変換による財務データの統計解析：売上高の場合

変換による財務データの統計解析
―売上高の場合―

地　道　正　行

要　旨

本稿では，データベース Osiris から抽出された世界160カ国における全上場企業（一般事業会社）93,836社の2015年の対数変換した売上高に対して非対称分布を当てはめた結果と，Box-Cox 変換後に正規分布を当てはめた結果を，可視化と赤池情報量規準を用いて比較検討することによってモデル選択を行った．この結果として，対数変換したものに非対称ティー分布を当てはめたものが最も当てはまりが良かった．さらに，変換前の売上高の粗データには，対数非対称ティー分布を当てはめた場合が最も良い結果を与えることもわかった．

キーワード：財務ビッグデータ（Financial Big Data），Box-Cox 変換（Box-Cox Transformation），（対数）非対称正規分布（(log-)skew-normal distribution），（対数）非対称ティー分布（(log-)skew-t distribution），尤度解析（likelihood analysis）

I　はじめに

家計の所得や企業の売上高などの社会科学の諸分野（特に，経済学，経営学，商学など）で扱われるデータは，対称性をもつ分布（例えば，正規分布）に従うものよりも，非対称で歪みをもつものが多いように思われる[1)]．このようなデータに，正規分布を前提とする統計学の理論と手法を直接応用する

1)　例えば，所得の分布に関する考察としては，Klein（1962）が詳しい．

ことは，何らかの齟齬を生むことが予想される．例えば，本稿で扱う売上高のように右に極端に歪んだ分布の（中心）位置をデータの平均値で推定することの意味を考えることは重要である．

このような問題に対して，Tukey（1957）はデータを変換するクラスを考察し，変換後のデータが対称な分布に従っているという結果に基づいて統計的に解析することの重要性について言及している．

本研究では，Bureau van Dijk（BvD）社[2)] のデータベース Osiris から抽出された世界160カ国の一般事業会社の全上場企業93,836社の主要財務情報に関する財務データ（粗データ）のうち，売上高の分布を考察する．その際，まず要約と可視化をすることによってデータが歪みを持つことを確認し，さらに対数変換したものも若干歪むという知見に基づいて（II 節），データを対数変換したものに非対称分布族の当てはめを行う（III 節）．さらに，データの Box-Cox 変換（Box-Cox transformation）後に正規分布を当てはめるために，最尤法によって母数を推定した後，その結果を可視化することによって正規分布への当てはめを検証する（IV 節）．さらに，対数変換したデータに正規分布と非対称分布を当てはめた場合と，Box-Cox 変換を行ったものに正規分布を当てはめた場合について赤池情報量規準（Akaike Information Criterion: AIC）の値を比較することによってモデル選択を行うとともに，（変換する前の）粗データに，対数正規分布，対数非対称分布，Box-Cox 変換に関する分布を当てはめた場合も AIC を比較することによってモデル選択を行う．（V 節）．

なお，付録 A では，正規分布と対数正規分布に対する尤度（likelihood）に関連する事項の関係を述べている．同様に，付録 B では，非対称正規分布（かつ非対称ティー分布）と対数非対称正規分布（かつ対数非対称ティー分布）に対する尤度に関連する事項の関係を述べている．また，Box-Cox 変換に関する尤度解析について付録 C に与えると共に，データの Box-Cox 変

2） https://www.bvdinfo.com/en-gb/

換に関するインタラクティブな機能を備えた可視化ツール（Web アプリケーション）の作成については，付録 D を参照されたい．

本研究は，Tukey（1977）によって提唱された探索的データ解析（Exploratory Data Analysis: EDA）にもとづいて行われる．本稿は，Leisch（2002）による Sweave を利用することによって動的に生成されており，再現可能性を確保している．

II　データの要約と可視化

データは，BvD 社の世界の全上場企業のデータベース Osiris から抽出されたものを利用する[3]．地道（2018-a, b），Jimichi *et al.*（2018）で行ったように，本稿でも，2015年の売上高を中心に可視化を行う．売上高（`sales2015`）のデータは，以下のようなものである：

$$\{x_1, \ldots, x_n\} = \{1859571, \ldots, 9052\}$$

ここで，標本の大きさ（データの個数）は，$n=30764$ である．このデータの数値的要約を表 1 に与える．この結果からデータは，平均値（mean）が中央値（median）よりも大きく右に位置することから，右に歪んだ分布か

表 1　売上高データの要約

	sales2015
Min.:	1
1st Qu.:	11428
Median:	63479
Mean:	1221878
3rd Qu.:	362194
Max.:	482130000

3）2018年 4 月に抽出されたデータセットを利用した．なお，このデータベースから抽出されたデータの前処理やデータラングリングなどの詳細については地道（2018-a）を参照されたい．

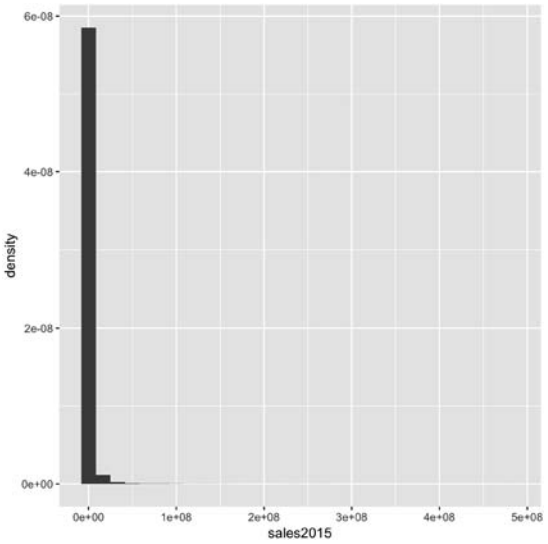

図１　売上高のヒストグラム：粗データ

らのデータであることがわかる[4]．

この結果をヒストグラムを描くことによって可視化し，検証しよう．

図１より，データは右に「歪んだ」分布から生成されていることがわかる．Jimichi *et al.*（2018）でも指摘されているが，このような歪みを補正するための代表的な変換は対数をとることである（Tukey（1977），Mosteller and Tukey（1977），Fox and Weisberg（2019）も参照のこと）．図２に対数スケールで描き直したものを与える．

このプロット（図２）から，売上高の対数（log(sales2015)）は，「正規分布」で近似できそうであるけれども，注意深くみると若干「歪んだ分布」となっていることがわかる．実際，歪度を求めると，$g_1 := m_3^2/m_2^{3/2} = -0.24 (< 0)$ となり，この結果（歪度が負の値をもつこと）から左に歪ん

4）社会科学系の研究論文では，数値的な要約を与える際に習慣として，標本の大きさ（size），平均値（mean），標準偏差（standard deviation），最小値（minimum），最大値（maximum）を表示しているものがあるが，非対称な分布からのデータである場合に，これらの統計量では，その知見を得ることは難しいことに注意しよう．

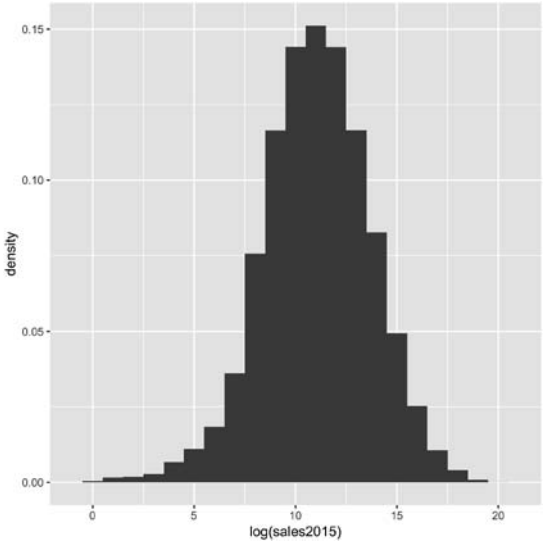

図２　売上高のヒストグラム：対数スケール

でいることがわかる．ここで，$m_j := \sum_{i=1}^{n}(x_i - \bar{x})^j/n$ はデータ $\{x_1, \ldots, x_n\}$ の平均 $\bar{x} = \sum_{i=1}^{n} x_i/n$ まわりの j 次モーメントである．

Jimichi *et al.*（2018）では，この歪みに対応するために，売上高の対数をとったものに非対称分布（skew-symetric distribution）を当てはめている．具体的には，非対称正規分布（skew-normal distribution）と非対称ティー分布（skew-t distribution）を当てはめ，赤池情報量規準（AIC）を利用して，モデル選択を行い，非対称ティー分布がデータに最も当てはまるという結果を得ている．

そこで，本稿では，2018年に再度抽出したデータで，これらのことが再現するかを確認するとともに，対数変換を拡張した Box-Cox 変換を利用した場合との比較を行う．

Ⅲ　対数変換データへの非対称分布の当てはめ

前節の可視化によって得られた知見，すなわち，本稿で扱っているデータ

の対数をとったものが若干歪んだ分布構造をもつことをふまえて統計モデリングを行う．（ここで行う統計モデリングは，地道（2017-a, b），Jimichi *et al*.（2018）の結果をベースとしている．）

