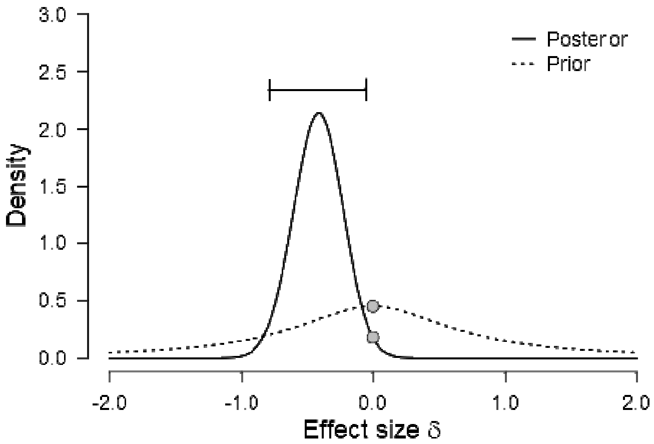
ワークショップ
「ベイズ統計とその外国語教育研究への応用」
配布資料

草薙　邦広（KUSANAGI, Kunihiro）
広島大学外国語教育研究センター

日本言語テスト学会（JLTA）
第 22 回（2019 年度）全国研究大会
2019/9/11
新潟青陵大学



講義 1 ベイズ統計の基本

0. 自己紹介

■**草薙邦広（くさなぎ　くにひろ）**

・外国語教育研究における**数理的アプローチ**

- 言語使用，学習，それらに関する行動の**数理モデリング**，研究方法論全般

1. はじめに

■ベイズ統計について率直なところ Q and A

Q. ベイズって**異端**なんでしょう？マニアックで少数のひとだけがやっているものでしょう？
A. どれだけ控えめにいっても，現在の統計分析や数理モデリング分野の主流派は，ベイズ主義を受け入れており，むしろ多数派だといわれています。

Q. 心理学などの他分野では，**統計革命**などと呼ばれる一大ムーブメントに発展しているように見えるけど，外国語教育研究でも同じように一大ムーブメントになるの？
A. おそらくそうはなりません。これまでの外国語教育研究が，数値化されたデータの記録，記述統計，統計的帰無仮説検定，実験計画法，多変量解析，統計改革などを**漸進的に**，**そして部分的に**受け入れてきたように，まったく同じような過程に沿ってベイズ統計も**ゆっくりと**浸透していくと思われます。

Q. これまでの外国語教育研究が使ってきた統計は**時代遅れ**になり，ベイズ統計ができない研究者の居場所は最終的になくなるの？
A. そうはなりません。**頻度主義（伝統的統計学）**として括られるこれまでの統計分析にも**長所**があり，ベイズ統計はこれらを否定するものではありません。ただし，非常に長期的にみると，現在主流の分析の多くは取って代わられるでしょう。

Q. ベイズの本を読むと**難しい数式ばかり**に見えるけど，一般的な外国語教育研究者にもベイズ統計が本当にできるの？
A. 少なくとも**数学の知識が必要となる**ことは間違いありませんが，ほとんどの統計ユーザーや教育実践者は，確率論や標準偏差の式，*t* 検定の *t* 値の式などを知らないまま研究を進めてきました。これをよいことと認めるわけではありませんが，ベイズ統計が特別というわけではありません。乱暴な言い方ですが，現在は高校数学で確率分布を詳細に扱うようになっており，長期的には世代交代によってこの点が克服されていくと思います。

Q. **敷居が高すぎて手がでない…。**
A. この認識は改めてほしいと思っています。敷居は高くありません。むしろ，どちらかというと「**玄関は広いが，そこからの階段が傾斜だ**」という認識の方が現実的です。

Q. ベイズ統計にすると，どんな**メリット**があるの？
A. これからご紹介しますが，外国語教育研究に関しては，ベイズ統計のメリットというよりは，これまでの方法の**問題点を回避する**，という観点が重要だと思っています。つまり，これまでの借金を返す方法を真剣に我々は考えているのであって，なにかに投資して新たな儲けを得ることについて考えているわけではありません。

Q. ベイズ統計は一枚岩ではないと聞いたけど？
A. ベイズ統計やそれに類する用語の定義にしても，実際の分析の仕方にしても，実に多様です。**確率論に関する態度**（主観的確率）を強調するときもあれば，ネイマン＝ピアソン式の**統計的帰無仮説検定の代替法**としての役割を強調するときもあります。また，MCMC を用いたりする**統計モデルの推定法**（ベイズ推定）としての技術的側面を強調するときもあります。広い意味で,「ベイズ統計はベイズの定理を基盤に置く，さまざまな分析の背景にある共通したロジックの 1 つ」だと理解するとよいでしょう。

Q. ベイズ統計が流行ると，外国語教育研究分野における理論面の発展はどうなる？
A. おそらく，**全面的に進歩が緩やかなものになる**と考えられます。同時に，一部分の理論は，修正を余儀なくされると思います。外国語教育研究においてベイズ統計は，これまでの方法に比べると，**柔軟で**，**いわば控え目な推論につながる**場面の方が多いと予想されます。これはよくいえば，勇み足な結論を積み重ねる研究に歯止めをかける力があるといえます。ベイズ統計は，科学研究上，多くの**可謬主義者**に好まれる考え方をもっています。

2. ベイズ統計の基本

■ベイズの定理の前におさらい

・確率というのは，一種の**約束事**の上に成り立っている（**公理的確率論**）

- 確率が満たすべき最低限の数学的条件を満たしているものは全部確率だと考える
- 素朴な自然言語の意味によって決まるものではない
- 確率をどう与えるか自体は確率であるということの問題ではない

・事象（event）A が起きる確率は P(A)と書く，**事象 A の確率**と読む

- 6 面サイコロにて 6 が出るという事象（事象 A）

$$P(A) = \frac{1}{6} \tag{1}$$

・サイコロを二回投げて，1 回目が 6，2 回目が 1 になる確率は？

- 確率は積をもとめることができる

$$P(A_1)P(A_2) = \frac{1}{36} \tag{2}$$

・ある事象が起きたという条件下で，別の事象が起きる確率を**条件付き確率**という
・P(A|B)は，事象 B が起きた条件下で事象 A が起きる確率
- 数学で　|　このマークは，右の「下での」という意味

・P(A)と P(B)が同時に起きる確率を**合同確率**という: P(A∧B)または P(A,B)
・合同確率は以下のように求められる

$$P(A \wedge B) = P(A|B)P(B) \tag{3}$$

- 事象 B が起きて，さらに事象 B の条件下で事象 A が起きる確率
- このとき，条件を持たない確率 P(B)を**周辺確率**という

$$\textbf{合同確率} = \textbf{条件付き確率} \times \textbf{周辺確率} \tag{4}$$

$$\mathrm{P}(A|B) = \frac{P(A \wedge B)}{P(B)} \tag{5}$$

$$\mathrm{P(B)} = \frac{\mathrm{P(A \wedge B)}}{P(A|B)} \tag{6}$$

■ベイズの定理の導出
・ここで，(3) 式に着目してみる
・(3) 式はこう書いても同じ

$$P(A \wedge B) = P(B|A)P(A) \tag{7}$$

$$P(A \wedge B) = P(A|B)P(B) \qquad (3) = \tag{8}$$

・なので，

$$P(B|A)P(A) = P(A|B)P(B) \tag{9}$$

・これを整理すると，

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{10}$$

・これが**ベイズの定理**

■ベイズの定理とベイズ統計

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \qquad (10) = (11)$$

・A を**データ**（D）と考える
・B を**仮説**（H）と考える，または母数（θ）

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \qquad (12)$$

・P(H|D)　データの下での仮説が正しい確率　**事後確率**という　分布とみなし**事後分布**ともいう
・P(D|H)　仮説の下でデータが得られる確率　**尤度**という
・P(H)　仮説が正しい確率　**事前確率**，**事前分布**という　しばしば**主観的確率**である
・P(D)　　データが得られる確率　しばしば無視できる，**正規化定数**という
・ここから　**事後確率∝尤度×事前確率**　と考えることができる
　- ∝は比例する，と読む

■事後分布

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

・データは既に得られて，もう変わらないのだから**定数**と見る
・**データという条件下**で，仮説が正しい確率，または**母数が分布**していると考える
・有意かどうか，そして母数とその誤差の平均的な大きさ，ではなく，母数の分布を直接考えられる
・母平均が 50 より大きい確率は 50%だろう

