Bayesian Data Analysis : Chapter1 Probability and inference

Andrew Gelman, John B.Carlin, Hal S.Stern,
David B.Dunson, Aki Vehtari, and Donald B.Rubin

山崎 遼也

情報学科 数理工学コース 4 回

2017/3/22～3/28

Table Contents

Key Words

1.1 ベイズデータ解析の 3 つのステップ

この本は観測する量, 詳しく知りたい量に対する確率モデルを用いてデータから推論を行うための実用的な方法に関連するものである. ベイズ手法の本質的な特徴は統計的データ解析に基づく推論の中で不確実性を量化するために確率を明示的に使用するということである.

ベイズデータ解析の過程は次の 3 ステップに分けることで理想化できる.

(1) 完全な確率モデル (問題における全ての観測, 未観測の量に対する結合確率分布) を立てる. モデルは潜在的な科学的な問題についての知識, データ収集過程で構成される必要がある.

(2) 観測データの条件付け. 観測データが与えられた下で, 事後分布 (最終的に関心のある未観測の量の条件付き確率分布) を計算し, 解釈する.

(3) モデルの適合, その結果の事後分布の含意を評価する. モデルがデータにどの程度適合しているか, 実質的な結論は妥当であるか, 結果は (1) のモデリング仮定にどの程度敏感であるか? これに応じて, モデルを変更または拡張して 3 つのステップを繰り返すことができます.

ここ 40 年でこれらの領域すべてで素晴らしい進展があり, これらの多くがこの本を通してレビューされ, 例で使われている. 私たちの治療は 3 つのステップすべてをカバーし, 第 2 のステップは計算方法論を含み, 第 3 のステップはテクニックと判断の微妙なバランスを問題の適用された状況に基づいてカバーします. 第 1 のステップは, 多くのベイジアン解析の大きな障害となっています. モデルはどこから来るのか? 適切な確率仕様を構築するにはどうすればよいか? これらの問題に関する手引きを提供し, モデルの適合性を遡及的に評価する第 3 のステップの重要性を示す.

第 2 のステップで条件付き確率分布を計算するために利用できる改良された技術に加えて, 第 3 のステップを実行することの利点は, 最初の試みで正しいモデル仕様を仮定する必要性をある程度緩和することである. 特に, 「主観的な」事前分布に対する結論の懸念される依存性を検討し, 調査することができる.

ベイジアン手法の利点, 仮説検定よりも区間推定

ベイジアン思考の主な動機は, 統計的結論の常識的な解釈を容易にすることである. 例えば, 厳密には, 繰り返し実践されるかもしれない類似の推論の系列に関してのみ解釈され得る頻度主義 (信頼) 間隔に対して, 興味のある未知の量に対するベイジアン (確率) 間隔は未知の量を含む高確率を持つものとして扱うことができる. 最近, 応用統計学では, 仮説検定ではなく区間推定に重点が置かれており, これは標準な信頼区間のほとんどのユーザに常識的なベイズ解釈を与える可能性が高いため, ベイズ視点に強い原動力を与える. この本の中の目標の 1 つは, 一般的な単純な統計的手順のベイズ解釈が正当化されるということを示すことです.

統計の基礎を議論するのではなく (基礎的な議論への言及についての, この章の終わりにある bibliographic note を参照), その柔軟性と一般性により複雑な問題に対処することのできるベイズの枠組みの実用的な利点に集中する. ベイズ推論の中心的な特徴は, 不確かさの直接的な定量化であり, 多くのパラメータと複雑な多層確率モデルを持つフィッティングモデルには原則的に障害がないことを意味します. 実際には, このような大きなモデルを使って設定し計算することが問題であり, このモデルの大部分は, これらのモデリングと計算上の課題を処理するための最近発展し, まだ発展中のテクニックに焦点を当てています. 複雑なモデルを設定する自由は, 3 章で詳しく議論するように, ベイジアンのパラダイムが複数のパラメータに対処する概念的に簡単な方法を提供するという事実から大部分が生じます.

1.2 統計的推論に対する一般的な表記

統計的推論は, 観察されていない量についての数値データからの結論を引き出すことに関係する. 例えば, 新しい抗がん剤の臨床試験は, 新薬を投与された集団における 5 年間の生存確率を, 標準的な治療を受けている集団のそれと比較するように設計されているかもしれない. これらの生存確率は, 多数の患者に対するもので, 母集団全体で実験することは実行可能でもなく, 倫理的にも受け入れられない. 従って, 真の確率についての推論, 特にそれらの相違は患者のサンプルに基づいていなければならない. この例では, 集団全体を 1 つまたは他の治療をすることが可能であったとしても, 両方の治療に誰かにすることは決して不可能であるため, 因果推論 (causal inference)(他の治療にさらされた場合の各患者における観察された結果とその患者の観察されない結果との比較) を評価するために統計的推論が依然として必要となる.

2 種類の推定量 (estimands)(統計的推論が行われる観測できない量), 1 つ目のプロセスの将来の観察や臨床試験の例では受けていない治療の結果など, 潜在的に観察可能な量と, 2 つ目の直接観察できない量, すなわち, 観察されたデータ (例えば, 回帰係数) に至る仮説的プロセスを支配するパラメータを区別する. これらの 2 種類の推定値の間の区別は必ずしも正確ではないが, 特定の問題の統計モデルが実際の世界にどのように適合するかを理解する方法として一般的に役立つ.

パラメータ, データ, 予測

一般的な表記として, θ で未観測のベクトル量または興味のある母集団パラメータ (臨床試験の例において, 集団の無作為に選択されたメンバーに対する各治療の下での生存確率), y で観測データ (生存者数や各治療群の死亡数など), $\tilde{y}$ で未知であるが, 潜在的に観測できる量 (他の治療を受けている患者の結果, またはすでに試験中の患者と同様の新たな患者のための各治療の結果) を表す. 一般に, これらの記号は多変量を表します. 一般的に, パラメータとしてギリシャ文字を使用し, 観測されたまたは観測可能なスカラーおよびベクトル (および時には行列) の小文字のローマ字, または観測された行列または観測可能な行列の大文字のローマ字を使用します. 行列表記法を使用する場合, ベクトルは全体にわたって列ベクトルとして考えられる. 例えば, u が n 個の成分を持つベクトルである場合, $u^{\mathrm{T}}u$ はスカラーであり, uu^{T} は $n \times n$ 行列となる.

観測データ

多くの統計学研究では, データは n 個のオブジェクトまたはユニットの集合のそれぞれに集められ, データをベクトル $y = (y_1, \cdots, y_n)$ として書くことができます. 臨床試験の例では, 患者 i が 5 年後に生存している場合には y_i を 1 に, 患者が死亡した場合には 0 としてラベル付けすることができる. いくつかの変数が各ユニットで測定される場合, 各 y_i は実際にはベクトルであり, データセット全体 y は行列です (通常は n 個の行を持つとみなされます). y 変数は結果と呼ばれ, 推論を行うときに, サンプリング過程, 母集団の自然変動のために変数の観測値が他に判明した可能性を許容したいという意味でランダムと見なされます.

