【セミナー】

母集団・標本・確率変数

――社会調査の推測統計（1）――

小林　久高
KOBAYASHI Hisataka

1　はじめに

かなり昔、『同志社社会学研究』において「これから数回のシリーズで、計量分析を理解する上で初心者がひっかかる点をシリーズとして論じる」と述べ、「ベクトルと行列の基礎」を書いた。しかしその後、他の用事にかまけてずっと続編を書かないでいた。そうこうしているうちにどんどん年を取り、このままでは学生や院生諸君にこういったことを告げぬまま大学を去ることになるのではという思いが強くなってきた。これでは申し訳ない。そういうわけで、これからはさぼらずこの仕事を続けていこうと考えた。

今回とりあげるのは母集団と標本にかかわることである。これは本シリーズの中でも最も基本的なものだ。なぜどんなテキストにでも書いてあるこのような問題を取り上げるのかというと、この問題が推測統計においてもっとも基本的な問題にもかかわらず、理解するためにはいくつかのハードルがあり、そのハードルについて多くのテキストは強調しておらず、結局、多くの学生はハードルを乗り越えられないまま、ぼんやりとした知識に基づいて、実際の検定や推定を行うことになってしまうからである。

学生時代の筆者はまさにそうだった。「母集団が確率分布であること」など考えもしなかった。「実際の標本と推測統計で考えられている標本の違い」についてもぼんやりとしか理解していなかった。「標本が確率変数のセットであること」も知らなかった。「確率変数の合成」というテーマは重要でない難しいテーマだと考えスルーした。「平均の平均」についてはさっぱりわからず、言い間違いかもしれないと考えた。それでも例を参考にしながら手続きにしたがって計算すれば検定も推定もできた。しかし、なんともかんとも気持ちの悪い状態であった。

その後、統計学の勉強を基礎からやり直したとき、この世界にはさまざまなハードルや落とし穴があることが見えてきた。「ああ、ここでつまずいたのだ」「ああ、ここでよくわからないまま先に進んで失敗したんだ」ということがわかってきたのである。テキストはそのわなについて強調せずさらりと書いている。真っ白な頭で考えるとそれはわなでもなんでもないからだ。しかし、多くの読者には常識があり、この常識がさまざまな場所にわなをしかける。多くの読者はわなにかかっていることも知らないまま進み、わけのわからないところへ行ってしまうのだ。

以下では、このわなを強調しながら母集団や標本について説明しようと思う。読み終わったときに読者の「なんとなくわからない、いや～な気分」が吹っ飛ぶことを望んでいる。

2　変数の分布と平均・分散

2.1 変数と値

異なる値をとりうるものを変数と言う。物の重

さや体積は異なる値をとりうるので変数である。身長や体重も異なる値をとりうるので変数である。収入や教育年数や勤続年数なども異なる値をとりうるので変数である。

「変数の分布」は、変数のそれぞれの値についてその頻度や割合がどうなっているのかを示している。たとえば、「体重が 50kg の人が 10 人、60kg の人が 20 人」や「収入 500 万円以下の世帯が○%」」という表現は変数の分布についての表現だ。

変数の分布について、平均や分散で分布の全体的なありようを示すことがよくある。平均は分布の中心的な傾向を示すものであり、分散は分布のちらばりの傾向を示すものだ。データの形式に応じてどのように平均や分散が算出されるのかをまず明らかにしよう。

2.2 平均と分散の基本式

（1）データの形式

表 1 は最も基本的な形式で表されたデータである。

表 1　各ケースの得点の表

ケース番号(i)	得点(x_i)
1	30 (x_1)
2	20 (x_2)
3	40 (x_3)
4	50 (x_4)
5	40 (x_5)
6	40 (x_6)
7	30 (x_7)
8	40 (x_8)
9	50 (x_9)
10	60 (x_{10})

（2）平均

このように n 個の値があるとき、その平均は次式で表される。i はケース番号。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right)$$

表のデータでは平均は下の式のようになる。

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$$
$$= \frac{1}{10}\left(\begin{matrix}30+20+40+50+40\\+40+30+40+50+60\end{matrix}\right) = 40$$

（3）分散

n 個の値があるとき、その分散は次式で表される。

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$
$$= \frac{1}{n}\left\{\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \cdots + \left(x_n - \bar{x}\right)^2\right\}$$

表の例では、分散は、上の平均＝40 を利用して次のように計算できる。

$$s^2 = \frac{1}{10}\sum_{i=1}^{10}\left(x_i - \bar{x}\right)^2$$
$$= \frac{1}{10}\left\{\begin{matrix}(30-40)^2 + (20-40)^2 + (40-40)^2\\+(50-40)^2 + (40-40)^2 + (40-40)^2\\+(30-40)^2 + (40-40)^2 + (50-40)^2\\+(60-40)^2\end{matrix}\right\}$$
$$= 120$$

分散は「値と平均の差の 2 乗」の総和をケース数 n で割ったものだ。それは、当該データの各ケースがどの程度中心傾向である平均値から離れているのかを（ケース数で割っているので）「平均的に」示す指標である。ケース数で割っているので、ケース数の異なるさまざまなデータにおいても、分散を用いて散らばりの程度が比較できる。分散が「（平均からの）ズレの 2 乗の平均」であること

は押さえておく必要がある。

2.3 値ごとのケース数をもとにした平均と分散

（1）データの形式

表 1 と同じデータは、表 2 のように値ごとのケース数として表されることもある。

表 2　得点ごとのケース数の表

得点(x_i)		ケース数(m_i)	
20	(x_1)	1	(m_1)
30	(x_2)	2	(m_2)
40	(x_3)	4	(m_3)
50	(x_4)	2	(m_4)
60	(x_5)	1	(m_5)
計		10	

iは得点のカテゴリー番号

（2）平均

x_1～x_t の値をとるものが、それぞれ m_1 個～m_t 個ある場合（全部で n 個）、その平均は次の式で表される。i はケース番号ではなくカテゴリー番号であることに注意する必要がある。ここを見誤るとわけがわからなくなる。

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{t}x_i m_i=\frac{1}{n}\left(x_1m_1+x_2m_2+\ldots+x_tm_t\right)$$

表の例では次のようになる。

$$\bar{x}=\frac{1}{10}\sum_{i=1}^{5}x_i m_i=\frac{1}{10}\begin{pmatrix}20\times1+30\times2\\+40\times4+50\times2\\+60\times1\end{pmatrix}=40$$

（3）分散

x_1～x_t の値をとるものが、それぞれ m_1 個～m_t 個ある場合（全部で n 個）、その分散は次の式で表される。

$$s^2=\frac{1}{n}\sum_{i=1}^{t}\left(x_i-\bar{x}\right)^2m_i$$
$$=\frac{1}{n}\begin{Bmatrix}\left(x_1-\bar{x}\right)^2m_1+\left(x_2-\bar{x}\right)^2m_2+\cdots\\+\left(x_t-\bar{x}\right)^2m_t\end{Bmatrix}$$

表の例で分散は次のようになる。

$$s^2=\frac{1}{10}\sum_{i=1}^{5}\left(x_i-\bar{x}\right)^2m_i$$
$$=\frac{1}{10}\begin{Bmatrix}(20-40)^2\,1+(30-40)^2\,2+(40-40)^2\,4\\+(50-40)^2\,2+(60-40)^2\,1\end{Bmatrix}$$
$$=120$$

2.4 値ごとの比率を用いた平均と分散

（1）データの形式

データは表 3 のように値ごとの比率で表されることもある。

表 3　得点ごとの比率の表

得点(x_i)		比率(p_i)	
20	(x_1)	0.1	(p_1)
30	(x_2)	0.2	(p_2)
40	(x_3)	0.4	(p_3)
50	(x_4)	0.2	(p_4)
60	(x_5)	0.1	(p_5)
計		1	

iは得点のカテゴリー番号

（2）平均

x_1～x_t の値をとるものについて、それぞれの個数が全体に占める比率として示されている場合、その比率が p_1～p_t（$\Sigma p_i=1$）ならば、値の平均は次の

式で表される。i はカテゴリー番号。

$$\bar{x} = \sum_{i=1}^{t} x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_t p_t$$

どうしてこのような式が成り立つかというとこうだ。2.3　(2) の「値ごとのケース数をもとにした平均」の式は、

$$\frac{1}{n}\sum_{i=1}^{t} x_i m_i$$

この 1/n はシグマの中に入れることができるから（補参照）、

$$\frac{1}{n}\sum_{i=1}^{t} x_i m_i = \sum_{i=1}^{t} x_i \frac{m_i}{n} = \sum_{i=1}^{t} x_i p_i$$

