慶應義塾大学ビジネス・スクール

統計学　ノート（2）

1　はじめに

これまで，標本分布から平均，標準偏差を求める方法を考えてきた。次に，母集団分布の特徴を調べることを考える。標本分布は対応する母集団分布を推定したものである。標本の大きさが大きいときには，標本分布は母集団分布のよい近似として得られる。

しかし，たいていの場合には標本の大きさはそれほど大きくなく，十分な精度で推定できない。そこで，他の情報源からの知識，例えば，過去の経験をあわせて利用することによって，母集団分布の一般的な型，特徴を想定することができる。このような想定を基にして確率分布または理論分布とよばれる分布が導出される。

確率分布は実際の度数分布に対する数学モデルである。標本分布とそれに対応して理論的に導かれた確率分布について議論する際，標本分布を経験分布とよぶ。

2　確率変数

研究の対象として選ばれる変数をxで表す。確率分布との関連で，xを確率変数という。

例えば2つのサイコロを振る繰り返し型の実験で，サイコロの出た目の和に興味があるとする。このとき，2 つのサイコロの出た目の和を確率変数xで表すと，このxは$2,3,\Lambda,12$のいずれか1つの値をとる。

xは標本空間における標本点のある関数であって，形式的な定義は次のように与えられる。

定義 1　確率変数は標本空間の上で定義された実数値関数である。

この種の変数に対して「確率変数」という語を用いるのはこの変数のとる値が不確定な実験の結果に依存することを表すためである。

いま，我々に関心があるのは実験に対して確率変数がとる値であって，実験の可能な結果全体ではないので，より簡単な新しい標本空間を構成し，その空間で考える。このとき，新しい空間の標本点に確率を割り当てることができれば，確率変数の確率分布が決まった

ことになる。確率変数の分布は常に確率分布であり，経験分布ではない。

以上より，確率分布は経験分布に対応する数学的モデルであり，一方，変数 x の経験分布は x の確率分布の近似と考えられる。

3　確率分布

3.1　確率分布の平均と分散

確率分布に対しても標本分布の場合と同様に，分布の平均と標準偏差を考える。標本分布との違いは相対度数の代わりに確率を用いることである。

まず，理論平均について考える。標本分布において標本平均は次の式で与えられていた。

$$\bar{x} = \sum_{i=1}^{k} x_i \frac{f_i}{n}$$

理論平均 μ は標本平均の標本相対度数 f_i/n を確率 $P\{x_i\}$ におきかえて次のように得られる。

$$\mu = \sum_{i=1}^{k} x_i P\{x_i\} \tag{1}$$

同様に，理論分散について考えると，標本分散は次の式で与えた。

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

この式で f_i/n を $P\{x_i\}$ で，$\bar{x}$ を μ でおきかえて理論分散を σ^2 で表すと，次のようになる。

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 P\{x_i\} \tag{2}$$

理論標準偏差 σ は理論分散 σ^2 の正の平方根である。標本分散を定義するときには n で割るか $n-1$ で割るかが問題となったが，理論分散では定義式に n は含まれないので問題はない。

なお，σ^2 の計算には (2) を展開して導いた次の式を用いるのが便利である。

$$\sigma^2 = \sum_{i=1}^{k} x_i^2 P\{x_i\} - \mu^2 \tag{3}$$

3.2　期待値

上で確率分布の平均と分散について考えたが，次により一般的な性質に用いられる期待値の概念について述べる。

一般に確率変数 x は $x_1, x_2, \ldots, x_k$ という値のいずれか 1 つをとりうるとし，これらの値を取る確率をそれぞれ $P\{x_1\}, P\{x_2\}, \ldots, P\{x_k\}$ とする。ここで，$\sum_{i=1}^{k} P\{x_i\} = 1$ である。このとき，確率変数 x の期待値を次の量で定義する。

$$E[x] = \sum_{i=1}^{k} x_i P\{x_i\} \tag{4}$$

(1) と (4) を比べると，$E[x]$ は確率変数 x の平均値 μ にほかならない。ここでは期待値という概念はむだのようにみえるが，さらに進んで期待値を考えてみる。確率変数 x そのものでなく，関数 $g(x)$ について期待値を考えると，次の公式で与えられる。

