【セミナー】

離散型確率変数とその分布

――社会調査の推測統計（2）――

小林　久高
KOBAYASHI Hisataka

1　はじめに

確率変数には離散型のものと連続型のものがある。離散型確率変数とは離散的な値をとる確率変数だ。たとえば、サイコロの目という確率変数は 1,2,3,4,5,6 という値をとるが、決して 1.25889…といった値をとることはない。それに対して重さや長さなどは離散的な値ではなく連続的な値をとる。こういった値をとる確率変数を連続型確率変数と言う。

今回は離散型確率変数について説明する。まず、離散型確率変数の表記法などについて述べ、次いで代表的な離散型確率分布である二項分布について説明する。この説明の中には大数の法則への言及も含まれる。最後に述べるのは超幾何分布である。ここでは、有限母集団補正項についても簡単に触れる。

2　離散型確率変数の基礎

2.1 確率分布

表 1 は、裏に 1 と書いてあるものが 1 枚、2 と書いてあるものが 2 枚、3 と書いてあるものが 1 枚の計 4 枚のカードからランダムに 1 枚選んだ場合、裏に書いてある数Xについての確率分布表である。

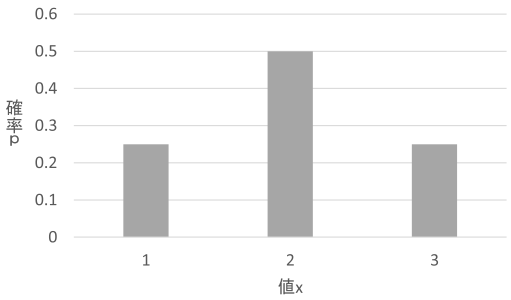
図 1 はそれをグラフで表したものだ。横軸が値、縦軸が確率になっている。離散型の確率変数についてはこのような形で確率分布を表すことができる。

表 1　X の確率分布表

確率変数X（カードに書かれた数）

値x	1	2	3	計
確率p	1/4	2/4	1/4	1

図 1　X の確率分布のグラフ



離散型確率変数については、次のような表記で確率を表す。

$$P(X = x_i) = f(x_i) \qquad (i = 1,2,\cdots,t)$$

これは言葉で言うと「X という確率変数が xi という値をとる確率が f(xi)である」となる。f(x)は確率質量関数と呼ばれる。4 枚のカードの例では次のようになる。

$$P(X = 1) = \frac{1}{4}$$

$$P(X=2)=\frac{2}{4}$$
$$P(X=3)=\frac{1}{4}$$

X は確率変数であるから、どの値についての確率も 0 以上であり、確率の計は 1 となる。

$$f(x_i)\geq 0 \qquad (i=1,2,\cdots,t)$$
$$\sum_{i=1}^{t} f(x_i)=1$$

2.2 期待値と分散

離散型確率変数 X の期待値と分散は次のようになる。

$$E(X)=\sum_{i=1}^{t} x_i p_i$$
$$V(X)=\sum_{i=1}^{t}(x_i-E(X))^2 p_i$$

4 枚のカードの例では次の通り。

$$E(X)=1\times\frac{1}{4}+2\times\frac{2}{4}+3\times\frac{1}{4}=2$$
$$V(X)=(1-2)^2\frac{1}{4}+(2-2)^2\frac{2}{4}+(3-2)^2\frac{1}{4}=\frac{1}{2}$$

