Baayen's Analyzing Linguistic Data

石田 基広 [*]　石田 和枝

2100/01/01

1　言語・文学研究と統計

本書 Analyzing Linguistic Data: A Practical Introduction to Statistics using R, Cambridge, 2008（以下 Baayen (2008) と表記する）は言語を対象データとした統計解析入門書である．また各種国際学会で広く使われているフリーの統計解析環境 R をメインツールとして，その利用方法を詳しく解説していることも本書の特色である．著者の R. Harald Baayen はオランダ出身の心理言語学であり，現在はカナダの Department of Linguistics, University of Alberta で教鞭を執っている[*1]．

Baayen の前著 Word Frequency Distributions,Kluwer Academic Publishers, 2001 （以下 Baayen (2001) と表記）は，計量・計算言語学の分野では広く知られた記念碑的著作である．ここで Baayen は言語（単語）の確率分布に関して Zipf 以来多数提案されてきた仮説を批判的に検証し，独創的な理論を構築している．通読するには相応の数学的知識が必要となるが，およそ言語（単語）を扱う研究者であれば避けて通ることのできない研究書である．

これに対して Baayen (2008) は語学・文学系の研究者を主な読者層とした言語データ解析のための入門書である．本書を通読することで，言語データの確率分布について基本的な知識を習得することができる．従って，研究者は言語データに適切な統計解析手法を選択することができるようになる．さらに実際の解析は，多数の強力な分析手法とグラフィックス作成機能を備えた R を駆使することで，個人パソコン上で実行することができる．これが本書の概要である．

最近では日本でもコーパス研究が浸透し，統計的な解析を応用した論文なども発表されている．しかし残念ながら，統計ソフトを使った結果を安易に掲載している論文も見受けられる．研究者の目的は言語と文学なのだから，何も統計学の基準に合わせる必要はないという反論もあるかもしれない．しかし統計的に問題のあるデータと解析手法から導かれた結論は，一般化できない．

何が問題であるかというと，まず正規分布を仮定した分析が広く行われている．あるいは選択された分析手法から判断する限り，データは正規分布していなければならないのだが，研究者がそのことの意味を深く自覚していないケースがある．要するに，対象とするデータの性質やその分布についての予備考察が欠けていることが非常に多い．

例を挙げよう．言語や文学の計量的研究では，対象が頻度であることが多い．語彙研究では，共起関係を調べるために z 値や t 値，あるいは MI 値が使われている．これらの指標はデータの正規性，あるいはランダム性を前提としている．しかしながら，語彙の選択は文法や文脈に依存しており，これらの仮定は明らかに成立

[*] ishida-m@ias.tokushima-u.ac.jp

[*1] http://www.ualberta.ca/ baayen/

していない．だからといって，これらの指標が言語研究にまったく役立たないとは思わないが，しかし個々の単語の z 値，あるいは t 値の大小を細かく議論するのは意味がないであろう．

またコーパス研究では数万以上のデータを対象に分析が行われる．ところが伝統的な統計検定は小規模データに対しては精緻な分析手法を用意しているが，大規模データに対しては無力なことが多い．データの規模が大きくなると，必ず有意差が出てしまうのである．

また分布の仮定と同じくらい重要なことであるが，語彙頻度の「効果」というのが，固定項なのかランダム項なのかについても議論が必要である（これについては後述する）．Baayen (2008) がこれらの問題すべてについて詳細に論じているわけではないが，少なくとも，言語データの性質や分布を考えるきっかけとなるのは間違いない．理解を深めるためにも Baayen (2001) をあわせて参照することをおすすめしたい．

2 Baayen (2008) の要約

話がやや脱線したが，ここで Baayen (2008) の内容をまとめておこう．第 1 章は R の利用方法である．R はプログラミング言語であるので，データの型や関数という概念を知る必要がある．R はコマンドラインでの操作が中心となる．これは GUI になれたユーザーには苦痛であろう．しかし統計プログラミングでは，ベクトルや行列の計算が重要となる．こうした計算やデータ処理を効率的に行うには，プログラミング技法の取得が欠かせない．

第 2 章はグラフィックス作成が中心話題である．グラフというと，プレゼンのための補助的要素と考えている読者もいるかもしれない．しかし現代のデータ解析では，グラフィックスは分析の手順の一つとして重要な役割を果たしている．それはデータの分布の図形化することで，分布の仮定が正しいか，外れ値があるかをチェックしやすくなるからである．この場合，Q–Q プロットや Box whisker プロットを利用した「診断」が行われるが，Baayen はその方法と理論，また注意点を詳しく述べている．

