第 4 章 期待値と分散

浅川 伸一

2006 年 04 月 26 日

3.3 二項分布の補足

パスカルの三角形から

$$ {}_nC_k = {}_{n-1}C_{k-1} + {}_{n-1}C_k \tag{1} $$

その他には

$$ \sum_{r=0}^{n} {}_nC_r = {}_nC_0 + {}_nC_1 + \cdots + {}_nC_n = 2^n \tag{2} $$

3.4 大数の法則 law of large numbers

ある事象 event が起こる確率は、独立試行の回数 n を十分に大きくすると、その事象の起こる確率あるいは割合 (%) は真の確率にいくらでも近付く。これを大数の弱法則という。

コイントスの場合、表 H の出る回数の割合は、何回も試行を繰り返すと、H の回数の割合は（一回の）成功確率に近づく。

どんな小さな値 ϵ に対しても、表の出る割合 r/n と真の確率 p との絶対値の差が 0 に近付く。

表の出る割合が真の確率に「近付く確率」が 1 である、を大数の強法則と言って区別することがある。脱線

3.5 実習

二項分布 $B(n, p)$ のパラメータ parameter をさまざまに変化させてグラフを描け。Excel の BINOMDIST() 関数

4　期待値と分散

4.1　期待値

ある値のとる確率が決まっているとき、その変数を確率変数 stochastic variable, random variable, or probabilistic variable と言う。例えばコインの表を 1 とし、裏を 0 とする場合、サイコロの目、など。ある確率変数 X が $x_1,\ x_2,\ \cdots,\ x_n$ のいずれか一つの値をとり、各値の確率が $p_1,\ p_2,\ \cdots,\ p_n$, $(p_1 + p_2 + \cdots p_n = \sum_{i=1}^{n} p_i = 1)$ で与えられている場合、

$$E(X) = \sum_{i=1}^{n} x_i p_i \tag{3}$$

を確率変数 X の期待値、あるいは平均 mean, average という。

表に 2、裏に 3 と書いてあるコインを投げて出た目の数の 1 万倍の金額を手に入れることができるとしよう。すると一回コイン投げをすると幾らもらえると期待できるだろうか。コインの裏、表の出る確率を 1/2 だとすると 2 万円 $\times 1/2 + 3$ 万円 $\times 1/2$ だから 2 万 5 千円と考えることができる。この場合 2 万円と 3 万円のちょうど中間の値になっている。

同様にして、サイコロをころがして出た目の数の 1 万倍の金額を受け取れるとすれば、幾らもらえるかを考えれば $1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6 = 3.5$ であるから 3 万 5 千円である。

起こりうる場合がすべて等しく n とおりあるのならば、その確率変数の期待値は

$$E(X) = \sum_{i=1}^{n} x_i p_i = \sum_{i=1}^{n} x_i \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4}$$

平均のこと、あるいは算術平均とも言う。実際の実験から得られたデータの平均を $\overline{X}$ と書いて標本平均と呼ぶことがある。一方、真の平均を μ と書いて区別する。$\overline{X}$ と μ の差の絶対値 $|\overline{X} - \mu|$ がいくらでも小さい値 ϵ よりも小さくなることを大数の弱法則 (前出) という。

1. データ数 n と平均 $\overline{X}$ の積は、データの総和 $\sum$ に等しい

$$n\overline{X} = \sum X \tag{5}$$

2. 各データ x_i と平均 $\overline{X}$ との差の総和は 0

$$\begin{aligned} \sum \left(x_i - \overline{X}\right) &= \underbrace{x_1 - \overline{X} + x_2 - \overline{X} + \cdots + x_n - \overline{X}}_{X \text{ が } n \text{ 個ある}} \\ &= \sum x_i - n\overline{X} \\ &= 0 \end{aligned}$$