	データ	母数
頻度主義	分布	定数
ベイズ統計	定数	分布

・頻度主義　　**母数→確率的にデータ発生**
・ベイズ統計　**データ→確率的に母数を推測**

・{54,58,45,34,46,57,67,45,76}というデータの下での母平均の事後分布（から得られる疑似乱数の分布）

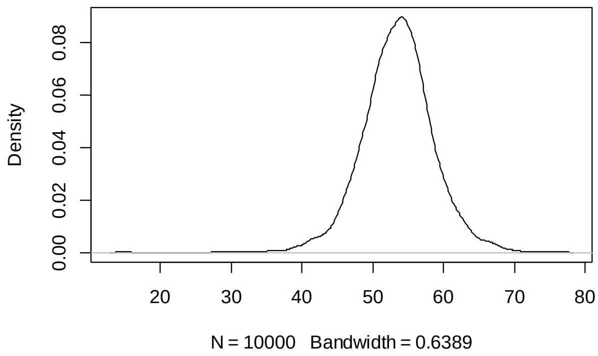


図 1. 観測の下での母数の事後分布のイメージ

■尤度

$$P(H|D) = \frac{\boldsymbol{P(D|H)}P(H)}{P(D)}$$

・ある仮説の下でデータが得られる確率
・こんな観測があったとき（**図 2**），平均 20 という仮説からこのデータが生まれる確率は低いだろう
- 平均 50 なら，このデータが生まれる確率は高そうだ
- それぞれのケースが，ある母数の下で得られる確率の総積（対数の和＝**対数尤度**）

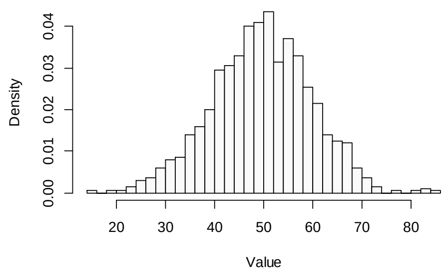


図 2. ある観測のイメージ

・**尤度関数**

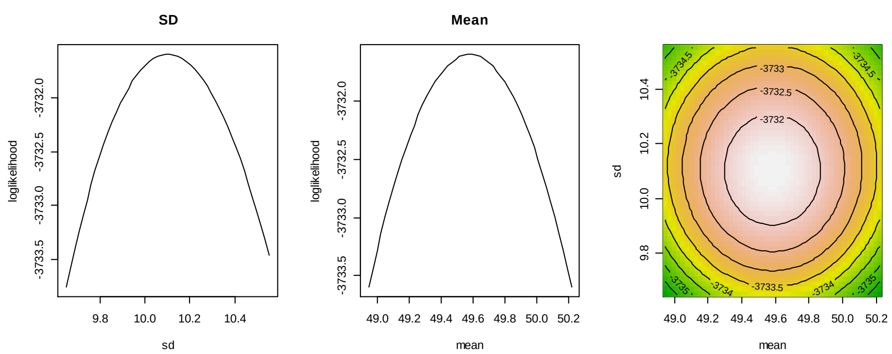


図 3. 尤度関数のイメージ

■事前分布

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

・データが得られる前の仮説への信念（**主観的確率**とも考えられる）
・「こんな分布で母数は分布しているのだろう」←**背景知識**が重要
・背景知識をもたない場合は，**無情報事前分布**や弱情報事前分布を使用する
・主観的とはいっても，取りうる**符号や最小値**，**最大値**を与えることはそれほど客観的でないわけではない

■例：ご好評頂いている**エリカの過去**

高校生のアツシくんは，クラスのマドンナであるエリカちゃんに首ったけ。高校生らしい（幸せな帰結を生むわけではないことをまだ知らぬがゆえの）好奇心で，エリカちゃんの男性遍歴をちょっとだけ知りたいと思ってしまった。アツシくんはやめておけばいいものを，給食の最中にエリカちゃんを含む女子グループの会話に聞き耳を立ててしまっている。

エリカちゃん「そういえば，私の 3 人目の彼がさぁ…」

どういうことやねん！ドキン！心臓の鼓動が止まらない。しかし，アツシくんは冷静さをすぐに取り戻し，心の中でこうつぶやいた**―「この観測の下で，エリカの元カレの数が分布している…」―。**

・まずは，事前分布を考える（**事前分布の設定**）
- 彼氏の数はカウントデータ　→　**離散分布**を考える
- エリカのことをよく知りたいが，今のところなにも情報がない
- 女子高生の男性遍歴からいって，彼氏が 10 人以上いるということはないだろう
- しかしそれ以上の情報はない　→　最小値 0，最大値 10 の**離散一様分布**を考えよう

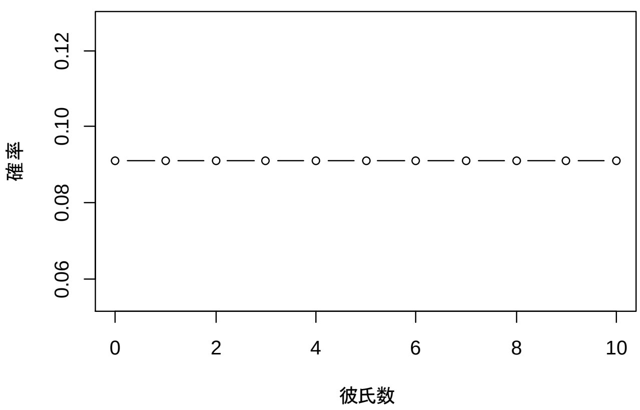


図 4. エリカの元カレ数に関する事前分布

・次に，モデルを考えよう（**尤度の計算**）

- 全彼氏数から，ランダム（均等確率）で 3 番目の彼氏が得られたと考える
- 彼氏数を n，言及された彼氏の序数を k とする
- 三番目がいるということは，0 人，1 人，2 人，3 人という確率は 0（最低 4 人はいる）
- 4 人いたとき, 3 番目が話題になる確率は 1/4, 5 人いたとき 3 番目が話題になる確率は 1/5…

$$p = \frac{1}{n} \quad (n > k) \tag{13}$$

$$p = 0 \qquad (n < k)$$

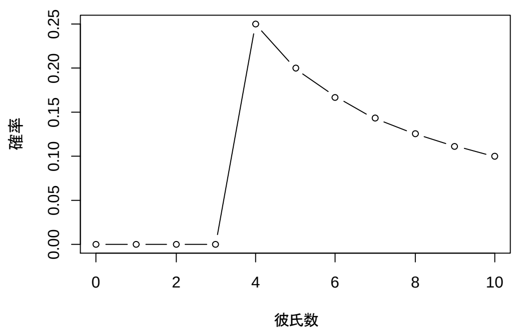


図 5. エリカの元カレ数に関する尤度

・事後確率を考えよう！

- **事前分布と尤度の積**（したあとに正則化定数で揃える）

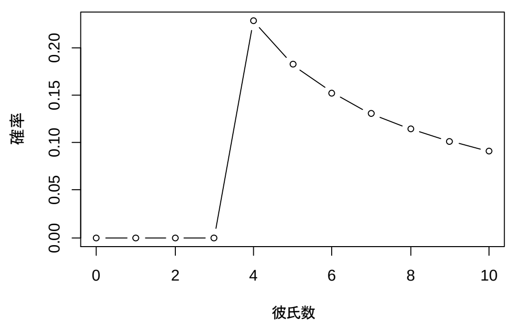


図 6. エリカの元カレ数に関する事後分布(1)