確率変数の期待値は、「とりうる値×確率（比率）」の総和で求められ、分散は「(とりうる値－期待値）の 2 乗×その確率」の総和で求められる。

分散はとりうる値の平均からのズレの 2 乗の期待値（平均）だから、確率変数と期待値だけを用いて分散を表すと次のようになる。

$$V(X)=E\left((X-E(X))^2\right)$$

2.3 離散型 2 次元確率分布

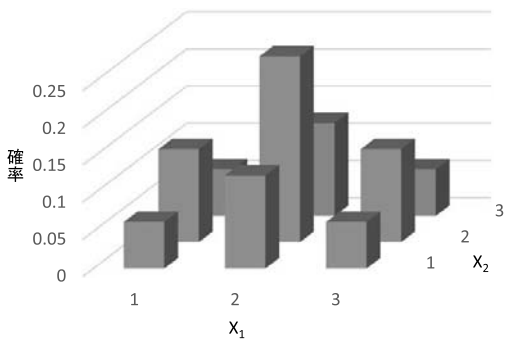
離散型の 2 次元確率分布についても簡単に触れておこう。上の 4 枚のカードから 2 枚のカードを復元抽出で引くことを考え、1 枚目のカードの数を X_1、2 枚目のカードの数を X_2 とする。そのとき、表 2 のような 2 次元確率分布表が作成できる。表の（1,2）とその下にある 2/16 は、1 枚目（X_1）が 1 であり 2 枚目（X_2）が 2 である確率が 2/16 であることを表している。X_1 と X_2 は独立なので、計の行や列にある両方の周辺確率（1/4 と 2/4）を掛け合わせたものがその確率になっている。表 2 をグラフで表したものが図 2 である。

表 2　（X_1,X_2）の 2 次元確率分布

		X_2			
		1	2	3	計
X_1	1	(1,1) 1/16	(1,2) 2/16	(1,3) 1/16	1/4
	2	(2,1) 2/16	(2,2) 4/16	(2,3) 2/16	2/4
	3	(3,1) 1/16	(3,2) 2/16	(3,3) 1/16	1/4
	計	1/4	2/4	1/4	1

上段：値 、下段：確率

図 2　離散型 2 次元確率分布のグラフ



2.4 標本平均の分布

ところで、上の X_1 と X_2 は表 1 の母集団 X からのサイズ 2 の標本とも考えられる。そして標本と母集団の間には、標本平均 $\overline{X}$ の期待値 $E(\overline{X})$ が母平均 μ に等しくなり、標本平均の分散 $V(\overline{X})$ が母分散を標本サイズで割ったもの（σ^2/n）に等しくなるという関係があると言われている。このことも確認しておこう。まず、標本平均は次のものだ。

$$\overline{X} = \frac{1}{2}(X_1 + X_2)$$

この標本平均の値と確率は表 3 のようになっており、それをまとめると表 4 になる。

表 3　標本平均の値と確率

		X_2			
		1	2	3	計
X_1	1	1 1/16	1.5 2/16	2 1/16	1/4
	2	1.5 2/16	2 4/16	2.5 2/16	2/4
	3	2 1/16	2.5 2/16	3 1/16	1/4
	計	1/4	2/4	1/4	1

上段：値（平均）　下段：確率

表 4　標本平均の確率分布

値	1	1.5	2	2.5	3	計
確率	1/16	4/16	6/16	4/16	1/16	1

これをもとに標本平均の期待値を計算すると、

$$E(\overline{X}) = 1\left(\frac{1}{16}\right) + 1.5\left(\frac{4}{16}\right) + 2\left(\frac{6}{16}\right)$$
$$+ 2.5\left(\frac{4}{16}\right) + 3\left(\frac{1}{16}\right) = 2$$

標本平均の分散は、

$$V(\overline{X}) = (1-2)^2\left(\frac{1}{16}\right) + (1.5-2)^2\left(\frac{4}{16}\right)$$
$$+ (2-2)^2\left(\frac{6}{16}\right) + (2.5-2)^2\left(\frac{4}{16}\right)$$
$$+ (3-2)^2\left(\frac{1}{16}\right) = \frac{4}{16} = \frac{1}{4}$$

となる。母平均と母分散は下のものだった。

$$\mu = E(X) = 2$$
$$\sigma^2 = V(X) = \frac{1}{2}$$

したがって、確かに標本平均 $\overline{X}$ の期待値 $E(\overline{X})$ は母平均 μ と等しく、標本平均の分散 $V(\overline{X})$ は母分散を標本サイズで割ったもの（σ^2/n）に等しい（証明については小林,2018 参照）。

3　二項分布

3.1 二項分布とは

さて、推測統計の世界では二項分布と呼ばれる離散型の確率分布はとても重要である。それは次のようなものだ。

1 回の試行で、注目している事象の生じる確率が p だとする。その試行を独立に n 回繰り返したとき、注目している事象の生じる回数を X とする。この X の分布が二項分布である。

二項分布に従う確率変数は、注目事象の出現回数という値をとり、その範囲は「0～試行回数」までの整数値である。二項分布は、Bin（n, p）と表記する。n は試行回数、p はその事象が 1 回の試行で生じる確率である。注目している事象が 1 回の試行で生じない確率は（1－p）である。

