III-2. 確率分布

III-2-1. 確率の計算と二項分布モデル

頻度主義の立場に立つと、はじめに、正確に確率どおりに現れる理想的な現象がどんな形（確率分布）を描くのかを考えなければなりません。そのために、例として出されるのが、理想的なコインとか理想的なサイコロです（コインやサイコロなど賭け事が理想的だと言ったのではありません。念のために）。ｎ回コインを投げてｋ回表になる。あるいは、ｎ回サイコロを投げてｋ回１の目が出る。その確率をP(k)としたとき（ｋとなる確率 Probability という感じでこの記号にしました。）、0 からｎまでのｋの値についてP(k)を計算することができます。ｋを横軸にしてP(k)の値をとったのが確率分布です。ｎ回コインを投げて、何回表が出るか、もっともありそうな回数は何回かと訊かれたら、たいていの人が$\frac{n}{2}$回と答えます。これをｋの期待値と言います。この場合は、直観的に、平均値にも中央値にもなっていることがわかります。ｋの分布は０から ｎ まで広がっていますから、分布の中心でここが一番確率が高いというわけです。サイコロで１が出る回数の場合は期待値は$\frac{n}{6}$で、分布が左右相称ではないので、真ん中という感じはしないかもしれませんが、ここがもっとも確率が高く、平均値かつ中央値になっていて、分布の中心です。このような分布を二項分布と言います。表が出ることの排反事象（表でないこと）は裏が出ることで、１の目が出ることの排反事象（１でないこと）が１以外の目が出ることで二項対立的だからです。

実際にｎ回コインを投げるかサイコロを振るかして、x回、表ないし１の目がでたら、その値をｋの値を表す横軸で見つけて、その点での確率P(x)が、理想的な確率分布上でxとなる確率です。 何故、二項分布で考えるのかというと、それ以外にうまいモデルが思いつかないからです。ただそれだけのことです。

次に、回数という不連続な値の確率に関するモデルを拡張して、背の高さとか、光の強さとか、値段とか、連続的な値にこの考え方に使える確率モデルを理論的につくります。さらに、そのような確率モデルで説明できる複数の確率変数の差や比なども確率変数と考えられるので、その変数の確率的変動を説明する確率モデルも作ります。それらのモデルの実際の形（平均値・ばらつき方）の推定のしかたを考えて、２章で説明した頻度主義の考え方で実際の判断（検定）をします。

この流れに従って、まず、二項分布モデルを数式として作るための作業をします。「組み合わせ」の数とか、積集合、和集合、確率の演算を説明するために、集合の概念とその演算の知識が必要になるので、使われる演算子の記述の仕方を含めて、基礎的な説明をまとめてします。

集合とはいくつかの要素をまとめたものです。普通は{　}で表します。たとえば、1,3,4,6,7を要素としてもつ数のまとまりを一つの集合として表せば、{1.3.4.6,7}です。{犬}ならば、すべての犬を表します。隣の犬も、自分のうちの犬も、セントバーナードも、チワワも要素として含まれます。{動物}ならば、犬や猫も要素として含まれていて、それらもまた集合として表せます。また、その集合の要素は、すべて{動物}の要素の中に含まれます。そのよ

うなものを部分集合と言います。ある集合の中に部分集合があれば。$\{A\} \subset \{B\}$のように、$\subset$で表します。$\{A\}$は$\{B\}$に部分集合として含まれるということです。$\{犬\} \subset \{動物\}$という風に表せます。これを包含関係と言います。$\{A\} \subset \{B\} \subset \{C\}$ならば、$\{A\} \subset \{C\}$であることは直感的に明らかです。ある要素がある集合に含まれるとき、 $e \in A$と$\in$を使って表します。「e は集合$\{A\}$の要素だ。」あるいは「集合$\{A\}$は要素として e を含む。」ということです。$\{隣の犬\} \in \{犬\}$です。ある集合$\{A\}$に含まれるがそれとは違う集め方をした別の集合$\{B\}$にも含まれるという場合、$A \cap B$と$\cap$で表します。積集合と言い、A かつ B と読みます。たとえば、$A = \{犬\}$で$B = \{雌\}$ならば$A \cap B = \{雌犬\}$です。Aである要素のすべてと、B である要素の全てを集めたものを和集合$A \cup B$と言います。 $A = \{犬\}$で$B = \{雌\}$ B ならば$A \cup B = \{人間を含めたすべての動物のメスと犬のすべて\}$です。ですから、この中に部分集合として$\{雌犬\}$が含まれます。$(A \cup B) \supset (A \cap B)$です。（なんだか、ウィーメエンズリブの人に怒られそうな文章になってしまいました。悪意があったわけではありません。犬が好きで、女性が好きなので、たまたま興味があるものを２つあげたらこうなってしまいました。）