・「三番目の彼氏がさあ」という観測の下で，エリカの元カレ数が分布している

・**事後期待値**を求めてみる（ベイズ統計における推定値，この分布の平均値）

- **エリカには 6.34 人くらい元カレがいる**

■事前分布の影響

・事前分布で設定した最大値が期待値を規定してしまっている

・仮に 100 人だと，事後分布は…

・期待値が 20 人台に

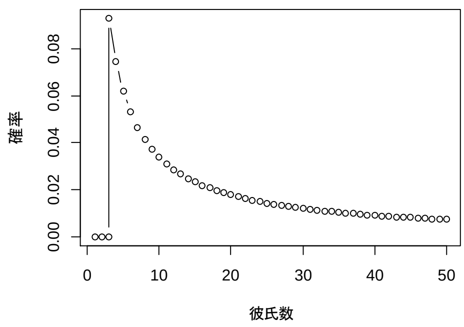


図 7. エリカの元カレ数に関する事後分布（2）

・しかし，こんなそれっぽい事前分布を設定（女子高生の平均的な元カレ数の分布？）すると…

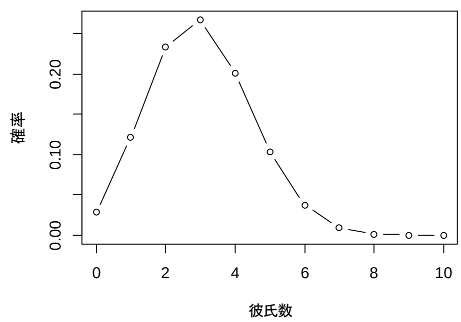


図 8. エリカの元カレ数に関する新しい事前分布

・ここでの事後分布は,

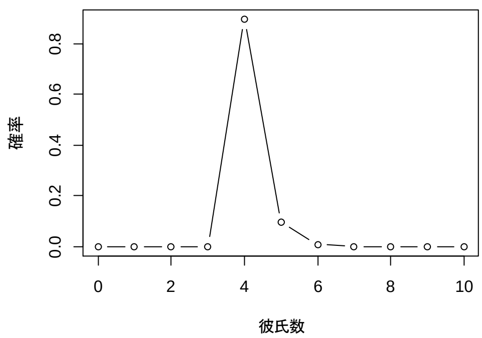


図 9. エリカの元カレ数に関する事後分布（3）

・元カレが 4 人（今彼が 5 人目）の確率は 90%だといえる
・95%の確率で 4 人か 5 人だといえる（**ベイズ信用区間**）

■**ベイズ更新**
・計算を終えたアツシくんの耳に「4 人目のときもさぁ…」という可愛らしいエリカの声
・このデータによって，**ベイズ更新**ができる（4 が死んで 5 の確率が上がった）

- 事前分布を図 10 とする（**昨日の事後分布は今日の事前分布**）
- **データが増えると事前分布の影響は小さくなっていく**

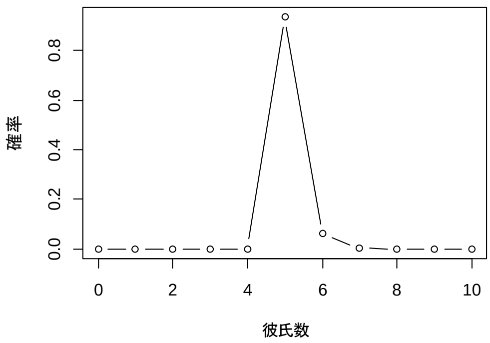


図 10. 新しい観測によってベイズ更新されたエリカの元カレ数の事後分布

■マルコフ連鎖モンテカルロ法（MCMC）
・しかしモデルが複雑になれば，上記のように簡単な数式で事後分布を計算できない
・MCMC はシミュレーションベースの数値解析的推定方法のひとつ
・**事後分布に比例する（従う）分布**から**乱数サンプル**を大量に作る
・得られるのは事後分布そのものではなく，あくまでも乱数
・この乱数を扱うことで，母数について手軽にさまざまな推測ができる
・最近では，３種類のアルゴリズムがよく知られている
- **メトロポリス・ヘイスティング法**
- **ギブス・サンプリング**
- **ハミルトニアン・モンテカルロ法**

・数値解析なので**初期値**が必要

■研究者が決めること・報告すること
・**モデル**
- **確率的プログラミング**という方法やその言語でモデルを定式化する

・**バーンイン区間（焼却区間）**
- 定常分布に至るまでの部分を捨てる

・**間引き区間**（thinning interval）
- 途中で間引きを行うことによって，安定した結果を得ようとすることもある

・MCMC **サンプル数（反復数）**
- 行なうシミュレーションの数，多いほうが安定しやすい

・各母数の**事前分布**
・**チェイン数**
- 初期値からシミュレーションを行なう一連のつながり

■**収束診断**

・ある程度，事後分布が安定していることを示す必要がある（**定常分布**）

・Gelman の収束診断，Geweke の収束診断，R^などを使って，収束したことを示す

・**トレース**と **MCMC サンプル**を示す

■母数の推定値や**ベイズ信用区間**

・MCMC サンプルの平均（**事後期待値; EAP**）は，**母数の推定値における代表値のひとつ**

・MCMC サンプルにおける 2.5%点から 97.5%点は 95%**ベイズ信用区間**とみなせる

- **パーセンタイル法**
- **最高密度区間**（HPD）という考えもある

■**ベイズ因子（BF）と情報仮説**

・（従来の統計がいうところの）帰無仮説と対立仮説の**もっともらしさの比**を考える

$$\mathrm{BF} = \frac{P(D|H_1)}{P(D|H_2)} \tag{14}$$

・たとえば，二群の平均値の比較において，標準化平均差 d

$$d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}} \tag{15}$$

について，

$$H_1 : d = 0 \tag{16}$$

$$H_2 : \infty < d < 0 \tag{17}$$

$$H_3 : -\infty > d > 0 \tag{18}$$

といった仮説（**情報仮説**）を立てることもできる

$$\mathrm{BF} = \frac{P(D|H(d \neq 0))}{P(D|H(d = 0))} \tag{19}$$

・効果量 d が 0 であるよりも，0 以外であるほうが 10 倍もっともらしいなど

・効果量の値が負であるよりも，正であるほうが 200 倍もっともらしいなど

- これらのベイズ因子の値に目安を設けることもある（が，慎重であるべき）

・従来の NHST とは異なり，**帰無仮説相当の主張を評価**することができる

・全面的に **p 値を廃止し**，**ベイズ因子の使用を推奨する**向きもある

講義 2：　ベイズ統計を使った外国語教育研究

1. テスト得点間の相関

■100 人の語彙テストと文法テストの相関係数を得る

■まずは，**頻度主義的な分析例**

表 1.

データの記述統計の例（*N* = 100）

	M	*SD*
語彙テスト	51.14	9.11
文法テスト	50.80	9.10

・分散共分散行列

$$\mathbf{\Sigma} = \begin{pmatrix} 82.96 & 46.19 \\ 46.19 & 82.73 \end{pmatrix} \tag{1}$$

・相関係数；$r = 0.56$ [.41, .68]

・無相関検定は有意，$t\,(98) = 6.64$，$p < .01$

■**ベイズ的分析**

・まずは，モデルの定式化：それぞれの観測が**多変量正規分布**に従うと考える

・多変量正規分布の母数は，**平均ベクトル μ** と**分散共分散行列 Σ** だ

・よって，観測値 x は,

$$[x_{i1}, x_{i2}] \sim \mathrm{MulitivariateNormal}([\mu_1, \mu_2], \mathbf{\Sigma}) \tag{2}$$

