確率論・統計学初学者におけるプログラミングを用いた指導方法の提案

佐藤　翔太郎*1
Email: n135217k@st.u-gakugei.ac.jp

*1：東京学芸大学教育学部人間社会科学類総合社会システム学科

◎Key Words プログラミング、大学生、確率論、統計学

1. はじめに

統計や、その学問である統計学は、様々な研究やビジネスの分野でデータ分析や分析に使用されている。これらで用いられる統計手法を学ぶ統計学は現在重要な学問であると言える。

統計を学ぶ上で確率は非常に重要である。なぜなら確率の有用な測度が統計で数量を扱う上で非常に重要だからである。過去において統計は近代確率論の結果を受けて発展しており、現代の統計の基礎は、そういった過去の統計の蓄積の上で成り立っている。こういったことから、統計と確率は非常に密接な関係にあると言える。

しかし、経済学部学生においては統計手法や統計学のみを扱う場合がおおく、確率論を理解できていないために、具体的な生データを用いた解析がうまく行えないことがあり、数学が苦手な割合の高い文系学生が確率論の理解に苦しむというケースが多く見られる。

そこで今回の研究は大学での教育における確率論の学習において、プログラミングを用いて、効率よく学習ができることを目的とする。

現在の確率論学習についての問題点は『確率論や統計学において必要である、「二項分布」「正規分布」「中心極限定理」「ランダムウォーク」「大数の法則」などを座学のみで理解することは大変難しく、実用レベルまでの育成に至らず、またそれぞれの学習内容の横の結びつきの理解が弱い。』ことである。

そのため学習の際に、プログラミング言語を用いて学習のサポートをすることで理解を深めることができると考えた。具体的には、R を用いて、実際のデータを用いて分散、標準偏差の仕組みを理解し、共分散、相関係数、２次元正規分布の同時密度関数への理解へと発展させ、複数のデータを用いる分析に必要な知識を養うことを目的とした指導方法の提案を行う。

２では、本研究における確率の意味を検討し、３では現在の確率論教育について整理し、４でプログラミングを用いた確率論教育の意義を検討し、５でプログラミングを用いた確率論教育方法について検討する。

2. 確率の意味

統計学を学ぶ文系学生において統計学の学び、なおかつ実際の調査で知識を使う上で、確率論の知識を理解していることは非常に有用である。そのため統計学を学ぶ前に確率論を学び、基礎的な知識を抑えることで「統計学で何を行うのか」「どういった理由でこの操作を行うのか」ということを理解し、実際の調査での活用で活かすことを目的とし、確率論の基礎的な論理を理解する必要がある。

また統計は一般に次のように言える。

「統計（statistics）という用語の基本的な意味は、集団を記述するということである。」

集団とは、集まりを数学的に言った言葉である。何らかの意味で同質とみなされ、同時に諸特徴・属性は均一ではなく、不規則に変動しているような個体の集まりを指す。

確率は一般に以下のように定義する。

① 全ての事象 A に対し、 $0 < P(A) < 1$

② $P(\Omega) = 1$

③ 互いに排反な事象 $A_1, A_2, \ldots \in F$ かつ
$A_i \cap A_j = \phi$ であるとき、
$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

ここで F は完全加法族である。事象 A の確率を P(A) とあらわす。事象は様々あるため、事象の集まりを F とすると、標本空間 S、確率 P（A）と合わせてこの３つで確率空間と呼ぶ。

3. 現在の確率論教育の検討

従来の統計教育では、検定・分析といった統計手法を別のものとして扱う。統計手法を理解する上で確率論の分野の理解は非常に重要なものであるが、文系学生で統計学を扱うことはあっても、確率論を扱う大学は非常に少ないことが現状としてあげられる。統計学では様々な統計手法を学ぶが、その際確率論の視点からの説明がされず小手先の技になってしまう学生が多数を占めてしまう。

また、統計ソフトを使うことにより、理論がわからなくても実際に検定・分析を行うことが可能である。

しかしこの場合理論を理解していないため、実際の結果に対する考察の欠如や、検定・分析方法の選択のミスが発生してしまう。

そこで実際の統計学では複数のデータを用いることが多いため、二つのデータの相関関係を二次元正規分布の確率密度関数の理解を深める必要がある。

実際に R を用いてデータを入力し、今日分散の値による変化を「可視化」できるような教育方法を提案する。

次章ではプログラミングを中心に教育する意義を述べる。

4. 確率論をプログラミングを中心として教育する意義

前章で述べた通り、プログラミングを用いる理由として、確率論の基礎的な内容の「可視化」をすることを最大の目的とした。確率論における理論がどのようにして実際用いられているのかということをしっかり理解することで、理論を理解し、それに基づいた統計手法を理解することができる。プログラミングを通して確率論の基礎を学ぶことで統計と確率の関係性を感覚的に理解しやすくなると考える。またプログラミングで実際の数値を使い学ぶことで統計ソフトで検定・分析を行う際、実際のデータを選別・収集し、統計手法においてどれが必要か、ということの理解にもつながると考える。