ところで、p は注目している事象の 1 回の試行での出現確率だから、注目している対象の母比率とも考えられる。そうすると、二項分布が述べているのは、注目対象の母比率が p である母集団から、サイズ n の標本を取り出したとき、標本の中にいる注目対象の個数 X についての確率分布ということになる。この視点はとても重要だ。

具体例を示そう。今、支持政党のある者が 2/5、ない者が 3/5 いる母集団があるとし、支持ありに注目しているとする。すなわち、注目している支持ありの母比率は 2/5 だ。ここでこの母集団から 3 人を復元抽出で抽出するすると、このサイズ 3 の標本における支持ありの人数 X は、二項分布 Bin（3, 2/5）に従い、X のとる値は 0 人、1 人、2 人、3 人ということになる。

3.2 二項分布の確率

上の X は確率変数であり、それぞれの値と確率の関係については次の式が成り立つ。

$$P(X=k)={}_nC_k p^k(1-p)^{n-k} \quad (k=0,1,2\cdots,n)$$

ここで、P（X=k）は X が k になる（標本で注目対象の数が k になる）確率を意味する。p は 1 回の試行での注目事象の出現確率（注目対象の母比率）、1－p は 1 回の試行での注目事象の非出現確率（非注目対象の母比率）である。また、n は試行回数（標本サイズ）、k は注目している事象の出現回数（標本での注目対象の数）であり、${}_nC_k$ は n 個から k 個を選ぶ組み合わせ数だ（組合せについては補を参照）。

上の支持政党の具体例で計算すると次のようになる。

$$P(X=0)={}_3C_0 p^0(1-p)^3=1\times\left(\frac{2}{5}\right)^0\left(\frac{3}{5}\right)^3=\frac{27}{125}$$

$$P(X=1)={}_3C_1 p^1(1-p)^2=3\times\left(\frac{2}{5}\right)^1\left(\frac{3}{5}\right)^2=\frac{54}{125}$$

$$P(X=2)={}_3C_2 p^2(1-p)^1=3\times\left(\frac{2}{5}\right)^2\left(\frac{3}{5}\right)^1=\frac{36}{125}$$

$$P(X=3)={}_3C_3 p^3(1-p)^0=1\times\left(\frac{2}{5}\right)^3\left(\frac{3}{5}\right)^0=\frac{8}{125}$$

すなわち、標本で支持ありの者が 0 人である確率は 27/125、1 人である確率は 54/125、2 人である確率は 36/125、3 人である確率は 8/125 ということになる。また、全体の確率の和は 1 になる。

$$P(X=0)+P(X=1)+P(X=2)+P(X=3)$$
$$=\frac{27}{125}+\frac{54}{125}+\frac{36}{125}+\frac{8}{125}=1$$

3.3 二項分布の式の意味

二項分布において、値（支持ありの人数）に対応する確率を求める式の意味を説明したものが表 5 だ。

表 5　二項分布の説明

パターン	1人目	2人目	3人目	確率
1	×	×	×	3/5×3/5×3/5＝27/125
2	○	×	×	2/5×3/5×3/5＝18/125
3	×	○	×	3/5×2/5×3/5＝18/125
4	×	×	○	3/5×3/5×2/5＝18/125
5	○	○	×	2/5×2/5×3/5＝12/125
6	○	×	○	2/5×3/5×2/5＝12/125
7	×	○	○	3/5×2/5×2/5＝12/125
8	○	○	○	2/5×2/5×2/5＝8/125
計				1

○支持あり、×支持なし

支持ありと支持なしを含む母集団から 3 人選ぶわけだから、そのパターンは 8 通りあり、それぞれ

の確率は表の右のようになる。

今、支持ありが 2 人の場合に注目しよう。それはパターン 5〜7 である。これら 3 パターンではいずれも支持ありが 2 人現れ、支持なしが 1 人現れるのだから、どのパターンの確率も、

$$\left(\frac{2}{5}\right)^2\left(\frac{3}{5}\right)^1$$

である。パターン 5〜7 の違いは、支持あり 2 人が、1 人目、2 人目、3 人目のどこに現れるかの違いであり、これが 3 パターンであることは、3 つの中から 2 つを選ぶ組み合わせ数（${}_3C_2$）が 3 であることに対応している。したがって、支持ありが 2 人になる確率全体は次の式で求められる。

$$P(X=2)={}_3C_2p^2(1-p)^1$$
$$=3\times\left(\frac{2}{5}\right)^2\left(\frac{3}{5}\right)^1=\frac{36}{125}$$