ここでは，売上高の対数（log(sales2015)）の分布を，Azzalini（1985）による非対称正規分布（skew-normal distribution）と Azzalini and Capitanio（2014）による非対称ティー分布（skew-t distribution）を使ってモデリングする．記号として，非対称正規分布を $\mathsf{SN}(\xi, \omega^2, \alpha)$，非対称ティー分布を $\mathsf{ST}(\xi, \omega^2, \alpha, \nu)$ と表す．ここで，$\xi(\in\mathbb{R})$ は位置母数（location parameter），$\omega(\in\mathbb{R}^+)$ は尺度母数（scale parameter）に対応し，$\alpha(\in\mathbb{R})$ は傾斜母数（slant parameter）または非対称母数（skew parameter）と呼ばれる．また，$\nu(\in\mathbb{R}^+)$ は自由度（degree of freedom）と呼ばれ，$\mathbb{R}$ は実数全体，$\mathbb{R}^+$ は正の実数を表す記号である．

これらの分布における母数をデータから最尤法（maximum likelihood method）によって推定する[5]と，まず非対称正規分布の母数の推定値は，

$$(\hat{\xi},\ \hat{\omega},\ \hat{\alpha}) = (13.01,\ 3.29,\ -1.15)$$

となり，ヒストグラムと統計モデル（p.d.f. の母数を最尤推定値で置き換えたもの）を重ね書きしたものが図 3 である．なお，非対称正規分布の p.d.f. については，付録 B の(11)式を参照されたい．

次に非対称ティー分布に対しては，

$$(\hat{\xi},\ \hat{\omega},\ \hat{\alpha},\ \hat{\nu}) = (12.57,\ 2.9,\ -0.83,\ 22.42)$$

となり，ヒストグラムと統計モデルを重ね書きしたものが図 4 である．なお，非対称ティー分布の p.d.f. については，付録 B の(19)式を参照されたい．

これらのプロットの結果を比較すると，図 3（非対称正規分布）と図 4（非対称ティー分布）の当てはまりの結果は優劣付けがたいことがわかる．この点については，後節で情報量規準によって選択する．

5） 非対称分布によるデータ解析を行うために sn パッケージを利用している．

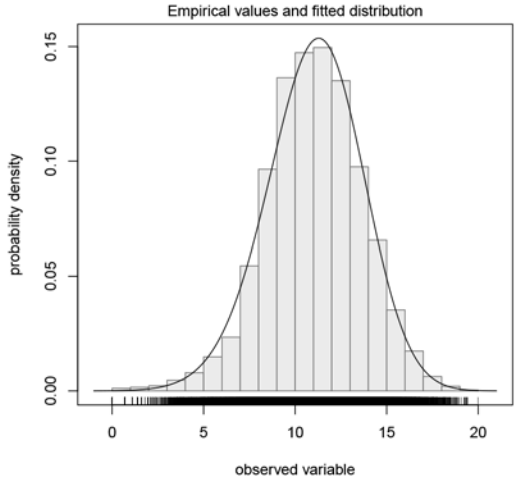

図 3　売上高の対数 $\log(\texttt{sales2015})$ のヒストグラムと統計モデル $f_{\mathrm{SN}}(\log(\texttt{sales2015})\,|\hat{\xi},\ \hat{\omega},\ \hat{\alpha})$

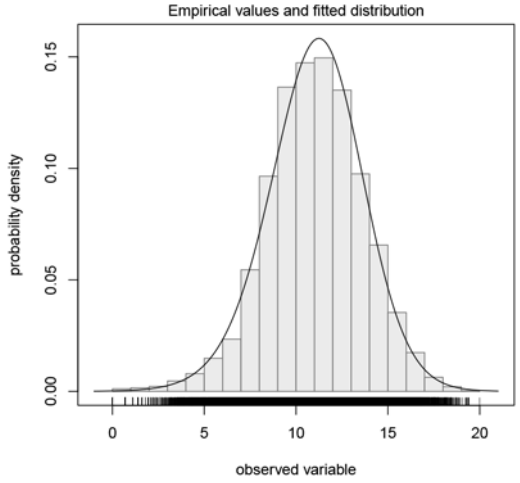

図 4　売上高の対数 $\log(\texttt{sales2015})$ のヒストグラムと統計モデル $f_{\mathrm{ST}}(\log(\texttt{sales2015})\,|\hat{\xi},\ \hat{\omega},\ \hat{\alpha},\ \hat{\nu})$

Ⅳ Box-Cox 変換よる正規分布の当てはめ

Box and Cox（1964）は，観測 $X(>0)$ の分布が非正規であり，非対称である場合に，以下のような変換を提案し，対称化し正規分布に近づけることを考えた：

$$X^{(\lambda)}:=\begin{cases}\dfrac{X^{\lambda}-1}{\lambda}, & \lambda\neq 0,\\ \log(X), & \lambda=0\end{cases} \tag{1}$$

この変換は Box-Cox 変換または，ベキ正規変換（power normal transformation）と呼ばれる．Box-Cox 変換の回帰分析への応用については，Carroll and Ruppert（1988），Draper and Smith（1998）の第13章，または，Fox and Weisberg（2019）の 3.4 節を参照されたい．

適当な λ の値にして，この変換によって以下が仮定される：

$$X^{(\lambda)}\sim \mathsf{N}(\mu,\ \sigma^2)\text{（；平均 }\mu\text{，分散 }\sigma^2\text{ の正規分布）} \tag{2}$$

図 5 は，前節で可視化された売上高のデータ（`sales2015`）に対する Box-Cox 変換を $\lambda=-1,\ -0.5,\ 0(\log),\ 0.5,\ 1.0$ の値に対して行ったものに対するボックスプロットである[6]．このプロットから，$\lambda=0$（対数変換）の近辺でデータの分布構造が対称に近いものとなっていることがわかる．

Box-Cox 変換 $X^{(\lambda)}$ において，λ は母数として扱われ，変換母数（transformation parameter）と呼ばれる．λ を推定する方法としては，最尤（Maximum Likelihood: ML）法がしばしば利用され，対数尤度（log-likelihood）関数を変換母数 λ に関して最大化することで達成される[7]．付録 C に Box-Cox 変換に関する尤度解析についての簡単な説明を与えるので参考にされたい．図 6 は，売上高のデータ（`sales2015`）に基づいて計算された対数尤度関数を変換母数を変化させることによって描いたものである[8]．

6） **car** パッケージの `symbox` 関数を利用して描画した．

7） より正確には，プロファイル対数尤度関数（profile log-likelihood function）である．付録 C の(39)式を参照のこと．

8） **car** パッケージの `boxCox` 関数を利用して描画した．

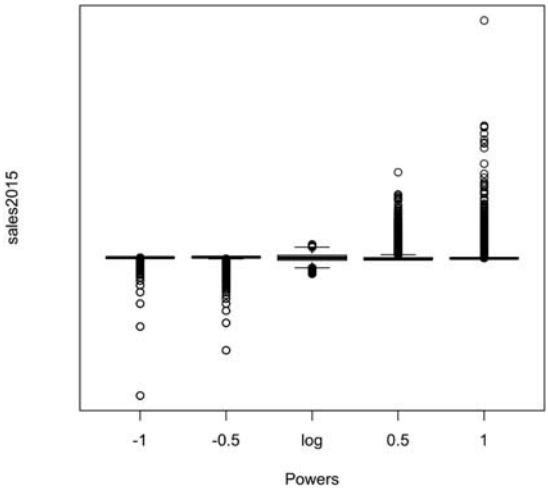

図5　売上高の Box-Cox 変換のボックスプロット：
$\lambda = -1,\ -0.5,\ 0(\log),\ 0.5,\ 1.0$ の場合

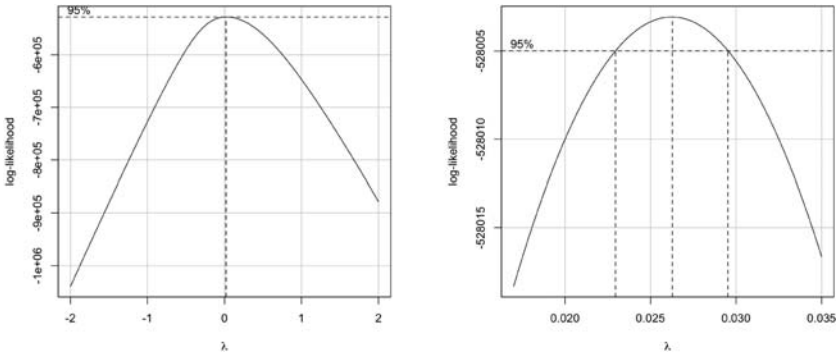

図6　売上高データに基づく対数尤度関数のプロット：
左図：$\lambda \in [-2,\ 2]$，右図：$\lambda \in [0.017,\ 0.035]$