$$E[x_i] = \sum_{i=1}^{k} g(x_i) P\{x_i\} \tag{5}$$

期待値の演算記号 E が次の性質をもつことが (5) から示される。ここで，c は任意の定数とする。

$$E[g(x) + c] = E[g(x)] + c \tag{6}$$

$$E[cg(x)] = cE[g(x)] \tag{7}$$

また，第 3 の性質は次の式で与えられる。

$$E[g(x) + h(y)] = E[g(x)] + E[h(y)] \tag{8}$$

ここで，x と y はいずれも任意の確率変数であり，g と h はそれぞれこれらの確率変数の任意の関数である。

確率変数の分散は平均と同様，期待値によって表される。つまり，(5) の $g(x)$ を $g(x) = (x - \mu)^2$ とおけばよいのである。したがって，

$$E[(x - \mu)^2] = \sum_{i=1}^{k} (x_i - \mu)^2 P\{x_i\}$$

これを (2) と比べれば，

$$\sigma^2 = E[(x - \mu)^2]$$

となることがわかる。したがって，平均と分散は確率変数の関数の期待値の特別な場合である。

期待値演算 E は，確率分布の平均，分散を特別な場合として含むような，平均値を一般化した概念として述べてきたが，これはまたある種の決定問題を解く上でも非常に有効である。

4　2 項分布

実験の可能な結果が事象 A が起こるかそうでないかの 2 つに分けられる場合について考える。便宜的に事象 A が起こるという結果を成功，起こらないという結果を失敗として分ける。ここで実験を繰り返すことを考え，得られる成功回数の総数，つまり事象 A の起こる回数を表す変数として確率変数 x を導入する。この種の確率変数を 2 項変数という。

まず，コインを投げる実験を考える。コインの表が出ることを成功と定義する。この実験を 3 回繰り返すとすると，確率変数 x は 3 回の投げで得られる表の数である。

表の出る確率と裏の出る確率が等しく，1/2 であるので，可能な結果はそれぞれ等しい確率 ($(1/2)^3 = 1/8$) で起こる。したがって，x の確率分布は次の表で与えられる。

表 1：コインを 3 回投げる実験の x の確率分布

x	0	1	2	3
確率	1/8	3/8	3/8	1/8

一方，2 つのサイコロを振る実験で出た目の和を x と定義すると，この実験では 2 項変数は生じない。なぜなら，サイコロの可能な結果は 6 通りであり，結果を成功か失敗の 2 つに分けていないからである。したがって，サイコロの 1 の目が出ることを成功と定義する。ここで，サイコロを 3 回振る実験を考える。そして，成功の起こる回数を x とする。可能な結果は結果の成功を S，失敗を F と書くことにすると，表 2 のように得られる。

表 2：サイコロを 3 回振る実験の結果と x の値

結果	SSS	SSF	SFS	FSS	SFF	FSF	FFS	FFF
x の値	3	2	2	2	1	1	1	0

それぞれの結果が起こる確率はコインを投げる実験とはかなり異なり等しくならず，乗法定理を用いて表 3 のように計算される。

表 3：サイコロを 3 回振る実験の結果とその確率

結果	SSS	SSF	SFS	FSS	SFF	FSF	FFS	FFF
確率	$\left(\frac{1}{6}\right)^3$	$\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)$	$\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)$	$\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)$	$\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2$	$\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2$	$\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2$	$\left(\frac{5}{6}\right)^3$

結果のそれぞれに割り当てられた確率を確率変数 x の値について和を考えることにより x に対する標本空間を導くことができる。次の表 4 は確率変数 x の確率分布を与える。

表 4：サイコロの実験のxの確率分布

x	0	1	2	3
$P\{x\}$	$\left(\frac{5}{6}\right)^3$	$3\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2$	$3\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)$	$\left(\frac{1}{6}\right)^3$

この 2 つの例で用いた方法はすべての 2 項分布の問題に適用できる。まず，全実験に対する標本空間を構成し，各標本点に対する確率を計算する。次に各標本点に確率変数 x の値を対応させ，確率変数の値ごとにその値に対応している点の確率を加えることによって x の各値に対する確率が求められる。これらの確率から x の確率分布が与えられる。

しかし，2 項分布の問題に対して，問題が起こるたびに上に述べたような計算を行うのは大変なので，この種のすべての問題に適用できる公式が必要である。そこで，一般的な 2 項分布の問題を考える。結果が 2 つ(成功と失敗)に分けられる実験を行うとする。成功が起こる確率が p と与えられているとすると，失敗が起こる確率を q で表すと，$p+q=1$ である。この実験を n 回繰り返すとする。n 回の実験で得られる成功の回数を x で表す。問題は x のとるそれぞれの値に対する確率を求めることである。$P\{x\}$ で 2 項変数の一般的な値に対する確率を表す。これらの確率によって定められる確率分布を 2 項分布といい，次の公式で与えられる。

$$P\{x\}=\frac{n!}{x!(n-x)!}p^x q^{n-x},\quad (x=0,1,2\cdots,n) \tag{9}$$

ここで，$n!/x!(n-x)!$ を 2 項係数とよぶが，n と x の値が大きいとき，この係数の値を求めるのは面倒である。

1 回の試行で成功の確率が p である実験の n 回の反復を一般にその実験の n 回の独立試行という。この言葉を使うと，$P\{x\}$ は 1 回の試行で成功の確率が p である実験の n 回の独立試行において，x 回の成功を得る確率であるといえる。