3.4 二項分布の期待値と分散

二項分布に従う X の期待値と分散については、次の式が成り立つ。

$$E(X)=np$$
$$V(X)=np(1-p)$$

上の例では次のようになる。

$$E(X)=np=3\left(\frac{2}{5}\right)=\frac{6}{5}$$
$$V(X)=np(1-p)=3\left(\frac{2}{5}\right)\left(\frac{3}{5}\right)=\frac{18}{25}$$

なぜこのような簡単な式が成り立つかは次の通りだ。まず、1 回の試行だけを考え、注目している A の生じる回数を値とする確率変数 Y を考える。1 回の試行では Y のとる値は 0 か 1 であり、A が生じる確率は p、生じない確率は 1−p である。Y の確率分布表は表 6 のようになる。

表 6　Y の確率分布表

確率変数Y

値y	1	0	計
確率p	p	1−p	1

期待値は「とりうる値×確率」の総和だから、この確率変数の期待値は次のようになる。

$$E(Y)=1p+0(1-p)=p$$

また、確率変数の分散は「(とりうる値−期待値) の 2 乗×確率」の総和であるから、次のようになる)。

$$V(Y)=(1-p)^2p+(0-p)^2(1-p)$$
$$=(1-p)^2p+p^2(1-p)$$
$$=p(1-p)((1-p)+p)$$
$$=p(1-p)$$

まとめておこう。

$$E(Y)=p$$
$$V(Y)=p(1-p)$$

次に、上の試行を n 回独立して行うと考え、それぞれの確率変数を $Y_1,Y_2,...,Y_n$ とする。$Y_1,Y_2,...,Y_n$ のそれぞれも、Y とまったく同じく 1 または 0 をとる確率変数であり、分布も同じである。したがって、

$$E(Y)=E(Y_1)=E(Y_2)=\cdots=E(Y_n)=p$$

$$V(Y)=V(Y_1)=V(Y_2)=\cdots=V(Y_n)=p(1-p)$$

ところで、1 回の試行で A の生じる回数についての確率変数が Y だったので、n 回の試行で A の生じる回数を X とするとき、次の式が成立する。

$$X=Y_1+Y_2+\cdots+Y_n$$

この期待値は次のようになる。

$$\begin{aligned}E(X)&=E(Y_1+Y_2+\cdots+Y_n)\\&=E(Y_1)+E(Y_2)+\cdots+E(Y_n)\\&=\underbrace{p+p+\cdots+p}_{n}=np\end{aligned}$$

また、確率変数 $Y_1,Y_2,...,Y_n$ はそれぞれ独立しているので、X の分散は次のようになる。

$$\begin{aligned}V(X)&=V(Y_1+Y_2+\cdots+Y_n)\\&=V(Y_1)+V(Y_2)+\cdots+V(Y_n)\\&=\underbrace{p(1-p)+p(1-p)+\cdots+p(1-p)}_{n}=np(1-p)\end{aligned}$$

それゆえ、期待値は np、分散は np(1-p)となるのである。まとめておこう。

$$E(X)=np$$
$$V(X)=np(1-p)$$

3.5 二項分布と標本比率

二項分布は標本における注目属性をもつものの数を値とする分布であった。したがって、二項分布は注目属性をもつものの標本比率にも関係している。そのあたりのことについて説明しよう。

まず、注目属性の標本比率 P'は「注目属性の標本での数」を「標本サイズ」で割ったものだから、次の式が成り立つ。

$$P'=\frac{X}{n}$$

ここでこの標本比率の期待値と分散は次のようになる。

$$E(P')=E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)=\frac{1}{n}np=p$$
$$V(P')=V\left(\frac{X}{n}\right)=\frac{1}{n^2}V(X)=\frac{1}{n^2}np(1-p)$$
$$=\frac{p(1-p)}{n}$$

したがって、p を母比率、P'を標本比率、n を標本サイズとするとき、標本比率 P'の期待値は p、分散は p(1－p)/n である。確率変数である標本比率 P'は母比率 p を中心に、分散 p(1－p)/n の広がりをもって分布するのである。これはとても重要なことなのできちんと書いておこう。

$$E(P')=p$$
$$V(P')=\frac{p(1-p)}{n}$$

たとえば支持政党のある者が 2/5 いる母集団から 3 人の標本を復元抽出で抽出するとき、支持ありの標本比率は次のように分布する。

$$E(P')=p=\frac{2}{5}$$

$$V(P')=\frac{p(1-p)}{n}=\frac{\frac{2}{5}\left(\frac{3}{5}\right)}{3}=\frac{6}{75}=\frac{2}{25}$$

