SSE-DP-2024-1

統計エキスパート演習 2023

国友直人・湯浅良太・西颯人・趙宇・中西正

統計数理研究所，統計数理研究所，一橋大学，東京理科大学，北海道大学

2024 年 1 月

SSE-DP(ディスカッションペーパー・シリーズ)は以下のサイトから無料で入手可能です。
https://stat-expert.ism.ac.jp/training/discussionpaper/

SSE-DP-2024-1

Proceedings of Statistics Expert Seminar 2023

by

Naoto Kunitomo, Ryota Yuasa, Hayato Nishi, Yu Zhao

and Tadashi Nakanishi

Institute of Statistical Mathematics, Institute of Statistical Mathematics, Hitotsubashi University, Tokyo University of Science and Hokkaido University

January 2024

(Summary)

At the core institution of Consortium for training experts in statistical sciences (https://stat-expert.ism.ac.jp/, the Institute of Statistical Mathematics (ISM)), we organized many classes and seminars in 2023. One of seminar group with a small number of participants re-examined some basic issues in statistics and applications. This report summarizes the results of investigations of the group and some R-programs and Python-programs which were developed by the participants.

統計エキスパート演習 2023 *

国友直人・湯浅良太・西颯人・趙宇・中西正 †

2023 年 12 月

鍵言葉 (Key Words)： 統計エキスパート養成プロジェクト, 基礎統計, 応用統計, 計算統計, R-プログラム, Python プログラム

要約

統計数理研究所が推進している統計エキスパート養成プログラムでは必ずしも統計学を専門としているわけではない各分野の若手研究者と統計家であるメンターにより統計エキスパート演習をおこなっている。2023 年度に実施したある一つのグループ演習では統計学の基礎と応用について基礎的ではあるがしばしば見逃しがちな幾つかの内容、幾つかの応用統計の話題ををとりあげた。統計学の専門的な研究とまではいかないが、大学・大学院などで統計学を教える機会が少なくない統計エキスパートにとり有益と考えられる基礎統計を巡る 5 つの話題および講義などに関連した応用統計の 3 つの話題についての報告をまとめて報告する。またとりあげた話題を検討する過程で新たに作成した R プログラム, Python プログラムなども掲載する。

*統計エキスパート養成事業における研修生とメンターによるあるグループ演習の報告書。参加メンバーは国友直人 (メンター), 湯浅良太, 西颯人, 趙宇, 中西正 が参加した。Slack 上での議論に対するメンターの清水邦夫教授・三輪哲久教授（統計数理研究所）のコメントに感謝する。

†統計数理研究所, 統計数理研究所, 一橋大学, 東京理科大学, 北海道大学

目次

はじめに

統計数理研究所が推進している統計エキスパート養成プログラムでは必ずしも統計学を専門としているわけではない各分野の若手研究者と統計学分野を専門としているメンターにより統計エキスパート演習をおこなっている。2023 年度に実施したある少人数グループによる統計エキスパート演習では統計学の基礎と応用について基礎的ではあるがしばしば見逃しがちな内容を検討した。一言で基礎統計と云ってもその内容は多岐にわたるが、統計学の専門的な研究とまではいかないが、統計エキスパートにとり有益と考えられる基礎統計を巡る議論、および 2023 年度に実施された講義に関連した応用統計の 3 つの話題について報告する。また議論の中で新たに作成した R プログラム, Python プログラムの幾つかを掲載する。

なお開発した計算プログラムは統計エキスパート養成プロジェクト参加者が個人の責任で開発したものであるから、統計数理研究所としてはその内容についての責任は無いと考える。あくまで個人の責任で講義やセミナーの為に一般に公開するが[1]、誤りや誤植などを見つけた方、あるいは内容についてコメントがある方々は kunitomo アット ism.ac.jp までご連絡いただければ幸いである。

この報告書では 2023 年度に統計エキスパートにおける社会科学系 (経済・経営・統計) のあるグループ演習で行った議論の中から、統計エキスパートなら誰でも知っているべきと思われる統計基礎の範囲からから 5 つの話題、多様な応用統計の話題のなかから 3 つ、合計 8 つの話題を取りあげた。

第 1 章では標準偏差の推定問題を考察する。統計的推測の例としてしばしば標本分散の不偏推定、最尤推定などが説明されているが、実際の応用では標準偏差を利用することが多いが、実は標本分散の不偏推定と標本標準偏差の不偏推定が異なるという、統計的問題を考察する。第 2 章では正規分布の仮定について考察する。多くの教科書では正規分布を仮定した上での統計的推測の理論と応用が説明されている。それでは正規分布の仮定が実際のデータ分析に置いてに妥当か否か、を検討する統計的方法について議論している。第 2 章に続いて第 3 章では現実に観察される事象の中でも滅多に起こることがないが時々は起きるが、一たび起きると社会や人間に大きな影響を与える、極値現象の統計学の入門である。正規分布とともに極値分布 (フレッシェ, グンベル, ワイブル) と一般化パレート分布などを説明する。第 4 章ではヒストグラムを議論する。統計的データ分析ではまずヒストグラムから出発することが多いが、それではヒストグラムをどう描くか、自明ではないことを説明するとともに赤池情報量規準 (AIC) を説明する。第 5 章では統計学ではしばしば二つの標本を比較して平均の差を見ることが行われる。この場合、分散を既知と仮定したり、分散が共通と仮定する場合について教科書では説明している。それでは分散が異なる場合はどうなるか？Behrens-Fisher-Welch 問題とも呼ばれているこの古くから統計家の間では知られているが、今もなお重要なこの問題を議論する。

応用統計として取りあげる三つの話題のなかでも第 6 章では「株価」などニュースなどの話題として日常よく登場する経済・金融データについて論じる。統計エキスパート

[1] 計算プログラムは GitHub（一定の期間の間はアドレス `https://github.com/hayato-n/stat_expert_ism` に掲示する予定）よりダウンロード可能とする予定である。

としてこうした社会・経済で身近に遭遇するデータについてどう見たらよいか、統計的課題について基礎的な議論を提供した。

最後の二つの話題は 2023 年 5 月〜7 月に実施された「統計エキスパート養成」における二つの講義「社会における統計科学」、「計算ベイズ」の単位取得のために提出されたレポートから話題を取りあげた。第 7 章では統計学で重要な標本調査における現代的課題についての考察を解説している。第 8 章では近年の統計分析では市民権を得ている重要な分析であるベイズ統計学の最近の発展を踏まえた消費者行動の統計分析を巡る話題をとりあげた。ここではたまたま二つの話題を今回取りあげたが、その他、「統計エキスパート養成」のために統計科学分野で活躍中の講師陣により様々な話題についての講義が用意されている。

なお各章はそれぞれ独立した話題を扱っており、参加各自の責任であり、ほとんど調整などはしていないことをお断りしておく。

ここでとりあげたテーマは「統計エキスパート」でとりあげたテーマのほんの一部分に過ぎない。しかしながら、例えば「統計学基礎」を理解することは大学で開講されている講義「統計学」の単位を履修するには十分かもしれないが、大学の一般教養やそこから一歩踏み出した専門課程や大学で統計学を教えたり、専門の研究分野や実際の社会で遭遇するデータを統計的に処理するには十分ではないであろう。実際の問題では統計学を使った対処法を各自が考え、教科書にはないかもしれない事例について処理していくことが求められる。「統計エキスパート」を目指す場合にはそうした努力を絶えず行っていく必要がある。2023 年という短い間に「統計エキスパート養成」プロジェクトのあるグループ研修で議論した事例を集めただけであるが、何らかの意味で様々な形で今後の議論の参考になれば幸いである。

国友直人 (著者代表)
2024 年 1 月

第 I 部: 統計基礎からの話題

第 1 章 標準偏差のほとんど不偏な推定 [2]

2023-8-18

国友直人・湯浅良太・西颯人

1. はじめに

統計検定 2 級の教科書「統計学基礎」3 章では多くの教科書と同様、統計的推定の基準として一致性と不偏性について議論、例として正規分布 $N(0,\sigma^2)$ における σ^2 の不偏分散推定を説明している。

ところが同章で利用している例、あるいは多くの実務家が日常的に利用しているのは分散ではなく標準偏差 σ が多い。そこで不偏分散推定量

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

を利用して[3]

$$\hat{\sigma} = \sqrt{s_n^2} = s_n$$

とすると、これは標準偏差の不偏推定量ではない。辞典 Wikipedia(2023.7.15) によると $(n-1)s_n^2/\sigma^2$ が自由度 $n-1$ の $\chi^2(n-1)$ にしたがうことから $\mathbf{E}[s_n] = \sigma c_4(n)$,

$$c_4(n) = \sqrt{\frac{2}{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}$$

とガンマ関数を利用して表現できる[4]。さらに Wikipedia では標準偏差のほとんど不偏な (A simple rule of thumb) 推定量

$$s_n^* = \sqrt{\frac{1}{n-1.5}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

が提案されている。分散の最尤推定量は

$$\hat{\sigma}_{ML}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

で与えられるが不偏性は持たずバイアスは $[(n-1)/n - 1]\sigma^2$ である。

標準的統計学における推定量の評価基準は推定の平均二乗誤差 (mean squared error) であるので、標準偏差の推定量について優劣があるか否か気にかかる。推定量の標準的基準として一般に母数 θ について推定量 $\hat{\theta}$ (例えば竹村 (2020)) の平均二乗誤差

$$\mathrm{MSE}(\hat{\theta}) = \mathbf{E}[(\hat{\theta} - \theta)^2] = [\mathbf{E}(\hat{\theta}) - \theta]^2 + \mathbf{Var}[\hat{\theta}]$$

[2] 統計エキスパート slack 上で行われた統計的推測の基礎についての素朴な疑問や議論の中で統計エキスパート養成事業にとり有用と思われる内容をまとめようと作成中のメモである。清水邦夫特任教授・三輪哲久特任教授のコメントに感謝する。

[3] 「統計学基礎」とは表記が異なる。

[4] 付論を参照。

を利用しよう。
分散についてはよく知られているように不偏分散はバイアスがないので $\mathrm{MSE}(s_n^2) = 2\sigma^4/(n-1)$, 最尤推定量のバイアス $\sigma^2[(n-1)/n-1]$ より $\mathrm{MSE}(\hat{\sigma}_{ML}^2) = ([1/n]^2 + 2(n-1)/n^2)\sigma^4 = (2n-1)/n^2\sigma^4$ となるので $n \geq 1$ から

$$\mathrm{MSE}(s_n^2) > \mathrm{MSE}(\hat{\sigma}_{ML}^2)$$

となる。この MSE の意味では最尤推定量が不偏分散推定量を一様に優越することは少なくとも統計理論家にはよく知られていて、不偏性の規準に対する問題、としてとりあげられることもある。

2. **標準偏差推定の MSE**

正規分布の標準偏差 σ の推定量として s_n, $\hat{\sigma}_{ML}$, s_n^* を比較してみると、次の結果が得られる。

$$\begin{aligned}
\mathrm{MSE}(s_n) &= \sigma^2\left[\frac{1}{2n} + (\frac{7}{16})\frac{1}{n^2} + o(\frac{1}{n^2})\right], \\
\mathrm{MSE}(\hat{\sigma}_{ML}) &= \sigma^2\left[\frac{1}{2n} + (\frac{7}{16})\frac{1}{n^2} + o(\frac{1}{n^2})\right], \\
\mathrm{MSE}(s_n^*) &= \sigma^2\left[\frac{1}{2n} + (\frac{10}{16})\frac{1}{n^2} + o(\frac{1}{n^2})\right].
\end{aligned}$$

この結果は本稿の計算プログラムによるシミュレーションの数値計算と整合的である。もっとも興味深いことは分散推定問題と異なり標準偏差の推定では最初の二つの推定量はほとんど差がない、という事実である。ほとんど不偏な標準偏差推定量は MSE 基準では僅かではあるが他より MSE が大きい。こうしたことは教科書には書かれていないが、実用的な意味は小さくないと考えられる。また同時にその理由を検討することは興味深い。

< **数理的導出** >

$X = (n-1)s_n^2/\sigma^2$ とすると $X \sim \chi^2(n-1)$ の密度関数は

$$g(x) = [1/\Gamma((n-1)/2)][\frac{x}{2}]^{(n-1)/2-1}\exp[-x/2](1/2)$$

であるから

$$\begin{aligned}
\mathbf{E}[s_n] &= [\frac{\sigma}{\sqrt{n-1}}]\mathbf{E}[X^{1/2}] \\
&= [\frac{\sigma}{\sqrt{n-1}}]\int_0^\infty x^{1/2}g(x)dx \\
&= [\frac{\sigma}{\sqrt{n-1}}\frac{1}{2^{(n-1)/2}}\frac{1}{\Gamma(\frac{n-1}{2})}]\int_0^\infty x^{\frac{n}{2}-1}e^{-x/2}dx \\
&= [\frac{\sigma}{\sqrt{n-1}}\frac{1}{2^{(n-1)/2}}\frac{1}{\Gamma(\frac{n-1}{2})}]2^{\frac{n}{2}}\int_0^\infty y^{\frac{n}{2}-1}e^{-y}dy \\
&= [\frac{\sigma}{\sqrt{n-1}}]2^{1/2}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}
\end{aligned}$$

となる。ここで Stirling の公式

$$\Gamma(x) = \sqrt{2\pi}x^{x-\frac{1}{2}}\exp(-x)[1 + \frac{1}{12x} + \frac{1}{288x^2} + \cdots]$$

を利用して (高木 (1960) p.258) ガンマ関数の x が大きい時を評価すると、例えば $\Gamma(n/2)/\Gamma((n-1)/2) = \sqrt{n/2}[1 - (3/4)n^{-1} - (7/32)n^{-2} + o(n^{-2})]$, $[1 - n^{-1}]^{-1/2} = 1 + (1/2)n^{-1} + (3/8)n^{-2} + o(n^{-2})$ より

$$\sqrt{\frac{2}{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} = 1 - \frac{1}{4n} - \frac{7}{32n^2} + o(\frac{1}{n^2})$$

となる。
一般に σ の推定量を $\hat{\sigma}_n = c_n\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}$ とすると MSE は

$$\mathrm{MSE}(\hat{\sigma}_n) = \sigma^2[c_n^2(n-1) + 1 - 2c_n\sqrt{2}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}]$$

となる。したがって $c_{1n} = 1/\sqrt{n-1}$, $c_{2n} = 1/\sqrt{n}$, $c_{3n} = 1/\sqrt{n-1.5}$ として評価すると結果が得られる。(例えば最初の場合は $2 - 2\sqrt{\frac{n}{n-1}}[1 - \frac{3}{4n} - \frac{7}{32n^2}] \sim \frac{1}{2n} + \frac{7}{16n^2}$, 2 番目は $\frac{n-1}{n} + 1 - 2[1 - \frac{3}{4n} - \frac{7}{32n^2}] \sim \frac{1}{2n} + \frac{7}{16n^2}$, 3 番目は $\frac{1-1/n}{1-(3/2n)} + 1 - 2[1 - \frac{3}{2n}]^{-1/2}[1 - \frac{3}{4n} - \frac{7}{32n^2}] \sim \frac{1}{2n} + \frac{10}{16n^2}$ となる。)

< 付論：非正規分布の場合 >

確率変数列 X_i $(i = 1, \cdots, n)$ が i.i.d., 期待値 $\mathbf{E}[X_i] = 0$, 分散 $\mathbf{E}[X_i^2] = \sigma^2$, 簡単化の為に $\mathbf{E}[X_i^6] < \infty$ とする。このとき $\mathbf{E}[\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2] = \sigma^2$ であるが

$$Y_n = \sqrt{n}[\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 - \sigma^2]$$

$Y_n^* = [1/\sqrt{n}]\sum_{i=1}^n (X_i^2 - \sigma^2)$ とすると、$\bar{X}_n \xrightarrow{p} 0$, Y_n^* に中心極限定理 (central limit theorem) を適用すると $Y_n \xrightarrow{w} N(0, (2 + \kappa_4)\sigma^4)$ となる。ただし $\kappa_4 = \mathbf{E}[X^4]/\sigma^4 - 3$ としたが、標本分散の分散の評価には 4 次積率が必要となる。(i.i.d. 系列なので $\mathbf{V}(X_i^2) = \mathbf{E}[(X_i^2 - \sigma^2)^2] = (\kappa_4 + 2)\sigma^4$ を利用すればよい。) そこで $[\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2]^{1/2} = [\sigma^2(1 + \frac{1}{\sigma^2\sqrt{n}}Y_n)]^{1/2}$ を形式的に展開すると

$$\begin{aligned} & [\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2]^{1/2} \\ = \quad & \sigma + \sigma(\frac{1}{2})\frac{Y_n}{\sigma^2\sqrt{n}} + \sigma(\frac{1}{2!})(\frac{1}{2})(-\frac{3}{2})[\frac{Y_n}{\sigma^2\sqrt{n}}]^2 + o_p(\frac{1}{n}) \end{aligned}$$

となる。(ここで剰余項の正当化は省略する。例えば積率条件は十分条件であり緩和できるだろう。) 形式的に期待値をとると

$$\mathbf{E}\left[\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}\right] = \sigma - \frac{1}{8}[\frac{1}{\sigma^3 n}]\mathbf{E}[Y_n]^2 + o(\frac{1}{n})$$

となる。そこで積率を評価すると[5]

$$\mathbf{E}\left[(\sum_{i=1}^{n}(X_i-\bar{X})^2)^2 \right] = \sigma^4\{\kappa_4[n(1-\frac{1}{n})^2]+2n(1-\frac{1}{n})^2+2n(n-1)(\frac{1}{n})^2+n^2(1-\frac{1}{n})^2\}$$

となるので $\mathbf{E}[Y_n^2] = \sigma^4[(\kappa_4+2)+\frac{2}{n-1}]$ となる。$\sqrt{(n-1)/(n-c)} \sim 1+(1/2)(c-1)/n+o(1/n^2)$, $[1+(1/2)(c-1)/n+o(1/n^2)][1-(1/8)(\kappa_4+2)/n+o(1/n^2)] = 1+(1/8)[4(c-1)-(\kappa_4+2)]/n+o(1/n^2)$ より実数 c に対して

$$\mathbf{E}\left[\sqrt{\frac{1}{n-c}\sum_{i=1}^{n}(X_i-\bar{X})^2} \right] = \sigma+\sigma\frac{4c-\kappa_4-6}{8n}+o(\frac{1}{n})$$

が得られるので $c^* = 3/2+\kappa_4/4$ とすると almost unbiased(ほとんど不偏) になる。例えば $X \sim N((\mu,\sigma^2)$ なら $\kappa_4 = 0$ より $c^* = 1.5$, $X \sim Poisson(\lambda)$ なら $\kappa_4 = 1/\lambda$, $c^* = 1.5+1/(4\lambda)$ となる。
ただしこの場合には標本分散のバイアスを補正するために 4 次積率が必要となるので実用性に乏しいように考えられる。実際のデータ分析では分散をよく推定するために高次積率の情報を利用するのは現実的ではないだろう。

3. シミュレーションプログラムと数値例

推定量の性質を調べるための標準的なシュミレーション・プログラム（R と Python 翻訳版) を開発した[6]。

このプログラムではデータ数 N, シミュレーション数 R としてシミュレーションで発生させた 3 つの推定量の標本分散、標本標準偏差の図、MSE、バイアスを比較したものである。以下では $R = 10^5$, $N = 10$ の場合を例示しておくが、出力表示を含めプログラムをほんの少し変更すれば自由に様々な状況について推定量の性質を調べることができる。例に示されている通り、シミュレーションの結果は数理的理論と整合的になっている。

---

[5]任意の $i,j = 1,\cdots,n$ に対して $p_{ii} = 1-1/n$,$p_{ij} = -1/n$ $(i \neq j)$ として評価すると
$\sum_{i=1}^{n}\mathbf{E}[X_i^4]p_{ii}^2+\sigma^4[\sum_{i_1=i_2\neq j_1=j_2}+\sum_{i_1=i_2\neq j_1=j_2}+\sum_{i_1=i_2\neq j_1=j_2}]$ より
$=(3+\kappa_4)\sigma^4\sum_{i=1}^{n}p_{ii}^2+\sigma^4[\sum_{i_1=i_2\neq j_1=j_2}+\sum_{i_1=j_1\neq i_2=j_2}+\sum_{i_1=j_2\neq i_2=j_1}]p_{i_1,j_1}p_{i_2,j_2}]$
$=\kappa_4\sigma^4 n(1-1/n)^2+\sigma^4[\sum_{i_1,j_1=1}^{n}p_{i_1,j_1}^2+\sum_{i_1,i_2=1}^{n}p_{i_1,i_1}p_{i_2,i_2}+\sum_{i_1,i_2=1}^{n}p_{i_1,i_2}p_{i_2,i_1}]$
を整理すればよい。

[6]最近では Python は工学系分野における一つの標準的な言語になっている。統計エキスパート・プロジェクトでは R と Python の知識はデータ分析の応用上で必須と見なしている。

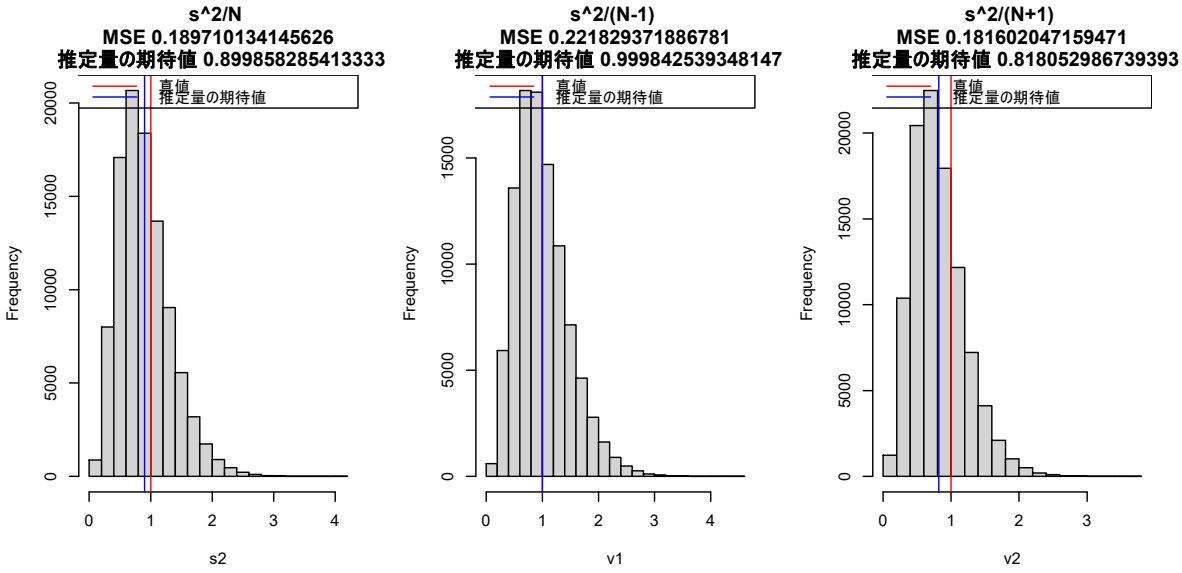

図 1: 分散推定シミュレーション

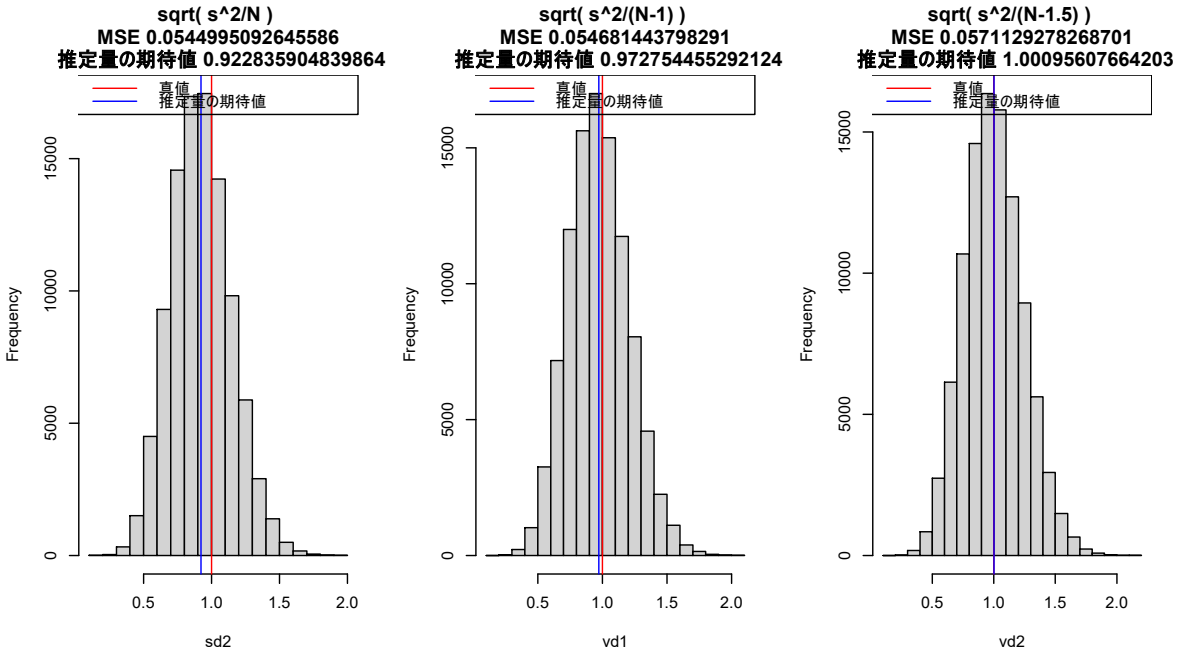

図 2: 標準偏差推定シミュレーション

4. 文献

[1] 統計学基礎 (改訂版), 日本統計学会編, 2015, 東京図書.

[2] 現代数理統計学 , 竹村彰通, 2020, 学術図書出版.

[3] 解析概論 (第三版), 高木貞治, 1960, 岩波.

[4 Unbiased Estimation of the Standard Deviation," Wikipedia, 2023-7-14.

```
### 2023-7-26 by Ryota Yuasa
### 分散の推定のMSE比較
R <- 10^5
N <- 10
s2 <- numeric(R)
v1 <- numeric(R)
v2 <- numeric(R)

for(i in 1:R){
  vX <- rnorm(N,0,1)
  s2[i] <- sum((vX - mean(vX))^2)/N
  v1[i] <- sum((vX - mean(vX))^2)/(N-1)
  v2[i] <- sum((vX - mean(vX))^2)/(N+1)
}

# MSE
mean((1-s2)^2)
mean((1-v1)^2)
mean((1-v2)^2)
# MSEの理論値
(2*N-1)/N^2
2/(N-1)
# 不偏性
mean(s2)
mean(v1)
mean(v2)

par(mfrow = c(1, 3))
hist(s2, breaks = 20, main = paste("s^2/N", "¥n", "MSE", mean((1-s2)^2), "¥n", "
推定量の期待値", mean(s2)))
abline(v = 1, col = 'red')
abline(v = mean(s2), col = 'blue')
legend("topright", legend = c("真値", "推定量の期待値"), col = c("red",
"blue"),lty=c(1,1))

hist(v1, breaks = 20, main = paste("s^2/(N-1)", "¥n", "MSE", mean((1-v1)^2),
"¥n", "推定量の期待値", mean(v1)))
abline(v = 1, col = 'red')
abline(v = mean(v1), col = 'blue')
legend("topright", legend = c("真値", "推定量の期待値"), col = c("red",
"blue"),lty=c(1,1))

hist(v2, breaks = 20, main = paste("s^2/(N+1)", "¥n", "MSE", mean((1-v2)^2),
"¥n", "推定量の期待値", mean(v2)))
abline(v = 1, col = 'red')
abline(v = mean(v2), col = 'blue')
legend("topright", legend = c("真値", "推定量の期待値"), col = c("red",
"blue"),lty=c(1,1))
par(mfrow = c(1, 1))

### 標準偏差のMSE比較
R <- 10^5
N <- 10
sd2 <- numeric(R)
vd1 <- numeric(R)
vd2 <- numeric(R)

for(i in 1:R){
```

```
  vX <- rnorm(N,0,1)
  sd2[i] <- sqrt(sum((vX - mean(vX))^2)/N)
  vd1[i] <- sqrt(sum((vX - mean(vX))^2)/(N-1))
  vd2[i] <- sqrt(sum((vX - mean(vX))^2)/(N-1.5))
}

# MSE
mean((1-sd2)^2)
mean((1-vd1)^2)
mean((1-vd2)^2)
# MSEの理論値
MSEs <- function(c,n){
  (c^2)*(n-1) + 1 - 2*c*sqrt(2)*gamma(n/2)/gamma((n-1)/2)
}
MSEs(1/sqrt(N), N)
MSEs(1/sqrt(N-1), N)
MSEs(1/sqrt(N-1.5), N)
#MSEs(1/sqrt(N-0.5), N)

# c=1/sqrt(N-x)として、MSEの理論値の最小化をするxを実験的に求める
MSEs2 <- function(x,n){
  MSEs(1/sqrt(n-x), n)
}
Ns <- c(3,5,7,10,15,20,30,50,100,200,300) #実験に用いるNたち
optx <- numeric(length(Ns))
for (i in 1:length(Ns))
optx[i] <- optimize(MSEs2, interval = c(-2,2), n=Ns[i])$minimum
optx # MSEの理論値の最小化をするx. Nを大きくするにつれ0.5に近づいている. ただし
MSEとして大きな差はなし

# MSEの理論値の近似値
1/(2*N) + (7/16)*(1/N^2)
1/(2*N) + (10/16)*(1/N^2)
# 不偏性
mean(sd2)
mean(vd1)
mean(vd2)

par(mfrow = c(1, 3))
hist(sd2, breaks = 20, main = paste("sqrt( s^2/N )", "¥n", "MSE",
mean((1-sd2)^2), "¥n", "推定量の期待値", mean(sd2)))
abline(v = 1, col = 'red')
abline(v = mean(sd2), col = 'blue')
legend("topright", legend = c("真値", "推定量の期待値"), col = c("red",
"blue"),lty=c(1,1))

hist(vd1, breaks = 20, main = paste("sqrt( s^2/(N-1) )", "¥n", "MSE",
mean((1-vd1)^2), "¥n", "推定量の期待値", mean(vd1)))
abline(v = 1, col = 'red')
abline(v = mean(vd1), col = 'blue')
legend("topright", legend = c("真値", "推定量の期待値"), col = c("red",
"blue"),lty=c(1,1))

hist(vd2, breaks = 20, main = paste("sqrt( s^2/(N-1.5) )", "¥n", "MSE",
mean((1-vd2)^2), "¥n", "推定量の期待値", mean(vd2)))
abline(v = 1, col = 'red')
abline(v = mean(vd2), col = 'blue')
legend("topright", legend = c("真値", "推定量の期待値"), col = c("red",
```

```
"blue"),lty=c(1,1))
par(mfrow = c(1, 1))
```

```
# %% 2023-7-25 by Nishi Hayato
# 使用するパッケージの読み込み
import numpy as np  # 数値計算・乱数生成用パッケージ
import matplotlib.pyplot as plt  # グラフ描画用パッケージ
from pathlib import Path  # ファイル操作用パッケージ

# 配色の設定（もしエラーになったらコメントアウトしてください）
plt.style.use("tableau-colorblind10")
# %%
##### Settings #####
R = 10**5
N = 10
Var_True = 10.0  # 母集団の分散（真の分散）
# DGP_Familyに設定した分布を母集団として乱数を生成します。
# GaussとPoissonを実装しています。
DGP_Family = "Gauss"
# DGP_Family = "Poisson"
# 乱数のシード値を固定することで、結果に再現性を持たせています
SEED_random = 123
# 画像を保存するフォルダを指定します。
IMG_path = Path("./img/")
####################
if DGP_Family.lower() in ("gauss", "normal"):
    def gen_data(rng):
        return rng.normal(loc=0, scale=np.sqrt(Var_True), size=(N, R))
    # N × R個の乱数をまとめて生成して、R個の分散推定値を一気に計算する仕様にし
ています。
    # N個の乱数を生成する -> 分散を計算する
    # というプロセスをR回反復するよりも、高速に実行できます。
elif DGP_Family.lower() == "poisson":
    def gen_data(rng):
        return rng.poisson(lam=Var_True, size=(N, R))
else:
    raise ValueError(f"{DGP_Family=} is not defined.")
# 以下で分散・標準偏差のMSEとバイアスを計算する関数を定義します。
def get_MSE_var(s2):
    return np.mean(np.square(s2 - Var_True))
def get_MSE_std(s):
    return np.mean(np.square(s - np.sqrt(Var_True)))
def get_Bias_var(s2):
    return np.mean(s2) - Var_True
def get_Bias_std(s):
    return np.mean(s) - np.sqrt(Var_True)
# %%
# 分散の推定のMSE比較
# 乱数生成器を設定
rng = np.random.default_rng(seed=SEED_random)
# 乱数をvXに格納
vX = gen_data(rng)
# 各種の分散推定値をR個ずつ計算
s2 = np.sum((vX - np.mean(vX, axis=0, keepdims=True)) ** 2, axis=0) / N
v1 = np.sum((vX - np.mean(vX, axis=0, keepdims=True)) ** 2, axis=0) / (N - 1)
v2 = np.sum((vX - np.mean(vX, axis=0, keepdims=True)) ** 2, axis=0) / (N + 1)
# %%
# MSE
print("s2: MSE =", get_MSE_var(s2))
print("v1: MSE =", get_MSE_var(v1))
print("v2: MSE =", get_MSE_var(v2))
```

```
# 不偏性
print("s2: Bias =", get_Bias_var(s2))
print("v1: Bias =", get_Bias_var(v1))
print("v2: Bias =", get_Bias_var(v2))
# %%
# グラフの描画
# グラフの描き方はいくつかある
# これはやさしい描き方ではないが、細かく設定できる方法
# ここでは詳しくは説明しない
fig, axes = plt.subplots(1, 3, figsize=(10, 3), sharex=True, sharey=True)
axes[0].hist(s2, bins=20)
axes[0].set_title("s^2/N¥nMSE {:.3f} (N={})".format(get_MSE_var(s2), N))
axes[0].axvline(Var_True, ymin=0, ymax=1, c="red", label="True value")
axes[0].axvline(np.mean(s2), ymin=0, ymax=1, c="blue", label="E[estimate]")
axes[0].legend()
axes[0].set_xlabel("s2")
axes[0].set_ylabel("Frequency")

axes[1].hist(v1, bins=20)
axes[1].set_title("s^2/(N-1)¥nMSE {:.3f} (N={})".format(get_MSE_var(v1), N))
axes[1].axvline(Var_True, ymin=0, ymax=1, c="red", label="True value")
axes[1].axvline(np.mean(v1), ymin=0, ymax=1, c="blue", label="E[estimate]")
axes[1].legend()
axes[1].set_xlabel("v1")
axes[1].set_ylabel("Frequency")

axes[2].hist(v2, bins=20)
axes[2].set_title("s^2/(N+1)¥nMSE {:.3f} (N={})".format(get_MSE_var(v2), N))
axes[2].axvline(Var_True, ymin=0, ymax=1, c="red", label="True value")
axes[2].axvline(np.mean(v2), ymin=0, ymax=1, c="blue", label="E[estimate]")
axes[2].legend()
axes[2].set_xlabel("v2")
axes[2].set_ylabel("Frequency")

fig.tight_layout()
fig.savefig(
    Path(IMG_path, f"{DGP_Family}-{N=}-Variance(sigma2={Var_True}).png"),
dpi=300
)
fig.show()
# plt.show()

# %%
# 標準偏差のMSE比較
# 分散と同様のプロセスを、標準偏差についても実施している。
# vd2はv2と分母が違うことに注意
rng = np.random.default_rng(seed=SEED_random)
vX = gen_data(rng)
sd2 = np.sqrt(np.sum((vX - np.mean(vX, axis=0, keepdims=True)) ** 2, axis=0) /
N)
vd1 = np.sqrt(np.sum((vX - np.mean(vX, axis=0, keepdims=True)) ** 2, axis=0) /
(N - 1))
vd2 = np.sqrt(
    np.sum((vX - np.mean(vX, axis=0, keepdims=True)) ** 2, axis=0) / (N - 1.5)
)
# %%
# MSE
print("sd2: MSE =", get_MSE_std(sd2))
```

```
print("vd1: MSE =", get_MSE_std(vd1))
print("vd2: MSE =", get_MSE_std(vd2))
# 不偏性
print("sd2: Bias =", get_Bias_std(sd2))
print("vd1: Bias =", get_Bias_std(vd1))
print("vd2: Bias =", get_Bias_std(vd2))
# %%
fig, axes = plt.subplots(1, 3, figsize=(10, 3), sharex=True, sharey=True)

axes[0].hist(sd2, bins=20)
axes[0].set_title("sqrt(s^2/N)¥nMSE {:.3f} (N={})".format(get_MSE_std(sd2), N))
axes[0].axvline(np.sqrt(Var_True), ymin=0, ymax=1, c="red", label="True value")
axes[0].axvline(np.mean(sd2), ymin=0, ymax=1, c="blue", label="E[estimate]")
axes[0].legend()
axes[0].set_xlabel("sd2")
axes[0].set_ylabel("Frequency")

axes[1].hist(vd1, bins=20)
axes[1].set_title("sqrt(s^2/(N-1))¥nMSE {:.3f} (N={})".format(get_MSE_std(vd1),
N))
axes[1].axvline(np.sqrt(Var_True), ymin=0, ymax=1, c="red", label="True value")
axes[1].axvline(np.mean(vd1), ymin=0, ymax=1, c="blue", label="E[estimate]")
axes[1].legend()
axes[1].set_xlabel("vd1")
axes[1].set_ylabel("Frequency")

axes[2].hist(vd2, bins=20)
axes[2].set_title("sqrt(s^2/(N-1.5))¥nMSE {:.3f}
(N={})".format(get_MSE_std(vd2), N))
axes[2].axvline(np.sqrt(Var_True), ymin=0, ymax=1, c="red", label="True value")
axes[2].axvline(np.mean(vd2), ymin=0, ymax=1, c="blue", label="E[estimate]")
axes[2].legend()
axes[2].set_xlabel("vd2")
axes[2].set_ylabel("Frequency")

fig.tight_layout()
fig.savefig(Path(IMG_path, f"{DGP_Family}-{N=}-Std(sigma2={Var_True}).png"),
dpi=300)
fig.show()
# plt.show()
# %%
```