となり、結局上の式が成り立つ。

この式を用いて表 3 から平均を計算すると、次のようになる。

$$\bar{x} = \sum_{i=1}^{5} x_i p_i = 20\times 0.1 + 30\times 0.2 + 40\times 0.4 + 50\times 0.2 + 60\times 0.1 = 40$$

比率さえわかれば平均は出せる。ケース数は平均の算出には必ずしも必要ではないということが重要だ。平均は「値×比率」の総和で計算できるのである。

（3）分散

x_1〜x_t の値をとるものについて、それぞれの個数が全体に占める比率として示されている場合、その比率が p_1〜p_t（Σp_i=1）ならば、その分散は次の式で表される。

$$s^2 = \sum_{i=1}^{t} (x_i - \bar{x})^2 p_i = \left\{ \begin{matrix} (x_1 - \bar{x})^2 p_1 + (x_2 - \bar{x})^2 p_2 \\ + \cdots + (x_t - \bar{x})^2 p_t \end{matrix} \right\}$$

なぜなら

$$\frac{1}{n}\sum_{i=1}^{t} (x_i - \bar{x})^2 m_i = \sum_{i=1}^{t} (x_i - \bar{x})^2 \frac{m_i}{n} = \sum_{i=1}^{t} (x_i - \bar{x})^2 p_i$$

表の例では次のようになる。

$$s^2 = \sum_{i=1}^{5} (x_i - \bar{x})^2 p_i = \left\{ \begin{matrix} (20-40)^2 0.1 + (30-40)^2 0.2 + (40-40)^2 0.4 \\ + (50-40)^2 0.2 + (60-40)^2 0.1 \end{matrix} \right\} = 120$$

平均の場合と同様、比率さえわかればケース数がわからなくても分散は出せる。分散は「(値－平均) の 2 乗×比率」の総和で計算できるのである。

われわれは常識として「平均は、値の総和をケース数で割って出すものだ」と思いこんでいる。しかしながら、変数の平均や分散はケース数がわからなくても計算できる。平均や分散はケース数には必ずしも関係しない概念なのである。このことを押さえておこう。

3　確率変数と確率分布

3.1 確率変数と確率分布

変数がとりうる値のそれぞれについて確率が決まっていることがある。そういう変数のことを確率変数という。確率変数は X など、大文字のアルファベットで表すことが多い。

この確率変数について、とりうる値とその確率

の対応関係を示したものが確率分布である。すべての値に対する確率の計は 1 になる。表 4 は「さいころの目の数」という確率変数 X についての確率分布表である。このように値に対する確率がそれぞれ明らかになっているのが確率分布だ。

表 4　確率分布表

確率変数X（さいころの目）

値(x)	1	2	3	4	5	6	計
確率(p)	1/6	1/6	1/6	1/6	1/6	1/6	1

値がまったくでたらめに出るような変数は確率変数ではない。確率変数では値の出現率が確率的に決まっていなければならないのである。

3.2 確率と比率

ところで、この確率分布の表（表 4）は、前に見た比率の表（表 3）はととても似ている。実は比率の表は確率分布の表とも解釈できるのである。

比率の表はそれぞれの値をとるものの比率を示している。ここで、この集団から 1 人をランダムに取り出すことを考えてみよう。このとき、もとの値の比率はそのまま生じる値についての確率になる。たとえば、表 3 で 30 点の者の比率は 0.2 であるが、この集団から 1 人取り出してその者が 30 点である確率も 0.2 となる。このように見るとき、比率の表におけるテストの点数は確率変数となり、得点は値で、比率が確率ということになる。比率の分布を確率分布として考えることができるということは押さえておく必要がある。

3.3 確率変数の期待値と分散

ある確率変数Xについての確率分布があるとき、その確率分布の平均を X の期待値と言い、E(X)で表す。E(X)を μ で表すこともある。

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

分散はこの期待値をもとに次のように表せる。分散を σ^2 で表すこともある。

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

これらの期待値と分散は、比率を用いた表のところで述べた平均と分散の式と全く同じである。すなわち、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。

分散が「値と平均の差の 2 乗」の平均であることは 2.2 で述べた。そこから、分散は「値と平均の差の 2 乗」の期待値だと言うことができる。したがって、確率変数を用いて分散を表すと次のようになる。

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

確率変数の期待値や分散は、確率変数ではなく何らかの決まった値になるということに注意が必要だ。2.4 で比率のデータをもとに平均や分散の具体的な値を算出したが、それと同じで確率分布の下では平均や分散は必ず決まった値になる。このことはとても重要だ。覚えておこう。

3.4 確率変数の独立性

ある確率変数 X と別の確率変数 Y が独立かどうかということはとても重要だ。たとえば 1,2,3,4 と書かれた 4 枚のカードから 2 枚のカードを引くひ

くことを考えよう。最初に出るカードの数字を確率変数 X、次に出るカードの数字を確率変数 Y とする。

このとき、1 枚目のカードを戻して 2 枚目のカードを引くときには（復元抽出）、1 枚目に何を引いても 2 枚目で何が出るかに影響しない。しかし、最初のカードを戻さない場合（非復元抽出）、1 枚目に引かれたカードが何かによって、2 枚目のカードがどうなるかが変わってくる。1 枚目に 2 を引くと 2 枚目には 2 は決して出ないのである。

前者のように、確率変数 X がどのような値をとっても、そのことで Y の確率が影響されない場合、X と Y は独立であると言う。独立の場合、たとえば上の例で（1 枚目,2 枚目）が（1,2）となる確率は次のようになる。

$$\frac{1}{4}\times\frac{1}{4}=\frac{1}{16}$$

これは、1 枚カードを引く場合の 1 の出る確率と 2 の出る確率を掛け合わせたものだ（補参照）。しかし、独立でない場合、こんな計算では（1,2）の出る確率は求められない（このとき（1,2）の出る確率は樹形図を書くとわかるが 1/6 になる）。

3.5 確率変数の合成

確率変数をいくつか組み合わせて新たな確率変数を作ることができる。たとえば、上で述べた 2 枚のカードの数値の合計である X＋Y は確率変数の合成変数だ。表 5 は、X と Y が独立している場合に生じうる X＋Y の値の表だ。値は、2〜8 までをとる。

表 5 からは X＋Y が 2 になるところは 1 つしかないのでそうなる確率は 1/16 であることがわかる。同様に、3 になるところは 2 つあるのでその確率は 2/16、4 になる確率は 3/16、5 になる確率は 4/16、6 になる確率は 3/16、7 になる確率は 2/16、8 になる確率は 1/16 であることが表からわかる。確率分布表は次の表 6 のようになる。

表 5　X＋Y の値の表

		Yの値			
		1	2	3	4
Xの値	1	2	3	4	5
	2	3	4	5	6
	3	4	5	6	7
	4	5	6	7	8

表 6　X＋Y の確率分布表

確率変数X+Y

値	2	3	4	5	6	7	8	計
確率	1/16	2/16	3/16	4/16	3/16	2/16	1/16	1

この X＋Y という合成変数について、期待値と分散を求めると次のようになる。

$$E(X+Y)=2\left(\frac{1}{16}\right)+3\left(\frac{2}{16}\right)+4\left(\frac{3}{16}\right)+5\left(\frac{4}{16}\right)$$
$$+6\left(\frac{3}{16}\right)+7\left(\frac{2}{16}\right)+8\left(\frac{1}{16}\right)=\frac{80}{16}=5$$

$$V(X+Y)=(2-5)^2\left(\frac{1}{16}\right)+(3-5)^2\left(\frac{2}{16}\right)$$
$$+(4-5)^2\left(\frac{3}{16}\right)+(5-5)^2\left(\frac{4}{16}\right)+(6-5)^2\left(\frac{3}{16}\right)$$
$$+(7-5)^2\left(\frac{2}{16}\right)+(8-5)^2\left(\frac{1}{16}\right)=\frac{40}{16}=2.5$$

3.6 確率変数の重要公式

確率変数についての公式も知っておく必要がある。次の公式はとても重要だ。c は定数を意味している（詳しくは補を参考にしてほしい）。

■期待値（平均）関係

$$E(c)=c$$
$$E(X+c)=E(X)+c$$
$$E(cX)=cE(X)$$

■分散関係

$$V(c)=0$$
$$V(X+c)=V(X)$$
$$V(cX)=c^2V(X)$$

この最後の式には注意が必要だ。

■合成変数関係

$$E(X+Y)=E(X)+E(Y)$$
$$E(XY)=E(X)E(Y) \quad \text{（独立のとき）}$$
$$V(X+Y)=V(X)+V(Y) \quad \text{（独立のとき）}$$

