外国語教育メディア学会 (LET) 関西支部メソドロジー研究部会第 10 号報告論集
草薙邦広 (pp. 1–40)

確率分布から見る外国語教育研究データ

草薙 邦広
広島大学

概要

本稿は，数理的アプローチによる研究実践と教育業務の改善への応用を念頭に置いた上で，観測に対してある確率分布の確率密度関数ないし確率質量関数をフィットさせ，外国語の運用と教育に関連する現象についての優れた数理的近似を得る手続きについて概観する。本稿では，離散確率分布である二項分布，ポアソン分布，幾何分布，負の二項分布，ゼロ過剰ポアソン分布，連続確率分布である正規分布，ガンマ分布，コーシー分布，レイリー分布，ワイブル分布，対数正規分布，指数正規合成分布，一般化極値分布，そして，数パターンの混合分布モデルを取り上げる。また，フィットに関連する手法として最尤推定について，さらに，マルコフ連鎖モンテカルロ法（MCMC）による事後分布のサンプリングについて紹介する。本稿では，全編に渡り，可読性と実用性を重視し，数理的な原理の説明を避け，その代わりに統計解析環境 R による解析コード例を併記した。

Keywords: 数理的アプローチ，確率分布，最尤推定，マルコフ連鎖モンテカルロ法（MCMC），研究方法論

1. 問題の所在

1.1 外国語教育研究における変数の多様化

外国語教育研究は，量的研究を主流とした学際的分野である。ただし，量的研究と一口にいっても，そこにあるアプローチ，姿勢，そして思想などといったものは，学際的ということばでは足りぬほど多岐に渡る。本稿の内容は，そのような多様性について思弁的に論じるものではない。しかしながら，その多様性は，それぞれの研究実践が取り扱うデータの種類にもはっきりと表れる。

2000年頃，第二言語習得研究（SLA）に浸透しつつあった認知主義の影響により，心理学実験において使用される各種の言語行動データが，国内の外国語教育研究に導入された。言語行動データには，判断課題などにおける正答率，反応時間データ，読解時間データ，そして視線計測データなどが含まれる。また，その後，脳神経科学の影響によって，脳機能イメージングデータがこれに加わった。時をほぼ同じくして，コーパス言語学や自然言語処理研究の影響によって，語の出現頻度や共起確率などといったテキストデータも，

有益なデータの一種とみなされるようになった。さらに，心理統計（psychometrics）の影響により，因子分析や項目反応理論といった潜在変数モデル，そして構造方程式モデリング（SEM）が分析方法として普及するようになり，テストや質問紙に対する回答データも，さらに一層重要な研究資源と考えられるように至った。

外国語教育研究の学際化とそれに伴う変数の多様化に関して，国内における 2010 年代の動きは，さらに慌ただしいものである。社会学などの影響により，社会経済的地位（SES）といった調査用変数も，マクロな教育政策的視点に関わるデータとして重要視されるようになってきている。また，高等教育研究や教育工学の影響によって，大量に記録される学習履歴データの利活用についても研究者の注目が集まり（草薙, 2017a），エデュケーショナル・データマイニングといった用語も知られるようになった（草薙・石井, 2016）。同時に，数多くの実践研究によって，教育業務の従事者としての視点から，比較的自由に導出された変数が多数もたらされるようになったことも特筆するべきであろう。

外国語教育研究がますます学際化し，そしてそれが取り扱う変数も同様に多様化していくことは，基本的に望ましいことである。しかし，取り扱う変数が多様化するにつれ，ひとつひとつの変数の数理的特性が十分に吟味されないようになるならば，それはある意味危険でもある。変数が多様化していく現在の状況だからこそ，かえって変数の数理的特性についての吟味が重要となるだろう。

1.2 外国語教育研究における数理的アプローチ

変数の数理的特性について吟味する，そうはいっても，それはかなり抽象的なことである。これを具体化するために，本稿は，観測と確率分布の関係というひとつの観点に着目する。しかし，この観点に触れる前に，本稿が基盤とする「数理的アプローチ」と筆者が呼ぶひとつの研究姿勢について簡単に紹介したい。数理的アプローチは，その実践例の数に比べ，その背景について述べられることが比較的少なかった。

数理的アプローチは，第一に，外国語教育研究の目的論に関して，帰結主義，なかでも功利主義を全面的に受け入れている。これは，外国語教育に関する研究実践の結果として生じる社会的および個人的効用を最大化することを，研究目標とするものである。特に，外国語教育に関する意思決定を媒介することによって，社会および個人に効用をもたらすことを前提としている。つまり，外国語運用に関わる認知機能を完全に解明する，といった自然主義的な課題は，その帰結が個人や社会に大きな効用をもたらすと期待されない限り，主たる研究目的にはならない。このような考え方は，認知科学というよりは，社会学，経済学，工学，または近年流行しているデータサイエンスなどに通じるものがある。

次に，観測が不可能な事象についての合理主義的な推論に対して過度な信頼を置かずに，むしろ不必要な要素を積極的に捨象する，という方針がある。程度の問題に帰すことができるとしても，これは，本質主義や認知主義にある意味反する姿勢である。さらに，

特に外国語運用に関わるなんらかの内的機構といった，直接的な観測が不可能な事象について，それがあたかも計算論的に再現可能であるというように捉えない。これは，行動主義や経験主義に見られる一種の強い態度であるとしても差し支えない。しかしながら，一般的な行動主義者がそうするように，認知主義者が仮定するような機構が，物理的にまたは自然主義的に存在し得ないなどといかなる場面でも含意することはなく，主たる研究の対象外とするだけである。

第三の点になるが，数理的アプローチは，外国語教育研究に見られる諸概念についての共訳不可能性を重く受けとめている。これに鑑みて，しばしば共訳不可能性を招きやすい自然言語による思弁的記述よりも，数理的および記号的な，すなわち形式的記述を優先する。これは，観測とそれに関する形式的記述こそが，研究者間のコミュニケーションにおいてもっとも効率的で，そしてもっとも確実性の高い手段であるという信念による。

このような姿勢は，現在の日本における外国語教育研究では，主流から明らかに外れるものであろう。しかし同時に，統計科学を援用する応用的学術分野のなかでは，もっともありふれた姿勢のひとつであると筆者は考えている。

数理的アプローチにおいて，観測された現象の優れた数理的近似を得ることは，そのもっとも基本的な研究方法であり，そして指針でもある。たとえば，公平なサイコロにおいて，任意の目が出る確率は，離散一様分布（discrete uniform distribution）に従う。離散一様分布という数学的概念は，いわばここではモデルであり，サイコロを振ることそれ自体ではない。ただし，サイコロを振ることの優れた数理的近似となっている。そのため，実際にサイコロを振って出る目を繰り返し観測するのではなく，このモデル，つまり，離散一様分布という概念について考えることによっても，さまざまな予測や意思決定が可能になる。この数理的近似によるさまざまな検証，予測，そしてそれらの帰結になる意思決定を通して，社会や個人はある種の効用を得るだろう。少なくとも，実際にサイコロを振るコストを大幅に削減できるという意味では，最低限の効用をもつことは明白である。

このとき，功利主義の下では，人間がサイコロを振って，ある目が出るという物理的現象に関するすべての要因や，それらの因果をかならずしも考慮しなくてよい，ということに注目されたい。たとえば，サイコロを握る手の形，手の中に配置されたサイコロの状態，重力，風圧，物体間の摩擦，気温，サイコロを投げる者の性格，人徳，そして運と呼ばれるものなどが，実際の現象に強く影響しているとしても，離散一様分布という概念が，実際に優れた数理的近似になっているのであれば，そして実際そのようになっているのであるが，上記の諸要素は積極的に捨象することができる。だからといって，この数理的近似こそが，サイコロを投げるという現象について実在性をもつなどと考えることはない。

また，数理的アプローチでは，このサイコロを振るという現象に関連する因果について，観測を経ずに合理主義的な推論を試みることは基本的にない。その上，「サイコロを投げて出る目は全部同じ」という知見をもってして，「同一確率仮説」や「等確率理論」

といった用語を立てるとか，または「風はサイコロの目が出る確率に影響する」，「性格はサイコロの目が出る確率と独立であるとみなせる」というような自然言語による命題を，整理がつかなくなるほど多数列記していくというようなことは避ける。これは現在，国内外の外国語教育研究において主流の方策に近い。しかし，数理的アプローチはむしろ，「公平なサイコロの目は，$n = 6$ の離散一様分布に従う」という簡潔でわかりやすい記述を好む。これは共訳不可能性を回避し，共訳可能性を担保するための方策である。

外国語教育研究における数理的アプローチの実践は，サイコロの例のように簡単なものではなく，やや複雑である。しかし，数理的アプローチによる国内の応用研究は，近年，筆者やその共同研究者による実践に限っても，多数見られるようになった。たとえば，反応時間や読解時間に対して指数正規合成分布（ex-Gaussian distribution）をフィットさせた Kusanagi（2014），Tamura and Kusanagi（2015a, 2015b），Tamura, Harada, Kato, Hara, and Kusanagi（2016），草薙（2017b），オンライン学習履歴データが従う分布を最尤推定によって検証した草薙（2017a），エッセイライティングにおける増加語数をポアソン分布（Poisson distribution）によってモデル化した川口・室田・後藤・草薙（2016），同種の情報を確率過程のひとつである隠れマルコフモデルによってモデル化した草薙・川口・阪上（to appear），単位時間における最大増加語数を一般化極値分布（generalized extreme value distribution）によってモデル化した草薙（2015）などがある。これらの研究が扱っている確率分布や確率過程などは，確かに一見複雑である。しかし，ある観測に対してその数理的近似を得て，その数理的近似について考察しているということが，サイコロの目の話とまったく同じである。つまり，観測がもつ数理的特性と，その特性を抽象する数理的近似を手がかりとして研究を進める，ということである。いうまでもなく，数理的近似を得ることは目標それ自体ではなく，重要な手段のひとつにすぎない。

このような数理的アプローチによる研究実践のもっとも基本的な方法は，観測に対して確率分布をフィットさせ，その母数を推定することである。この方法こそが，まさに本稿がこれ以降，具体的に紹介していくものである。

1.3 観測に対して確率分布をフィットさせる手続きの利点

得られた観測に対して確率分布をフィットさせる手続きが，数理的アプローチの基本である，といったところで，この方法はなにも本稿で述べるところの数理的アプローチに限ったものではない。一般的にいって，明らかに正規分布に従わない変数に対して，別の確率分布をフィットさせ，その母数を報告する手続きは，記述統計の方法としても，より優れたものである。具体的にいえば，確率分布のフィットは，中心傾向に囚われない意思決定を可能にする。たとえば，ファットテールと呼ばれる裾の重い分布に従う変数の裾のほうの値を予測する場合，適切な確率分布をフィットさせることによって，予測精度が大幅に向上することがある。そして，これこそが重要なことであるが，外国語教育研究では