第 2 章 確率分布の積率・裾と正規性 [7]

2023-8-22

国友直人・湯浅良太

1. はじめに

統計検定 2 級の教科書「統計学基礎」2 章では多くの教科書と同様、確率分布と期待値・分散の説明がある。その後、積率 (モーメント) から歪度 (skewness)・尖度 (kurtosis) および幾つかの確率分布の説明があるが、なぜこうした積率が意味があるのか曖昧な説明と思われる。ここでは確率分布・積率と統計的データ分析について統計エキスパートなら理解すべきと思われる論点について説明しよう。
一次元の記号として確率分布関数 $F(x)$, 確率関数 $p(x)$, 密度関数 $f(x)$ $(= \frac{dF(x)}{dx})$、積率を $\mathbf{E}[X^k]$, $\mu_k = \mathbf{E}[(X - \mathbf{E}(X))^k]$ としよう。

例 : 平均 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ を考える。未知の母平均 μ の推定量としては標本平均 $\hat{\mu}$ $(= \bar{X}_n) = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ が自然である。母分散 σ^2 の推定量としては標本不偏分散 $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$ が最も良く用いられる。χ^2- 分布の性質から $\mathbf{E}[s_n^2] = \sigma^2$ となる。

期待値や分散は 1 次積率・2 次積率であるが、と同様に 3 次積率・4 次積率 (moments) の推定量として平均周りの高次の標本積率

$$m_3 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^3 \ , m_4 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^4$$

などが用いられる。正規分布を仮定すると 3 次・4 次積率 $\mathbf{E}[(X-\mu)^3] = 0, \mathbf{E}[(X-\mu)^4] = 3\sigma^4$ となる。したがって歪度 (skewness)$\kappa_3 = \mathbf{E}[(X-\mu)^3]/\sigma^3 = 0$, 尖度 (kurtosis)$\kappa_4 = \mathbf{E}[(X-\mu)^4]/\sigma^4 - 3 = 0$ となる。(なお尖度を $\kappa_4^* = \mathbf{E}[(X-\mu)^4]/\sigma^4$ で定義することもある。) このことより母集団分布が正規分布であることの妥当性は統計量として

$$b_1 = \frac{\sqrt{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^3}{[\sqrt{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}]^3} \ , b_2 = \frac{n\sum_{i=1}^{n}(X_i - \bar{X}_n)^4}{[\sum_{i=1}^{n}(X_i - \bar{X}_n)^2]^2} - 3\frac{(n-1)}{(n+1)}$$

などが利用される。標本数 n が大きければ中心極限定理 (CLT) により近似的に

$$b_1 \overset{a}{\sim} N(0, \frac{6(n-2)}{(n+1)(n+2)})$$

$$b_2 \overset{a}{\sim} N(0, \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)})$$

[7] 統計エキスパート slack 上で行われた統計的推測の基礎についての素朴な疑問や議論の中で統計エキスパート養成事業にとり有用と思われる内容をまとめようと作成中のメモである。

となる[8]。データ分析では $\kappa_4 > 0$ のとき分布の裾は正規分布より厚い、$\kappa_4 < 0$ のとき裾確率は正規分布より薄い、などと解釈されている。データから計算される b_2 は有限値であるが、直ちに $\mu_4 < \infty$ を意味するわけではない。
一般に大きさ n の独立標本 $X_1, X_2, \cdots, X_n$ とする。n 個の標本から計算することのできる統計量、未知母数 θ とすると、その推定量としては一般には様々な方法が考えられるが、母数に対応する標本積率を利用するのが積率法 (moment method) と呼ばれている。この積率に基づく方法についての一つの問題は真の確率分布が存在するが、未知の状況では、例えば中心 0, スケール 1 の Cauchy 分布

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2} \ \ (-\infty < x < +\infty)$$

のように一見すると正規分布に似ていても積率が存在するとは限らないことである。コーシー分布は極端な例ではあるが $\mathbf{E}[|X|] = +\infty$ となり、$|x| \to +\infty$ のときに密度関数が $c\,x^{-2}$(c は定数)、分布の裾 (tail) がベキ的になる、すなわち極端な値が時々生じ得る[9] ことが最大の特徴と云える。

連続型の確率変数 $|X|$ の分布関数を $F(z)$ とすると、期待値 $\int_0^\infty z df(z)$ が有限であれば $z = \int_0^z 1 \cdot dx$ より積分範囲 ($(0 \le x \le z, 0 \le z)$ は $(x \le z, 0 \le x)$ を意味する。したがって

$$\begin{aligned}\mathbf{E}[|X|] &= \int_0^\infty \int_0^z dx dF(z) \ (= \int_0^\infty \int_x^\infty dF(z)dx \ ) \\ &= \int_0^\infty \mathrm{P}\left(|X(\omega)| > x\right) dx \ .\end{aligned}$$

この等式より期待値が存在する為には裾確率 $P(|X| > z)$ は $|z|$ が大きいとき十分なスピードで減衰する必要がある。高次のモーメント $\mathbf{E}[|X|^r]$ $(r \ge 2)$ が存在するには、より急速に裾が減衰する必要があることも分かる。

正規分布の密度は $c\ \exp[-(1/2)x^2]$ であるから急減少関数の一種であり任意の次数の積率 $\mathbf{E}[|X|^k] < +\infty$ が存在するので、統計的方法を教育するには都合の良い分布と云えよう。しかし現実の課題についてデータ分析する際に想定することが妥当か否かは全くの別の問題である。次節で説明する極端な事象のデータ分析はこうした問題が応用に関わる重要な例と云えよう。その前段階として本節ではデータから視覚的に正規性を調べる統計的方法を説明しておこう。

2. 正規分布の妥当性

統計学基礎では正規分布および正規分布から派生する統計量の説明が重視されている。これは「近代的な統計分析は W.Gosset による t-分布の研究から始まる」という意味の説明を R.A.Fisher が述べたとされるなど統計学の発展の歴史的経緯による事情もあると思われる。

他方、実際のデータ分析ではデータが少なかろうと多かろうとも、正規分布からの標本とみなすことができるか、あるいは近似的に妥当であるか否かは重要な論点なので、様々な統計的方法が開発されている。ここでは「統計学基礎」でごく簡単に言及されている

[8] Kendall and Stuart,"The Advanced Theory of Statistics," 4th edition, 1977, Vol.1, Charles Griffin & Company Limited, p.325-326 に基づき説明したが 高次の積率の存在条件などを仮定する必要がある。国友 (2015) Page 109 の b_2 の漸近分散を僅かに修正した。

[9] R 上でのコーシー乱数 rchaucy(乱数の数) を発生させて確かめてみると良い。

積率, 正規 QQ プロットなど経験分布に基づくグラフィカルな方法、特に QQ プロットなどの統計的な方法について言及しておく。また回帰分析への応用であるR計算プログラムの利用法も合わせて説明しておく[10]。

統計分析では実際に得られる観測値が独立な確率変数列 X_i $(i = 1, \cdots, n)$ が同一のある確率分布 $F(\cdot)$ からの標本と考えることから出発する[11]。この確率分布 $F(\cdot)$ を正規分布と見なしてよいかどうか調べるには、得られるデータから計算できる経験分布

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(X_i \leq x)$$

が基本的である。ここで指標関数 $\mathbf{I}(\omega) = 1$ (ωが実現), 0 (ωが実現しない) である。$n \to +\infty$ のとき大数の (弱) 法則により任意の x に対して $F_n(x) \xrightarrow{p} F(x)$ となる。例えばRでは経験分布関数
data=rnorm(100)
Fn=ecdf(data)
plot(Fn)
により確かめることができる。
同様に正規 QQ プロットは
qqnorm(data)
により実行できる。

経験分布については中心極限定理より $\sqrt{n}[F_n(x) - F(x)] \xrightarrow{w} N(0, F(x)(1 - F(x)))$ となる。実用上では視覚的に分かりやすい QQ プロットが利用されることが多いが、信頼区間を構成するには順序統計量 (ordered statistics) の評価が必要である。互いに独立・同一分布にしたがう確率変数例 X_i $(i = 1, \cdots, n)$ から大きさで順位をつけた順位統計量 $X_{(1)} \leq \cdots \leq X_{(n)}$ としよう。分布関数 F が連続型の場合には変換した確率変数列 $F(X_k)$ は $P(F(X_k) \leq x) = P(X_k \leq F^{-1}(x)) = F(F^{-1}(x)) = x$ より一様分布にしたがうが、順序統計量を変換すると $F(X_{(k)})$ はベータ分布 $B(k, n+1-k)$ にしたがう。そこでベータ分布の α 分位点を b_α とすると、$P(b_{\alpha/2} \leq F(X_{(k)} \leq b_{1-\alpha/2}) = 1 - \alpha$ となるので $X_{(k)}$ の $1 - \alpha$ 信頼区間は $[F^{-1}(b_{\alpha/2}), F^{-1}(b_{1-\alpha/2})]$ で与えられる[12]。x 軸として $F^{-1}(\mathrm{rank}(X_i)/(n+1))$（理論的分位点）, y 軸には X_i の経験分位点を図にしたものが QQ プロットである。特に F として正規分布とすると正規 QQ プロットが得られる。

< 順序統計量の分布 >

順序統計量 (order statistc) は分位点 (quantile) などの分析を通じて統計学の発展で重要な役割をはたしていることもあり簡単に説明しておこう。例えば最大値、最小値、中央値 (median) などは特別な順序統計量である。例えば竹内 (1963) が説明しているが、第 k 順序統計量 $X_{(k)}(k = 1, \cdots, n)$ について $P(X_{(k)} \leq x)$ は X_i の中で少なくとも k 個が x

[10] x 軸と y 軸に分布 F と分布 G の分位点をとり作成する図。説明は例えば
https : //en.wikipedia.org/wiki/Q − Q_plot を参照。

[11] 統計的時系列分析、統計的確率過程分析、統計的時空間分析などでは独立性が成り立たない状況を明示的に議論している。

[12] F^{-1} は逆関数であるが left-continupus inverse 右連続逆関数 $F^{\leftarrow}(x) = \inf\{y|F(y) \geq x\}$ を用いて類似の議論ができる。

を超えない確率であるから、h 個 ($h \geq k$) が x を超えない確率の和として、

$$F_k(x) = \sum_{h=k}^{n} {}_nC_h [F(x)]^h [1 - F(x)]^{n-h}$$

である。密度関数は微分して整理すると

$$\begin{aligned} f_k(x) &= \sum_{h=k}^{n} {}_nC_h \frac{d}{dx} [F(x)]^h [a - F(x)]^{n-h} \\ &= \sum_{h=k}^{n} {}_nC_h f(x) [h(F(x))^{h-1}(1 - F(x))^{n-h} \\ &\qquad -(n-h)(F(x))^h (1 - F(x))^{n-h-1}] \\ &= \frac{1}{B(k, n+1-k)} f(x) [F(x)]^{k-1} [1 - F(x)]^{n-k} \end{aligned}$$

となる。一様分布なら $f(x) = 1$, $F(x) = x$ $(0 \leq x \leq 1)$ よりベータ分布となる。

<Kolmogorov-Smirnov 統計量 >
コルモゴロフ・スミルノフ統計量は分布の妥当性、確率論との関連において重要な意味がある。経験分布 F_n は大数の (強・弱) 法則より n が大きい時に真の分布に収束する。しかし一般には確率分布関数は x の関数なので各点で収束することは必ずしも全体が収束するとは限らない。この問題については、

$$D_n = \sqrt{n} \sup_x |F_n(x) - F(x)|$$

が $n \to$ のとき $D^* = \sup_{t \in [0,1]} |B^{(o)}(t)|$ (ここで $B^o(t)$ は Brownian Bridge と呼ばれる $[0, 0]$ から $[1, 0]$ への経路を持つ正規連続確率過程, 定義は付論を参照) に弱収束することが知られている。確率変数 D^* の分布関数は

$$G(x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp[-2k^2 x^2]$$

で与えられる。
この結果 (極限分布は元の分布 F に依存していない) は確率論における弱収束 (weak convergence, invariance principle) の理論を用いた[13]統計学への最初の応用例と思われる。その後に Anderson-Darling 統計量を始め様々な統計的方法が開発されるなど統計学の展開に大きな影響があった。

なおデータ系列 data が与えられたとき、R では
ks.test(ksdata,"pnorm",mean=mean(data),sd=sd(data))
とすれば Kolmogorov-Smirnov 検定が可能である。また二つのデータ系列 data1, data2 の同等性は
ks.test(data1,data2)
とすれば簡単に実行でできる。この kolmogorov-Smirnov 検定はノンパラメトリック検

[13]例えば Billingsley, P. (1999),"Convergence of Probability Measures," Wiley, 2nd edition, Page 103, ただし同書では和の範囲に誤植がある。

定の一種であるが、単に検定統計量を求めるよりも基準化して経験分布を比較する方がデータ分析には重要と思われる。

3. 応用例：回帰残差分析のプログラム例

回帰分析の残差から誤差項の正規性について QQ プロットを用いて調べることができるので、R-プログラムを例示しておこう。分散が均一 σ^2 正規分布にしたがい、p 個の説明変数がある時、残差系列 e_i $(i = 1, \cdots, n)$ は $e_i \sim N(0, \sigma^2(1 - p_{ii}))$ $(i = 1, \cdots, n)$ となる。(ここで p_{ii} は説明変数から構成される射影行列（$\mathbf{P}_Z$: $((p+1) \times (p+1))$ ベキ等行列）の第 (i, i) 成分 $(i = 1, \cdots, n)$ である。このとき $[\sum_{j=1}^{n} e_j^2 - e_i^2/(1 - p_{ii})]/\sigma^2$ は e_i とは独立に自由度 $n - p - 1$ の χ^2 分布にしたがうので、スチューデント化した残差

$$t_i = \frac{\sqrt{n-p-1}e_i}{\sqrt{\sum_{j=1}^{n} e_j^2 - e_i^2/(1-p_{ii})}\sqrt{1-p_{ii}}} \ (i = 1, \cdots, n)$$

は自由度 $n - p - 1$ の t 分布にしたがう[14]ことなど標準的な仮説検定を巡る問題については同書を参照されたい。

ここでは視覚的に妥当性を調べるために R-プログラムを作成した。このプログラムは次に挙げる https(2023.8.22 現在) 上にある説明にもとづくものである。

https://lbelzile.github.io/lineaRmodels/qqplot.html

文献

[1] 統計学基礎 (改訂版), 日本統計学会編, 2015, 東京図書.

[2] (応用をめざす) 数理統計学 , 国友直人, 2015, 朝倉.

[3] 数理統計学, 竹内啓, 1963, 東洋経済.

4. 付論：ブラウン運動・ブラウン橋について

1 次元標準ブラウン運動 (Brownian Motion)$\{B_t\}$ とは連続時間の確率過程（初期条件 $B_0 = 0$）であり条件 (i) 任意の区間分割 $0 \leq t_0 < t_1 < \cdots < t_N < +\infty$ に対して，確率変数列 $B_{t_0}, B_{t_1} - B_{t_0}, \cdots, B_{t_N} - B_{t_{N-1}}$は互いに独立、(ii) $B_{t+s} - B_s \sim N(0, t)$ により与えられる。

さらに、$[0, 1]$ 上のブラウン橋 (Brownian Bridge) $B_t^{(0)}$ とは $B_t - tB_1$ で表現される確率過程である。$\{B_t(\omega)\}$ は ω を固定したときの時間 t についての経路が (ジャンプなしの) 連続過程であり、こうした確率過程が存在するか否かは数理的には証明が必要であり、自明ではないが、自然科学を始め、近年では統計学やファイナンス経済学などでも利用されている。

[14]例えば佐和隆光 (1979)「回帰分析」, Page134 を参照。

```
### QQ plot
### https://lbelzile.github.io/lineaRmodels/qqplot.html
### prepared by R. Yuasa (2023.8.23)

# Rにあるmtcarsというデータで回帰分析を行う
ols <- lm(mpg ~ wt, data = mtcars)
n <- length(mtcars$wt)
#スチューデント化残差
esr <- rstudent(ols)

#真の分布がt分布のときの理論的分位点
emp_quant <- qt(rank(esr)/(n + 1),  df = n - 3)
#各点ごとに信頼区間を計算する関数
#引数はn,distの他にも、...の部分に違うものを与える事が出来る
confint.qqplot.ptw <- function(n, dist = "norm", ...){
  #sapply関数は第1引数にデータ、第2引数に関数を与える
  #1:nというデータをそれぞれ、この場で定義した関数に代入し、その結果を与える
  t(sapply(1:n, function(i){
    #dist='tの場合、'paste0('q','t')は文字列'q'と't'の結合を行い、"qt"を返す
    #do.callは引数のリストを関数に与えて実行する
    #qbeta関数によりBeta(i,n-i+1)の2.5%点と97.5%点を計算
    #dist='tの場合、qtという関数にBeta(i,n-i+1)の2.5%点と97.5%点と...の部分(下で
はdf=n-3)を与える
    do.call(paste0('q', dist), list(qbeta(c(0.025, 0.975), i, n - i + 1), ...))
  }))
}

#上の関数を用いる
confint_lim <- confint.qqplot.ptw(n = n, dist = "t", df = n - 3)
#経験分位点に沿って信頼区間をプロットする
#matplotは2つの行列に対し、1つ目の行列のk列目の値をx軸、2つ目の行列のk列目の値を
y軸に取るような線を引く
#kは1から行列の列数
#今回は1つ目の行列として与えているのがベクトルで、x軸にはベクトルの値が用いら
れ、y軸には2つ目の引数で与えられている行列の各列の値をとるような線が引かれる
matplot(sort(emp_quant), confint_lim, type = "l", lty = 2, col="grey",
        main = "Q-Q plot", xlim = c(-2, 2), ylim = c(-2, 2),
        xlab = "理論的分位点", ylab = "経験分位点")
#切片0傾き1の理論直線のあてはめ
abline(a = 0, b = 1)
#観測をプロットする
points(esr, emp_quant, pch = 20)
```

第 3 章　確率分布の裾と極値現象の分析 [15]

2023-11-29 国友直人

1. はじめに

我々を取り巻く自然現象、経済現象、生命現象の中には稀にしか起こらないが、一たび起きると大きな災害として我々の社会や日常生活に深刻な影響を及ぼすことがある。幾つかの身近な例を挙げておくと、大雨や台風などの風水害、大きな地震、などの自然現象、企業や国の信用力の低下による債務不履行や倒産、株価の急落、などの経済現象はそれほど頻繁ではないが観察されている。一例として 2011 年 3 月 11 日直後の日本において経験した事情の一端を示した 2 枚の写真を下に掲示しておく[16]。こうした自然現象や経済現象を初等統計学で議論されている正規分布をそのまま利用してデータ分析することには基本的に無理があるが、実はその理由や確率分布を考察することから (統計的) 極値論という分野が発展してきている。ここでは統計学を学んだが統計的極値論 (statistical extreme value theory, SEVT) をあまり聞いたことがない応用分野の若手研究者を主な対象に、統計的極値論においてこれまで議論されてきている論点と幾つかの基本事項を解説しよう。

図 1:2011.3 東日本大震災 1

図 2:2011.3 東日本大震災 2

なお統計検定 2 級の教科書「統計学基礎」2 章では多くの教科書と同様、確率分布と期待値・分散を説明、その後、積率 (モーメント) から歪度 (skewness)・尖度 (kurtosis) および幾つかの確率分布の説明がある。なぜ期待値と分散だけではない積率が意味があるのか、説明は曖昧と思われる。また、確率分布の裾 (tail) と極値 (extreme value) に関する統計的分析法についての記述は全く見受けられない。そこで確率分布・積率と統計的データ分析について統計エキスパートなら理解しておくべきと思われる確率分布の裾と極値を巡る論点について「2 章　確率分布の積率・裾と正規性」でとりあげた話題の続きとしてあまく長くならない範囲で説明を試みる。
一次元の記号として確率分布関数 $F(x)$, 確率関数 $p(x)$, 密度関数 $f(x)$ $(= \frac{dF(x)}{dx})$、積率を $\mathbf{E}[X^k]$, $\mu_k = \mathbf{E}[(X - \mathbf{E}(X))^k]$ $(k = 1, 2, \cdots)$ とする。

議論の出発点として「確率分布の積率・裾と正規性」から次の等式を引用しておこう。連

[15] 統計エキスパート slack 上で行われた統計的推測の基礎についての素朴な疑問や議論の中で統計エキスパート養成事業にとり有用と思われる内容をまとめようと作成したメモである。

[16] 2011 年東日本大震災を経験したにも関わらずごく一部の統計家を除き、日本では「統計検定の教科書」を含め、問題の理解や関心があまり高くないようなので本稿を書いた。

続型の確率変数 $|X|$ の分布関数を $F(z)$ とすると、期待値 $\int_0^\infty zdF(z)$ が有限であれば

$$\mathbf{E}[|X|] = \int_0^\infty \mathrm{P}\left(|X(\omega)| > x\right) dx \,. \tag{1}$$

この等式から期待値が存在する為には裾確率 $P(|X| > z)$ は $|z|$ が大きいとき十分なスピードで減衰する必要がある。高次のモーメント $\mathbf{E}[|X|^r]$ $(r \geq 2)$ が存在するには、より急速に裾が減衰する必要があることも分かる。

正規分布の密度は $c\ \exp[-(1/2)x^2]$ であるから急減少関数の一種であり任意の次数の積率 $\mathbf{E}[|X|^k] < +\infty$ が存在するので、統計的方法を教育するには都合の良い分布と云えよう。しかし現実の課題についてデータ分析する際に想定することが妥当か否かは全くの別の問題である。次に述べる極端な事象のデータ分析はこうした問題が応用に関わる重要な例である。

数学的期待値 $\mathbf{E}[|X|] < \infty$, あるいは発散、分散 $\mathbf{E}[|X|] < \infty$, または発散、という両方の状況が極値分析では重要である。しかしながら期待値や分散が発散する場合は中心極限定理が成り立たないなど、普通の統計学の道具立ては使えない[17]。したがって、統計的方法の開発は困難、したがってこの間、一部の理論家の関心が高かったが日本の多くの統計家とはかなり距離がありそうである。他方、極値のデータ解析が応用上でどこまで有用であるかについては意見が分かれる所だろう。ここでは極値統計学とロバスト統計学に関する一つの論点を 3 節で簡単に言及しておく。背景となる確率論の基礎的事実を付論で簡単に言及するにとどめる。

2. 極端な事象とリスク評価

入門統計学では確率変数 X がある範囲 $(a, b]$ に入る確率を計算する方法として正規分布を利用して

$$\mathbf{P}\left(a < X \leq b\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

と評価する統計的方法が説明されている。この種の計算では $[\mu - k\sigma, \mu + k\sigma]$ の確率は $k = 1, 2, 3$ に対し 68%,95%,99.9% などである。区間外の確率は $k = 5$ では 10^{-6},$k = 6$ では 10^{-23} などとなり裾確率は非常に小さいのが特徴である。例えば標本算術平均 $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ とすると中心極限定理 $\frac{\sqrt{n}}{\sigma}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0,1)$ により正規分布の利用を正当化されることが多い。正規分布はよく知られているように裾確率は図のように急速に減衰する。正規分布と好対照に裾の厚い確率分布としてパレート分布の密度関数 $f(x) = k/x^{k+1}$ $(x \geq 1)$ を図 4 に示しておく。

統計的分析では突発的な現象や極端な現象 (extreme events) を分析することも重要である。例えば伝統的な (河川・海岸・建築・腐食・破壊現象) 工学、あるいは損害保険の分野では風水害、地震、事故など時々しか観察されない稀な事象 (rare events) の分析が議論されている。また経済・金融でも銀行・保険会社では VaR(value-at-risk, バリューアットリスク) と呼ばれる統計的金融リスク管理が金融業務上で必要である。この VaR 管理では通常は 1 日単位で金融機関が保有する価値の損失分布の左側 1%、5% の管理を問題とするが、実際にデータ上で 1% の確率事象が実現することは頻繁には起こらない。観察される日々のデータから稀な事象が分析対象であり、正規分布の利用により 1990 年

[17] この場合については確率論においてかなり調べられている。例えば無限分解可能分布や安定分布についての基礎的事実は付論を参照されたい。

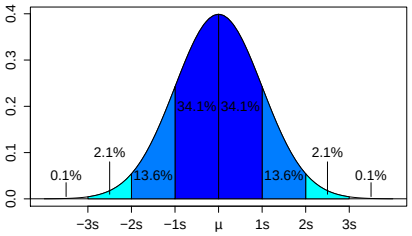

図 3:正規分布と正規確率

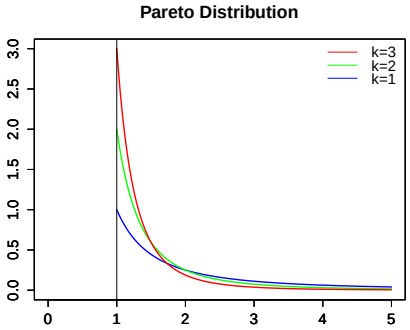

図 4:パレート分布と裾

代のバブル崩壊時には金融機関や規制当局における金融リスクの過小評価という重大な問題が生じたと考えられる。

稀に起こる現象の説明として時々「新しいベキ法則を発見した！」と主張されることがあることに、統計エキスパートとしては注意する必要がある。ベキ分布は少なくとも V. Pareto (1848-1923)、古典的極値論は少なくとも Fisher-Tippet の論文 (1927 年) まで遡る。統計学における稀な現象の分析は統計的極値論 (statistical extreme value theory) と呼ばれいるが、オランダにおける 1950 年代の堤防の決壊により国土のほぼ 1/3 が大災害に見舞われ、堤防の修復の必要性などが大きな契機となり理論が発展した。その時の重要な課題は災害を防ぐ為には堤防の高さをどの程度にすることが科学的と言えるかであった。一方で堤防の高さを十分に高くすれば良いと考えられるが、他方では堤防を作るにはかなりな予算が必要となる。こうした問題は建物の耐久性、ダムや河川管理の問題などの工学分野を中心に幅広い応用の可能性がある。

最大値データの裾確率の評価に関連して 3 つのタイプの極値分布 (extreme value distribution) があることが Fisher-Tippet の定理として知られている[18]。i.i.d. 確率変数列 X_i $(i = 1, \cdots, n)$ とすると最大値 $M_n = \max\{X_1, \cdots, X_n\}$ に対して適当な数列 a_n, b_n が存在して確率 $P((M_n - a_n)/b_n \leq z)$ が分布関数 $G(z)$ に収束するときには 3 タイプの極値分布 (フレッシェ, グンベル, ワイブル) しかありえないことを示したことが定理の内容

[18]正確な数理的内容は de Haan, L. and Ferreira, A. (2006), "Extre,me Value Theory : An Introduction," Springer を参照。

である。極値分布は期待値の周りに左右対称で釣鐘型、裾が急速に減衰していく正規分布とはかなり異なる確率分布であり、これらの確率分布を利用すると裾確率の評価はかなり異なりうる。3 つのタイプの極値分布をまとめたのが、位置母数 μ, スケール母数 σ を含めて一般化極値分布 (GEVD, generalized extreme value distribution)

$$G(x) = \exp\left\{-[1+\xi(\frac{x-\mu}{\sigma})]^{-1/\xi}\right\} \tag{2}$$

である。(しばしば $\mu=0$,$\sigma=1$ と基準化されて表現される。) スケール $\sigma=1$ として裾指数 ξ について $\xi \to 0$ とするとグンベル型分布

$$G(x) = \exp[-\exp(-x)] \quad (-\infty < x < +\infty) \tag{3}$$

が得られる。こうした極値分布の利用は極値統計学 (statistical extreme value theory) の出発点である。数理的には正則変動関数論 (regulary-varing function) によりかなり詳細な研究が行われてきている。中心極限定理は確率変数の和の挙動についての確率法則であったが、極値分布は確率変数の最大値・最小値の挙動についての確率法則であり、極端な事象、稀に起きる事象のリスク評価などに有用と考えられる[19]。

極値分布の例

例えば指数分布 Exp(1) に従う確率変数の場合には、$F(x) = 1-e^{-x}$ $(0 \le x)$ であるが、$a_n = 1, b_n = \log n$ とおけば

$$\begin{aligned}\mathbf{P}\left(\frac{M_n - b_n}{a_n} \le x\right) &= [F(x+\log n)]^n \\ &= [1-\exp(-(x+\log n))]^n = \left[1-n^{-1}\exp(-x)\right]^n \to \exp(-e^{-x})\end{aligned}$$

となる。極限分布は Gumbel 分布であり裾指数 $\xi = 0$ となる。

パレート分布に従う確率変数の場合には、$F(x) = 1-[k/(k+x)]^\alpha$ $(0 \le x; \alpha > 0, k > 0)$ とすると、$a_n = kn^{1/\alpha}, b_n = 0$ とおけば

$$\mathbf{P}\left(\frac{M_n - b_n}{a_n} \le x\right) = \left[F(kn^{1/\alpha}x)\right]^n = \left[1-(\frac{1}{1+n^{1/\alpha}x})^\alpha\right]^n \to \exp[-x^{-\alpha}]$$

となる。極限分布は Frechet 分布であり裾指数 $\xi = 1/\alpha$ となる。

一様分布 $U(0,1)$ に従う確率変数の場合には、$F(x) = x$ $(0 \le x \le 1)$ であるが、$a_n = 1/n, b_n = 1$ とおけば

$$\mathbf{P}\left(\frac{M_n - b_n}{a_n} \le x\right) = \left[F(n^{-1}x+1)\right]^n = \left[1+\frac{x}{n}\right]^n \to e^x$$

となる。(2) において $\xi = -1$, $\sigma = 1$, $\mu = -1$ とすると Wiebull 分布 $\exp[-(-x)^\alpha]$ $(x \le 0, \alpha = 1)$ に収束する。この極限分布は右裾が有限となる。

閾値極値論

3 つのタイプの極値分布を応用することは可能であるが、グループ別の最大値がデータと

[19] 統計的極値理論 (extreme value theory, EVT) の基本事項については高橋・志村 (2016) がある。定評のある英語の書籍は Resnick (2021) などかなりある。