上では X＋Y の確率分布表を作成して期待値や分散を計算したが、そんなことをしなくても、X と Y それぞれの期待値と分散さえわかれば、X＋Y の期待値や分散はわかる。

表 7　X と Y の確率分布表

確率変数X

値(x)	1	2	3	4	計
確率(p)	1/4	1/4	1/4	1/4	1

確率変数Y

値(y)	1	2	3	4	計
確率(p)	1/4	1/4	1/4	1/4	1

すなわち、表 7 をもとに、

$$E(X)=E(Y)=2.5$$
$$V(X)=V(Y)=1.25$$

を計算し、合成変数の公式を使うと X＋Y の期待値や分散は簡単に計算できたのである。すなわち、

$$E(X+Y)=E(X)+E(Y)=2.5+2.5=5$$
$$V(X+Y)=V(X)+V(Y)=1.25+1.25=2.5$$

4　母集団と標本

4.1 推測統計についての常識的誤解

さて、準備が整ったので母集団と標本について述べていくことにしよう。母集団や標本、そして推測統計について、読者はどう考えているだろうか。

まず、母集団についてこう考えていないだろうか。「推測統計で母集団というのはたとえば日本の有権者というようにサイズの決まったものである」「母集団の平均や分散は知られていないから、その値は決まった値というよりも、確率的なものだ」。

標本について、こう考えていないだろうか。「標本分布とは、具体的な標本での変数の値の分布であり、それをもとにして標本平均や標本分散が計算できる」。

さらに、推測統計についてこう考えていないだろうか。「推測統計は、わかっている標本分布からわからない母集団の分布を推測するわけだから、いつも標本→母集団という道筋で考えていく」。

これらは常識的な観点からすると正しそうに思える。しかし「基礎的な推測統計」の観点からするとどれも普通の考え方ではない。それゆえ、こ

ういった常識をもとにテキストを読んでいくと「わけがわからない」という事態が生じてしまうのである。母集団、標本、推測統計の原理について順に論じていこう。

4.2 母集団

母集団という言葉を聞くと、われわれはすぐ人の集まりのような気がする。しかし、本来はある変数の値の集まりである。そして、基礎的な推測統計においては、せいぜい 1 つか 2 つの変数しか問題にしない。だから、たとえば、大学生という母集団ではなく、大学生の身長、あるいは大学生の身長と体重という母集団があるということになる。

さて、こういった母集団について、基礎的な推測統計では 2 つの仮定を置いて議論をスタートさせる。それらは（1）母集団の変数は確率変数、(2) 母集団は無限母集団ということだ。

母集団の変数が確率変数であるということに違和感を持つ読者もいるだろう。しかし、そう考えなければさまざまな混乱が生じる。推測統計のテキストではしばしば「さいころの目」などを例として議論を進める。この「さいころの目」の確率分布が母集団の分布であり、標本はそのさいころを振ったときに出る数を意味している。2 回振るとき、標本のサイズは 2 となる。こういった議論を「大学生の身長」などの議論に整合的につなげるためには、母集団の「大学生の身長」といった変数も確率変数と考える必要がある。それは難しいことではなく、母集団を比率からとらえるというだけのことだ。すなわち、2.4 の「値ごとの比率を用いた平均と分散」のデータのようなものとして母集団を考えるのである。このとき、比率はその母集団から 1 ケースを選ぶときの確率に対応する。

母集団の変数が確率変数であったとしても母集団の平均や分散は決まった値になることには注意が必要だ。2.4 の「値ごとの比率を用いた平均と分散」のところで、比率のデータから決まった値の平均と分散が計算できた。それと同じで、確率変数の平均や分散は決まった値になる。

次に無限母集団について。母集団を無限母集団とみなすことの眼目は、標本抽出過程で母集団の分布は変わらないということだ。実際に母集団の大きさが無限である必要はない。そうではなく分布が変わらないことが重要なのである。だから、標本対象を 1 つ抽出しては返し、また次を抽出するという復元抽出の方法をとる場合、たとえ母集団の大きさが有限でも無限母集団の性質を持っていると言える。復元抽出するとき 10 人の（身長の）母集団から 100 人の（身長の）標本を抽出することも可能だ。

母集団を無限母集団と考えるのは「基礎的な推測統計」に限定された話である。さらに進んだ統計の世界では有限母集団を取り扱うことも普通にある。しかしながら、常識的な有限母集団の考えを持ったまま推測統計の基礎を学ぶとわけのわからない状態になってしまう。

4.3 標本

母集団から抽出される母集団の一部分を標本と言う。基礎的な推測統計で標本について忘れてはならないのはこの用語が 2 つの意味で用いられるということだ。第 1 は、母集団から現実に抽出される標本という意味、第 2 は、母集団からどんな標本が抽出されるのだろうと理論的に考えるときの標本という意味である。ここではそれぞれを現実の標本、理論的標本と呼ぶことにしたい。

標本のサイズを n とする場合、現実の標本は母集団から抽出された n 個の値の集合である。1 つの変数（たとえば大学生の身長）を問題としてい

る場合、現実の標本の値の分布は 1 次元の軸上に並ぶ n 個の値であり、この値をもとに平均や分散などが計算できる。

一方、理論的標本とは母集団から抽出される n 個の独立した確率変数の集合とみなされる。n 個はすべて同じ無限母集団から抽出されるので、それぞれの確率変数の確率分布は母集団の確率分布と同じになる。また、n 個それぞれが 1 つの確率分布に従うので、標本全体としては n 次元の確率分布をする。

理論的標本については少しわかりにくいと思うので例で説明しよう。1、2、3 が等確率で出るルーレットがあるとし、出る数字を X という確率変数で表そう。この X の分布が母集団の確率分布となる。母集団の確率分布は次の表 8 のようになる。

表 8　X の確率分布表

確率変数X（ルーレットの数）

値	1	2	3	計
確率	1/3	1/3	1/3	1

このルーレットを 2 回し、1 回目を X_1、2 回目を X_2 とすると、標本は（X_1, X_2）という確率変数のペアと考えられ、その確率分布の値は（1,1），(1,2)，(1,3）…というように 2 次元となる。

表 9　２次元確率変数（X,Y）の確率分布表

		X_2			
		1	2	3	計
X_1	1	(1,1) 1/9	(1,2) 1/9	(1,3) 1/9	1/3
	2	(2,1) 1/9	(2,2) 1/9	(2,3) 1/9	1/3
	3	(3,1) 1/9	(3,2) 1/9	(3,3) 1/9	1/3
	計	1/3	1/3	1/3	1

上段：値 、下段：確率

表 9 はこの 2 次元の確率分布についての確率分布表である。たとえば、(1,3)の下にある 1/9 は(1,3)となる確率、すなわち、1 回目に 1、2 回目に 3 の出る確率を意味している。

ルーレットを n 回し、その確率変数を $X_1,X_2,\cdots,X_n$ とするとき、標本は（$X_1,X_2,\cdots,X_n$）という集合になり、その確率分布は n 次元になる。

ここで、標本の確率分布と標本をもとに作った合成変数の確率分布は異なることに注意が必要だ。標本サイズを 2 とするとき標本の確率分布は表 9 のように 2 次元確率分布になるが、X_1+X_2 といった合成変数の確率分布は表 6 で見たように 1 次元の確率分布になる。標本が（$X_1,X_2,\cdots,X_n$）と n 次元であっても合成変数である $X_1+ X_2+\cdots+X_n$ は 1 次元の確率変数なのである。

推測統計では、現実の標本ではなく確率変数の集合である理論的標本をもとに議論を展開する。標本という言葉から現実の標本しか思い浮かべないなら、落とし穴に落ちること必定である。

4.4 母数と標本統計量

（1）確率変数の期待値（平均）と分散

母集団についての平均や分散などを母数と言う。一方、標本の平均や分散は標本統計量と呼ばれる。母数と標本統計量については混乱が生じやすいのできちんと整理しておこう。

確率変数 X の期待値（平均）と分散の一般式は次のようになる（x_i は値、p_i は確率、t は値のとりうる範囲）。

■期待値（平均）と分散の一般式（離散型）

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

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

これらは確率変数ではなく決まった値になる。X が確率分布をしていても、その期待値や分散は確率変数ではなく決まった値になるのである。

（2）母平均と母分散

母集団と標本という観点をとって、母集団を確率変数 X の分布と見るとき、X の平均を母平均 μ と呼び、X の分散を母分散 σ^2 と呼ぶ。それらは次のようになる。