3. 期待値とは、各データと期待値との差の 2 乗和を最小にする値である。任意の値 c を使って

$$\sum (x_i - E(X))^2 < \sum (x_i - c)^2 \tag{6}$$

が成り立つ。右辺のカッコ内で $E(X)$ を引いて、足して、も答えは変わらない。それを展開すると

$$\begin{aligned}
&\sum (x_i - c)^2 \\
&\quad = \sum (x_i - E(X) + E(X) - c)^2 \\
&\quad = \sum \{(x_i - E(X)) + (E(X) - c)\}^2 \\
&\quad = \sum \left\{ (x_i - E(X))^2 + 2\,(x_i - E(X))\,(E(X) - c) + (E(X) - c)^2 \right\} \\
&\quad = \sum (x_i - E(X))^2 + \sum 2\,(x_i - E(X))\,(E(X) - c) + \sum (E(X) - c)^2 \\
&\quad = \sum (x_i - E(X))^2 + 2\,(E(X) - c) \sum (x_i - E(X)) + n\,(E(X) - c)^2
\end{aligned}$$

ここで $\sum (x_i - E(X)) = 0$ なので第 2 項が消える。よって

$$\sum (x_i - c)^2 = \sum (x_i - E(X))^2 + n\,(E(X) - c)^2 \tag{7}$$

上式はすべての項が 2 乗だから正になるので

$$\sum (x_i - c)^2 > \sum (x_i - E(X))^2 \tag{8}$$

が $E(X)$ 以外のいかなる c に対しても成り立つ。

5 分散

期待値の他に分布を表現する数値として分散 variance が挙げられる。分散とは、各データから期待値を引いた値を 2 乗したものの総和の期待値である。

$$V(X) = E\left((X - E(X))^2 \right) = \sum_{i=1}^{n} (x_i - E(X))^2\, p_i \tag{9}$$

分散とは、分布の散らばり方を表す指標である。期待値との差を 2 乗することで各データが期待値からどれほど離れているかを数値している。2 乗しないと $(1/n)\sum (x_i - E(X)) = 0$ となって意味を成さない。

5.1 用語の定義

期待値と同じく、分散も標本分散 s^2 と母分散 σ^2 とを別けて表記する。$\sum (x_i - E(X))^2$ を偏差平方和 sum of square という。分散とは偏差平方和をデータの数で割った値である。

別の表現方法を使うと、

$$\begin{aligned} V(X) &= E\left((X-E(X))^2\right) && (10)\\ &= E\left\{X^2-2X\,E(X)+(E(X))^2\right\} && (11)\\ &= E\left(X^2\right)-E\left\{2X\;E(X)\right\}+E\left\{(E\,(X))^2\right\} && (12)\\ &= E\left(X^2\right)-2E(X)E(X)+(E(X))^2 && (13)\\ &= E\left(X^2\right)-(E(X))^2 && (14)\end{aligned}$$

分散とは 2 乗の期待値から、期待値の 2 乗を引いたものである。標本平均と標本分散を用いれば

$$s^2=\frac{1}{n}\sum_{i=1}^{n}x^2-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2 \tag{15}$$

と書ける。ただし数学的に等しいだけで、数値計算の上では桁落ちに対する注意が必要。

分散の平方根 square root を標準偏差 standard deviation と呼ぶ

$$s=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2}=\sqrt{\frac{\left(x_1-\overline{X}\right)^2+\left(x_2-\overline{X}\right)^2+\cdots+\left(x_n-\overline{X}\right)^2}{n}} \tag{16}$$

6　二項分布の期待値と分散

コインを 5 回トスして、表 H の出る回数を x とした 2 項分布 $B(5,1/2)$ の期待値は

$$\begin{aligned} E(X) &= \sum_{x=0}^{5}x\cdot{}_5C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{5-x} && (17)\\ &= (0\cdot{}_5C_0+1\cdot{}_5C_1+2\cdot{}_5C_2+\cdots+5\cdot{}_5C_5)\cdot\left(\frac{1}{2}\right)^5 && (18)\\ &= (0\cdot1+1\cdot5+2\cdot10+3\cdot10+4\cdot5+5\cdot1)\cdot\left(\frac{1}{2^5}\right) && (19)\\ &= (0+5+20+30+20+5)\cdot\frac{1}{32} && (20)\\ &= 80/32 && (21)\\ &= 2.5 && (22)\end{aligned}$$