・X ~ Y は，**X が Y 分布（母数）に従う**と読む

・分散共分散行列 **Σ** は，相関係数 ρ を使って（22）のように表現することもできる

- ρ は，それぞれの標準偏差に対する共分散の比であるから

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{3}$$

・ベイズ統計では，モデル内の母数や係数も確率変数として扱うので…

・平均ベクトル（μ_1，μ_2）と σ_1，σ_2，ρ の分布を考える必要がある（**事前分布**）

・以下のように**設定**した（これは分析者が設定する）

・これはほぼ**無情報**であるといってもよい

$$\mu_1, \mu_2 \sim \mathrm{Normal}(0, 1000) \quad (4)$$

$$\sigma_1, \sigma_2 \sim \mathrm{Uniform}(0, 1000) \quad (5)$$

$$\rho \sim \mathrm{Uniform}(-1, 1) \quad (6)$$

・この事前分布を可視化すると…

・これは，データを取る前の研究者の主観を表しているともいえる

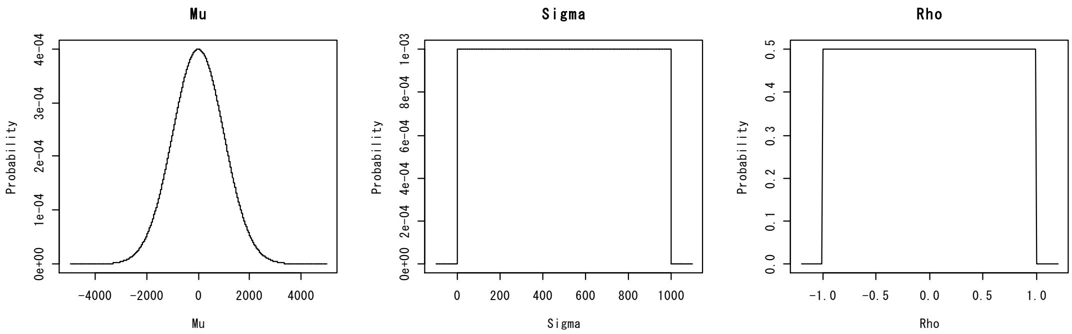


図 1．事前分布の可視化の様子

・ベイズ統計の基本的方針は，この**主観をデータによって更新して，事後分布を構築する**こと

$$事後分布 \propto 事前分布 \times 尤度 \quad (7)$$

・MCMC によって事後分布に近似するサンプルを得る

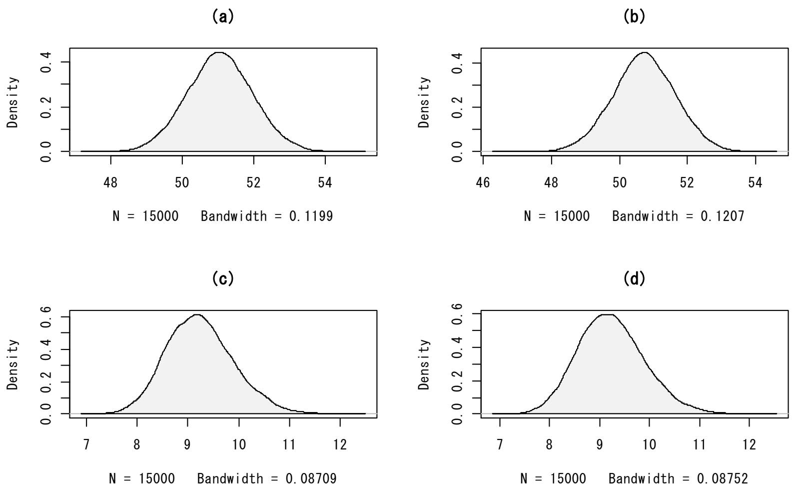


図 2．平均ベクトル（μ_1，μ_2）と σ_1，σ_2 の事後分布に近似するサンプルの可視化

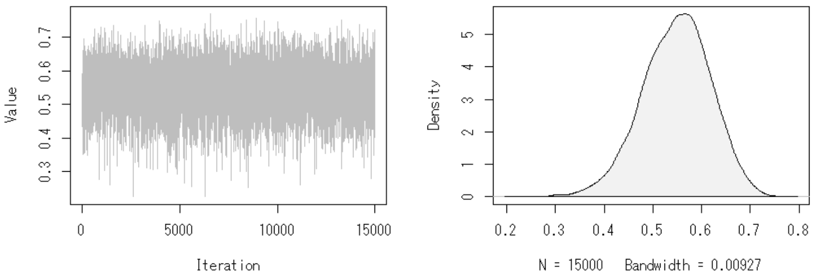


図 3.　相関係数 ρ のトレース図と事後分布に近似するサンプルの可視化

表 2.

MCMC によって得られたサンプルの記述統計

母数	事後期待値	事後標準偏差	2.5%点	25%点	50%点	75%点	97.5%点
μ_1	51.06	0.92	49.26	50.45	51.06	51.67	52.88
μ_2	50.73	0.92	48.91	50.11	50.73	51.35	52.52
σ_1	9.25	0.67	8.04	8.78	9.20	9.67	10.66
σ_2	9.24	0.67	8.02	8.77	9.20	9.66	10.67
ρ	0.55	0.07	0.40	0.50	0.55	0.59	0.68

・相関係数の 95%ベイズ信用区間を最高密度区間（HDI）によって構築すると，[0.41, 0.69]だった

・相関係数が正の値を取る確率は，100%

・相関係数が 0.5 以上の値を取る確率は，76%であるなどと推論できる

- 実際は MCMC サンプルを数え上げるだけでいい

2. 事前事後の平均値

■処遇の効果検証

・30 人に処遇を施し，その事前事後のテストスコアを比較する

・まずは，頻度主義的な分析例

表 3.

記述統計の例

	M	*SD*
事前	49.93	3.30
事後	55.40	3.35
差得点	5.47	1.81

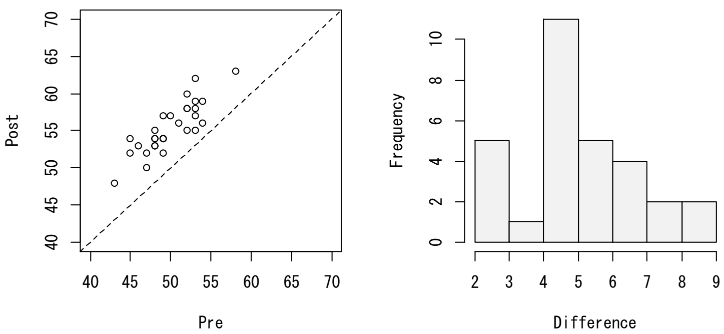


図 4. 散布図と差得点のヒストグラム

・対応ありの t 検定；$t(29) = 16.50, p < .01$
・差得点の 95%信頼区間；[4.79, 6.14]
・標準化平均差；$d = 3.03$

■ベイズ的分析
・差得点の平均値 μ_d と差得点の標準偏差 σ_d の事後分布を直接考えてもよい
・事前＋差得点 ＝ 事後
・ここでは標準化平均差 δ は，差得点の平均値 μ_d と差得点の標準偏差 σ_d の**生成量**だと考える

$$d_i = y_i - x_i \tag{8}$$

$$\delta = \frac{\mu_d}{\sigma_d} \tag{9}$$

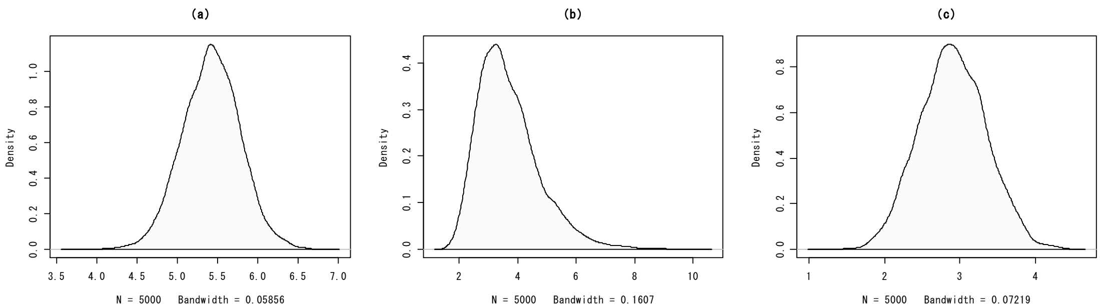


図 5. 事後分布に近似するサンプルの可視化；それぞれ μ_d，σ_d，δ

・たとえば，μ_d の 95%信用区間は[4.71, 6.11]
・**95%の確率で単純効果量はこの間に入るだろう**，と推論できる
・また，δ の 95%信用区間は[2.07, 3.77]
・**95%の確率で標準化効果量はこの間に入るだろう**，と推論できる

3. 重回帰モデル

・中間テストの得点と出席回数から期末試験の点数を予測する

・200 人のデータ

・まずは，可視化

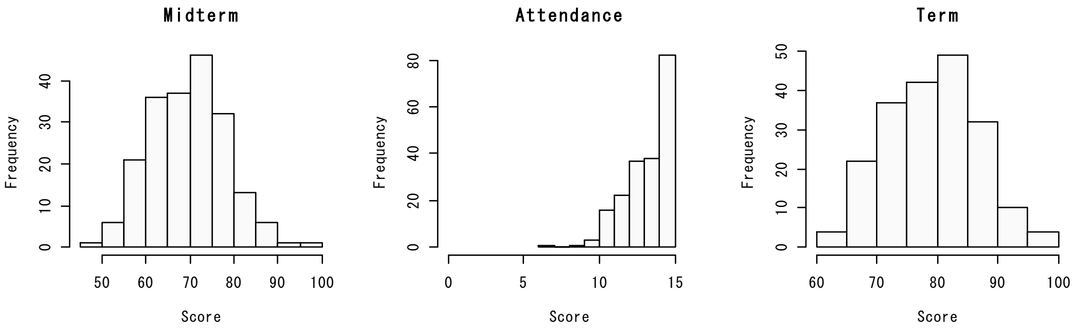


図 6. データのヒストグラム

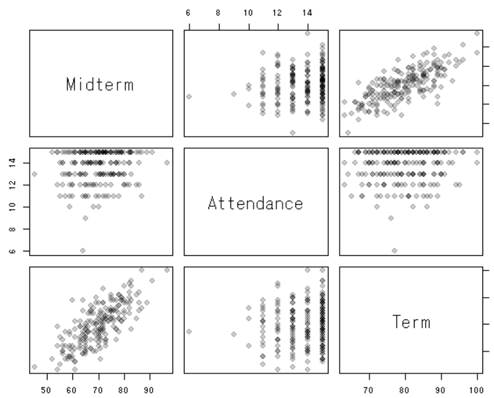


図 17. 散布図行列

・相関係数行列

$$\boldsymbol{\rho} = \begin{bmatrix} 1.00 & .14 & .75 \\ .14 & 1.00 & .09 \\ .75 & .09 & 1.00 \end{bmatrix} \tag{10}$$