交換可能性

統計解析の通常の出発点は, n 個の値 y_i が交換可能 (exchangeable) とみなされる (しばしば暗黙の) 仮定であり, これは, インデックスの順列について不変である結合確率密度 $p(y_1, \cdots, y_n)$ として不確実性を表すことを意味する. 結果に関連する情報が説明変数ではなく結合インデックスで伝達されるならば, 交換不可能なモデルが適切であろう (下記参照). 交換可能性の考え方は統計にとって基本的なものであり, 本の中でそれを繰り返しふりかえる.

一般に, 分布 $p(\theta)$ に関する未知のパラメータベクトル θ を与えられたときに, 独立して同一分布 (iid) として交換可能な分布からのデータをモデル化する. 臨床試験の例では, 結果 y_i を, 生存率の未知の確率 θ を与えて, iid としてモデル化することができる.

説明変数

ランダムにモデル化するのに困らないように, 各ユニットの観測をするのが一般的です. 臨床試験の例では, そのような変数は各患者の年齢や以前の健康状態を含むだろう. この変数の第 2 のクラスを説明変数または共変量 (covariates) と呼び, x とラベルを付ける. すべての n 個のユニットについて説明変数の集合全体を表すために X を使用する. k 個の説明変数がある場合, X は n 行, k 列の行列である. X をランダムとして扱うと, 交換可能性の概念は, $(x, y)_i$ の n 個の値の分布をインデックスの任意の順列によって変更しないことを要求するように拡張することができる. インデックスがランダムに割り当てられていると考えることができる X に十分な関連情報を組み込んだ後に交換可能なモデルを仮定することは常に適切である. 交換可能性の仮定から, 2 つのユニットが同じ x の値を持つならば, y の分布は同じであるという意味で, x が与えられた場合の y の分布は研究のすべてのユニットについて同じである. 任意の説明変数 x は, それらをモデル化したい場合 y カテゴリに移動できる. 8 章では, アンケート, 実験, 観察研究, および回帰モデルの文脈における本書の後半部分の分析の文脈において説明変数 (予測変数とも呼ばれる) の役割について詳細に説明する.

階層モデル

5 章以降の章では, 階層モデル (hierarchical models)(マルチレベルモデル (multilevel models) とも呼ばれる) に焦点を当てる. これは, 情報が複数の異なるレベルの観測単位で利用可能な場合に使用されます. 階層モデルでは, ユニットの各レベルで交換可能性について議論することができる. 例えば, 別々の無作為化された実験において, 2 つの医療処置がいくつかの異なる都市の患者に適用されると仮定する. そして, 他の情報が入手できない場合は, 各都市の患者を交換可能なものとして扱い, また交換可能な異なる都市の結果を扱うことは妥当であろう. 実際には, 都市レベルの説明変数として, 各都市に関連する情報があれば, また個々のレベルで前に説明した説明変数を含めると意味があり, これらの説明変数が与えられた条件付分布は交換可能である.

1.3 ベイズ推論

パラメータ θ, または未観測データ $\tilde{y}$ についてのベイジアン統計の結果は確率表記で出来上がる. これらの確率表記は y の観測値の条件に基づき, 表記は簡単に $p(\theta|y)$ や $p(\tilde{y}|y)$ で書き表される. また, 共変量 x の既知の値を暗黙的に条件付ける. ベイズ推論が多くの教科書に記載されている統計的推論へのアプローチからはずれているのは, 観察されたデータの条件付けの基本レベルです. これは, θ の真の未知の値を条件として, 可能な y 値の分布に対する θ(または $\tilde{y}$) を推定するために使用される手順の遡及的評価に基づく. この違いにもかかわらず, 多くの単純な分析では, 表面的に類似の結論が統計的推論への 2 つのアプローチの結果であることがわかる. しかし, ベイズ手法を用いて得られた解析は, より複雑な問題に容易に拡張することができる. この節では, ベイジアン推論の基本的な数学と表記を紹介し, 次の節では遺伝学の例を示す.

確率表記

この時点で表記法に関するいくつかのコメントが必要となる. まず, $p(\cdot|\cdot)$ は文脈によって決定される引数を用いた条件付き確率密度を表し, 同様に $p(\cdot)$ は周辺分布を表す. 用語「分布」と「密度」を同じ意味で使用する. 同じ表記法は, 連続密度関数と離散確率質量関数に使用される. 同じ方程式 (または式) における異なる分布は, 例えば, 以下の (1.1) のように, $p(\cdot)$ で表される. 標準の数学的記法を悪用していますが, この方法はコンパクトであり, 離散事象の確率に対して $p(\cdot)$ を使用する標準的な実用に似ている. 文脈によっては, 混乱を避けるため, 事象の確率に $\Pr(\cdot)$ という表記を使用することがある. 例えば, $\Pr(\theta > 2) = \int_{\theta>2} p(\theta)d\theta$. 標準的な分布を使用する場合は, 分布の名前に基づく表記法を使用する. 例えば, θ が平均 μ と分散 σ^2 の正規分布を持つ場合, $\theta \sim \mathrm{N}(\mu, \sigma^2)$ または $p(\theta) = \mathrm{N}(\theta|\mu, \sigma^2)$ と書くか, より明示的に $p(\theta|\mu, \sigma^2) = \mathrm{N}(\theta|\mu, \sigma^2)$ である. ここでは, 確率変数には $\mathrm{N}(\mu, \sigma^2)$, 密度関数には $\mathrm{N}(\theta|\mu, \sigma^2)$ などの表記を使用する. いくつかの標準的な分布に関する記法と公式は, 付録 A に記載されている.

また, すべての正のランダム変数 θ に対して, 次の式を使用することもある. 変動係数は $\mathrm{sd}(\theta)/\mathrm{E}(\theta)$, 幾何平均は $\exp(\mathrm{E}[\log(\theta)]$, 幾何標準偏差は $\exp(\mathrm{sd}[\log(\theta)])$ と定義される.

ベイズ則

y が与えられた下での θ についての確率表記をするためには, θ と y の結合確率分布を提供するモデルから始めなければならない. 結合確率質量または結合密度関数は, しばしば事前分布 $p(\theta)$ およびサンプリング分布 (sampling distribution)(またはデータ分布 (dara distribution))$p(y|\theta)$ と呼ばれる 2 つの密度の積として書くことができる.

$$p(\theta, y) = p(\theta)p(y|\theta)$$

ベイズ則 (Bayes' rule) として知られている条件付き確率の基本特性を使用して, データ y の既知の値へのシンプルな条件付けで, 事後密度が得られる.

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{1.1}$$

ここで, $p(y) = \sum_\theta p(\theta)p(y|\theta)$ で, 総和は θ の取りうる値のすべてでとられる (または, 連続な θ の場合では $\int p(\theta)p(y|\theta)d\theta$).