して得られる必要がある。例えば毎年繰り返される河川の氾濫、海岸に打ち寄せられる波浪など一部の自然現象ではこうしたデータが得られることはある。しかしながら、金融データのような明確な周期性の発生が期待できない場合には、グループ・データの最大値はあまり分析に適しているとは考えにくいので、ある閾値を超えたデータ、閾値極値の解析が考えられている。

ここで確率変数 X が分布関数 F にしたがうとする。超過分布関数を

$$F_u(x) = \mathbf{P}\left(X \leq x + u | X > u\right) \quad (x \geq 0)$$

により定める。このとき条件付確率より

$$F_u(x) = \frac{F(x+u) - F(u)}{1 - F(u)}$$

である。仮に元の確率分布 $F(u)$ が一般化極値分布にしたがうとすれば、u が大きいとき裾確率を

$$1 - F(u) \sim [1 + \xi(\frac{u-\mu}{\sigma})]^{-1/\xi}$$

と表現できる。同様に $1 - F(u+x) \sim [1 + \xi(\frac{u+x-\mu}{\sigma})]^{-1/\xi}$ である。特に位置母数 $\mu = 0$ として条件付確率を適当な母数として $\sigma(u)$ $(= \sigma + \xi(u - \mu))$ とすると

$$P(X > u + x | X > u) \sim \quad \frac{[1 + \xi(\frac{u+x}{\sigma})]^{-1/\xi}}{[1 + \xi(\frac{u}{\sigma})]^{-1/\xi}} = \left[1 + \xi\frac{x}{\sigma(u)}\right]^{-1/\xi}$$

と表現できる。このことをまとめると Balkema-deHaan の定理[20]として次のような結果として知られている。

定理 1： $X_1, \cdots, X_n$ を独立で同一の確率変数列で分布を F とする。このとき基準化数列 $a_n > 0, b_n \in \mathbf{R}$ とある極値分布 $H_{\xi,\mu,\sigma}$, 右裾 $x_F = +\infty$ とする。このときある非負関数 $\sigma(u)$ が存在して

$$\lim_{u \to x_F} |F_u(x) - G_{\xi,\sigma(u)}(x)| = 0$$

であり極限分布は $\xi \neq 0$ のとき

$$G_{\xi,\beta}(x) = 1 - \left[1 + \xi\frac{x}{\sigma(u)}\right]^{-\frac{1}{\xi}} \quad (\xi \neq 0) \tag{4}$$

である。また $\xi \to 0$ のときには

$$G_{\xi,\beta}(x) = 1 - \exp\left[-\frac{x}{\sigma(u)}\right] \tag{5}$$

となる。

ここで分布関数 G は一般化パレート分布 (generalized Pareto distribution, 略して GPD) と呼ばれるが、確率分布の裾がパレート分布、すなわちべき関数的に減衰することを意

[20]正確には Balkema, A. and de Haan, L. (1974), "Residual life time at great age," 2-5, *Annals of Probability* を参照。

味するからである。例えば指数分布に従う確率変数の場合には、$F(x) = 1 - e^{-x}$ $(0 \leq x)$ であるので

$$F_u(x) = \frac{F(x+u) - F(u)}{1 - F(u)} = \frac{1 - e^{-(x+u)} - (1 - e^{-u})}{1 - (1 - e^{-u})} = 1 - e^{-x}$$

となるので $\xi \to 0$ の場合に対応する。また一様分布 $U(0,1)$ に従う確率変数の場合には、$F(x) = x$ $(0 \leq x \leq 1)$ より

$$F_u(x) = \frac{F(x+u) - F(u)}{1 - F(u)} = \frac{x + u - u}{1 - u} = \frac{x}{1 - u}$$

となる。

分位点と母数の推定

一般化パレート (GPD) 分布 $G_{\xi,\sigma}(x)$ については $\mu = 0$ としておくと、$X \sim G_{\xi,\sigma}(x)$ のとき $F_u(x) = G_{\xi,\sigma+\xi u}(x)$ となることは有用である。また平均超過関数 (mean excess function) を $\mathbf{E}[X - u | X > u]$ により定義すると

$$\mathbf{E}[X - u | X > u] = \frac{\sigma + \xi u}{1 - \xi} \quad (0 \leq \xi < 1)$$

より閾値 u の線形関数となるという性質がある[21]。ここで仮に $F_u(x)$ に対する極限分布 $G_{\xi,\sigma(u)}(x)$ の $\sigma(u)$ $(= \sigma + \xi u)$ を一定値 β で近似できるとき極値分布を利用して高分位点の推定法として閾値 u を用いて

$$\frac{F(x+u) - F(u)}{1 - F(u)} \sim 1 - \left[1 + \xi \frac{x}{\beta}\right]^{-\frac{1}{\xi}}$$

が利用できる。すなわち $1 - F(x+u) \sim [1 - F(u)] \left[1 + \xi \frac{x}{\beta}\right]^{-\frac{1}{\xi}}$ より、データ n 個の中で閾値 u を超える超過個数データが N_u 個であったとすると、裾確率を $1 - F(u)$ を N_u/n で推定すれば

$$F(x+u) \sim 1 - \frac{N_n}{n} \left[1 + \xi \frac{x}{\beta}\right]^{-\frac{1}{\xi}} \tag{6}$$

である。したがって、$\xi > 0$ であれば $1 - p$ が小さいときに $F(x+u) = p$ を利用して

$$\hat{x}_p = \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{n}{N_u}(1-p) \right)^{-\hat{\xi}} - 1 \right]$$

より高分位点を母数 ξ,β の推定値を利用すると $u + \hat{x}_p$ により推定できる。ここで $\xi > 0$ の場合はフレッシェ型の極値分布であり、元の確率分布が裾が厚い場合の極値分布に対応していることが興味深い。($\xi < 0$ は右裾が有限でワイブル型の場合である。) また (6) において $\xi \to 0$ とすると裾指数 ξ の値によらないグンベル型分布を利用して高分位点を推定することもできる。

[21] 確率変数 Y の密度関数を $h(y) = [1 + (y\xi/\tilde{\sigma})]^{-1/\xi - 1}(1/\tilde{\sigma})$ とする (右裾は ∞ の場合を扱う)。変換 $Z = 1 + Y\xi/\tilde{\sigma}$ より $\mathbf{E}[Y] = \int_0^\infty y h(y) dy = (\tilde{\sigma}/\xi^2) \int_1^\infty [z^{-1/\xi} - z^{-1/\xi - 1}] dz = [\tilde{\sigma}/\xi^2][-\xi/(\xi - 1) - \xi] = \tilde{\sigma}/(1 - \xi)$ となる。したがって $\tilde{\sigma} = \sigma + \xi u$ より結果が得られる。

応用例としては先ほど言及した挙げた 1954 年にオランダで起きた堤防の決壊をきっかけに堤防の高さに関する議論がされたことが特筆されよう。(詳しくは De Haan and Ferreira (2006)1.1.4 節に説明されている。) 当時の議論は 1 万年に 1 度 (10^{-4}) というレベルの堤防の高さをいかなる科学的基準から設計しようと云うものであるが、これは一種の外挿の問題である。この議論から閾値統計学が発展したという事実が興味深いことを指摘しておく。同書にはその他の例も挙げているが、他の分野としては損害保険業におけるソルベンシーの国際規制と呼ばれている問題なども重要な応用と言えるだろう。

極値問題では裾指数 ξ の推定がもっとも重要な問題であるが、例えば一般化パレート分布を仮定して（極値分布が正しいと仮定して）最尤推定を行うことなどが考えられるが、実際の数値計算上では $\xi = 0$ の周辺に注意する必要がある。
分布の裾が厚く母数 $\xi > 0$ のとき裾確率 $P(X > x) \sim cx^{-\alpha}$ ($\alpha = 1/\xi > 0$ と表現するとセミパラメトリック推定法として順序統計量 $X_{(n)} \geq X_{(n-1)} \geq \cdots \geq X_{(1)}$ を利用した

$$H_{n,k} = \frac{1}{k} \sum_{j=0}^{k-1} \log X_{(n-j)} - \log X_{(n-k)}$$

となる Hill 推定量が知られている。また Pickand 推定量は

$$P_{n,k} = \frac{1}{\log 2} \log \left[\frac{X_{(n-k)} - X_{(n-2k)}}{X_{(n-2k)} - X_{(n-4k)}} \right]$$

で与えられる。
Hill 推定量については $\bar{F}(x) = cx^{-\alpha} \quad (x > x_0 > 0, \alpha > 0)$ のとき $k(n) \to \infty$ かつ $k(n)/n \to 0$ などの一定の正則条件の下で漸近的に一致性 $H_{n,k} \xrightarrow{p} 1/\alpha$, さらに $\sqrt{k}[H_{n,k} - (1/\alpha)]$ の $N(0, \alpha^{-2})$ への漸近正規性が成立する[22]。したがって $\sqrt{k}[H_{n,k}^{-1} - \alpha] \xrightarrow{w} N(0, \alpha^2)$ となる。ただし、Hill 推定量のバイアスは小さくなく、また推定の際に利用する順序統計量の数 k が重要な問題である。観測されるデータから k を選択する方法はかなり難しいが幾つかの方法が提案されている。

3. 極値統計学 vs. ロバスト統計学

実際のデータを生成している確率分布が正規分布と見なせない場合、選択肢は少なくとも 2 通りあると思われる。データの一部がはずれち値 (outlier, 異常値) と見なせる典型例としては数値の記入ミスなどがあるだろう。実際、公的統計ではしばしば記入ミスをチェックするためのデータ審査という過程を経て公表値が作成されているが、多数のデータの中に記入ミスが皆無とは言い切れない。こうした outlier が存在する場合の推定問題ではかりに outlier が存在しても頑健な (robust) 結果が得られる様々な統計的方法が開発されている。例えばデータの中心 (center) についてのロバストな推定法として標本平均 (arithmetic mean) ではなく中央値 (median) の利用が提唱されることがある。(ロバスト統計学については例えば藤澤 (2017) を挙げておく。)

これに対して公的統計、愛知県の企業調査を行うと必ず得られる 1 つの outlier の扱いが良い例を提供していると言えるだろう。統計調査の目的が愛知県の設備投資を推計したい場合には外れ値をあらかじめデータから取り除くと大きなバイアスが生じるだろう。この例はデータ上で観察される極端なデータの統計的分析はデータ分析の目的に依存す

[22]証明はかなり複雑になる。例えば Resnick(2021)9 章を参照されたい。

ることを示唆している、と云えよう。

4. 極値分布を巡るプログラム例

R の中には Parero 分布にしたがう乱数は用意されていないので $X \sim \text{Pareto}(a, b)$ のとき $U = F(X)$ は一様乱数にしたがうことを利用して次のように実行すればよい。

a=1.5
b=1.0
data=runif(1000)
data2=$b/(1 - \text{data})^{1/a}$
hist(data2)

とする。なお Pareto 分布は $a > 1$ なら $\mathbf{E}[X] = ab/(a-1)$, $a > 2$ なら $\mathbf{Var}[X] = ab^2/[(a-1)^2(a-2)]$ である。

Resnick にしたがい裾係数 a の次の推定プログラムにより Hill 推定量を行う。次の関数 Hilla(x) はデータ・セット x から a を推定するために Hill プロットを与える関数である。

```
Hilla =function(x)
{
ordered = rev(sort(x))
ordered = ordered[ordered[] ¿ 0]
n = length(ordered)
loggs = log(ordered)
hill = cumsum(loggs[1:(n - 1)])/(1:(n - 1)) - loggs[2:n]
hill = 1/(hill)
plot(1:length(hill), hill, type = ”l”,
xlab = ”順序統計量の数”,
ylab = ”a の Hill 推定値”, main=”Hill プロット”)
}
```

なお統計的極値論を利用するための R プログラムの幾つかは高橋・志村 (2016), Resnick (2021) に説明、掲載 (ダウンロード可能) されている。

文献

[1] 統計学基礎 (改訂版), 日本統計学会編, 2015, 東京図書.

[2] (応用をめざす) 数理統計学, 国友直人, 2015, 朝倉書店.

[3] 極値現象の統計分析 (Heavy Tail Phenomena), S. Resnick（国友・栗栖, 翻訳), 2021, 朝倉書店.

[4] 極値統計学, 高橋倫也・志村隆彰, 2016, 近代科学社.

[5] ロバスト統計 (外れ値への対処の仕方), 藤澤洋徳, 2017, 近代科学社.

[6] de Haan and Ferreira (2006), *Extreme Value Theory, An Introduction*, Springer.

付論：安定分布と無限分解可能分布

2 項分布の近似に端を発した中心極限定理 (CLT) は歴史的には様々な方向に拡張が試みられた。しかし、例えば確率変数の分散 $\mathbf{E}[Z^2] = \infty$ のときは和の分布はどうなるのであろうか？例えばコーシー分布をとると $\mathbf{E}[|Z|] = \infty$ で与えられる。この場合には特性関数[23] は $\varphi(t) = \mathbf{E}[\exp(itZ)] = e^{-|t|}$ となる（なお $i^2 = -1$ である)。したがって、もし独立な確率変数 Z_i $(i = 1, \cdots, n)$ がコーシー分布にしたがっていれば和 $S_n = \sum_{i=1}^n Z_i$ の特性関数は $\varphi_{S_n}(t) = e^{-n|t|}$ であるから確率変数 $Y_n = S_n/n$ の特性関数は $\varphi_{Y_n}(t) = e^{-|t|}$ となり再びコーシ分布にしたがう。ここで注目すべき点は基準化は $S_n/\sqrt{n}$ ではなく $S_n/n^{1/\alpha}$ $(\alpha = 1)$ となることである。これに対して中心極限定理 (central limit theorems, CLT) は $\alpha = 2$ の場合に対応する。

コーシー分布は期待値も発散する極端な確率分布と考えられるのでもう少し現実的に期待値は定義できるが分散 $\mathbf{E}[Z^2] = \infty$ となる場合を考察しよう。

例 A.1： 確率分布として $P(Z > z) = P(Z < -z)$ $(z > 0)$ および $P(|Z| > z) = z^{-\alpha}$ $(z \geq 1, 0 < \alpha < 2)$ を取りあげる。特性関数 $\varphi(t)$ を考察すると

$$\begin{aligned} 1 - \varphi(t) &= \int_1^\infty (1 - e^{itz}) \frac{\alpha}{2} z^{-\alpha-1} dz + \int_{-\infty}^{-1} (1 - e^{itz}) \frac{\alpha}{2} |z|^{-\alpha-1} dz \\ &= \alpha \int_1^\infty \frac{1 - \cos(tz)}{z^{\alpha+1}} dz \\ &= t^\alpha \alpha \int_t^\infty \frac{1 - \cos(u)}{u^{\alpha+1}} du \end{aligned}$$

となる。ここで $\int u^{-\alpha-1} du$ は可積分なので定数 $C = \alpha \int_0^\infty (1 - \cos(u))/u^{\alpha+1} du$ とすると $t \to 0$ のとき $1 - \varphi(t) \sim C|t|^\alpha$ と近似できる。したがって、確率変数 Z_i $(i = 1, \cdots, n)$ が互いに独立にこの分布にしたがっているとき、和 $S_n = \sum_{i=1}^n Z_i$ の特性関数は

$$\begin{aligned} \mathbf{E}[\exp(it \frac{S_n}{n^{1/\alpha}})] &= \left[\varphi(\frac{t}{n^{1/\alpha}}) \right]^n \\ &= \left[1 - (1 - \varphi(\frac{t}{n^{1/\alpha}})) \right]^n \\ &\to e^{-C|t|^\alpha} \quad (n \to \infty) \end{aligned}$$

となる。こうした評価より分布の裾が厚い場合には一般に確率変数和の極限分布の特性関数は定数 C_2, C_3 を利用して

$$\varphi_Z(t) = \exp\left[\quad C_2 \int_1^\infty (e^{itz} - 1) \frac{dz}{z^{\alpha+1}} + C_3 \int_{-\infty}^{-1} (e^{itz} - 1) \frac{dz}{|z|^{\alpha+1}} \right]$$

となることが期待される。

定義 A.1： 確率変数 Z が無限分解可能分布 (infinitely divisible distribution) にしたがうとは、和の分布が $\mathcal{L}[Z_{n1} + \cdots + Z_{nn}] = \mathcal{L}[Z]$ となる互いに独立・同一分布にしたがう

[23]特性関数は実数 t に対する複素関数 $\varphi(t) = \mathbf{E}[\exp(itZ)]$ で与えられ、分布関数と 1-1 対応する。コーシー分布の特性関数を求めるには複素積分を利用する必要があるが、詳しくは確率論の教科書を参照されたい。

確率変数列 Z_{ni} $(i = 1, \cdots, n)$ が存在することである。

原点での漸近挙動を考慮した表現は次のようにまとめることができる。

定理 A.1： 無限分解可能分布にしたがう確率変数 Z の特性関数は C_0, C_1 (> 0) を定数として

$$\varphi_Z(t) = \exp\left[itC_0 - \frac{C_1}{2}t^2 + \int (e^{itz} - 1 - itz1_D)\nu(dz)\right] \tag{7}$$

と表現される。ここで $D = \{|z| \leq 1\}$, $\int (z^2 \wedge 1)\nu(dz) < \infty, \nu(\{0\}) = 0$ である。($\nu(\cdot)$ は Levy 測度と呼ばれている。)

注意 ：Lévy(レヴィ―) 測度は異なる形で表現されることがある。無限分解可能分布はジャンプ確率過程と密接に関係する[24]。一般に無限分解可能分布は重要な確率分布を含んでいる。例で利用した確率分布の特性関数は安定分布 (stable distribution) の特性関数の形となるが、例では中心部分を無視したことに注意する必要がある。分散が有限でない場合には分布の中心部分の評価と裾の評価を別々に行い、最終的に組み合わせる必要が生じる。安定分布は次のよう定義され、特性関数で表現されることが知られているが詳細は省略する。

定義 A.2： 独立で同一分布にしたがう確率変数列 Z_i $(i, \cdots, n)$ が確率変数 Z の分布と同一であり、ある定数 a_n, b_n が存在し和の分布 $\mathcal{L}[Z_1 + \cdots + Z_n] = \mathcal{L}[a_n Z + b_n]$ となるとき Z の分布は安定分布 (stable distribution) と呼ばれる。

定理 A.2： (i) 安定分布にしたがう確率変数 Z の特性関数は $C_1, C_2, C_3, 0 < \alpha \leq 2$ を定数として次の形で与えられる。

$$\begin{aligned}\varphi_Z(t) \quad = \quad & \exp\left[itC_0 - \frac{C_1 t^2}{2} + C_2 \int_0^\infty (e^{itz} - 1 - itz1_D)\frac{dz}{z^{\alpha+1}}\right. \\ & \left. + C_3 \int_{-\infty}^0 (e^{itz} - 1 - itz1_D)\frac{dz}{|z|^{\alpha+1}}\right]\end{aligned} \tag{8}$$

ただし $D = \{|z| \leq 1\}$ である。
(ii) 特性関数は C_0, d $(d > 0), \theta$ $(-1 \leq \theta \leq 1)$ を定数として次の形になる。
(i) $0 < \alpha < 1, 1 < \alpha \leq 2$ のとき

$$\varphi_Z(t) = \exp\left[itC_0 - d|t|^\alpha \left(1 + i\theta \frac{t}{|t|}\tan\frac{\pi}{2}\alpha\right)\right] , \tag{9}$$

(ii) $\alpha = 1$ のとき

$$\varphi_Z(t) = \exp\left[itC_0 - d|t|\left(1 + i\theta\frac{t}{|t|}\frac{2}{\pi}\log|t|\right)\right] . \tag{10}$$

[24]例えば佐藤健一「加法過程」(紀伊国屋数学叢書) が詳しい。

第 4 章　ヒストグラム再訪 [25]

2023-12-21

国友直人・西颯人

1. はじめに

統計検定 2 級の教科書「統計学基礎」は多くの教科書と同様に最初に記述統計の方法による 1 次元のデータ分析から始まるが、その中でもヒストグラを説明することは重要項目である。実際に n 個の 1 次元データからヒストグラムを作成しようとすると n 個のデータを分類し、階級値、階級幅、階級数を決める必要がある。実は階級数をどうとるか自明なことではなく、ヒストグラムからデータについてグラフィカルな情報はかなりこの設定に依存している。そこで統計家は幾つかの基準を提案しているが、すべての統計家を満足する基準は今のところ存在しない。歴史的にはスタージャスの公式をきっかけに幾つかの方法が提案されてきている。

近年では PC を利用すると統計分析するデータから簡単にヒストグラムや基本統計量が計算できるようになっているので、この問題は統計分析の基本と云えよう。しかし多くの統計パッケージでは自動的にあるルールにしたがって決定しているようである。(特に説明がしばしばない) 暗黙のルールを理解、問題を適切に対処することは「統計エキスパート」の第一歩と云えるだろう。本章はこの問題を取りあげる。

2. スタージェスの公式

非負の整数値をとる n 個のデータ $x_1, \ldots, x_n$ からヒストグラムを作成する方法として，分割の個数 k_n を

$$k_n = 1 + 3.32 \log_{10} n \tag{1}$$

と定めるスタージェスの公式が知られている。Sturges(1926, Journal of the American Statistical Association) の議論を次のように 2 項分布の近似として解釈することができる。なお確率や 2 項分布も基礎事項については「統計学基礎」に説明されているので省略する。2 項分布の定義から

$$\sum_{i=0}^{k_n-1} {}_{k_n-1}C_i \Big(\frac{1}{2}\Big)^i \Big(\frac{1}{2}\Big)^{k_n-1-i} = \Big(\frac{1}{2}+\frac{1}{2}\Big)^{k_n-1} = 1$$

が成り立つので，これを書き換えると $\sum_{i=0}^{k_n-1} {}_{k_n-1}C_i = 2^{k_n-1}$ となる。n 個のデータが k_n 個の階級に分割されるとき，各階級に入る個数を ${}_{k_n-1}C_0, {}_{k_n-1}C_1, \ldots, {}_{k_n-1}C_{k_n-1}$ とすると，$n = \sum_{i=0}^{k_n-1} {}_{k_n-1}C_i$ とおくことができる。従って，$n = 2^{k_n-1}$ という方程式に書き換えられるので，これを解くことによってスタージェスの公式が得られる。

スタージェスの公式は一つの方法であるが，離散分布について二項分布に関するある種の対称性を仮定していると解釈するとある種の妥当性はあるものの、統計的には必ず

[25] 統計エキスパート slack 上で行われた統計的推測の基礎についての素朴な疑問や議論の中から統計エキスパート養成事業にとり有用と思われる内容をまとめようと作成したメモである。「統計学」(久保川達也・国友直人, 2016, 東京大学出版会) 2 章の説明を利用した。

しも最適な方法というわけではない。ここで真の確率分布は連続型と仮定すると、n 個のデータよりヒストグラムを作成する問題は未知の密度関数 $f(x)$ を観測される n 個のデータより推定する問題とみなすことができる。ヒストグラムを密度関数の推定値 $\hat{f}_n(x)$ で表すと，$f(x)$ を $\hat{f}_n(x)$ で推定するときの誤差は

$$MISE = \int \mathbf{E}[\{\hat{f}_n(x) - f(x)\}^2]dx$$

で測ることができる。この量は平均積分 2 乗誤差 (Mean Integrated Squared Error) と呼ばれているが、未知の密度関数 $f(x)$ とデータから計算する密度関数 $\hat{f})x)$ の差なので何らかの意味で小さくすることが望ましい。例えば Scott (1979, Biometrika) は，この量を最小化する問題を考え，区間幅 h_n として

$$h_n = 3.49\frac{s_x}{n^{1/3}}$$

とすることを提案している。ここで s_x はデータの標準誤差である。この選択は真の分布が正規分布である場合には n が大きいときに漸近的に最適な選択になる。他方，Freedman and Diaconis (1981, Probability Theory and Related Fields) は，四分位範囲 IQC を用いて $h_n = 2\text{IQC}/n^{1/3}$ とすることが漸近的に最適であることを主張している。

これら二つの方法はいずれもデータ数 n が大きいときに推定効率が良いことによる正当化に基づいている。また通常の密度関数の推定問題では $n^{1/5}$ が議論されることがあるのに対して $n^{1/3}$ となることが興味深いが、ヒストグラム（階段関数）を用いることから生じるバイアス項の影響によると考えられる[26]。関連する密度関数の推定については例えば国友 (2015, 7 章) に基本的な議論がある。

3. AIC 最小化法

観測されたデータの背後に何らかの真の分布の存在を仮定するとしても、離散分布を考えるアプローチも応用上では有力である。例えば多くの経済・社会・人間について標本として得られる統計データは有限母集団 (finite population) の実現値と見なすことができるだろう。特定の離散確率分布 (discrete probability distribution) についての情報がない場合には多項分布 (multinomial distribution) を想定することが自然である。多項分布はセル数と各セルの確率で定義されるが、観測される n 個のデータから分布の未知母数を決定してデータへフィットした結果をどう評価するかが課題となる。

ここでヒストグラム (度数分布) を多項分布と見なして階級数 c とすると、ある範囲に度数 n_i $(i = 1, \cdots, c)$(総度数 n について $\sum_{i=1}^{c} n_i = n$) が観察される確率は

$$P(n_1, \cdots, n_c | p_1, \cdots, p_c) = \frac{n!}{\prod_{i=1}^{c} n_i!} \prod_{i=1}^{c} p_i^{n_i} \tag{2}$$

で与えられる。特に確率の和が 1 になる条件 $\sum_{i=1}^{c} p_i = 1$ のみを課して尤度関数を最大化する問題はラグランジュ乗数 λ とすると関数

$$L(p_1, \cdots, p_c) = \sum_{i=1}^{c} n_i \log[p_i] + \log n! - \sum_{i=1}^{n} \log n_i! - \lambda[\sum_{i=1}^{c} p_i - 1]$$

[26] こうした理論的論点については例えば Kogure (1990) がまとまって議論している。

の最適化を行えば良いので $n_i/p_i = \lambda$ $(i = 1, \cdots, c)$ より $n = \sum_{i-1}^{c} n_i = \lambda$ を得る。したがって確率 p_i $(i = 1, \cdots, c)$ の最尤推定量は $\hat{p}_i = n_i/n$ である。そこで対数尤度関数 $l_n = \sum_{i=1}^{c} n_i \log[p_i] - \sum_{i=1}^{n} \log n_i! + \log n!$ に代入すると、AIC(赤池情報量規準) は $-2l_n(\hat{p}_1, \cdots, \hat{p}_c) + 2$(パラメター数) であるから、定数項を無視すれば

$$\text{AIC} = -2[\sum_{i=1}^{c} n_i \log \frac{n_i}{n}] + 2(c-1) \tag{3}$$

で与えられる。この AIC を最小化する c を選ぶ方法は AIC 最小化法と呼ばれている。

なおこの方法ではヒストグラムについての様々な制約を組み入れることは容易である。例えば対称性は $c = 2m+1$ のとき $p_i = p_{2m+1-i}$ $(i = 1, \cdots, m)$ として組み入れられ、この制約条件の元で最尤推定 $\hat{p}_i^{RML}$ $(i = 1, \cdots, m)$, パラメター数 m, $c = 2m+1$ として AIC を定義すれば (c が偶数なら $c = 2m$) 元の AIC と比較可能である。

赤池情報量規準 (Akaike's Information Criteria) は一般的なモデル選択 (model selection) として一つの標準的な方法となっている[27]。なお、多項分布を利用する AIC 最小化法は比較的容易に 2 次元データやカテゴリカルデータの分析[28] にも拡張可能である。

4. 実験結果[29]

上記の式 (3) に基づいてヒストグラムの AIC を計算し、AIC を最小化する Python プログラムを開発したので、シミュレーション結果とともに提示する。なお、坂本・石黒・北川 (1982) では $n_i = 0$ となるヒストグラムのビン i について、ここの個数 n_i を $1/e$ で置換する方法を提案している。この方法は $n_i \log[n_i]$ が最小となる n_i が $1/e$ であることが根拠となっているが、$\hat{p}_i$ が最尤推定量ではなくなるので、理論的には妥当性に疑問がある。一方で応用上は、ビンの個数 c が過大になることして強いペナルティがかかるようになると予想される。

実装では、真偽値をとる引数 posit を用意して、posit=True の場合に上記の置換方法を適用するようになっている。なお (3) 式に $n_i = 0$ をそのまま代入すると対数が発散してエラーとなるので、右側極限が $n_i \log[n_i] \to 0$ となることから $n_i \log[n_i] = 0$ と定義することで対応している。

以下は北川の銀河データに対してプログラムを適用した 結果を示している。$c = 28, 14, 7$ とした場合、表 1 を見ると式 (3) 通りに計算した AIC(posit=False) の基準では、もっともビンが多い $c = 28$ が最適となる。一方、$1/e$ への置換を行った AIC(posit=True) の基準では、$c = 14$ が最適となる。実際にヒストグラムを図示した図 1 であり、どちらの AIC に基づく選択が良かったのかは明らかでない。

[27]例えば坂本・石黒・北川 (1982)、AIC の基礎理論や拡張については北川 (2020) などを参照されたい。統計数理研究所から Web 上で公開されている RS-DECOMP ではデフォルトとして「その他の手法」の中の「ヒストグラム」で自動計算される。RS-DECOMP は net 上からすぐに利用できる。

[28]統計数理研究所で開発された R 上で動くカテゴリー・データの統計解析プログラム CATDAP は https://jasp.ism.ac.jp/ism/catdap/ から利用可能である。

[29]このシミュレーションは、北川源四郎 (2019) 数理手法 VII 講義資料 4 ($https://elf-c.he.u-tokyo.ac.jp/courses/382/files/9300?module_item_id=5044$) の例と合致するように作成している。しかしながら、$c = 28$ のときの AIC(posit=False) を除き、この講義資料に掲載された AIC と本稿での計算結果は一致していない。このため、ビンの境界の位置などに違いがないか検討が必要と考えられる。

表 1 : ビンの個数 c と AIC

c	AIC(posit=False)	AIC(posit=True)
28	432.38	488.07
14	440.23	458.18
7	450.37	460.36

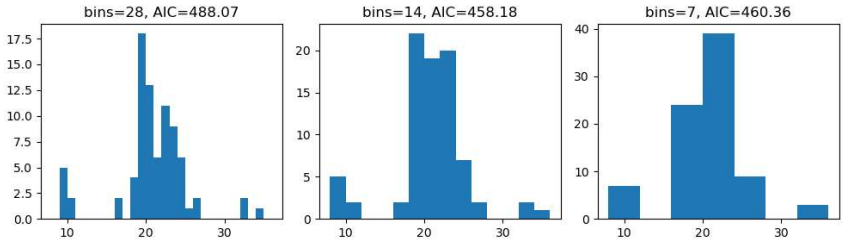

図 1:ヒストグラム
(注：ビンの個数を変えたときのヒストグラム, AIC は $n_i = 1/e$ と置換する方法を採用)

さらにビンの個数 c を細かく動かした結果を図 2 に示す。これを見ると、AIC の変動はジグザグとしており、AIC が最小となるビン数が必ずしも明瞭にわかるわけではないことが確認できる。また (3) 式をそのまま適用した AIC(posit=False)（図 2 左）では、ビン数が増大してもあまり AIC が増加せず、ビン数の選択が難しい。一方で AIC(posit=True)（図 2 右）では AIC の増加が顕著であるため、5 から 11 程度のビン数を選択することになると思われる。

5. ヒストグラムを巡る課題

「統計学基礎」を始め多くの統計学の教科書の第一章ではデータが与えられた時にまずデータのヒストグラムを調べることを説明していることが多い。それではどの様にしてヒストグラムを作るか、実は明確な答えは 2023 年でも統計家の間での合意は存在していない。仮に (ブラックボックス的に) エクセルを利用するなど出来合いの計算プログラムを使うとしても、実は出力を自前の PC がどの様に計算しているかを理解しているデータサイエンティスト、実務家は意外に多くはないのではないだろうか？

この問題をデータの背後に真の密度関数、あるいは多項分布が存在すると仮定して統計的推定としてとらえると、実は幾つかの基本的な統計的課題が存在する。例えば離散的な実データが利用可能だとして、ある区間にデータが存在しない場合にどの様に扱ったら良いか定かではない。(密度関数の推定では通常は密度がどこかでゼロとなる場合は仮定されない。) また区間幅を一様にとることが適切でない場合もあるだろうし、データが有界な場合に端の区間をどうとるか、実務的には色々な工夫が行われている。さらに二次元データではどうしたらよいか、確率の非負制約 ($p_i \geq 0$ $(i = 1, \cdots, c)$) を LASSO (least absolute shrinkage and selection operator) 的に入れたら結果は異なるだろうか？

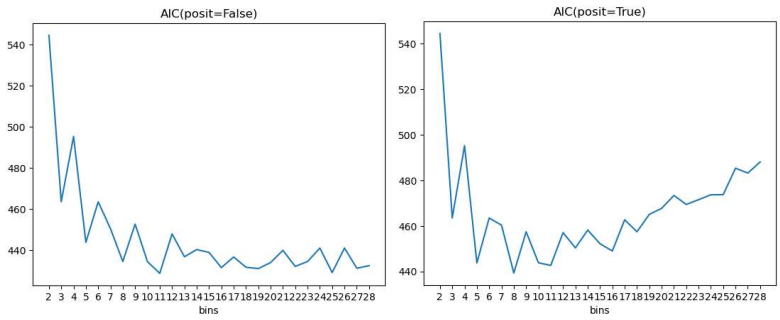

図 2: ヒストグラム
(注：ビンの個数を変えたときのヒストグラム, AIC は $n_i = 1/e$ と置換する方法を採用)

こうした問題には AIC 最小化法は容易に拡張可能と考えられるが今のところ既存の研究は見いだせない。

統計エキスパートにとり考慮すべき課題はなお少なくない。

文献

[1] 統計学基礎 (改訂版), 日本統計学会編, 2015, 東京図書.

[2] (応用をめざす) 数理統計学, 国友直人, 2015, 朝倉書店.

[3] 情報量統計学 , 坂本慶行・石黒真木夫・北川源四郎, 1982, 共立出版.

[4] (R による) 時系列モデリング入門 , 北川源四郎, 2020, 岩波書店.