■ベイズ的分析

・まずは，モデルを立てる

M_1：
$$y_i = \beta_0 + \varepsilon_i \tag{11}$$

M_2：
$$y_i = \beta_0 + \beta_1 x_{1_i} + \varepsilon_i \tag{12}$$

M_3：
$$y_i = \beta_0 + \beta_2 x_{2_i} + \varepsilon_i \tag{13}$$

M_4：
$$y_i = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \varepsilon_i \tag{14}$$

・それぞれにおいて，以下のように事前分布を設定

$$\beta_j \sim \mathrm{Normal}(0, 10) \quad (15)$$

$$\varepsilon_j \sim \mathrm{Normal}(0, \sigma^2) \quad (16)$$

$$\sigma^{-2} \sim \mathrm{Gamma}(\frac{1}{2000}, \frac{1}{2000}) \quad (17)$$

・このモデルの母数を MCMC で推定する
・周辺尤度を推定し，**ベイズ因子**によってモデル比較を行う

表 4.
対数ベイズ因子表

分子	対数周辺尤度	分母			
		M_1	M_2	M_3	M_4
M_1	-1166.08	0	-498	-373	-512
M_2	-668.17	498	0	125	-15
M_3	-793.91	373	-125	0	-139
M_4	-653.61	512	15	139	0

・4 つ目のモデルが**もっともらしそうだ**→MCMC サンプルを使って各母数について検討

表 5.
各母数の事後分布に近似する MCMC サンプルの記述統計

	事後期待値	事後標準偏差	2.5%点	50%点	97.5%点
β_0	1.48	0.98	-0.48	1.47	3.41
β_1	1.16	0.19	0.77	1.03	1.55
β_2	0.89	0.04	0.81	0.86	0.96
σ^2	34.39	3.54	28.06	34.14	42.06

M_4：

$$y_i = \beta_0+\beta_1 x_{1_i} + \beta_2 x_{2_i} + \varepsilon_i \quad (18) = (14)$$

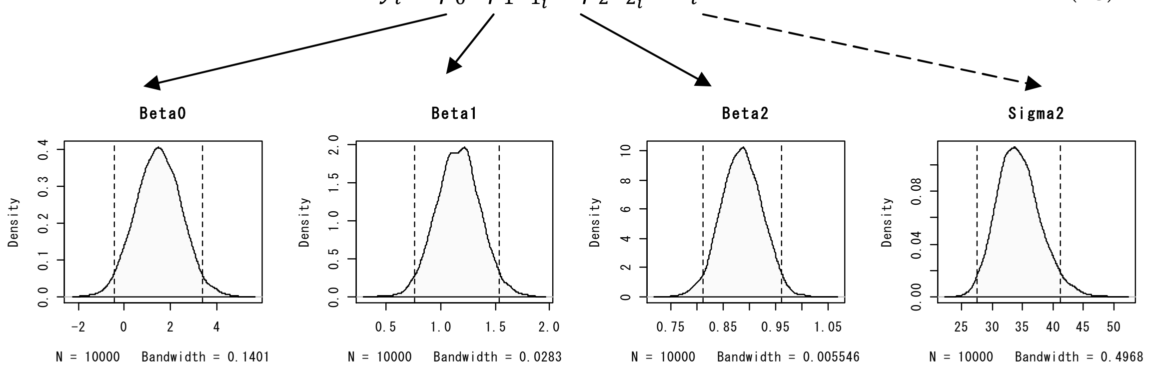


図 8. 事後分布の概観

ハンズオン：　ベイズ統計を試してみる

1. JASP ハンズオン

■JASP とは

・"A Fresh Way to Do Statistics"
・SPSS 風の洗練された UI をもつ統計ソフトウェア
・無料・無償
・マルチプラットフォーム（Windows, MacOS, Linux）
・現在も開発・サポートが活発
・基本的な分析のベイズ代替的手法を実装
・現在は Version 0.10.2
・現在サポートする主な機能

伝統的分析手法	ベイズ的分析手法
t 検定，分散分析，回帰分析，相関，クロス集計表の分析，因子分析，主成分分析，メタ分析，ネットワーク分析，共分散構造分析	t 検定，分散分析，回帰分析，相関，クロス集計表の分析，因子分析，主成分分析

・可視化にも非常に優れる

■インストール

1. Website へアクセス

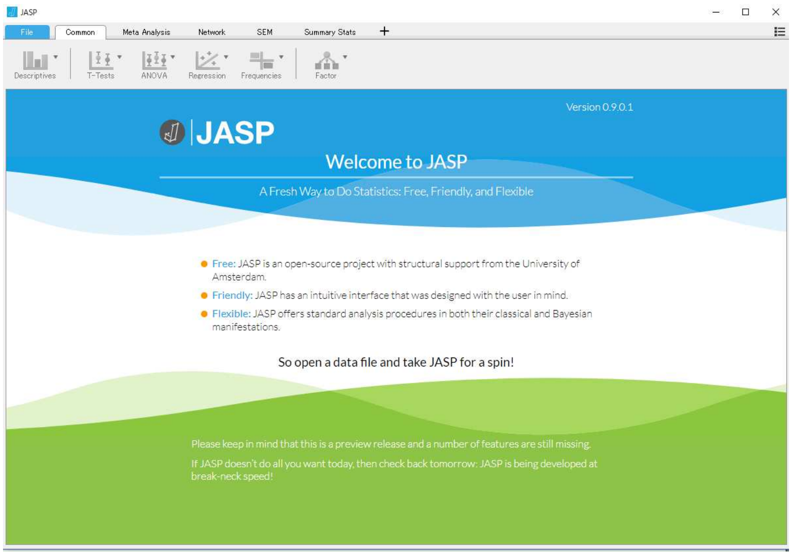
https://jasp-stats.org/

2. ダウンロード

3. それぞれの OS を選択

- JASP for Windows
- JASP for Mac
- Linux Installation Guide

4. インストーラウィザードに従いインストール
5. 起動
6. メイン画面



・左上に File，Common などがある

7. ファイルを読み込み

・File を選択　→　Computer →　BayesWS →　independent_ttest.csv を選択
・基本は.csv 形式が便利

サポートするファイル形式

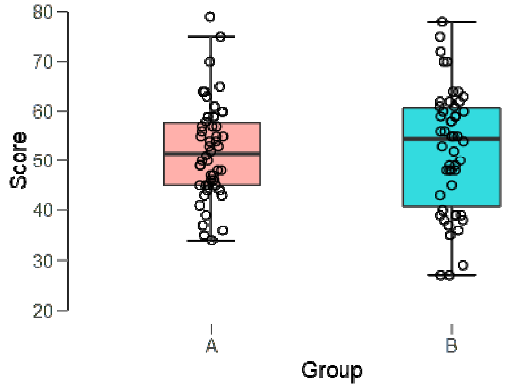
.jasp（独自形式），.csv，.txt，.sav

・Open Science Framework によるデータをインポートすることも可能
・大量のデータ例も付属（Data Library）

8. 記述統計と可視化

・メニュー　→　Common →　Descriptives →　Descriptive Statistics →　Variable に Score を入れる
→　Split にグループ化変数である Group を入れる
・Plot →　Distribution plots，Boxplots などに適宜チェックを入れる
・Statistics →　必要な統計量にチェックを入れる

・表はそのままコピペもできるし，LaTeX 出力もできる
・グラフは画像として保存することもできる

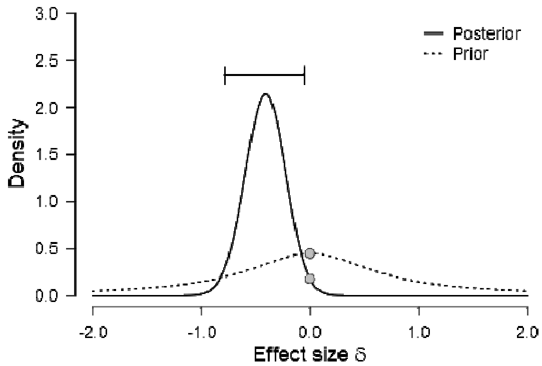


9. *t* 検定
・メニュー　→　Common →T-Tests →　Independent T-Test
・Dependant Variable に Score を Grouping Variable に Group を入れる
・Welch，Effect Size，Descriptives，Descriptive plots などに適宜チェック
・なにもいえないという結果