最尤推定の結果，得られた推定値（最尤推定値）は，$\hat{\lambda} = 0.026$ となる．この推定値を使って変換された売上高データ

$$\{x_1^{(\hat{\lambda})},\ \cdots,\ x_n^{(\hat{\lambda})}\} = \left\{\frac{x_1^{\hat{\lambda}}-1}{\hat{\lambda}},\ \cdots,\ \frac{x_n^{\hat{\lambda}}-1}{\hat{\lambda}}\right\}$$

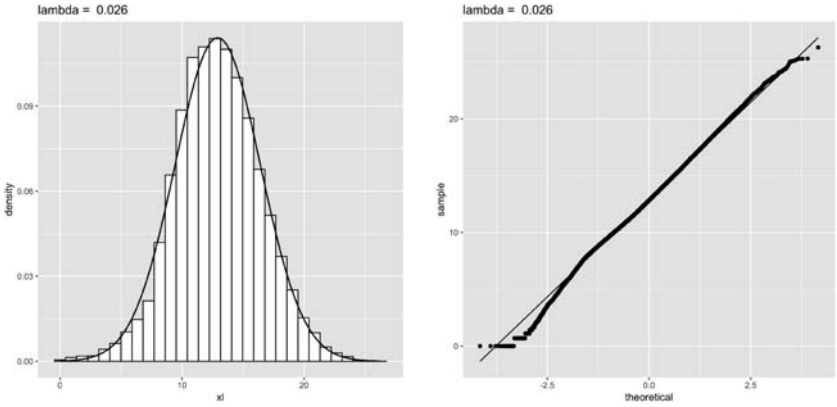

図 7　Box-Cox 変換された売上高データのヒストグラム（統計モデル付き）と正規 Q-Q プロット

のヒストグラムに，正規分布 $\mathrm{N}(\hat{\mu}(\hat{\lambda}), \hat{\sigma}^2(\hat{\lambda}))$ の確率密度関数（；統計モデル）

$$f(x_1^{(\hat{\lambda})}) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}(\hat{\lambda})} \exp\left\{ -\frac{(x^{(\hat{\lambda})} - \hat{\mu}(\hat{\lambda}))^2}{2\hat{\sigma}^2(\hat{\lambda})} \right\}$$

を重ね書きしたものと，正規 Q-Q プロットを図 7 に与える．ここで，統計モデルにおける母数の推定値は，

$$\hat{\mu}(\hat{\lambda}) := \frac{1}{n} \sum_{i=1}^{n} x_i^{(\hat{\lambda})}, \quad \hat{\sigma}^2(\hat{\lambda}) := \frac{1}{n} \sum_{i=1}^{n} (x_i^{(\hat{\lambda})} - \hat{\mu}(\hat{\lambda}))^2$$

で与えられる．

図 7 から，Box-Cox 変換後のデータは正規分布で近似できそうであるが，左裾のあたりで若干当てはまりが悪いことが分かる．

以上の可視化された結果（図 3，4，7 参照）からは，当てはまりの良さの比較を行うことが難しいため，次節で赤池情報量規準を用いて数値的にモデル選択を行う．

V　赤池情報量規準によるモデル選択

ここでは，売上高の対数に対して，正規分布，非対称正規分布，非対称ティー分布を当てはめたときと，Box-Cox 変換後に正規分布を当てはめたときの赤池情報量規準[9] AIC を比較することによってモデル選択を行う．なお，Box-Cox 変換後に正規分布を当てはめた場合に関する AIC の定義については，付録 C における注意 1 を参照のこと．

表 2　売上高の対数変換と Box-Cox 変換後のデータに各種の分布を当てはめたときの AIC 値

モデル	dim	AIC
正規分布（対数変換後）	2	146768
非対称正規分布（対数変換後）	3	146533
非対称ティー分布（対数変換後）	4	146461
正規分布（Box-Cox 変換後）	3	164335

表 2 における dim はそれぞれのモデルの母数ベクトルの次元を表す．非対称ティー分布の場合の AIC が最も小さく，このモデルがここで考察されている中で最も良いという結果が与えられた．よって，Box-Cox 変換後のデータに正規分布を当てはめるよりも，対数変換後に非対称ティー分布を当てはめる場合が良く，地道（2017-b），Jimichi *et al.*（2018）の結果を肯定する結果となった．

ただし，Box-Cox 変換後のデータに正規分布を当てはめた結果は，変換母数の推定値 $\hat{\lambda}$ を，（変換する前の）売上高の粗データにもとづいたアルゴリズムによって計算されたものとなっている（注意 1 参照）ので，このことに結果が依存する可能性がある．実際，Box-Cox 変換は対数変換を含むように拡張されているにも関わらず，対数変換後に正規分布を当てはめた AIC の値よりも Box-Cox 変換後に正規分布を当てはめた場合が「過大に評価」さ

9）赤池情報量規準については，例えば，Akaike（1973），Konishi and Kitagawa（2008）を参照されたい．

れている感がある．この観点から，売上高の粗データに対数正規分布，対数非対称正規分布，対数非対称ティー分布，Box-Cox 分布[10] を当てはめた場合の比較を表 3 に与える．なお，これらの AIC の値は，変換前と変換後の AIC に関する関係式(10)，(18)，(26)，(43)を利用して求めている．

表 3　売上高のデータに各種の分布を当てはめたときの AIC 値

モデル	dim	AIC
対数正規分布	2	825645
対数非対称正規分布	3	825410
対数非対称ティー分布	4	825338
Box-Cox 分布	3	825398

表 3 より，対数正規分布よりも Box-Cox 分布が当てはまりが良いという「直感」に従った結果が得られており，さらに，Box-Cox 分布は，対数非対称正規分布よりも若干当てはまりが良いという結果が得られていることは興味深い．さらに，最も当てはまりが良いのは，（変換後の結果と同様に）対数非対称ティー分布であることもわかる．

VI　おわりに

本稿では，売上高の対数に非対称分布を当てはめた結果と，Box-Cox 変換後に正規分布を当てはめた結果を可視化と AIC を用いて比較検討することによってモデル選択を行った．結果としては，対数変換したものに非対称ティー分布を当てはめたものが最も当てはまりが良いという結果となり，地道（2017-b），Jimichi *et al.*（2018）の結果が肯定されるものとなった．さらに，変換前の売上高の粗データには，Box-Cox 分布を当てはめた結果も悪くないが，対数非対称ティー分布を当てはめた場合が最も良い結果を与えることもわかった．

10）標本 X の Box-Cox 変換 $X^{(\lambda)}$ 後の分布が正規分布になる場合に，変換前の標本 X の分布を便宜上「Box-Cox 分布」と呼んでいる．

今後の課題としては，地道（2017-b），Jimichi *et al.*（2018）で考察されている両対数モデルを当てはめた場合と，Box-Cox 変換を応用した場合との比較・検討をすることである．また，注意 1 で言及した Box-Cox 変換後のデータへの尤度法を適用する際の問題点を考察することや，Box-Cox 変換後のデータにティー分布を当てはめることを考えることも興味深いテーマであろう．

最後に，本稿の冒頭で述べた売上高の（中心）位置をデータの平均値で推定することの意味を考える．平均値は

$$\bar{x}=1221877.64$$

で与えられ，この値は，表 1 にあるように第 3 四分位点（$Q_3=362193.75$）の値を大きく超えている．すなわち，平均値は対象となる全企業の75％の企業に対する売上高よりも遙かに大きくなり，平均値をデータの中心位置として考えること自体が理解に苦しむ結果となる．統計学の基礎では，このような右に歪んだデータの場合は，中央値：

$$x_{\mathrm{med}}=63479$$

が中心位置としては適切であると指摘されており，その性質（全体の50％の位置にある値）から妥当な値となっているように思われる．しかしながら，中央値は数学的な扱いが難しいこと[11]が知られている．これらのことを折衷する特性値として，例えば，売上高の対数に非対称ティー分布を当てはめたときの位置母数 ξ の最尤推定値を，データの単位へもどすために，対数関数の逆関数（指数関数）で変換した以下のものを考えることができる：

$$\exp(\hat{\xi})=62038.78$$

この値は，中央値と近い値をとり，さらに最尤推定量の漸近的な性質（漸近正規性からその指数関数で変換したものは漸近的に対数正規分布に従うこと）を援用することが可能であり，統計的推測などの観点から扱いやすい．以上のことから，歪んだ分布から生成されたデータの統計解析は，単純なようで難しい問題を数多くもつことがわかる．（筆者は関西学院大学商学部教授）

11）例えば，標本中央値の漸近的な挙動や，相対的な効率の計算などを行うためには複雑な計算が必要となることが知られている．

謝辞

本研究の一部は以下の助成を得ていることに感謝の意を表する：

- 科学研究費基盤研究 C：「グラフィカル・データ・アナリシスによる格差研究と社会環境会計による解決方法の提案」（2016年～2018年），課題番号：16K04022，研究代表者：阪智香
- 科学研究費基盤研究 C：「共有価値創造（CSV）のための社会環境会計の構築」（2019年～2021年），課題番号：19K02006，研究代表者：阪智香
- 平成29年度学際大規模情報基盤共同利用・共同研究拠点（JHPCN）課題：「財務ビッグデータの可視化と統計モデリング」，課題番号：jh171002-NWJ，研究代表者：地道正行
- 平成30年度学際大規模情報基盤共同利用・共同研究拠点（JHPCN）課題：「財務ビッグデータの可視化と統計モデリング」，課題番号：jh181001-NWJ，研究代表者：地道正行
- 平成31年度学際大規模情報基盤共同利用・共同研究拠点（JHPCN）課題：「財務ビッグデータの可視化と統計モデリング」，課題番号：jh191002-NWJ，研究代表者：地道正行
- 関西学院大学図書館図書費 B，個人研究費

また，BvD 社の増田歩氏にはデータの抽出に関して多大なるご協力いただいた．ここに感謝の意を表する．

参考文献

[1] Akaike, H.（1973）Information theory and an extension of the maximum likelihood principle, *Proceedings of the 2nd International Symposium on Information Theory,* Petrov, B. N., and Caski, F.（eds.）, Akadimiai Kiado, Budapest: pp. 267-281.

