$e=2.71828\cdots$
- このような分布を「**ポアソン分布**」といい、平均で分布が決まる
  - 極めてまれにしか起こらない現象を一定期間観測した場合の分析に使われる（交通事故の死亡者数、病気の死亡者数、電話の呼び出し数など）

$$
\mathrm{P}(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}
$$

  - x : 発生回数
  - λ : 平均の発生回数
  - つまり、二項分布において試行回数 n が極めて大きい場合（ $n \to \infty$ ）に用いる

ポアソン分布の特徴

- 平均 λ が大きくなると、左右対称な分布になる
- 平均も分散も、λ になる

正規分布と関連する分布

自然界や一般に観察できる多くのものについて、その分布は、平均値を中心に左右対称の**釣り鐘状**の分布になっていることがあります。

- 生物現象、毎年の雨量など
- 身長や体重、標準的なテストの成績など

正規分布 (normal distribution)

正規分布とは

- 平均値を中心とした左右対称の釣り鐘状になる分布を、「**正規分布**」という

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

  - π は円周率で、$\pi=3.14159\cdots$
  - e は自然対数の底で、$e=2.71828\cdots$
  - μ : 平均値
  - σ : 標準偏差
  - σ^2 : 分散
- 確率密度関数が上の式になるとき、『x は、平均が μ で標準偏差が σ の正規分布にしたがう』といい、$N(\mu,\sigma^2)$ と表す
  - つまり、平均値と分散 (標準偏差) から分布が決まる
- さまざまな推定や検定に使われる

正規分布の特徴

- 分布の中心は平均値 μ で、最も高い値 (極大値) をとる
  - 平均値が変化すると、分布が左右に移動する
- $\mu\pm\sigma$ で変曲点 (曲線の凹凸の変わり目) になる
  - 標準偏差が変化すると、分布の高さや広がりが変化する
- 平均値、中央値、最頻値は一致する
- 平均値と標準偏差から、分布の割合 (曲線とx軸に囲まれる面積) が決まる
  - $\mu\pm\sigma$ の範囲 : 全体の約 68.26% を含む
  - $\mu\pm2\sigma$ の範囲 : 全体の約 95.44% を含む
  - $\mu\pm3\sigma$ の範囲 : 全体の約 99.73% を含む
  - それ以外の範囲 : 全体の約 0.27% を含む

標準正規分布

標準正規分布とは

- 正規分布で、平均値が 0 、標準偏差が 1 になるように、正規分布の確率密度関数の変数 x を次のように変換する (変数変換)

$$z=\frac{x-\mu}{\sigma}$$

- すると、標準偏差の式は次のように変形される

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{\frac{z^2}{2}}$$

- 確率密度関数が上の式になるとき、『z は、標準正規分布にしたがう』という
  - 『z は、平均が 0 で標準偏差が 1 の正規分布にしたがう』ことになり、$N(0,1^2)$ と表すことができる
  - つまり、変数だけから分布が決まる
  - さまざまな推定や検定に使われる

標準化（基準化）(standardization)

- 先ほどの変数変換を、「**標準化**」または「基準化」という

$$z=\frac{x-\mu}{\sigma}$$

- (感覚としては) 分布を平均値の分だけ 0 まで移動し、分布の広がりを $\frac{1}{\sigma}$ にする
  - 単位の異なるデータや平均値･分散が異なるデータを比較するときに使う (英語と数学のテストの成績の比較)
- ちなみに**偏差値**は、標準化を応用したもので、次のような式になる

$$z=50+10\times\frac{x-\mu}{\sigma}$$

標準正規分布の特徴

- 基本的な特徴は、正規分布と同じ
- 分布の曲線とx軸で囲まれた全体の面積は1になり
  - すなわち、面積が確率 (割合) を表すことになる

カイ二乗分布

カイ二乗分布とは

- 変数 $z_1,z_2,\cdots,z_r$ が標準正規分布にしたがい、互いに独立であるとする
- 次のようになるとき、『z は、自由度 r のカイ二乗 (X^2) 分布にしたがう』という

$$\begin{aligned}X^2&=z_1{}^2+z_2{}^2+\cdots+z_r{}^2\\&=\sum_{i=1}^{r}z_i{}^2\\&=\sum_{i=1}^{r}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2\end{aligned}$$