裾の重い分布形状を見せる変数こそが，比較的多数であると考えられる。

さらに近年は，（a）一般化線形モデル（GLM），（b）一般化線形混合効果モデル（GLMM），（c）階層ベイズモデルなどが，外国語教育研究に導入されるようになってきている。これらのモデルでは，さまざまな確率分布を扱うため，観測に対して確率分布をフィットさせる手続きに親しむことは，上記のモデルを適切に使用するための最初の足がかりになる。そのなかでもベイズ統計に関しては，事前分布の設定について確率分布の知識が不可欠になるのであるから，なおさらである。詳しくは後述するが，一般化線形モデルと最尤推定による母数の推定は，手法上非常に似通っている。たとえば，リンク関数を恒等関数（identity function）とし，誤差がポアソン分布に従うとした切片だけの一般化線形モデルにおいて，切片がもつ回帰係数の値と，最尤推定による母数 λ の推定値は一致する。このように，確率分布のフィットについて理解することは，一般的な意味で解析精度を上げることに関しても，間接的に役立つと考えられる。

2. 確率分布のフィット

2.1 概論

さて，確率分布のフィットとは，観測，つまり所与のデータであるところの確率変数をもっともよく代表する関数とその母数を探す手続きのことである。ここでいう関数は，確率密度関数（probability density function, PDF）や確率質量関数（probability mass function, PMF）を指す。前者は連続確率分布の関数であり，後者は離散確率分布の関数である。データをよく代表する，とは，観測と関数ないしモデルがよく近似している状態である。具体的には，（a）観測と関数による期待値のずれが総合的に見て十分に小さい，（b）データの条件下で，ある関数が十分もっともらしい，といった性質や，確率的なもっともらしさとモデルがもつ複雑さのバランスなど，さまざまな方法によって特徴づけられる。

通常，確率分布をフィットさせる手続きは，（a）確率分布の選択，（b）母数の推定，（c）誤差や適合性などの検討，という3つの工程でおこなわれる（e.g, Ricci, 2005）。以下に，それぞれの工程について概要を述べていく。

2.2 確率分布の選択

最初の工程は，確率分布の選択である。しかし，フィットさせる確率分布を選択する前に，得られた観測の特性を見極めるべきである。分布形状に限っていえば，以下のような点が参考になる。

（a） 連続確率分布か離散確率分布か
（b） 値域は非負か
（c） 分布は対称型か非対称型か
（d） 極端な値はどの程度あるか，負方向に位置するか正方向に位置するか

たとえば，図 1 のようなデータを観測したとする。ここからは，統計解析環境 R（R Core Team, 2016）のコードを併記しながら説明をおこなう。下記のコードを R に入力することで，まったく同じ結果が得られるはずである。

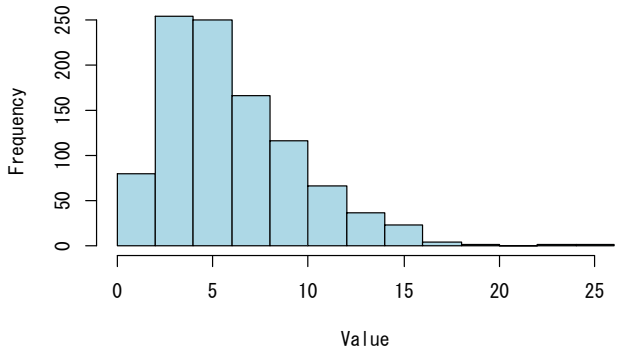

図 1. 観測を表すヒストグラムの例

```
#図１のデータを作成し，ヒストグラムにより可視化
set.seed(0); dat<-rgamma(1000, shape=3,scale=2)
hist(dat,main="",xlab="Value",col="lightblue")
```

この観測は，連続確率分布に属するようであり，さらに非対称型であることが見て取れる。また，極端な値，つまり中心から離れた値は正の方向に偏っていることがわかる。後述するが，このような形状を示すデータは，対数正規分布（log-normal distribution），ガンマ分布（Gamma distribution），ワイブル分布（Weibull distribution），レイリー分布（Rayleigh distribution），指数正規合成分布によって適切にモデル化できる場合が多い。この要領で，あらかじめ観測の特性を考え，単一ないし複数の確率分布を選択しておく。

これらの特性は観測の可視化によって確認することもできるが，データの発生メカニズムについての知識を導入することによって確率分布を選択してもよい。たとえば，単位時間あたりのある事象の発生回数のデータは，ポアソン分布に従うことが知られている。

本稿では，外国語教育研究データと潜在的な関連が深そうな，二項分布（binomial distribution），ポアソン分布，幾何分布（geometric distribution），負の二項分布（negative binomial distribution），ゼロ過剰ポアソン分布（zero-inflated Poisson distribution），正規分布，ガンマ分布，コーシー分布（Cauchy distribution），レイリー分布，ワイブル分布，対数正規分布，指数正規合成分布，一般化極値分布，混合分布モデル（mixture distribution model）を取り上げる。これらの分布の特徴を，上記の観点にあわせて，表 1 にまとめる。この表は，非常に簡略的なものであるため，ときに不正確であるかもしれない。詳しくは蓑谷（2003）など，さまざまな確率分布をカバーする文献を参考にされたい。

表 1.

さまざまな分布の特徴

	タイプ	値域	対称性	中心傾向／極端な値
二項分布	離散	非負	対称	中心に集まる
ポアソン分布	離散	非負	主に対称	中心に集まる
幾何分布	離散	非負	非対称	正に極端な値
負の二項分布	離散	非負	非対称	ほとんどが正に極端な値
ゼロ過剰ポアソン分布	離散	非負	非対称	0 が特殊
正規分布	連続	正負	対称	極端な値は非常に少ない
ガンマ分布	連続	非負	非対称	正に極端な値
コーシー分布	連続	正負	対称	極端な値は少ない
レイリー分布	連続	非負	非対称	正に極端な値
ワイブル分布	連続	非負	非対称	正に極端な値
対数正規分布	連続	非負	主に非対称	正に極端な値
指数正規合成分布	連続	正負	非対称	正に極端な値
一般化極値分布	連続	正負	主に非対称	正に極端な値
混合分布モデル	主に連続	主に正負	主に非対称	場合による

確率分布は，その母数の値によって，形状がかなり変わるものであるが，それぞれの分布における確率密度関数および確率質量関数の典型的な例を，図 2 に示す。このような図によってイメージをもっておくとよい。図 2 に関する R のコードは本稿では省略した。

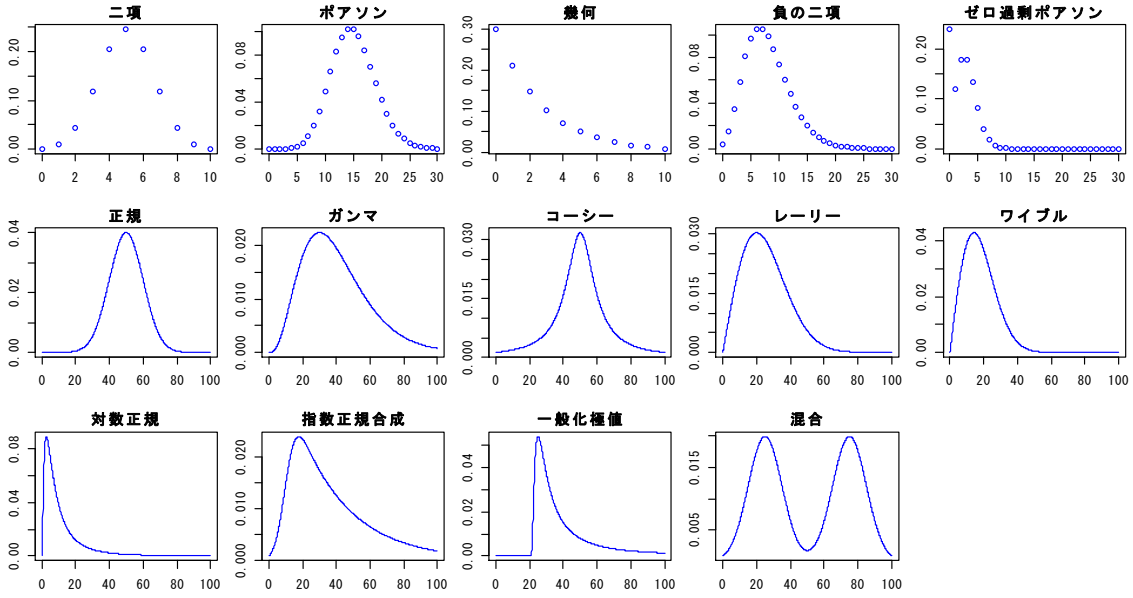

図 2. さまざまな分布のイメージ

当該の研究における先行研究など，過去の研究実践において，すでに使用されている分布を選択することも，十分に有益な方策である。また，その確率分布がもつ特性が，明確に対象とする現象にフィットする場合もある。たとえば，表 2 のような例は，外国語教育研究やその関連分野において，ある程度異論なく使用されると考えられるものである。このような例も参考に，確率分布を選択するとよい。

表 2.
さまざまな分布とその適用例

	適用の例
二項分布	テストの正答回数，反復を伴う課題における成功回数など
ポアソン分布	単位時間におけるエッセイライティングの増加語数，単位時間におけるオンライン教材へのアクセス回数など
幾何分布	一度成功するまで反復を続ける課題において，成功に至るまでの回数など
負の二項分布	複数回成功するまで反復を続ける課題において，成功に至るまでの回数，文中における語や節の数など
ゼロ過剰ポアソン分布	単位時間におけるオンライン教材へのアクセス回数などで，なんらかの原因でアクセスがされない特段の理由があるときなど
正規分布	誤差の分布など
ガンマ分布	オンライン学習プログラムなどにおける学習時間，回答時間，ときにテスト成績，反復を伴う課題に要する時間，資産や年収など
コーシー分布	正規分布よりも明らかに裾が重くない場合
レイリー分布	オンライン学習プログラムなどにおける学習時間，テスト成績
ワイブル分布	オンライン学習プログラムなどにおける学習時間，回答時間，判断課題の反応時間，読解時間など
対数正規分布	オンライン学習プログラムなどにおける学習時間，回答時間，判断課題の反応時間，読解時間，文中における語や節の数，テスト成績，資産や年収など
指数正規合成分布	判断課題の反応時間，読解時間など
一般化極値分布	反復を伴う課題の最高成績，個人が単位時間あたりに書く最大語数など
混合分布モデル	複数の離散的な過程などによる結果が同一変数に混入した場合