[2] Azzalini, A.（1985）A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics,* Vol. 12, No. 2, pp. 171-178.

[3] Azzalini, A. with the collaboration of A. Capitanio（2014）*The Skew-Normal and Related Families,* Cambridge University Press, Institute of Mathematical Statistics Monographs.

[4] Carroll, R. J. and D. Ruppert（1988）*Transformation and Weighting in Regression,* Chapman and Hall, Monographs on Statistics and Applied Probability.

[5] Draper, N. R. and H. Smith（1998）*Applied Regression Analysis,* Third Edition, Wiley Series in Probability and Statistics, Wiley Interscience.

[6] Fox, J. and S. Weisbrerg（2019）*An R Companion to Applied Regression,* Third Edition, Sage.

[7] 地道正行（2017-a）R による対数非対称正規線形モデルによる財務データの統計モデリング，商学論究，第64巻，第５号，pp. 159-185，関西学院大学商学研究会．

［8］ 地道正行（2017-b）Rを利用した対数非対称分布族にもとづく財務データの統計モデリング，経済学論究，第71巻，第2号，pp. 141-174，関西学院大学経済学部研究会．
［9］ Jimichi, M., D. Miyamoto, C. Saka, and S. Nagata（2018）Visualization and statistical modeling of financial big data: Double-log modeling with skew-symmetric distributions, *Japanese Journal of Statistics and Data Science,* Vol. 1, No. 2, pp. 347-371, https://doi.org/10.1007/s42081-018-0019-1
［10］ 地道正行（2018-a）探索的財務ビッグデータ解析―前処理，データラングリング，再現可能性―，商学論究，第66巻，第1号，pp. 1-31，関西学院大学商学研究会．
［11］ 地道正行（2018-b）探索的財務ビッグデータ解析―データ可視化，統計モデリング，モデル選択，モデル評価，動的文書生成，再現可能研究―，商学論究，第66巻，第2号，pp. 1-41，関西学院大学商学研究会．
［12］ Klein, L. R.（1962）*An Introduction to Econometrics,* Prentice Hall.
［13］ Konishi, S. and G. Kitagawa（2008）*Information Criteria and Statistical Modeling,* Springer.
［14］ Leisch, F.（2002）*Sweave: Dynamic generation of statistical reports using literate data analysis,* In Wolfgang Härdle and Bernd Rönz, editors, Compstat 2002 - Proceedings in Computational Statistics, pp. 575-580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9.
［15］ Mosteller, F. and J. W. Tukey（1977）*Data Analysis and Regression: A Second Course in Statistics,* Addison-Wesley, Reading Mass.
［16］ Patil, DJ（2012）*Data Jujitsu: The Art of Turning Data into Product,* An O'Reilly Radar Report, O'Reilly.
［17］ Tukey, J. W.（1957）On the comparative anatomy of transformations, *The Annals of Mathematical Statistics,* Vol. 28, No. 3, pp. 602-632.
［18］ Tukey, J. W.（1977）*Exploratory Data Analysis,* Addison-Wesley Publishing Co.
［19］ 梅津雄一，中野貴広（2018）『Rと Shinyで作る Webアプリケーション』，C&R研究所．
［20］ Unwin, A.（2015）*Graphical Data Analysis with R,* Chapman and Hall/CRC

付録

ここでは，正規分布と対数正規分布，非対称分布，Box-Cox変換に関する尤度解析に関する事項，さらにBox-Cox変換をデータに当てはめることを可視化するWebアプリケーションの作成について述べる．なお，各分布で扱われる母数ベクトルは，本来であればそれぞれ別の記号を使う必要がある

が，煩雑になることを避けるために，一部（正規分布の場合）を除いて $\boldsymbol{\theta}$ を利用していることに注意しよう．

付録 A　正規分布と対数正規分布

A.1　正規分布

確率変数 X が確率密度関数（probability density function: p.d.f.）：

$$f_{\mathsf{N}}(x\,|\,\boldsymbol{\theta}_{\mathsf{N}})=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{3}$$

をもつとき，確率変数 X は正規分布 $\mathsf{N}(\mu,\ \sigma^2)$ に従うと呼ばれ，

$$X\sim\mathsf{N}(\mu,\ \sigma^2)$$

と書かれる．ここで，

$$\mu\in\mathbb{R},\quad \sigma\in\mathbb{R}^+:=(0,\ \infty)$$

は未知母数であり，$\boldsymbol{\theta}_{\mathsf{N}}:=[\mu,\ \sigma^2]'$ は母数ベクトル[12] である．

A.2　対数正規分布

確率変数 X に対して，その対数 $\log(X)$ が正規分布 $\mathsf{N}(\mu,\ \sigma^2)$ に従うとき，X は対数正規分布 $\mathsf{LN}(\mu,\ \sigma^2)$ に従うといわれる：

$$X\sim\mathsf{LN}(\mu,\ \sigma^2)\overset{\text{def.}}{\iff}\log(X)\sim\mathsf{N}(\mu,\ \sigma^2)$$

対数正規分布の p.d.f. は以下のように与えられる：

$$f_{\mathsf{LN}}(x\,|\,\boldsymbol{\theta}_{\mathsf{N}})=\frac{1}{\sqrt{2\pi\sigma^2}\,x}\exp\left\{-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right\},\quad x\in\mathbb{R}^+ \tag{4}$$

A.3　対数正規分布の対数尤度と正規分布の対数尤度の関係

対数正規分布と正規分布の p.d.f. には以下の関係が成り立つ：

$$f_{\mathsf{LN}}(x\,|\,\boldsymbol{\theta}_{\mathsf{N}})=f_{\mathsf{N}}(\log(x)\,|\,\boldsymbol{\theta}_{\mathsf{N}})\frac{1}{x} \tag{5}$$

12）プライム（′）は行列・ベクトルの転置を表す．

この関係を利用すると，無作為標本 $\{X_1, \ldots, X_n\}$ が対数正規分布 $\mathsf{LN}(\mu, \sigma^2)$ に従うとき，その同時確率密度関数が，

$$\prod_{i=1}^{n} f_{\mathsf{LN}}(x_i \mid \boldsymbol{\theta}_{\mathsf{N}}) = \prod_{i=1}^{n} f_{\mathsf{N}}(\log(x_i) \mid \boldsymbol{\theta}_{\mathsf{N}}) \frac{1}{x_i} \tag{6}$$

となることがわかる．

さらに，この結果から，対数正規分布に従う無作為標本 $\{X_1, \ldots, X_n\}$ にもとづく対数尤度は，

$$\begin{aligned} \ell_{\mathsf{LN}}(\boldsymbol{\theta}_{\mathsf{N}} \mid \boldsymbol{x}) &= \log \prod_{i=1}^{n} f_{\mathsf{LN}}(x_i \mid \boldsymbol{\theta}_{\mathsf{N}}) = \log \prod_{i=1}^{n} f_{\mathsf{N}}(\log(x_i) \mid \boldsymbol{\theta}_{\mathsf{N}}) \frac{1}{x_i} \\ &= \ell_{\mathsf{N}}(\boldsymbol{\theta}_{\mathsf{N}} \mid \log \boldsymbol{x}) - \sum_{i=1}^{n} \log x_i \end{aligned} \tag{7}$$

となる．これは，対数正規分布に従う無作為標本 $\{X_1, \ldots, X_n\}$ にもとづく対数尤度 $\ell_{\mathsf{LN}}(\boldsymbol{\theta}_{\mathsf{N}} \mid \boldsymbol{x})$ と対数変換された無作為標本 $\{\log(X_1), \ldots, \log(X_n)\}$ が正規分布に従う場合の対数尤度 $\ell_{\mathsf{N}}(\boldsymbol{\theta}_{\mathsf{N}} \mid \log(\boldsymbol{x}))$ の間の関係を表している．

このことから，

$$\max_{\boldsymbol{\theta}_{\mathsf{N}}} \ell_{\mathsf{LN}}(\boldsymbol{\theta}_{\mathsf{N}} \mid \boldsymbol{x}) = \max_{\boldsymbol{\theta}_{\mathsf{N}}} \ell_{\mathsf{N}}(\boldsymbol{\theta}_{\mathsf{N}} \mid \log \boldsymbol{x}) - \sum_{i=1}^{n} \log x_i \tag{8}$$