■母平均と母分散

$$\mu = E(X)$$
$$\sigma^2 = V(X)$$

これらは、上の「期待値（平均）と分散の一般式」から求められ、確率変数ではなく決まった値になる。母集団を確率分布ととらえても、その平均や分散は確率変数ではなく決まった値になるからだ。

上のルーレットにおいて、母平均と母分散は次のようになる。

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

（3）現実の標本の平均と分散

現実の標本の平均や分散は、現実の標本において実際に出た値から計算する。サイズ n の現実の標本での平均と分散は次の式で表される（x_i はそれぞれのケースの得点）。

■現実の標本の平均と分散

$$\bar{x} = \sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$
$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

現実の標本の平均や分散はもちろん決まった値になる。上のルーレットの話で、現実に 1 と 3 が出れば、

$$\bar{x} = \frac{1}{2}(1+3) = 2$$
$$s^2 = \frac{1}{2}\left((1-2)^2 + (3-2)^2\right) = 1$$

となる。

（4）標本平均と標本分散

理論的標本の平均、分散のことを標本平均、標本分散と言う。サイズ n の標本についての標本平均と標本分散は、次の式で表される。

■標本平均と標本分散

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$
$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

これらは確率変数であることに注意が必要である。

上のルーレットの場合、標本平均と標本分散は次のようになる。

$$\bar{X} = \frac{1}{2}(X_1 + X_2)$$
$$S^2 = \frac{1}{2}\left((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2\right)$$

理論的標本は確率変数（X_1～X_n）の集合だが、標本平均や標本分散などはそこにある確率変数を合成して作られた新たな変数である。そして合成された変数である標本平均や標本分散はそれ自体確率変数である。

(5) 注意すべきこと

母平均、母分散などの母数は決まった値であり、標本平均や標本分散などの標本統計量は確率変数である。ここが一番重要なところである。

統計学のテキストには「標本統計量の分布」という表現が見られ、ここでわけがわからなくなる学生がいる。現実の標本の平均や分散は分布を持たない 1 つの値だから、それしか念頭に置いていないと「標本統計量の分布」という表現はわけのわからないものになるのである。理論的標本において、標本は確率変数の集合であり、標本平均や標本分散はそこから合成された確率変数であり、決まった値を持つものではないということを強調しておこう。

4.5 基礎的な推測統計の原理

推測統計は現実の標本から母集団の状況を推測するためにある。すなわち標本→母集団という方向で推測するのである。しかしながら、この矢印の方向だけを信じていると、基礎的な推測統計は理解できない。というのは標本→母集団の推測を成り立たせているのは母集団→標本の原理だからだ。

わかりやすく言うとこうなる。基礎的な推測統計では「○○といった母集団があるとするならば××という標本が生じる」という定理をまず明らかにし、この定理と「△△といった標本が得られた」という事実の両者から「母集団は○○でないかもしれない」などと推論するものだ。すなわち、全体としての推測は標本→母集団という方向性を持っているのだが、用いられる定理は母集団→標本という方向性を持っているのだ。

基礎的な推測統計はこういった定理がなぜ成り立つのかを議論するところから始まる。だから、定理を理解する上では母集団→標本という方向で考えていかなければならない。そうしてはじめて推測統計の基礎はマスターできるのである。

以上で基本的な説明は終わりである。以下では今まで述べてきたことをもとに、初学者が躓きやすい問題について述べていく。

5　標本平均の分布

「標本平均の平均は母平均であり、標本平均の分散は母分散を標本サイズで割ったものだ」。これは推測統計で一番大事なことと言っていい。しかし、学生時代に統計学の勉強をしていて最初にわからなくなったのは実にこの一番大事な話だった。そして、そこからどんどんわけがわからない世界に突入していったのだった。読者の中には同じ経験をしている者も多いのではないだろうか。

「平均の平均」や「平均の分散」という言葉が理解できなかったのも仕方がなかったと今は思う。これらの概念の理解には、確率変数の期待値、確率変数の分散、確率変数の合成の知識が前提とされるが、筆者はそれらについてあまりにも無知だったからである。

読者がここまで述べてきた内容を理解しているならば、「平均の平均」や「平均の分散」については少しの説明だけで理解できるはずだ。

まず、何が問題とされているのかをきちんと知っておく必要がある。ここでの問題は、母集団から抽出された標本についてその平均がどのようなものでありうるのかということである。順を追っ

て説明していこう。

まず、母集団の変数 X が確率分布をし、母平均が μ、母分散が σ^2 だとする。

$$\mu = E(X)$$
$$\sigma^2 = V(X)$$

そこからサイズ n の標本を復元抽出で取り出すことを考えると、標本は X_1〜X_n の確率変数の集合になる。そしてこれらの確率変数をもとに、平均 $\overline{X}$ という新しい確率変数を作ることができる。

$$\overline{X} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)$$

この平均 $\overline{X}$ は確率変数の合成変数である。「平均の平均」や「平均の分散」という言葉が述べているのは、この「合成された確率変数としての平均 $\overline{X}$」の期待値（平均）と分散のことである。そして、次のような関係が成り立つと述べているのである。

$$E\left(\overline{X}\right) = \mu$$
$$V\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

すなわち、標本平均 $\overline{X}$ の期待値（平均） $E\left(\overline{X}\right)$ は母平均 μ に等しくなり、標本平均の分散 $V\left(\overline{X}\right)$ は母分散 σ^2 を標本サイズ n で割ったものになるというのが話の結論である。

平均という言葉が、1 つの確率変数の平均（期待値）という意味と、複数の確率変数の合成変数としての平均という意味の両者で用いられていることには注意が必要だ。整理すると次のようになる（表 10）。

表 10　母平均・標本平均・標本平均の平均

$E(X)$	母平均（期待値）	決まった値
$\overline{X}$	標本平均	確率変数
$E(\overline{X})$	標本平均の平均（期待値）	決まった値

ここで言われていることの意味はまだわかりにくいかもしれない。例をもとに具体的なイメージをつかみ、母集団と標本において本当にそんな関係が成り立つのかを見てみよう。

2 か 6 のいずれかが等確率で出るルーレットがあるとする（2 と 6 にしたのはこの値が計算に都合よいからで、別の値でもまったくかまわない）。このルーレットの数を値とする確率変数 X の確率分布表は表 11 のようになる。

表 11　X の確率分布表

確率変数X（ルーレットの数）

値	2	6	計
確率	1/2	1/2	1

この分布は母集団の分布と考えられ、母平均と母分散は次のようになる。

$$\mu = E(X) = 2\left(\frac{1}{2}\right) + 6\left(\frac{1}{2}\right) = 4$$
$$\sigma^2 = V(X) = (2-4)^2\left(\frac{1}{2}\right) + (6-4)^2\left(\frac{1}{2}\right) = 4$$

すなわち、母平均と母分散はともに 4 である。

ここで、このルーレットを 4 回したとき生じうる結果を示したものが表 12 である。この表は生じうるすべての可能性を網羅しており、結果は a から p までの 16 パターンの内のどれかに必ずなる。

表 12　各標本の標本平均

標本記号	1回目	2回目	3回目	4回目	標本平均
a	2	2	2	2	2
b	2	2	2	6	3
c	2	2	6	2	3
d	2	2	6	6	4
e	2	6	2	2	3
f	2	6	2	6	4
g	2	6	6	2	4
h	2	6	6	6	5
i	6	2	2	2	3
j	6	2	2	6	4
k	6	2	6	2	4
l	6	2	6	6	5
m	6	6	2	2	4
n	6	6	2	6	5
o	6	6	6	2	5
p	6	6	6	6	6

これら a〜p はそれぞれもとの母集団から選ばれうるサイズ 4 の標本と考えられる。a〜p のどの標本も、生じる確率は 1/16 である。

この表の標本平均の列に注目すると、標本平均の値が 2 や 6 になるのはそれぞれ 1 回、3 や 5 になるのはそれぞれ 4 回、4 になるのは 6 回であることがわかる。確率で言いなおすと標本平均が 2 や 6 になる確率はそれぞれ 1/16、3 や 5 になる確率はそれぞれ 4/16、4 になる確率は 6/16 ということになる。これを確率分布表で示すと表 13 のようになる。