2.3 母数の推定

フィットさせる確率分布が決定したら，次は，その確率分布における母数を観測より推定する工程になる。母数とは，一般に θ（シータ）などと表記されるが，これは確率分布を特徴づける値のことである。ガンマ分布を例にあげると，ガンマ分布は 2 つの母数をもつ。ひとつは形状母数である k，もうひとつは尺度母数である θ である。または，形状母数を α，逆尺度母数ないし比率母数を β として扱う場合もある。後者は，ベイジアンに好まれる母数化のようである。本稿では，都合上，両方の場合を使い分けているため，十分に注意されたい。たとえば，ガンマ分布の確率密度関数において，観測にもっとも適合する 2 つの母数の組み合わせを探すことが，母数の推定である。

2.3.1 最尤推定による母数の点推定

非常に大雑把な説明になるが，母数の推定とは，観測と関数による期待値のずれや，観測の条件下における関数のもっともらしさを，最小化・最大化する値の組み合わせを，ほとんどの場合，機械的な計算によってもとめる手続きであると理解しておけばよい。つまり，結局のところ，観測に対して任意の関数における数理的近似の度合いを最大化する母数の値をもとめることである。この手順をフィットともいう。具体的には，モーメント法や最尤推定が使用されることが多い。ここでは，最尤推定を例として概説する。

最尤推定について概説する前に，まずは尤度（ゆうど）について説明する。尤度とは，観測からみた場合の関数の値のもっともらしさである。この値を最大化する方法が，最尤推定である。x を観測を示す確率変数，θ を母数とすると，

$$f(x|\theta) \tag{1}$$

は，母数の下での観測の起こりやすさであるから，これは確率密度関数であるが，逆に x を所与のものとしてみたとき，これは母数のもっともらしさであるともいえる。このことから，一般に，尤度関数を L として書くと

$$f(x|\theta) = L(\theta|x) \tag{2}$$

という関係をもつことがわかる。尤度は，ある母数の条件下において，確率密度関数から与えられるそれぞれの観測の確率における積として計算できる。たとえば，図 3 のような正規分布に従う観測が得られたとする。

```
#図 3 のデータを作成し，ヒストグラムにより可視化
set.seed(0); dat<-rnorm(1000, 50,10)
hist(dat,main="",xlab="Value",col="lightblue")
```

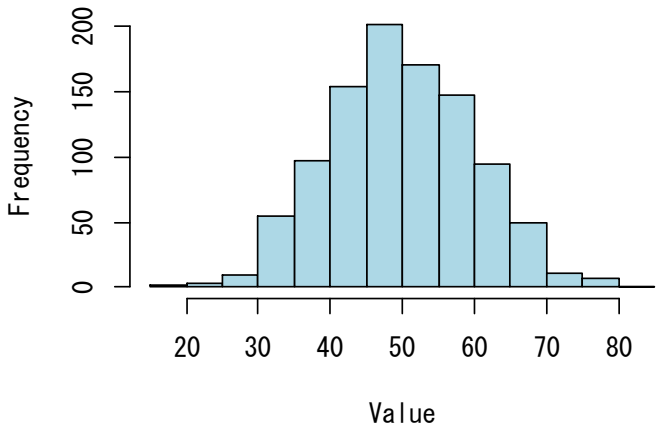

図 3. 正規分布に従う観測の例

この観測 *x* という条件下において，たとえば，母数 θ（μ = 20, σ = 10）の尤度は，以下のように計算できる。ここでは，一般的に行われているように，それぞれの確率の対数の和をもとめる方法を使用している。

```
#尤度を計算
L<-sum(log(dnorm(dat,20,10)))
L
```

このデータセットの値では，-8171.67 である。一方，この観測 *x* という条件下における母数 θ（μ = 50, σ = 10）の尤度は，以下のように計算できる。

```
#尤度を計算
L2<-sum(log(dnorm(dat,50,10)))
L2
```

この母数における尤度は，-3719.16 である。ここでは，母数 θ が，μ = 20, σ = 10 であるよりも，μ = 50, σ = 10 であるほうが，断然もっともらしいと考えることができる。参考として，図 4 に，μ = 20, σ = 10 の確率密度曲線と μ = 50, σ = 10 の確率密度曲線を描き足す。μ = 50, σ = 10 の確率密度曲線のほうが，ヒストグラムにより適合していることが視覚的にもわかるだろう。

```
#確率密度曲線の描画
x<-seq(0,100,.1)
hist(dat,main="",xlab="Value",col="lightblue",freq=F,xlim=c(0,100))
lines(x,dnorm(x,20,10),lwd=2,lty=2,col="pink")
lines(x,dnorm(x,50,10),lwd=2,lty=2,col="lightgreen")
```

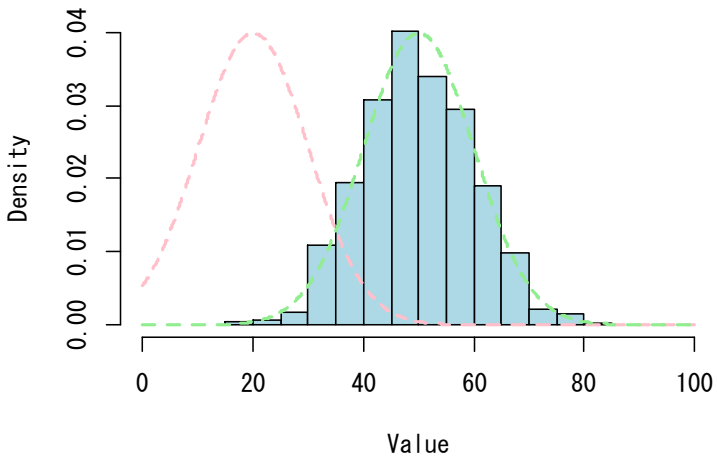

図 4. 正規分布に従う観測と 2 つの確率密度曲線

また，便宜的に，母標準偏差を 10 に固定した母平均の尤度関数は，図 5 のようになる。ここでは，母平均を 0 から 100 までの間について，小数点 2 桁刻みで計算している。

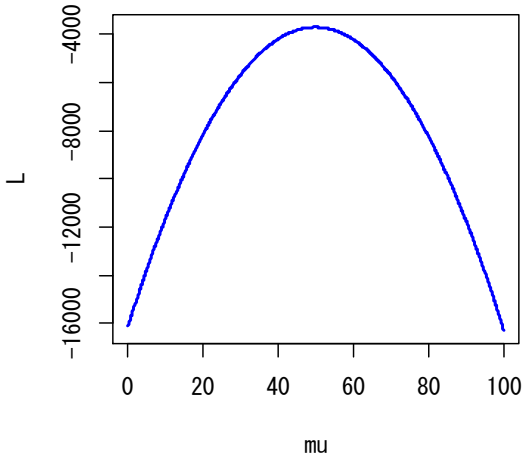

図 5. 母平均についての尤度曲線の例

```
#尤度関数の描画
L<-numeric(1000);mu<-numeric(1000)
    for(i in 1:1000){
    mu[i]<-i/10
    L[i]<-sum(log(dnorm(dat,mu[i],10)))
}

result<-data.frame(mu,L)
plot(result,type="l",lwd=2,col="blue")
```

図 5 からわかるように，母平均が 50 近くであるほうが，そこから離れた値よりも，よりもっともらしいことがわかる。最尤推定とは，このようにして，もっともっともらしい値を探索する方法である。実際の計算については，準ニュートン法（quasi-Newton method），特に BFGS 法やそれに類するものなど，最適化問題の解を見つける反復計算アルゴリズムによって行われる。R では，準ニュートン法をサポートする汎用最適化関数の `optim` 関数などが用意されている。また，そのラッパー関数である `mle` 関数や，bbmle パッケージ（Ben Bolker and R Development Core Team, 2016）の `mle2` 関数などがある。下に，上記のデータについて，正規分布を選択し，母平均と母標準偏差について，`mle` 関数と `mle2` 関数によって最尤推定をするコードを記す。これらの関数では，デフォルトとして BFGS 法を使用している。もちろん，直接 `optim` 関数によって推定することもできるが，本稿ではこれを省略する。

```
#mle による最尤推定の例
eval<-function(mu,sigma){ L<--sum(log(dnorm(dat,mu,sigma)));L}
fit.mle<-mle(eval,start=list(mu=50,sigma=10))
fit.mle

#mle2 による最尤推定の例
library(bbmle)
eval<-function(mu,sigma){ L<--sum(log(dnorm(dat,mu,sigma)));L}
fit.mle2<-mle2(eval,start=list(mu=50,sigma=10))
fit.mle2
```

最尤推定をするときは，基本的に，初期値を指定する必要がある。ここでは，それぞれ母平均を 50，母標準偏差を 10 として指定しているが，初期値の指定については，ノーフ

リーランチ定理のように，これといって万能な方法はないはずである。ただ，モーメント法を先に使用してその推定値を初期値とする，または記述統計を利用する，といった方法で十分な結果が得られることが多い。さらに，後述する MASS パッケージ（Venables, Ripley, 2002）の `fitdistr` 関数や，fitdistrplus パッケージ（Delignette-Muller & Dutang, 2015）の `fitdist` 関数などでは，分布によっては，初期値を自動で設定してくれるため，一般的な使用をするときには，特にこだわらなくてもよい場合も多い。

さて，ここではそれぞれの関数によって，μ = 49.84，σ = 9.98 と点推定された。このときの対数尤度は，-3719.03 であった。これらの推定値は，母数の最尤推定量であるともいえる。この推定値を母数とみなして，当該の観測に確率密度曲線を描き足すと，図 6 のようになる。この母数における確率密度曲線は，観測に対してとてもよい数理的近似になっていることが見て取れる。

```
#確率密度曲線の描画
x<-seq(0,100,.1)
hist(dat,main="",xlab="Value",col="lightblue",freq=F,xlim=c(0,100))
lines(x,dnorm(x,49.84,9.98),lwd=2,lty=2,col="blue")
```

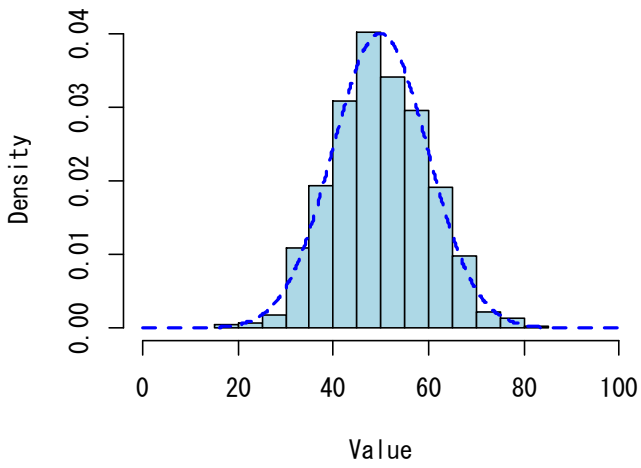

図 6. 正規分布に従う観測と最尤推定した母数の値による確率密度曲線

2.3.2 比較的間便な方法

実際のところ，MASS パッケージの `fitdistr` 関数や，fitdistrplus パッケージの `fitdist` 関数などを使用することによって，尤度関数を自分で用意しなくとも，そしてほとんどの場合初期値を設定しなくても，容易に最尤推定をおこなうことができる。まず，MASS パッケージの `fitdistr` 関数による最尤推定の例を示す。

```
#fitdistr 関数による最尤推定
library(MASS)
fit.fitdistr<-fitdistr(dat,densfun="normal")
coef(fit.fitdistr)
```