[5] "*Optimal Cells for a Histgram*" (1990), Atsushi Kogure, The Fukushima University Research Series, Hassakusha Ltd.

```
###################################
#  A program written by H. Nishi
#  2023-12-25
###################################
# 下の方に変数positがあるので、その値をTrueにするかFalseにするかで設定を変更できる


# %%
import numpy as np  # 数値計算ライブラリ
from scipy import optimize, special  # 科学計算ライブラリ
import matplotlib.pyplot as plt  # グラフ描画ライブラリ

# %%

_BINS_NUMPY = ["auto", "fd", "doane", "scott", "stone", "rice", "sturges",
"sqrt"]


def aic_hist(hist, bin_edges, posit=False):
    # ヒストグラムに対してAICを計算する関数
    # posit=Trueとすると、n=0となるビンに1/e個のデータが入っていたと見なす（坂本・石黒・北川 1982)
    k = len(hist)
    n = np.sum(hist)

    if posit:
        hist = np.maximum(hist, 1 / np.e)

    # C = special.gammaln(n + 1) # 定数項を正確に計算するとこうなる
    C = 0
    len_bin = bin_edges[1:] - bin_edges[:-1]
    p = hist / n / len_bin

    llik = C + np.sum(hist[hist > 0] * np.log(p[hist > 0]))

    aic = -2 * llik + 2 * (k - 1)
    # print(k, n, bins, llik, len_bin)

    return aic


def search_hist_num(
    x, method="aic", posit=False, engine="brute", bins_max=None, hist_range=None
):
    # AICが最小となるビン数を選択する関数
    if method is None:
        method = "aic"

    if method.lower() in _BINS_NUMPY:
        # use numpy function
        bin_edges = np.histogram_bin_edges(x, bins=method, range=hist_range)
        bins = len(bin_edges) - 1
        res = None

    elif method.lower() == "aic":
        if bins_max is None:
            bins_max = 2 * int(np.floor(np.sqrt(len(x)))) - 1
```

```
        bins_max = min(len(x), bins_max)
        bins_min = 2

        # function to minimize
        def f(bins):
            hist, bin_edges = np.histogram(x, bins, range=hist_range)
            aic = aic_hist(hist, bin_edges, posit)
            return aic

        if engine.lower() == "brute":
            # brute force method
            res = optimize.brute(
                lambda b: f(b[0]),
                ranges=(slice(bins_min, bins_max + 1, 1),),
                finish=None,
                full_output=True,
            )
            bins = int(res[0])
            bin_edges = np.histogram_bin_edges(x, bins=bins, range=hist_range)

        else:
            raise ValueError(
                "Unknown optimization method. Currently only 'brute' is
available."
            )

    return bins, bin_edges, res


# Galaxy data
galaxy = np.array(
    [
        9.172,
        9.350,
        9.483,
        9.558,
        9.775,
        10.227,
        10.406,
        16.084,
        16.170,
        18.419,
        18.552,
        18.600,
        18.927,
        19.052,
        19.070,
        19.330,
        19.343,
        19.349,
        19.440,
        19.473,
        19.529,
        19.541,
        19.547,
        19.663,
        19.846,
```

```
        19.856,
        19.863,
        19.914,
        19.918,
        19.973,
        19.989,
        20.166,
        20.175,
        20.179,
        20.196,
        20.215,
        20.221,
        20.415,
        20.629,
        20.795,
        20.821,
        20.846,
        20.875,
        20.986,
        21.137,
        21.492,
        21.701,
        21.814,
        21.921,
        21.960,
        22.185,
        22.209,
        22.242,
        22.249,
        22.314,
        22.374,
        22.495,
        22.746,
        22.747,
        22.888,
        22.914,
        23.206,
        23.241,
        23.263,
        23.484,
        23.538,
        23.542,
        23.666,
        23.706,
        23.711,
        24.129,
        24.285,
        24.289,
        24.366,
        24.717,
        24.990,
        25.633,
        26.690,
        26.995,
        32.065,
        32.789,
        34.279,
]
```

```
)

# %%
hist_range = (8, 36)

# 【設定可能項目】ここのpositを変えることで、坂本他の置換方法を適用するかを選べる
posit = True
# print(np.histogram(galaxy, 28, range=hist_range))
print(f"{posit=}")

print(f"AIC for galaxy data (n={len(galaxy)})")
print("Bin Size| AIC")
print("-" * 30)
# ビン数が[28, 14, 7]の場合のAICを計算する
for bins in [28, 14, 7]:
    # print(np.histogram(galaxy, bins, range=hist_range)[0])
    print(
        f"   {bins}¥t|{aic_hist(*np.histogram(galaxy, bins, range=hist_range), posit=posit)}"
    )

# %%
# スタージェスの公式と、AIC最小の場合のビン数の比較
print("sturges = ", search_hist_num(galaxy, method="sturges", hist_range=hist_range)[0])
print(
    "MAIC bins=",
    search_hist_num(
        galaxy, method="aic", posit=posit, engine="brute", hist_range=hist_range
    )[0],
)
# print(
#     "MAIC bins=",
#     search_hist_num(galaxy, method="aic", engine="optuna", hist_range=hist_range)[0],
# )

# %%
# numpyで選べる自動的なビン数の選択法を比較
for m in _BINS_NUMPY:
    print(
        f"{m} = ",
        search_hist_num(galaxy, method=m, hist_range=hist_range)[0],
    )

# %%
# ビン数が[28, 14, 7]の場合のヒストグラムとAIC
plt.figure(figsize=(10, 3))
for i, b in enumerate([28, 14, 7]):
    plt.subplot(1, 3, i + 1)
    plt.hist(galaxy, bins=b, range=hist_range)
    plt.title(
        "bins={}, AIC={:.2f}".format(
            b, aic_hist(*np.histogram(galaxy, b, range=hist_range), posit=posit)
        )
    )
plt.tight_layout()
```

```
plt.show()

# %%
# ビン数を変化させたときのAICの変化をグラフに表示
bins = list(range(2, 29))
plt.plot(
    bins,
    [aic_hist(*np.histogram(galaxy, b, range=hist_range), posit=posit) for b in
bins],
)
plt.xticks(bins)
plt.xlabel("bins")
plt.title(f"AIC({posit=})")
plt.show()

# %%
# AIC最小となるビン数のときのヒストグラム
b = search_hist_num(
    galaxy, method="aic", posit=posit, engine="brute", hist_range=hist_range
)[0]
plt.hist(galaxy, bins=b, range=hist_range)
plt.title(
    f"bins={b}, AIC={aic_hist(*np.histogram(galaxy, b, range=hist_range),
posit=posit)}"
)
plt.show()
```

第 5 章 Fisher-Behrens-Welch 再訪 [30]

2023-8-18

国友直人・湯浅良太・西颯人

1. はじめに

統計検定 2 級の教科書「統計学基礎」3 章では多くの教科書と同様に t 検定の説明がある。2 標本問題についてまず 2 標本の分散が等しい場合の t 検定の方法を説明した後に 2 標本の分散が等しくない場合について説明、統計量

$$W = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}}$$

を定義, 自由度

$$f = \frac{(g_1 + g_2)^2}{\frac{g_1^2}{m-1} + \frac{g_2^2}{n-1}}$$

の t 分布を利用する検定方法を説明している。ここでは記号 $s_x^2 = [1/(m-1)]] \sum_{i=1}^m (X_i - \bar{X})^2$, $s_y^2 = [1/(n-1)] \sum_{i=1}^n (Y_i - \bar{Y})^2$, $g_1 = s_x^2/m$, $g_2 = s_y^2/n$ を用いた。さらに「実数値をとる自由度 f の t 分布を利用すると仮説の下で近似的に t 検定になることが知られている」、と説明している。

ここでの説明について、(i) 実際の 2 標本データでは各グループの分散が等しいという、教科書の最初の例はほどんどあり得ないこと、(ii) 自由度 f は実は s_x^2,s_y^2 という確率変数 (つまりランダム) に依存する、という二つの問題があるように感じられた。ここでの説明は古典的な Behrens-Fisher 問題についての Welch の方法に対応するが、これが何を意味しているのかもう少し正確に議論する必要があるだろう。

Welch の方法は確率変数 $Y = s_x^2/m + s_y^2/n$ の分布を自由度 ν の χ 二乗分布の期待値と分散に一致させて

$$\frac{2}{\nu} = \frac{\mathbf{Var}(Y)}{[\mathbf{E}(Y)]^2}$$

により

$$\frac{1}{\nu}[\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}]^2 = \frac{\sigma_x^4}{m^2(m-1)} + \frac{\sigma_y^4}{n^2(n-1)}$$

に推定値を代入する方法と解釈できる。(竹内 (1975))。

以上の議論から実際的な設定の下で「統計学基礎」で最初に説明している t 検定を利用すると実際には確率や棄却域がどれだけずれるのか、気になるところである。

2. 計算プログラムと数値例

統計量 $v = W$ の帰無仮説の下での確率分布は Kabe (1966) に与えられている。母数 $\lambda = \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}$ により変数を基準化して確率変数 $U = [\bar{X} - \bar{Y} - (\mu_x - \mu_y)]/\sqrt{\lambda}$ と $Z = [s_x^2/m + s_y^2/n]/\lambda$ の同時分布から、$v = [U/\sqrt{Z}]$ の精密分布は超幾何級数 (hypergeometric series)

$$F(\alpha, \beta; \gamma; x) = \sum_{r=0}^{\infty} \frac{\Gamma(\alpha + r)}{\Gamma(\alpha)} \frac{\Gamma(\beta + r)}{\Gamma(\beta)} \frac{\Gamma(\gamma)}{\Gamma(\gamma + r)} \frac{x^r}{r!}$$

[30] 統計エキスパート slack 上で行われた統計的推測の基礎についての素朴な疑問や議論の中から統計エキスパート養成事業にとり有用と思われる内容をまとめようと作成中のメモである。

により次のように表現できる。

$$f(v) = c(\alpha_1^2 + v^2)^{-(p_1+p_2+\frac{1}{2})} F(p_2, p_1+p_2+\frac{1}{2}; p_1+p_2; \frac{\alpha_1-\alpha_2}{\alpha_1+v^2}) \ ,$$

$$c = \alpha_1^{p_1}\alpha_2^{p_2}\Gamma(p_1+p_2+\frac{1}{2})[\sqrt{\pi}\Gamma(p_1+p_2)]^{-1} \ ,$$

ただし、ここで記号 $N_1 = m, N_2 = n, \sigma_1 = \sigma_x, \sigma_2 = \sigma_y$ に対して $N_1 > N_2$ のとき $\alpha_i = \frac{N_i(N_i-1)}{\sigma_i^2}[\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}], p_i = (N_i-1)/2, \ (i = 1, 2)$ と置いた。特に 2 標本の標本数 $N_1 = N_2$、分散 $\sigma_1^2 = \sigma_2^2$ であれば $\alpha_1 = \alpha_2$ となるので普通の t 分布が導かれる。
超幾何関数による上の表現は一見すると複雑そうに見えるが、t 分布 ($\alpha_1 - \alpha_2 = 0$ の場合, 自由度 $N_1 + N_2 - 2$) の周りでの項 $[1 - \alpha_2/\alpha_1]/[1 + v^2/\alpha_1]$ による展開と見ると、その解釈は容易だろう。例えば母数

$$\eta = [\frac{N_1(N_1-1)}{\sigma_1^2} - \frac{N_2(N_2-1)}{\sigma_2^2}]\lambda$$

が大きくなるにつれて Kabe 分布の t 分布からの乖離度が大きくなることが分かる。

推定量の性質を調べるために超幾何分布による厳密分布（Kabe）と t 検定による結果について比較を行う為に超幾何関数の R ライブラリー library("hypergeo") を利用した計算プログラムを作成した。(なお作成した R-プログラムの Python 翻訳版も作成した[31]。)
分散が同一とは限らない場合に Welch 統計量の厳密分布と t 分布の密度関数の比較、分位点と棄却域の差を求めることができる。以下では $m = 10, n = 3, \sigma_x = 5, \sigma_y = 10$ と $m = 100, n = 30, \sigma_x = 5, \sigma_y = 10$ という母数が所与の場合の密度関数の図を例示しておくが、出力表示を含めプログラムをほんの少し変更すれば自由に様々な状況について検定統計量の分布をしらべることができる。

なお、ここで示している例 1 では 2 標本で標準偏差は 2 倍、データ数は 10,3 とそれぞれ小さいので厳密分布と t 分布の差は小さくない。例 2 はデータ数 100,30 の 2 標本なので密度関数はかなり t 分布に近い。いずれの場合も等分散を仮定して t 検定を行うとサイズがかなりずれることに注意する必要がある。

さらに実際のデータ分析では分散母数は未知であるから有限標本では Welch 統計量, t 統計量の帰無分布も未知母数に依存する。未知母数をデータから推定すると推定誤差の影響により有意水準や棄却域の領域がかなり変化する可能性があることにも注意が必要である。特に分散の均一性を仮定した 2 標本平均の検定は有意水準や棄却域の領域を大きく間違える可能性があると云えよう。

最後になるが、データ数がそれほど多くない場合には未知分散に依存しないような正確検定は 2023 年時点においても知られていないようである。データ数が多い例 2 の場合には標本分散は母分散に確率収束し、標本平均に中心極限定理 (CLT) を利用、と云う漸近理論を適用すると、統計量の分布は標準正規分布 N(0,1) によりかなり良く近似できる。このことを利用すると例 2 の状況では検定についての十分な精度が保証できるようである。二つの標本数がどの程度まで大きいと、N(0,1) により精度を保証できるかは興味深い課題であろう。

[31] 最近では Python は学系分野における一つの標準的な言語になっている。統計エキスパート・プロジェクトでは R と Python の知識はデータ分析の応用上で必須と見なしている。

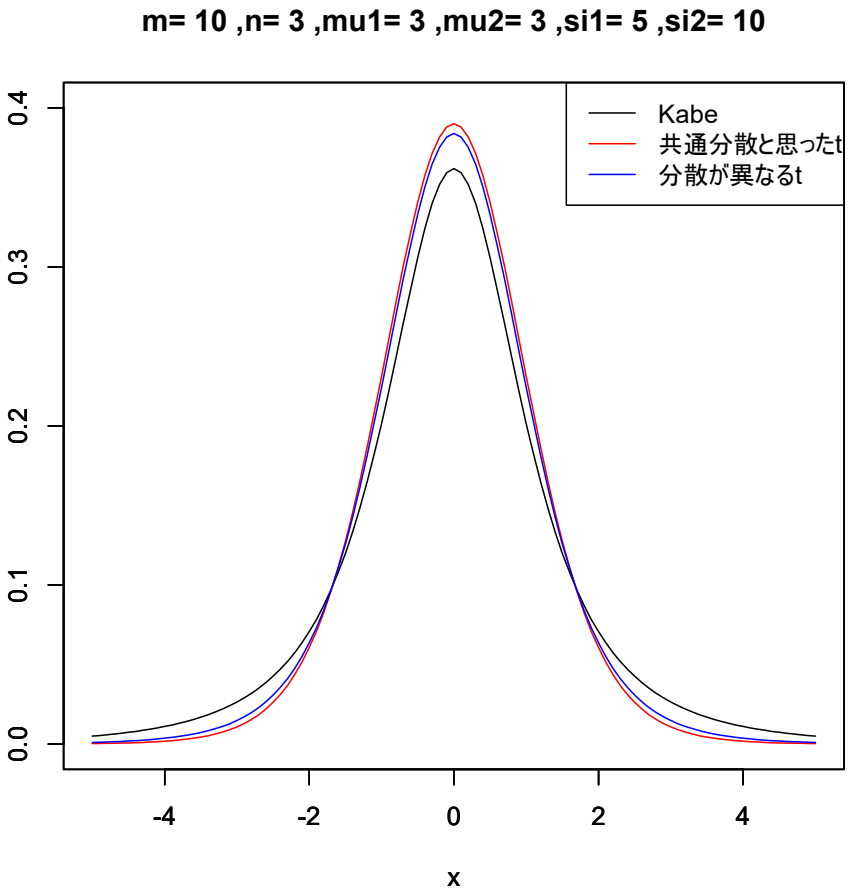

図 1:Welch 統計量と t 近似

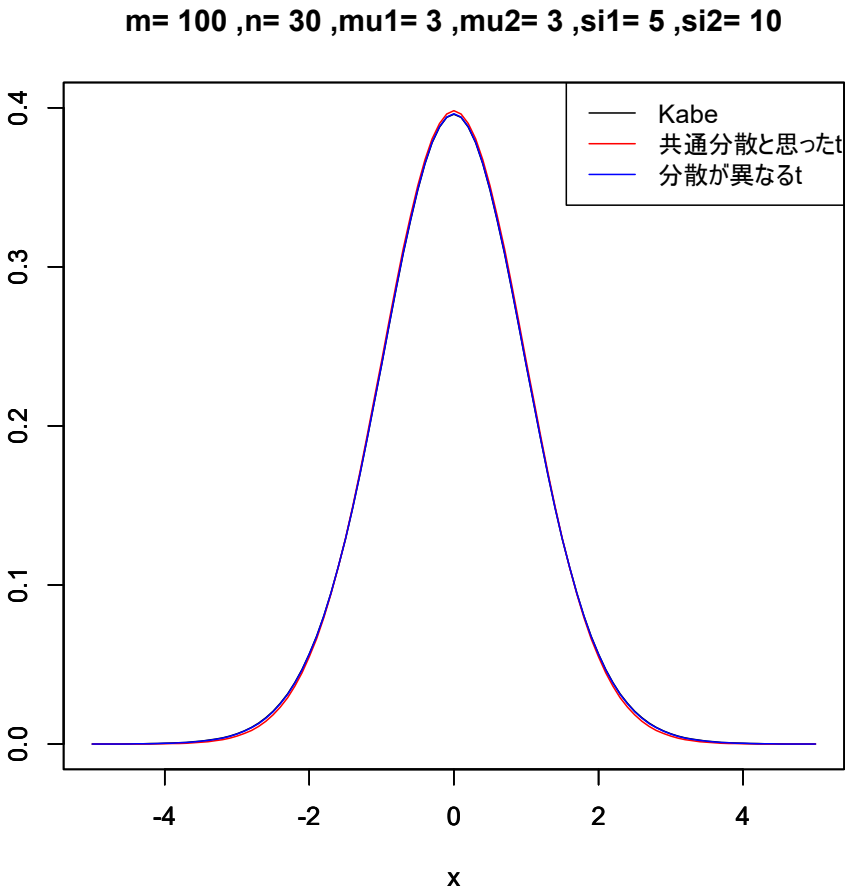

図 2 : Welch 統計量と t 近似

3. 文献

[1] 統計学基礎 (改訂版), 日本統計学会編, 2015, 東京図書.

[2] 現代数理統計学 , 竹村彰通, 2020, 学術図書出版.

[3] 確率分布と統計解析, 竹内啓, 1975, 日本規格協会.

[4 On the exact distribution of the Fisher-Behrens-Welch statistic," Kabe, D.G., 1966, Metrika, 10-1, 13-15..

```
### Fisher-Behren's-Welch
### 2023-8-18 prepared by R. Yuasa
### 未知パラメータを既知とした場合に計算できるKabeによるものとの比較

library("hypergeo")

## パラメータの設定
N1 <- 10 # m
N2 <- 3 # n
mu1 <- 3
mu2 <- 3
si1 <- 5
si2 <- 10

## 分散が異なる場合のtのためのデータ
X1 <- rnorm(N1, mu1, sd=si1)
X2 <- rnorm(N2, mu2, sd=si2)

s1 <- sqrt( sum(( X1 - mean(X1) )^2) / (N1-1) )
s2 <- sqrt( sum(( X2 - mean(X2) )^2) / (N2-1) )

## Kabeによる密度関数, N1>N2
alp1 <- ( N1*(N1-1)/si1^2 ) * ( si1^2/N1 + si2^2/N2 )
alp2 <- ( N2*(N2-1)/si2^2 ) * ( si1^2/N1 + si2^2/N2 )
p1   <- (N1-1)/2
p2   <- (N2-1)/2
C <- (alp1^p1)*(alp2^p2)*gamma(p1+p2+1/2)*(pi^(1/2)*gamma(p1+p2))^(-1)
pdfv <- function(v){
  C*(alp1+v^2)^(-(p1+p2+1/2))*Re(hypergeo(p2, p1+p2+1/2, p1+p2,
(alp1-alp2)/(alp1+v^2)))
}

## 密度関数の比較
## 分散が等しいと仮定するかどうかで統計量が変わるのでtどうしで違うのは妥当ではある
## Kabeと分散が異なるtの違いはそのまま検定に影響を与えうる
plot(pdfv, -5,5, xlim=c(-5,5), ylim=c(0,0.4), ylab="", main = paste("m=", N1,
",n=", N2, ",mu1=", mu1, ",mu2=", mu2, ",si1=", si1, ",si2=", si2))
par(new=T)
curve(dt(x,df=N1+N2-2), -5,5,col="red", xlim=c(-5,5), ylim=c(0,0.4), ylab="")
par(new=T)
curve(dt(x,df= (s1^2/N1+s2^2/N2)^2/( (s1^2/N1)^2/(N1-1) + (s2^2/N2)^2/(N2-1) )),
-5,5,col="blue", xlim=c(-5,5), ylim=c(0,0.4), ylab="")
legend("topright", legend=c("Kabe", "共通分散と思ったt", "分散が異なるt"),
lty=1, col=c("black","red","blue"))

## 棄却確率を比較するためにKabeによるものから棄却域を求める
lev <- 0.05
optfunc_v <- function(cc){
  (integrate(pdfv, -cc, cc)$value - (1-lev))^2
}
cc <- optimize(optfunc_v, interval=c(0.2,5))$minimum

## 受容確率の比較を行う
R <- 10^5
accept1 <- 1-numeric(R)
accept2 <- 1-numeric(R)
accept3 <- 1-numeric(R)
```

```
# %%
### Fisher-Behren's-Welch
### 2023-8-18 prepared by H. Nishi
### 未知パラメータを既知とした場合に計算できるKabeによるものとの比較

import numpy as np
import matplotlib.pyplot as plt

# グラフ描画パッケージ（密度推定など）
import seaborn as sns

# 特殊関数の読み込み
#  （超幾何関数を計算する関数を含む）
# library("hypergeo")
from scipy import special

# 統計関数の読み込み
from scipy import stats

# 積分・最適化のための関数
from scipy import integrate, optimize


rng_seed = 123

## パラメータの設定
# N1 <- 100 # m
# N2 <- 30 # n
# mu1 <- 3
# mu2 <- 3
# si1 <- 5
# si2 <- 10
N1 = 10  # m
N2 = 3  # n
mu1 = 3
mu2 = 3
si1 = 5
si2 = 10


## 分散が異なる場合のtのためのデータ
# X1 <- rnorm(N1, mu1, sd=si1)
# X2 <- rnorm(N2, mu2, sd=si2)

# s1 <- sum(( X1 - mean(X1) )^2) / (N1-1)
# s2 <- sum(( X2 - mean(X2) )^2) / (N2-1)
rng = np.random.default_rng(rng_seed)
X1 = rng.normal(loc=mu1, scale=si1, size=N1)
X2 = rng.normal(loc=mu2, scale=si2, size=N2)

s1 = np.sqrt(np.sum((X1 - np.mean(X1)) ** 2) / (N1 - 1))
s2 = np.sqrt(np.sum((X2 - np.mean(X2)) ** 2) / (N2 - 1))


## Kabeによる密度関数, N1>N2
# alp1 <- ( N1*(N1-1)/si1^2 ) * ( si1^2/N1 + si2^2/N2 )
# alp2 <- ( N2*(N2-1)/si2^2 ) * ( si1^2/N1 + si2^2/N2 )
# p1   <- (N1-1)/2
```

第II部: 応用統計からの話題

第6章　論説：株価を統計的に予測する？(資産価格の基本定理を巡って) [32]

2023年12月

国友直人・中西正

第7章 論文紹介「ビッグデータの統計的パラドックスについて」

2023年12月

湯浅良太

("Statistical paradise and paradoxes in big data (I) Law of large populations, big data paradox and 2016 US presidential election" by X. Meng, *Annals of Applied Statistics*, 2018, Vol.12-2, 685-726 [33]

第8章 報告" 階層ベイズロジットモデルと異質な消費行動"　[34]

2023月12月

趙宇

[32]統計エキスパート slack 上に投稿されたある研修生の発見に触発されて書かれた原稿である。ただし「統計学」(久保川達也・国友直人,2016, 東京大学出版会) の第 14 章の一部、「数理ファイナンスの基礎」(国友直人・高橋明彦, 2003, 東洋経済新報社) の第 1 章の一部を元に本稿を作成した。

[33]2023年4月〜5月に実施された統計エキスパート養成事業における連続講義シリーズ「社会における統計科学」におけるレポートの改訂稿である。

[34]2023年4月〜5月に実施された統計エキスパート養成事業における講義「計算ベイズ」における期末レポートの改訂稿である。

論説：株価を統計的に予測する？(資産価格の基本定理を巡って) [1]

国友直人・中西正

2023-11-30
2024-1-17(一部改訂)

1. はじめに

理系学部で統計学を学び、経済・金融データを観察すると、「統計学を応用すると将来の株価を統計的に予測できるはず？」と考える大卒者は（かなり昔から）研究者を含めかなり存在している。最近ではこうしたデータサイエンティストに加えて、「機械学習 (machine learning) を応用すると将来の株価を予測できるはず？」と考える研究者もちらほら、とはある部外者の想像に過ぎない。統計エキスパート養成事業では 2022 年 4 月にある証券会社のネット広告を見たある研修生による感想を巡ってメンターを含め自由討論で興味深い幾つかの議論が行われた。現代の日本では様々な場面でこうした記事 (あるいは耳寄りな話儲け話など) に遭遇するが、理系・文系の様々な分野での「統計エキスパート」を目指す諸氏はこうした内容について「統計学の視点」から的確に判断できることが望ましいと言えよう。

本稿ではこの例を題材に内外の株価、国債、外国為替レート、金利など、経済・金融市場と統計分析の基礎を議論することとしよう。単純明快な解答を得ることは金融ビジネスを生業としている関係者が介在する実社会の中、実は容易ではないのだが、簡単な例を通じて「統計エキスパート」による今後の金融データ・経済データの分析のあり方を検討することは有意義、問題を理解することは重要と考えられる。本稿は「今まで経済・経営・金融などにおける統計的問題を学んだことがない」理系出身の若手研究者を主な対象として、幾つかの基本的視点を指摘する小論である。

2. 日米の株価を巡って

ある証券会社が 2022 年 4 月にネット上で発行した日米株価に関するパンフレットの内容を統計学的観点より実例としてをここで取りあげてみよう。(念のために本稿の最後に付録として掲載しておく。) 普通の一般人がたまたま銀行や証券会社の窓口に出かけると (優良な顧客と見なされれば) アドバイザーが対応、日米の株価の動向を「内外の投資環境」として説明、最後には親切にアドバイスすることは普通に行われている業務である。その際、金融データの動向を図・グラフ、その他の統計的用語、例えば相関係数などが有用と判断すれば、ごく普通に

[1]「統計学」(久保川達也・国友直人,2016, 東京大学出版会) の第 14 章の一部、「数理ファイナンスの基礎」(国友直人・高橋明彦, 2003, 東洋経済新報社) の第 1 章の一部を元に本稿を作成した。例えば「統計学」同章前半部の random walk をはじめとする統計的時系列モデルの概要の説明は統計エキスパートの講義「統計的時系列解析」で議論したので省略した。

利用している。こうしたアドバイザー業務を行う従業員はフィナンシャル・プランナーという資格を持っているのが通常である。むろん証券会社は様々な金融資産の取引を仲介することにより得られる手数料収入を生業としていることを考慮していても、このパンフレットに説明しているような 2012 年〜2022 年の間の日米の株価に強い相関が見られるという指摘は興味深い。そこでもう少し期間を長くとり、この間の日本の日経 225, 米国ダウ, 日米為替レートを調べてみよう。

便宜上で 1990 年代 (1990 年 1 月から)、2000 年代、2010 年代、2020 年代 (2023 年 11 月まで) と時代区分をとり 2010 年代の日米の代表的な株価インデックスとして Nikkei225, ダウ (円換算) を取りあげ、x 軸に Nikkei225,y 軸にダウ (円換算) の二次元プロット図を回帰直線 (赤) と共に示したのが図 1 である[2]。確かに「統計学基礎・第 5 章」に出てくる例と同様な線形関係を示しているいるように判断する研修生がいてもおかしくはなさそうである。ここでダウ (円換算) という意味

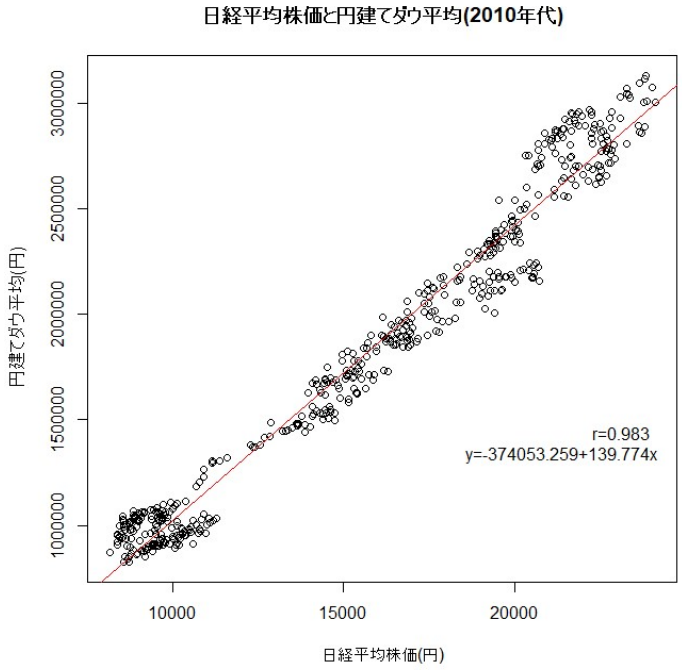

図 1: 日米株価 (2010 年代)

であるが、ニューヨークの取引所ではドルで取引が行われるので日本の投資家か

[2]データの出所：investing.com, データ・タイプ：日経 225,NY ダウは株価インデックス, 週次データ, 各市場の終値, 円建て NYD は，ドル円× NYD(USD) とし，終値のデータから作成した。ただし、一か所データが欠損；Nikkei225 の 2019/4/28 分のデータが連休のため存在しないので前週終値をそのまま 2019/4/28 として採用した。実際には日米の株式市場には約 1 日の時差がある。

らみるとドル・円為替レートで円換算して判断する必要がある、という意味である。したがって元々の時系列は Nikkei225(円)x, ダウ (ドル)y, ドル円レート z という 3 つの時系列から x と y/z を取り出して円で換算した日米株価を 2 次元プロットしたのが図 1 なのである。ところで先ほどの図は 2010 年代のデータから作成したが、ここでもう少し長い時間におけるデータとして 1990 年-2023 年の間に実際に取引されている円とドルで測った株価インデックスという 2 系列 (単位が異なることに注意して) を時系列プロットしたのが図 2 である。

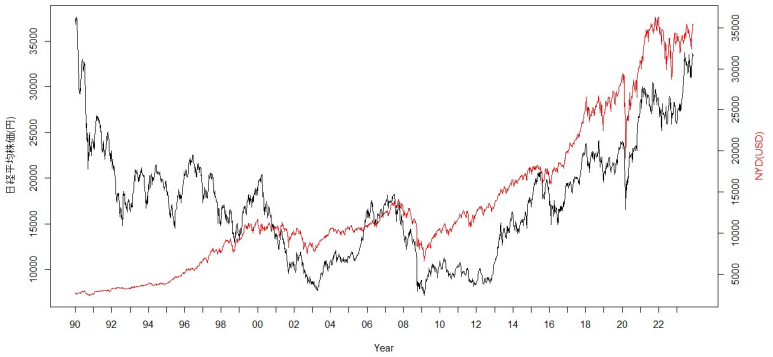

図 2: 日米株価 (1990-2023)

図 2 で示されている約 30 年間における日米の株価と為替レートの水準を眺めると様々なマクロ経済変動が思い浮かぶが、その個々の事象の説明はエコノミストやマクロ経済学者に任せるとして、ここでは統計学的観点から見ると興味深い特徴があることを指摘しておこう。図から直ぐに株価の時系列は時間の経過とともにギザギザした上下動をしつつ、時期によりある方向 (ドリフトと呼ばれる) に変化する動きを見せている。これは時間の経過とともに（高校物理の図などで目にする）滑らかに変動する系列とは全く異なるタイプの変動と云える。ここでは省略するが、データの観測間隔を 1 週間から 1 日、1 時間、さらに 1 分などと短くしても株価の場合にはこうした特徴は維持されている。

次に図 1 に示される二つの価格の関係は興味深いので、1990 年-2023 年の間を 10 年刻み（2020 年代はほぼ 4 年間）で図 1 と同じようなプロットをして図 3・図 4 を作成した。すると 2010 年代に見られる (正係数で) ほぼ直線的な関係は実はその期間のみに観察された日米間の 3 つの時系列の間における二つの価格データ間の関係であることが分かる。(なお図中 r はデータから計算した相関係数、最小二乗法により推定された直線も記入しておいた。) この 10 年間の日本・米国における株式市場、円ドル為替市場においては「事後的に見ると円換算の日米株価」はかなり安定的な関係を維持しつつ推移していたことが確認できる。しかし同時

に他の時期ではかなり異なる様相を呈していることも分かるだろう。

株価の変動をめぐり観察されるこうした現象は他の時期、あるいは他の先進諸国の資本市場において大部分の金融市場で観測される価格データ、株価、為替レート、国際価格、金利などでかなり一般的に観測されている。ここでとりあげた実

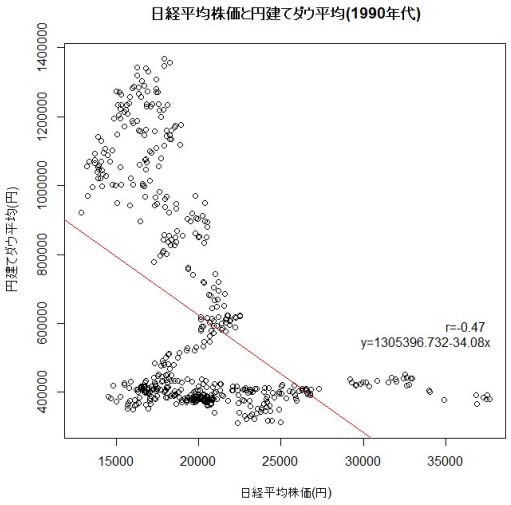

図 3: 日米株価 (1990 年代)