第 3 章は確率分布の話である．統計的解析では，データの分布を仮定する．分布を考える上で重要なのが，データが離散的か連続的かという区別である．頻度などのデータは一つ二つと数えるものであり，これは離散的なデータである．一方，音声を聞いてから反応するまでの時間であれば連続的である．コーパスから得られるデータのほとんどは前者であろう．しかし Baayen の専門は言語心理学であるため，後者のデータも対象とされる．そのため，Baayen (2008) では二つのデータのそれぞれについて代表的な確率分布を取り上げている．簡単に言えば，前者については二項分布やポアソン分布が，後者については正規分布が論じられている．

第 4 章では基本的な統計理論と解析手法について詳しく述べられている．基本事項なので，一見，目新しい事項はない．しかし R で統計解析を行う場合，関数コードを書くことになるため，分析手法の裏側にある仕組みや理論を自然と意識することになる．特に統計モデルを R で実現する方法や，解析目的であるグループ間の差違を明示するための Treatment Coding などの話題は，一般的な統計入門書で取り上げられることは少ない．しかし，この仕組みを知ることで，統計解析の意味がより鮮明になるであろう．

第 5 章は分類手法であり，日本のコーパス研究者の間でもなじみの話題であろう．かつては相当の計算量を要求した分析手法であるが，今では手もとのパソコンで簡単に実現できる．取り上げられているのは主成分分析や対応分析，クラスター分析と決定木，サポートベクターマシンといったよく知られた手法である．これらの手法はすべて R で実現することができる．しかし，より精度の高い分析を実現するためには，パラメータなどを変えた上で，試行を繰り返すことが欠かせない．また，多数の試行の中から適切な結果を判定することも重要になる．判定はブートストラップやクロスバリデーションによって行われることが多いが，言語系ではあまり積極的に利用されていないようである．Baayen (2008) を通じて，結果の検証の重要さが納得できるであ

ろう．

第 6 章は回帰が扱われている．回帰は言語系では，教育や心理分野をのぞいてあまり利用されていないようである．それは，単回帰モデルが本質的に連続型（あるいは誤差が正規分布に従う）のデータを対象としているからでもある．ただし頻度などの離散型データを回帰するモデルもある．それどころか，現代では分散分析などを含めた解析手法を，「一般化線型モデル」として全体的に考察するのが普通である．この章では，単純な回帰モデルから，ポアソン分布などを対象としたモデルまでが詳しく説明されている．特に後半では，Baayen の代表的な研究業績である語彙の増加曲線をモデル化する手法が取り上げられている．語彙の頻度は一般的な確率分布には従っていない．さらには重要なのは，ある特定のコーパスで語彙頻度を考える場合，そのコーパスではたまたま出現していない語があり，それは別のコーパスであれば出現してもおかしくない．語彙の分布を一般的に考えるのであれば，出現していない語彙についても考慮されなければならない．これは Baayen (2001) で数学的に厳密に議論されているが，Baayen (2008) でも簡単に述べられている．特に語彙の分析を行っている読者には参考になるであろう．

3　言語データの誤差と効果

最後に第 7 章であるが，重要な内容を含む章であるので詳しく論じよう．Baayen は言語心理学の専門家である．言語心理学では，被験者の反応時間，たとえば語を認識するまでの反応時間を計り，その意味を探ろうとする．この場合，誤差と効果をどのように分離するかという問題が生じる．効果とは，たとえば簡単な例を挙げれば，語がカタカナ語か漢字かである．この場合，語という因子に二つの水準，カタカナと漢字があることになる．ところが，カタカナ語として何を選ぶか，あるいは漢字として何を選ぶかは，それぞれの母集団からランダムに選ぶことになる．カタカナ語を特定してしまえば，分析結果は，そのカタカナ語に固有の結果であり，カタカナ語一般に適用できない．あるカタカナ語と別のカタカナ語では反応時間に差があるかもしれないが，この分析で関心があるのは，その二つのカタカナ語の差違ではなく，カタカナ語全体としての効果である．

また被験者によっても反応時間に差があるだろう．この差は，被験者を変えれば，当然変わってくる．分析者に関心があるのは，山田太郎君や加藤一郎君の反応ではない．すなわち，被験者を変えることで生じる誤差を検討しなければならない．さらには，ある被験者はあるカタカナ語へは素早く反応したが，別のカタカナ語への反応は鈍かったという相乗効果（交互作用）もあるだろう．これは被験者が人間の場合に限る問題ではない．たとえば，ある一人作家について調べようとすれば，その作家のテキストごとの誤差についての考慮が必要となるだろう．