また，fitdistrplus パッケージの `fitdist` 関数の例は以下のとおりである。

```
#fitdist 関数による最尤推定
library(fitdistrplus)
fit.fitdist<-fitdist(dat,"norm")
coef(fit.fitdist)
```

`fitdistr` 関数は，ベータ分布（beta distribution），コーシー分布，カイ二乗分布，指数分布，*f* 分布，ガンマ分布，幾何分布，対数正規分布，ロジスティック分布，負の二項分布，正規分布，ポアソン分布，*t* 分布，そしてワイブル分布をサポートしている。また，`fitdist` 関数は，確率密度関数や確率質量関数，そしてそれらの累積分布関数などが与えられていれば，どのような分布についても推定することができる。この関数は，前者の関数とは異なり，モーメント法や最大適合度推定（maximum goodness-of-fit estimation）もサポートしており，さらに誤差や適合度の評価などに関しても，非常に便利な機能と連携させることができる。以降，本稿では，この関数を中心としてコードを記していく。

2.4　誤差や適合性などの検討

母数を推定したら，次は誤差や適合性を検討する必要がある。ここでは，`fitdist` 関数を中心として，母数の推定後の手続きについて概説する。ガンマ分布を例として取り上げる。

```
#数値例の作成
set.seed(0)
dat2<-rgamma(1000,shape=2,rate=1/10)

#最尤推定
fit<-fitdist(dat2,"gamma")
```

ここで生成されたオブジェクト `fit` は，そのまま可視化することができる。これを可視化した様子が図 7 である。

```
#可視化
plot(fit)
```

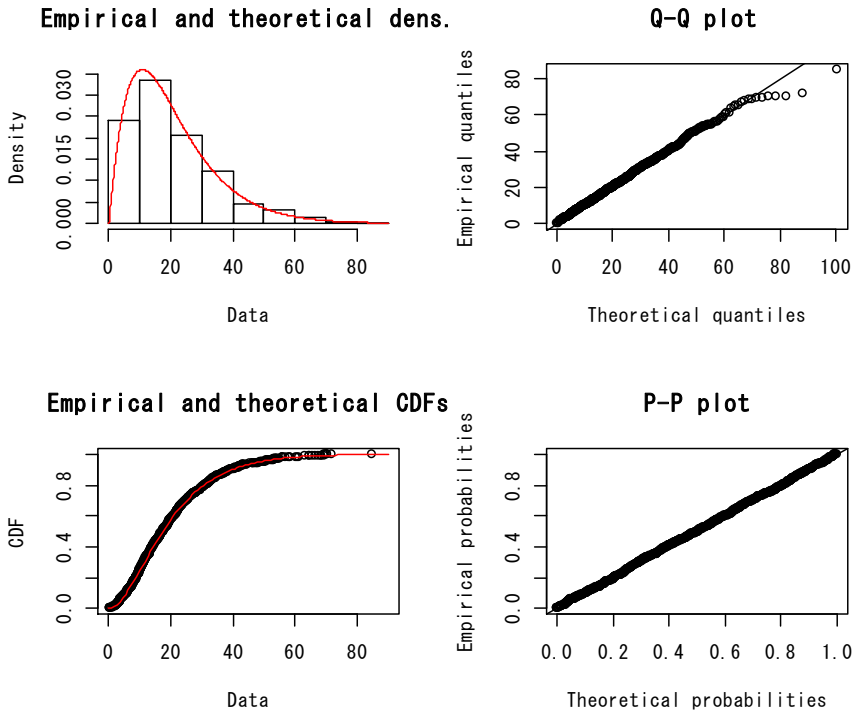

図 7. `fitdistrplus` パッケージによる可視化の例

まず，左上に位置する図は，観測を表すヒストグラムと，推定した母数による確率密度曲線を重ね描きしたものである。右上に位置する図は，Q-Q プロットと呼ばれるものである。これは，理論上の確率密度曲線と観測を比較するために使用される。この図では，横軸が理論上の確率密度曲線によって得られる分位数，縦軸が観測の分位数である。もしも，この確率密度曲線が観測のよい数理的近似になっているのであれば，プロットされる点を結ぶと直線に近くなるはずである。左下に位置する図は，理論上の累積分布関数と経験累積分布関数（ECDF）を重ねて描いたものである。右下の図は，P-P プロットと呼ばれるもので，Q-Q プロットと同じ要領によって，累積確率の理論的期待値と観測の順位による累積確率をプロットしたものである。これらの図によって，視覚的にフィットの度合いを把握することが可能になる。

次に，`fitdist` 関数によって得られるオブジェクトを `summary` 関数に渡すことによって，各母数の点推定値のみならず，その標準誤差，対数尤度，赤池情報量基準（AIC），ベイズ情報量規準（BIC），そして各母数の相関行列などを知ることができる。

```
#要約
summary(fit)
```

標準誤差，対数尤度や各種の情報量基準は，以下のようにすることによっても取り出すことができる。

```
#各情報の取り出し
#母数の標準誤差
fit$sd
#対数尤度
fit$loglik
#AIC
fit$aic
#BIC
fit$bic
```

対数尤度プロット（log-likelihood plot）と呼ばれる機能も便利である。fitdistrplus パッケージでは，`llplot` 関数によって，母数の組み合わせにおける尤度をヒートマップで表現できる。当該の例における対数尤度プロットを，図 8 に示す。

```
#対数尤度プロット
llplot(fit)
```

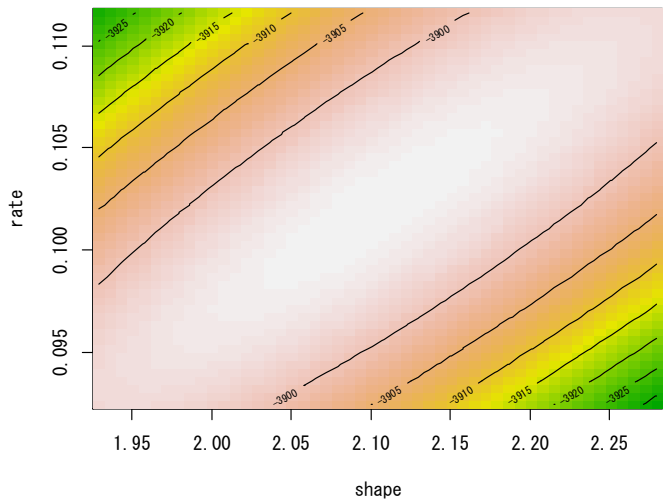

図 8. 対数尤度プロットの例

場合によっては，複数ある母数のうちのひとつ，またはいくつかを固定した状態で，残りの母数の値のみを推定したい場合がある。たとえば，形状母数の値がなんらかの理由

によって既知である状況などがあてはまる。また，それが実質科学的によい方法であるかについては難しい問題であるが，結果的に母数を倹約できるため，適合度の観点において，適切な制約を与えることがよい場合もありえる。いずれにせよ，`fitdist` 関数では，以下のようにして母数を固定した上で，残りの自由母数の値のみを推定できる。

```
#形状母数を 2 に固定して比率母数を推定
fit2<-fitdist(dat,"gamma",fix.arg=list(shape=2))
#対数尤度プロット
llplot(fit2)
```

ちなみに，ひとつの母数のみを推定した場合，対数尤度プロットはヒートマップではなく，図 9 のように尤度曲線を返す。これは，図 5 の場合と同じ要領である。

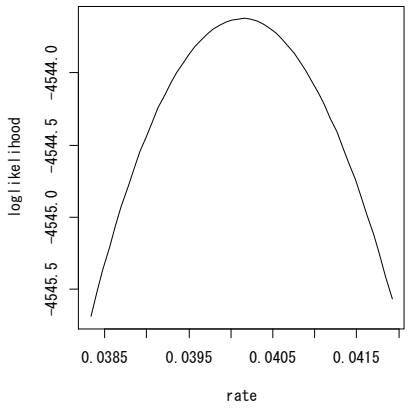

図 9. ひとつの母数を固定した上での尤度曲線の例

さて，誤差や信頼区間の検討には，ブートストラップ法を適用することもできる。外国語教育研究におけるブートストラップ法の概説については，草薙（2014）などがあるため，手法自体についてはこちらを参照されたい。

fitdistrplus パッケージには，`bootdist` 関数という専用の関数があり，パラメトリック・ブートストラップおよびノンパラメトリック・ブートストラップの両方をサポートしている。この関数は，`fitdist` 関数が返すオブジェクトを使用する。ここでは，B = 1,000 としたノンパラメトリック・ブートストラップ法によって，ブートストラップ・パーセンタイル信頼区間を構築する。なお，$\alpha = .05$ とした。

```
#B = 1,000，ノンパラメトリック法でブートストラップ
#環境によっては数分の計算時間がかかる場合もある
boot.fit<-bootdist(fit,bootmethod="nonparam",niter=1000)

#ブートストラップ結果の要約
summary(boot.fit)
boot.fit$CI
```

ブートストラップによって作成されたオブジェクトを summary 関数に渡すと，それぞれの母数について，ブートストラップによって得られた分布の中央値，2.5%点，97.5%点を返す。この 2.5%点と 97.5%点は，それぞれノンパラメトリック・ブートストラップ法によるパーセンタイル信頼区間の下限と上限にあたる。

さらに，以下のようにすると，ブートストラップによって得られた分布，つまりブートストラップ標本を直接的に可視化することができる。これを図 10 に示す。

```
#ブートストラップ推定値
boot.shape<-boot.fit$estim[,1]
boot.rate<-boot.fit$estim[,2]

#これをヒストグラムで描いて，中央値，パーセンタイル信頼区間を描き入れる
par(mfrow=c(1,2))
hist(boot.shape,col="lightblue",main="Shape",xlab="Estimate")
abline(v=quantile(boot.shape,c(0.025,.5,.975)),col=2)
hist(boot.rate,col="lightblue",main="Rate",xlab="Estimate")
abline(v=quantile(boot.rate,c(0.025,.5,.975)),col=2)
```