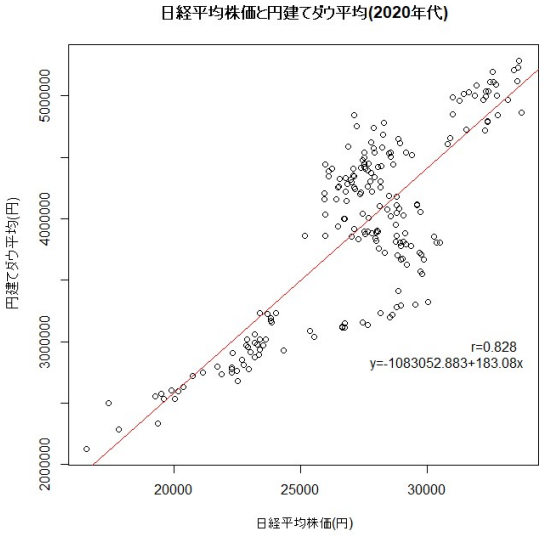

図 4: 日米株価 (2020 年代)

例における図 1〜図 4 を眺めてどの様に理解・判断するか、あるいは銀行や証券会社の窓口で説明を受けた顧客がどう判断するか、むろん各自の自己責任である[3]。しかし少なくとも「統計エキスパート」なら図 1〜図 4 の意味を理解、相関係数や回帰直線の妥当性などを非専門家に説明できることが望ましいと言えるだろう。念のため第 3 変数の円ドル為替レートの動向を図 5 に図示しておく。この時代の日米を取り巻く経済情勢を反映してかなりの変動が観察されている。

[3]実際に顧客が窓口で金融商品を発注するときには、一昔はともかく現代の日本では将来発生するであろうリスク (risk) についての説明が義務づけられている。

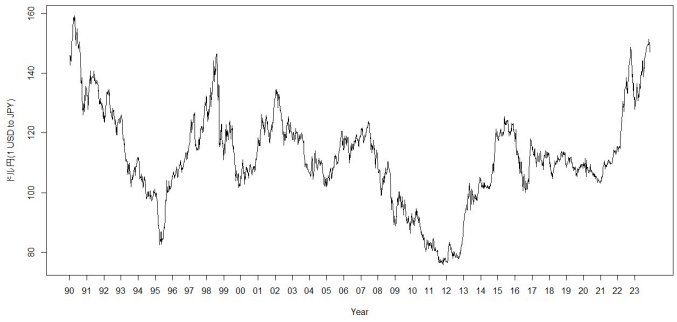

図 5: 円ドル為替レート (1990-2023)

3. 市場と確率

経済時系列の中でも金融時系列の見方について近年では統計学的な観点からの理解も深まっている。まずはかなり遠回りに見えるが、単純化した例により金融市場で観察される金融時系列の変動を理解することに資する幾つかの重要な論点を説明しよう。

有限数の参加者により二つの事象のどちらかに賭けることが繰り返し行われ、(参加者の損得の集計値がゼロとなる) ゼロ・サム・ゲームにより例示される市場取引の意味の考察から始めよう。賭けゲームの市場において毎回の賭けゲームにおける (事前には結果がわからない) 不確実性な事象を仮に H(表) と T(裏) とする。各参加者は初期資産 1 より毎回賭けゲームに参加し、単位あたり利得は H であれば 1、T であれば-1 というルールにより参加者の掛け金を毎回分配するというゲームを考え、ゲームを運営する胴元のコストはゼロ、各参加者はなるべく自己の利得を最大化するような戦略をとると仮定する。第 i 回の利得 (1 か-1) を x_i $(i = 1, ..., n)$ として、毎回資産の小さな部分 α だけ事象 H に投資する個人 K 君 (α はあとで選ぶ) は資産がゼロとなると賭けゲームより撤退するか、あるいは他の参加者 (含む胴元) から資金を借りてゲームに参加し続けられる、とする。

仮に長い間ではこの H の回数が T よりほんの少し有利で n が十分に大きいとき、「[条件 I] : (H の回数)/n $\to 1/2 + \epsilon$ $(\epsilon > 0)$」であったとする。このときゲームに参加している時点 n における K 君の資産額を Y_n とすると

$$Y_n = 1 \cdot (1 + \alpha x_1) \cdot (1 + \alpha x_2) \cdots (1 + \alpha x_n) \tag{1}$$

である。資産の対数をとり $n \to \infty$ のときの資産価値 (例えば $0 < \alpha < 1/2$ として Y_n を評価すると n が大きい時には大数の強法則 (strong law of large numbers)[4]より $(1/n) \sum_{i=1}^{n} x_i \longrightarrow (1/2 + \epsilon) - (1/2 - \epsilon) = 2\epsilon$ となる確率は 1 であるから、

[4]大数の強法則については確率論の本格的な教科書、例えば Billingsley (1995),*Probability and*

確率 1 で

$$\begin{aligned}\log Y_n &= \sum_{i=1}^n \log[1+\alpha x_i] \\ &> \sum_{i=1}^n [\alpha x_i - \alpha^2 x_i^2] \\ &= n\alpha[\tfrac{1}{n}\sum_{i=1}^n x_i - \alpha] \longrightarrow n\alpha[2\epsilon - \alpha]\end{aligned}$$

となる。ここで α は任意の実数とできれば (仮に幾ら小さくとも正実数であれば)、ϵ の値が分かれば $2\epsilon - \alpha > 0$ とすると、確率 1 で右辺は n が大きくなるにつれて幾らでも大きくできる。また仮に $\epsilon < 0$ であれば T に投資する戦略を考えればよい。もしこの市場ゲームの参加者の数が有限なら一方で資産がゼロとなり市場ゲームから撤退する参加者が発生するが、他方で利益をあげることができる K 君もいる。時間の経過とともに市場ゲームが存立できなくなる。したがって [条件 I] が成立する場合には「[条件 II] : 市場ゲームはそのうち必ず破綻する」ことになる。したがってこの命題の対偶をとると次の結果が導かれる。

定理 1 : 「この市場ゲームが必ず破綻するとは限らない」ならば

$$\lim_{n\to\infty} \frac{(\text{H の回数})}{n} = \frac{1}{2} \tag{2}$$

である。

ここではたとえ各自がそれぞれ個人的確率から出発してゲームに参加していたとしても、このゲームが市場として存立する為には (経済学の用語を用いると一種の均衡と解釈) 確率と見なせる数値が得られること興味深い。ここでの議論は例えば公平な (fair) サイコロの目についても成立するなど状態数が有限個であればかなり一般的な設定で成立する。

さてここで得られる市場で成立する公平な価格 (fair-price) としての確率を市場確率 Q と呼べば、$n-1$ 時点で Y_{n-1} の資産を持つ K 君の n-1 期と n 期の資産の間には $Y_n = Y_{n-1}[1+\alpha x_n]$ が成り立ち、$v_n = Y_{n-1}\alpha x_n$ と置けば

$$Y_n = Y_{n-1} + v_n \ \ (n > 0) \tag{3}$$

および $v_n = \alpha Y_{n-1}$ (確率 1/2), $v_n = -\alpha Y_{n-1}$ (確率 1/2) である。このランダム・ウオーク (酔歩) モデルはもともと賭ゲームから生じたと言われている。Y_n の辿る過程は過去の情報 $Y_{n-1}, Y_{n-2}, \cdots$ が与えられたとき Y_n の確率測度 Q についての条件付期待値について

$$\mathbf{E}^Q[Y_n | Y_{n-1}, \cdots, Y_0] = Y_{n-1} \tag{4}$$

が成立する。この条件を満たす確率過程は martingale(マルティンゲール) とも呼ばれている[5]。

Measure, 3rd edition, Wiley などがある。確率 1 での収束 (概収束, almost-sure convergence) は大数の弱収束における確率収束 (convergence in probability) より強い概念となる。

[5] Y_n が (4) を満たせば $v_t = Y_t - Y_{t-1}$ は無相関のランダムな系列となる。すなわち $\mathbf{E}^Q[v_t v_s] = 0$ $(s \neq t)$ となる。

ここで0時点における資産額 $K_0 = 1$ をゲームに参加する権利価格と解釈すると、n 時点における K_n は市場で評価される価格、(4) はマルチンゲールの価格モデルと解釈できるだろう。ランダム・オークでは時点 $n-1$ においては、それまでの情報を幾ら利用しても時点 n における Y_n の値が増加する事象も減少する事象も同等に確からしい。したがって、時点 $n-1$ における時点 n の値 Y_n の予測値は Y_{n-1} そのものが妥当となる。この式で表現される系列をある初期値 Y_0 から人工的に発生させると、Y_n の実現する系列はギザギザした経路をとり、過去の情報から将来の経路の予測はかなり困難である。このことを現在の価格が将来の価格の情報を反映している、市場が効率的 (efficient) である、などと表現されるが、ランダム・ウオーク・モデルはそうした状況を表現するのに役立つことになっている。

「No Free Lunch」条件

時刻 $0, 1, \cdots, l$ $(2 < l < \infty)$ という離散時間にリスクを伴う2つの証券 S_1, S_2 が市場で取引されている経済を考えよう。ここで後の議論を簡単化する為に、証券 S_i $(i = 1, 2)$ を保有することよって生じうる配当は最終期 $t = l$ のみに発生することとして、その途中では様々な理由から価格が変動している状況を想定する。

ここで数値例として仮に時刻 $t = 0$ においては証券 S_1, S_2 の価格がともに 10 であり、時刻 $t = 1$ においては2つの不確実な自然の状態 ω_1, ω_2 に対して証券 S_1 と証券 S_2 の価格が次の表で与えられていると想定してみよう。時刻 $t = 0$ にお

表 1

	証券 S_1	証券 S_2
(自然の状態)ω_1	6	12
(自然の状態)ω_2	6	18

いて初期資産をゼロとして、証券 S_1 と証券 S_2 の可能な保有量を $y_1,\ y_2$ としておく。ただし、ここで負値の保有（"short sales"の可能性）も許すことにするので y_i $(i = 1, 2)$ は任意の実数とする。この設定のもとで仮に二つの証券をそれぞれ (y_1, y_2) だけ持っていたとすると、時刻 $t = 1$ において自然の状態 ω_1 が実現すれば

$$6y_1 + 12y_2 \geq 0$$

のときに時刻 $t = 1$ での利得が非負となる。同じ設定の下でもし状態 ω_2 が実現してしまえば時刻 $t = 1$ では

$$6y_1 + 18y_2 \geq 0$$

のときに利得が非負となる。したがってこれらの領域を表せば時刻 $t = 1$ において状態 ω_1 と状態 ω_2 のいずれが実現しても非負の収益を得ることが出来るためには (y_1, y_2) の領域として原点に対して凸な領域となる必要があることがわかる。こ

こで (y_1, y_2) はリスクを伴う 2 つの証券を時刻 $t = 0$ で購入（あるいは売却）する量をそれぞれ表して直線

$$10y_1 + 10y_2 = 0$$

を書いてみよう。このとき領域 $10y_1 + 10y_2 < 0$ は時刻 $t = 0$ において証券保有の価値が負となる領域、領域 $10y_1 + 10y_2 > 0$ は時刻 $t = 0$ において証券保有の価値が正となる領域をそれぞれ表していることに注意する。そして、例えば直線上の点（-10, 10）を時刻 0 のポートフォリオとして選べば、このような市場価格が時刻 $t = 0$ で成立している経済では、明らかに”free lunch”（タダの昼食）[6]を享受できることが確認できよう。ここで”free lunch”とは「裁定機会の利用」の意味である。

次に表 1 の状態 ω_2 における価格を変化させて 2 つのリスクを伴う証券の価格が次の表で与えられる経済を考えよう．この経済では時刻 $t = 1$ において仮に状

表 2

	証券 S_1	証券 S_2
(自然の状態)ω_1	6	12
(自然の状態)ω_2	18	6

態 ω_2 が実現すれば非負の収益の得られる範囲は

$$18y_1 + 6y_2 \geq 0$$

によって与えられる。したがって、この経済においては状態 ω_1 と状態 ω_2 のいずれが実現しても非負の収益を得ることが出来る (y_1, y_2) の領域を表現すると、今度はネット・ゼロ・ポジションを表す直線 $10y_1 + 10y_2 = 0$ とは原点を除いて交わることはないことがわかる。すなわち，今度は時刻 $t = 0$ で成立している市場価格を所与とすると、この経済では ”free lunch ”を享受することはできなくなっている。

ここで表 2 で表現されている ”free lunch ”が存在しない経済について次の事実に注目しよう。自然の状態 ω_1 と ω_2 に対する 2 つの非負の収益線に関する法線ベクトルはそれぞれ $(6, 12)', (18, 6)'$ で与えられている。ここで時刻 $t = 0$ の価格ベクトルは $(10, 10)'$ であるのでこの 3 つのベクトルの間には

$$\begin{pmatrix} 10 \\ 10 \end{pmatrix} = q_1 \begin{pmatrix} 6 \\ 12 \end{pmatrix} + q_2 \begin{pmatrix} 18 \\ 6 \end{pmatrix} \tag{5}$$

[6]著名な経済学者のミルトン・フリードマンによる「世の中にただの昼飯 free lunch などは存在しない」という独特の説明から採った言葉である。むろん、大学や職場などでも上司や同僚にお昼をおごってもらう (free lunch) こともあるが、そうした時には結局は代金よりも高くつくことを頼まれることも少なくない。経済学では「free lunch」が存在すると、「合理的な経済人」から成り立つ市場経済ではミクロ的均衡は成立しないと考えるのが常識的である。むろん、様々な理由から現実の市場で裁定機会が存在しているように見えることもあるようである。もし、本当に裁定機会が存在したならば価格調整を阻む何らかの制度的・経済的要因がなければ、そうした機会は急速に解消されるはずであろう。

を満足する非負ベクトル $q' = (q_1, q_2)'$ が存在していることが見てとれよう。

ここで状態 ω_1 が起きれば 1, 起きなければ 0 となる証券を仮想的に考えれば、その価格が q_1 であると解釈することができる。同様に状態 ω_2 が起きれば 1, 起きなければ 0 となる証券を考えれば価格が q_2 と解釈されよう。こうした状態に依存する仮想的証券を Arrow=Debreu 証券、状態の価格は状態価格（state price）と呼ばれている。(5) を解けば幸いにも（数値例として作ったゼロ金利状態に対応して）基準化則 $q_1 + q_2 = 1$ も同時に満足して $(q_1, q_2) = (2/3, 1/3)$ となる。そこで q_1 と q_2 をそれぞれ時刻 $t = 1$ における自然の状態 ω_1, ω_2 に対する評価確率と考えれば、時刻 $t = 1$ における価値にリスクを伴っている 2 つの証券の (時刻 $t = 0$ における) 期待金額はそれぞれ（Z_1の期待価格）$= 6q_1 + 12q_2 = 10$,（Z_2の期待価格）$= 18q_1 + 6q_2 = 10$ となっている。すなわち時刻 $t = 1$ における S_1 の期待価格と S_2 の期待価格は時刻 $t = 0$ において市場で成立している 2 つの証券の価格水準 10 に等しいことが分かる。すなわち、これら 2 つの証券価格は均衡（理論）価格であると考えることができる。

ここでの数値例ではとりあえず ”no free lunch ” 条件のみを用いて均衡価格が導かれていることに注目しておこう。

4. 資産価格の基本定理

経済においてリスクの伴う証券の数が有限であって、さらに不確実な自然の状態も有限個しかない場合には前節の数値例で説明したことを簡単に拡張することができる。これまで説明した数値例では時刻 $t = 0$ における価格ベクトルを $\mathbf{p}' = (p_1, p_2) = (10, 10)$, 証券 S_1, S_2 の購入量を $\mathbf{y}' = (y_1, y_2)$, 時刻 $t = 1$ における市場価格が

$$\mathbf{Z} = (Z_{ij}) = \begin{pmatrix} 6 & 12 \\ 18 & 6 \end{pmatrix}$$

とおいたと見なすことにしよう。ここで $\mathbf{p}$ と $\mathbf{y}$ は縦に実数を並べたベクトルとするので、$\mathbf{p}'$ は縦ベクトルを横に並べた横ベクトルを表す。

ここでユークリッド空間におけるベクトル間の不等式記号を定義しておこう。二つの m 次元ベクトル $\mathbf{x} = (x_i)$ と $\mathbf{y} = (y_i)$ に対して (i) 「 $\mathbf{x} \geq \mathbf{y}$ 」とは、「任意の i について $x_i \geq y_i$ が成り立つ」ことを意味するものとしよう。(ii) 同様に「 $\mathbf{x} > \mathbf{y}$ 」とは、「任意の i について $x_i \geq y_i$ かつ、ある j が存在して $x_j > y_j$ が成り立つ」ものとする。さらに、(iii) 「 $\mathbf{x} \gg \mathbf{y}$ 」とは、「任意の i について $x_i > y_i$ が成り立つ」ことをそれぞれ意味するものとしよう。

一般に自然の状態が m 個、証券の数が n 個の場合にも（m, n ともに有限とする）同じ記号を用いることにすれば時刻 $t = 1$ において非負の収益が存在する領域はベクトル

$$\mathbf{Zy} \geq \mathbf{0} \tag{6}$$

によって表すことができる。ここで $\mathbf{Z} = (z_{ij})$ は (m, n) 行列で z_{ij} は自然の状態が ω_i の時に（配当込みで定義される）第 j 証券の (非負) 価格を並べた利得行列、ゼ

ロ・ベクトル $\mathbf{0}' = (0, \cdots, 0)$, $\mathbf{p} = (p_i)$ は $n \times 1$（非負）ベクトル、$\mathbf{y} = (y_i)$ は $n \times 1$ ベクトルでそれぞれ時刻 $t = 0$ において既に分かっている各証券の価格と保有額を表している。このとき時刻 $t = 0$ における価値ゼロとなる証券の組み合わせは二つのベクトル $\mathbf{p}$ と $\mathbf{y}$ の内積

$$\mathbf{p}'\mathbf{y} = \sum_{j=1}^{n} p_j y_j = 0 \tag{7}$$

で与えられる。このとき $m = n = 2$ のときの数値例から類推すると次の命題が成立が予想される。定理 2・3 の証明は国友・高橋 (2003)1 章にある。

定理 2 ：次の命題は同値である.
(i) $\mathbf{p}'\mathbf{y} = \sum_{i=1}^{n} p_i y_i \ < \ 0$ ならば $\mathbf{Z}\mathbf{y} \geq \mathbf{0}$ となる実ベクトル $\mathbf{y}$ は存在しない。
(ii) $\mathbf{q}'\mathbf{Z} = \mathbf{p}'$ なる非負ベクトル $\mathbf{q} = (q_i) \geq 0$ が存在する.

ここで条件 (III) $\mathbf{p}'\mathbf{y} < 0, \mathbf{Z}\mathbf{y} \geq \mathbf{0}$ が成り立つときを「第 2 種の裁定機会が存在する」と云うことにする。すなわち、上の定理 1 の条件 (i) は「第 2 種の裁定機会が存在しない」ことと同等である。さらに、条件 (IV) $\mathbf{p}'\mathbf{y} \leq 0, \mathbf{Z}\mathbf{y} > \mathbf{0}$ が成り立つときに「第 1 種の裁定機会が存在する」と云うことにしよう[7]。このとき次の定理が成り立つことがわかる。

定理 3： 方程式

$$\mathbf{q}'\mathbf{Z} = \mathbf{p}' \tag{8}$$

を満足する正ベクトル $\mathbf{q} \gg \mathbf{0}$ が存在する必要十分条件は「第 1 種及び第 2 種の裁定機会が存在しない」ことと同等である。

すなわち、裁定価格に関する第 1 命題として知られている定理 3 は 2 種類の「裁定機会が存在しない」ことと「状態価格が存在する」ことが同値であることを意味している。

ここで数理的な構造を見ると定理 1 は線形代数、あるいは凸解析の分野においては Farkas=Minkowski(ファルカス・ミンコフスキー) の補題として知られている。数値例を元とした図からこの問題の数学的構造を考察すると、この命題は分離定理（separation theorem）と呼ばれているより一般的な数学的命題の特殊な場合となつている[8]。

次に以上で用いた記号を少し修正して証券の理論価格を見てみよう。証券 S_1 の理

[7]第一種の裁定機会は今は元手が例えゼロであっても、確率 1 で将来に必ず正の利益が得られる、第二種の裁定機会は今は元手ななく借り入れを行っているとしても、将来には必ず借入額は返済できる、ことを意味する。

[8]自然の状態が無限個、時間が連続的、取引はいつでも可能、など一般の場合は数理的にはかなり抽象的な問題となる。既に約 30 年前になるが数理経済学、確率論の数学者により詳しいい研究がなされた。主要な結果については例えば Delbaen and Schachermayer (2006), *The Mathematics of Arbitrage*, Springer を参照されたい。

論価格 ψ_1 を関数 $\psi(\cdot)$ と見て $\psi_1 = \psi(S_1) = \sum_{i=1}^{m} q_i z_{i1}$ と書くことにしよう。同様に証券 S_j $(j = 2, \cdots, n)$ についても $\psi_j = \psi(S_j)$ としておくと、複数の証券 S_j より作られるポート・フォリオ $X = \sum_j y_j S_j$ に対して $\psi(\cdot)$ を $\psi(X) = \psi(\sum_j y_j S_j) = \sum_j y_j \psi(S_j)$ によって定めることができる。すなわち関数 $\psi(\cdot)$ は確率変数 X に対する線形汎関数となる。一般に派生証券の価格を確率変数 X で表せば、確率 $P(X = x_j)$ に一対一に対応する評価値を表す正数 $\{q_j\}$ により $\psi(X) = \sum_i q_i x_i$ で与えられる。

オプション契約の数値例

前節で用いた $m = n = 2$ の数値例を利用して可能な限り簡単な応用を考えて見よう。時刻 $t = 1$ における証券 S_1 と S_2 の価格をそれぞれ $S_1(1), S_2(1)$ として、時刻 $t = 1$ において権利行使価格 6 により $\max(S_1(1), S_2(1))$ を受け取ることのできるオプション契約が存在するものとしよう。このオプション契約の時刻 t = 0 における理論価格は前節の議論から容易に計算できる。ここで満期におけるペイ・オフ関数は $X = \max(S_1(1), S_2(1)) - 6$ なので、既に求めた理論確率 $q' = (q_1, q_2)$ を用いて $\psi(X)$ を計算すれば

$$\psi(X) = \mathbf{E}^Q[\max(S_1(1), S_2(1)) - 6] = (12 - 6)\frac{2}{3} + (18 - 6)\frac{1}{3} = 8$$

で与えられる。そこで、このオプション契約を証券 S_3 として表 1.2 のペイ・オフ行列を拡張してみると次のような表が得られる。

オプション契約が存在するこの経済で時刻 $t = 0$ における 3 つの証券 S_1, S_2, S_3

表 3

	証券 S_1	証券 S_2	証券 S_3
(自然の状態)ω_1	6	12	6
(自然の状態)ω_2	18	6	12

の価格がそれぞれ $S_1(0) = S_2(0) = 10, S_3(0) = 8$ としてみよう。このとき拡張されたこの経済においてもやはり”free lunch”は存在しないことを確かめることができよう。ここで (y_1^*, y_2^*) を連立方程式

$$\begin{pmatrix} 6 \\ 12 \end{pmatrix} = y_1^* \begin{pmatrix} 6 \\ 18 \end{pmatrix} + y_2^* \begin{pmatrix} 12 \\ 6 \end{pmatrix}$$

の解とすれば、一意的に解 $(y_1^*, y_2^*) = (\frac{3}{5}, \frac{1}{5})$ が存在することに注意しよう。このことは証券 S_1 を 3/5 単位、S_2 を 1/5 単位保有することにより第 3 の証券をコスト 8 $(= 10 \times (3/5) + 10 \times (1/5))$ で複製 (あるいはヘッジ) できることを意味している。すなわち、仮に時刻 $t = 0$ においてある金融機関がこのオプションを販売したとしても、あらかじめ時刻 $t = 1$ になって発生するリスクをコスト 8 で必ず回避できることを意味している。

ここで説明したオプション契約の価値はある確率測度に関する数学的期待値となったことに注意しておく。この確率測度はオプションの評価法として知られているリスク中立化法（risk neutralized method）における確率測度（リスク中立化確率測度と呼ばれることがある）に対応している。また、ここでまでの議論で議論された理論価格は効用関数等で表現される人々の選好に依存していないこと注意する必要があろう。近年の経済学の標準的議論においては不確実性の存在する経済では均衡においては、人々のリスクに対する選好に応じてリスク・プレミアムが決定されと考えるのが一般的であろう。すなわち、リスクを伴う証券の均衡価格は人々の選好に依存すると一般的には考えられる。これに対して、派生証券の標準理論では、かなり一般的な仮定の下で人々の選好とは独立に確率測度が決まり、派生証券の価格が決定されることが大きな特色である。

3 期間の数値例

前節の数値例を少し複雑にして、リスクを伴い途中では配当がない二つ証券 S_1, S_2 について $t = 0, 1, 2$ の 3 期間における価格変動を考えよう。いま時刻 $t = 0, 1, \cdots$ における証券 S_1, S_2 の価格ベクトルを $\mathbf{S}(t)' = (S_1(t), S_2(t))$ $(t = 0, 1, \cdots, l; l \geq 2)$ とおくときに、これが以下のような数値をとるものとする。

< 図 6 >

$t=0$	$t=1$	$t=2$	(基本事象)
		$\nearrow \begin{pmatrix} 4 \\ 10 \end{pmatrix}$	ω_1
	$\begin{pmatrix} 6 \\ 12 \end{pmatrix} \rightarrow$	$\begin{pmatrix} 6 \\ 12 \end{pmatrix}$	ω_2
	$\nearrow$	$\searrow \begin{pmatrix} 7 \\ 13 \end{pmatrix}$	ω_3
$\begin{pmatrix} 10 \\ 10 \end{pmatrix}$			
	$\searrow$	$\nearrow \begin{pmatrix} 20 \\ 8 \end{pmatrix}$	ω_4
	$\begin{pmatrix} 18 \\ 6 \end{pmatrix} \rightarrow$	$\begin{pmatrix} 18 \\ 6 \end{pmatrix}$	ω_5
		$\searrow \begin{pmatrix} 16 \\ 4 \end{pmatrix}$	ω_6

この数値例における時刻 $t = 0$ における時刻 $t = 1$ で実現するであろう事象を対象とした ”no free lunch ” 条件については既に説明した。その結果としては状態

価格（理論確率）$q_1 = 2/3, q_2 = 1/3$ が既に得られた。この場合は金利はゼロの場合に対応する。

次にいま仮に時刻 $t = 1$ において価格が $S(1)' = (6, 12)$ となっているものとしよう。この価格が市場でついている時には時刻 $t = 2$ において意図的に 3 つの可能性 $(\omega_1, \omega_2, \omega_3)$ を考えたので少し複雑になっている。$S(1)' = (6, 12)$ からそれぞれの事象に進む確率をそれぞれ $q_{1|1}, q_{2|1}, q_{3|1}$ とすると、再び ”no free lunch ” 条件（(8) 式）と確率の基準化則 $q_{1|1} + q_{2|1} + q_{3|1} = 1$ を満たす確率として、例えば $q_{1|1} = 1/4, q_{2|1} = 1/4, q_{3|1} = 1/2$ とすればよい。全く同様の議論を時刻 $t = 1$ において価格が $S(1)' = (18, 6)$ の場合にも適用してみよう。この場合 3 つの事象に進む確率をそれぞれ $q_{1|2}, q_{2|2}, q_{3|2}$ とおけば、例えば $q_{1|2} = q_{2|2} = q_{3|2} = 1/3$ と置けば整合的であることも確認されよう。なお、ここで例として挙げた確率 $q_{i|j}$ $(i = 1, 2, 3; j = 1, 2)$ は条件付確率であることに注意しておこう。この数値例で表される経済における基本事象を図のように $(\omega_1, \cdots, \omega_6)$ とすれば、基本事象の確率は条件付確率の公式を利用すると

$$Q(\omega_1) = Q(\omega_2) = \frac{1}{6}\,, Q(\omega_3) = \frac{1}{3}\,, Q(\omega_4) = Q(\omega_5) = Q(\omega_6) = \frac{1}{9}$$

となることがわかる。

なお、ここで例示した数値でなくとも事象 j を条件とする事象 i の条件付確率についての関係 $2q_{1|1} = q_{3|1}$, $q_{2|1} = 1 - q_{1|1} - q_{3|1}$, $q_{1|2} = q_{3|2}$, $q_{2|2} = 1 - q_{1|2} - q_{3|2}$, $q_{i|j} > 0$ $(i = 1, 2, 3; j = 1, 2)$ を満足していれば「裁定機会が存在しない条件」(8) 式）と整合的となっている。

さてここで数値例として求めた確率測度を Q としよう。この確率測度 Q によって証券価格 $\mathbf{S}(t)$ の動き $(t = 0, 1, 2)$ を特徴づけることができる。いま時刻 $t = 1$ において $t = 0$ の価格ベクトル $\mathbf{S}(0)$ を所与とすると $\mathbf{S}(1)$ の (Q に関する) 条件付期待値は

$$\mathbf{E}^Q[\mathbf{S}(1)|\mathbf{S}(0)] = \frac{2}{3}\begin{pmatrix} 6 \\ 12 \end{pmatrix} + \frac{1}{3}\begin{pmatrix} 18 \\ 6 \end{pmatrix} = \mathbf{S}(0)$$

となる。さらに時刻 $t = 0, 1$ の証券価格 $\mathbf{S}(0), \mathbf{S}(1)$ を所与とするとき時刻 $t = 2$ における証券価格 $\mathbf{S}(2)$ の条件付期待値を $\mathbf{E}^Q[\mathbf{S}(2)|\mathbf{S}(1), \mathbf{S}(0)]$ で表せば、これが $\mathbf{S}(1)$ に一致することがわかる。一般に条件付期待値について確率 1 で

$$\mathbf{E}^Q[\mathbf{S}(t)|\mathbf{S}(t-1), \cdots, \mathbf{S}(0)] = \mathbf{S}(t-1) \tag{9}$$

を満足するときこの $\{\mathbf{S}(t)\}$ は離散時間のマルチンゲール（martingale）と呼ばれる。マルチンゲールは 2 節の t 例のような「公平な賭（fair gamble）」の概念を一般化した確率過程（stochastic process）であるが、 本節では無裁定条件 (no aribitrager conditiobn) から $\mathbf{S}(t)$ $(t = 0, 1, 2)$ のマルチンゲール性を説明したわけである。

ここでとりあげた No-Free-Lunch をいう命題の内容は近年のファイナンス分野では「資産価格の基本定理」(Fundamental Theorems of Asset Prices) と呼ばれ、しばらく前からファイナンス (金融経済・企業金融) 分野の基礎になっている[9]。

5. 保険と期待値原理

事象 (event, 集合) に対して確率 (probability) が定義されるとリスク (risk) が定義される。リスクを市場で取引しようとすると事象の価値、あるいはリスクの市場価値を評価する必要がある。この問題は歴史的には保険 (insurance) の意味に関する統計的議論などから考察されてきたが次の例が一つの解釈を与えている。

毎回の結果として仮に H(表) か T(裏) という二つの事象の中のどちらかが起き、3 回繰り返されると云う毎回の支払い (payoff) が 1,-1 となる市場ゲームを考える。H か T の市場確率をそれぞれ 1/2 として、3 回のコイン投げの結果の事象 $\mathbf{\Omega} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ はどれも同等に確からしい。この市場に参加できるある保険会社が、市場にはアクセスできないある個人 K に対して 3 回ともに T が起きる事象ときに 1 単位保証する保険 $X(\omega)$ を販売することを考える。(つまり $X(TTT) = 1$, その他の ω については $X(\omega) = 0$ である。) 会社の運営費を無視すると、この保険の販売に際して前もって請求すべき保険料は期待値

$$\mathbf{E}[X(\omega)] = 1 \times \frac{1}{8} + 0 \times \frac{7}{8} \tag{10}$$

であれば良い。

このことは次のように考えれば納得しやすいであろう。個人 K より保険料として 1/8 を徴収した会社は T に投資する。もし最初の結果が H であればこの保険会社は事象 TTT は起きることがないので事業は終了である。もし最初の結果が T であればこの保険会社の収入は 1/8 あるので資産は 1/8+1/8=1/4 に増加する。この場合には 2 回目にやはり T に 1/4 を投資すると、同様の議論より 2 回目に H が出れば終了、T が出れば資産は 1/4+1/4=1/2 となる。3 回目も同様にするとこの場合には 3 回目にもやはり T に 1/2 を投資すると、同様の議論より 3 回目に H が出れば終了、T が出れば資産は 1/2+1/2=1 となり、この 1 を事象 TTT となることを恐れて保険契約に加入していた個人に対して支払いを行えばすべての契約が履行できる。ここで保険会社は契約上で支払う可能性に対して複製戦略 (replicating strategy) を構成していることに注目しよう。この戦略はファイナンスにおけるデリバティブ (derivatives, 派生証券) の議論と同一である。この簡単な例では可能であれば個人 K が直接に市場に参加して自分で自分のリスクをヘッジしてもよいが、保険会社は契約により生じるリスクを徴収した保険料を利用して 100% ヘッジ (回避) しているのである。

[9]例えば「数理ファイナンスの基礎」(国友直人・高橋明彦,2003, 東洋経済)1 章〜3 章。本格的な数学系の書籍としては Delbaen F. and W. Schachermayer (2006),"The Mathematics of Arbitrage," Springer などがある。連続時間の確率過程の応用に関心があればまずは前者を参照されたい。

このような数学的期待値 $\mathbf{E}[X]$ に基づく保険契約の評価法を保険計算の分野では統計的原理、あるいは期待値原理 (expectation principle) と呼ばれている。