(1.1) の等価形は, θ に依存せず, 固定された y を有する係数 $p(y)$ を省略して, (1.2) の右辺である非正規化事後密度 (unnormalized posterior density) をもたらす定数と考えることができる.

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{1.2}$$

この式の第 2 項, $p(y|\theta)$ はここでは y の関数ではなく θ の関数として取られます. これらの簡単な数式は, ベイズ推論の技術的なコアをカプセル化します. 特定の応用の主なタスクは, モデル $p(\theta, y)$ を開発し, 適切な方法で $p(\theta|y)$ を要約する計算を実行することである.

予測

しばしば予測推論と呼ばれる, 未知の観測可能な量についての推論を行うため, 似た論理に従う. データ y が考慮される前に, 未知であるが観測可能な y の分布は

$$p(y) = \int p(y, \theta) d\theta = \int p(\theta) p(y|\theta) d\theta \tag{1.3}$$

である. これは y の周辺分布と呼ばれるが, より情報を含む名前は事前予測分布 (prior predictive distribution) である. プロセスの観測の前では条件付きではないため事前で, それが観測可能な量についての分布であるから予測である.

データ y が観測された後, 同じプロセスで未知の観測可能な $\tilde{y}$ を予測することができる. 例えば, $y = (y_1, \cdots, y_n)$ は, スケール上で n 回重み付けされた物体の記録された重みのベクトルであり, $\theta = (\mu, \sigma^2)$ は, 物体の未知の真の重みとスケールのばらつきであり, $\tilde{y}$ は, 予定された新しい計量における物体のまだ記録されていない重みであり得る. $\tilde{y}$ の分布は, 観測された y に対して条件付きであるため事後で, 観測可能な $\tilde{y}$ に対する予測であるため予測で, 事後予測分布 (posterior predictive distribution) と呼ばれる.

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y)d\theta = \int p(\tilde{y}|\theta, y)p(\theta|y)d\theta = \int p(\tilde{y}|\theta)p(\theta|y)d\theta \quad (1.4)$$

2, 3 個目は, θ の事後分布に対する条件付き予測の平均として事後予測分布 (posterior predictive distribution) を表す. 最後は, θ を与えられた y と $\tilde{y}$ の仮定された条件付き独立から従う.

尤度

選ばれた確率モデルにベイズ則を使うことは, データ y が $p(y|\theta)$ を通してのみ事後推論 (1.2) に影響を与えることを意味する. ここで, y に対して, θ の関数としてみなされるとき, 尤度関数と呼ばれる. このようにベイズ推論は, 所与のデータサンプルに対して, 同じ尤度関数を有する 2 つの確率モデル $p(y|\theta)$ が θ について同じ推論をもたらすという尤度原理と呼ばれることもある.

尤度原理 (likelihood principle) は合理的であるが, 特定の分析のために採用されたモデルの枠組みまたはモデル族の中だけで有効である. 実際には, 選択したモデルが正しいとはほとんど確信できない. 6 章では, サンプリング分布 (データの再現を想像して想像すること) がモデルの仮定を調べる上で重要な役割を果たすことを見ることができる. 実際, 適用されたベイズ統計学者の考えは, 可能な様々なモデルの下でベイズ則を適用しようとするものである.

尤度比, オッズ比

与えられたモデルでの点 θ_1 と θ_2 で評価された事後密度 $p(\theta|y)$ の比は, θ_2 と対する θ_1 の事後オッズ (posterior odds) と呼ばれる. この概念の最もよく知られた応用は離散パラメータであり, θ_2 は θ_1 の補集合であると考えられる. オッズは確率の代替表現を提供し, ベイズ則はそれらの観点から表現されるときに特に単純な形式を取るという魅力的な特性を有する.

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)p(y|\theta_1)/p(y)}{p(\theta_2)p(y|\theta_2)/p(y)} = \frac{p(\theta_1)p(y|\theta_1)}{p(\theta_2)p(y|\theta_2)} \tag{1.5}$$

言い換えれば, 事後オッズは, 事前オッズに尤度比 (likelihood ratio)$p(y|\theta_1)/p(y|\theta_2)$ を掛けたものに等しい.

1.4 離散確率の例: 遺伝学とスペルチェック

次に, ベイズの定理について, 目先の目的が母集団全体を記述するパラメータの推定というより, 特定の離散量に関する推論であるような 2 つの例を示します. これらの離散的な例では, 事前, 尤度, 事後確率を直接見ることができる.

遺伝状態についての推論

人間の男性は X 染色体を 1 つ, Y 染色体を 1 つ持っていますが, 女性は 2 つの X 染色体を持ち, 各染色体は両親からそれぞれ継承しています. 血友病は, X 染色体に関連した劣性遺伝を示す疾患であり, X 染色体上に病気を引き起こす遺伝子を遺伝する男性が罹患しているのに対して, 2 つの X 染色体のうちの 1 つにのみ遺伝子を持つ女性は影響を受けない. この疾患は, 2 つのそのような遺伝子を継承する女性にとっては一般的に致命的であり, これはまれである. なぜなら, この遺伝子の出現頻度はヒト集団において低いからである.

事前分布

罹患した兄弟を持つ女性を考えてみましょう. 彼女の母親は, 血友病遺伝子のキャリアでなければならないことを意味します. 彼女の父親は影響を受けていないとも言われています. したがって, 女性自身がその遺伝子を有する可能性が 50%である. 女性の状態である未知の関心量には, ただ 2 つの値しかありません. 女性は遺伝子のキャリア ($\theta = 1$) かそうでないか ($\theta = 0$) のいずれかである. これまでに提供された情報に基づいて, 未知 θ の事前分布は簡単に $\Pr(\theta = 1) = \Pr(\theta = 0) = \frac{1}{2}$ として表すことができる.

データモデルと尤度

事前情報を更新するために使用されるデータは, 女性の息子の影響状態からなる. 彼女には 2 人の息子がいて, どちらも影響を受けていないとします. y_i =1 または 0 は, それぞれ罹患したまたは罹患していない息子を示すとする. 2 人の息子の結果は交換可能であり, 未知の θ の条件付きは独立している. 私たちは息子が一卵性双生児ではないと仮定します. 2 つの独立したデータ項目は, 以下の尤度関数を生成する.

$$\Pr(y_1 = 0, y_2 = 0|\theta = 1) = (0.5)(0.5) = 0.25$$
$$\Pr(y_1 = 0, y_2 = 0|\theta = 0) = (1)(1) = 1$$

これらの表現は, 女性がキャリアである場合, 各息子は遺伝子を継承して影響を受ける確率は 50%であり, キャリアでなければ 1 に近い確率で彼女の息子は影響を受けません. (実際には母親がキャリアではなくても影響を受ける確率はゼロではありませんが, このリスクは (突然変異率は小さく), この例では無視できます).