2 項分布の性質

計算により，一般の 2 項分布の平均と分散に対する公式を導くことができる。ここでは結果だけを与える。この公式により平均，分散を求める計算が容易になる。

$$\begin{aligned}\mu &= np\\ \sigma &= \sqrt{npq}\end{aligned} \tag{10}$$

5 正規分布

連続型変数の確率分布の中で最も頻繁に用いられる分布がこの節で述べる正規分布である。

正規分布はほぼ左右対称で釣鐘型をした度数分布に対して非常に有効であると知られ，また他の理由からもきわめて重要であると見られている理論分布である。正規分布はその曲線を表す次の式によって定義される。

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

ここで，μ と σ はそれぞれ正規分布の平均と標準偏差を表す。

分布の平均の幾何学的な意味は分布のヒストグラムが均衡を保つような x 軸上の点であった。この性質からヒストグラムがある垂直軸に関して対称であれば，平均は明らかに x 軸上の対称点にあるはずである。これは，標本の大きさを限りなく増やし，階級の幅を小さくしていったときの平均の極限値に対しても同様に成り立つ。平均の極限値(また

は理論値) を μ で表す。

ある度数分布に対するヒストグラムの極限形が正規曲線であるとする。このとき，s の極限値 σ はこの正規曲線に関して次のような幾何学的な意味をもつことが数学的に証明できる。

1. $\mu-\sigma$ と $\mu+\sigma$ の間の正規曲線下の面積は全面積の約 68％である。
2. $\mu-2\sigma$ と $\mu+2\sigma$ の間の正規曲線下の面積は全面積の約 95％である。
3. $\mu-3\sigma$ と $\mu+3\sigma$ の間の正規曲線下の面積は全面積の約 99.7％である。

正規曲線の主要な性質の 1 つは曲線の位置と形が μ と σ によって完全に決まるということである。μ の値は曲線の中心を決め，σ の値は分布の広がりの程度を与える。理論度数分布を表しているすべての正規曲線の全面積は 1 であるので，σ の値が大きくなれば曲線の高さは減少し，分布の広がりが大きくなる。正規曲線の形はその平均と標準偏差によって完全に決まることから，すべての正規曲線は簡単な変数変換によってある標準的な正規曲線に変えることができる。

取り扱いがもっとも簡便な正規曲線は平均が 0，分散が 1 の正規曲線であり，この曲線が表す分布を標準正規分布という。そして，任意の正規曲線はこの標準正規曲線に変形することができる。一般に，平均 μ，標準偏差 σ の正規曲線の横軸上の 1 点 x が標準正規曲線の横軸上の 1 点 z に対応するならば，x は μ から，標準偏差の z 倍のところにある。したがって対応する 2 点間の関係は公式

$$x = \mu + z\sigma$$

で与えられる。z を x によってあらわすと，次の標準化の公式が得られる。

$$z = \frac{x-\mu}{\sigma} \tag{11}$$

標準化の公式は任意の変数 x を平均 0，標準偏差 1 の変数 z に変換するものである。これは変数 x が正規変数でなくても成り立つ。この公式によって変数 x が変数 z に変換されたとき，変数 x は標準単位で測定されたという。

6　2 項分布の正規近似

2 項分布に関する問題は試行回数 n が大きくない場合，簡単に解ける。しかし，n が大きいとき公式 (9) による計算は非常に面倒になる。したがって，簡単でよい近似法が望まれる。そのような近似法の 1 つとして，正規分布によるものがある。

ヒストグラムを正規分布で近似するとき，当てはめた正規分布の広がりが増し，高さが減少してしまうことも考えられる。このようなヒストグラムの変化を避けるために，元の変数を標準単位の変数に変換するのがよい。つまり，変数

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

に対するヒストグラムを図示する。

(10) から標準変数 z は

$$z = \frac{x - np}{\sqrt{npq}}. \tag{12}$$

変数 z は平均 0，標準偏差 1 の分布にしたがうから，z に対するヒストグラムは n が大きくなっても全体としてはほぼそのままの形にとどまり，広がってしまうことはない。

p を一定に保ち，n を次第に大きくしていけば，z の分布は平均 0，標準偏差 1 の分布に次第に近づくことが数学的に証明できる。実際的には，n が，$p \leq 1/2$ ならば $np > 5$ を，$p > 1/2$ ならば $nq > 5$ を満たす限り，経験的にこの近似がかなりよいといえる。