10. 二群の *t* 検定のベイズ的代替分析（ベイズ因子）
・メニュー　→　Common →T-Tests →　Bayesian Independent Samples T-Test
・Dependant Variable に Score を Grouping Variable に Group を入れる
・BF01，Prior and Posterior などに適宜チェック
・BF01 は，帰無仮説（群間の効果量が 0）が対立仮説（0 以外）とする仮説よりもどれほどもっともらしいかを示す
・(BF10) にすると反対
・ここでは **4 倍程度帰無仮説の方がもっともらしい**ということがわかった
・ここでは効果量の事前分布にコーシー分布を指定している
・この事前分布のパラメーター（*r*）を変えることで結果は変わる
・Results →Remove All

11. 対応ありの *t* 検定のベイズ的代替分析（ベイズ因子）
・File →Paired_ttest.csv
・可視化（散布図）をしてみましょう
・メニュー　→　Common →T-Tests →　Bayesian Paired Samples T-Test
・T1 と T2 の変数を投入
・同様
・やや対立仮説のほうがもっともらしいようだ

・効果量の符号に関する仮説も立てることができる



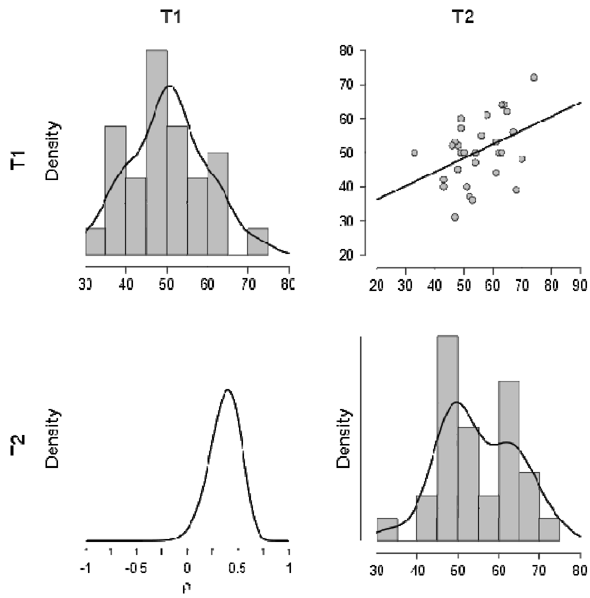
12. 相関分析

・Common　→　Regression →　Correlation Matrix →２つの変数を投入

・Common　→　Regression →　Bayesian Correlation Matrix →２つの変数を投入

→　Correlation Matrix，Posteriors Under H1 にチェック

・ここでは無相関検定 r = 0 を帰無仮説としている



13. 一元配置分散分析

・File → oneway_anova.csv を読み込む

・Common →　ANOVA →　Dependant Variable に Score，Fixed Factors に Group を投入

・適宜欲しい情報にチェックを入れる

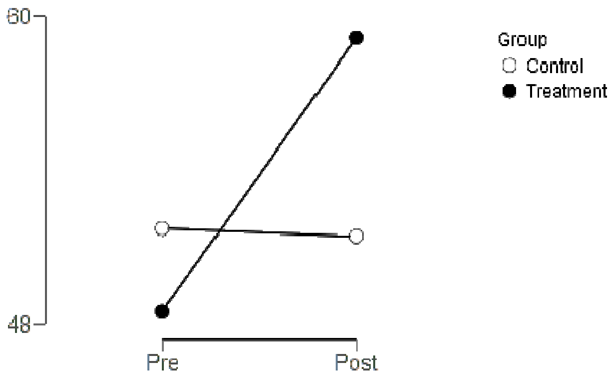
・Common →　Bayesian ANOVA →　Dependant Variable に Score，Fixed Factors に Group を投入

・ここでは，差がないモデルの方が 2.7 倍程度もっともらしい

・Post Hoc Tests に Group を投入
・各組み合わせの差についてベイズ因子が表示される

14. 混合計画の分散分析

・File → twoway_anova.csv を読み込む
・Common → Repeated Measures ANOVA
・Repeated Measures Factor に Pre と Post を投入
・Between Subject Factor に Group を投入
・適宜欲しい情報にチェックを入れる



・Common → Bayesian Repeated Measures ANOVA
・ここでは，すべてのモデルについてベイズ因子の比較が行われている

Models	P(M)	P(M|data)	BF_M	BF_{10}	error %
Null model (incl. subject)	0.200	4.578e -6	1.831e -5	1.000	
RM Factor 1	0.200	3.534e -4	0.001	77.194	1.149
Group	0.200	2.202e -6	8.808e -6	0.481	0.706
RM Factor 1 + Group	0.200	1.872e -4	7.488e -4	40.882	2.090
RM Factor 1 + Group + RM Factor 1 ＊ Group	0.200	0.999	7304.284	218319.489	3.154

Note. All models include subject.

M1：要因の影響はなし
M2：事前事後の影響のみ
M3：群の影響のみ
M4：事前事後および群の影響のみ
M5：二要因の交互作用

	分母				
分子	M1	M2	M3	M4	M5
M1	1.00	0.01	2.08	0.02	0.00
M2	77.19	1.00	160.49	1.89	0.00
M3	0.48	0.01	1.00	0.01	0.00
M4	40.88	0.53	84.99	1.00	0.00
M5	281319.49	2828.19	453886.67	5340.24	1.00

15. 重回帰モデル

・File → regression.csv を読み込む

・Common →　Linear Regression

・Dependent Variable に Reading，Covariates に Grammar と Vocabulary

・適宜チェック

・Common → Bayesian Regression

・基本的に ANOVA と同様

ハンズオン： その他の環境や文献案内

ベイズ統計に関する計算環境など

- JASP
- R
- BayesFactor（R パッケージ）：JASP に関連した分析が R でできます。スクリプトをご参照ください
- MCMCPack（R パッケージ）
- brms（R パッケージ）
- BEST（R パッケージ）
- rstan（R パッケージ）
- Python（pyMC など）
- Julia（Mamba.jl）
- JAGS（MCMC 用ソフトウェア）
- WinBUGS（MCMC 用ソフトウェア）
- Stan（MCMC 用ソフトウェア）

6. ベイズ統計に関する書籍

・近年は，爆発的に書籍が出回っており，どれを手にとっても基本的に似たような内容になっています
・ここでは以下のような本をおすすめします

- 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
- 豊田秀樹（2017）『実践ベイズモデリング―解析技法と認知モデル―』朝倉書店
- 豊田秀樹（2018）『たのしいベイズモデリング：事例で拓く研究のフロンティア』北大路書房　**（草薙も，外国語教育の事例について分担執筆しております！ぜひお買い求めください！）**
- **馬場真哉**（2019）『R と Stan **ではじめるベイズ統計モデリングによるデータ分析入門』講談社**
- 渡辺澄夫（2012）『ベイズ統計の理論と方法』コロナ社
- 渡部洋（1999）『ベイズ統計学入門』福村出版
- Gelman et al. (2013). *Bayesian data analysis (third edition)*. Chapman and Hall/CRC
- Krushchke, J. *Doing Bayesian analysis: A tutorial with R, JAGS and Stan*. Academic Press.（翻訳書あり）
- Lee, M. D., Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical Course*. CUP. （翻訳書あり）

7. 総まとめ

講義３：　より高度なモデリングへ（おまけ）

形態素習得順序の普遍性

■Dulay and Burt（1972）

・第二言語習得研究全体の歴史の中でももっとも重要な文献のひとつ

・第二言語における形態素の習得において，その順序に**普遍性**があることを主張（**自然順序**）

・スペイン語母語話者（n = 60）と中国語母語話者（n = 55）を対象

・口頭算出における GSM（group score method）のスコアを 11 の形態素毎に算出

・スペイン語母語話者と中国語母語話者間のスコアの順位相関係数が高かった（r = .91 [.68, .98], p < .01）

・相関係数が十分に高いことをもって普遍性を主張

表 1.