他方、経済・金融の分野では不確実性に直面する経済人は期待値ではなく期待効用を最大化するように不確実性を評価するとの考え方が一般的である。不確実性を含むある金融契約の価値を確率変数 X, 効用関数 $U(X)$ として、例えば X が (宝くじの賞金額)、C $(> E[X])$ を (宝くじの購入額)、宝くじの購入者はかりに不確実性 (リスク) がなかったとしたときの評価 $CQ(X)$ としよう。このとき $CQ(X) < E(X)$ のときに宝くじを購入するのが合理的と解釈できる。宝くじでは C は $E(X)$ よりかなり高いが、購入者にとっては期待効用 $\mathbf{E}[U(X)]$ $(> CQ(X))$ より宝くじにより満足が得られることになる。このように $CQ(X) < \mathbf{E}[X]$ と評価する人を危険愛好的 (risk lover)、また $CQ(X) > \mathbf{E}[X]$ と評価する人を危険回避的 (risk averter)、$CQ(X) = \mathbf{E}[X]$ と評価する人を危険中立的 (risk neutral)、と呼ばれている。こうした期待効用をより大きくするという行動原理は期待効用最大化原理と呼ばれるが、期待値は低くとも宝くじを購入したり、ギャンブルを行う行動、あるいは逆にリスクを回避するために保険を契約する行動などが共存する社会・経済におけるリスクに関わる現象を説明可能としている。

なお実際に期待値を計算するには確率、確率分布が必要であり、経済・経営の分析では主観確率 (subjective probability)、個人確率 (personal probability)、により評価と解釈されることが多い。また、実際の金融市場や保険市場は情報や不確実性の源泉が不完全な場合が多いので (市場の非完備性と云われることがある) 期待効用原理が保険の期待値原理と必ずしも矛盾するとは言えないのである。

6. 市場と連続時間の確率過程

金融市場で観察される価格変動は近年では短い時間間隔においても観察されるようになっている。こうした細かな時間間隔、その極限としての連続時間の確率過程の問題を考えよう。

区間 $[0,T]$ 上を TN 等分して各 $i/(TN)$ $(i = 1, \cdots, TN)$ 時点で確率変数 Z_i のある確率測度 P についての期待値 $\mathbf{E}_Q[Z_i] = 0$, 分散 $\mathbf{V}_Q[Z_i] = \sigma^2/(TN)$ とする。すなわち微少な区間での確率変数の区間当たりの分散を $\sigma^2(1/TN)$ とする。簡単化して $T = 1$ とおき、$k(n)$ を n に依存させて各時刻 $(k(N)/N) = t(k_N) \in [0,1]$ 上の $k(N)$ 個の確率和 $X_N(t(k_N)) = (\frac{1}{\sigma}) \sum_{i=1}^{k(N)} Z_i$ として ($[a]$ は a を超えない最大の整数)、$[0,1]$ 上の任意の時刻 t において補間により $X_N(t(k_N))$ が連続経路 (つまり $t(k_N)$ について X が連続) をとるようにしよう。このとき $N \to \infty$ につれて $t(k_N) \to t \in [0,1]$ に対して確率分布の意味で収束するはずなのでこれを

$$X_N(t(k_N)) \xrightarrow{d} X(t)\ , \tag{11}$$

と表現しよう。ここで極限として表れる確率過程 $X(t)$ は中心極限定理より $X(t) \sim N(0,t)$ となるはずである。またこの確率過程は任意の $0 \le s < t \le 1$ に対し $X(s)$

と $X(t) - X(s)$ は独立な確率変数となるはずである。

ここでの議論は一種の中心極限定理の応用であるが、確率変数 $X(t)$ $(t \in [0,1])$ はブラウン運動 (Brownian Motion) と呼ばれている連続時間上の確率過程として数学的に意味を持つことが知られている。ブラウン運動の実現経路は時間軸上の経路は不規則な変動を示す。任意の t に対して正規分布 $N(0,t)$ にしたがい、t について連続な経路をとる確率変数列 $X(t)$ が存在するのである。こうした確率変数列が意味を持つか否かは自明ではないが、このブラウン運動は任意の ω に対して非有界変動、二乗変動は有界で $E[(X(t))^2] = t$ となることが知られている。連続時間の確率過程であるブラウン運動はマルチンゲール性を持つので、ブラウン運動やその関数 (汎関数と呼ばれる) は資産価格の動的経路を分析する手段として広く利用されている。

なお価格過程は非負であるのである時刻 t における価格 P_t の対数をとり $Y_t = \log P_t$ とすると、$Y_t - Y_0 \sim N(\mu t, \sigma^2 t)$ となる幾何ブラウン運動と呼ばれる統計モデルは連続時間の金融時系列における標準となっている。

7. 金融市場の計量分析

古くから株価、外国為替レート、国債など債券価格、金利などを始め金融市場で観察される価格変動は短期的には激しく変動するため予測が困難であることが観察されている。仮にたまたまある時点において近い将来の価格水準の予測できたとしても、しばらく先の時点で同様のことが起きる可能性は大きくないのである。一見すると逆説的に聞こえるが、金融市場などで観察される価格”水準”データでは将来値の予測力が大きくないことが市場経済における”正常な状態”と云ってよい。実はこうした経験則は時系列の統計分析を学ぶことにより確認できるとともに、さらに新たな展望が得られるのである。統計分析においてランダムオーク・モデル、単位根が存在する場合は非定常確率過程と呼ばれている。こうした非定常時系列の性質を議論するには階差系列、(対数) 収益率系列 (すなわち配当などを無視して単純化した場合には、ある時刻 n(離散時間 $n = 1.2, \cdots, T$ とする) の価格 P_n より $Y_n = \log[P_n/P_{n-1}]$ で定義される）などがほぼ無相関な時系列や定常系列と見なしうる系列に変換した後に統計分析を行うことが、今日では金融データの統計学的分析においてきわめて常識的なのである。また秒単位の高頻度金融データの解析もかなり行われている。

近年では金融時系列のリスク指標として収益率の分散をボラティリティと呼び、派生証券 (derivatives) の取引と共にボラティリティが時間とともに変動する確率過程・統計モデルの利用も行われている。

時系列データと理論価格

ここで No-Free-Lunch 条件から導かれるマルチンゲール性と現実にデータとして観察される金融価格との関係について考えてみよう。この問題は実は理論的に導

きられるマルチンゲール性の確率測度 Q と実際に観察される確率 P との関係と見ることができる。確率測度 Q で評価される事象 A が $Q(A) = 0$ *or* $Q(A) = 1$ のとき、別の確率測度 P についてそれぞれ $P(A) = 0$ *or*$P(A) = 1$ を満足するとき確率測度 Q と P は絶対連続 (absolutely continuous) と呼ばれている。したがってここでの問題は Q についてのマルチンゲール性と絶対連続な確率過程は何か？という数理的問題なのである。例えば $Z_n = Y_n - Y_{n-1}$, $Z_n(n = 1, \cdots, T)$ を $\mathbf{E}[Z_n] = 0$,$\mathbf{E}[Z_n] = 1$ の互いに独立な正規過程とすると、

$$X_n = Z_n + \theta_T Z_{n-1} \ , \tag{12}$$

ただし ($\theta_n = c/\sqrt{T}, c$ は定数) とすると。このとき移動平均過程 MA(1) にしたがう X_n の同時分布は Z_n の系列と絶対連続となることを示すことができる[10]。したがって $P_n = P_{n-1} + Z_n$ は ARIMA(0,1,1) にしたがう。

すなわち、金融データから計算される階差データやリターン・データに僅かな自己相関が見られたとしても、無裁定条件に矛盾するとは限らないことになる。ただしこの場合には階差データの自己相関はあまり大きくないので実際の多くのデータ分析の結果と符合すると考えられる。

8. 幾つかの注意点

ここで本稿の 3 節-6 節で述べた「資産価格の基本定理」の視点から見た時の実際の日米株価データの統計分析についての幾つかの論点を述べておこう。

(i)「統計学基礎 5 章」で説明している線形モデルは誤差が互いに独立に正規分布 $N(0, \sigma^2)$ にしたがうと仮定された統計モデルであり、係数の有意性、安定性・フィットの良さ、などの統計量の説明はそうした仮定の下で正当化されている。日米の株価の片方を説明変数、他方を被説明変数と見なすには無理がある。例えば日本の投資家（米国の投資家も同様であるが）は日本と米国の株式市場で時差はあるがネットを利用すればほぼ同時に取引できる。なお図 2 では x 軸に日経 225,y 軸にダウ (円建て) をプロットしているが、これは同日データと云っても日本の日付は米国の日付より僅かに先行する為である。実証的には前日の米国の株価 y から翌日の日本の株価 x への影響の方が同日における日本の株価 x から米国の株価 y への影響よりも大きいことが知られている。

(ii) 与えられた統計データは歴史的に得られた 3 つの時系列であり、互いに独立な確率変数の実現系列データとは考えられない。また (日本の投資家にとり重要と思われる) 円で測った日米の株価の分析では第 3 の変数である円・ドル為替レートの変動分析は欠かせない。

(iii) 一般にマルチンゲール系列は高い自己相関 (autocorrelation) [11]をを示すことが

[10] こうした内容についての文献はあまり見かけないが、測度論的確率論の議論を利用する必要があり、詳細は他の機会に譲る。

[11] 時刻 t, s の確率変数 X_t と X_s の相関係数、互いに独立なら 0 となるが、マルチンゲールならデータから計算される相関は 1 に近い。$cor(X_t, X_s)$ は時間差 $|t - s|$ にのみ依存する場合は (弱) 定常 (stationary) と呼ばれている。詳しくは統計的時系列分析の書籍を参照されたい。

知られていて、この事実をもとに Granger-Newbold (1974, Journal of Econometrics) は「Sprious Regression(見せかけの回帰)」を提唱、様々な議論が行われている。他方、株価などの金融時系列では差分 (原系列 Y_n とすると $X_n = Y_n - Y_{n-1} (n = 1, \cdots, n, n = 0$ は初期値)) をとると自己相関がかなり小さくなることが知られており、金融時系列では差分系列や収益率系列の統計分析が盛んにおこなわれている。金融時系列の水準系列間の関係については Granger-Engle (1987, Econometrica) 以来、共和分 (co-integration) を巡る議論が時系列計量経済 (time series econometrics) では様々な議論が行われている。

(vi)2 節のデータは株価インデックスの週次データであるが近年では株価については日次、1 分、1 秒あるいは全ての取引データが利用可能[12] であり、こうしたデータ分析は高頻度金融データ分析と呼ばれている。こうしたデータ分析と連続時間確率モデルとの関係の分析[13]も研究されている。

(v) 例えば投資家にとり現在・過去のデータによる統計分析は将来の行動への指針として大きな意味があると考えられる。統計的時系列分析では現在・過去のデータ分析から将来の動向についての何らかの意味で予測できるという論点が重要であり、統計的時系列論の主要な論点は「将来の動向の予測」の観点から開発されているという経緯が重要である。マルチンゲールを含む多次元時系列分析は多次元非定常時系列分析 (multivariate non-stationary time series analysis) と呼ばれ今なお研究が行われている。むろん様々な方法が提案されているが、その詳細は別の機会としよう。

(付録：参考としたパンフレットと注意)

2022/4/8 の記事（トウシル Rakuten 楽天証券「3 分でわかる！今日の投資戦略〔平日毎朝 8 時掲載〕日経平均株価の上値は？「円建て NY ダウ」の行方から占う」）
https://media.rakuten-sec.net/articles/print/36778
はたまたまネットで公開されていたパンフレットであり、2024 年 1 月 17 日にも検索可能であった。削除されることもあり得るが、その場合には (時代とともに内容は変化するが) 近くの証券会社や銀行の窓口で相談すると類似の記事が得られるはずなので参考情報を得られたい。

[12]東京証券取引所をはじめ主要な取引所データは有料ではあるが一般に公開されている。

[13]例えば Kunitomo, Sato and Kurisu (2018),*Separating Information Maximum Likelihood Method for Hight Frequency Data* Springer などを参照されたい。

2023.12
統計数理研究所 湯浅 良太

統計エキスパート演習
ビッグデータの統計的パラドックスについての論文紹介

本報告の概要

大学統計教員育成研修の中で、統計を用いる多分野の専門家による講演が連続講義として行われた. その中で日経リサーチ技術顧問、統計研究研修所客員教授などを務める鈴木督久先生による「世論調査と統計学」に関する講演が行われた. その中で重要な論点として、偏った集団からの多数のデータとランダムサンプリングに基づく少数のデータによる統計的推測の比較があった. 本報告では、この話題に関連する論文である Meng (2018) "Statistical paradises and paradoxes in big data (I) Law of large populations, big data paradox, and the 2016 US presidential election" を紹介する.
「母集団の 1% に対するランダムサンプリングによる回答率 60% のデータと、母集団の 80% に対してだが偏った集団を対象としたデータではどちらの方が良いと言えるか」「偏った集団を対象とした調査であっても $50\%, 60\%, 70\%, 80\%, 90\%, 95\%, 99\%$ と調査対象の割合を増やすと、いつランダムサンプリングより信頼できると言えるか」といった疑問に取り組む事を目的とした論文である. そのためのフレームワークが有限母集団の枠組みの中で導入され、2016 年のアメリカ大統領選を例にその適用がなされている.
本報告は連続講義に際して提出したレポートに加筆したものである. いくつかの数学的な証明は補足において与えている.

問題設定と要となる等式

有限母集団を考え、母集団に属する人は N 人とする.
$X_j, j = 1, \dots, N$ が、j 番目の人に関するデータとする.
適当な関数 $G : X \mapsto G(X) \in \mathbb{R}$ により、興味のある値が表されるとする. $G_j \coloneqq G(X_j), j = 1, \dots, N$ とし、この平均 $\bar{G}_N \coloneqq \sum_{j=1}^N G_j/N$ を推定することに興味があるとする.
サンプルは $\{X_j, j \in I_n\}$、ただし、I_n は大きさ n の $\{1, \dots, N\}$ の部分集合とする. $\bar{G}_N$ の推定量は

$$\bar{G}_n = \frac{1}{n} \sum_{j \in I_n} G_j = \frac{\sum_{j=1}^N R_j G_j}{\sum_{j=1}^N R_j},$$

ただし、$j \in I_n$ なら $R_j = 1$、$j \notin I_n$ なら $R_j = 0$ とする. $\sum_{j=1}^n R_j = n$ である事に注意.

$\{A_1, \dots, A_N\}$ を数の集合とする. 確率変数 J が $\{1, \dots, N\}$ 上で定義されるとき、この J によって確率変数 A_J を定める. もし J が一様分布なら、$E_J(A_J) = \sum_{j=1}^N A_j/N$ である.
J が一様分布に対し、$f = E_J(R_J) = n/N, \rho_{R,G} = \mathrm{Corr}_J(R_j, G_j), \sigma_G$ を G_j の標準偏差とすると、$\mathrm{Var}(R_J) = E[R_J^2] - E[R_J]^2 = f - f^2$ であり、

$$\bar{G}_n - \bar{G}_N = \frac{E_J(R_J G_J)}{E_J(R_J)} - E_J(G_J) = \frac{\mathrm{Cov}_J(R_J, G_J)}{E_J(R_J)} = \rho_{R,G} \times \sqrt{\frac{1-f}{f}} \times \sigma_G. \tag{1}$$

この式がこの論文における要となる式である.
$(1-f)/f$ の部分はデータの多さの影響を表している. サンプリング率 f が 1 に近いと精度が良く、0 に近づくと精度が悪化する事が分かる.
σ_G は問題の難しさを表している. もし G_J が一定ならデータが 1 つでも推定は上手く行く.
$\rho_{R,G}$ はデータの質を表している. 例えば大きな値ばかりサンプリングされていると $\rho_{R,G}$ は 1 に近づき推定精度は下がる事を示す. 満遍なくサンプリングされていれば 0 に近づき推定精度が高くなる事を示す.

また、平均二乗誤差は

$$E_{\boldsymbol{R}}[(\bar{G}_n - \bar{G}_N)^2] = E_{\boldsymbol{R}}[\rho_{R,G}^2] \times \frac{1-f}{f} \times \sigma_G^2 =: D_I \times D_O \times D_U. \tag{2}$$

ただし、$\boldsymbol{R} = \{R_1, \ldots, R_N\}$ に対し、$E_{\boldsymbol{R}}$ が $\sum_{j=1}^N R_j = n$ を満たす $\boldsymbol{R}$ の何らかの分布に関する期待値を表す.
$D_I = E_{\boldsymbol{R}}[\rho_{R,G}^2]$ は、data defect index と呼ばれ、偏った集団からのサンプリングであるかどうかを示す.
$D_O = (1-f)/f$ は、dropout odds と呼ばれ、データの量が多いか少ないかを示す.
$D_U = \sigma_G^2$ は、degree of uncertainty と呼ばれ、問題の難しさを示す. 追加の情報を得る事で減らす事が出来る.

D_I は、分析の目的 (G の選択)、解析手法、実際に得られるデータ (R の構造) などによって影響を受ける. D_I に関していくつかの考察を行う.
N 個から n 個を重複なく選ぶ ${}_NC_n$ 通りそれぞれが等確率でサンプリングされる、単純ランダムサンプリング (SRS, simple random sampling) の下、$E_{SRS}(\bar{G}_n) = \bar{G}_N$ であり、

$$V_{SRS}(\bar{G}_n) = \frac{1-f}{f}\frac{1}{N-1}\sigma_G^2 = \frac{1-f}{n}\frac{N}{N-1}\sigma_G^2 =: \frac{1-f}{n}S_G^2 \tag{3}$$

で、これは $\bar{G}_n$ の平均二乗誤差は分散と一致するので、(2) の左辺に代入すれば $D_I = 1/(N-1)$ となる.

得られたサンプルは全て $R_J = 1$ となるので、D_I はサンプルから推定する事はできない. ただし、例えば選挙後に D_I を推定する事は無理ではない. 過去のデータや類似のデータから D_I や $\rho_{R,G}$ の事前分布を構成して用いうるかもしれない. この点は重要な論点の 1 つであるが、少なくとも紹介論文の本文中ではあくまで枠組みの提示までが主となっており、特に推定方法について深入りはされていない.

j 番目の人がトランプを支持するならば $G_j = 1$、そうでないなら $G_j = 0$ とする. $p_G = P_J(G_J = 1), O_G = p_G/(1-p_G)$ とすると、Hoeffding の等式と Fréchet bounds により、

$$-\min\left\{\sqrt{\frac{D_O}{O_G}}, \sqrt{\frac{O_G}{D_O}}\right\} \le \rho_{R,G} \le \min\left\{\sqrt{O_G D_O}, \frac{1}{\sqrt{O_G D_O}}\right\}. \tag{4}$$

質を量で補えるか？

(1), (3) から、

$$Z_{n,N} \coloneqq \frac{\bar{G}_n - \bar{G}_N}{\sqrt{V_{SRS}(\bar{G}_n)}} = \frac{\rho_{R,G}\sqrt{\frac{1-f}{f}}\sigma_G}{\sqrt{\frac{1-f}{n}S_G^2}} = \sqrt{N-1}\rho_{R,G}. \tag{5}$$

つまり、$E_{\boldsymbol{R}}(\rho_{R,G}) \neq 0$ を共通とする調査の中で、$\bar{G}_n$ の誤差は SRS の下での誤差をベンチマークとして、母集団のサイズ N が増加するにつれ $\sqrt{N}$ のレートで増加する. これをこの論文では大母集団の法則 (LLP, Law of Large Population) と呼んでいる. 母集団のサイズが大きくなるとその分推定誤差が大きくなることを示している. こうした状況をビッグデータにおける統計的パラドックスと本論文では呼んでいる.
この (5) から $\log Z_{n,N} = \log \rho_{R,G} + 0.5\log(N-1)$ である. もしサンプリングバイアスが存在する場合 N が変化しても $\rho_{R,G}$ は安定した大きさの値をとる. 一方で、サンプリングバイアスがなければ $\log \rho_{R,G}$ は $\log(N-1)$ を打ち消すような値をとる. よって、

$$\log|Z_{n,N}| = \alpha + \beta \log N$$

という回帰を考え、β が 0 に近い値をとればサンプリングバイアスが弱く、β が 0.5 に近い値をとればサンプリングバイアスが強いという分析を行うことができる.

また、(2), (3) から、(lack-of-)design effect と呼ばれる量が次のように表される.

$$\text{Deff} = \frac{E_{\boldsymbol{R}}[(\bar{G}_n - \bar{G}_N)^2]}{V_{SRS}(\bar{G}_n)} = (N-1)E_{\boldsymbol{R}}(\rho_{R,G}^2) = (N-1)D_I. \tag{6}$$

(3), (6) から、明らかに次の定理が成立する.

Theorem 1. サンプリング率 $0 < f < 1$, 問題の難しさ $D_U = \sigma_G^2$ を固定した下、サイズ $\{N_\ell\}$ が増加する母集団の列で、$\lim_{\ell\to\infty} N_\ell = \infty$ となるとする. このとき、サンプルサイズの列 $n_\ell = fN_\ell \to \infty$ である. $A_N = O(B_N)$ はそれぞれ $\limsup_{\ell\to\infty}(|A_N|/|B_N|) < \infty$ を表すとする. 次の 3 条件は同値である.
(1) $\mathrm{Deff} = O(1)$
(2) $\mathrm{MSE}_{\boldsymbol{R}}(\bar{G}_n) = O(n^{-1})$
(3) $D_I = O(N^{-1})$

よって、D_I を N^{-1} のレート、つまり $\rho_{R,G}$ を $N^{-1/2}$ のレートでコントロールする事で、平均二乗誤差が n^{-1} のレートとなる. 2016 年のアメリカ大統領選の投票者数は $N \approx 1.4 \times 10^8$ なので、$\rho_{R,G} \approx 8.4 \times 10^{-5}$ が求められる.
例えば $p_G = P_J(G_J = 1) = 1/2$ のとき、(4) から

$$D_I \leq \min\left\{\frac{f}{1-f}, \frac{1-f}{f}\right\} = \min\left\{\frac{n}{N-n}, \frac{N-n}{n}\right\} \tag{7}$$

なので、サンプリング率が 0 か 1 に偏っているときには $D_I = O(N^{-1})$ が成立する. ただしこれは定理 1 のサンプリング率が固定されているという条件とは異なるため、あくまで D_I の挙動に関するイメージを掴むための性質である.

(2) から $D_I D_O$ をコントロールすることが重要である. $\bar{G}_n$ の MSE と、サンプルサイズ n_{eff} の SRS の下での推定量の MSE が等しくなるものとして、有効サンプルサイズ n_{eff} を定める. つまり、(2), (3) から有効サンプルサイズ n_{eff} は次を満たす.

$$D_I D_O = \left(\frac{1}{n_{\mathrm{eff}}} - \frac{1}{N}\right)\left(\frac{N}{N-1}\right). \tag{8}$$

$n_{\mathrm{eff}}^* = (D_O D_I)^{-1}$ とすると $n_{\mathrm{eff}} = n_{\mathrm{eff}}^*/\{1 + (n_{\mathrm{eff}}^* - 1)N^{-1}\}$ なので、

$$n_{\mathrm{eff}} \leq n_{\mathrm{eff}}^* = \frac{f}{1-f} \times \frac{1}{D_I} = \frac{n}{1-f} \times \frac{1}{ND_I}$$

定理 1 の条件下では $ND_I = O(1)$ で N の有効サンプルサイズへの効果は D_I で相殺される. もし $D_I = O(1)$ であれば ND_I は増加し n_{eff} は大きく減ってしまうことになる. $D_I = E_{\boldsymbol{R}}(\rho_{R,G}^2) \geq [E_{\boldsymbol{R}}(\rho_{R,G})]^2$ より、$E_{\boldsymbol{R}}(\rho_{R,G}) \neq 0$ で N が大きくなっても消えないなら $D_I = O(1)$ である. ビッグデータと呼ばれる状況の多くはこれである.

モチベーションとなった疑問への回答

「母集団の 1% に対するランダムサンプリングによる回答率 60% のデータと、母集団の 80% に対してだが偏った集団を対象としたデータではどちらの方が良いと言えるか」に答える. ランダムサンプリングによるものを $f_s = n_s/N$、 $D_I^{(s)} = \mathrm{Deff}/(N-1), D_O = (1 - rf_s)/(rf_s)$、ただし r は回答率とする. ビッグデータによるものを $D_O^{\mathrm{BIG}} = (1-f)/f$ とする. $n_{\mathrm{eff}}^{\mathrm{BIG}} > n_{\mathrm{eff}}$ は $D_I^{\mathrm{BIG}} D_O^{\mathrm{BIG}} < D_I D_O$ と同値.

$$\mathscr{O} = \frac{D_O}{D_O^{\mathrm{BIG}}} = \frac{1 - rf_s}{rf_s} \times \frac{f}{1-f}$$

とする. $D_I \approx \rho_{R,G}^2$ なので $D_I^{\mathrm{BIG}} D_O^{\mathrm{BIG}} < D_I D_O$ は $|\rho_{R,G}^{\mathrm{BIG}}| \leq \sqrt{\mathscr{O}}|\rho_{R,G}|$ となる.
$f >> rf_s$ なら $\sqrt{\mathscr{O}} >> 1$ である. $f_s = 0.01, r = 0.6, f = 0.8$ なら $\sqrt{\mathscr{O}} \approx 26$ である. サンプリングの偏りの構造が大きく異なりはしないなら、ビッグデータの方が信頼に値する.
一方、2016 年アメリカ大統領選の有権者数 $N \approx 231,55,000$ のとき、SRS の下で $|\rho_{R,G}^{(s)}| \approx \sqrt{2/\pi}(N-1)^{-1/2} = 5.2 \times 10^{-5}$ で、もしビッグデータの場合にサンプリングに偏りが大きく、$\rho_{R,G} \approx 5 \times \rho_{R,G}^{(s)} = 2.6 \times 10^{-4}$ とすると $26 \times \rho_{R,G} \approx 0.0068$ である. アメリカ大統領選の 2016 年のデータから先の等式に基づき $\rho_{R,G}$ を推定する事により、比較を行うと 80% に対する調査ではまだビッグデータが信用に値するが、50% となると母集団の 1% に対するランダムサンプリングによる回答率 60% のデータの方が信頼に値する事が示される.

補足

補足として、本報告で現れたいくつかの等式や不等式に対する証明を与える.

(3) の証明

まず $E_{SRS}(\bar{G}_n) = \bar{G}_N$ を示す.

$$E_{SRS}(\bar{G}_n) = \frac{1}{n} E_{SRS}\left(\sum_{j=1}^{N} R_j G_j\right) = \frac{1}{n}\sum_{j=1}^{N} E_{SRS}\left(R_j G_j\right) = \frac{1}{n}\sum_{j=1}^{N} G_j P(R_j = 1) = \frac{1}{n}\sum_{j=1}^{N} G_j \frac{{}_{N-1}C_{n-1}}{{}_{N}C_{n}}$$
$$= \frac{1}{n}\frac{n}{N}\sum_{j=1}^{N} G_j = \frac{1}{N}\sum_{j=1}^{N} G_j = \bar{G}_N.$$

次に $E_{SRS}(\bar{G}_n^2)$ を計算する.

$$E_{SRS}(\bar{G}_n^2) = \frac{1}{n^2} E_{SRS}\left(\sum_{j=1}^{N} R_j G_j^2\right) + \frac{1}{n^2} E_{SRS}\left(\sum_{j=1}^{N}\sum_{i\neq j} R_i R_j G_i G_j\right)$$
$$= \frac{1}{n^2}\frac{n}{N}\sum_{j=1}^{N} G_j^2 + \frac{1}{n^2}\frac{{}_{N-2}C_{n-2}}{{}_{N}C_{n}}\sum_{j=1}^{N}\sum_{i\neq j} G_i G_j$$
$$= \frac{1}{nN}\sum_{j=1}^{N} G_j^2 + \frac{1}{n^2}\frac{n(n-1)}{N(N-1)}\left\{\left(\sum_{j=1}^{N} G_j\right)^2 - \sum_{j=1}^{N} G_j^2\right\}$$
$$= \frac{1}{nN}\frac{N-n}{N-1}\sum_{j=1}^{N} G_j^2 + \frac{(n-1)N}{n(N-1)}(\bar{G}_N)^2.$$

以上より、(3) を示す.

$$V_{SRS}(\bar{G}_n) = E_{SRS}(\bar{G}_n^2) - E_{SRS}(\bar{G}_n)^2 = \frac{N-n}{n(N-1)}\left(\frac{1}{N}\sum_{j=1}^{N} G_j^2 - (\bar{G}_N)^2\right) = \frac{N-n}{n(N-1)}\sigma_G^2 = \frac{1-f}{f}\frac{1}{N-1}\sigma_G^2.$$

Hoeffding の等式

Hoeffding の等式 $\mathrm{Cov}(X,Y) = \iint\{F_{X,Y}(x,y) - F_X(x)F_Y(y)\}dxdy$ を示す.
$(X_1,Y_1),(X_2,Y_2)$ が (X,Y) と同一の分布にそれぞれ独立に従うとする. $X_1 - X_2 = \int 1_{u\leq X_1} - 1_{u\leq X_2} du$ であることを用いて、

$$2\mathrm{Cov}(X,Y) = 2\{E(X_1Y_1) - E(X_1)E(Y_1)\} = E[(X_1 - X_2)(Y_1 - Y_2)]$$
$$= E\left[\iint\{1_{u\leq X_1} - 1_{u\leq X_2}\}\{1_{v\leq Y_1} - 1_{v\leq Y_2}\}dudv\right]$$
$$= \iint 2F_{X,Y}(u,v) - 2F_X(u)F_Y(v)dudv.$$

Fréchet bounds

Fréchet bounds $\max\{F_X(x) + F_Y(y) - 1, 0\} \leq F_{X,Y}(x,y) \leq \min\{F_X(x), F_Y(y)\}$ を示す. まず上限を示す.

$$F_{X,Y}(x,y) \leq P(X\leq x, Y\leq y) \leq \min\{P(X\leq x, Y\leq y)\} = \min\{F_X(x), F_Y(y)\}.$$

次に下限を示す. $F_X(x)+F_Y(y)-1 \leq F_{X,Y}(x,y)$ を示せば明らか.

$$\begin{aligned} F_X(x)+F_Y(y)-1 &= P(X\leq x)+P(Y\leq y)-1 \\ &= P(X\leq x)-P(Y>y) \\ &\leq P(X\leq x)-P(X\leq x, Y>y) \\ &= P(X\leq x, Y\leq y)=F_{X,Y}(x,y). \end{aligned}$$

(4) **の証明**

$\rho_{R,G}=\mathrm{Cov}(R_J,G_J)/(\sigma_R\sigma_G), \sigma_R=\sqrt{f-f^2}, \sigma_G=\sqrt{p_G-p_G^2}$ であることに注意する.
また、R_J, G_J が $0,1$ の値をとるものなので、積分範囲は 0 から 1 で考える. $\int F_{R_J}(x)dx = P_J(R_J=0) = 1-f, \int F_{G_J}(y)dy = 1-p_G$ に注意して計算すると

$$\begin{aligned} \mathrm{Cov}(R_J,G_J) &= \iint \{F_{R_J,G_J}(x,y)-F_{R_J}(x)F_{G_J}(y)\}dxdy \\ &\leq \iint \min\{F_{R_J}(x)(1-F_{G_J}(y)), F_{G_J}(y)(1-F_{R_J}(x))\}dxdy \\ &\leq \min\left\{\iint F_{R_J}(x)(1-F_{G_J}(y))dxdy, \iint F_{G_J}(y)(1-F_{R_J}(x))dxdy\right\} \\ &= \min\{(1-f)p_G, (1-p_G)f\}. \end{aligned}$$

よって、

$$\rho_{R,G} \leq \min\left\{\sqrt{\frac{1-f}{f}\frac{p_G}{1-p_G}}, \sqrt{\frac{1-p_G}{p_G}\frac{f}{1-f}}\right\} = \min\left\{\sqrt{O_G D_O}, \frac{1}{\sqrt{O_G D_O}}\right\}.$$

この等号は $R_J=G_J$ のとき成立する.
次に下限を示す.

$$\begin{aligned} \mathrm{Cov}(R_J,G_J) &= \iint \{F_{R_J,G_J}(x,y)-F_{R_J}(x)F_{G_J}(y)\}dxdy \\ &\geq \iint [\max\{F_{R_J}(x)+R_{G_J}-1, 0\}-F_{R_J}(x)F_{G_J}(y)]dxdy \\ &= \iint \max\{-(1-F_{R_J}(x))(1-F_{G_J}(y)), -F_{R_J}(x)F_{G_J}(y)\}dxdy \\ &\geq \max\left\{\iint -(1-F_{R_J}(x))(1-F_{G_J}(y))dxdy, \iint -F_{R_J}(x)F_{G_J}(y)dxdy\right\} \\ &= \max\{-fp_G, -(1-f)(1-p_G)\} \\ &= -\min\{fp_G, (1-f)(1-p_G)\} \end{aligned}$$

よって、

$$\rho_{R,G} \geq -\min\left\{\sqrt{\frac{f}{1-f}\frac{p_G}{1-p_G}}, \sqrt{\frac{1-f}{f}\frac{1-p_G}{p_G}}\right\} = -\min\left\{\sqrt{\frac{O_G}{D_O}}, \sqrt{\frac{D_O}{O_G}}\right\}.$$

この等号は $R_J=1-G_J$ のとき成立する.

参考文献

Meng, X.-L. (2018) Statistical paradises and paradoxes in big data (I) Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics* **12**, 2, 685-726.