記号であらわすと、

$$A = \{a_1, a_2 \cdots a_l, c_1, c_2, \cdots c_n\} \text{ and } B = \{b_1, b_2, \cdots b_m, c_1, c_2, \cdots c_n, \ \},$$

の場合

$$D = A \cup B = \{a_1, a_2 \cdots a_l, b_1, b_2, \cdots b_m, c_1, c_2, \cdots c_m\}$$

式 1

$$C = A \cap B = \{\, c_1, c_2, \cdots c_m\}$$

式 2

A が起きて、その条件のもとにBが起きることをB|Aと表します。$(A \cap B) = (B \cap A)$ですが、$(A|B) \neq (B|A)$です。これらを使うと、何回かサイコロを振って出てくるもの組み合せを集合としてとらえて、その組み合わせの数やそのような組み合わせになる確率を考えることができます。たとえばサイコロを２回振って目が１とそれ以外の数になる組み合わせを考えます。Aが１になる。Bが 1 以外になることにします。組み合わせの数をNとしてその組み合わせになる確率をPとします。1 回の試行であることが起きる確率をpとします。それが起こらない確率はqで$p + q = 1$です。コインの場合は表になる確率$p = \frac{1}{2}$、表にならない確率$q = \frac{1}{2}$、サイコロの場合は１の目が出る確率$p = \frac{1}{6}$、１の目が出ない確率$q = \frac{5}{6}$ということです。互いに排反事象ですから$p + q = 1$ということは理解できますね。

１回サイコロを振る場合は、A かB しかないので

$N(A) = 1, N(B) = 1, N(A \cup B) = 2, P(A) = \frac{1}{6}, P(B) = \frac{5}{6}, P(A \cup B) = 1,$

ですね。

サイコロを２回振る場合は、$A|A, B|A, A|B, B|B$の四通りがあって、これらは互いに排反事象（あることが起きた場合には他のことは起こらない）です。

A とＢが背反事象ならば

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$$

式３

$\mathrm{N}(\mathrm{A}|\mathrm{A}) = 1, \quad \mathrm{N}(\mathrm{B}|\mathrm{A}) = 1, \mathrm{N}(\mathrm{A}|\mathrm{B}) = 1, \mathrm{N}(\mathrm{B}|\mathrm{B}) = 1, \mathrm{N}(\mathrm{A}|\mathrm{A}) \cup (\mathrm{B}|\mathrm{A}) \cup (\mathrm{A}|\mathrm{B}) \cup (B|B) = 4$
それぞれの確率は

$$\mathrm{P}(\mathrm{A}|\mathrm{A}) = \frac{1}{6} \cdot \frac{1}{6}, \quad \mathrm{P}(\mathrm{B}|\mathrm{A}) = \frac{1}{6} \cdot \frac{5}{6}, \mathrm{P}(\mathrm{A}|\mathrm{B}) = \frac{5}{6} \cdot \frac{1}{6}, \mathrm{P}(\mathrm{B}|\mathrm{B}) = \frac{5}{6} \cdot \frac{5}{6}$$

$$\mathrm{P}\big((\mathrm{A}|\mathrm{A}) \cup (\mathrm{B}|\mathrm{A}) \cup (\mathrm{A}|\mathrm{B}) \cup (B|B)\big) = 1$$

B|AはＡが起きたという条件のもとにＢが起きるのですから、与えられた条件を式にすると

$$P(B|A) = P(A)P(B|A)$$

と書くべきですが、この場合は、ＡにかかわりなくＢがおきるので、$P(B)$は一定で

$$P(B|A) = P(B)$$

です。たがいに独立したあること（Ａ）とあること（Ｂ）が同時に起きる確率はそれら個々の確率の積ですね。この関係を式で表すと次のようになります。

$$\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{A})\mathrm{P}(\mathrm{B})$$