が成り立つので，結局，それぞれの最尤推定値（ベクトル）は一致する：

$$\hat{\boldsymbol{\theta}}_{\mathsf{N}} = \arg\max_{\boldsymbol{\theta}_{\mathsf{N}}} \ell_{\mathsf{LN}}(\boldsymbol{\theta}_{\mathsf{N}} \mid \boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}_{\mathsf{N}}} \ell_{\mathsf{N}}(\boldsymbol{\theta}_{\mathsf{N}} \mid \log \boldsymbol{x}) \tag{9}$$

これらの結果から，赤池情報量規準に関して以下のことが成り立つ：

$$\begin{aligned} \mathrm{AIC}_{\mathsf{LN}} &:= -2\ell_{\mathsf{LN}}(\hat{\boldsymbol{\theta}}_{\mathsf{N}} \mid \boldsymbol{x}) + 2 \dim \boldsymbol{\theta}_{\mathsf{N}} \\ &= -2\ell_{\mathsf{N}}(\hat{\boldsymbol{\theta}}_{\mathsf{N}} \mid \log \boldsymbol{x}) + 2 \sum_{i=1}^{n} \log x_i + 2 \dim \boldsymbol{\theta}_{\mathsf{N}} \\ &= \mathrm{AIC}_{\mathsf{N}}(\log \boldsymbol{x}) + 2 \sum_{i=1}^{n} \log x_i \end{aligned} \tag{10}$$

ここで，$\dim \boldsymbol{\theta}_{\mathsf{N}}(=2)$ は，母数ベクトル $\boldsymbol{\theta}_{\mathsf{N}}$ の次元（dimension）である．

付録 B　非対称分布

B.1　非対称正規分布

非対称正規分布の定義は以下のように与えられる：

定義 1（非対称正規分布）確率変数 X が p.d.f.

$$f_{\mathsf{SN}}(x\,|\,\boldsymbol{\theta}) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\frac{x-\xi}{\omega}\right),\ x\in\mathbb{R}=(-\infty,\ \infty) \tag{11}$$

をもつとき，確率変数 X は非対称正規分布 $\mathsf{SN}(\xi,\ \omega^2,\ \alpha)$ に従うと呼ばれ，

$$X\sim\mathsf{SN}(\xi,\ \omega^2,\ \alpha)$$

と書かれる．ここで，

$$\xi\in\mathbb{R},\quad \omega\in\mathbb{R}^+,\quad \alpha\in\mathbb{R}$$

は未知母数であり，$\boldsymbol{\theta}=[\xi,\ \omega,\ \alpha]'$ は母数ベクトルである．また，

$$\phi(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right),\quad \Phi(z)=\int_{-\infty}^{z}\phi(x)dx\quad (z\in\mathbb{R})$$

は，それぞれ，標準正規分布 $\mathsf{N}(0,\ 1)$ の p.d.f. と c.d.f.（累積分布関数）である．なお，$(\xi,\ \omega^2,\ \alpha)$ は直接母数（direct parameters）と呼ばれる．

B.2　対数非対称正規分布

確率変数 X に対して，その対数 $\log(X)$ が非対称正規分布 $\mathsf{SN}(\xi,\ \omega^2,\ \alpha)$ に従うとき，X は対数非対称正規分布 $\mathsf{LSN}(\xi,\ \omega^2,\ \alpha)$ に従うといわれる．

$$X\sim\mathsf{LSN}(\xi,\ \omega^2,\ \alpha)\overset{\text{def.}}{\iff}\log(X)\sim\mathsf{SN}(\xi,\ \omega^2,\ \alpha)$$

（Azzalini and Capitanio（2014）の p. 53 を参照のこと．）

対数非対称正規分布の p.d.f. は以下のように与えられる：

$$f_{\mathsf{LSN}}(x\,|\,\boldsymbol{\theta}) = \frac{2}{\omega x}\phi\left(\frac{\log(x)-\xi}{\omega}\right)\Phi\left(\alpha\frac{\log(x)-\xi}{\omega}\right),\ x\in\mathbb{R}^+ \tag{12}$$

B.3　対数非対称正規分布の対数尤度と非対称正規分布の対数尤度の関係

ここで扱う対数尤度の関係は，A.3 で考察した，対数正規分布の対数尤度

と正規分布の対数尤度の関係と同様の結果が，対数非対称正規分布と非対称正規分布の間にも成り立つことであり，ほぼ，同様の議論展開であるが，結果をたどることにする．

まず，対数非対称正規分布と非対称正規分布の p.d.f. には以下の関係が成り立つ：

$$f_{\mathrm{LSN}}(x\,|\,\boldsymbol{\theta}) = f_{\mathrm{SN}}(\log(x)\,|\,\boldsymbol{\theta})\frac{1}{x} \tag{13}$$

この関係を利用すると，無作為標本 $\{X_1, \ldots, X_n\}$ が対数非対称正規分布 $\mathsf{LSN}(\xi, \omega^2, \alpha)$ に従うとき，その同時確率密度関数が，

$$\prod_{i=1}^{n} f_{\mathrm{LSN}}(x_i\,|\,\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{\mathrm{SN}}(\log(x_i)\,|\,\boldsymbol{\theta})\frac{1}{x_i} \tag{14}$$

となることがわかる．

さらに，この結果から，対数非対称正規分布に従う無作為標本 $\{X_1, \ldots, X_n\}$ にもとづく対数尤度は，

$$\begin{aligned}\ell_{\mathrm{LSN}}(\boldsymbol{\theta}\,|\,\boldsymbol{x}) &= \log \prod_{i=1}^{n} f_{\mathrm{LSN}}(x_i\,|\,\boldsymbol{\theta}) = \log \prod_{i=1}^{n} f_{\mathrm{SN}}(\log(x_i)\,|\,\boldsymbol{\theta})\frac{1}{x_i} \\ &= \ell_{\mathrm{SN}}(\boldsymbol{\theta}\,|\,\log \boldsymbol{x}) - \sum_{i=1}^{n} \log x_i \end{aligned} \tag{15}$$

となる．これは，対数非対称正規分布に従う無作為標本 $\{X_1, \ldots, X_n\}$ にもとづく対数尤度 $\ell_{\mathrm{LSN}}(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ と対数変換された無作為標本 $\{\log(X_1), \ldots, \log(X_n)\}$ が非対称正規分布に従う場合の対数尤度 $\ell_{\mathrm{SN}}(\boldsymbol{\theta}\,|\,\log(\boldsymbol{x}))$ の間の関係を表している．

このことから，

$$\max_{\boldsymbol{\theta}} \ell_{\mathrm{LSN}}(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = \max_{\boldsymbol{\theta}} \ell_{\mathrm{SN}}(\boldsymbol{\theta}\,|\,\log \boldsymbol{x}) - \sum_{i=1}^{n} \log x_i \tag{16}$$

が成り立つので，結局，それぞれの最尤推定値（ベクトル）は一致する：

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_{\mathrm{LSN}}(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \ell_{\mathrm{SN}}(\boldsymbol{\theta}\,|\,\log \boldsymbol{x}) \tag{17}$$

これらの結果から，赤池情報量規準に関して以下のことが成り立つ：

$$\begin{aligned}\mathrm{AIC}_{\mathsf{LSN}} &:= -2\ell_{\mathsf{LSN}}(\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{x}) + 2\dim\boldsymbol{\theta} \\ &= -2\ell_{\mathsf{SN}}(\hat{\boldsymbol{\theta}}\,|\,\log\boldsymbol{x}) + 2\sum_{i=1}^{n}\log x_i + 2\dim\boldsymbol{\theta} \\ &= \mathrm{AIC}_{\mathsf{SN}}(\log\boldsymbol{x}) + 2\sum_{i=1}^{n}\log x_i \end{aligned} \tag{18}$$

B.4 非対称ティー分布

定義 2（非対称ティー分布）確率変数 X が p.d.f.：

$$f_{\mathsf{ST}}(x\,|\,\boldsymbol{\theta}) = \frac{2}{\omega} f_{\mathrm{t}}\left(\frac{x-\xi}{\omega}\,\middle|\,\nu\right) F_{\mathrm{t}}\left(\alpha\frac{x-\xi}{\omega}\sqrt{\frac{\nu+1}{\left(\frac{x-\xi}{\omega}\right)^2+\nu}}\,\middle|\,\nu+1\right),\quad x\in\mathbb{R} \tag{19}$$

をもつとき，確率変数 X は非対称ティー分布 $\mathsf{ST}(\xi,\ \omega^2,\ \alpha,\ \nu)$ に従うと呼ばれ，

$$X\sim\mathsf{ST}(\xi,\ \omega^2,\ \alpha,\ \nu)$$

と書かれる．ここで，

$$\xi\in\mathbb{R},\quad \omega\in\mathbb{R}^+,\quad \alpha\in\mathbb{R},\quad \nu\in\mathbb{R}^+$$