  - 自由度とは分布の形状に影響を及ぼす値で、自由度の値が変わると分布の形状も変化する

カイ二乗分布の特徴

- 平均は r 、分散は $2r$ になる
- 適合度や独立性の検定など、分散の推定や検定に利用される
- 自由度が大きくなると、対称な分布に近づく

t分布

t分布とは

- 標準正規分布にしたがう確率変数 x と、自由度 r のカイ二乗分布 ${\chi_r}^2$ にしたがう確率変数 y があり、互いに独立であるとする
- 次のようになるとき、『x は、自由度 r のt分布にしたがう』という

$$t = \frac{x}{\sqrt{y/r}} = \frac{x}{\sqrt{\frac{{\chi_r}^2}{r}}}$$

- (感覚としては) 正規分布にしたがう統計量を標準化したものの分布を表す
- この分布を発表した William Gosset のペンネームから「スチューデント (Student) のt分布」とも呼ぶ

t分布の特徴

- 平均は 0 、分散は $\frac{r}{r-2}$ になる
- 自由度が大きくなると、標準正規分布に近づく
    - 自由度が小さいときは、正規分布よりも裾の長い分布になる
- 母平均の推定や検定、平均値の差の検定や検定など、多くの統計的推定で利用される

F分布

F分布とは

- 確率変数 x と確率変数 y が、それぞれ自由度 m と n のカイ二乗分布にしたがい、互いに独立であるとする
- 次のようになるとき、『x は、第一自由度が m で第二自由度が n のF分布にしたがう』という

$$F = \frac{x}{m} \Big/ \frac{y}{n} = \frac{{\chi_m}^2}{m} \Big/ \frac{{\chi_n}^2}{n}$$

- (感覚としては) 分散の比率についての分布を表す

F分布の特徴

- 2つの自由度を持つ
- 分散検定や分散分析 (分散比の分布を調べる) に利用される
- 平均は $\frac{n}{n-2}$ 、分散は $\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ になる

標準化を用いた確率の計算

標準得点（標準正規分布の復習）

平均が μ 、分散が σ^2 の**正規分布**から、**標準正規分布**を導くときに、次の式を用いて**標準化**を行います。

$$z = \frac{x-\mu}{\sigma}$$

このときの z を、**標準得点**（standardized score）といいます。

標準得点は、平均が0、分散が1の標準正規分布 $N(0,1)$ にしたがいます。

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}}$$

標準得点を使った確率の計算

正規分布とみなされるデータを標準化すれば、標準正規分布表を用いて、確率を計算することができます。

例題

> 高校3年生のAさんの身長は175cmである。Aさんが入学する、B大学の学生の身長について、平均は182cmで、標準偏差は8.3cmである。このとき、B大学の学生がAさんより身長が高い確率を求める。

確率の求め方

1. まず、B大学の学生の身長を x として、標準得点を計算する。

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{x-182}{8.3} \end{aligned}$$

2. 標準得点 z は標準正規分布にしたがうので、標準正規分布表を用いて、確率を求める。
「身長が175cmより大きい」ということは、標準得点が次のようになるということである。

$$\begin{aligned} z &> \frac{175-182}{8.3} \\ &> -0.843\cdots \\ &\simeq -0.84 \end{aligned}$$

3. したがって、「身長が175cmより大きい」確率 $P(z>-0.84)$ は、標準正規分布表から $z=-0.84$ の値を求めればよい。

$$\begin{aligned} P(x>175) &\simeq P(z>-0.84) \\ &= 1-P(z\leq 0.84) \\ &\simeq 1-0.2005 \\ &\simeq 0.80 \end{aligned}$$

つまり、「B大学の学生がAさんより身長が高い確率」は約80%（0.8）となる。