階層ベイズロジットモデルと異質な消費者行動

趙宇

1　はじめに

複数のブランドが競合する市場において、消費者が特定のブランドを選択せざるを得ない状況に直面することは珍しくない。この選択問題は、個人の選択結果を目的変数とし、選択行動に影響を与えると考えられる様々なマーケティング変数（例えば価格やプロモーションなど）を説明変数とするブランド選択モデルを用いて定式化できる。マーケティングにおいては、ロジットモデルやプロビットモデルなどのブランド選択モデルが多く用いられる。これらのモデルは、選択の結果が離散的な変数で表されるため、離散選択モデルとも呼ばれる。また、モデルのパラメータが個人に共通する値を持ち、推定法としてはロジットモデルでは最尤法が、プロビットモデルでは（数値解析を回避するため）MCMC を利用されることが一般的である。

一方、効果的なマーケティング施策を展開するためには、すべての消費者に一様にアプローチするだけでなく、消費者個々の反応の異質性を理解することが重要である。例えば、値引きに対する消費者の反応は個人差があり、価格に敏感な人ほど値引きに強く反応すると容易に想像できる。消費者ごとに異なるパラメータを持つブランド選択行動をモデル化する際には、ベイズの定理を用いて消費者の異質性を表現できる。ここで、個人 i に関するブランド選択の結果を $\boldsymbol{y}_i$、パラメータを $\boldsymbol{\beta}_i$ とし、個人 i に依存しないパラメータを $\boldsymbol{\theta}$ とすると、ベイズの定理により以下の関係式が成り立つ。

$$p(\boldsymbol{\beta}_i, \boldsymbol{\theta}|\boldsymbol{y}_i) = \frac{p(\boldsymbol{y}_i|\boldsymbol{\beta}_i, \boldsymbol{\theta})p(\boldsymbol{\beta}_i, \boldsymbol{\theta})}{p(\boldsymbol{y}_i)}$$

共通性を表す $\boldsymbol{\theta}$ を仮定する理由は、後述の階層ベイズロジットモデルで詳しく見られるように、異質性を推定するのに不足する情報をデータが持つ共通情報で補うことができるからである。上記の式において、分母の $p(\boldsymbol{y}_i)$ は $\boldsymbol{\beta}_i, \boldsymbol{\theta}$ に依存しないため、定数として扱うことができる。したがって、

$$p(\boldsymbol{\beta}_i, \boldsymbol{\theta}|\boldsymbol{y}_i) \propto p(\boldsymbol{y}_i|\boldsymbol{\beta}_i, \boldsymbol{\theta}) \times p(\boldsymbol{\beta}_i|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) \tag{1}$$

となる。ここで、$p(\boldsymbol{\beta}_i|\boldsymbol{\theta}), p(\boldsymbol{\theta})$ はそれぞれ $\boldsymbol{\beta}_i$ と $\boldsymbol{\theta}$ の事前分布であり、$p(\boldsymbol{\beta}_i|\boldsymbol{\theta})$ は $\boldsymbol{\theta}$ に対する尤度関数としても解釈できる。$p(\boldsymbol{y}_i|\boldsymbol{\beta}_i, \boldsymbol{\theta})$ はパラメータ $\boldsymbol{\beta}_i$ に対する尤度関数である。式 (1) では、パラメータ $\boldsymbol{\beta}_i, \boldsymbol{\theta}$ の推定はすべて事後分布である $p(\boldsymbol{\beta}_i, \boldsymbol{\theta}|\boldsymbol{y}_i)$ に基づいて行われる。この推定において問題となるのは、明らかに事前分布の設定方法である。階層ベイズロジットモデルにおける事前分布の設定については第 4 章で詳しく説明する。

本稿では、はじめに消費者のブランド選択行動をモデル化する際に広く用いられるロジットモデルと、その拡張形である階層ベイズロジットモデルについて解説する。次に、階層ベイズロジットモデルを推定する手法であるマルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo: MCMC）について、ギブス（Gibbs）サンプリングとメトロポリス・ヘイスティングス（Metropolis-Hastings: MH）サンプリングの基本的な挙動をシミュレーション実験を通じて確認する。最後に、化粧品の購買履歴データを用いた事例分析について説明する。

2　ブランド選択モデル：ロジットモデル

この章では、マーケティングでよく使われる離散選択モデルの一つである、ロジットモデルの概要を整理する。

いま、U_{itj} を消費者 i の時点 t におけるブランド j に対する効用とし、効用関数 [1] を

$$U_{itj} := V_{itj} + \varepsilon_{itj} \tag{2}$$

と定義する。ここで、V_{itj} は効用の確定部分であり、ε_{itj} は不確実性を表す確率変数である。消費者 i が時点 t で選択可能なブランドの集合を J とする。効用最大化の原理によると、消費者 i が時点 t でブランド j を選択する確率 p_{itj} は、

$$p_{itj} = P(y_{it} = j) = P\left(V_{itj} + \varepsilon_{itj} > \max_{j \neq k; j,k \in J}(V_{itk} + \varepsilon_{itk})\right) \tag{3}$$

になる。ただし、確率の公理により、$0 \leq p_{itj} \leq 1, \forall j \in J, \sum_{j \in J} p_{itj} = 1$ を満たす。この誤差項 ε_{itj} の確率分布を Gumbel 分布と仮定した場合、(3) 式からロジットモデルが導出できる。まず、Gumbel 分布の確率密度関数は以下の式で与えられる。

$$f(\varepsilon_{itj}|\omega, \eta) := \omega \exp(-\omega(\varepsilon_{itj} - \eta)) \exp(-\exp(-\omega(\varepsilon_{itj} - \eta)))$$

ここで、ω は尺度母数、η は位置母数と呼ばれる。さらに、$\max_{j \neq k; j,k \in J}(V_{itk} + \varepsilon_{itk}) := V^*_{it} + \max_{j \neq k; j,k \in J}(\varepsilon_{itk})$ とおき、ε_{itj} が時点 t に関して独立に同一の $\omega = 1, \eta = 0$ のパラメータ値をもつ Gumbel 分布に従うものとすると、$V_{itk} + \varepsilon_{itk}$ は $\omega = 1, \eta = V_{itk}$ の Gumbel 分布に従い、$\max_{j \neq k; j,k \in J}(V_{itk} + \varepsilon_{itk})$ は $\omega = 1, \eta = \ln \sum_{j \neq k; j,k \in J} \exp(V_{itk})$ の Gumbel 分布に従う。したがって、$V^*_{it} = \ln \sum_{j \neq k; j,k \in J} \exp(V_{itk})$ になる。Gumbel 分布の性質より

$$P\left(V_{itj} + \varepsilon_{itj} > \max_{j \neq k; j,k \in J}(V_{itk} + \varepsilon_{itk})\right) = P\left(\max_{j \neq k; j,k \in J}(\varepsilon_{itk}) - \varepsilon_{itj} < V_{itj} - V^*_{it}\right)$$
$$= \frac{1}{1 + \exp\left(-V_{itj} + \ln \sum_{j \neq k; j,k \in J} \exp(V_{itk})\right)} = \frac{\exp(V_{itj})}{\sum_{k \in J} \exp(V_{itk})}$$

が導かれる。これをロジットモデルと呼ぶ。特に、選択可能なブランド数 $|J| = 2$ の場合のロジットモデルを二項ロジットモデル、$|J| \geq 3$ の場合のロジットモデルを多項ロジットと呼ぶ、

V_{itj} に関して線形性を仮定すると、

$$U_{itj} = \sum_{h=0}^{p} \beta_h x^h_{itj} + \varepsilon_{itj} \tag{4}$$

と書ける。ただし、$x^0_{itj} := 1, x^1_{itj}, \ldots, x^p_{itj}$ は切片項と説明変数である。企業のマーケティング活動が短期的であるとすれば、β_0 は短期的なマーケティング活動と無関係な独立の長期的なベースライン部分（すなわち、ブランド価値)、$\beta_1, \ldots, \beta_p$ は短期的効果と解釈できる。また、この長期的な特性 β_0 をブランドごとに固有な価値として考えることもできる。この場合、β_0 を β_{0i} に修正すると良い。

ここで、２つのブランド j, k が選択される相対的な確率を考えよう。

$$\frac{p_{itj}}{p_{itk}} = \frac{P(y_{it} = j|\beta_h, x_{it})}{P(y_{it} = k|\beta_h, x_{it})} = \frac{\exp(\sum_{h=0}^{p} \beta_h x^h_{itj})}{\exp(\sum_{h=0}^{p} \beta_h x^h_{itk})} = \exp\left(\sum_{h=0}^{p} \beta_h (x^h_{itj} - x^h_{itk})\right)$$

つまり、ロジットモデルには、候補となるブランドの集合 J が変化しても 2 つのブランドの選択比率の比は変わらないという問題がある (IIA: independence from irrelevant alternatives ともいう)。IIA を回避する方法としてネステッドロジットモデルなどがある。詳細については文献 [7] を参照されたい。

3　階層ベイズロジットモデル

第 2 章のロジットモデル (2) を階層ベイズロジットモデルに拡張するには、まず、式 (4) を以下のように修正する。

$$U_{itj} = \sum_{h=0}^{p} \beta_{hi} x^h_{itj} + \varepsilon_{itj} \tag{5}$$

このとき、$\boldsymbol{\beta}_i := (\beta_{0i}, \beta_{1i}, \ldots, \beta_{pi})^\top$ は消費者ごとのパラメータである。式 (5) は、個体内モデル（within-subject）とも呼ばれる。異質性を表す $\boldsymbol{\beta}_i$ を推定する際には、消費者の共通性の情報を加味する。$\boldsymbol{z}_i := (z_{1i}, \ldots, z_{li})^\top$ を顧客の属性（年齢、性別、その他の購買行動特性など）を表すベクトルとすると、

$$\beta_{hi} = \boldsymbol{z}_i^\top \boldsymbol{\theta}_h + e_{hi}, \quad h = 0, \ldots, p \tag{6}$$

という個体間の関係を表現する回帰モデルを構成できる。これを個体間モデル（between-subject）と呼ぶ。ここで、$\boldsymbol{\theta}_h := (\theta_{1h}, \ldots, \theta_{lh})$ は変数 h に対する回帰係数ベクトルであり、e_{hi} は $Cov(e_{hi}, e_{mi}) = v_{hm} \neq 0$ を満たす回帰の誤差項である。モデル (6) は以下の行列表現と等価である。

$$\boldsymbol{\beta}_i = \Theta^\top \boldsymbol{z}_i + \boldsymbol{e}_i, \quad \boldsymbol{e}_i \sim N_{p+1}(\boldsymbol{0}, \boldsymbol{V}_\beta) \tag{7}$$

ただし、$\Theta := [\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p]$ は $l \times (p+1)$ とし、$\boldsymbol{e}_i := (e_{0i}, e_{1i}, \ldots, e_{pi})^\top$ とする。記号 $\boldsymbol{V}_\beta$ は多変量正規分布の分散共分散行列を表す。モデル (7) の仮定より、$\boldsymbol{\beta}_i$ に対する事前分布として

$$\boldsymbol{\beta}_i \sim N_{p+1}\left(\Theta^\top \boldsymbol{z}_i, \boldsymbol{V}_\beta\right) \tag{8}$$

となる。ちなみに、モデル (7) は消費者 i の回帰方程式であるが、これを N 人分まとめて表現すると、

$$B_{N\times(p+1)} = Z_{N\times l}\Theta_{l\times(p+1)} + E_{N\times(p+1)} \tag{9}$$

が得られる。つまり、モデル (6) は多変量回帰の形に帰着できる。

パラメータの推定値 $\hat{\boldsymbol{\beta}}_i$ が得られたら、消費者 i が時点 t でブランド j を選択する確率 p_{itj} は、

$$\hat{p}_{itj} = \frac{\exp(\hat{V}_{itj})}{\sum_{k\in J}\exp(\hat{V}_{itk})} = \frac{\exp(\sum_{h=0}^{p}\hat{\beta}_{hi}x_{itj}^h)}{\sum_{k\in J}\exp(\sum_{h=0}^{p}\hat{\beta}_{hi}x_{itk}^h)} \tag{10}$$

と計算できる。

4　階層ベイズロジットモデルの推定

式 (1) に示すように、事後分布を用いて推測を行うときに、$p(\boldsymbol{y}_i|\boldsymbol{\beta}_i, \boldsymbol{\theta}), p(\boldsymbol{\beta}_i|\boldsymbol{\theta}), p(\boldsymbol{\theta})$ を定式化しなければならない。$p(\boldsymbol{y}_i|\boldsymbol{\beta}_i, \boldsymbol{\theta})$ は第 3 章で示した個体内モデルが対応する。$p(\boldsymbol{\beta}_i|\boldsymbol{\theta})$ は個体間モデルが対応する。個体間モデルには、分散共分散パラメータ $\boldsymbol{V}_\beta$ がある。これの事前分布として、以下のように設定する。

$$\boldsymbol{V}_\beta \sim \text{inverse-Wishart}(\nu_0, \boldsymbol{V}_0) \tag{11}$$

この仮定は、inverse-Wishart 分布が分散共分散行列の準共役事前分布であることを利用している。inverse-Wishart 分布のパラメータについては、たとえば R の bayesm パッケージではデフォルトの設定として $\nu_0 = (p+1)+3, \boldsymbol{V}_0 = \nu_0 I_{p+1}$ と設定する（この設定は、[4] の第 5 章の説明に準じる）。

$p(\boldsymbol{\theta})$ については、パラメータ $\boldsymbol{\theta}$ の事前分布として

$$\text{vec}(\Theta|\boldsymbol{V}_\beta) \sim N(\text{vec}(\bar{\Theta}), A^{-1} \otimes \boldsymbol{V}_\beta) \tag{12}$$

が設定できる。ここで、vec は行列の列を縦につなげてベクトル化することを示す記号である。$\otimes$ はクロネッカー積を示す記号である。式 (12) のパラメータについては、R の bayesm パッケージのデフォルトの設定として $\bar{\Theta} = \boldsymbol{0}, A = 0.01 I_l$ を利用する。ただし、$A = 0.01 I_l$ という設定は多くの事例においては小さすぎる傾向が見られ、必ずしも適切な選択とは限らない [4]。

これらの事前分布の下で、第 3 章で示した階層ベイズロジットモデルを推定するアルゴリズムを構築できる。

まず、事後分布は次のように表現できる。

$$
\begin{aligned}
&p(\{\boldsymbol{\beta}_i\}, \Theta, \boldsymbol{V}_\beta | \{\boldsymbol{y}_{it}\}, \{\boldsymbol{z}_i\}, \{\boldsymbol{x}_{it}\}) \\
&\propto p(\Theta|\boldsymbol{V}_\beta) p(\boldsymbol{V}_\beta) \prod_{i=1}^{N} p(\boldsymbol{\beta}_i|\Theta, \boldsymbol{V}_\beta, \boldsymbol{z}_i) \prod_{t=1}^{T_i} p(\boldsymbol{y}_{it}|\boldsymbol{\beta}_i, \boldsymbol{x}_{it})
\end{aligned} \tag{13}
$$

次に、マルコフ連鎖モンテカルロ法（MCMC）を用いてパラメータの推定アルゴリズムを導出する。

(1) パラメータ $\{\boldsymbol{\beta}_i\}$ の条件付き事後分布は、式 (13) より

$$
p(\boldsymbol{\beta}_i|\Theta, \boldsymbol{V}_\beta, \boldsymbol{z}_i, \{\boldsymbol{y}_{it}\}, \{\boldsymbol{x}_{it}\}) \propto p(\boldsymbol{\beta}_i|\Theta, \boldsymbol{V}_\beta, \boldsymbol{z}_i) \prod_{t=1}^{T_i} p(\boldsymbol{y}_{it}|\boldsymbol{\beta}_i, \boldsymbol{x}_{it}), \quad \forall i \tag{14}
$$

となる。この関係が共役とならないため、メトロポリス・ヘイスティングス法（M-H 法）を用いてサンプリングする。この方法は、基本的に事後確率が大きくなる方向にマルコフ連鎖が動いていくが、ある確率（採択確率もしくは受理確率）で事後確率が小さくなるほうへも動く。このため、パラメータの空間全体を動き回ることができるようになる。M-H 法の特徴としては、パラメトリックなモデルであればどのような統計モデルでも適用可能な点が挙げられる。ロジットモデルに応用される M-H 法の代表的なアルゴリズムとして、ランダムウォーク・サンプラーがある（Appendix A を参照されたい）。具体的には、$\boldsymbol{\beta}_i$ を発生するために以下のアルゴリズムを用いればよい。

1. $\boldsymbol{\beta}_i^{(0)}$ を初期値とし、ランダムに設定する
2. $\boldsymbol{v} \sim N_{p+1}(0, \sigma_{\boldsymbol{v}}^2 I_{p+1})$ を (行ベクトルとして) 生成し、$\boldsymbol{\beta}_i^* = \boldsymbol{\beta}_i^{(r-1)} + \boldsymbol{v}^\top, r \geq 1$ とする
3. $u \sim \text{Uniform}[0,1]$ を生成し、以下のように確率的選択を行う

$$
\boldsymbol{\beta}_i^{(r)} = \begin{cases} \boldsymbol{\beta}_i^*, & u \leq \alpha\left(\boldsymbol{\beta}_i^{(r-1)}, \boldsymbol{\beta}_i^*\right) := \min\left\{1, \frac{p(\boldsymbol{\beta}_i^*|\Theta, \boldsymbol{V}_\beta, \boldsymbol{z}_i, \{\boldsymbol{y}_{it}\}, \{\boldsymbol{x}_{it}\})}{p(\boldsymbol{\beta}_i^{(r-1)}|\Theta, \boldsymbol{V}_\beta, \boldsymbol{z}_i, \{\boldsymbol{y}_{it}\}, \{\boldsymbol{x}_{it}\})}\right\} \\ 2 \text{ へ戻る}, & \text{それ以外の場合} \end{cases}
$$

(2) パラメータ Θ の条件付き事後分布は、式 (13) より

$$
p(\Theta|\{\boldsymbol{\beta}_i\}, \boldsymbol{V}_\beta, \{\boldsymbol{z}_i\}) \propto p(\Theta|\boldsymbol{V}_\beta) \prod_{i=1}^{N} p(\boldsymbol{\beta}_i|\Theta, \boldsymbol{V}_\beta, \boldsymbol{z}_i) \tag{15}
$$

となる。これは、多変量回帰モデル (9) の回帰係数行列をベイズ推測する問題として解釈できる。回帰係数行列の条件付き事後分布の導出はやや煩雑なのでここでは省く（たとえば [3] を参照されたい）。以下は結果だけを示す。

$$
\text{vec}(\Theta) \sim N\left(\tilde{\boldsymbol{\delta}}, \boldsymbol{V}_\beta \otimes (Z^\top Z + A)^{-1}\right) \tag{16}
$$

ここで、$\tilde{\boldsymbol{\delta}} = \text{vec}(\tilde{D}), \tilde{D} = (Z^\top Z + A)^{-1}(Z^\top Z\hat{D} + A\bar{D}), \hat{D} = (Z^\top Z)^{-1}Z^\top B, \bar{D} = \mathbf{0}$ である。

(3) パラメータ $\boldsymbol{V}_\beta$ の条件付き事後分布は、式 (13) より

$$
p(\boldsymbol{V}_\beta|\{\boldsymbol{\beta}_i\}, \Theta, \{\boldsymbol{z}_i\}) \propto p(\boldsymbol{V}_\beta) \prod_{i=1}^{N} p(\boldsymbol{\beta}_i|\Theta, \boldsymbol{V}_\beta, \boldsymbol{z}_i) \tag{17}
$$

が得られる。これは、多変量回帰モデル (9) の分散共分散行列に対するベイズ推測である。前項と同じ理由で、以下は結果だけを示す。

$$
\boldsymbol{V}_\beta \sim \text{inverse-Wishart}(\nu_0 + N, \boldsymbol{V}_0 + S^\top) \tag{18}
$$

ここで、$S^\top = \sum_{i=1}^{N}(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_i)(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_i)^\top, \bar{\boldsymbol{\beta}}_i = \Theta^\top \boldsymbol{z}_i$ である。

また、式 (15) も式 (17) も共役になるのでギブスサンプラー（Appendix B を参照されたい）を用いてサンプリングできる。

5　シミュレーション

MCMC による階層ベイズロジットモデルの推定に関して、以下の設定に基づく数値実験を行った。

$$P(y_{it}=j)=\frac{\exp\left(\sum_{h=1}^{3}\beta_{hi}x_{ith}^{j}\right)}{\sum_{k=1}^{2}\exp\left(\sum_{h=1}^{3}\beta_{hi}x_{ith}^{k}\right)},\quad i=1,\ldots,300;\quad t=1,\ldots,10;\quad j=1,2$$

$$\boldsymbol{z}_i,\boldsymbol{x}_{it}\sim\text{Uniform}(0,1),\ \forall i,t$$

$$\boldsymbol{\beta}_i\sim N(\Theta^{\top}\boldsymbol{z}_i,\boldsymbol{V}_{\beta})+\boldsymbol{\epsilon}_i,\ \forall i$$

$$\Theta\sim N(0,I_3)$$

$$\boldsymbol{V}_{\beta}\sim\text{inverse-Wishart}(7,I_3)$$

$$\boldsymbol{\epsilon}_i\sim 0.6N(-3,1)+0.4N(1.5,1.2),\ \forall i$$

このシミュレーション設定では、混合正規分布を使用して、個々の顧客の選好に異質性を導入している。また、データの生成において階層構造を持たせている。事後分布の推定には R の bayesm パッケージを利用している。

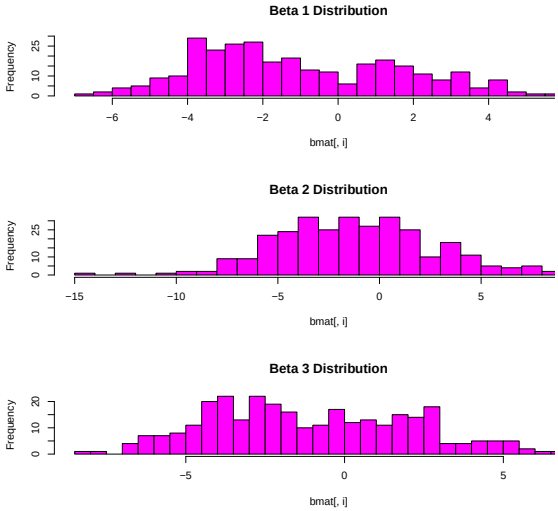

図 1: $\boldsymbol{\beta}_i$ の事前分布により生成した値の分布

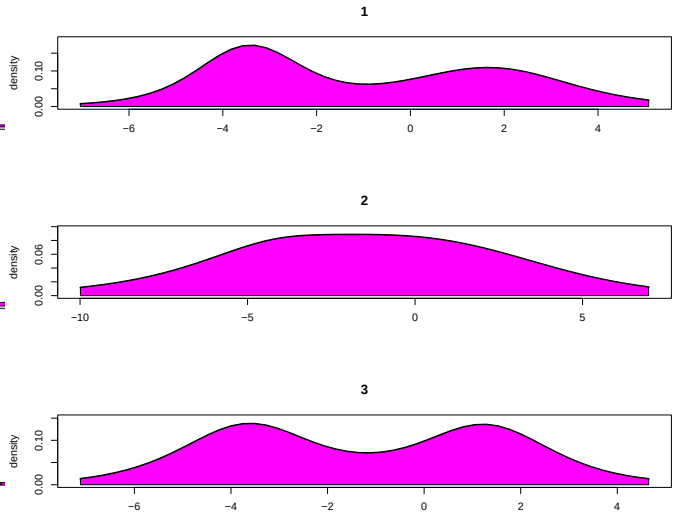

図 2: 個人ごと事後平均の分布（密度）

図 1 は $\boldsymbol{\beta}_i$ の事前分布から生成された値のヒストグラムである。図 2 では、$\boldsymbol{\beta}_1$ と $\boldsymbol{\beta}_3$ の事後分布が双峰性を示しており、これは異質性を捉えていることを示唆している。ここで、$\boldsymbol{\beta}_1$ から $\boldsymbol{\beta}_3$ までの事後平均の分布は、シミュレーションによる値の分布と類似していることが確認できる。$\boldsymbol{\beta}_i$ のリサンプリングの結果は図 3 で示されている。サンプリングされた値の自己相関では、良い収束の兆候が見られる。また、対数尤度の変化は、図 4 で確認できる。

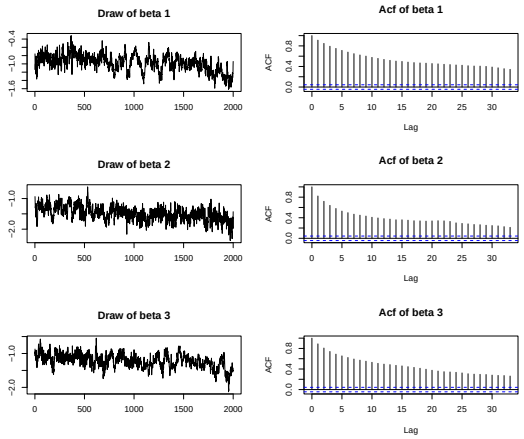

図 3: $\boldsymbol{\beta}_i$ のリサンプリング結果

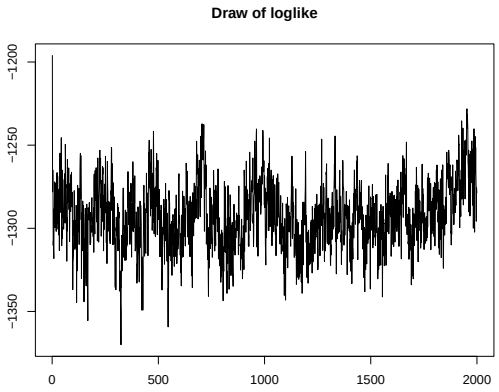

図 4: 対数尤度の変化

図 5 と図 6 は Θ のサンプリングの結果である。自己相関では、良い収束の兆候が見られる。

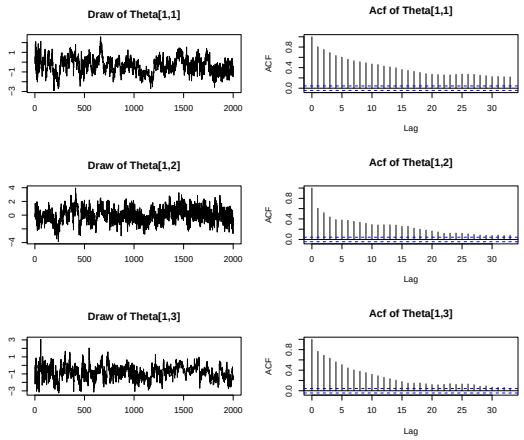

図 5: Θ のリサンプリング結果 1

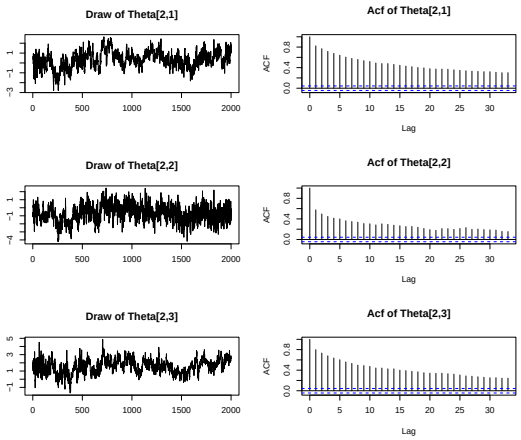

図 6: Θ のリサンプリング結果 2

6　事例：化粧品のブランド選択

この章では、「消費者購買履歴データ QPR」[1]から抽出された化粧品に関連するデータ（2021 年度）を用いてモデルの挙動を観察する。このデータセットには、消費者のデモグラフィック属性（性別、年代、未婚または既婚、家族構成、年収）や化粧品の購買履歴のデータが含まれている。ここでは簡便性を考慮し、大手メーカーである資生堂、花王、コーセーの化粧品の購買履歴データセットから利用可能な変数を主に使用し、それ以外のアンケート調査でしか得られないデータについては他の変数や統計量で代用することとする。各化粧品メーカーは異なるブランドラインを持っているが、ここでは単純化のため、化粧水を購入する消費者のみに焦点を当てる。実際に購入される化粧水は多種多様であるが、その違いは考慮しないことにする。つまり、分析の目的は、「化粧水というカテゴリーの商品を購入する消費者はどのブランドを選ぶのか」に絞る。また、通販や実店舗などの購入経路があるが、ここでは Amazon.com を利用する顧客に限定する。

階層ベイズロジットモデルの推定に用いる変数を下表にまとめる。

表 1. 変数の定義

[1] https://www.macromill.com/service/digital-data/consumer-purchase-history-data/

変数	定義
y_{it}	消費者 i の時点 t での選択
x^1_{itj}	消費者 i の時点 t のブランド j の価格
x^2_{itj}	消費者 i の時点 t のブランド j の購入時間帯（夜 1、昼 0）
x^3_{itj}	消費者 i の時点 t のブランド j の購入曜日（土日 1、平日 0）
x^4_{ijt}	消費者 i の時点 t のブランド j の配達エリア（首都圏 1、それ以外 0）
z^1_i	消費者 i の年齢
z^2_i	消費者 i の個人収入
z^3_i	消費者 i の家族人数

図 7 は $\{\boldsymbol{y}_{it}\}$ と $\{\boldsymbol{x}_{it}\}$ を格納する「xdata.csv」ファイルの構造を示している。ID という変数は、消費者を識別するための固有の番号である。「brand」という変数はブランドの選択結果に対応している。残りの変数は各ブランドの $\{\boldsymbol{x}_{it}\}$ に対応している。このファイルとは別に、$\{\boldsymbol{z}_{it}\}$ を格納する「zdata.csv」ファイルも用意されている。これら 2 つのファイルのデータの次元は異なるため、分けて管理するのが便利である。

P9

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	brand	PriceSh	PriceKa	PriceKo	TimeSh	TimeKa	TimeKo	WeekSh	WeekKa	WeekKo	AreaSh	AreaKa	AreaKo
2	3588087	1	2350	767	794	0	0	0	0	1	0	0	0	0
3	13121789	2	2000	793.5	794	1	0	0	1	0	0	0	0	0
4	7959228	1	2350	1980	794	0	0	0	0	1	0	0	0	0
5	3315322	1	2350	592	794	0	1	0	0	1	0	0	0	0
6	4162745	1	2350	780	794	0	1	0	0	1	0	0	1	0
7	3094151	2	1206	793.5	794	1	0	0	1	0	0	0	0	0
8	19268468	3	2350	793.5	597	0	0	0	0	0	1	0	0	1
9	3082414	3	2350	793.5	626	0	0	1	0	0	0	0	0	0
10	7959228	3	2350	793.5	1210	0	0	1	0	0	0	0	0	0
11	1204558	1	2350	1980	794	0	0	0	0	1	0	0	0	0
12	25192064	3	2350	793.5	798	0	0	1	0	0	1	0	0	0
13	26656544	2	1628	793.5	794	1	0	0	1	0	0	0	0	0
14	7165847	1	2350	2530	794	0	0	0	0	1	0	0	0	0
15	20823406	2	1166	793.5	794	1	0	0	1	0	0	0	0	0
16	15077563	3	2350	793.5	1155	0	0	0	0	0	1	0	0	0
17	8197246	1	2350	503	794	0	1	0	0	1	0	0	0	0

図 7: データファイル「xdata.csv」の構造

モデルは第 3 章で説明した階層ベイズロジットモデルを用いる。事前分布に関しては第 4 章で示した設定を用いる。モデル推定のための R コードを Appendix C に掲載する。

表 1 で示されている変数に基づいて、以下の 5 つのモデルを考えてモデルの選択を行う。

(M1) 式 (5) に対して価格、購入時間帯、購入曜日、配達エリアを説明変数として用いる

(M2) 式 (5) に対して価格、購入曜日、配達エリアを説明変数として用いる

(M3) 式 (5) に対して価格、購入時間帯、配達エリアを説明変数として用いる

(M4) 式 (5) に対して価格、購入時間帯、購入曜日を説明変数として用いる

(M5) 式 (5) に対して価格のみを説明変数として用いる

ベイジアンモデリングでは DIC（偏差情報量規準）を用いてモデル選択を行うことができる。DIC については例えば文献 [8] の補論が詳しい。ここでは詳細を省くが、一般には DIC の最も小さいモデルが良いモデルであると考えられる。DIC の評価結果を図 8 に示す。

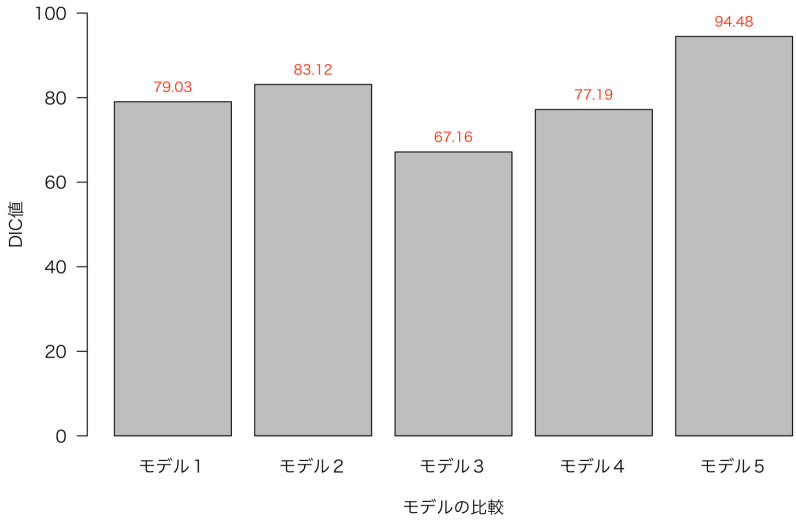

図 8: DIC の計算結果

すなわち、モデル (M3) が最もよいモデルと考えられる。以下の考察はモデル (M3) を用いて説明する。モデル (M3) においては、式 (5) に対応する表現は以下のとおりとなる。ここで、定数項のダミーについては、最初の二つのブランドに限って考慮するという点に注意したい。

$$
\begin{aligned}
U_{it1} &= \beta_{1i} x_{it1}^2 + \beta_{2i} x_{it1}^2 + \beta_{4i} x_{it1}^4 + \beta_{5i} + \varepsilon_{it1} \\
U_{it2} &= \beta_{1i} x_{it2}^2 + \beta_{2i} x_{it2}^2 + \beta_{4i} x_{it2}^4 + \beta_{6i} + \varepsilon_{it2} \\
U_{it3} &= \beta_{1i} x_{it3}^2 + \beta_{2i} x_{it3}^2 + \varepsilon_{it1}
\end{aligned}
$$