二項分布の期待値、分散の一般形を示す。「道具としての統計学」 52 ページは複雑、偏微分など未知の数学を用いてる。「推測統計学はじめの一歩」47 ページは論理に飛躍がある。そこで以下のように考える。

二項分布にしたがう確率変数 x の確率密度関数 $f(x)$ は定義により

$$f(x) = {}_nC_x p^x \,(q)^{n-x} \tag{23}$$

である。この期待値は

$$\begin{aligned} E(x) &= x_0 f(x_0) + x_1 f(x_1) + \cdots + x_n f(x_n) && (24)\\ &= \sum_{x=0}^{n} x_i f(x_i) && (25)\\ &= \sum_{x=0}^{n} x \left({}_nC_n p^x q^{n-x}\right) && (26) \end{aligned}$$

ここで、次のような二項式を展開して

$$\begin{aligned} (pt+q)^n &= {}_nC_0 \,(pt)^0 \, q^n + {}_nC_1 \,(pt)^1 \, q^{n-1} + {}_nC_2 \,(pt)^2 \, q^{n-2} + \cdots \\ &\quad + \cdots + {}_nC_x \,(pt)^x \, q^{n-x} + \cdots + {}_nC_n \,(pt)^n \, q^0 \end{aligned}$$

この式の両辺を t で微分すると、右辺第 1 項は定数であるから微分すると 0 になることに注意して

$$\begin{aligned} &np(pt+q)^{n-1} = \\ &\quad 1\left({}_nC_1 p^1 1 q^{n-1}\right) t^0 + 2\left({}_nC_2 p^2 q^{n-2}\right) t^1 + \cdots \\ &\quad \cdots + x\left({}_nC_x p^x q^{n-x}\right) t^{x-1} + \cdots + n\left({}_nC_n p^n q^0\right) t^{n-1} \end{aligned} \tag{27}$$

ここで $t=1$ とおくと左辺は np になる。右辺は $\sum x\,({}_nC_n p^x q^{n-x})$ の形をしているので $E(x)$ である。従って

$$E(X) = np \tag{28}$$

ちょっと手抜き。分散は

$$\begin{aligned} V(x) &= \sum_{i=0}^{n} (x_i - E(X))^2 f(x_i) = \sum_{i=0}^{n} x_i^2 f(x_i) - \{E(x)\}^2 \\ &= \sum_{i=0}^{x} x_i^2 \left({}_nC_x p^x q^{n-x}\right) - \{E(x)\}^2 \end{aligned}$$

ここで式 (27) の両辺に t をかけて

$$\begin{aligned} &npt(pt+q)^{n-1} \\ &\quad = 1\left({}_nC_1 p^1 1 q^{n-1}\right) t + 2\left({}_nC_2 p^2 q^{n-2}\right) t^2 \cdots \\ &\quad \cdots + x\left({}_nC_x p^x q^{n-x}\right) t^x + \cdots + n\left({}_nC_n p^n q^0\right) t^n \end{aligned} \tag{29}$$

この式の両辺を t で微分すると

$$\begin{aligned} &\cancel{np}(pt+q)^{n-1} + n(n-1)p^2 t(pt+q)^{n-2} \\ &\quad = 1^2\left({}_nC_1 p^1 q^{n-1}\right) t^0 + 2^2\left({}_nC_2 p^2 q^{n-2}\right) t^1 + \cdots \\ &\quad \cdots + x^2\left({}_nC_x p^x q^{n-x}\right) t^{x-1} + \cdots + n^2\left({}_nC_n p^n q^0\right) t^{n-1} \end{aligned} \tag{30}$$

となる。ここでまたもや $t=1$ とおけば、左辺は $np+n(n-1)p^2$ となる。右辺は $\sum x_i^2\,({}_nC_x p^x q^{n-x}) = E(x^2)$ の形になっている。このことと、分散の定義は 2 乗の期待値から期待値の 2 乗であることから

$$\begin{aligned} V(x) &= np+n(n-1)p^2-(E(x))^2 \\ &= np+n(n-1)p^2-n^2p^2 \\ &= np+n^2p^2-np^2-n^2p^2 \\ &= np-np^2 \\ &= np(1-p)=npq \end{aligned}$$