は未知母数であり，$\boldsymbol{\theta}=[\xi,\ \omega,\ \alpha,\ \nu]'$ は母数ベクトルである．

$$f_{\mathrm{t}}(z\,|\,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+\frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}},\quad F_{\mathrm{t}}(z\,|\,\nu) = \int_{-\infty}^{z} f_{\mathrm{t}}(x\,|\,\nu)\,\mathrm{d}x$$

は，それぞれ，自由度 ν のティー分布の p.d.f. と c.d.f. である．なお，$(\xi,\ \omega^2,\ \alpha,\ \nu)$ は直接母数と呼ばれる．

B.5 対数非対称ティー分布

確率変数 X に対して，その対数 $\log(Y)$ が非対称ティー分布 $\mathsf{ST}(\xi,\ \omega^2,\ \alpha,\ \nu)$ に従うとき，X は対数非対称ティー分布 $\mathsf{LST}(\xi,\ \omega^2,\ \alpha,\ \nu)$ に従うといわれる．

$$X\sim\mathsf{LST}(\xi,\ \omega^2,\ \alpha,\ \nu)\overset{\text{def.}}{\Longleftrightarrow}\log(X)\sim\mathsf{ST}(\xi,\ \omega^2,\ \alpha,\ \nu)$$

対数非対称ティー分布の p.d.f. は以下のように与えられる：

$$f_{\mathsf{LST}}(x\,|\,\boldsymbol{\theta})=\frac{2}{\omega x}f_{\mathrm{t}}\left(\frac{\log(x)-\xi}{\omega}\middle|\nu\right)F_{\mathrm{t}}\times\left(\alpha\frac{\log(x)-\xi}{\omega}\sqrt{\frac{\nu+1}{\left(\frac{\log(x)-\xi}{\omega}\right)^2+\nu}}\middle|\nu+1\right),\quad x\in\mathbb{R}^+\tag{20}$$

B.6　対数非対称ティー分布の対数尤度と非対称ティー分布の対数尤度の関係

ここで扱う対数尤度の関係も，A.3 で考察した，対数正規分布の対数尤度と正規分布の対数尤度の関係と同様の結果が，対数非対称ティー分布と非対称ティー分布の間にも成り立つことであり，ほぼ，同様の議論展開であるが，結果をたどることにする．

対数非対称ティー分布と非対称ティー分布の p.d.f. には以下の関係が成り立つ：

$$f_{\mathsf{LST}}(x\,|\,\boldsymbol{\theta})=f_{\mathsf{ST}}(\log(x)\,|\,\boldsymbol{\theta})\frac{1}{x}\tag{21}$$

この関係を利用すると，無作為標本 $\{X_1,\ \ldots,\ X_n\}$ が対数非対称ティー分布 $\mathsf{LST}(\xi,\ \omega^2,\ \alpha,\ \nu)$ に従うとき，その同時確率密度関数が，

$$\prod_{i=1}^{n}f_{\mathsf{LST}}(x_i\,|\,\boldsymbol{\theta})=\prod_{i=1}^{n}f_{\mathsf{ST}}(\log(x_i)\,|\,\boldsymbol{\theta})\frac{1}{x_i}\tag{22}$$

となることがわかる．

さらに，この結果から，対数非対称ティー分布に従う無作為標本 $\{X_1,\ \ldots,\ X_n\}$ にもとづく対数尤度は，

$$\ell_{\mathsf{LST}}(\boldsymbol{\theta}\,|\,\boldsymbol{x})=\log\prod_{i=1}^{n}f_{\mathsf{LST}}(x_i\,|\,\boldsymbol{\theta})=\log\prod_{i=1}^{n}f_{\mathsf{ST}}(\log(x_i)\,|\,\boldsymbol{\theta})\frac{1}{x_i}$$

$$= \ell_{\mathrm{ST}}(\boldsymbol{\theta} \mid \log \boldsymbol{x}) - \sum_{i=1}^{n} \log x_i \tag{23}$$

となる．これは，対数非対称ティー分布に従う無作為標本 $\{X_1, \ldots, X_n\}$ にもとづく対数尤度 $\ell_{\mathrm{LST}}(\boldsymbol{\theta} \mid \boldsymbol{x})$ と対数変換された無作為標本 $\{\log(X_1), \ldots, \log(X_n)\}$ が非対称ティー分布に従う場合の対数尤度．$\ell_{\mathrm{ST}}(\boldsymbol{\theta} \mid \log(\boldsymbol{x}))$ の間の関係を表すことに注意しよう

このことから，

$$\max_{\boldsymbol{\theta}} \ell_{\mathrm{LST}}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \max_{\boldsymbol{\theta}} \ell_{\mathrm{ST}}(\boldsymbol{\theta} \mid \log \boldsymbol{x}) - \sum_{i=1}^{n} \log x_i \tag{24}$$

が成り立つので，結局，それぞれの最尤推定値（ベクトル）は一致する：

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_{\mathrm{LST}}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \ell_{\mathrm{ST}}(\boldsymbol{\theta} \mid \log \boldsymbol{x}) \tag{25}$$

これらの結果から，赤池情報量規準に関して以下のことが成り立つ：

$$\begin{aligned} \mathrm{AIC}_{\mathrm{LST}} &:= -2\ell_{\mathrm{LST}}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{x}) + 2\dim \boldsymbol{\theta} \\ &= -2\ell_{\mathrm{ST}}(\hat{\boldsymbol{\theta}} \mid \log \boldsymbol{x}) + 2\sum_{i=1}^{n} \log x_i + 2\dim \boldsymbol{\theta} \\ &= \mathrm{AIC}_{\mathrm{ST}}(\log \boldsymbol{x}) + 2\sum_{i=1}^{n} \log x_i \end{aligned} \tag{26}$$

付録 C　Box-Cox 変換に関する尤度解析

ここでは，Box-Cox 変換に関する尤度解析について述べる．詳細については，Box and Cox (1964) を参照されたい．

まず，無作為標本 $\{X_1, \ldots, X_n\}$ について，$X_i^{(\lambda)} = (X_i^{\lambda} - 1)/\lambda \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$ が成り立つとき，その同時確率密度関数は，

$$\prod_{i=1}^{n} f_{\mathrm{BC}}(x_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f_{\mathsf{N}}(x_i^{(\lambda)} \mid \boldsymbol{\theta}_{\mathsf{N}}) \times x_i^{\lambda - 1} \tag{27}$$

で与えられる．ここで，$\boldsymbol{\theta} := [\boldsymbol{\theta}_{\mathsf{N}}', \lambda]' = [\mu, \sigma^2, \lambda]'$ であり，この結果から，無作為標本 $\{X_1, \ldots, X_n\}$ にもとづく対数尤度は，

$$\begin{aligned}\ell_{\mathrm{BC}}(\boldsymbol{\theta}) &:= \ell_{\mathrm{BC}}(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = \log \prod_{i=1}^{n} f_{\mathrm{N}}(x_i^{(\lambda)}\,|\,\boldsymbol{\theta}_{\mathrm{N}}) \times x_i^{\lambda-1} \\ &= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{(x_i^{(\lambda)}-\mu)^2}{2\sigma^2}\right\} \times x_i^{\lambda-1} \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(x_i^{(\lambda)}-\mu)^2}{2\sigma^2} + (\lambda-1)\sum_{i=1}^{n}\log x_i \\ &= \ell_{\mathrm{N}}(\boldsymbol{\theta}_{\mathrm{N}}\,|\,\boldsymbol{x}^{(\lambda)}) + (\lambda-1)\sum_{i=1}^{n}\log x_i \end{aligned} \tag{28}$$

で与えられる．ここで，

$$\ell_{\mathrm{N}}(\boldsymbol{\theta}_{\mathrm{N}}\,|\,\boldsymbol{x}^{(\lambda)}) := -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(x_i^{(\lambda)}-\mu)^2}{2\sigma^2} =: \ell_{\mathrm{N}}(\boldsymbol{\theta}_{\mathrm{N}}) \tag{29}$$

は，Box-Cox 変換後の無作為標本 $\{X_1^{(\lambda)}, \ldots, X_n^{(\lambda)}\}$ にもとづく対数尤度関数である．

通常の最尤推定法は，対数尤度(28)を母数ベクトル $\boldsymbol{\theta}$ で偏微分したものを $\mathbf{0}$ とおくことによって得られる方程式（尤度方程式）

$$\frac{\partial \ell_{\mathrm{BC}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{30}$$

を，$\boldsymbol{\theta}$ に関して解くことによって与えられる．しかしながら，この場合尤度方程式の計算が複雑になる[13] ことから，以下のような方法によって行われる．

まず，変換母数 λ が与えられたものと考え，(28)式の Box-Cox 変換後の無作為標本 $\{X_1^{(\lambda)}, \ldots, X_n^{(\lambda)}\}$ にもとづく対数尤度関数に関する尤度方程式

$$\frac{\partial \ell_{\mathrm{N}}(\boldsymbol{\theta}_{\mathrm{N}})}{\partial \boldsymbol{\theta}_{\mathrm{N}}} = \begin{bmatrix} \dfrac{\partial \ell_{\mathrm{N}}(\boldsymbol{\theta})}{\partial \mu} \\ \dfrac{\partial \ell_{\mathrm{N}}(\boldsymbol{\theta})}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{31}$$