表 13　標本平均の確率分布表

確率変数 $\overline{X}$

値	2	3	4	5	6	計
確率	1/16	4/16	6/16	4/16	1/16	1

この確率分布の表をもとに標本平均の期待値（平均）と標本平均の分散を計算すると次のようになる。

$$E(\overline{X}) = 2\left(\frac{1}{16}\right) + 3\left(\frac{4}{16}\right) + 4\left(\frac{6}{16}\right) + 5\left(\frac{4}{16}\right) + 6\left(\frac{1}{16}\right) = 4$$

$$V(\overline{X}) = (2-4)^2\left(\frac{1}{16}\right) + (3-4)^2\left(\frac{4}{16}\right) + (4-4)^2\left(\frac{6}{16}\right) + (5-4)^2\left(\frac{4}{16}\right) + (6-4)^2\left(\frac{1}{16}\right) = 1$$

この結果を見ると、なるほど次のことが成立していることがわかる。

$$\mu = 4 = E(\overline{X})$$
$$\frac{\sigma^2}{n} = \frac{4}{4} = 1 = V(\overline{X})$$

すなわち、標本平均の期待値（平均）は母平均に等しくなっており、標本平均の分散は母分散を標本サイズ n で割ったものになっているのである。これが、「標本平均の平均は…」で言いたかったことなのだ。

こういったことはこの例に限らずいつでも成り立つ。標本平均の期待値と母平均の間には次のような関係があるからだ。

$$E(\overline{X}) = E\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n}\left[E(X_1) + E(X_2) + \cdots + E(X_n)\right] = \frac{1}{n}\left[\underbrace{\mu + \mu + \cdots + \mu}_{n}\right] = \frac{1}{n}n\mu = \mu$$

この証明を理解するためには、母集団における X の確率分布と標本を構成する X_1〜X_n それぞれの確

率分布がすべて同じ確率分布であることに気づくとともに、次の 2 つの公式を思い出す必要がある（補を参照）。

$$E(cX)=cE(X)$$
$$E(X+Y)=E(X)+E(Y)$$

標本平均の分散については次のようになる。

$$V(\bar{X})=V\left[\frac{1}{n}(X_1+X_2+\cdots+X_n)\right]$$
$$=\frac{1}{n^2}V[X_1+X_2+\cdots+X_n]$$
$$=\frac{1}{n^2}[V(X_1)+V(X_2)+\cdots+V(X_n)]$$
$$=\frac{1}{n^2}\left[\underbrace{\sigma^2+\sigma^2+\cdots+\sigma^2}_{n}\right]$$
$$=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n}$$

こちらでは、母集団における X の確率分布と標本を構成する X_1～X_n それぞれの確率分布がすべて同じ確率分布であることに加え、X_1～X_n は独立だということに気づく必要がある。そして、次の 2 つの公式を思い出す必要もある（補を参照）。

$$V(X+Y)=V(X)+V(Y)\quad\text{（独立のとき）}$$
$$V(cX)=c^2V(X)$$

ここで明らかになった、次の 2 つの式は推測統計で最も大切な式と言ってもいい。もう一度書いておく。

$$E(\bar{X})=\mu\qquad V(\bar{X})=\frac{\sigma^2}{n}$$

これらは、復元抽出で（あるいは無限母集団から）取りだされた独立した $X_1,X_2,\cdots,X_n$ からなる標本と母集団の間で成り立つ関係である。

6　２種類の分散と不偏推定量

本稿の最初の部分において、分散は次の式で定義されている。

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

しかし、多くのテキストにおいて、分散は次の式で定義されている。

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

2 つの式を見ると、総和を n で割るのか n−1 で割るのかの違いがあることがわかる。この 2 つの分散はどう違うのか。これもまた初学者を悩ませる問題だ。ここでは標本の分散について、n で割ったものを標本分散（S^2）、n−1 で割ったものを不偏分散(U^2)と呼び、それらの性質について見ていく。

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$
$$U^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

上の式は、サイズ n の理論的標本の標本分散と不偏分散を表している。ここで S^2 や U^2 というように大文字を使うのは、それらが確率変数であることを強調するためである。U^2 をなぜ不偏分散と言うのかは後に明らかになる。

前節で使った 2 と 6 の出るルーレットをここで

も使おう。それぞれの標本について、標本平均だけでなく標本分散と不偏分散も計算すると表 14 のようになる。

表 14　各標本の標本分散と不偏分散

標本記号	1回目	2回目	3回目	4回目	標本平均 $\bar{X}$	標本分散 S^2	不偏分散 U^2
a	2	2	2	2	2	0	0
b	2	2	2	6	3	3	4
c	2	2	6	2	3	3	4
d	2	2	6	6	4	4	16/3
e	2	6	2	2	3	3	4
f	2	6	2	6	4	4	16/3
g	2	6	6	2	4	4	16/3
h	2	6	6	6	5	3	4
i	6	2	2	2	3	3	4
j	6	2	2	6	4	4	16/3
k	6	2	6	2	4	4	16/3
l	6	2	6	6	5	3	4
m	6	6	2	2	4	4	16/3
n	6	6	2	6	5	3	4
o	6	6	6	2	5	3	4
p	6	6	6	6	6	0	0

この表の標本平均の列について、期待値（平均）は母平均μに等しくなることはすでに見た。

$$E\left(\overline{X}\right)=4=\mu$$

ここで、標本分散の期待値がどうなるかを考えてみよう。標本分散の列を確率分布表にまとめると表 15 のようになる。

表 15　標本分散の確率分布表

確率変数S^2

値	0	3	4	計
確率	2/16	8/16	6/16	1

この表 15 をもとに標本分散の期待値を出すと次のようになる。

$$E\left(S^2\right)=0\left(\frac{2}{16}\right)+3\left(\frac{8}{16}\right)+4\left(\frac{6}{16}\right)=3$$

すでに明らかにしたように、母分散σ^2は 4 であった。ここから、標本分散 S^2 の期待値は母分散σ^2と等しくないことがわかる。

$$E\left(S^2\right)\neq\sigma^2$$

一方、不偏分散についてはどうか。不偏分散の列を確率分布表にまとめると表 16 のようになる。

表 16　不偏分散の確率分布表

確率変数U^2

値	0	4	16/3	計
確率	2/16	8/16	6/16	1

ここから不偏分散 U^2 の期待値は次のように計算できる。

$$E\left(U^2\right)=0\left(\frac{2}{16}\right)+4\left(\frac{8}{16}\right)+\left(\frac{16}{3}\right)\left(\frac{6}{16}\right)=4$$

不偏分散 U^2 の期待値は母分散σ^2と等しくなっているのである。

$$E\left(U^2\right)=4=\sigma^2$$

標本平均や標本分散などの標本統計量から母平均や母分散などの母数を推定する場合、標本統計量の持つ性質が問題とされる。そして、標本統計量の期待値（平均）が母数に等しくなるような性

質は望ましいとされ、そのような性質を持つ標本統計量は母数の不偏推定量と言われる。なぜ望ましいのかというと、現実の標本において計算された平均や分散などで母数を推定すると間違いが生じるが、その推定は「平均的には」母数に等しくなる、という特徴を持っているからである（だから「不偏」なのである）。

標本平均は母平均の不偏推定量である。n−1 で割った不偏分散も母分散の不偏推定量である。しかしながら n で割った標本分散は母分散の不偏推定量ではない。

$$E\left(\overline{X}\right)=\mu \qquad E\left(U^2\right)=\sigma^2 \qquad E\left(S^2\right)\neq\sigma^2$$

わかりにくいのは、母分散自体は n（n は母集団の大きさ）で割った分散の式（正確に言えば、この式に対応する比率を用いた分散の式）で求められるのに対し、この母分散を推定するのに n−1（n は標本サイズ）で割った分散を用いるという点である。集団の分散そのものを計算する場合は n で割る式、その集団を標本と考えそこから母分散を推測する場合は n − 1 で割る式を使うと覚えておくといい。

不偏分散の期待値が母分散に等しくなるというのは、復元抽出で（あるいは無限母集団から）取りだされた独立した $X_1,X_2,\cdots,X_n$ からなる標本においての話である。非復元抽出ではこのことは成り立たない。