事後分布

ベイズ則は, データ内の情報を事前確率と組み合わせるために使用される. 特に, 関心は, 女性がキャリアである事後確率に焦点を当てる可能性が高い. 結合データ (y_1, y_2) を表すために y を使うと, これはシンプルに

$$\begin{aligned}\Pr(\theta = 1|y) &= \frac{p(y|\theta = 1)\Pr(\theta = 1)}{p(y|\theta = 0)\Pr(\theta = 0) + p(y|\theta = 1)\Pr(\theta = 1)} \\ &= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1)(0.5)} = \frac{0.125}{0.625} = 0.20\end{aligned}$$

である. 直感的には, 女性に罹患していない子供がいる場合, 彼女がキャリアである可能性は低く, ベイズ則は補正の程度を決定する正式なメカニズムを提供することは明らかである. 結果は, 事前オッズおよび事後オッズの観点からも記述することができる. キャリアである女性の事前オッズは $0.5/0.5 = 1$ である. 彼女の 2 人の罹患していない息子についての情報に基づく尤度比は $0.25/1 = 0.25$ であるので, 事後確率は $0.25/1 = 0.25$ である. 確率に戻すと, 以前のように $0.25/(1 + 0.25) = 0.2$ が得らる.

データの追加

ベイズ分析の重要な側面は, 逐次分析を実行することが容易であることです. 例えば, その女性にも影響を受けていない第三の息子がいるとします. 計算全体をやり直す必要はないのだ. むしろ, 以前の事後分布を新しい事前分布として使用して

$$\Pr(\theta = 1|y_1, y_2, y_3) = \frac{(0.5)(0.20)}{(0.5)(0.20) + (1)(0.8)} = 0.111$$

を得る. あるいは, 第 3 の息子が影響を受けていると仮定すると, キャリアである女性の事後確率が 1 になる (突然変異の可能性は無視する) ことが容易に確認できます.

スペル訂正

言葉の分類は不確実性を管理する問題である. たとえば, 誰かが 'radom' と入力したとします. どう読むべきだろうか? それは, 'random' または 'radon' または他のもののスペルミスまたはミスタイプである可能性があります. もしくは, (または, この段落での最初の使用のように)'radom' の意図的な入力である可能性があります. 'radom' が実際に 'random' を意味する確率どのくらいだろうか? y をデータ, θ を人が入力しようとしていた単語とすると,

$$\Pr(\theta|y = \text{'radom'}) \propto p(\theta)\Pr(y = \text{'radom'}|\theta) \tag{1.6}$$

この積は非正規化された事後密度である. この場合, 簡単化のためにありそうな単語 θ に対して 3 つの可能性 (random,randon,ranom) しか考えていないとすると, まず θ の 3 つすべての値に対して非正規化された密度を計算し, 次に正規化をすることで, 興味のある事後確率を計算できる.

$$p(\text{random}|\text{'radom'}) = \frac{p(\theta_1)p(\text{'radom'}|\theta_1)}{\sum_{j=1}^{3} p(\theta_j)p(\text{'radom'}|\theta_j)}$$

ここで, θ_1 =random, θ_2 =radon, θ_3 =radom である.

事前確率 $p(\theta_j)$ は, 最も単純には, ある大きなデータベース内のこれらの単語の頻度から, 理想的には手元の問題に適合したもの (例えば, 問題の単語が電子メールのような文書に現れている場合, 最近の学生の電子メールのデータベース) から来ることができる. 尤度 $p(y|\theta_j)$ は, 疑わしい単語を特定するために電子メールを書いた後に人々がフォローアップされたいくつかの調査を使用して, おそらく適合する, スペルとタイピングエラーのいくつかのモデリングから来る可能性があります.

事前分布

他の文脈がなければ, いくつかのデータベースにおいて, これらの 3 つの単語の相対頻度に基づいて事前確率 $p(\theta_j)$ を割り当てることは意味がある. ここでは Google の研究者が提供する確率,

θ	$p(\theta)$
random	7.60×10^{-5}
radon	6.05×10^{-6}
radom	3.12×10^{-7}

を用いる. これらの可能性のみを検討しているので, 3 つの数値を合計して 1 にすること ($p(\text{random}) = \frac{760}{760+60.5+3.12}$) ができますが, 調整は (1.6) の比例定数に吸収されるため, 必要はありません.

上記の表に戻って, 私たちは, コーパスの ‘radom’ がそれほど高い確率であることに驚きました. 私たちは Wikipedia でその言葉を調べ, それが中規模の都市であることを発見しました. ‘ポーランド最大かつ最高の参加型エアショーの本拠地です... また, ポーランドのデザインの半自動 9mm パラピストルの人気のある非公式の名前です...’ 私たちが遭遇する文書では, ‘radom’ の相対的確率は非常に高いようです. 上記の確率が当社のアプリケーションに適切ではないように見える場合, これは, モデルにまだ含まれていない事前の情報または信念があることを意味します. この例のモデルの意味を理解した後, この点に戻ります.

尤度

Google のスペルと入力エラーのモデルの条件付き確率は

θ	$p(\text{‘radom’}\mid\theta)$
random	0.00193
radon	0.000143
radom	0.975

である. この尤度関数は確率分布ではないことを強調します. むしろ, 3 つの異なる確率分布からの特定の結果 (‘radom’) の条件付き確率の集合であり, 未知パラメータ θ の 3 つの異なる可能性に対応する.

これらの特定の値は, この特定の 5 文字の単語が正しく入力される確率は 97%, 文字を誤って’random’ から打ち漏らすことによる確率は 0.2%で, それよりも低い確率で ‘radon’ の最後の文字を打ち間違える. Google はこれらの確率について強い直観を持っておらず, Google のエンジニアをここで信頼します.

事後分布

事前確率と尤度を掛け合わせて確率を計算し, 次に再正規化して事後確率を得る.

θ	$p(\theta)p(\text{`radom'}\|\theta)$	$p(\theta\|\text{`radom'})$
random	1.47×10^{-7}	0.325
radon	8.65×10^{-10}	0.002
radom	3.04×10^{-7}	0.673

したがって, モデル上の条件付きでは, 型付き単語 'radom' は, 'random' の誤植であるように約 2 倍の正解であり, 誤った 'radon' の例である可能性は非常に低い. 完全な分析には, これらの 3 つの単語を超える可能性が含まれますが, 基本的な考え方は同じです.

意思決定, モデル検査, モデル改善

私たちはここから 2 つの方向を想像することができます. 最初のアプローチは, 単語が正しく入力された確率の 3 分の 2 を受け入れることです. あるいは, 最初のパスで 'radom' を正しいと宣言するだけです. 2 つ目の選択肢は, 例えば, 'radom' がタイプミスのように見えること, およびそれが正しいと推定される確率が非常に高いように言うことによって, この確率に疑問を呈することであろう.

私たちが事後分布の主張に異議を申し立てるとき, そのモデルはデータに適合しない, あるいはこれまでモデルに含まれていなかった追加の事前情報があると言っています. この場合, 我々は 1 つの単語だけを調べているので, 適合の欠如は問題ではない. したがって, 事後的な紛争は, 事前または尤度のいずれかで, 追加情報の主張に対応しなければならない.