Dulay and Burt を再現したデータ

	Functors	Chinese		Spanish	
		k	*n*	*k*	*n*
(a)	Pronoun Case	55	55	58	60
(b)	Article	45	55	56	60
(c)	Progressive	40	55	55	60
(d)	Contractible Copula	44	54	52	58
(e)	Plural	28	55	53	60
(f)	Contractible Auxiliary	24	45	32	49
(g)	Past Regular	20	43	32	47
(h)	Past Irregular	25	55	35	60
(i)	Long Plural	21	53	34	58
(j)	Possessive	21	55	28	60
(k)	Third Person	6	49	25	53

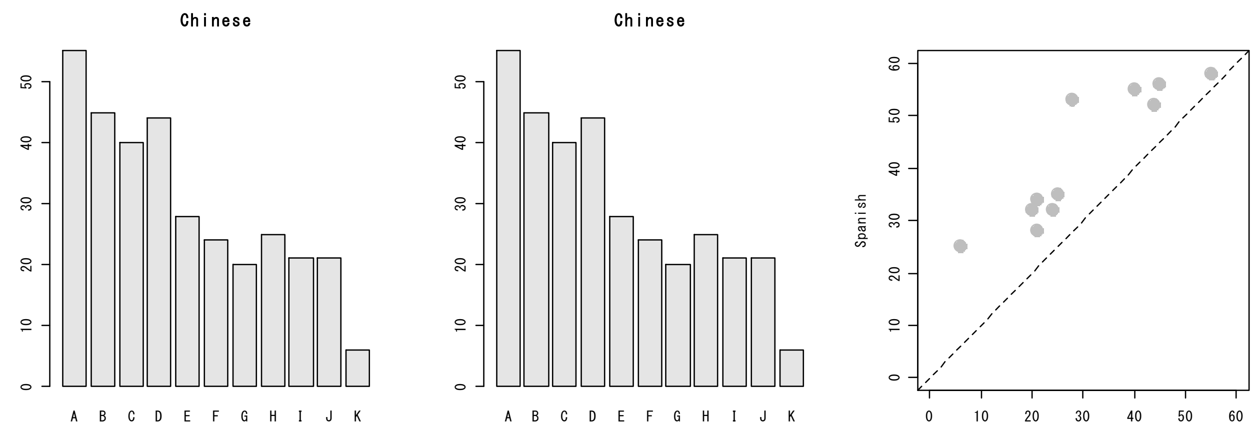


図 1. 当該データ

■普遍性のその後
・論理的な意味における「普遍性」に関する誤認識は（好意的に？）問われないままだったが…
- その主張に反して，母語の影響の大きさは後の研究で強く認識されるものとなった
- **「形態素の習得順序に，ある一定のパターンが見られる」**といった弱めの主張は否定されなかった

■自然順序を階層ベイズで分析する
・我々の観測は正用数であり，我々がモデリングしているのは形態素の習得順序ではなく，**任意の集団と，所与の形態素の集合における確率的順序関係**
・**観測の生成過程**を数理的に考える
・正用数を k，必然事例数を n とすると，k は n および θ（確率）を母数とする二項分布に従う
- 添字 i（$1 < i < 11$）は，各形態素

$$k_i \sim \mathrm{Binomial}(n_i, \theta_i) \tag{1}$$

・ここで，二項分布の確率質量関数は,

$$f(k|\theta, n) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{2}$$

・i 個の形態素は，それぞれ θ_i をもつ
・母数 θ_i は，母数 α および母数 β のベータ分布に従うと仮定する
- ベータ分布は 0 から 1 の間にある値を柔軟に表現できる

$$\theta_i \sim \mathrm{Beta}(\alpha, \beta) \tag{3}$$

・ベータ分布の確率密度関数は，B をベータ関数とすると，

$$f(\theta_i|\alpha, \beta) = \frac{\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}}{\mathrm{B}(\alpha, \beta)} \tag{4}$$

・実数 α および β を項にとるベータ関数は,

$$\mathrm{B}(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta \tag{5}$$

・さらに，ベータ分布の母数 α と母数 β に事前分布を与える；それぞれ無情報事前分布とした

$$\alpha \sim \mathrm{Uniform}(-\infty, \infty) \tag{6}$$

$$\beta \sim \mathrm{Uniform}(-\infty, \infty) \tag{7}$$

■事後分布を得て分析を加える

■優越率行列

・このモデルから，ある形態素の母数 θ_i が，別の形態素の母数 θ_j よりも高い値を取る確率を計算することができる；これをここでは優越率 d とする；これは**ある形態素が別の形態素よりも相対的に順位が高い確率**を示す

・これは一種の**サーストンモデル**の応用ともいえる

・すべての形容詞の組み合わせについて，これを正方行列で表現することができる；これを優越率行列 $\boldsymbol{d}$ とする

・これは**モデルの生成量**

$$\boldsymbol{d} = \begin{bmatrix} P(\theta_1 \leq \theta_1) & \dots & P(\theta_{11} \leq \theta_1) \\ \vdots & \ddots & \vdots \\ P(\theta_1 \leq \theta_{11}) & \dots & P(\theta_{11} \leq \theta_{11}) \end{bmatrix} \tag{8}$$

■順位の事後確率

・このモデルから，**各形態素が取りうる順位の事後確率**を考えることができる

■所与の順序が正しい確率

・このモデルから，ある与えられた順序が正しい確率を評価することができる

■Ulam 距離

・このモデルから，2 群の **Ulam 距離**（編集距離のひとつ，一つの要素を抜き，スライドさせて適切な場所に挿入するという編集工程数）の事後分布を計算できる（が，詳しくは割愛）

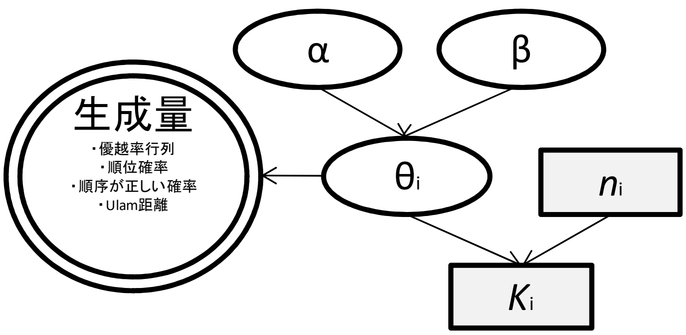


図 2. プレート表現（グラフィカルモデル）による本モデルの概要

■データ

・Dulay and Burt は元データを報告していないので，図から画像認識技術を使いデータを再現

・ハミルトニアン・モンテカルロ法（反復数 = 10,000，チェイン数 = 3，焼却区間 = 1,000，間引きなし）

・Rhat の値で MCMC の収束を判断

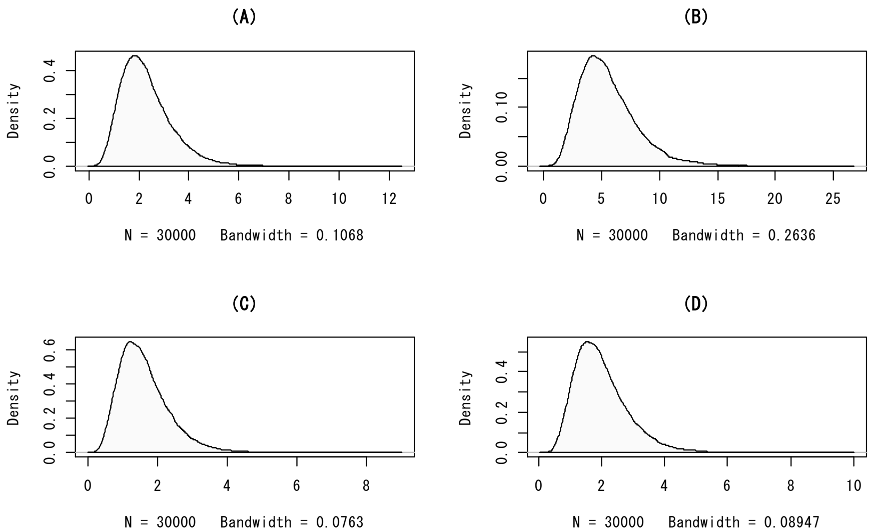


図 3. 事後分布；（A）中国語母語話者の α，（B）スペイン語母語話者の α

（C）中国語母語話者の β，（D）中国語母語話者の β

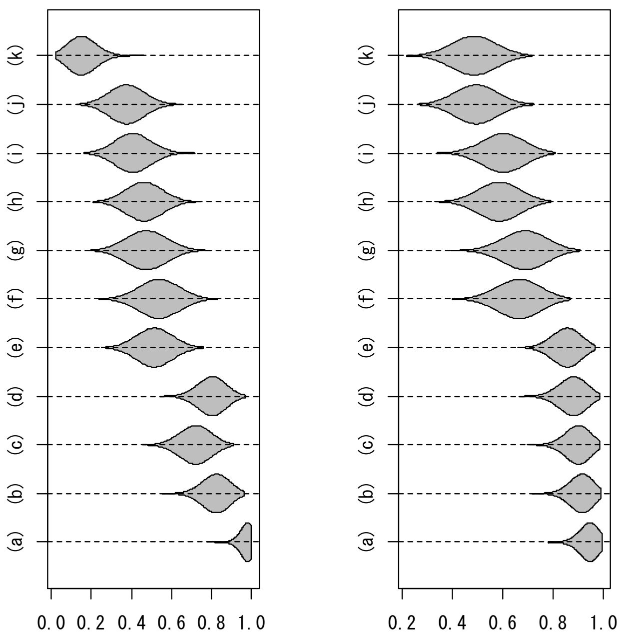


図 4. θ の事後分布；左列は中国語母語話者，右列はスペイン語母語話者

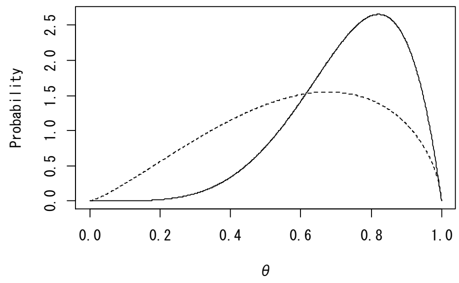


図 5. 事後期待値によるベータ分布；破線は中国語母語話者，実線はスペイン語母語話者

表 2.