5. プログラミングを中心とした教育方法の検討

5.1 確率論の教育方法

確率論を理解する上で重要なのは、「分散」と「標準偏差」をしっかり理解することである。またその２つの値からの２つの確率変数の共分散を取り、共分散を用いて、相関係数を求め、２次元正規分布の同時密度関数へと発展させることで統計手法につながる確率論のより深い理解ができると考える。

分散の定義の式は

$\mathrm{V(X)=E}((X-\mu)^2)$

と表される。またこの分散の値の平方根を撮ったものが標準偏差である。この２つの値の変化を可視化することにより次の２次元正規分布の確率密度関数への理解につながる。

5.2 正規分布への発展

前項で学習者の分散と標準偏差の理解の方法について考察した。この項ではその分散と標準偏差の理解を確率分布に結びつける。

一般の正規分布において、確率密度関数の式は平均をμ、標準偏差σとしたとき、

$$f(x)=\frac{1}{\sqrt{2}\pi\sigma}exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

と表すことができる。一般の正規分布は平均（期待値）μと分散σ^2が決まれば一つに決まる。このときに分散と標準偏差により正規分布のグラフが形作られるということを理解することが重要である。

5.3 正規分布から２次元正規分布の確率密度関数への発展

前述の正規分布は、確率変数が１つの場合によるものであるが、統計学において２つ以上のデータを扱うことが多い。その際、その二つのデータの相関関係をしっかり分析できることが重要である。この節ではそのために２次元正規分布の確率密度関数における共分散と相関係数をしっかり理解することに重点を置く。

確率変数Xと Yとの間の共分散、$Cov(X,Y)$はXの平均をμ_x、Yの平均をμ_yとして連続型確率変数の場合は、

$$\mathrm{Cov(X,Y)}=\left[(X-\mu_x)(X-\mu_y)\right]$$
$$=\iint(x-\mu_x)(y-\mu_y)f(x,y)dxdy$$

と定義する。またこの共分散を用いて、X と Y の間にどれだけ強い線形の関係があるかという相関係数pを定義できる。

$$p=\frac{\mathrm{Cov(X,Y)}}{\sqrt{\sigma_x^2\sigma_y^2}}=E\left[\frac{X-\mu_x}{\sigma_x}\cdot\frac{Y-\mu_y}{\sigma_y}\right]$$

この値が大きいということは、X が大きいときに Y も大きい値をとるということである。相関係数は$-1<p<1$の範囲でとり、１に近いほど相関性が高いと言える。

分散の求め方を学習した上で共分散と相関係数を学ぶことにより学習者は正規分布における複数のデータの相関関係について深く理解することができると考える。

5.4 教育方法における教育効果への考察

前述の教育方法を行うことにより、学習者は実際のデータを扱う上で複数のデータの相関関係の把握が容易になると考える。問題点であった統計手法の誤用や分析のミスといったものが、共分散と相関係数のより深い理解により分析や検定がしっかり行うことが可能になると考える。

6. おわりに

本研究では、現在の統計・確率教育の問題をプログラミングを用いて解決できないかということを考え、教育方法の検討を行った。現在、文系の学生であっても統計学を扱うことは増えており様々な統計解析ソフトにより調査をもとに統計学手法を用いた研究が数多く行われている。

しかし、前述の問題点にもあげた通り、統計学の理解力と実践における総合的な理解、つまり統計と確率を結びつけて理解をすることを重要と考えそのうち一つの例を「共分散と相関係数をプログラミングにより可視化させ、理解をサポートする」という教育方法の提案を行った。データの違いにより分散の値や標準偏差の値が変わることを R によってプロットした２次元正規分布の確率密度関数を見ることで違いを把握させることにより学習者の理解がより深まると考えている。

本研究ではプログラミングの教育方法、正規分布から中心極限定理、大数の法則への指導方法の発展、統計手法にどのように結びつけていくか、といった点が不十分であると考える。

今後はプログラミングを用いて現段階の教育方法をどう実践していくかという検討をし、学習者は統計・確率を総合的に理解できるか、統計手法をしっかり検定・分析で利用できるか、という点について授業実践を通しさらなる研究をしていきたい。

参考文献

(1) 竹内啓：“統計学大辞典”，東洋経済新報社，(1989)．
(2) 松原望：“松原望の確率過程超！入門” 東京図書 (2011)
(3) 牧厚志，和合肇，西山茂，人見光太郎，吉川肇子，吉田栄介，濱岡豊：“経済・経営のための統計学”，有斐閣アルマ (2005)