この問題については, 尤度を批判する特別な理由はありません. 一方で, 事前確率は, 文脈依存性が非常に高い. 'ランダム' という言葉は当然のことながら, 私たち自身の統計書に頻繁に掲載されています. 私達にとって 'radon' はまったく新しいものでしたが, 'radon' が時々発生します (9.4 節を参照). 高い確率の 'radom' での私たちの驚きは, 私たちの特定の問題に関連する追加の知識を表しています.

モデルは, 事前確率に文脈情報を含めることによって, 最も直ちに詳述することができる. 例えば, 調査中の文書が統計書である場合, その人が 'random' と入力する可能性が高くなります. x がモデルによって使用される文脈情報としてラベル付けされると, ベイジアン計算は,

$$p(\theta|x, y) \propto p(\theta|x)p(y|\theta, x)$$

となる.

最初の近似では, 最後の項を $p(y|\theta)$ に単純化することで, 特定の誤りの確率 (つまり, 意図された単語 θ を与えた特定の文字列 y を入力する確率) は文脈に依存しない. これは完全な仮定ではありませんが, モデリングと計算の負担を軽減できます.

ベイジアン推論の実際の課題には, これらの確率をデータから推定するためのモデルの設定が含まれます. その時点で, 上に示したように, ベイズ則は, 手元の問題に対するモデルの意義を決定するために簡単に適用することができます.

1.5 不確実性の尺度としての確率

我々はすでに確率密度などの概念を用いてきたが, 実際には, 読者が基本的な確率理論にかなり精通していると仮定している (しかし, 1.8 節では, ベイズ解析でよく起こる確率計算の簡単な技術的レビューを提供する). しかし, ベイズの枠組み内での確率の使用は非ベイズ統計よりもはるかに広いので, より詳細な統計的例を検討する前に, 確率の概念の基礎を少なくとも簡単に検討することが重要である. 確率の数学的定義について, 読者側で共通の理解を当然とらえています. 確率は相互に排他的な結果に対して非負の加算値である ‘結果’ の集合に定義された数量であり, 相互に排他的な可能性のあるすべての結果に対して 1 になる.

ベイズ統計では, 確率は, 不確実性の基本的な尺度として使用されます. このパラダイムの中で, コイントスが表である確率を議論するのと同じように, サッカーワールドカップでの「明日の雨」の確率やブラジルの勝利の可能性について議論することは同じように正当です. したがって, ある範囲にあるサイズ 100 の既知の固定母集団からの 10 項目の無作為標本の平均がある範囲内にある確率を考慮するように, 未知の推定値が特定の値の範囲にある確率を考慮することは自然なこととなる.

これらの 2 つの確率のうちの 1 つ目は, データが取得された後により重要であるが, 2 つ目はデータが取得される前により関連性が高い. ベイズ手法は, ある種の状況や ‘自然の自由’(観測不可能であるか, まだ観測されていない) に関する利用可能な部分的知識 (尺度としての確率) を体系的な方法で尺度として利用することを可能にする. 指針の原則は, 未知のものに関する知識の状態が確率分布によって記述されるということです.

不確実性の数値的尺度は何を意味していますか? 例えば, コイン投げの表の確率は $\frac{1}{2}$ になることが広く認められています. なぜこれはそうですか? 2 つの正当化が一般的に与えられているようです.

(1) 対称性または交換可能性の議論. 同様の可能性を前提として,

$$\text{確率} = \frac{\text{満たされた場合の数}}{\text{可能性の数}}$$

コイントスの場合, これは実際には, コインが落ちる方法を決定する際に働く力についての仮定, ならびにトスの初期の物理的条件に基づいた, 物理的な議論である.

(2) 頻度の議論. 確率=トスの長いシーケンスで得られた相対的な頻度. 物理的には互いに独立して同じ方法で実行されると仮定される.

上記の議論は両方とも, コインの性質と投げ方についての判断を必要とする点で主観的であり, 同様に起こりうる事象の意味, 同一の測定, 独立についての意味論的議論を伴う. 頻度議論は, 同一のトスの長いシーケンスの仮説を含むという点で, ある特別な困難を有すると知覚されるかもしれない. 厳密に言えば, この考え方は, 同一の事象の長いシーケンスにおいて, 少なくとも概念的に, 埋め込まれることのない単一のコイン投げに対する確率の記述を可能にしない.

次の例は, 確率判断がますます主観的になりうる方法を示しています. まず, 以下の修正されたコイン実験を考えてみましょう. 特定のコインが, さらに情報が提供されないで, 両面表か両面裏かのいずれかであると述べられているとします. まだ表の確率について議論することはできますか? 一般的な言い分では確かにできることは明らかです. おそらく, この新しい確率をどのように評価するかははっきりしていませんが, 多くの場合, おそらくラベル表と裏の交換可能性に基づいて, $\frac{1}{2}$ の同じ値に同意します.

次にいくつかの例を考えてみましょう. 明日のサッカーでコロンビアがブラジルを試合をするとしましょう. コロンビアが勝利する確率は? 明日の雨の確率は? 明日雨が降ったら, コロンビアが勝つ確率は? ロケット打ち上げが失敗する確率はどのくらいであろうか? これらの質問のそれぞれは, 常識的には合理的であるように見えますが, 参照される確率についての強い頻度の解釈を検討することは困難です. しかし, 通常は周波数解釈が構築でき, これは統計上非常に有用なツールです. 例えば, 将来のロケット打ち上げを, 同じ種類の打ち上げ予定の母集団からのサンプルと見なし, 失敗した過去の打ち上げの頻度を見てみることができます (この例に関する詳細については, この章の最後の書誌注記を参照してください). このようなことを科学的に行うということは, 確率モデル (あるいは, 少なくとも類似の事象の参照集合 (reference set)) を作成することを意味し, これは簡単なコイントスに類似した状況に戻ります. これは, 問題の結果を交換可能であるとみなし, したがって同様に可能性があると考える必要があります.

確率が不確実性を定量化する合理的な方法なのはなぜですか? 以下の理由はしばしば進歩している.

(1) 類推によって. 物理的なランダム性は不確実性を招くので, ランダムな事象の言語における不確実性を記述することは妥当と思われる. 一般的な発言は, 「おそらく」や「起こりそうもない」などの多くの用語を使用しており, より正式な確率計算を科学的推論の問題に拡張するための使用法と一致しているようです.

(2) 公理的または規範的アプローチ. このアプローチでは, すべての統計的推論を意思決定のコンテキストで利得と損失に置きます. 合理的公理 (秩序, 推移など) は, 不確実性を確率の形で表現しなければならないことを意味する. この規範的論理的根拠は, 示唆的だが魅力的ではないと考える.