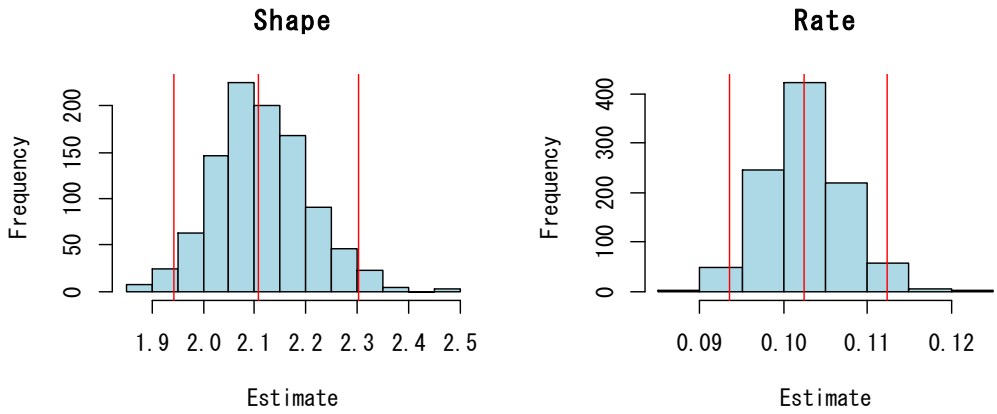

図 10. ブートストラップ標本の可視化

実際の研究実践では，ひとつの観測に対して複数の確率分布をフィットさせ，それら確率分布ごとの統計量や情報量基準を比較することがある。このような比較をするためには，fitdistrplus パッケージの `gofstat` 関数を使用すると便利である。この関数は，(a) コルモゴロフ・スミルノフ検定の統計量，(b) クラメール・フォンミーゼス検定の統計量，(c) アンダーソン・ダーリング検定の統計量，(d) 赤池情報量基準，(e) ベイズ情報量規準をもとめる。ここでは，当該のガンマ分布に従うデータに対して，(a) ガンマ分布，(b) 正規分布，(c) 対数正規分布の 3 つをフィットさせ，それらの適合度を比較する手続きをおこなう。このコードによって，表 3 のような結果が得られる。

```
#それぞれの分布をフィットさせる
gam<-fitdist(dat2,"gamma")
norm<-fitdist(dat2,"norm")
logn<-fitdist(dat2,"lnorm")

#比較（リストで入れる）
gofstat(list(gam, norm, logn), fitnames=c("gamma", "normal",
"lognormal"))
```

表 3.
適合度を示す統計量および情報量基準の比較

統計量および情報量基準	ガンマ分布	正規分布	対数正規分布
コルモゴロフ・スミルノフ統計量	0.02	0.11	0.05
クラメール・フォンミーゼス統計量	0.06	4.06	0.85
アンダーソン・ダーリング統計量	0.41	24.14	5.71
赤池情報量基準	7791.70	8188.35	7883.08
ベイズ情報量規準	7801.51	8198.17	7892.90

これらの指標は，すべて小さい値を取るもののほうが，優れた適合度を示すことになる。ここでは，当然ながら，総合的に見てガンマ分布がもっともよく適合しているといえそうである。

統計的帰無仮説検定の有意性によって，二値判断的に適合度の度合いを調べることは，かならずしもよい方法ではないが，観測に対してモデルが適合しているかどうかを一標本コルモゴロフ・スミルノフ検定などによって判定する場合もある。参考までに，一標本コルモゴロフ・スミルノフ検定に関するコードを示しておく。

```
#正規分布の場合
set.seed(0)
dat.ks.norm<-rnorm(100,0,1)
ks.test(dat.ks.norm,"pnorm")

#特定の平均と標準偏差をもつ正規分布
set.seed(0)
dat.ks.norm2<-rnorm(100,0,1)
ks.test(dat.ks.norm2,"pnorm",0,1)

#ガンマ分布
set.seed(0)
dat.ks.gamma<-rgamma(100,2,3)
ks.test(dat.ks.gamma,"pgamma",2,3)

#ワイブル分布
set.seed(0)
dat.ks.weibull<-rweibull(100,2,3)
ks.test(dat.ks.weibull,"pweibull",2,3)
```

3. 確率分布のフィットにおける実際

以上が，観測に対して確率分布をフィットさせる手続きの基本であった。ここからは，外国語教育研究データへの応用を念頭に置いた上で，その実際について要点を述べていく。

3.1 研究実践における報告の仕方

日本の外国語教育研究において，観測に対して確率分布をフィットさせる手続きは，筆者やその共同研究者による研究実践を中心に数例見られるものの，未だ，業界全体に広く受け入れられている方法とはいいがたい。ここでは，実際の研究実践において，この手法による結果をどのように報告したらよいか，一定の指針を示したい。

まずは，以降どのような処理をするにせよ，得られた観測自体について適切に記述することがもとめられることは変わりない。データの記述については，まずは 4 次程度までのモーメント（平均，分散，歪度，尖度）と，いくつかの分位点を報告するとよいだろう。この節では，形状母数（α）4，比率母数（β）0.1 を母数にもつガンマ分布に従う 1,000 個のデータを例にする。

```
#ここでの数値例の作成
set.seed(1)
dat3<-rgamma(1000,shape=4,rate=1/10)
```

最初はモーメントの計算であるが，ここでは moments パッケージ（Komsta & Novomestky, 2015）を使用する。

```
#モーメントの計算
library(moments);moments.dat<-numeric(0)
moments.dat[1]<-mean(dat3)
moments.dat[2]<-var(dat3)
moments.dat[3]<-skewness(dat3)
moments.dat[4]<-kurtosis(dat3)
moments.dat
```

次に分位点である。ここでは，`quantile` 関数を使用し，最小値，第一四分位数，中央値，第三四分位数，最大値をもとめる。これらの値は五数要約値などともいわれる。

```
#分位数の計算
quantiles.dat<-quantile(dat);quantiles.dat
```

あくまでも例であるが，これらの情報を元に，表 4 と表 5 の要領でそれぞれの値を報告するとよい。

表 4.

観測のモーメント（N = 1,000）

	平均	分散	歪度	尖度
観測	38.94	390.16	0.91	4.06

表 5.

観測の五数要約値（N = 1,000）

	最小値	第一四分位数	中央値	第三四分位数	最大値
観測	2.63	24.34	35.07	50.16	142.58

もちろん，このデータをヒストグラム，カーネル密度曲線，箱ひげ図などによって適宜可視化してもよい。作成した可視化の例を図 11 に示す。また，経験累積分布関数を描くことも有益である。図 12 に経験累積分布関数の例を示す。

```
#画面の分割
par(mfrow=c(3,1))
#ヒストグラム
hist(dat3,main="ヒストグラム",col="lightblue",xlab="Value")
#カーネル密度曲線
x<-seq(0,150,.1)
plot(density(dat3),col="blue",lwd=2,main="カーネル密度曲線
",xlab="Value")
#箱ひげ図
boxplot(dat3,horizontal=T,ylim=c(0,150),col="lightblue",xlab="Value
",main="箱ひげ図")
#経験累積分布関数
par(mfrow=c(1,1))
plot(ecdf(dat3),main="ECDF",col="blue",lwd=3,xlab="Value")
```

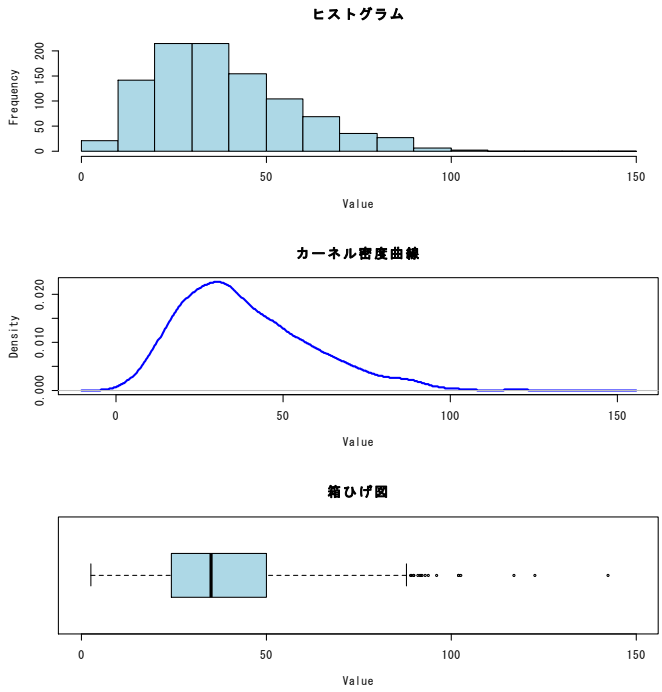

図 11.さまざまな方法による可視化の例

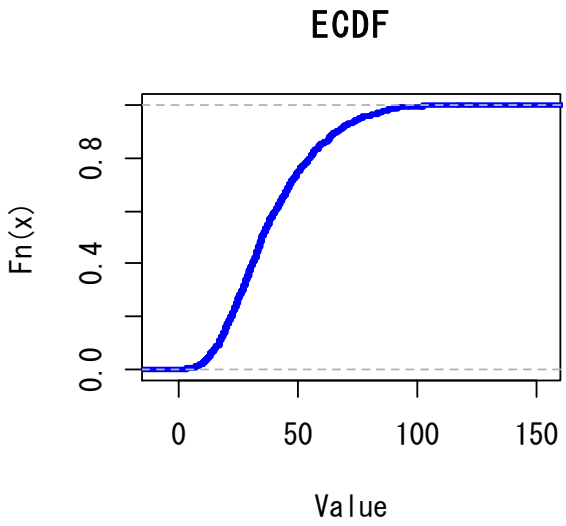

図 12. 経験累積分布関数の例

次に，確率分布を選択する。確率分布の選択については，紙幅が許すならば，それなりにもっともらしい理由を付記すべきである。たとえば，「図 11 を見てわかるように，この観測の分布形状は，明らかに正規分布を逸脱するものであると考えられる。この観測は，正方向に重い裾をもつ連続変数であるから，そのような特性を表現しうる（a）ガンマ分布，（b）対数正規分布，（c）ワイブル分布の 3 つによるフィットを試みることとした」といった文言である。

また，実際にフィットを試みる際には，推定方法や初期値の設定方法について，かならず言及すべきである。例をあげるならば，「それぞれの分布における母数の推定法は，最尤推定であった」，「初期値は $\alpha = 10$，$\beta = 1$ とした」などである。ただし，初期値の設定によって，推定結果が劇的に変わるような場合は，基本的にそのフィットは十分なものではないことが多い。

その後，実際に推定した母数を報告するか，またはその前に，あらかじめ選択したうちの，どの分布がもっともよく観測に適合しているかについて記述する。ここでは，後者を先におこなうと仮定して例をあげる。まずは，上記の通り，当該のデータに対して，ガンマ分布，対数正規分布，ワイブル分布の 3 つをフィットさせる。これらの適合度を表 6 にまとめる。

```
#それぞれの分布をフィットさせる
gam<-fitdist(dat3,"gamma")
lnorm<-fitdist(dat3,"lnorm")
weib<-fitdist(dat3,"weibull")

#比較（リストで入れる）
gofstat(list(gam,lnorm,weib),fitnames=c("gamma","log-
normal","Weibull"))
```

表 6.