母比率と標本比率の間に見られるこの関係は、

標本比率から母比率を推定するときに利用される。

さて上の関係であるが、別の観点からも見ておくことにしよう。この考え方は母数と標本平均の関係についてのよい復習になると思う。

3.4 の二項分布の期待値と分散の解説では、1 回の試行で 0,1 という値をとる確率変数 Y を考えた。この期待値と分散は次のようなものだった。

$$E(Y) = p$$
$$V(Y) = p(1-p)$$

そして、次の X の分布を考え、それが結局二項分布に従うことを示したのである。

$$X = Y_1 + Y_2 + \cdots + Y_n$$

ここで、次のような確率変数 $\bar{Y}$ を考えてみよう。

$$\bar{Y} = \frac{1}{n}\left(Y_1 + Y_2 + \cdots + Y_n\right)$$

この $\bar{Y}$ は、母集団 Y からの標本 $Y_1, Y_2, \ldots, Y_n$ の標本平均と考えられる。

ここで「標本平均の期待値が母平均に等しくなり、標本平均の分散が母分散を n で割ったものに等しくなる」という知識をもとにすると、次が成り立つ。

$$E(\bar{Y}) = E(Y) = p$$
$$V(\bar{Y}) = \frac{V(Y)}{n} = \frac{p(1-p)}{n}$$

一方、この標本平均 $\bar{Y}$ は二項分布において注目されているものの標本比率 P'とも考えられる。というのは、次の式が成り立つからだ。

$$\bar{Y} = \frac{1}{n}\left(Y_1 + Y_2 + \cdots + Y_n\right) = \frac{X}{n} = P'$$

したがって、次の式が成り立つ。

$$E(P') = E(\bar{Y}) = E(Y) = p$$
$$V(P') = V(\bar{Y}) = \frac{V(Y)}{n} = \frac{p(1-p)}{n}$$

こういうわけで、こちらの考え方からも、p を母比率、P'を標本比率、n を標本サイズとするとき、標本比率 P'の期待値は p であり、分散が p(1－p)/n であることが示せる。

3.6 大数の法則

上の標本比率の議論は有名な大数の法則にも関係する。大数の法則とは、標本サイズ n を大きくしていくことで、標本比率が母比率に等しくなっていくことを言うものだ。すなわち、二項分布に従う X についての次の式が大数の法則を示している（n は標本サイズ、p は母比率である。等式左の X/n が確率変数なのに対し、右の p は決まった値であることに注意しよう）。

$$\lim_{n\to\infty}\frac{X}{n} = p$$

上で見たように標本比率（P'＝X/n）の期待値は p（母比率）、分散は p(1－p)/n であった。ここで n をどんどん大きくしていくとどうなるだろうか。

X は決まった値ではなく確率変数なので、X/n が 0 に近づいていくなどと考えてはいけない。そうではなく、標本比率 X/n=P'の分散である p(1－p)/n がどんどん 0 に近づいて行くのである。このことは標本比率 P'の分布の広がりがどんどん狭くなって行き、最後には広がりが 0 になってしまうことを

意味する。それゆえ結局、標本比率 P'は、その期待値である p（母比率）に等しくなるわけだ。

大数の法則をはじめて知ったとき、筆者はそんなことは当たり前じゃないかと思った。標本サイズをどんどん大きくしていくと最後には母集団に含まれるすべてのものが標本になるので、比率は母集団と同じになるに決まっていると思ったのである。しかし話はそういうことではない。ここでは無限母集団からの標本、あるいは復元抽出法での標本を考えているのである。このあたり、間違えないようにすること。

4　超幾何分布

4.1 超幾何分布とは

二項分布に関連の深い確率分布として超幾何分布がある。二項分布は無限母集団あるいは母集団から復元抽出した標本に関する分布であるのに対し、超幾何分布は有限母集団から非復元抽出した標本に関する分布である。基礎的な推測統計は無限母集団を対象とするので、超幾何分布は基礎的なテキストでは紹介されない。本稿も無限母集団を対象に議論を進めているのだが、例外的にここで紹介しておくことにしよう。