(3) 賭けの一貫性. が発生した場合に$1 の返品に対して交換する (すなわち賭け)$$p$ の部分 ($p \in [0, 1]$) としてイベント E に付随する確率 p を定義する. つまり, E が発生すると, $(1 - p)$ を得ることができます. E の補修号が出れば, $$p$ を失う. 例えば,

・コイントス. コイントスを公正な賭けと考えることは, $p = \frac{1}{2}$ に対応する均等オッズを示唆している.

・ゲームに対するオッズ. B チームに対して 10 対 1 のオッズで試合に勝つために A チームに賭けることを望むなら (すなわち, 勝って 1 が 10 になる), A チームの勝者の確率は少なくとも $\frac{1}{11}$ です.

一貫性の原則は, すべての可能性のある出来事に確率を割り当てることは, あなたと賭けて明確な利益を得ることができないようなものでなければならないと述べています. この原理で構築された確率は, 確率論の基本的な公理を満たさなければならないことが証明できる. 賭けの根拠には基本的な困難がいくつかあります.

・すべてのイベントに対して, いずれの方向にもベットする意思がある正確なオッズが必要です. あなたが確信が持てない場合は, 正確なオッズをどのように割り当てることができますか?

・もしある人があなたと賭けたいと思っていて, あなたが持っていない情報があれば, 賭けをするのは賢明ではないかもしれません. 実際には, 確率は不完全な (必要ではあるが十分ではない) 賭けのガイドです.

これらのすべての考慮事項は, 適用統計値の不確実性を要約するための確率が合理的なアプローチである可能性を示唆していますが, 究極の証拠はアプリケーションの成功にあります. この本の残りの章では, 確率は統計アプリケーションにおける不確実性を処理するための豊かで柔軟な枠組みを提供することを示しています.

主観性と客観性

確率を使用するすべての統計的方法は, 世界の数学的理想化に頼っているという意味で主観的です. ベイズ手法は, 事前分布に依存しているために, 特に主観的であると言われることがありますが, ほとんどの問題では, モデルの「尤度」と「事前」の両方を指定する科学的判断が必要です. 例えば, 線形回帰モデルは, 一般に, 少なくとも回帰パラメータについて想定される事前分布の少なくとも疑わしいものである. 一般的な原則はここで働いています. 複製が存在するときはいつでも, 観察される多くの交換可能ユニットの意味において, データからの確率分布の特徴を推定し, 分析をより客観的にする範囲がある. 実験全体が数回複製された場合, 5章で説明したように, 事前分布のパラメータ自体をデータから推定することができます. しかし, いずれの場合でも, 科学的判断を必要とする特定の要素, 特に解析に含まれるデータの選択, 分布について想定されているパラメトリックな形, モデルがチェックされる方法が残るであろう.

1.6 例: フットボールのポイントスプレッドについての確率

経験的データともっともらしい実証的仮定を用いて確率を割り当てる方法の例として, 私たちは専門的な (アメリカン) フットボールの試合における一定の成果の確率を見積もる方法を検討する. これは, ベイズ推論ではなく, 確率割り当ての一例です. フットボールの試合結果に確率を割り当てる多くのアプローチが示されています. 観測データに基づく経験的確率を用いて主観的評価を行い, パラメトリック確率モデルを構築する.

フットボールのポイントスプレッドと試合結果

フットボールの専門家は, 2 つのチームの能力の違いの尺度として, すべてのフットボールの試合にポイントスプレッドを与えている. 例えば, チーム A は 3.5 ポイント差でチーム B を倒すかもしれない. このポイントスプレッドの意味は, 好きなチーム A がチーム B を負け, 弱者を 4 ポイント以上打ち負かすという命題が公正な賭けとみなされるということです. すなわち, A が 3.5 ポイント以上勝つ確率は $\frac{1}{2}$ である. ポイントスプレッドが整数である場合, チーム A はポイントスプレッドよりも多くのポイントで勝つ可能性が高いということは, ポイントスプレッドよりも少ないポイントで勝つ (または失う) ためです. A がポイントスプレッドだけで勝つという肯定的な確率があり, その場合どちらの側も払われない. ポイントスプレッドの割り当て自体は, 確率論的推論における興味深い演習です. 1 つの解釈は, ポイントスプレッドがゲームの可能な結果についてのギャンブル人口の信念の分布の中央値であるということである.

1981, 1983, 1984 年のシーズンに行われたプロフットボールの 672 試合のポイントスプレッドと実際のゲーム結果は, 図 1.1 にグラフで示されています. (労働争議のため 1982 年のシーズンの多くは中止された.) 散布図の各点は, ポイントスプレッド x と, 実際の結果 (お気に入りのスコアから弱者のスコアを差し引いたもの)y を表示します. (ポイントスプレッドがゼロのゲームでは, 「お気に入り」と「弱者」というラベルがランダムに割り当てられていました). 小さなランダムジッタがグラフの各点の x 座標と y 座標に加算され, 複数の点が正確に重なり合わないようにします.

観測された頻度に基づいて確率を割り当てる

特定のイベントに確率を割り当てることは重要です. Pr(お気に入りの勝利), Pr(お気に入りの勝利|ポイントスプレッドは 3.5 ポイント), Pr(ポイントスプレッド以上のお気に入り勝利), Pr(ポイントスプレッド以上の好きな勝利は 3.5 ポイント|ポイントスプレッドは 3.5 ポイント) など. 我々は, 新聞を読んでフットボールの試合を見ることによって集められた非公式の経験に基づいて, 主観的な確率を報告するかもしれない. 優勝チームがゲームに勝つ確率は確かに 0.5 より大きく, おそらく 0.6 から 0.75 の間でなければなりません. より複雑なイベントでは, 私たちの面でより直感的な知識が必要です. より体系的なアプローチは, 図 1.1 のデータに基づいて確率を割り当てることです. 繋がれたゲームを半分の勝利と半分の敗北として数え, ポイントスプレッドがゼロであるゲームを無視する (従ってお気に入りはない) とすると, 以下のような経験的な見積もりが得られる.

- Pr(お気に入りの勝利)$=\frac{410.5}{655} = 0.63$
- Pr(お気に入りの勝利$|x = 3.5$)$=\frac{36}{59} = 0.61$
- Pr(ポイントスプレッド以上でのお気に入りの勝利)$=\frac{308}{655} = 0.47$
- Pr(ポイントスプレッド以上でのお気に入りの勝利$|x = 3.5$)$=\frac{32}{59} = 0.54$

これらの経験的確率の割り当てはすべて, 知識のあるフットボールファンの直感と一致するので賢明であるように見える. しかし, このような確率の割り当ては, 直接的に関連するデータ点の少ないイベントでは問題になります. 例えば, 8.5 ポイントのお気に入りは, この 3 年間で 5 回のうち 5 回を獲得しましたが, 9 ポイントのお気に入りは 20 回のうち 13 回勝ったのです. しかし, 私たちは, 8.5 ポイントのお気に入りよりも 9 ポイントのお気に入りのほうが勝つ確率が高くなることを現実的に予想しています. ポイントスプレッドが 8.5 の小さなサンプルサイズは, 不正確な確率の割り当てにつながります. 我々は, パラメトリックモデルを使用する代わりの方法を検討する.