最後に、標本分散の期待値がどうなるのかということについて述べておこう。

$$S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

$$U^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

したがって、不偏分散に n−1 をかけて n で割ると標本分散になる。上で見たように、

$$E\left(U^2\right)=\sigma^2$$

だから、標本分散の期待値は次のようになる。

$$E\left(S^2\right)=\frac{n-1}{n}\sigma^2$$

この標本分散の期待値に関わる最後の式について証明しておく。次のよく使う式を利用する（この証明は「補」にある）。

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

■証明

$$E\left(S^2\right)=E\left(\frac{1}{n}\Sigma\left(X_i-\overline{X}\right)^2\right)$$

$$=\frac{1}{n}E\left(\Sigma\left({X_i}^2-2X_i\overline{X}+\overline{X}^2\right)\right)$$

$$=\frac{1}{n}E\left(\Sigma{X_i}^2-\Sigma2X_i\overline{X}+\Sigma\overline{X}^2\right)$$

$$=\frac{1}{n}E\left(\Sigma{X_i}^2-2\overline{X}\Sigma X_i+\Sigma\overline{X}^2\right)$$

$$=\frac{1}{n}E\left(\Sigma{X_i}^2-2\overline{X}n\overline{X}+n\overline{X}^2\right)$$

$$=\frac{1}{n}E\left(\Sigma{X_i}^2-2n\overline{X}^2+n\overline{X}^2\right)$$

$$=\frac{1}{n}E\left(\Sigma{X_i}^2-n\overline{X}^2\right)$$

$$=\frac{1}{n}\left(E\left(\Sigma{X_i}^2\right)-E\left(n\overline{X}^2\right)\right)$$

$$=\frac{1}{n}\Sigma\left(E\left({X_i}^2\right)\right)-E\left(\overline{X}^2\right)$$

ここで「よく使う式」を使って

$$= \frac{1}{n}\Sigma\left(V(X)+(E(X))^2\right)-\left(V\left(\bar{X}\right)+\left(E\left(\bar{X}\right)\right)^2\right)$$
$$= \frac{1}{n}\Sigma\left(\sigma^2+\mu^2\right)-\left(\frac{1}{n}\sigma^2+\mu^2\right)$$
$$= \frac{1}{n}\left(\Sigma\sigma^2+\Sigma\mu^2\right)-\left(\frac{1}{n}\sigma^2+\mu^2\right)$$
$$= \frac{1}{n}\left(n\sigma^2+n\mu^2\right)-\left(\frac{1}{n}\sigma^2+\mu^2\right)$$
$$= \sigma^2+\mu^2-\left(\frac{1}{n}\sigma^2+\mu^2\right)=\sigma^2-\frac{1}{n}\sigma^2$$
$$= \frac{n-1}{n}\sigma^2$$

7　おわりに

以上で母集団・標本・確率変数についての話は終わる。今までのもやもや感が解消され、すっきりした気分になった読者がいればうれしいのだが、学生時代の自分のことを考えると、なかなかそう簡単ではないのかもしれない。細かいことはいいが、基礎的な推測統計のイメージだけはきちんとつかんでおいてほしい。

とりわけ重要なのは、確率変数と決まった値をきちんと区別することだ。母集団は確率分布としてとらえられているが、母数は決まった値であること、標本には理論的標本と現実の標本があり、理論的標本は確率変数の集合であること、標本統計量は確率変数の合成変数であり、なんらかの確率分布に従うことなどを知っておく必要がある。また、この認識をもとに、標本統計量と母数についての「標本平均の平均は母平均」という表現や「標本平均の分散は母分散を標本サイズで割ったもの」という表現の意味も理解しておく必要がある。

今回は推測統計の最も基本的なところに限定して話をすすめてきた。そこでは標本は、無限母集団からのものか、有限の母集団から復元抽出で得られるものとされている。しかし、現実の調査での標本は、サンプリングの仕方にもよるが、有限母集団から非復元抽出で選ばれていると考えられるものもある。そういったデータについては今回説明してきたことは役に立たないと思うかもしれないが、それは違う。今回述べた推測統計の原理はそういったデータの分析においても基礎となる考え方なのである。この問題については本シリーズの最後で説明したいと思う。

文献

本稿で扱った内容だけでなく、基礎から推測統計をきちんと勉強するためには、次の文献にあたるのがいいと思う。

■高校数学の復習として
馬場敬之ほか, 2003-2004『初めから始める数学 1〜数学 C』（全 6 冊）マセマ

■大学数学の基礎として
馬場敬之・高杉豊, 2003『微分積分キャンパスゼミ』マセマ
馬場敬之・高杉豊, 2003『線形代数キャンパスゼミ』マセマ

■統計学
小寺平治,2002『ゼロから学ぶ統計解析』講談社
馬場敬之・久池井茂, 2003『確率統計キャンパスゼミ』マセマ

補：シグマ・期待値・分散の公式

1　シグマ

1.1 高校時代に習ったシグマ

高校では下のようにシグマを使って数列の和を表現した。

$$\sum_{i=1}^{n} i = 1+2+\cdots+n = \frac{1}{2}n(n+1)$$
$$\sum_{i=1}^{n} i^2 = 1^2+2^2+\cdots+n^2 = \frac{1}{6}n(n+1)(2n+1)$$
$$\sum_{i=1}^{n} i^3 = 1^3+2^3+\cdots+n^3 = \left\{\frac{1}{2}n(n+1)\right\}^2$$

一方、統計学で扱うのは次のような形のシグマである。i の場所に注意して混乱しないように。

$$\sum_{i=1}^{n} x_i = x_1+x_2+\cdots+x_n$$

1.2 項が 1 つの場合

$$\boxed{\sum_{i=1}^{n} x_i = x_1+x_2+\cdots+x_n}$$

これは箱が n 個あって、そこに x_1〜x_n が入っていて、それを全部足すと考えるといい。

$$\boxed{\sum_{i=1}^{n} 1 = 1+1+\cdots+1 = n}$$

これは次のように考えるといい。

$x_1 = 1,\quad x_2 = 1,\quad \cdots,\quad x_n = 1$

n 個の箱には全部 1 が入っているから全部足すと n。

$$\boxed{\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i}$$

添え字に依存しない文字は外に出せる。

なぜなら、

$$\sum_{i=1}^{n} cx_i = cx_1+cx_2+\cdots+cx_n$$
$$= c(x_1+x_2+\cdots+x_n) = c\sum_{i=1}^{n} x_i$$

$$\boxed{\sum_{i=1}^{n} c = nc}$$

なぜなら

$$\sum_{i=1}^{n} c = c\sum_{i=1}^{n} 1 = c(1+1+\cdots+1) = cn = nc$$

n 個の箱の中に c が入っていると考えるといい。

1.3 項が 2 つ以上の場合

$$\boxed{\sum_{i=1}^{n}(x_i+y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

なぜなら

$$\sum_{i=1}^{n}(x_i+y_i) = (x_1+y_1)+(x_2+y_2)+\cdots+(x_n+y_n)$$
$$= (x_1+x_2+\cdots+x_n)+(y_1+y_2+\cdots+y_n)$$
$$= \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

分配できるのである。

$$\boxed{\sum_{i=1}^{n}(x_i+c) = \sum_{i=1}^{n} x_i + nc}$$

なぜなら

$$\sum_{i=1}^{n}(x_i+c) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} c = \sum_{i=1}^{n} x_i + nc$$

たとえば

$$\sum_{i=1}^{n}(x_i+3)=(x_1+3)+(x_2+3)\cdots+(x_n+3)$$

$$=\sum_{i=1}^{n}x_i+3n$$

1.4 二重のシグマ

（1）1 変数で添え字が 2 つの場合

$$\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}=x_{11}+\cdots+x_{1j}+\cdots+x_{1n}$$
$$+x_{21}+\cdots+x_{2j}+\cdots+x_{2n}$$
$$+\cdots$$
$$+x_{i1}+\cdots+x_{ij}+\cdots+x_{in}$$
$$+\cdots$$
$$+x_{m1}+\cdots+x_{mj}+\cdots+x_{mn}$$

m×n 個の箱に入っている x_{ij} を全部足す。

（2）異なる添え字の 2 変数の掛け算

$$\sum_{i=1}^{m}\sum_{j=1}^{n}x_iy_j=\sum_{i=1}^{m}x_i\sum_{j=1}^{n}y_j=\sum_{j=1}^{n}y_j\sum_{i=1}^{m}x_i$$

なぜなら

$$\sum_{i=1}^{m}\sum_{j=1}^{n}x_iy_j=(x_1y_1+x_1y_2+\cdots+x_my_n)$$
$$=x_1(y_1+y_2+\cdots+y_n)$$
$$+x_2(y_1+y_2+\cdots+y_n)$$
$$+\cdots+x_m(y_1+y_2+\cdots+y_n)$$
$$=(x_1+x_2+\cdots+x_n)(y_1+y_2+\cdots+y_n)$$
$$=\sum_{i=1}^{m}x_i\sum_{j=1}^{n}y_j$$