N を母集団の大きさ、M をその中で注目する属性をもっているものの数、n を標本サイズとし、X を当該属性のあるものの標本での出現数とするとき、X＝k における超幾何分布は次のように表せる。

$$P(X=k)=\frac{\left({}_{M}C_{k}\right)\left({}_{N-M}C_{n-k}\right)}{\left({}_{N}C_{n}\right)}$$

この式は複雑に見えるが、高校時代の確率の問題にあったようなものだ。少し説明しておこう。

4.2 超幾何分布の式の意味

今、大きさ 5（N＝5）の母集団を考え、そこに属性 A をもつものが 2 個（A_1,A_2）、属性 B をもつものが 3 個（B_1,B_2,B_3）含まれているとする（図 3）。ここで A の属性に注目すると M＝2 となる。

この母集団からサイズ 2（n=2）の標本を抽出するとして、それぞれの個体が出現する場合を○、出現しない場合を×とするとき、標本のパターンは表 7 のように 10 通りになる。それぞれのパターンの出現確率はすべて同じだ。

図 3　母集団

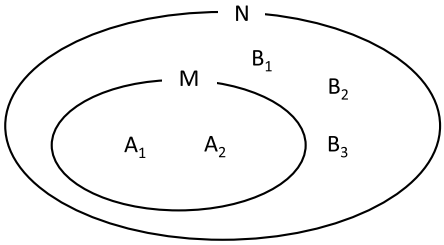


表 7　サイズ 2 の標本のパターン

パターン	A_1	A_2	B_1	B_2	B_3
1	○	○	×	×	×
2	○	×	○	×	×
3	○	×	×	○	×
4	○	×	×	×	○
5	×	○	○	×	×
6	×	○	×	○	×
7	×	○	×	×	○
8	×	×	○	○	×
9	×	×	○	×	○
10	×	×	×	○	○

パターン 1 では注目している属性のものが 2 個出現しているので k＝2、パターン 2～7 では k＝1、パターン 8～10 では k＝0 となる。ここから次のこ

とがわかる。

$$P(X=0)=\frac{3}{10}$$
$$P(X=1)=\frac{6}{10}$$
$$P(X=2)=\frac{1}{10}$$

ここで、注目属性をもつものが標本で 1 個出現する場合（k＝1）の式について考えよう。この確率についての式は次のものだった。

$$P(X=1)=\frac{({}_2C_1)({}_3C_1)}{({}_5C_2)}$$

この式の分母の ${}_5C_2$ はNの中から2個選ばれる組み合わせ数であり、表 7 のパターンが全部で 10 になることを示している。分子の ${}_2C_1$ は M の中（すなわち A_1,A_2）から 1 個選ばれる組み合わせ数（1 個で組み合わせというのも妙だが）であり 2 通りである。また、分子の ${}_3C_1$ はM以外（すなわち B_1,B_2,B_3）から 1 個選ばれる組み合わせ数であり 3 通りである。そして、これらを掛け合わせた 2×3 の 6 通りが、パターン 2〜7 の 6 通りに対応する。したがって、式が表しているのは、この 6 パターンを全体の 10 パターンで割るということであり、それは、注目する属性のものが 1 個出現する確率になるのだ。

$$P(X=1)=\frac{({}_2C_1)({}_3C_1)}{({}_5C_2)}=\frac{2\times 3}{10}=\frac{6}{10}$$

さて、二項分布の説明に際しては「支持政党のある者が 2/5、ない者が 3/5 いる母集団があり、ここから 3 人の標本を復元抽出で抽出する」という例を出したが、ここではこれに似た「支持政党ありの者が 20 人いる 50 人の母集団から非復元抽出で 3 人選ぶ」状況を見てみよう。

このとき、標本で支持政党ありが k 人（0〜3）になる確率は超幾何分布の式で決まる。支持政党ありが 2 人いる場合の確率を計算すると、下のように 0.290 になる。

$$P(X=2)=\frac{({}_{20}C_2)({}_{50-20}C_{3-2})}{({}_{50}C_3)}=0.290\cdots$$

前の二項分布の例で X＝2 となる確率は 36/125＝0.288 だったから、二項分布と超幾何分布では確率がわずかに異なることがわかる。

4.3 超幾何分布の期待値と分散

超幾何分布の期待値と分散は次のようになる。

$$E(X)=\frac{nM}{N}$$
$$V(X)=\left(\frac{nM}{N}\right)\left(1-\frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)$$