2 項変数 x を標準化した変数の分布が標準正規分布に近づくという事実は，2 項変数 x のヒストグラムが n が大きいとき適当な正規曲線で当てはめられることを意味する。この正規曲線は (10) で与えられる平均と標準偏差をもつ。

n 回の試行において，成功の回数 x より成功の割合 x/n を用いたほうが便利な場合も多い。x/n を $\hat{p}$ とかくと，z は次のように書ける。

$$z = \frac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}} \tag{13}$$

この z は式を書き直しただけなので，近似的に平均 0，標準偏差 1 の分布にしたがう。n が大きいとき，成功の割合 $\hat{p}$ のヒストグラムは

$$\mu_{\hat{p}} = p$$
$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \tag{14}$$

の平均と標準偏差をもつ正規曲線でうまく当てはめられる。

しかし，n がかなり大きいならば，2 項分布に関するすべての問題が正規近似によって十分に処理できると考えてはならない。確率変数 x が負の値を取り得ない場合などには，分布は対称でなくなってしまうからである。

7 問題

1. ハートの 2，3，4 の 3 枚のトランプ札が入った箱がある。2 回目の抽出をする前に最初取り出した札をもとに戻しておくという方法で，2 枚の札をこの箱から取り出し，得られた札の上の数字の和を x で表す。場合の数を数え上げる方法で確率変数 x の分布を導け。

2. 最初に取り出した札はもとに戻さないとして，問題 1 を解け。

3. 問題 1 で得た分布に対して μ と σ の値を求めよ。

4. 問題 2 で得た分布に対して μ と σ の値を求めよ。

5. 硬貨を 3 回投げる。少なくとも 2 回表が出たら，サイコロを転がすことが許されて，出た目の数だけのドルがもらえる。このゲームを 1 度だけ行うとすると，このゲームで何ドルもらえると期待できるか。

6. 硬貨を 3 回投げるゲームを考える。このゲームでは，表が出るたびに 5 ドルがもらえる。そして，3 回とも裏のときにはもう 1 回硬貨を投げることが許されて，このときは結果がどうであろうとも 2 ドルがもらえるだけでなく，表が出たらさらに 12 ドルもらえる。このゲームで勝ちとれる期待金額を求めよ。

7. 1 回のブリッジで勝つ確率を 1/4 とし，5 回の勝負をするものとする。公式を用いて各 x に対する $P\{x\}$ の値を計算せよ。

8. 問題 7 の変数 x の平均と標準偏差を計算し，それが公式から求めた結果と一致することを示せ。

9. 2 項分布の平均の公式を確かめよ。（ヒント：まず，$\sum_{x=0}^{n} xP\{x\}$ を項別に書き，np を共通因子としてくくり出したものがそれぞれの項が $n-1$ 回の試行に対する 2 項分布になっている。）

10. 2 項分布の分散の公式を確かめよ。（他のテキストなどで調べてみよ。）

11. ある試験である学生がとった点数が標準単位zで 0.8 であったとする。試験の点数は正規分布に従うと仮定して，この試験を受けた学生の何％がこの学生より高い点数をとったといえるか。

12. ある試験の得点分布は平均 130 点，標準偏差 20 点の正規分布にほぼ近い形をしていた。100 点以上を合格とするとき，この試験で不合格になる学生は何％か。

13. あるゲームで勝つ確率が 0.6 のとき，このゲームを 7 回やって 4 回以上勝つ確率を正確な方法と，正規近似による方法の両方で求めよ。

14. ある多肢選択式試験が 20 題の問題からなるとする。各問題には 4 つの選択肢があり，正解はその中の 1 つである。このとき，ある学生がどの問題も当て推量で答えをえらぶとき，正解を少なくとも 8 題当てる確率はいくらか。

15. $n = 20, p = 1/10$ の 2 項分布で，$P\{0\}$ を求め，この値に基づいてこの 2 項分布には正規分布がうまく当てはまらない理由を考えよ。また，正規曲線による近似法による $P\{0\}$ を計算し，もとの結果と比較せよ。

参考文献

[1] 初等統計学 原書第 4 版，P. G. ホーエル，培風館 (1981)。

[2] 統計学入門，東京大学教養学部統計学教室編，東京大学出版会 (1991)。

慶應義塾大学ビジネス・スクール

Contents Works Inc.