添え字に依存しない文字は外に出せるのである。

■例：m=2,　n=3

$x_1=1,\ x_2=3,\quad y_1=2,\ y_2=5,\ y_3=4$

$$\sum_{i=1}^{2}\sum_{j=1}^{3}x_iy_j$$
$$=x_1y_1+x_1y_2+x_1y_3+x_2y_1+x_2y_2+x_2y_3$$
$$=1\times2+1\times5+1\times4+3\times2+3\times5+3\times4$$
$$=2+5+4+6+15+12=44$$
$$\sum_{i=1}^{2}x_i\sum_{j=1}^{3}y_j=(x_1+x_2)(y_1+y_2+y_3)$$
$$=(1+3)(2+5+4)=4\times11=44$$

（3）異なる添え字の 2 変数の足し算

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(x_i+y_j)$$
$$=\sum_{i=1}^{m}\sum_{j=1}^{n}x_i+\sum_{i=1}^{m}\sum_{j=1}^{n}y_j$$
$$=n\sum_{i=1}^{m}x_i+m\sum_{j=1}^{n}y_j$$

■例：m=2,　n=3

$x_1=1,\ x_2=3,\quad y_1=2,\ y_2=5,\ y_3=4$

$$\sum_{i=1}^{2}\sum_{j=1}^{3}(x_i+y_j)$$
$$=(x_1+y_1)+(x_1+y_2)+(x_1+y_3)$$
$$+(x_2+y_1)+(x_2+y_2)+(x_2+y_3)$$
$$=(1+2)+(1+5)+(1+4)$$
$$+(3+2)+(3+5)+(3+4)$$
$$=3+6+5+5+8+7=34$$
$$\sum_{i=1}^{2}\sum_{j=1}^{3}(x_i+y_j)=\sum_{i=1}^{2}\sum_{j=1}^{3}x_i+\sum_{i=1}^{2}\sum_{j=1}^{3}y_j$$
$$=\sum_{i=1}^{2}x_i\sum_{j=1}^{3}1+\sum_{i=1}^{2}1\sum_{j=1}^{3}y_j$$
$$=3\sum_{i=1}^{2}x_i+2\sum_{j=1}^{3}y_j$$
$$=3\times(1+3)+2\times(2+5+4)$$
$$=3\times4+2\times11=12+22=34$$

（4）添え字のついた変数がない場合

$$\boxed{\sum_{i=1}^{m}\sum_{j=1}^{n}1=1+1+\cdots+1=mn}$$

これは、m×n 個の箱に入っている 1 を足すということ。

$$\sum_{i=1}^{2}\sum_{j=1}^{3}c=c+c+c+c+c+c$$

これは、6 個の箱に入っている c を足すということ。

$$\boxed{\begin{aligned}&\sum_{i=1}^{m}\sum_{j=1}^{n}x_i\\&=\sum_{i=1}^{m}\sum_{j=1}^{n}x_i\cdot 1=\sum_{i=1}^{m}x_i\sum_{j=1}^{n}1=\sum_{i=1}^{m}x_i n\\&=n\sum_{i=1}^{m}x_i\end{aligned}}$$

$$\boxed{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_i+c\right)=\sum_{i=1}^{m}\sum_{j=1}^{n}x_i+\sum_{i=1}^{m}\sum_{j=1}^{n}c=n\sum_{i=1}^{m}x_i+mnc}$$

なぜなら

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_i+c\right)=\sum_{i=1}^{m}\sum_{j=1}^{n}x_i+\sum_{i=1}^{m}\sum_{j=1}^{n}c$$

$$=\sum_{i=1}^{m}x_i\sum_{j=1}^{n}1+c\sum_{i=1}^{m}\sum_{j=1}^{n}1=n\sum_{i=1}^{m}x_i+mnc$$

2　確率変数の期待値と分散

2.1 一般式

$$E\left(X\right)=\mu$$

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

2.2 離散型確率変数

（1）分布

x	x_1	x_2	$\cdots$	x_i	$\cdots$	x_t	計
p	p_1	p_2	$\cdots$	p_i	$\cdots$	p_t	1

（2）期待値

$$E\left(X\right)=\mu=\sum_{i=1}^{t}x_i p_i$$

（3）分散

$$V\left(X\right)=\sigma^2=\sum_{i=1}^{t}\left(x_i-\mu\right)^2 p_i$$

2.3 連続型確率変数

（1）期待値

$$E\left(X\right)=\mu=\int_{-\infty}^{\infty}xf\left(x\right)dx$$

$f\left(x\right)$ は確率密度関数。

（2）分散

$$V\left(X\right)=\sigma^2=\int_{-\infty}^{\infty}\left(x-\mu\right)^2 f\left(x\right)dx$$

3　期待値と分散に関する公式

$$\boxed{E\left(c\right)=c}$$ （c は定数）

$$\boxed{E\left(X+c\right)=E\left(X\right)+c}$$

$$\boxed{E\left(cX\right)=cE\left(X\right)}$$

$$\boxed{V\left(c\right)=0}$$

$$\boxed{V\left(X+c\right)=V\left(X\right)}$$

$$\boxed{V\left(cX\right)=c^2V\left(X\right)}$$

なぜなら

$$V\left(cX\right)=\sum_{i=1}^{n}\left(cx_i-c\mu\right)^2 p_i$$

$$=\sum_{i=1}^{n}\left(c\left(x_i-\mu\right)\right)^2 p_i=c^2\sum_{i=1}^{n}\left(x_i-\mu\right)^2 p_i$$

$$=c^2V\left(X\right)$$

別解

$$
\begin{aligned}
&V(cX)\\
&=E\left[\{cX-E(cX)\}^2\right]=E\left[\{cX-cE(X)\}^2\right]\\
&=E\left[\{c(X-E(X))\}^2\right]=E\left[c^2\{X-E(X)\}^2\right]\\
&=c^2E\left[\{X-E(X)\}^2\right]=c^2V(X)
\end{aligned}
$$

4　合成変数の値と確率

4.1 X と Y の同時確率

合成変数の値と確率という問題は躓きやすい。以下離散 2 変数の合成変数について表示し証明をしているが、連続変数でも公式は成り立つ。

	y_1	y_2	…	y_j	…	y_t	計
x_1	p_{11}	p_{12}	…	p_{1j}	…	p_{1t}	$p_{1\bullet}$
x_2	p_{21}	p_{22}	…	p_{2j}	…	p_{2t}	$p_{2\bullet}$
⋮	⋮	⋮		⋮		⋮	⋮
x_i	p_{i1}	p_{i2}	…	p_{ij}		p_{it}	$p_{i\bullet}$
⋮	⋮	⋮		⋮		⋮	⋮
x_s	p_{s1}	p_{s2}	…	p_{sj}	…	p_{st}	$p_{s\bullet}$
計	$p_{\bullet 1}$	$p_{\bullet 2}$		$p_{\bullet j}$		$p_{\bullet t}$	1

上の表は 2 変数の同時確率分布の確率を示したものだ。たとえばサイコロを 2 回ふる場合、x_1～x_6、y_1～y_6 にそれぞれ 1～6 の値が入り、最初に 2、次に 3 が出る確率は p_{23} である。計の行および列は X、Y それぞれの確率分布を表している（周辺確率分布）。

X と Y が独立であるとき、次の式が成り立つ。

$$p_{ij}=p_{i\bullet}p_{\bullet j}\quad\text{（独立のとき）}$$

また、ここでは次の式が成り立つ。

$$E(X)=\sum_{i=1}^{s}x_ip_{i\bullet}\qquad E(Y)=\sum_{i=1}^{t}y_jp_{\bullet j}$$

4.2 合成変数 X＋Y の値の表

次の表は 2 つの確率変数の合成変数 X＋Y の値を示したものだ。サイコロを 2 回ふる場合、x_2+y_3 のところが 5 になる。

	y_1	y_2	…	y_j	…	y_t
x_1	x_1+y_1	x_1+y_2	…	x_1+y_j	…	x_1+y_t
x_2	x_2+y_1	x_2+y_2	…	x_2+y_j	…	x_2+y_t
⋮	⋮	⋮		⋮		⋮
x_i	x_i+y_1	x_i+y_2	…	x_i+y_j		x_i+y_t
⋮	⋮	⋮		⋮		⋮
x_s	x_s+y_1	x_s+y_2	…	x_s+y_j	…	x_s+y_t

期待値は次の式で求められる。

$$E(X+Y)=\sum_{i=1}^{s}\sum_{j=1}^{t}(x_i+y_j)p_{ij}$$

$$=(x_1+y_1)p_{11}+(x_1+y_2)p_{12}+\cdots$$
$$+(x_i+y_j)p_{ij}\cdots+(x_s+y_t)p_{st}$$