試合結果とポイントスプレッドの違いのパラメトリックモデル

図 1.2a に, フットボールデータセットの試合の, 観測された試合結果とポイントスプレッドの差 $y - x$ がプロットされています. (もう一度, ランダムジッタが両方の座標に追加されました.) このプロットは, x とは無関係に $y - x$ の分布をモデル化するのがおおよその可能性があることを示唆しています. (演習 6.10 を参照) 図 1.2b は, フィッティングされた正規密度が重ねられたすべての試合の差 $y - x$ のヒストグラムです. このプロットは, 確率変数 $d = y - x$ の周辺分布を正規分布で近似することが合理的であることを示唆しています. d の 672 個の値の標本平均は 0.07 であり, 標本標準偏差は 13.86 であり, フットボールゲームの結果はポイントスプレッドおよび標準偏差にほぼ等しい平均で約 14 点 (2 回のコンバーテッドタッチダウン) であることを示唆している. 議論の残りの部分については, d の分布を, x およびノーマルとは無関係にし, 各 x について平均ゼロおよび標準偏差 14 を用いる. つまり, 図 1.2b で示されるように,

$$d|x \sim \mathrm{N}(0, 14^2)$$

である. 割り当てられた確率モデルは完全ではありません. それはデータに正確に適合せず, 実際のデータでよく見られるように, フットボールのスコアもポイントスプレッドも連続値の量ではありません.

パラメトリックモデルを餅入り確率の割当

それにもかかわらず, モデルは事象に確率を割り当てるのに使用できる便利な近似を提供する. d が平均ゼロの正規分布を持ち, ポイントスプレッドとは無関係であれば, 好きなチームがポイントスプレッド以上に勝つ確率は, ポイントスプレッドの任意の値を条件とし, したがって無条件に $\frac{1}{2}$ である. ノーマルモデルが $\Pr_{\text{norm}}$ として得た確率を表すと, ノーマルモデルを仮定して, x ポイントのお気に入りがゲームに勝つ確率は, 以下のように計算することができる.

$$\Pr_{\text{norm}}(y > 0|x) = \Pr_{\text{norm}}(d > -x|x) = 1 - \Phi\left(-\frac{x}{14}\right)$$

ここで, Φ は標準正規累積分布関数である. 例えば,

- $\Pr_{\text{norm}}$(お気に入りが勝つ$|x = 3.5) = 0.60$
- $\Pr_{\text{norm}}$(お気に入りが勝つ$|x = 8.5) = 0.73$
- $\Pr_{\text{norm}}$(お気に入りが勝つ$|x = 9.0) = 0.74$.

3.5 ポイントのお気に入りの確率は先に与えた経験値と一致しますが, 8.5 ポイントと 9 ポイントのお気に入りの確率は, 小さなサンプルに基づく経験値よりも直感的になります.

1.7 例: レコードリンケージの精度の推定

私たちは, データから推定される別の例を用いて, 本質的に経験的な (主観的または個人的な) 確率の性質を強調する.

レコードリンケージとは, 同じ個人に対応する異なるデータベースのレコードを識別するためのアルゴリズム技術の使用を指します. レコードリンケージ技術は, さまざまな設定で使用されます. ここに記載されている作業は, 米国との間のレコードリンケージの中で策定され, 最初に適用された. 国勢調査と大規模なポスト列挙調査は, 母集団のサブグループのために国勢調査の適用範囲を評価するために実施された広範な照合作業の第一歩です. この最初のステップの目的は, コンピュータによって可能な限り多くのレコードをコンピュータで「一致」させ, 過度のエラー率なしに宣言し, 一致したと宣言されていないすべてのレコードに対して手作業による処理のコストを回避することです.

潜在的な一致にスコアを割り当てる既存の方法

レコードリンケージの文献では, 多変量レコードの個々の情報フィールドに「重み」を割り当て, 2 つのレコード間の一致の近さをまとめた y という複合スコアを得るという問題に多くの注意が払われています. ここでは, これらのルールが選択されているという意味でこのステップが完了したと仮定します. 次のステップは, 一致する候補ペアの割り当てであり, 各レコードのペアは, それぞれのデータベースからの最良の潜在的な一致からなる. 指定された重み付けルールは, 次に, 一致する候補ペアを順序付けます. 国勢調査局の動機づけ問題では, 代替案の「宣言の一致」と「フォローアップへの送信」との間でバイナリの選択が行われます. ここでは, レコードが一致すると宣言されたカットオフのスコアが必要です. 次に, 誤一致率は, 誤って一致した対の数を, 一致した対の数で除した数として定義される.

このような決定問題に特に関係するのは, マッチした候補ペアがそのスコアの関数として正しいマッチである確率を評価するための正確な方法である. スコアを確率に変換するための単純な方法が存在するが, これらは非常に不正確で, 一般的には楽観的な, 誤った一致率の推定につながる. 例えば, 10^{-3} から 10^{-7} の範囲の名目偽一致確率 (すなわち, 一致とほぼ確実とみなされる対) を有するレコードのセットの手動チェックは, 1%の範囲に近い実際の偽の一致率を見出した. 名目偽一致確率が 1%の記録は, 実際の偽一致率が 5%であった.

フットボールの例と同じように, ベイズ手法を使用して, これらを再較正して, 与えられた決定ルールの客観的な確率を得ることができます. 過去のデータを使用して, ポイントスプレッドに条件付きの異なるゲーム結果の確率を見積もった. 私たちのアプローチは, スコア y で作業し, y の関数として一致の確率を経験的に推定することです.

一致確率の経験的な推定

混合モデリングを使用して, 正確な一致確率を得ることができ, これについては 22 章で詳しく説明します. 候補マッチの以前に得られたスコアの分布は, 真のマッチのスコアの分布と不一致の分布の「混合」とみなされる. 混合モデルのパラメータは, データから推定される. 推定されたパラメータは, スコア上の任意の所定の決定閾値についての偽の一致の確率の推定値 (真の一致とは異なると宣言されたペア) を計算することを可能にする. 実際に使用された手順では, 混合モデルのいくつかの要素 (正規分布の混合を許容するのに必要な最適な変換など) は, 既知の一致ステータスのトレーニングデータを (校正手順を適用したデータとは別に) 使用して適合されましたが, 詳細はこちら. 代わりに, 未知の一致状況を持つ一連のデータで手法をどのように使用するかに焦点を当てます.

このアプローチのサポートは, 図 1.3 に示されています. これは, 1990 年の国勢調査の 2 年前に単一の地方で実施された 'test Census' 調査からの 2300 の記録から得られた特定のデータセットにおける一致および不一致のスコアの分布を表示します. 2 つの分布 $p(y|\text{match})$ と $p(y|\text{non-match})$ は大部分が区別されます. つまり, ほとんどの場合, 候補をマッチとして識別することも, スコアだけを指定することもできません. しかし, いくつか重複している.