本書にならって具体的にみてみよう．Keune et al.: Social, Geographical, and Register Variation in Dutch (2005) では，オランダとフラマン地域で発行されているオランダ語新聞七種のコーパスから，形容詞・副詞語尾 -lijk （英語の -ly，独語の -lich) の頻度が研究されている．ここでの関心は，頻度に地域差があるか，あるいは新聞のジャンル (register) によって異なるか．さらには，そこには交互作用があるかというものである．この場合，地域とジャンルが固定項となる．この二つの効果を確かめる方法はいくつもあるだろう．記述的な目的であれば，本書の前半で紹介される主成分分析やクラスター分析などの手法も有効だろう．

しかし予測あるいは一般化を目的とする場合，効果に有意な差があるかを調べる必要がある．クラスター分析でも，例えばブートストラップやクロスバリデーションの方法を使えないわけではないが，より効率的に検査できる方法が望ましい．

ところで，この問題を従来の枠組で分析しようとするならば，地域とジャンルがカテゴリであるから，これ

を効果とする二元配置分散分析を単純に適用してしまうであろう．それどころか，重回帰分析も検討されるかもしれない．

しかし，この二つの古典的な分析手法を利用すると，どちらの変数にも効果は認められない．Baayen (2008) では，この二つの分析アプローチはそもそも行われていないが，筆者が分析してみた結果を以下に（一部省略して）引用する．

```
summary(writtenVariationLijk.lm)

Call:
lm(formula = Count ~ Country * Register, data = writtenVariationLijk)

Residuals:
    Min      1Q  Median      3Q     Max
-80.088 -52.013 -35.475  -2.541 636.912

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        64.4750    12.0910   5.332 1.41e-07 ***
CountryNetherlands                 18.6125    17.0992   1.088    0.277
RegisterQuality                    -2.8875    17.0992  -0.169    0.866
RegisterRegional                   10.0000    14.8084   0.675    0.500
CountryNetherlands:RegisterQuality  0.7875    24.1819   0.033    0.974
CountryNetherlands:RegisterRegional -16.5250    22.6201  -0.731    0.465
---

Residual standard error: 108.1 on 554 degrees of freedom
Multiple R-squared: 0.0047,Adjusted R-squared: -0.004283
F-statistic: 0.5232 on 5 and 554 DF,  p-value: 0.7588

> summary(writtenVariationLijk.aov)
                 Df  Sum Sq Mean Sq F value Pr(>F)
Country           1   18009   18009  1.5399 0.2152
Register          2    3413    1707  0.1459 0.8643
Country:Register  2    9172    4586  0.3921 0.6758
Residuals       554 6479214   11695
```

念のため出力の意味を説明する．

以下，説明を補う

これを見るとデータの残差がほとんど説明されていない．仮に効果があるとしても（実はあるのだが），それは単純に「誤差」に組み込まれてしまっている．すなわち，このモデルは役に立たない．

ここで分析の目的は，地域とジャンルに効果があるかどうかを知りたいわけであるが，データに含まれる

-lijk で終わる語はそれぞれ固有の使われ方をしている．しかし我々の関心は，ある特定の語と別の特定の語の差にあるわけではない．これらの語は，たまたま抽出されたサンプルに過ぎず，別のコーパスを対象とすれば．選ばれる単語は異なってくる．

そこで，単語がランダムであることを明示したモデルが必要となる．さらに同じ単語でも，地域によって使われ方に違いがあるかもしれない．この二つをランダム項としてモデルに組み込んだ混合モデルの分析が検討されるべきである．

ところで，順序が前後するが，ここで目的変数は頻度であるので，その分布を仮定する必要がある．言語データの頻度の分布については，Baayen (2001) に詳しいので詳細は省略するが，少なくとも語彙頻度が正規分布することは考えられない．語の頻度を近似する分布としては，二項分布が使われることが多い．これは，ある単語が選ばれたか，選ばれなかったかを，壺から玉を取り出すモデルに重ね合わせる考え方である．

このデータの頻度は，各新聞コーパスの最初から 150 万語までの範囲で抽出されたものである．つまり，ある区間内で観測された頻度である．この場合は，ある時間範囲の間に観測された地震や飛行機事故の回数の分布に使われるポアソン分布の利用が検討される．すなわち，ここでの頻度はポアソン分布に従うものと仮定する．この仮定が必ずしも正しくないことは，分析の過程で明らかにされるが，ここでは省略する．