を考える．ここで，母数 μ に関する方程式は，

$$\frac{\partial \ell_{\mathrm{N}}(\boldsymbol{\theta})}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i^{(\lambda)}-\mu) = 0 \tag{32}$$

より，

13）変換母数 λ の微分に関する計算箇所が複雑になる．

$$\mu=\frac{1}{n}\sum_{i=1}^{n}x_i^{(\lambda)}=:\hat{\mu}(\lambda) \tag{33}$$

という解が得られる．これを，λ が適当に固定されたものとでの μ の最尤推定値と考える．なお，$\hat{\mu}(\lambda)$ は，変換母数 λ に依存するので，厳密には推定値ではないことに注意されたい．また，母数 σ^2 に関する方程式は，

$$\frac{\partial \ell_{\mathrm{N}}(\boldsymbol{\theta})}{\partial \sigma^2}=-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i^{(\lambda)}-\mu)^2=0 \tag{34}$$

より，

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i^{(\lambda)}-\mu)^2 \tag{35}$$

が得られる．上式に(33)式を代入したものを

$$\hat{\sigma}^2(\lambda):=\frac{1}{n}\sum_{i=1}^{n}(x_i^{(\lambda)}-\hat{\mu}(\lambda))^2 \tag{36}$$

とおき，λ が適当に固定されたものとで σ^2 の最尤推定値と考える．

以上の計算によって得られた推定値(33)，(36)から，$\boldsymbol{\theta}_{\mathrm{N}}$ に対する最尤推定値のベクトルは，

$$\hat{\boldsymbol{\theta}}_{\mathrm{N}}(\lambda):=\begin{bmatrix}\hat{\mu}(\lambda)\\ \hat{\sigma}^2(\lambda)\end{bmatrix}=\begin{bmatrix}\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}x_i^{(\lambda)}\\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i^{(\lambda)}-\hat{\mu}(\lambda))^2\end{bmatrix} \tag{37}$$

によって与えられる．

さらに，このベクトルを使って，変換母数 λ 以外の母数ベクトル $\boldsymbol{\theta}$ の推定ベクトル

$$\hat{\boldsymbol{\theta}}(\lambda):=\begin{bmatrix}\hat{\boldsymbol{\theta}}_{\mathrm{N}}(\lambda)\\ \lambda\end{bmatrix}=\begin{bmatrix}\hat{\mu}(\lambda)\\ \hat{\sigma}^2(\lambda)\\ \lambda\end{bmatrix}=\begin{bmatrix}\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}x_i^{(\lambda)}\\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i^{(\lambda)}-\hat{\mu}(\lambda))^2\\ \lambda\end{bmatrix} \tag{38}$$

を構成し，これを対数尤度方程式(28)に代入することによって得られる以下の式をプロファイル対数尤度方程式と呼ぶ：

$$\begin{aligned}\ell_{\text{profile}}(\lambda) &:= \ell_{\text{BC}}(\hat{\boldsymbol{\theta}}(\lambda)\,|\,\boldsymbol{x}) \\ &= \ell_{\text{N}}(\hat{\boldsymbol{\theta}}_{\text{N}}(\lambda)\,|\,\boldsymbol{x}^{(\lambda)}) + (\lambda-1)\sum_{i=1}^{n}\log x_i \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log(\hat{\sigma}^2(\lambda)) + (\lambda-1)\sum_{i=1}^{n}\log x_i \\ &= c - \frac{n}{2}\log(\hat{\sigma}^2(\lambda)) + (\lambda-1)\sum_{i=1}^{n}\log x_i \end{aligned} \tag{39}$$

ここで，定数 $c := -n\log(2\pi)/2 - n/2$ は λ に依存していないため，実際の数値計算を行うときには無視されることがある．

変換母数 λ の推定は，プロファイル対数尤度(39)を適当に選択された λ の値（メッシュ）に対してプロットすることによって最大値を与える値（$\hat{\lambda}$ とおく）を探すことによって行われる．さらに，μ, σ^2 の推定値は，$\hat{\lambda}$ を(33)，(36)式に代入することによって得られ，母数ベクトル $\boldsymbol{\theta}$ に対する最尤推定値ベクトルは以下のように与えられる：

$$\hat{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}}(\hat{\lambda}) := \begin{bmatrix} \hat{\mu}(\hat{\lambda}) \\ \hat{\sigma}^2(\hat{\lambda}) \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^{(\hat{\lambda})} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} (x_i^{(\hat{\lambda})} - \hat{\mu}(\hat{\lambda}))^2 \\ \hat{\lambda} \end{bmatrix} \tag{40}$$

赤池情報量規準については，

$$\begin{aligned}\text{AIC}_{\text{BC}} &:= -2\ell_{\text{BC}}(\hat{\boldsymbol{\theta}}(\hat{\lambda})\,|\,\boldsymbol{x}) + 2\dim\boldsymbol{\theta} \\ &= -2\ell_{\text{N}}(\hat{\boldsymbol{\theta}}_{\text{N}}(\hat{\lambda})\,|\,\boldsymbol{x}^{(\hat{\lambda})}) + 2\dim\boldsymbol{\theta} - 2(\hat{\lambda}-1)\sum_{i=1}^{n}\log x_i \\ &= n\log(2\pi\hat{\sigma}^2(\hat{\lambda})) + n + 2\dim\boldsymbol{\theta} - 2(\hat{\lambda}-1)\sum_{i=1}^{n}\log x_i \end{aligned} \tag{41}$$

となる．ここで，$\dim\boldsymbol{\theta}=3$ であることに注意しよう．

注意 1 ここまでの議論は，無作為標本 $\{X_1, \ldots, X_n\}$ の分布にもとづく最尤推定を考えてきたが，Box-Cox 変換後の無作為標本 $\{X_1^{(\lambda)}, \ldots, X_n^{(\lambda)}\}$ にもとづく対数尤度は，

$$\tilde{\ell}(\boldsymbol{\theta}) := \ell_{\text{N}}(\boldsymbol{\theta}_{\text{N}}) = \ell_{\text{N}}(\boldsymbol{\theta}_{\text{N}}\,|\,\boldsymbol{x}^{(\lambda)})$$

で与えられる（㉙式参照）．よって，本来であれば，Box-Cox 変換後の標本 $\{X_1^{(\lambda)}, \ldots, X_n^{(\lambda)}\}$ にもとづく対数尤度 $\tilde{\ell}(\boldsymbol{\theta})$ を母数ベクトル $\boldsymbol{\theta}=[\mu, \sigma^2, \lambda]'$ に関して最大化する方法によって推定値ベクトルを得る方法が良いと考えられが，尤度方程式が複雑となり，初期値の取り方によっては収束しない等の問題がある．

このことから，無作為標本 $\{X_1, \ldots, X_n\}$ にもとづいて㊵式で推定された $\hat{\boldsymbol{\theta}}$ を利用した場合の，無作為標本 $\{X_1^{(\lambda)}, \ldots, X_n^{(\lambda)}\}$ の情報量規準 AIC は以下の式で計算する方法が考えられる：

$$\begin{aligned}\mathrm{AIC}_{\mathrm{N}}(\boldsymbol{x}^{(\hat{\lambda})}) &:= -2\tilde{\ell}(\hat{\boldsymbol{\theta}}) + 2\dim\boldsymbol{\theta} \\ &= -2\ell_{\mathrm{N}}(\hat{\boldsymbol{\theta}}_{\mathrm{N}} \mid \boldsymbol{x}^{(\hat{\lambda})}) + 2\dim\boldsymbol{\theta} \\ &= n\log(2\pi\hat{\sigma}^2(\hat{\lambda})) + n + 6 \end{aligned} \tag{42}$$

なお，変換される前の無作為標本にもとづく赤池情報量規準㊶との関係は以下のように与えられる：

$$\begin{aligned}\mathrm{AIC}_{\mathrm{BC}} &= -2\ell_{\mathrm{N}}(\hat{\boldsymbol{\theta}}_{\mathrm{N}}(\hat{\lambda}) \mid \boldsymbol{x}^{(\hat{\lambda})}) + 2\dim\boldsymbol{\theta} - 2(\hat{\lambda}-1)\sum_{i=1}^{n}\log x_i \\ &= \mathrm{AIC}_{\mathrm{N}}(\boldsymbol{x}^{(\hat{\lambda})}) - 2(\hat{\lambda}-1)\sum_{i=1}^{n}\log x_i \end{aligned} \tag{43}$$

付録 D　Box-Cox 変換の可視化のための Web アプリケーションの作成

Box-Cox 変換をデータに対して行う際に，変換母数 λ を変化させるにつれて分布の形状がどのように変化するかをインタラクティブに可視化することができれば，どのような母数の値が正規分布に近いかを視認しながら確かめることが可能となる．ここでは，RStudio 社[14] が開発している Shiny[15] を利