すなわち、（値×確率）の総和である。

分散は次の式で求められる

$$V(X+Y)=E\left(\left((X+Y)-E(X+Y)\right)^2\right)$$
$$=\Sigma\Sigma\left(\left(x_i+y_j\right)-E(X+Y)\right)^2p_{ij}$$

すなわち「（値－期待値）の 2 乗×その確率」の総和で求められる。

4.3 合成変数 XY の値の表

次の表は 2 つの確率変数の合成変数 XY の値を示したものだ。サイコロを 2 回ふる場合、x_2y_3 のところが 6 になる。

	y_1	y_2	…	y_j	…	y_t	
x_1	x_1y_1	x_1y_2	…	x_1y_j	…	x_1y_t	
x_2	x_2y_1	x_2y_2	…	x_2y_j	…	x_2y_t	
⋮	⋮	⋮		⋮		⋮	
x_i	x_iy_1	x_iy_2	…	x_iy_j		x_iy_t	
⋮	⋮	⋮		⋮		⋮	
x_s	x_sy_1	x_sy_2	…	x_sy_j	…	x_sy_t	

期待値は次の式で求められる。

$$\boxed{E(XY)=\sum_{i=1}^{s}\sum_{j=1}^{t}(x_iy_j)p_{ij}}$$

すなわち、（値×確率）の総和である

分散は次の式で求められる。

$$V(XY)=E\left(((XY)-E(XY))^2\right)$$
$$=\Sigma\Sigma\left((x_iy_j)-E(XY)\right)^2p_{ij}$$

すなわち「（値－期待値）の 2 乗×その確率」の総和で求められる。

4.4 合成変数の期待値と分散についての公式

$$\boxed{E(X+Y)=E(X)+E(Y)}$$

なぜなら

$$E(X+Y)=\sum_i\sum_j(x_i+y_j)p_{ij}$$
$$=\sum_i\sum_j(x_ip_{ij}+y_jp_{ij})$$
$$=\sum_i\sum_j x_ip_{ij}+\sum_i\sum_j y_jp_{ij}$$
$$=\sum_i x_i\sum_j p_{ij}+\sum_j y_j\sum_i p_{ij}$$
$$=\sum_i x_ip_{i\bullet}+\sum_j y_jp_{\bullet j}=E(X)+E(Y)$$

X,Y が独立のとき次が成り立つ。

$$\boxed{E(XY)=E(X)E(Y)}\quad\text{（独立のとき）}$$

なぜなら

$$E(XY)=\sum_i\sum_j x_iy_jp_{ij}=\sum_i\sum_j x_iy_jp_{i\bullet}p_{\bullet j}$$
$$\left(\because p_{ij}=p_{i\bullet}p_{\bullet j}\right)$$
$$=\sum_i x_ip_{i\bullet}\sum_j y_jp_{\bullet j}=E(X)E(Y)$$

分散や期待値の証明では次の公式がよく使われる。

$$\boxed{V(X)=E(X^2)-[E(X)]^2}$$

なぜなら

$$V(X)=\sum(x_i-\mu)^2p_i$$
$$=\sum(x_i^2-2\mu x_i+\mu^2)p_i$$
$$=\sum x_i^2p_i-2\mu\sum x_ip_i+\mu^2\sum p_i$$
$$=\sum x_i^2p_i-2\mu\mu+\mu^2=\sum x_i^2p_i-\mu^2$$
$$=E(X^2)-\mu^2=E(X^2)-[E(X)]^2$$

別解

$$V(X)$$
$$=E\left[(X-E(X))^2\right]=E\left[(X-\mu)^2\right]$$
$$=E(X^2-2\mu X+\mu^2)$$
$$=E(X^2)-E(2\mu X)+E(\mu^2)$$
$$=E(X^2)-2\mu E(X)+\mu^2$$
$$=E(X^2)-2\mu\mu+\mu^2$$
$$=E(X^2)-\mu^2$$
$$=E(X^2)-[E(X)]^2$$

X,Y が独立のとき次が成り立つ。

$$\boxed{V(X+Y)=V(X)+V(Y)}\quad\text{（独立のとき）}$$

この証明には上の公式を使う。

$$V(X+Y)=E\left((X+Y)^2\right)-(E(X+Y))^2$$
$$=\sum_i\sum_j(x_i+y_j)^2p_{ij}-\left(\sum_i\sum_j(x_i+y_j)p_{ij}\right)^2$$

$$= \sum_i \sum_j \left({x_i}^2 p_{ij} + {y_j}^2 p_{ij} + 2 x_i y_j p_{ij} \right)$$
$$- \left(\sum_i \sum_j x_i p_{ij} + \sum_i \sum_j y_j p_{ij} \right)^2$$

$$= \sum_i \sum_j {x_i}^2 p_{ij} + \sum_i \sum_j {y_j}^2 p_{ij} + \sum_i \sum_j 2 x_i y_j p_{ij}$$
$$- \left(\sum_i x_i \sum_j p_{ij} + \sum_j y_j \sum_i p_{ij} \right)^2$$

$$= \sum_i {x_i}^2 \sum_j p_{ij} + \sum_j {y_j}^2 \sum_i p_{ij} + \sum_i \sum_j 2 x_i y_j p_{ij}$$
$$- \left(\sum_i x_i p_{i\bullet} + \sum_j y_j p_{\bullet j} \right)^2$$

$$= \left(\begin{array}{l} \sum_i {x_i}^2 p_{i\bullet} + \sum_j {y_j}^2 p_{\bullet j} \\ + \sum_i \sum_j 2 x_i y_j p_{ij} \end{array} \right)$$
$$- \left(\begin{array}{l} \left(\sum_i x_i p_{i\bullet} \right)^2 + \left(\sum_j y_j p_{\bullet j} \right)^2 \\ + \sum_i \sum_j 2 x_i y_j p_{i\bullet} p_{\bullet j} \end{array} \right)$$

$$= \sum_i {x_i}^2 p_{i\bullet} - \left(\sum_i x_i p_{i\bullet} \right)^2$$
$$+ \sum_j {y_j}^2 p_{\bullet j} - \left(\sum_j y_j p_{\bullet j} \right)^2$$
$$\left(\because \ p_{ij} = p_{i\bullet} p_{\bullet j} \right)$$

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

別解

$$V(X+Y)$$
$$= E\left[(X+Y)^2\right] - \left[E(X+Y)\right]^2$$
$$= E\left(X^2 + 2XY + Y^2\right) - \left[E(X) + E(Y)\right]^2$$

$$= E\left(X^2\right) + 2E(XY) + E\left(Y^2\right)$$
$$- \left[\left(E(X)\right)^2 + 2E(X)E(Y) + \left(E(Y)\right)^2\right]$$

$$= E\left(X^2\right) - \left(E(X)\right)^2 + E\left(Y^2\right) - \left(E(Y)\right)^2$$
$$+ 2E(XY) - 2E(X)E(Y)$$

$$= \left[E\left(X^2\right) - \left(E(X)\right)^2\right] + \left[E\left(Y^2\right) - \left(E(Y)\right)^2\right]$$
$$\left(\because \ E(XY) = E(X)E(Y) \right)$$

$$= V(X) + V(Y)$$

5　シグマと確率変数の計算で注意すること

ある確率分布に従う確率変数 X があり、同一の確率分布に従う確率変数 X_1、X_2 があったとする。このとき、

$$\boxed{X^2 \neq X_1 X_2}$$

下の、X^2 の分布と X_1X_2 の分布の違いを見ること。

X の分布は

X

値	1	2	計
確率	1/2	1/2	1

X^2 の分布は

X^2

値	1	4	計
確率	1/2	1/2	1

X_1X_2 の値と確率は

X_1X_2

		X_2		
		1	2	
X_1	1	1(1/4)	2(2/4)	1/2
	2	2(1/4)	4(1/4)	1/2
		1/2	1/2	1

値（確率）

X_1X_2 の分布は

X_1X_2

値	1	2	4	計
確率	1/4	2/4	1/4	1

同じ分布に従う確率変数であっても、異なる確率変数は違うものとして取り扱う。

シグマと期待値の計算は似ているところが多いので間違えてしまうこともある。以下のことには注意が必要だ。

$$E(X+Y)=E(X)+E(Y)$$
$$\sum(x_i+y_i)=\sum x_i+\sum y_i$$

$$E(X)=\mu$$
$$\sum x_i=n\bar{x}$$

$$E(c)=c$$
$$\sum c=nc$$

シグマについてわけがわからなくなったら、まずはシグマをとった式を書いてみること。

確率変数の期待値や分散についてわからなくなったらシグマの式に直してみること。