式４

したがって

$$P(B|A) = P(A)P(B|A) = P(A)P(B)$$

式 5

と書けます。

ところで、P(A∩B)はA とＢが同時に起きるという意味ですが、実際の時間の中で「同時」ということではなくて、数学的に考えた場合に「同時」ととらえられる、一つの試行の中でということです。「同時」ということよりは「互いに独立した」という条件の方が重要なのです。

組み合わせの数を整理すると
AA=｛Ａが２回｝
AB=｛Ａが１回｝
BB=｛Ａが０回｝
と整理できて、
AA= A|A
AB= (B|A) ∪ (A|B)
BB= B|B
$\mathrm{N}(\mathrm{AA}) = 1, \quad \mathrm{N}(\mathrm{AB}) = 2, \mathrm{N}(\mathrm{B}|\mathrm{B}) = 1, \mathrm{N}(\mathrm{AA}) \cup (\mathrm{AB}) \cup (BB) = 4$
１試行について２通り、そのそれぞれについて次の試行で２通りの結果があるのだから、

その２つの試行を１セットとして、１試行と考えれば４通りあるということになります。１試についてn_1とおりの結果があり次の試行でそのそれぞれに結果についてn_2通りの結果があれば、２回の試行を１セットの試行と考えれば、あらわれる結果には$n_1 \cdot n_2$とおりの組み合わせがあります。つまり、２つを組み合わせた場合の数は一つひとつの場合の数の掛け算なのです。
実際に、考えてみます。サイコロを振って、１の目が出るのがＡ、それ以外の目が出るのをＢ とします。４回投げるのが１試行だとすると、一回目にサイコロの目が１で、残りの ３回が１以外になるのは、$(|B|B|B|A)$と表されて、

$$P(|B|B|B|A) = \mathrm{P(A \cap B \cap B \cap B)} = \mathrm{P(A)P(B)P(B)P(B)}$$

$$= pqqq = \frac{1}{6}\frac{5}{6}\frac{5}{6}\frac{5}{6} = \frac{125}{1296}$$

ここで、サイコロの目が１回だけ １ で、後の３回は１以外という組み合わせをすべて書くと、$(B|B|B|A)$、$(B|B|A|B)$、$(B|A|B|B)$、$(A|B|B|B)$の４通りがあります。

サイコロの目が１回だけ１で、後の３回は１以外になる確率は、p^1q^3の４倍、

$4pq^3$=$4\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^3$=$\frac{500}{1296}$になります。ここで$(p+q)^n$という式の展開を考えます。

n=1 ならば、$(p+q)^1 = p+q$
n=2 ならば　$(p+q)^2 = p^2+2pq+q^2$
n=3 ならば　$(p+q)^3 = p^3+3\mathrm{p}^2q+3pq^2+q^3$
n=4 ならば　$(p+q)^4 = p^4+4\mathrm{p}^3q+6p^2q^2+4pq^3+q^4$
n-4 の場合の式の後ろから２項目を見てください。サイコロの目が１回だけ１で、後の３回は１以外になる確率は、p^1q^3の４倍、$4pq^3$と同じになっています。式を展開していく過程で生じる$pqqq$、$qpqq$、$qqpq$、$qqqp$の４つの項をpq^3の項として一つにまとめたのですから、当たり前だといわれれば、それまでです。しかし、そうだとすれば、n 回サイコロを振って、１の目が k 回出る確率は、後ろから k+1 番目の項だということになります。前から数えると、$\mathrm{n}-(k-1)$番目の項です。一番先頭が、n 回振って、n 回１の目が出る（k=n になる）確率で、一番最後が一回も１の目が出ない（k=0 になる）確率です。$p+q$=1 だから、それを何乗しても１で、常に確率の総和は１だということも確認できます。各項の k の値を横軸にして、それぞれの確率を縦軸にプロットしたものを二項分布といいます。各項の前の部分に数字がありますが、この数字を二項係数と言いい次のような記号で表します。

$$\binom{n}{k}$$

$(p+q)^n$を展開したときの前から$\mathrm{n}-(k-1)$番目の項の係数という意味です。

$$(p+q)^4 = \binom{4}{4}p^4 + \binom{4}{3}\mathrm{p}^3 q + \binom{4}{2}p^2q^2 + \binom{4}{1}pq^3 + \binom{4}{0}q^4$$