当該の例における適合度を示す統計量および情報量基準の比較

統計量および情報量基準	ガンマ分布	対数正規分布	ワイブル分布
コルモゴロフ・スミルノフ統計量	0.02	0.04	0.04
クラメール・フォンミーゼス統計量	0.03	0.46	0.43
アンダーソン・ダーリング統計量	0.23	3.25	2.70
赤池情報量基準	8644.39	8705.13	8670.85
ベイズ情報量規準	8654.21	8714.94	8680.66

このように複数の基準を示した上で，「各種の検定統計量や情報量基準を総合的に評価すると，ガンマ分布が観測に対してもっとも優れた適合を示したと考えられる」というように結論づける。また，観測の分布を示すヒストグラムに，各分布の確率密度曲線を重ね描きすると一見してわかりやすくなる。これは図 13 のようになる。ここでは，青がガンマ分布，赤が対数正規分布，緑がワイブル分布の確率密度曲線である。青のガンマ分布がほかに比べてよい近似になっていることがわかる。もちろん，Q-Q プロットなどを示すことも有益である。

```
#描画
x<-seq(0,150,.1)
hist(dat3,col="lightblue",main="",xlab="Value",freq=F,ylim=c(0,.04)
,breaks=20)
lines(x,dgamma(x,coef(gam)[1],coef(gam)[2]),lwd=2,col="blue")
lines(x,dlnorm(x,coef(lnorm)[1],coef(lnorm)[2]),lwd=2,col="red")
lines(x,dweibull(x,coef(weib)[1],coef(weib)[2]),lwd=2,col="green")
```

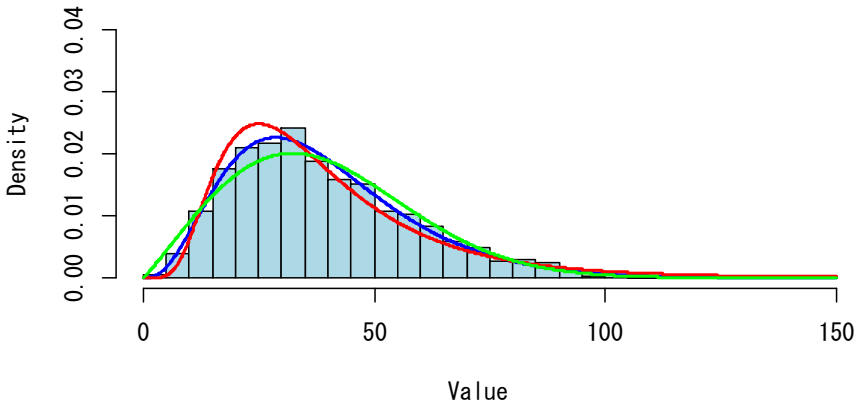

図 13. 複数の確率分布をフィットさせた場合における比較の例

次は，フィットさせた分布についての点推定値，誤差，信頼区間などの報告である。点推定値の報告のみであれば，「観測に対してガンマ分布をフィットさせたところ，その母数の推定値は，α = 3.80，β = 0.10 であった」というような簡潔な表現がよい。また，誤差や信頼区間を報告するのであれば，その方法を，「α = .05，B = 1,000 とし，ノンパラメトリック・ブートストラップ法によってパーセンタイル信頼区間を構築した」というように明示化した上で，点推定値と併せて，「ガンマ分布における母数の推定値は，α = 3.80 [3.51, 4.12] ，β = 0.10 [0.09, 0.11]であった」などと記すとよいだろう。

3.2 さまざまな確率分布をフィットさせるためのコード

ここでは，さまざまな確率分布をフィットさせるためのコード例を，順に示していく。分布をフィットさせる観測用のデータとして，事前にそれぞれの分布に従う擬似乱数を作成している。外国語教育研究を念頭に置いた数値シミュレーションについては，草薙（2016）を参考にされたい。

3.2.1 二項分布

二項分布がもつ母数は，試行回数 n と成功確率 p であるが，成功確率 p の最尤推定量は，成功回数を m としたとき，

$$p = \frac{m}{n} \tag{3}$$

である。このように成功確率 p をもとめること自体は，非常に簡単であるため，ここでは便宜的に当該のコードを省略する。

3.2.2 ポアソン分布

ポアソン分布の母数は，λ である。以下のようなコードでフィットさせる。

```
#ポアソン分布に従う数値例の作成
set.seed(0)
dat.poisson<-rpois(1000,4)
#フィット
fit.poisson<-fitdist(dat.poisson,"pois")
fit.poisson
```

第 1 節にて触れたように，この母数は，一般化線形モデルを使って以下のような要領でもとめることもできる。観測が階層データであるときは，一般化混合効果モデルなどを使って切片の変量効果について推定してもよい。ほかの分布についても，同様の方法によって推定できる場合が多い。

```
#一般化線形モデルにおける切片の推定
fit.poisson2<-glm(dat.poisson~1,family=poisson(identity))
coef(fit.poisson2)
#またはこちらでもよい
fit.poisson2<-glm(dat.poisson~1,family=poisson)
exp(coef(fit.poisson2))
```

3.2.3 幾何分布

幾何分布の母数は，成功確率 p である。以下のようなコードでフィットさせる。

```
#幾何分布に従う数値例の作成
set.seed(0)
dat.geom<-rgeom(1000,.5)
#フィット
fit.geom<-fitdist(dat.geom,"geom")
fit.geom
```

3.2.4 負の二項分布

負の二項分布の母数は，成功回数 r，成功確率 p である。以下のようなコードでフィットさせる。

```
#負の二項分布に従う数値例の作成
set.seed(0)
dat.negbin<-rnegbin(1000,5,.5)
#フィット
fit.negbin<-fitdist(dat.negbin,"nbinom")
fit.negbin
```

3.2.5 ゼロ過剰ポアソン分布

ゼロ過剰ポアソン分布の母数は，λ（または μ）および σ である。以下のようなコードでフィットさせる。ここでは，gamlss パッケージ（Rigby & Stasinopoulos, 2005）を使用している。

```
#ゼロ過剰ポアソン分布に従う数値例の作成
library(gamlss)
set.seed(0)
dat.ZIP<-rZIP(1000,8,.2)
#フィット
fit.ZIP<-gamlss(dat.ZIP~1,family=ZIP)
fit.ZIP
#なお，係数は変換する必要がある
```

3.2.6 正規分布

正規分布の母数は，μ と σ である。必要性がない場合も多いが，以下のようなコードでフィットさせることもできる。

```
#正規分布に従う数値例の作成
set.seed(0)
dat.norm<-rnorm(1000,50,10)
#フィット
fit.norm<-fitdist(dat.norm,"norm")
fit.norm
```

3.2.7 ガンマ分布

ガンマ分布の母数は，本稿で繰り返し述べているように，k と θ または，α と β である。さらにこれも繰り返しになるが，以下のようなコードでフィットさせる。

```
#ガンマ分布に従う数値例の作成
set.seed(0)
dat.gamma<-rgamma(1000,5,.1)
#フィット
fit.gamma<-fitdist(dat.gamma,"gamma")
fit.gamma
```

3.2.8 コーシー分布

コーシー分布は，観測へフィットさせるというよりは，ベイズ統計において事前分布として使用する場合が多い。この分布の母数は，一母数 x_0 と尺度母数 γ である。以下のようなコードでフィットさせる。

```
#コーシー分布に従う数値例の作成
set.seed(0)
dat.cauchy<-rcauchy(1000,0,1)
#フィット
fit.cauchy<-fitdist(dat.cauchy,"cauchy")
fit.cauchy
```

3.2.9 レイリー分布

レイリー分布の母数はひとつのみであり，その母数は σ である。以下のようなコードでフィットさせる。ここでは，VGAM パッケージ（Yee, 2010）を使用している。また，初期値に 4 を入れている。

```
#レイリー分布に従う数値例の作成
library(VGAM);set.seed(0)
dat.ray<-rrayleigh(1000,5)
#フィット
fit.ray<-fitdist(dat.ray,"rayleigh",start=list(4))
fit.ray
```

3.2.10 対数正規分布

対数分布の母数は，対数平均（log μ）と対数標準偏差（log σ）である。以下のようなコードでフィットさせる。

```
#対数正規分布に従う数値例の作成
set.seed(0)
dat.lnorm<-rlnorm(1000,1,10)
#フィット
fit.lnorm<-fitdist(dat.lnorm,"lnorm")
fit.lnorm
```

3.2.11 指数正規合成分布

指数正規合成分布は，おそらく国内の外国語教育研究において，奇しくも正規分布の次に頻繁に使用されている分布であるかもしれない。この分布の母数は μ，σ，そして指数成分である τ である。retimes パッケージ（Massidda, 2013）を使用し，以下のようなコードでフィットさせる。

```
#指数正規合成分布に従う数値例の作成
library(retimes)
set.seed(0)
dat.exgauss<-rexgauss(1000,2000,1000,500)
#フィット
fit.exgauss<-timefit(dat.exgauss)
fit.exgauss
```

3.2.12 一般化極値分布

一般化極値分布は，母数の値によってそれぞれガンベル型，フレシェ型，そしてワイブル型などと分類され，最大値などが従う分布として知られている。ismev パッケージ（Original S functions written by Janet E. Heffernan with R port and R documentation provided by Alec G. Stephenson, 2016）を使用し，以下のようなコードでフィットさせる。

```
#一般化極値分布に従う数値例の作成
library(ismev)
dat.gev<-numeric(1000)
for(i in 1:1000){dat.gev[i]<-max(rnorm(100,50,10))}
#フィット
fit.gev<-gev.fit(dat.gev)
fit.gev$mle
```

3.2.13 混合分布モデル

混合分布モデルは，一般的な確率分布とはやや種類が異なるものである。ただ，単変量の混合正規分布モデルや混合ガンマ分布モデルなどは，使用目的によっては，ほかの分布とまったく同様に扱うこともできる。混合分布モデルの母数の推定には，一般的に EM アルゴリズムと呼ばれる方法が使用される。

まずは，要素数を 2 とした単変量の混合正規分布モデルについて取り扱う。この場合の母数は，混合比 λ，μ_1，μ_2，σ_1，σ_2 の 5 つである。ここでは mixtools パッケージ（Benaglia, Chauveau, Hunter, & Young, 2009）を使い，EM アルゴリズムにて，混合分布モデルをフィットさせるコードを紹介する。

```
#要素数 2 の混合正規分布モデルに従う数値例の作成
library(mixtools)
set.seed(0)
dat.mixnorm2<-c(rnorm(100,50,10),rnorm(200,120,20))

#フィット
fit.mixnorm2<-normalmixEM(dat.mixnorm2)
fit.mixnorm2
```