Baayen (2008) が適用したモデルとその結果を引用する（ただし，これは筆者が 2009 年 3 月に R-2.8.1 上で実行した結果であり，Baayen (2008) に掲載の数値と異なる）．

```
> writtenVariationLijk.lmer1
Generalized linear mixed model fit by the Laplace approximation
Formula: Count ~ Country * Register + (1 + Country | Word)
   Data: writtenVariationLijk
  AIC  BIC logLik deviance
 2857 2895  -1419     2839
Random effects:
 Groups Name               Variance Std.Dev. Corr
 Word   (Intercept)        0.87433  0.93505
        CountryNetherlands 0.40269  0.63458  -0.356
Number of obs: 560, groups: Word, 80

Fixed effects:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        3.62081    0.10576   34.24  < 2e-16 ***
CountryNetherlands                 0.28381    0.07421    3.82 0.000131 ***
RegisterQuality                   -0.04582    0.01992   -2.30 0.021446 *
RegisterRegional                   0.14419    0.01667    8.65  < 2e-16 ***
CountryNetherlands:RegisterQuality 0.02022    0.02649    0.76 0.445267
CountryNetherlands:RegisterRegional -0.22598    0.02432   -9.29  < 2e-16 ***
---

Correlation of Fixed Effects:
```

```
            (Intr) CntryN RgstrQ RgstrR CnN:RQ
CntryNthrln -0.369
RegistrQlty -0.092  0.131
RegistrRgnl -0.110  0.157  0.584
CntryNth:RQ  0.069 -0.175 -0.752 -0.439
CntryNth:RR  0.075 -0.191 -0.400 -0.685  0.534
```

ここでも出力を補足しておこう．

以下，説明を補う

結果を見ると，地域とジャンルが頻度に影響を及ぼしているのは明らかである．先に引用した単純な二元配置の分散分析や重回帰を利用した場合と比較されたい．

言語は優れて人間的な現象であり，非常に大きな誤差に加えて，系統的な偏りがある．言語データを対象とした分析を行う場合，少なくとも一般化を目的とするのであれば，誤差や偏りを適切にモデルに組み込んだ分析手法が必要となるだろう．

4 結論

以上のように Baayen (2008) は，単に統計学と R に関する入門書というだけではなく，言語の計量分析の有効な指針を示した本でもある．あえて難をいえば，取り上げられている分析手法がやや古典的なものに限られていることである．残念ながら，最近応用例の増えてきたベイズ分析がほとんど取り上げられていない（混合モデルの検証手法として簡単に紹介されてはいる）．ベイズ分析の特徴は，事前に特定の確率分布を仮定してしまうことなく，データに即した分析を行えることにある．たとえば

言語データのように，確率分布の特定の困難な分野では，今後はベイズ的手法が積極的に用いられるであろう．

また意外に思えるかもしれないが，最近の遺伝子工学分野では，言語データの解析で扱われてきた理論や手法が，遺伝子配列解析のために積極的に導入され，独自の発展も進んでいる（例えば String Kernel を使った手法）．言語は文字の連なりであり，また DNA も塩基の配列という意味で同じである．後者の研究分野では Zipf 以来の確率分布やその分析手法が広く用いられているのである．こうした発展も言語研究に改めてフィードバックされるべきであろう．たとえば

また本書では R での解析に Baayen が独自に開発した languageR パッケージが中心的な役割を果たしている．確かに便利なパッケージであるが，本書で行われる分析を実現するのに必須のパッケージというわけではない．むしろ，ほとんどは R にデフォルトで備わっている関数で十分に実現可能である．また languageR パッケージには，計算のアルゴリズムに疑問を感じる実装部分が一部に認められる（これについては，筆者は Baayen に個人的に指摘している）．さらに R 本体はバージョンアップが頻繁なことで知られているが，その更新に languageR パッケージの機能が十分に追いついておらず，実行結果が Baayen (2008) の掲載内容と異なってしまう部分も多々ある．

そのような点も目につくが，しかし人文系ではデータそのものに基づく厳密な経験的分析と，その結果を踏襲した（飛躍しない）論述が十分には浸透していない．そのような状況を考えると，Baayen (2008) が，人文系，特に言語・文学系の研究パラダイムを大きく変えるきっかけとなることを期待したいところである．

最後になるが，本書と内容あるいは目的の重なる英書が時期を同じくして 2 冊上梓されている．あわせて紹

介しておく.

Keith Johnson: Quantitative Methods In Linguistics, Wiley, 2008

Gries Stephan: Quantitative Corpus Linguistics with R: A Practical Introduction, Routledge, 2008