と表せます。一般化して書くと

$$(p+q)^n = \sum_{k=0}^{n}\binom{n}{k}\mathrm{p}^{n-k}q^k$$

となります、

k なる確率についてだけ取り出して書くと

$$\mathrm{P(k)} = \binom{n}{k}\mathrm{p}^{n-k}q^k$$

ですが、$q = 1-p$なので、pだけの式に書き換えると

$$\mathrm{P(k)} = \binom{n}{k}\mathrm{p}^{n-k}(1-p)^k$$

となります。１回の確率がpであること n 回繰り返した時の二項分布を

$$\mathrm{B}(n,p)$$

と書きます。

さて、残された問題は$\binom{n}{k}$をどのように計算するかです。式１は

$$(p+q)^n = \sum_{k=0}^{n}\binom{n}{k}\mathrm{p}^{n-k}q^k$$

となっていてpとqのべき乗数の和$(n-k)+\mathrm{k}$はどの項も n です。$pppqqq$のように書くと、n この文字が列を作っています。もしこれが a,b,c,d,e,f のようにn個の文字が並んでいて、その順番を入れ替えたときに、何個の並べ方があるかを考えます。最初に来る可能性があるのは n 個あって、その次は先頭になったもの以外、その次は先頭とその次目以外の文字が並ぶと考えていけば、それぞれについて、組み合わせの数は、n、$n-1$、$n-2$…のようになって、最後は一つだけになります。独立した組み合わせ同士を組み合わせた組み合わせの数は、組み合わせ数同士の掛け算ですから、組合わせの総数は$\mathrm{n}(n-1)(n-2)\cdots 3\cdot 2\cdot 1$つまり、n!になります。さて、文字が n 個あるときに、k 個を p 組、$\mathrm{n}-\mathrm{k}$個を q 組に分けたとします。同じように考えると、p 組だけの並び方の総数は k!で q 組だけの並び方の総数は(n-k)!です。組み合せ数同士の掛け算が、組み合わせた場合の組み合わせ数の総数なのですから、

p 組の内部の組み合わせ数×q 組の内部の組み合わせ数×組内の組み合わせを考慮しない p と q の組の違いだけの組み合わせ数＝総組み合わせ数

$$\mathrm{k!\,(n-k)!}\binom{n}{k} = n!$$

となります。これを変形すると、

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

式 6

という公式が導き出せます。ここまでやってきたことを言い変えると、n個の文字があったものから、k個を選び出して P 組にしたのです。n個の中からk個取り出す組み合わせの数はコンビネーションという名前がついていて、　${}_nC_k$と表します

以上を整理すると

$$\binom{n}{k} = \quad {}_nC_k = \frac{n!}{k!\,(n-k)!}$$

となります。

二項分布は$\mathrm{B}(n,p)$と表します。確率pで起こる現象を n 回繰り返す二項分布ということです。$\mathrm{B}(n,p)$について、確率pでおこる事象が起こる回数は

$$\mathrm{P(k)} = \binom{n}{k}\mathrm{p}^{n-k}(1-p)^k = \quad {}_nC_k\mathrm{p}^{n-k}(1-p)^k = \frac{n!}{k!\,(n-k)!}\mathrm{p}^{n-k}(1-p)^k$$