ちなみに，このモデルの確率密度関数は以下のように定義できる。

```
#要素数 2 の混合正規分布モデルにおける確率密度関数
dnormmix<-function(x,lambda, mu1, mu2, sigma1, sigma2){
    y<-lambda*dnorm(x,mu1,sigma1)+(1-lambda)*dnorm(x,mu2,sigma2)
    y
}
```

次に，要素数を 2 とした混合ガンマ分布モデルについて取り扱う。この場合の母数は，λ，α_1，α_2，β_1，β_2 の 5 つである。先述と同様に，この混合分布モデルをフィットさせるコードを紹介する。

```
#要素数 2 の混合ガンマ分布モデルに従う数値例の作成
set.seed(0)
dat.mixgamma2<-c(rgamma(100,4,1),rgamma(300,10,.2))

#フィット
fit.mixgamma2<-gammamixEM(dat.mixgamma2)
fit.mixgamma2
```

なお，このモデルの確率密度関数は以下のように定義できる。

```
#要素数 2 の混合ガンマ分布モデルにおける確率密度関数
dmixgamma<-function(x,lambda,a1,a2,b1,b2){
    y<-lambda*dgamma(x,a1,b1)+(1-lambda)*dgamma(x,a2,b2)
     y
}
```

3.3 マルコフ連鎖モンテカルロ法

ここまでは，主として，最尤推定による分布母数の点推定について報告してきた。しかし，これは頻度主義にもとづく方法であり，「母数はただひとつの真の値を取る」という見方に依拠している。しかし，ベイズ推定の見方では，頻度主義とは対極的に，母数は確率分布をなすと捉える。後者のほうが，外国語教育における実務的状況に適している場合もあり，解析精度がよいなど，さまざまな面において優れることもある。

ベイズ推定の概略については，本稿の範囲ではないが，ここでは，マルコフ連鎖モンテカルロ法（MCMC）によって，母数の事後分布（posterior distribution）からサンプルを得る方法について紹介する。ベイズ推定や MCMC 自体については，豊田（2015, 2016），松浦（2016）などを参考にされたい。また，昨今は，草薙（2017b），草薙・徳岡（2016），草薙・石井（2016），草薙（to appear）など，ベイズ統計を応用した外国語教育に関する研究実践や学会主催のワークショップなどが見られるようになってきた。

さて，以下のような観測において，正規分布を仮定し，その母平均と母分散についてベイズ推定の要領によって検討したいとする。

```
#正規分布に従うデータ例の作成
set.seed(0)
dat.mcmc1<-rnorm(1000,50,10)
```

R において，母平均値と母分散の事後分布よりサンプルを得るもっとも簡単な方法のひとつは，MCMCpack パッケージ（Martin, Quinn, & Park, 2011）の `MCMCregress` 関数と coda パッケージ（Plummer, Best, Cowles, & VInes, 2006）を使うことであろう。`MCMCregress` 関数は本来，ギブス・サンプリングによって一般線形モデルにおける係数の事後分布をもとめる関数であるが，これを援用して，母平均と母分散の事後分布からサンプリングをおこなうことができる。この関数では，回帰係数の事前分布は多変量正規分布，条件つき誤差分散の事前分布は逆ガンマ分布に設定されている。バーンイン区間を 1,000，その後の反復回数を 10,000 とし，事前分布の形状については，本関数のデフォルトに従う場合，以下のようなコードで MCMC 計算をおこなうことができる。

```
#MCMC の例
library(MCMCpack);library(coda);set.seed(0)
posterior1<-MCMCregress(dat.mcmc1~1,burnin=1000,mcmc=10000)
```

この事後分布のサンプルを要約すると，以下のような情報が得られる。母数のサンプルについての 2.5%点および 97.5%点は，それぞれ 95%ベイズ信用区間の下限と上限とみなせるため，この値を信用区間として論文などで報告するとよい。

```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
            Mean    SD Naive SE Time-series SE
(Intercept) 49.84 0.3139 0.003139      0.003139
sigma2      99.77 4.4784 0.044784      0.046521
2. Quantiles for each variable:
            2.5%   25%   50%    75%  97.5%
(Intercept) 49.23 49.63 49.84  50.05  50.46
sigma2      91.34 96.66 99.67 102.69 108.87
```

MCMC の結果を可視化するために，このオブジェクトをそのまま `plot` 関数に渡すことも便利である。その結果が図 14 である。左側の図は，MCMC 計算における各母数のトレースを表し，右側の図は，事後分布より得た各母数の分布をカーネル密度曲線で表したものである。

```
summary(posterior1)
plot(posterior1)
```

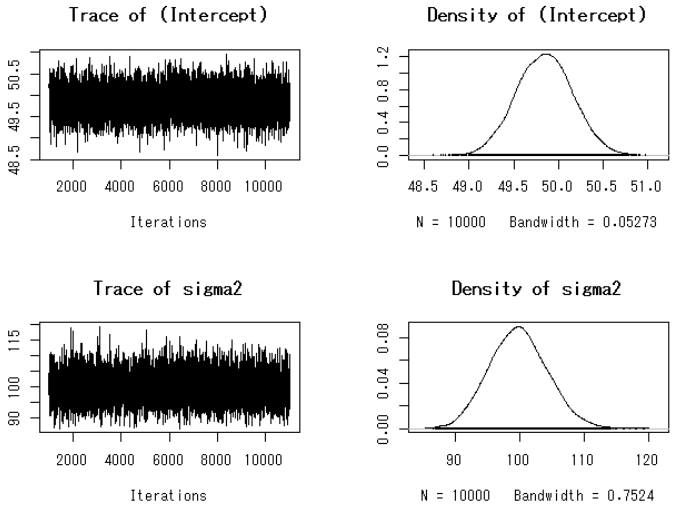

図 14. MCMC における各母数のトレースと事後分布（正規分布の例）

MCMC 計算をおこなう際には，収束診断が不可欠である。ここでは Geweke の収束診断によって判断する。Geweke の収束診断は，マルコフ連鎖における前後の値の差を検討するもので，慣習的に，$[Z] < 1.96$ であるとよいとされる（e.g., Plummer, Best, Cowles, & VInes, 2006）。この例では，両方の母数とも問題なく収束したと判断できる。

```
geweke.diag(posterior1)
```

ポアソン分布をフィットさせる場合も同様に，`MCMCpoisson` 関数を使用して母数 λ における事後分布からサンプルを得ることができる。

```
#ポアソン分布の場合
set.seed(0);dat.mcmc2<-rpois(1000,5)
posterior2<-MCMCpoisson(dat.mcmc2~1,burnin=1000,mcmc=10000)
summary(exp(posterior2))
```

そのほかの分布についても，尤度関数を自分で用意することによって，上の例と同様に，母数の事後分布について検討することができる。ここでは，メトロポリス法による MCMCmetrop1R 関数を使って，指数正規合成分布の母数について検討する。ここでは初期値として正解を設定した。このコードでは，事前分布を無情報事前分布と設定したことになり，バーンイン区間は 1,000，反復回数は 50,000 回である。また，図 15 に MCMC 計算の結果を可視化する。

```
#指数正規合成分布に従う数値例を作成
set.seed(0)
dat.mcmc3<-rexgauss(1000,3000,1000,1000)

#関数を準備
llf<-function(beta,x){
    sum(log(dexgauss(x,beta[1],beta[2],beta[3])))
}

#MCMC 計算
posterior3<-
MCMCmetrop1R(llf,theta.init=c(3000,1000,1000),x=dat.mcmc3,
mcmc=50000,burnin=1000)
plot(posterior3)
```

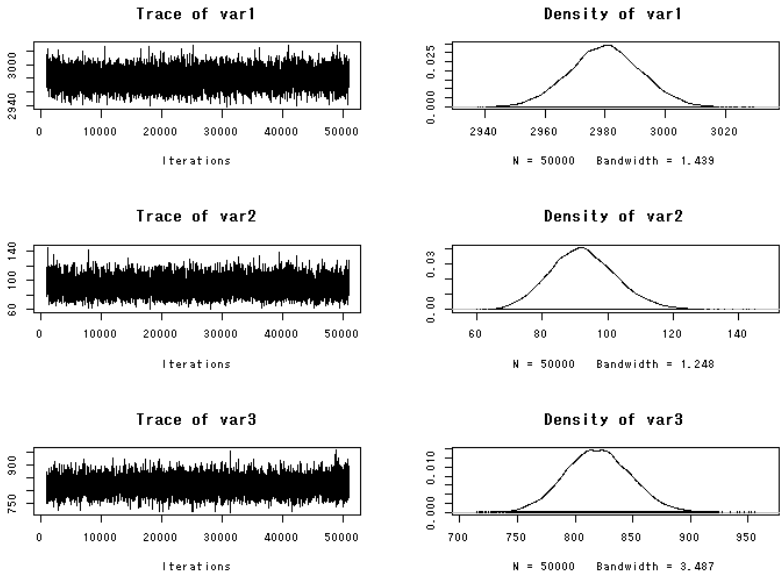

図 15. MCMC における各母数のトレースと事後分布（指数正規合成分布の例）

この例については，その必要性がかなり薄いが，事後分布のサンプルについては，推定する値が多くなるときに箱ひげ図などで示すこともある。それが図 16 である。ただし，これは紙幅を節約できるという点では都合がいいが，それぞれの母数のスケールが大きく離れているときなどは，ときに不親切な図になることに注意が必要である。

```
#箱ひげ図
post.df<-as.data.frame(posterior3)
boxplot(post.df,names=c("mu","sigma","tau"),col="lightblue",
  horizontal=T,xlab="Estimate")
```

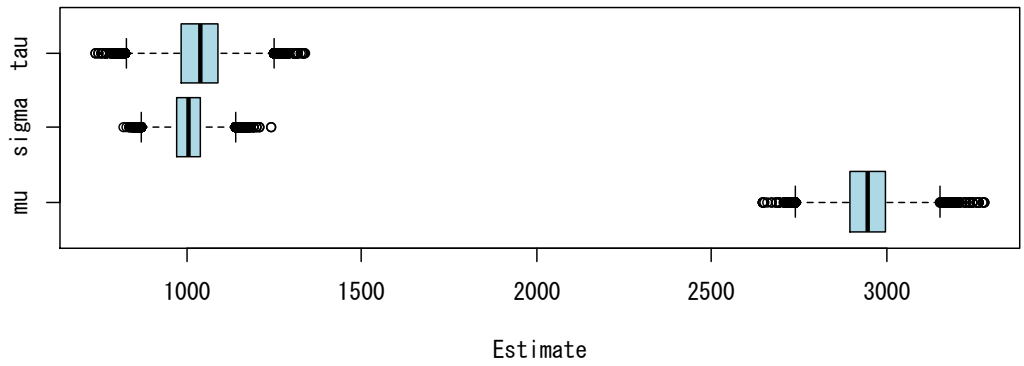

図 16. 各母数の事後分布を示す箱ひげ図

このように，最尤推定だけではなく，一種のベイズ推定の要領によっても，確率分布の母数について柔軟に検討することができる。これはベイズ推定の利点のごく一部にしか過ぎず，むしろ，得られる点推定値や推定区間については，最尤推定やその後のブートストラップ法による信頼区間の構築と大差ない。しかし，ベイズ推定の主な利点は，階層ベイズモデルのような，より複雑なモデルを構築できること，そして母数の事前分布について，研究者が自由に設定できることである。これらの点に関しては別の機会に譲りたい。