中国語母語話者の優越率行列

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
(a)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(b)	.00	1.00	.91	.61	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(c)	.00	.09	1.00	.15	.99	.97	1.00	1.00	1.00	1.00	1.00
(d)	.00	.39	.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(e)	.00	.00	.01	.00	1.00	.41	.66	.71	.88	.93	1.00
(f)	.00	.00	.03	.00	.59	1.00	.73	.78	.91	.95	1.00
(g)	.00	.00	.00	.00	.34	.27	1.00	.54	.75	.84	1.00
(h)	.00	.00	.00	.00	.29	.22	.46	1.00	.73	.83	1.00
(i)	.00	.00	.00	.00	.12	.09	.25	.27	1.00	.64	1.00
(j)	.00	.00	.00	.00	.07	.05	.16	.17	.36	1.00	1.00
(k)	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.00

表 3.

スペイン語母語話者の優越率行列

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
(a)	1.00	.76	.84	.91	.96	1.00	1.00	1.00	1.00	1.00	1.00
(b)	.24	1.00	.62	.74	.86	1.00	1.00	1.00	1.00	1.00	1.00
(c)	.16	.38	1.00	.63	.79	1.00	1.00	1.00	1.00	1.00	1.00
(d)	.09	.26	.37	1.00	.67	1.00	1.00	1.00	1.00	1.00	1.00
(e)	.04	.14	.21	.33	1.00	.99	.98	1.00	1.00	1.00	1.00
(f)	.00	.00	.00	.00	.01	1.00	.39	.82	.76	.97	.98
(g)	.00	.00	.00	.00	.02	.61	1.00	.88	.84	.98	.99
(h)	.00	.00	.00	.00	.00	.18	.12	1.00	.42	.85	.87
(i)	.00	.00	.00	.00	.00	.24	.16	.58	1.00	.89	.91
(j)	.00	.00	.00	.00	.00	.03	.02	.15	.11	1.00	.54
(k)	.00	.00	.00	.00	.00	.02	.01	.13	.09	.46	1.00

表 4.

中国語母語話者における順位の事後確率

	1	2	3	4	5	6	7	8	9	10	11
(a)	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
(b)	.00	.59	.33	.07	.04	.00	.00	.00	.00	.00	.00
(c)	.00	.04	.17	.76	.00	.00	.00	.00	.00	.00	.00
(d)	.00	.37	.50	.13	.00	.00	.00	.00	.00	.00	.00
(e)	.00	.00	.00	.01	.28	.29	.21	.13	.06	.02	.00
(f)	.00	.00	.00	.02	.43	.26	.15	.08	.04	.01	.00
(g)	.00	.00	.00	.00	.14	.20	.23	.21	.14	.08	.00
(h)	.00	.00	.00	.00	.09	.17	.23	.25	.17	.09	.00
(i)	.00	.00	.00	.00	.02	.05	.11	.20	.33	.29	.00
(j)	.00	.00	.00	.00	.01	.02	.06	.13	.26	.52	.00
(k)	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.00

表 5.

スペイン語母語話者における順位の事後確率

	1	2	3	4	5	6	7	8	9	10	11
(a)	.66	.19	.19	.04	.01	.00	.00	.00	.00	.00	.00
(b)	.21	.35	.24	.14	.05	.00	.00	.00	.00	.00	.00
(c)	.09	.25	.30	.24	.13	.00	.00	.00	.00	.00	.00
(d)	.04	.14	.23	.32	.26	.00	.00	.00	.00	.00	.00
(e)	.01	.06	.12	.25	.53	.01	.02	.00	.00	.00	.00
(f)	.00	.00	.00	.01	.02	.34	.52	.04	.07	.00	.00
(g)	.00	.00	.00	.00	.00	.36	.28	.14	.19	.01	.01
(h)	.00	.00	.00	.00	.00	.18	.11	.29	.34	.04	.04
(i)	.00	.00	.00	.00	.00	.09	.05	.35	.28	.13	.11
(j)	.00	.00	.00	.00	.00	.02	.01	.14	.10	.39	.34
(k)	.00	.00	.00	.00	.00	.00	.00	.04	.03	42	.50

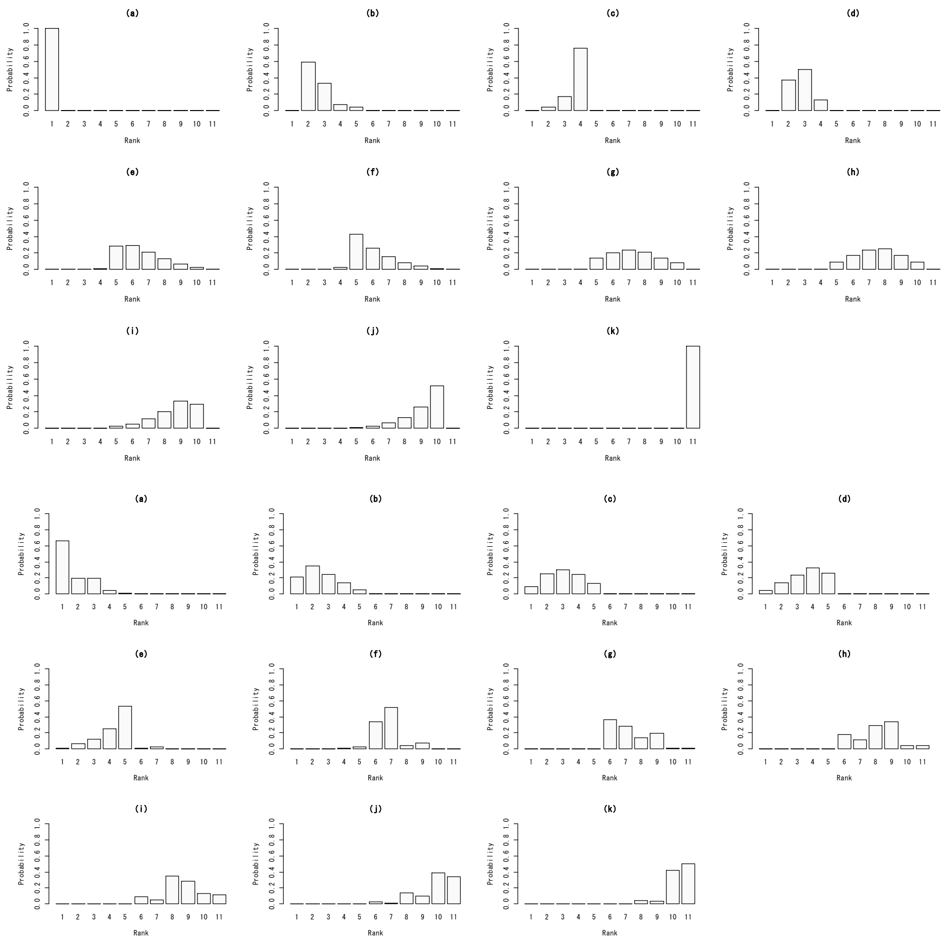


図 6. 順位の事後確率；上段は中国語母語話者，下段はスペイン語母語話者

・Dulay and Burt が宣言した順序仮説（p. 51）が，厳密に中国語・スペイン語の 2 群でともに正しい確率は，およそ 1/40,000
・せいぜいが，「2 群を Ulam 距離で見るとき，2 から 3 の間に入る類似性だといえる」というのが数理的に安全な言明
・しかしながら，11!がおよそ 4,000 万であることを考えると，この研究は，大幅に我々の知識を更新したといえる
・真実において普遍的か，そうでないかという観点は置いておくとしても，ベイズモデリングはこのように，柔軟かつ簡便な方法で，圧倒的に実用性のある情報をもたらしてくれる