14）https://www.rstudio.com/

15）Shiny は R を用いて簡単に Web アプリケーションをつくるための環境である．R パッケージ Shiny として配布されている．詳しくは，https://www.rstudio.com/products/shiny/ を参照されたい．なお，梅津，中野（2018）も丁寧な説明があるので参照されたい．

用することによって，インタラクティブ性をもつ Web アプリケーションを作成する．以下にその手順を与える：

（S1）リスト 1 で与えられるスクリプトを Rmd ファイル InteructiveBoxCox.Rmd[16] として保存

（S2）RStudio[17] で InteructiveBoxCox.Rmd ファイルを開く（図 8）

（S3）▶ Run Documant ボタンをクリック

以上の手順によって，レンダリングが行われ，インタラクティブ機能をもつ専用のウィンドウが開く（図 9）．

Listing 1　インタラクティブ・グラフィックスを作成するための Rmd ファイル：InteructiveBoxCox.Rmd

```
---
title: "Interactive␣Box-Cox␣Visualization"
runtime: shiny
output: html_document
---
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r echo=FALSE}
firmfin <- readRDS("../../0DataSet/firmfinC2018.frame.rds")
colnames(firmfin) <- c("firm", "firmID", "year", "month", "country", "SIC.code", "
    SIC.name", "sales", "employees", "assets.total")
require(dplyr)
require(ggplot2)
firmfin2015 <- firmfin %>%
  filter(year == 2015, sales > 0, employees > 0, assets.total > 0, month == 12)
```

Histogram and Q-Q plot by Box-Cox Trasformed Sales Data
```{r histogram, echo=FALSE}
inputPanel(
  selectInput("n_breaks", label = "Number␣of␣bins:",
              choices = c(10, 20, 30, 40, 50), selected = 20),
  selectInput("year", label = "year:",
              choices = seq(1985, 2016), selected = 2015),
  sliderInput("lambda", label = "lambda:",
```

16）Rmd（RMarkdown）ファイルは Markdown 言語に R のコードを埋め込むことを可能にしたものである．

17）RStudio は，RStudio 社で開発されている R の統合開発環境（Integrated Development Environment: IDE）である．以下の URL を参照のこと：https://www.rstudio.com/products/rstudio/
```

```
              min = -0.1, max = 0.1, value = 0, step = 0.001)
)
sliderValues <- reactive({
  library(e1071)
  library(car)
  tmp <- firmfin %>% filter(year == input$year, sales > 0, employees > 0, assets.
     total > 0, month == 12)
   data.frame(
     Name = c("Skewness",
              "Kurtosis"),
     Value = as.character(c(skewness(bcPower(tmp$sales,lambda = as.numeric(
         input$lambda))),
                            kurtosis(bcPower(tmp$sales, lambda = as.numeric(
                                input$lambda))))),
     stringsAsFactors = FALSE))
  })
  # Show the values in an HTML table ----
renderTable({
    sliderValues()
  })
renderPlot({
  tmp <- firmfin %>% filter(year == input$year, sales > 0, employees > 0, assets.
      total > 0, month == 12)
  tmp %>% ggplot(aes(x = bcPower(sales, lambda = as.numeric(input$lambda)))) +
    geom_histogram(bins = as.numeric(input$n_breaks),aes(y=..density..), fill="
        white", color="black") +
    stat_function(
       fun = dnorm,
       args = list(mean = mean(bcPower(tmp$sales, lambda = as.numeric(input$lambda)
           )),
                   sd = sd(bcPower(tmp$sales, lambda = as.numeric(input$lambda)))),
       lwd = 0.5
    )
  })
```

```{r normalQQplot, echo=FALSE}
renderPlot({
  tmp <- firmfin %>% filter(year == input$year, sales > 0, employees > 0, assets.
      total > 0,  month == 12)
  tmp %>% ggplot(aes(sample = bcPower(sales, lambda = as.numeric(input$lambda))))
      +
   stat_qq() + stat_qq_line()
  })
```

さらに，専用ウィンドウ（図９）の Open in Browser ボタンをクリックすることによって，Web ブラウザが開き，インタラクティブ性をもつグラフィックスを表示することができる．このことは，Box-Cox 変換を売上高のデータに対して，ダイナミックに当てはめ，その結果をリアルタイムに可視化するための Web アプリケーションが作成されたことを表している．

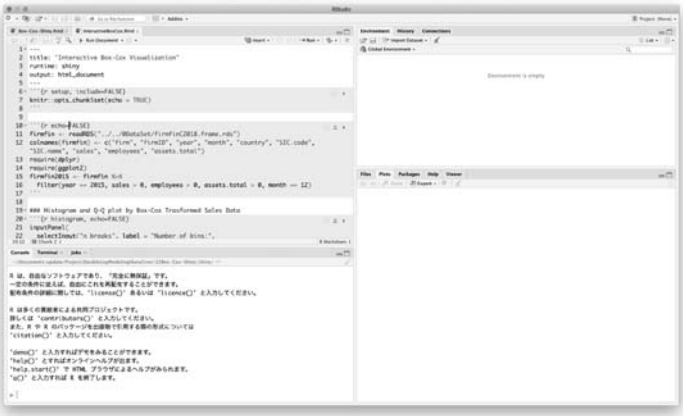

図 8　RStudio：ソースコードペイン上で **Rmd** ファイル（**InteructiveBoxCox.Rmd**）を開いたところ

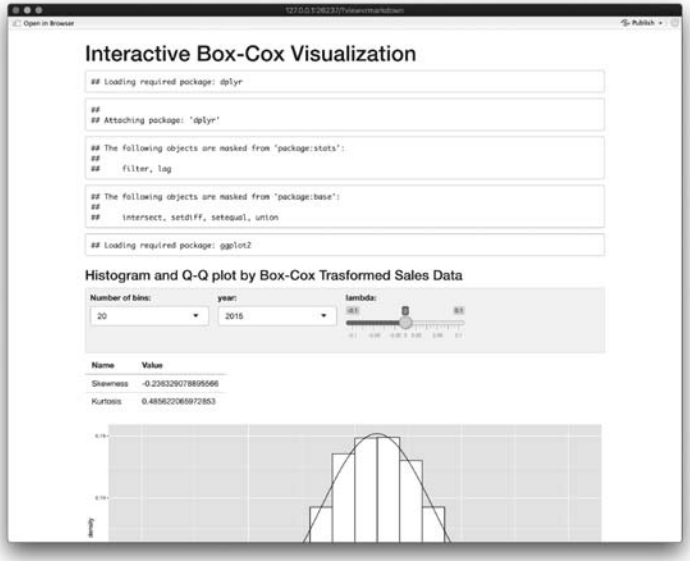

図 9　専用ウィンドウ：インタラクティブ付き

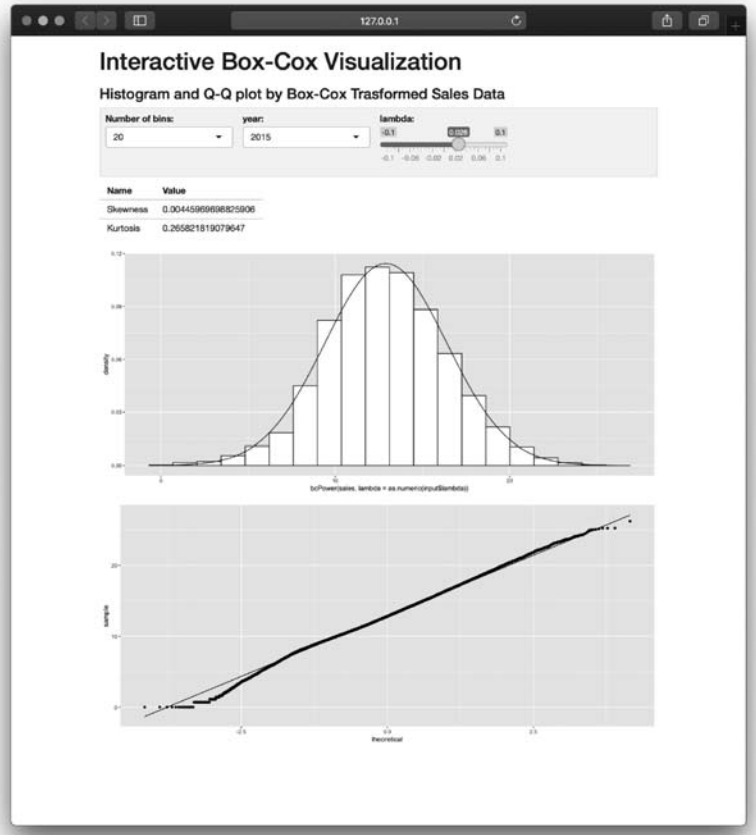

図10 売上高データの Box-Cox 変換に関する Web アプリケーション：ヒストグラムの階級数（**Number of bins**）と年（**year**）をドロップダウンメニューから選択でき，変換母数 λ（**lambda**）はスライダーバーを調整することによって選ぶことができる．これらの選択された値に対するヒストグラム（統計モデル付き）と正規 Q-Q プロットがダイナミックに描画される．また，変換後のデータに対する歪度（**Skewness**）と尖度（**Kurtosis**）の値も表で与えられている．