3.4 確率分布の推定母数と期待値や分散

観測に対して確率分布をフィットさせる方法は，従来の記述統計に取って代わるものではない。本稿で述べたように，観測を従来のやり方で記述することが重要でなくなることはない。しかし，仮にだが，観測に対してある確率分布が十分にフィットしている場合，その確率分布の推定母数から，期待値や分散を計算できることも理解しておくべきである。たとえば，ガンマ分布に従う変数 X の期待値は，k と θ による母数化のとき，

$$E(X) = k\theta \tag{4}$$

であり，α と β の母数化のときは，

$$E(X) = \frac{\alpha}{\beta} \tag{5}$$

である。また，分散は，k と θ による母数化のとき，

$$V(X) = k\theta^2 \tag{6}$$

であり，α と β の母数化のときは，

$$V(X) = \frac{\alpha}{\beta^2} \tag{7}$$

である。歪度は（8）式，尖度は（9）式のようになる。ここで k は α と読み替えてもよい。

$$skewness = \frac{2}{\sqrt{k}} \tag{8}$$

$$kurtosis = \frac{6}{k} \tag{9}$$

ガンマ分布はもっとも簡単な例であるが，ほかの分布についても同様に期待値などを計算できる。ちなみに，ガンマ分布の母数より，平均，分散，歪度，尖度を計算するコードは以下の通りである。

```
gammadescriptive<-function(shape,scale){
      m<-shape*scale
      v<-shape*(scale^2)
      s<-2/sqrt(shape)
      k<-6/shape
      result<-list("mean"=m,"variance"=v,"skew"=s,"kurtosis"=k)
      result
}
```

このように，観測に対してある確率分布がよく適合しており，さらにその母数が報告されている研究実践については，仮に記述統計が欠落したとしても，事後的にそのおおよその値を計算することが可能である。しかし，平均，分散，歪度，尖度のみから任意の分布の母数を正確に推定することは，けっして容易ではない。

3.5 乱数生成と再現可能性

上記の点に少し関連するが，確率分布をフィットさせる方法は，シミュレーション研究とも強い関連をもっている。観測をよく代表する関数があれば，その関数に従うデータを生成することが可能になるのであるから，ほとんどの場合，観測に対して確率分布をフィットさせる実践は，計算上の再現性を高める行為にもつながる（草薙, 2016）。また，発展的にその関数によって，さまざまなシミュレーション研究が可能になるということも重要な利点のひとつである。外国語教育研究に関するシミュレーション研究は，いまだに幕が開いたとはいえない状況であるが，観測に対して確率分布をフィットさせる実践は，それ自体がもしも瑣末なものであっても，まったく新しい種類の研究を導く可能性もある。

4. 総括

本稿では，外国語教育研究における数理的アプローチについて概説し，外国語の運用と教育に関する現象に対して，その優れた数理的近似を得る具体的な方法の基礎を紹介した。より具体的にいえば，最尤推定によって，観測に対して任意の確率分布をフィットさせ，その母数を推定する方法を中心に，種々のその周辺的技法についても述べた。

本稿が紹介した方法は，いずれも階層性をもたない単変量の場合のみに適用されるものである。しかし，本来の数理的アプローチは，階層性や時系列性などを主な対象とするため，このアプローチ全体から見て，本稿の内容は非常に限定的なものでしかない。

さらに，数理的アプローチは，合理主義的な推論というよりは，確率分布や確率過程といった数理的特性を手がかりとして，経験主義的に研究を進めるものである。これは，観測に確率分布をフィットをさせ，母数の推定値を得る実践のみに終始し，実質科学的な議論をまったくしない，というものではない。むしろ，観測がどのように生成されるか，その仕組みを明らかにすることが，究極的な目標のひとつである。

さて，本稿の最初に述べたことに還るのであるが，昨今，外国語教育研究の学際化に伴い，外国語教育研究に関する変数も多様化している。新しく研究に取り入れられるようになった変数は，もちろん，従来の変数のような数理的特性をもたないかもしれない。また，そもそも数理的特性がまったく不明な変数も多い。これは，研究分野の発展において自然なことである。しかし，そのような比較的新しい変数に対して，従来の正規分布のみに依拠した分析がミスフィットになってしまうことも，容易に予想できることである。学際化が激しく進む今の現状だからこそ，慎重な目をもってデータをつぶさに見つめ，その

数理的特性を丹念に吟味する姿勢こそが必要になるのではないだろうか。本稿がこれまでに紹介した手法やその背景にある考え方は，そのような姿勢をもつ者たちの共通言語になるものだと考えている。

大風呂敷を広げたような本稿のタイトルに反して，外国語教育研究データは，現状においてあまり確率分布という観点からは見られておらず，外国語教育研究データの分布は，大局的に見るとほとんど不明の状態であるといってもよい。本稿が，現実の観測データをまったく扱わなかったことは，まさに本稿の限界点であり，筆者の力不足であるが，これは現状をそのまま表しているともいえる。今後，さまざまな研究対象を専門にする研究者によって，それぞれが扱う具体的な変数の数理的特性についての報告がもたらされ，いずれ，本稿のタイトルのような形によってその成果がまとめられればよい，と筆者は前向きに考えている。そのような試みは，少なくとも筆者の信念の下で，外国語教育研究をこれまで以上に発展させると見込まれている。

参考文献

Ben Bolker and R Development Core Team (2016). *bbmle: Tools for General Maximum Likelihood Estimation*. R package version 1.0.18. https://CRAN.R-project.org/package=bbmle

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, *32(6)*, 1–29.

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, *64(4)*, 1–34.

川口勇作・室田大介・後藤亜希・草薙邦広 (2016). 「エッセイライティングにおける増加語数の時系列推移傾向とエッセイ評価の関係―モデルフィッティングを用いた検討―」*Language Education & Technology*, *52*, 319–343.

Komsta, L, & Novomestky, F. (2015). *moments: Moments, cumulants, skewness, kurtosis and related tests*. R package version 0.14. https://CRAN.R-project.org/package=moments

Kusanagi, K. (2014). Speeded effect on accuracy, sensitivity, response bias and reaction time of L2 learners' grammaticality judgments: Using signal detection theory. *JABAET Journal*, *18*, 37–54.

草薙邦広 (2014).「外国語教育研究におけるブートストラップ法の応用可能性」『外国語教育メディア学会（LET）関西支部メソドロジー研究部会報告論集』*5*, 1–15.

草薙邦広 (2015).「一般化極値分布をもちいた単位時間内における最大語数のモデリング」第 55 回外国語教育メディア学会全国研究大会（公募シンポジウム）．千里ライフサイエンスセンター.

草薙邦広 (2016).「外国語教育研究における数値シミュレーションの基礎―さまざまな分布の下での乱数生成によるデータの復元―」『外国語教育メディア学会（LET）関西

支部メソドロジー研究部会報告論集』*9*, 1–19.

草薙邦広 (2017a).「オンライン学習履歴データの統計的取り扱い」『広島外国語教育研究』*20*, 231–244.

草薙邦広 (2017b).「外国語の読解時における相の強制現象―ベイズ統計によるアプローチ― 『中部地区英語教育学会紀要』*46*, 33–38.

草薙邦広 (to appear).「外国語教育研究者のためのベイズ統計入門」第 57 回外国語教育メディア学会全国研究大会（ワークショップ）. 名城大学.

草薙邦広・石井雄隆 (2016).「量的研究の最前線―ベイズ統計とデータマイニング―」第 42 回全国英語教育学会埼玉研究大会（ワークショップ）. 獨協大学.

草薙邦広・徳岡大 (2016).「外国語における形態統語的鈍感性とそのエビデンス：ベイズ統計学による再検証」『ことばの科学研究』*17*, 61–83.

草薙邦広・川口勇作・阪上辰也 (to appear). 「隠れマルコフモデルによるライティング過程の把握とその形成的評価への援用」第 57 回外国語教育メディア学会全国研究大会. 名城大学.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*. *42(9)*, 1–21.

Massidda, D. (2013). *retimes: Reaction Time Analysis*. R package version 0.1-2. https://CRAN.R-project.org/package=retimes

松浦健太郎 (2016).『Stan と R でベイズ統計モデリング（Wonderful R）』共立出版.

蓑谷千凰彦 (2003).『統計分布ハンドブック』朝倉書店.

Original S functions written by Janet E. Heffernan with R port and R documentation provided by Alec G. Stephenson. (2016). *ismev: An Introduction to Statistical Modeling of Extreme Values*. R package version 1.41. https://CRAN.R-project.org/package=ismev

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*, 7–11.

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ricci, V. (2005). *Fitting distributions with R*. Contributed Documentation available on CRAN.

Rigby R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, *54(30)*. 507–554.

Tamura, Y. & Kusanagi, K. (2015a). Asymmetrical representation in Japanese EFL learners' implicit and explicit knowledge about the countability of normal/material nouns. *Annual Review of English Language Education in Japan*, *26*, 253–268.

Tamura, Y., & Kusanagi, K. (2015b). Measuring Japanese learners' explicit and implicit knowledge of constraints on verb semantics: A case of assertive predicates in English as a foreign

language. *International Journal of Curriculum Development and Practice*, *17*, 25–37.

Tamura, Y., & Harada, Y., Kato, D., Hara, K., & Kusanagi, K. (2016). Unconscious but slowly activated grammatical knowledge of Japanese EFL learners: A case of *tough* movement. *Annual Review of English Language Education in Japan*, *27*, 169–184.

豊田秀樹 (2015).『基礎からのベイズ統計学：ハミルトニアンモンテカルロ法による実践的入門』朝倉書店.

豊田秀樹 (2016).『はじめての統計データ分析―ベイズ的（ポスト *p* 値時代）の統計学―』朝倉書店.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S. Fourth Edition*. Springer, New York.

Yee, T. M. (2010). The VGAM Package for categorical data analysis. *Journal of Statistical Software, 32(10)*, 1–34.