式 7

二項分布全体について書くと以下のようになります。

$$(p+q)^n = \binom{n}{n}p^nq^0 + \binom{n}{n-1}p^{n-1}q^1 + \cdots + \binom{n}{n-k}p^{n-k}q^k + \cdots + \binom{n}{0}p^0q^n$$

$$= \sum_{k=0}^{n}\binom{n}{n-k}p^{n-k}q^k$$

式 8

これらを、視覚的に表したものが下図で、これをパスカルの三角形と言います。

$$\binom{0}{0}$$

$$\binom{1}{1} \quad \binom{1}{0}$$

$$\binom{2}{2} \quad \binom{2}{1} \quad \binom{2}{0}$$

$$\binom{3}{3} \quad \binom{3}{2} \quad \binom{3}{1} \quad \binom{3}{0}$$

$$\binom{4}{4} \quad \binom{4}{3} \quad \binom{4}{2} \quad \binom{4}{1} \quad \binom{4}{0}$$

$$\binom{5}{5} \quad \binom{5}{4} \quad \binom{5}{3} \quad \binom{5}{2} \quad \binom{5}{1} \quad \binom{5}{0}$$

これらを具体的に計算すると次のようになります。

1

1　1

1　2　1

1　3　3　1

1　4　6　4　1

1　5　10 10　5　1

1　6　15　20　15　6　1

全体として以下の式の様に対称形になりますが

$$\binom{n}{k} = \binom{n}{n-k}$$

上下にみると次のような関係があります。

$$\binom{n}{k} = \binom{n-k}{k-1} + \binom{n-k}{k}$$

1

1　1

1　2　1

1　3　3　1

1　4　6　4　1

1　5　10 10　5　1

1　6　15　20　15　6　1

$2 = 1 + 1$

$3 = 1 + 2$

$4 = 1 + 3,\ 6 = 3 + 3$

１回の試行である事象が起こる確率をpとして、それらを n 回繰り返すことを、二項分布の記号として、$B(n,p)$と表します。この時、ある事象が k 回起きる確率を $p(k)$と すると、$p(k)$ は二項分布にしたがうといくことで、次のように表します。

$$p(k) \sim B(n,p)$$

図 5 には、サイコロを振って何回 1 が出るかという確率 $p(k) \sim B\left(n, \frac{1}{6}\right)$ を （ $1 \leq n \leq 10$ ）について、図示しました。n の増加によって、二項分布が、次第にシンメトリックな、正規分布に近づいていくことがおぼろげにわかります。

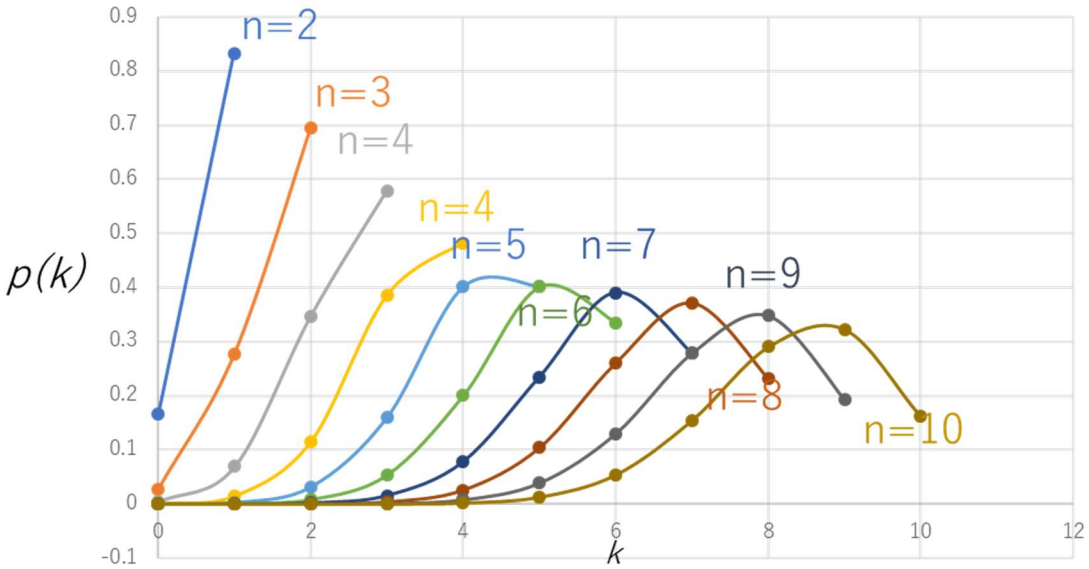
p(k)
k
n=2
n=3
n=4
n=4
n=5
n=7
n=6
n=9
n=8
n=10
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
0
2
4
6
8
10
12

図. 5. 二項分布