図 9 は、消費者ごとの事後分布の推定結果を示している。ここで、Coef1〜Coef5 はそれぞれ $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_4, \boldsymbol{\beta}_5, \boldsymbol{\beta}_6$ に対応している。パラメータを確率変数として考えることは、ベイジアンモデリングの大きな特徴と言える。最尤推定は、消費者の異質性を考慮していない。ここで、第 2 章で説明したロジットモデルを最尤推定し、その係数の点推定値とそれらの分布を比較すれば、異質性の程度を把握することができる。実際、モデル (4) に最尤法を適用したところ、$\hat{\beta}_1 = -0.36, \hat{\beta}_2 = 0.70, \hat{\beta}_4 = -11.71, \hat{\beta}_5 = 0.66, \hat{\beta}_6 = 0.26$ が得られた。これらの数値と係数の事後平均の分布を比較すると、価格と配達エリアに対する最尤推定値は分布の中心からかなり離れているため、異質性の程度が大きいと考えられる。一方、購入時間帯に対する最尤推定値は事後分布の中心付近に位置しているが、事後分布が双峰性を示しているため、異質性の程度が小さいとは考えにくい。

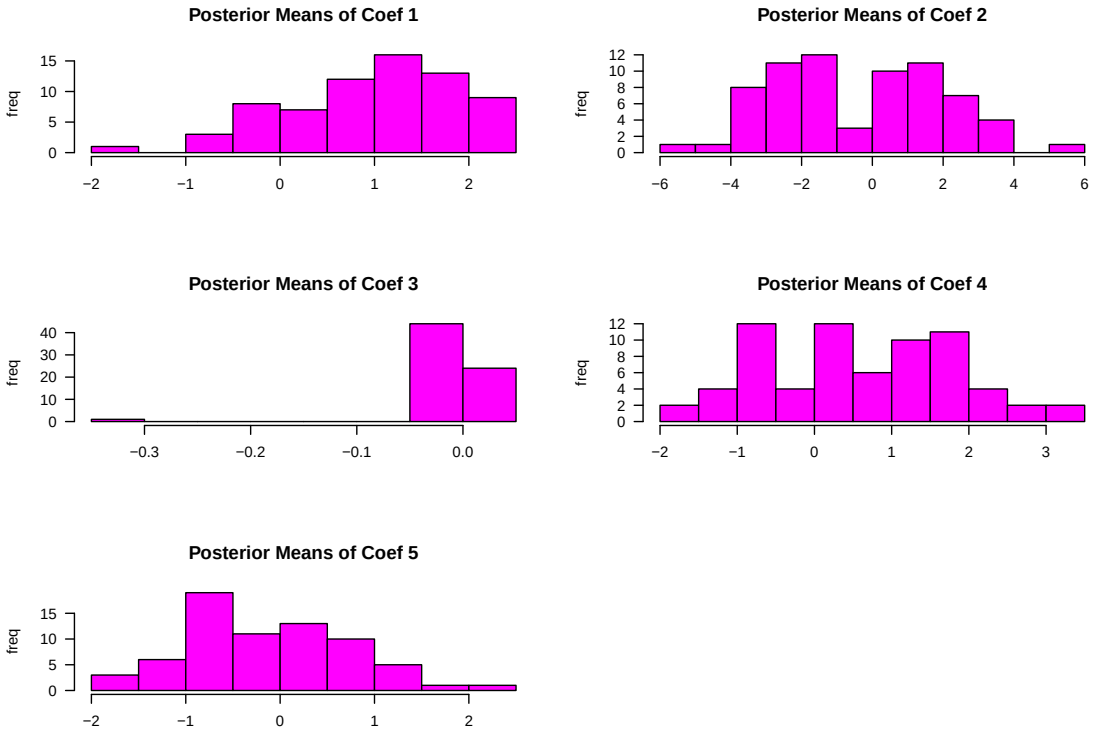

図 9: 消費者ごとの事後平均の分布

図 10 はさらに 76 名の消費者のそれぞれの係数の事後分布の箱ひげ図を示している。この図を用いて消費者ごとの考察ができる。

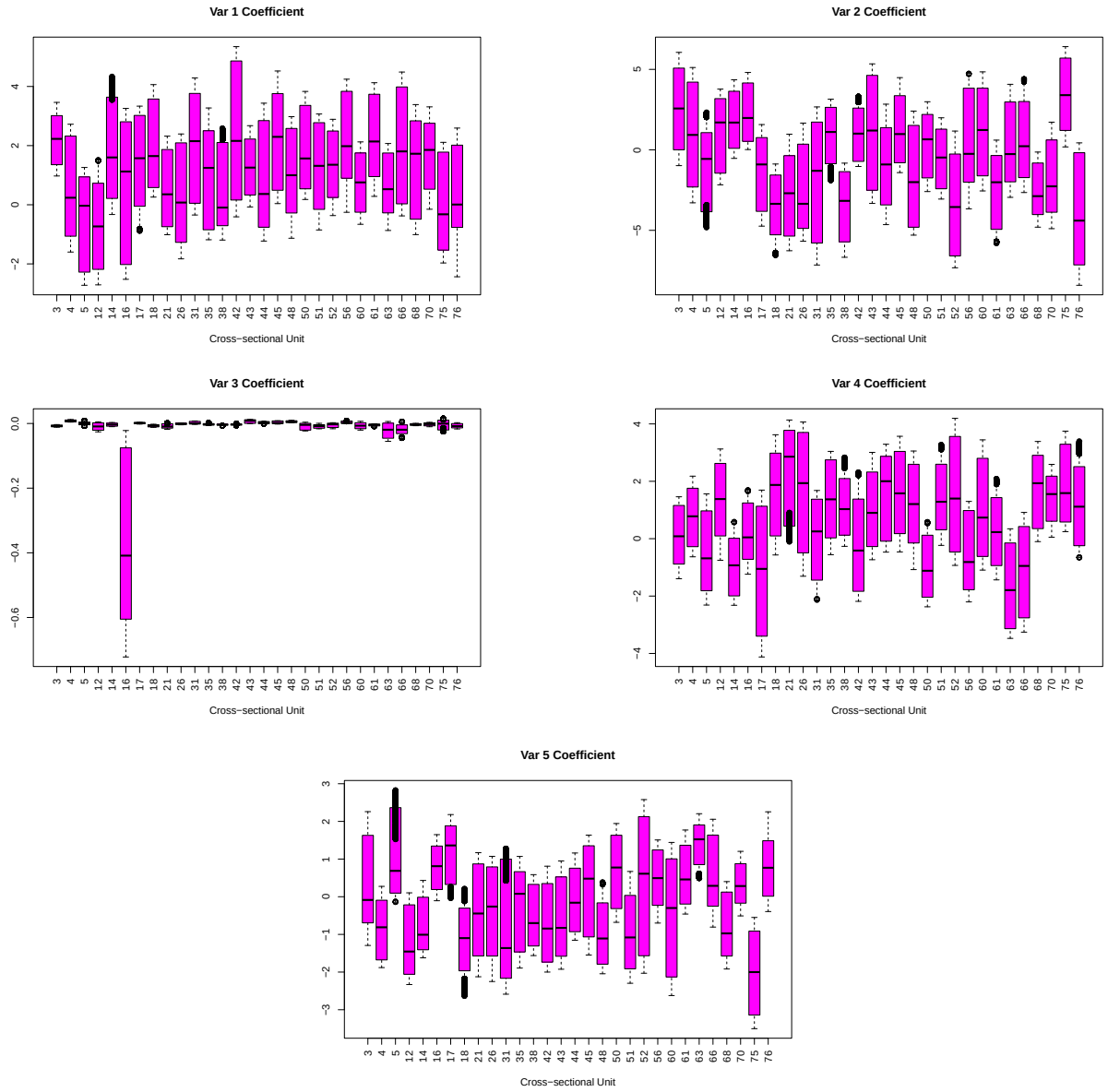

図 10: $\boldsymbol{\beta}_i$ の事後分布の箱ひげ図

最後に、各係数の有意性について確認する。ここでは文献 [7] で説明した疑似 t-値（事後平均/事後標準偏差）を用いて有意性を判定する。具体的には、$|t-$値$|>1.96$ ならば有意、そうでない場合は有意でないと判定する。この結果を表 2 に示す。この結果によると、どの係数も半数以上の消費者が有意な結果となっていることがわかる。

表 2. 事後平均の平均と有意な顧客数

パラメータ	事後平均の平均	有意な人数
価格	1.2602	58
購入時間帯	-0.9544	52
配達エリア	0.0124	60
定数項 1	0.6451	53
定数項 2	-0.2156	51

7　おわりに

MCMC は、複雑な統計モデルの事後分布からのサンプリングに非常に強力な手法として知られている。実際のデータ分析にこの手法を適用する際には、収束性の検証とアルゴリズムの効率性評価が必要不可欠である。本稿で行われた数値実験では、サンプリングされた値の自己相関を視覚的に検証した。しかし、より確かな収束判断を行うためには、Gelman & Rubin の収束診断などの追加的な手法を適用する必要がある。また、事例分析においては、リサンプリング用のパラメータ設定の微調整などを行った。これらの分析を通じて、MCMC を使用して消費者のブランド選択行動をベイズモデルで柔軟にモデリングできることが示された。なお、本稿の作成にあたり、以下の文献を主に参照した：[1, 2, 3, 4, 5, 6, 7, 8]。

A　ランダムウォーク M-H 法の補足

M-H 法にランダムウォークを応用する発想はとても自然で、文献 [2] で紹介された例をもとに R でシミュレーションを行った。具体的には、区間 $[-\delta, \delta]$（ここで、$\delta = 0.05, 1, 15$）上の一様分布を提案分布として、ランダムウォークアルゴリズムにより 10000 個のサンプルを作成し、目標分布である $N(0,1)$ を生成した。下図の 1 行目はシミュレーションデータの分布と標準正規分布の確率密度関数のグラフを重ねたものである。2 行目は連鎖の最後の 1000 回の反復をプロットしたものである。3 行目は自己相関関数のプロットである。この例では、適切な δ の選択が重要であることが見て取れる。

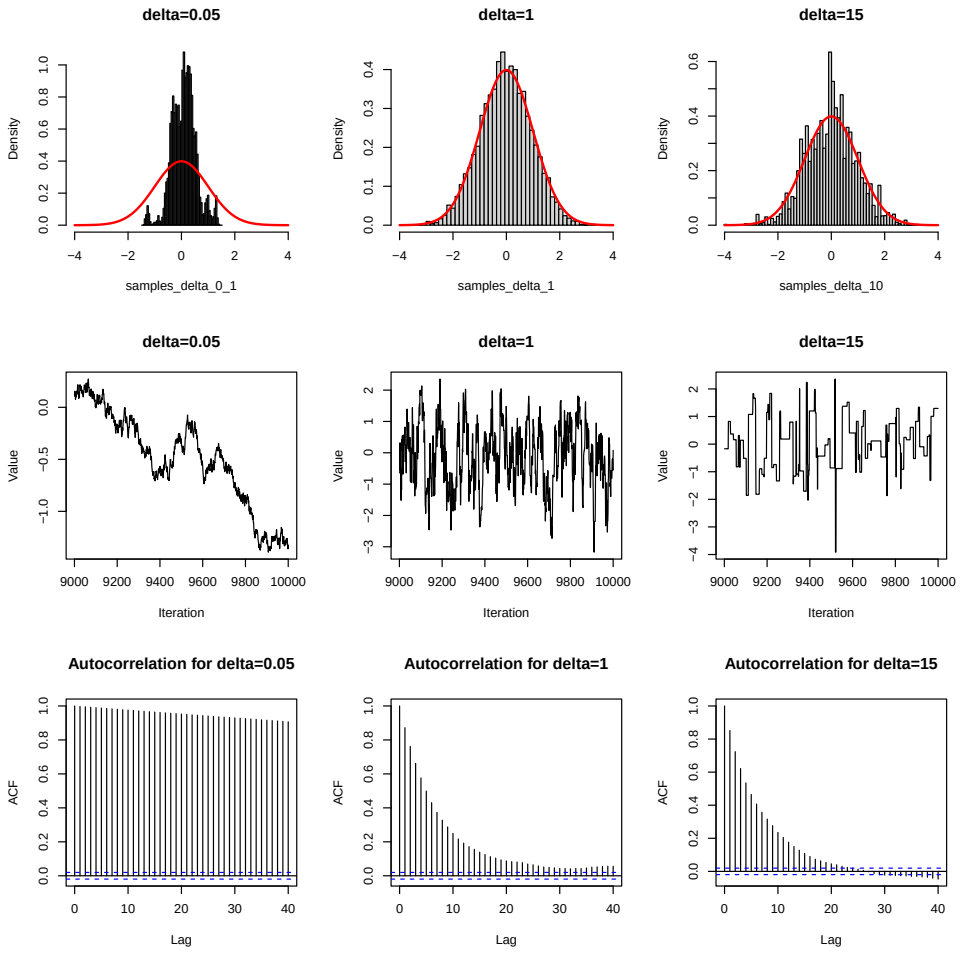

図 11: ランダムウォーク M-H 法。目標分布：標準正規分布、提案分布：一様分布

B　ギブスサンプラーの補足

ギブスサンプラーは MCMC の中で最もよく知られている方法と言っても過言ではない。分布が既知であればギブスサンプラーは簡単に実装できる。ここで、2 変量正規分布に基づいた簡単なシミュレーション例を用いてギブスサンプラーの挙動を確認しよう。

2 変量正規分布の確率密度関数は以下のように定義される。

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right] \right\}$$

ここで、μ_1, μ_2 は期待値、σ_1, σ_2 は標準偏差、ρ は相関係数を表す記号である。各変数の条件付き分布（完全条件付き事後分布ともいう）はそれぞれ

$$X_2|X_1 = x_1 \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1), (1-\rho^2)\sigma_2^2\right)$$
$$X_1|X_2 = x_2 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2), (1-\rho^2)\sigma_1^2\right)$$

と導かれる。ギブスサンプラーでは、$x_1^{(t)}$ に対して、

$$X_2^{(t+1)}|X_1^{(t+1)} = x_1^{(t)} \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1^{(t)}-\mu_1), (1-\rho^2)\sigma_2^2\right)$$
$$X_1^{(t+1)}|X_2^{(t+1)} = x_2^{(t+1)} \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2^{(t+1)}-\mu_2), (1-\rho^2)\sigma_1^2\right)$$

のようにサンプリングを行い、$x_2^{(t)}$ に対しても似たようなサンプリングを行う（上式にある記号の下付き添字について 1 と 2 を入れ替えればよい）。いま、初期値 $(x_1^{(0)}, x_2^{(0)}) = (-7, -7)$ と適当に指定して、上記のデータ生成過程において $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) = (1.1, 1.8, 3^2, 4^2, 0.6)$ と指定した上で、2 変量についてそれぞれ 1000 個のデータをサンプリングした。つまり、

$$\begin{aligned} x_1^{(0)} \to & X_2^{(1)}|X_1^{(1)} = x_1^{(0)} \to X_1^{(1)}|X_2^{(1)} = x_2^{(1)} \to \\ & X_2^{(2)}|X_1^{(2)} = x_1^{(1)} \to X_1^{(2)}|X_2^{(2)} = x_2^{(2)} \to \\ & \cdots \\ & X_2^{(1000)}|X_1^{(1000)} = x_1^{(999)} \to X_1^{(1000)}|X_2^{(1000)} = x_2^{(1000)} \end{aligned}$$

のプロセスで生成した 1000 個のデータと

$$\begin{aligned} x_2^{(0)} \to & X_1^{(1)}|X_2^{(1)} = x_2^{(0)} \to X_2^{(1)}|X_1^{(1)} = x_1^{(1)} \to \\ & X_1^{(2)}|X_2^{(2)} = x_2^{(1)} \to X_2^{(2)}|X_1^{(2)} = x_1^{(2)} \to \\ & \cdots \\ & X_1^{(1000)}|X_2^{(1000)} = x_2^{(999)} \to X_2^{(1000)}|X_1^{(1000)} = x_1^{(1000)} \end{aligned}$$

のプロセスで生成した 1000 個のデータである。これらのデータ（緑色の点）をプロットしたのが Figure B.2 である。青い等高線は指定したバラメータの下での 2 変量正規分布を示している。黒い点（初期値）から出発する赤い線は、マルコフ連鎖の最初の 50 ステップを示している。

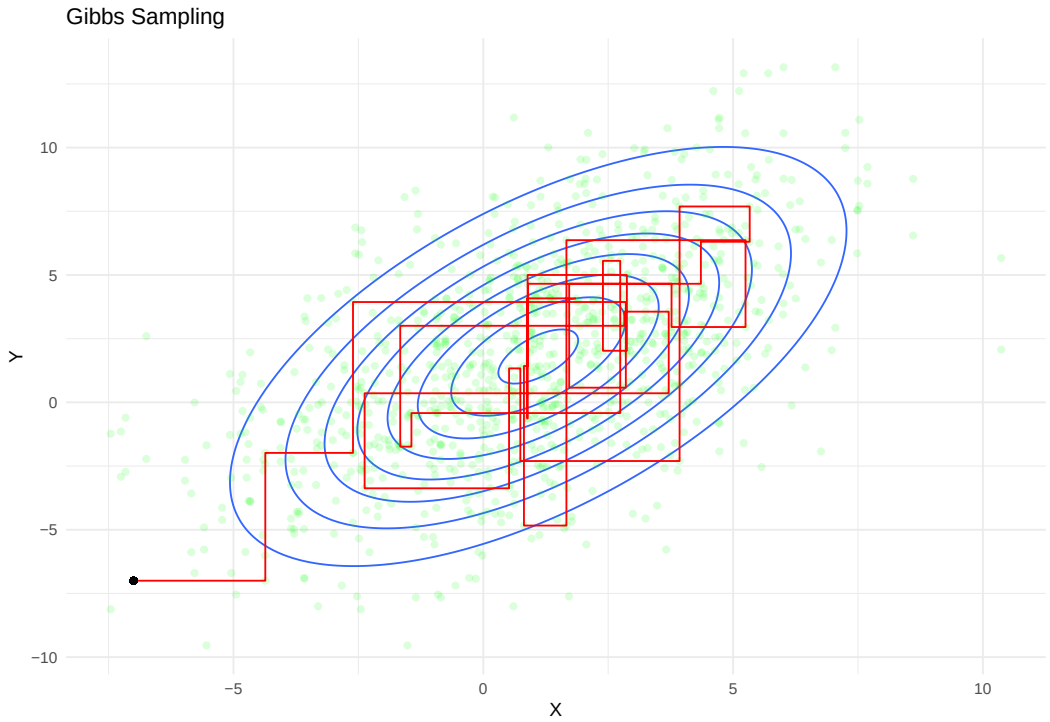

図 12: ギブスサンプラー。パラメータ：$(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) = (1.1, 1.8, 3^2, 4^2, 0.6)$、データ数：1000、初期値：$(x_1, x_2) = (-7, -7)$

ギブスサンプラーに基づき発生したデータの分布は正規分布の形と似ていることが確認できる。Shapiro-Wilk normality test の結果（p 値 $= 0.1304$）によりこのデータが 2 変量正規分布に従っていることが言える。また、生成したデータに基づいたパラメータの点推定値が $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\rho}) = (1.0, 1.7, 3^2, 4^2, 0.6)$ と計算され、パラメータの真値 $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) = (1.1, 1.8, 3^2, 4^2, 0.6)$ にかなり近い数値となっていることが確認できる。

C　R コード

図 11 の R コードを以下に示す。

```
# ランダムウォークメトロポリス・ヘイスティングス法
# 目標分布
target <- function(x) {
  return(dnorm(x, mean=0, sd=1))
}

# ランダムウォークMH法
rw_mh <- function(n, init, delta) {
  # 初期化
  x <- numeric(n)
  x[1] <- init

  for (i in 2:n) {
    # 提案分布からサンプリング
    y <- runif(1, min=x[i-1]-delta, max=x[i-1]+delta)
    # 受理確率の計算
    alpha <- min(1, target(y) / target(x[i-1]))
    # 受理判定
    if (runif(1) < alpha) {
      x[i] <- y # 受理
    } else {
      x[i] <- x[i-1] # 棄却
```

```
    }
  }
  return(x)
}


# サンプリング
set.seed(123)
nn = 10000
delta_005 <- rw_mh(nn, init=0, delta=0.05)
delta_1 <- rw_mh(nn, init=0, delta=1)
delta_15 <- rw_mh(nn, init=0, delta=15)

par(mfrow=c(3,3))
# ヒストグラム
library(KernSmooth)
h1 <- dpih(delta_005)
bins1 <- seq(min(delta_005)-0.1, max(delta_005)+0.1+h1, by=h1)
h2 <- dpih(delta_1)
bins2 <- seq(min(delta_1)-0.1, max(delta_1)+0.1+h2, by=h2)
h3 <- dpih(delta_15)
bins3 <- seq(min(delta_15)-0.1, max(delta_15)+0.1+h3, by=h3)

hist(delta_005, freq=FALSE, breaks=bins1,xlim=c(-4, 4), main="delta=0.05")
curve(dnorm(x, mean=0, sd=1), add=TRUE, col="red", lwd=2)
hist(delta_1, freq=FALSE, breaks=bins2,xlim=c(-4, 4), main="delta=1")
curve(dnorm(x, mean=0, sd=1), add=TRUE, col="red", lwd=2)
hist(delta_15, freq=FALSE, breaks=bins3,xlim=c(-4, 4), main="delta=15")
curve(dnorm(x, mean=0, sd=1), add=TRUE, col="red", lwd=2)

# 折れ線グラフ
plot((length(delta_005)-999):length(delta_005), delta_005[(length(delta_005)-999):length(delta_
    005)], type="l", main="delta=0.05", xlab="Iteration", ylab="Value")
plot((length(delta_1)-999):length(delta_1), delta_1[(length(delta_1)-999):length(delta_1)], type="l"
    , main="delta=1", xlab="Iteration", ylab="Value")
plot((length(delta_15)-999):length(delta_15), delta_15[(length(delta_15)-999):length(delta_15)], type
    ="l", main="delta=15", xlab="Iteration", ylab="Value")

# 自己相関関数
acf(delta_005, main="Autocorrelation for delta=0.05")
acf(delta_1, main="Autocorrelation for delta=1")
acf(delta_15, main="Autocorrelation for delta=15")
```

図 12 の R コードを以下に示す。

```
library(mvtnorm)
library(ggplot2)

set.seed(1)

mu <- c(1.1, 1.8)
sigma <- c(3, 4)
rho <- 0.6
sigmas <- matrix(c(sigma[1]^2, rho*sigma[1]*sigma[2], rho*sigma[1]*sigma[2], sigma[2]^2), nrow = 2)

x <- seq(-15.0, 15.0, 0.1)
y <- seq(-15.0, 15.0, 0.1)
grid <- expand.grid(x=x, y=y)
Z <- dmvnorm(grid, mean=mu, sigma=sigmas)

px_cond_y <- function(y) {
  px_mean <- mu[1] + rho*(sigma[1]/sigma[2])*(y-mu[2])
  px_scale <- sigma[1]*sqrt(1-rho^2)
  return(rnorm(1, mean=px_mean, sd=px_scale))
}

py_cond_x <- function(x) {
  py_mean <- mu[2] + rho*(sigma[2]/sigma[1])*(x-mu[1])
  py_scale <- sigma[2]*sqrt(1-rho^2)
  return(rnorm(1, mean=py_mean, sd=py_scale))
}

gibbs_sampling <- function(steps=1000, x_init=c(0, 0)) {
  samples <- matrix(0, nrow = steps+1, ncol = 2)
  samples[1,] <- x_init
```

```
  x <- x_init[1]
  y <- x_init[2]

  for (i in 2:(steps+1)) {
    if (i %% 2 == 0) {
      x <- px_cond_y(y)
    } else {
      y <- py_cond_x(x)
    }
    samples[i, ] <- c(x, y)
  }

  return(samples)
}

x_init <- c(-7.0, -7.0)
samples <- gibbs_sampling(steps=1000, x_init=x_init)

df <- data.frame(x = grid$x, y = grid$y, z = Z)
samples_df <- data.frame(x = samples[, 1], y = samples[, 2])
p <- ggplot(df, aes(x = x, y = y)) +
  geom_contour(aes(z = z)) +
  geom_point(data = samples_df, aes(x = x, y = y), alpha = 0.15, color = 'green') +
  geom_path(data = samples_df[1:50, ], aes(x = x, y = y), color = 'red') +
  geom_point(aes(x = x_init[1], y = x_init[2]), color = 'black') +
  labs(title = "Gibbs Sampling", x = "X", y = "Y") +
  theme_minimal()

print(p)

library(mvnormtest)
mshapiro.test(t(samples_df[-1,]))
#平均値
colMeans(samples_df[-1,])
#標準偏差
sqrt(cov(samples_df[-1,]))
#相関係数
cor(samples_df[-1,])
```

シミュレーションの R コードを以下に示す。

```
library(bayesm)
library(MASS)

set.seed(123)

n <- 300 #データ数
nx <- 3 #説明変数 x の数
nz <- 2 #属性 z の数
nrep <- 10 #時点 t の長さ
ncat <- 2 #ブランドの数

#顧客属性 z を生成
Z <- matrix(runif(n * nz), nrow = n, ncol = nz)
Z = t(t(Z) - apply(Z, 2, mean))

#Theta を生成
Theta <- matrix(rnorm(nz * nx), nrow = nz, ncol = nx)

#inverse-Wishart から V_beta を生成
Vbeta <- rWishart(1, df = nx + 4, Sigma = diag(nx))[,,1]

#混合分布用のパラメータ
weights <- c(0.6, 0.4)
means <- c(-3, 1.5)
sds <- c(1, 1.2)
mix <- function(weights, means, sds) {
  selected <- sample(1:length(weights), 1, prob = weights)
  return(rnorm(1, mean = means[selected], sd = sds[selected]))
}

#beta を生成
betaDGP <- function(z, Theta, Vbeta) {
  mbeta <- t(Theta) %*% z
```

```
  return(mvrnorm(1, mbeta, Vbeta) + mix(weights, means, sds))

}

#X,y を生成
ndata <- vector("list", n)
cate <- 1:ncat
for (i in 1:n) {
  beta0 <- betaDGP(Z[i,], Theta, Vbeta)
  beta <- matrix(beta0, nx, ncat)
  X <- matrix(runif(nrep * nx), nrep, nx)

  XX <- matrix(NA, nrow = nrep * ncat, ncol = nx)
  for (j in 1:ncat) {
    XX[((j - 1) * nrep + 1):(j * nrep), ] <- X
  }

  Xbeta <- XX %*% beta
  j <- nrow(Xbeta) / nrep
  Xbeta <- matrix(Xbeta, byrow = TRUE, ncol = j)
  prob <- exp(Xbeta) / as.vector(exp(Xbeta) %*% c(rep(1,j)))

  y <- integer(nrep)
  for (k in 1:nrep) {
    yp <- rmultinom(1, 1, prob[k, ])
    y[k] <- cate %*% yp
  }

  ndata[[i]] <- list(y = y, X = XX, beta = matrix(beta0, ncol = 1))
}

Data1 <- list(p = ncat, lgtdata = ndata, Z = Z)

#事前分布の設定
Prior1 <- list(ncomp = 2)

#MCMC のパラメータ設定
R <- 10000
keep <- 5
Mcmc1 <- list(R = R, keep = keep, nprint = 0)

out1 <- rhierMnlRwMixture(Data = Data1, Prior = Prior1, Mcmc = Mcmc1)

#図 1
bmat <- matrix(0, n, nx)
for (i in 1:n) {
  bmat[i,] <- Data1$lgtdata[[i]]$beta
}
par(mfrow = c(nx, 1))
for (i in 1:nx) {
  hist(bmat[,i], breaks = 30, col = "magenta",
       main = sprintf("Beta %d Distribution", i))
}
#図 2
plot(out1$betadraw)
#図 3
par(mfrow = c(3, 2))
beta_means <- t(apply(out1$betadraw, c(2,3), mean))
for (i in 1:3) {
  plot(beta_means[, i], type = "l", xlab = "", ylab = "", main = sprintf("Draw of beta %d", i))
  acf(beta_means[, i], type = "correlation", main = sprintf("Acf of beta %d", i))
}
#図 4
par(mfrow = c(1, 1))
plot(out1$loglike,type="l",xlab="",ylab="", main="Draw of loglike")
#図 5と図 6
par(mfrow = c(3, 2))
for (i in 1:3) {
  for (j in 1:2) {
    idx <- (i-1)*2 + j
    plot(out1$Deltadraw[,idx], type="l", xlab="", ylab="", main=sprintf("Draw of Theta[%d,%d]", i,
         j))
    acf(out1$Deltadraw[,idx], type="correlation", main=sprintf("Acf of Theta[%d,%d]", i, j))
  }
```

```
}
```

事例分析の R コードを以下に示す。

```
library(bayesm)
set.seed(123)

Dat1 <- read.table("xdata.csv", header=TRUE, sep=",", na.strings="NA",dec=".",strip.white=
    TRUE)
IndAttr <- read.table("zdata.csv", header=TRUE, sep=",", na.strings="NA",dec=".",strip.white=
    TRUE)
reg=levels(factor(Dat1$ID))
nreg=length(reg)

p=3
na=3
nz=3

lgtdata <- list()
for (j in 1:nreg){
  y=Dat1$brand[Dat1$ID==reg[j]]
  Xa=cbind(Dat1[Dat1$ID==reg[j], c('PriceSh', 'PriceKa', 'PriceKo', 'TimeSh', 'TimeKa', 'TimeKo
      ', 'AreaSh', 'AreaKa', 'AreaKo')])
  X=createX(p, na=na, nd=NULL, Xa=Xa, Xd=NULL, DIFF=FALSE, base=3)
  lgtdata[[j]]=list(y=y,X=X)
}

Z=t(t(as.matrix(IndAttr))-apply(IndAttr,2,mean))
Data3=list(p=p,lgtdata=lgtdata,Z=Z)
Prior3=list(ncomp=1)

Mcmc3=list(R=50000,sbeta=0.01,keep=1)
out3=rhierMnlRwMixture(Data=Data3, Mcmc=Mcmc3, Prior=Prior3)

PD <- max((out3$loglike)[-c(1:45000)]) - mean((out3$loglike)[-c(1:45000)])
DIC3 <- -2*mean(out3$loglike) + 2*PD
print(DIC3)

s=45001
t=50000
beta.mean = beta.sd = beta.t = matrix(0, nrow=nreg, ncol=5)

for(i in 1:nreg){
  for(j in 1:5){
    beta.mean[i,j] <- mean(out3$betadraw[i,j,s:t])
    beta.sd[i,j] <- sd(out3$betadraw[i,j,s:t])
    beta.t[i,j] <- beta.mean[i,j] / beta.sd[i,j]
  }
}
Delta.mean = Delta.SD = Delta.t = matrix(0, nrow=1, ncol=10)
for(i in 1:10){
  Delta.mean[1,i] <- mean(out3$Deltadraw[s:t,i])
  Delta.SD[1,i] <- sd(out3$Deltadraw[s:t,i])
  Delta.t[1,i] <- Delta.mean[1,i] / Delta.SD[1,i]
}

beta.mean
beta.sd
beta.t
Delta.mean
Delta.SD
Delta.t

# summary(out3$Deltadraw)
summary(t(out3$betadraw[1,,]),burnin=45000)
# plot(out3$Deltadraw)
plot(out3$betadraw)
```

参考文献

[1] Peter M Guadagni and John DC Little. A logit model of brand choice calibrated on scanner data. *Marketing science*, Vol. 2, No. 3, pp. 203–238, 1983.

[2] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[3] Peter D Hoff. *A first course in Bayesian statistical methods*, Vol. 580. Springer, 2009.

[4] P Rossi, G Allenby, and R McCulloch. Bayesian statistics and marketing. no. 13 in wiley series in probability and statistics, 2005.

[5] 照井伸彦, Wirawan Dony Dahana, 伴正隆. マーケティングの統計分析. シリーズ統計科学のプラクティス / 小暮厚之, 照井伸彦編, No. 3. 朝倉書店, 2009.

[6] 照井伸彦, 佐藤忠彦. 現代マーケティング・リサーチ：市場を読み解くデータ分析. 有斐閣, 新版, 2022.

[7] 佐藤忠彦. マーケティングの統計モデル. 統計解析スタンダード / 国友直人, 竹村彰通, 岩崎学編. 朝倉書店, 2015.

[8] 小西貞則, 越智義道, 大森裕浩. 計算統計学の方法：ブートストラップ・EM アルゴリズム・MCMC. シリーズ予測と発見の科学, No. 5. 朝倉書店, 2008.

おわりに

統計数理研究所が推進している統計エキスパート養成プログラムでは必ずしも統計学を専門としているわけではない各分野の若手研究者とメンターにより統計エキスパート演習を行っているが、直接の関係者以外はこの統計エキスパート養成事業の中で具体的にどの様な講義や議論が行われているかはよく分からないのではないかと思われる。2023 年に実施した一つの少人数グループによる演習では、統計学の基礎と応用について基礎的ではあるがしばしば見逃しがちな内容を検討した。統計学の専門的な研究とまではいかないが、統計エキスパートにとり有益と考えられる基礎統計を巡ってある意味では初歩的な議論かもしれないが、真面目に議論し、文案をまとめた。そうした議論の過程に置いて、関連する基礎的な書籍の一つである「統計学基礎」(統計検定 2 級の教科書) に書かれている内容を確かめる、あるいは深めるために新たに作成した R プログラム, Python プログラムを作成した。また統計エキスパートプログラムでは様々な講義が行われているが、単位取得のために研修生は多くのレポートを書いている。そうした中から二つのレポート、およびメンターが提供した話題を加えて応用統計から三つの話題を取りあげた。むろん単なる社会科学系（経済・経営・統計）の一つのグループ演習の例であるから取りあげた話題は全体から見れば偏った選択とならざるを得ないことをお断りする。

統計数理研究所では文部科学省の支援を受け、統計人材育成コンソーシアム (https://stat-expert.ism.ac.jp/) における事業の一環として統計エキスパート養成計画を 2021 年より実施している。統計エキスパート人材育成の事業内容は多岐に及び、この報告書はそのごく一部分に過ぎないが、あるグループがこれまでエキスパート演習として検討した議論を例として示しておくことも意味があるのではないかと判断した。

日本における統計科学分野における統計エキスパート人材の必要性が叫ばれている中で日本における統計科学および応用の諸分野における統計エキスパート養成の教育として妥当な内容は何か、今で

もなお模索中である。今後の教育内容を改善する参考の為にも、各界の統計科学や応用諸分野などの関係者からの忌憚のないコメントを期待したい。

国友直人 (著者代表)
2024 年 1 月