N＝50、M＝20、n＝3 の上の例では次のようになる。

$$E(X)=\frac{3\times 20}{50}=\frac{6}{5}$$
$$V(X)=\left(\frac{3\times 20}{50}\right)\left(1-\frac{20}{50}\right)\left(\frac{50-3}{50-1}\right)=\frac{846}{1225}$$
$$=0.690$$

二項分布の例での期待値は次の通りだった。

$$E(X)=np=3\left(\frac{2}{5}\right)=\frac{6}{5}$$
$$V(X)=np(1-p)=3\left(\frac{2}{5}\right)\left(\frac{3}{5}\right)=\frac{18}{25}=0.72$$

ここから、超幾何分布の期待値は二項分布の期待値と同じであり、超幾何分布の分散は二項分布の分散より小さいことが予測できる。

この予測が正しいことを確認するためには、超幾何分布の母集団について、人数ではなく比率をもとに考えてみればいい。すなわち、注目する属性をもつものの数ではなく、注目する属性を持つものの母集団での比率 p をもとに考えるのである。

$$\frac{M}{N} = p$$

このとき、超幾何分布の期待値と分散の式は次のようになる。

$$E(X) = np$$
$$V(X) = np(1-p)\left(\frac{N-n}{N-1}\right)$$

この式を見ると、超幾何分布の期待値は二項分布と同じであり、分散は二項分布の分散に

$$\frac{N-n}{N-1}$$

をかけたものになっていることがわかる。したがって、超幾何分布の分散は二項分布の分散より通常小さくなることがわかる。

このことは次のことを意味している。注目している属性のものが何割か含まれる母集団があるとし、そこからの復元抽出または非復元抽出で標本を抽出するとき、標本でのその属性をもつものの数の期待値はどちらの抽出法でも同じだが、その分散は非復元抽出法の方が小さくなる。

n が一定の場合、N がどんどん大きくなると(N－n)/(N－1)は 1 に近づき、超幾何分布の分散は二項分布の分散に近づいて行く。超幾何分布の二項分布による近似は、$n < 0.1N$ を満たせば用いることができると言われている（蓑谷 2003:442）。

(N－n)/(N－1)は有限母集団修正項と呼ばれるものだ。これは二項分布と超幾何分布の関係だけにかかわるものではない。たとえば有限母集団からの非復元抽出標本において、標本平均の分散は次のようになる。

$$V(\bar{X})_{yugen} = \frac{N-n}{N-1} V(\bar{X})_{mugen}$$

読者はこの有限母集団修正項に、本シリーズの最後でまた出会うことになるはずである。

5　おわりに

統計学のテキストに登場する二項分布は、決して難しいものではないが、きちんと順を追って読んでいかないとわけがわからなくなる。それで、「ここはまあいいか」と考え、次の章に進むなんてことがよくある。

しかしながらこの分布は重要だ。これについての知識がなければ比率についての推定や検定は理解できない。また標本サイズの決定についての理屈もわからない。実に重要な確率分布なのである。わかりやすく書いたつもりなので理解してほしい。

文献

小林久高,2018「母集団・標本・確率変数」『同志社社会学研究』22
蓑谷千凰彦, 2003『統計分布ハンドブック』朝倉書店

補：順列と組合せ

1　順列

異なるn個のものからr個を取り出して1列に並べたものの総数。

$$ {}_nP_r = \underbrace{n(n-1)(n-2)\cdots(n-r+1)}_{r} = \frac{n!}{(n-r)!} $$

例
異なる5個のものから3個を並べた順列の総数。

$$ {}_5P_3 = 5\times4\times3 = 60 $$

2　組合せ

異なるn個のものから順序を問題とせずr個を取り出して1組にしたものの総数。

$$ {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n-r)!} $$

また、次の式が成り立つ。

$$ {}_nC_r = {}_nC_{n-r} \qquad {}_nC_n = {}_nC_0 = 1 \qquad 0! = 1 $$

表記については、下の右のようにする場合もある。

$$ {}_nC_r = \binom{n}{r} $$

例
異なる5個のものから3個の組み合わせの総数。

$$ {}_5C_3 = \frac{5\times4\times3}{3\times2\times1} = 10 $$