アプリケーションデータセットでは, 一致状況がわかりません. したがって, 我々は, 2 つの成分分布および各成分に属するスコアの集団の割合を推定する単一の組み合わせヒストグラムに直面する. 混合モデルの下では, スコアの分布は

$$p(y) = \Pr(\text{match})p(y|\text{match}) + \Pr(\text{non-match})p(y|\text{non-match}) \qquad (1.7)$$

と書ける. 混合確率 (Pr(match)) と一致と不一致の分布のパラメータは, 一致状態が未知のデータから結合されたヒストグラムに適用された (22 章で記述されるような) 混合モデルアプローチを用いて推定される.

この手法を使用してレコードリンケージの決定を行うには, 判定閾値の関数としての誤一致率を与える曲線を作成します. この値を超えるスコアは一致すると宣言されます. 所与の決定閾値について, 確率分布は, 分布 $p(y|\text{non-match})$ に由来する閾値より上のスコア y である, 偽の一致の確率を推定するために (1.7) の確率分布を使用することができる. 閾値が低いほど, より多くのペアが一致として宣言します. より多くの一致を宣言するにつれて, エラーの割合が増加します. ここで説明するアプローチは, 各しきい値について客観的な誤差の推定量を提供する必要があります. (次の段落の検証を参照してください). 意思決定者は, より多くのマッチを自動的に宣言し, 間違いを少なくするという目的の間に, 許容できるバランスを提供する閾値を決定することができます.

テストデータを使用した確率の外部検証

上記のアプローチは, 一致状態が分かっているデータを使用して外部から検証されました. この方法は, 1988 年の国勢調査の 3 つの異なる場所からのデータに適用され, その方法の 3 つのテストが可能であった. 我々は 1 つの詳細な結果を提供する. 他の 2 つの結果も同様でした. 混合物モデルを, 試験部位の全ての候補対のスコアに適合させた. 次に, 推定されたモデルを用いて図 1.4 の線を作成しました. この線は, 一致すると宣言された症例の割合に関して予想される偽の一致率 (および不確かさの範囲) を示します. (ほとんどの候補ペアをマッチと宣言します). 誤一致の割合は, 宣言された一致の数の関数であり, 意味があります. グラフを右に移動するにつれて, 弱く弱いケースが一致すると宣言しています.

図 1.4 の線は, モデルから推定された誤一致率の予想割合と誤一致率の 95%事後境界を表示します. (これらの境界は, 誤一致率が 95%の事後確率が存在する推定範囲を与える. 事後間隔の概念については, 次の章で詳しく説明します). グラフの点には実際の誤一致率が表示されます. これはモデルとよく一致します. 特に, このモデルは, 一致した場合の 90%未満を宣言し, 誤った一致の大部分を避けるために, 他の 10%程度をあきらめるよう勧告し, ドットは同様のパターンを示す.

大量のファイルをほとんどエラーなく一致させることは可能です. また, 候補一致の質は, 誤一致率が加速するある時点で劇的に悪化する. 図 1.5 は, 偽の一致率が加速する関心領域における較正手順の挙動を強調するために, 前のディスプレイに拡大鏡を表示しています. 予測される擬似一致率曲線は, 観察された偽マッチ率曲線が急峻に上昇する点の近くで上方に曲がり, これは較正方法の特に有利な特徴である. 較正手順は, 実際の確率に近い予測された確率および有益でありかつ真の値を含む区間推定を提供するという観点から良好に機能する. 比較すると, 経験較正なしで重みを掛けることによって構築された一致確率の元の推定値は非常に不正確であった.

1.8 確率理論によるいくつかの有用な結果

読者は確率と確率分布を含む基本操作に精通していると仮定する. 特に, 本の主要部分について十分に理解されなければならない基本的な確率の背景には, 結合密度の操作, 単純なモーメントの定義, 変数の変換, およびシミュレーションの方法が含まれる. この節では, これらの前提条件を簡単に見直し, 本書の残りの部分で使用されるいくつかのさらなる表記法を明確にします. 付録 A は, 一般的に使用される確率分布に関する情報を提供します.

1.3 節で紹介したように, 我々は一般に, それぞれの変数に与えられた名前を反映する仮引数を用いて, 結合確率質量または結合密度関数による結合分布を表す. したがって, 2 つの量 u と v について, 結合密度を $p(u, v)$ と書く. 特定の値を参照する必要がある場合, この表記法は, 例えば $p(u, v = 1)$ のようにさらに濫用されます.

結合密度 $p(u, v)$ に関するベイズ計算では, $p(u|v)$ のような条件付き分布または密度関数, および $p(u) = \int p(u, v)dv$ のような周辺密度に関係する. この表記法では, u と v のどちらかまたは両方をベクトルにすることができます. 典型的には, 文脈から明らかなように, 後者の表現における積分の範囲は, 統合された変数の全範囲を指す. 周辺密度と条件密度の積として結合密度を因数分解することもしばしば有用である. 例えば, $p(u, v, w) = p(u|v, w)p(v|w)p(w)$ である.

たとえば, $\pi(\theta)$, $f(y|\theta)$ のように, パラメータや観測値の分布に異なる表記を使用する著者もいますが, すべての確率分布がベイジアン推論で同じ論理状態を持つということはあいまいです. しかし, 適切な条件付けを示すためには常に注意する必要があります. 例えば, $p(y|\theta)$ は $p(y)$ とは異なる. しかし, 簡潔にするために, 私たちの表記法は, 全体にわたって保持する仮説の条件付けを隠します. つまり, 空白で判断を下すことはできません. さらに明示的にするには, 次のような表記を使用します.

$$p(\theta, y|H) = p(\theta|H)p(y|\theta, H)$$

ここで, H は, モデルを定義するために使用される一連の仮説または仮定を指す. また, 既知の説明変数 x に対して明示的な条件付けを省略することもあります.

平均, 分散に対して標準的な表記 E(·), var(·) を用いる.

$$\mathrm{E}(u) = \int up(u)du, \ \mathrm{var} = \int (u - \mathrm{E}(u))^2 p(u)du$$

ベクトルパラメータ u については, 平均の式は同じであり, 共分散行列は

$$\mathrm{var}(u) = \int (u - \mathrm{E}(u))(u - \mathrm{E}(u))^{\mathrm{T}} p(u)du$$

u は列ベクトルと見なされます. (分散行列と共分散行列という用語を同じ意味で使用しています). E(u) と var(u) は実際には変数 u ではなく分布関数 $p(u)$ の関数なので, この表記法はやや不正確です. 期待を伴う式では, 条件変数として明示的に現れない変数は全て期待値に統合されていると仮定される. 例えば, E($u|v$) は, v を固定した状態での u の条件付き期待値, すなわち v の関数としての条件付き期待値であり, E(u) は u の期待値である.(v と同じように)

条件付き確率を用いるモデリング

有用な確率モデルは, より複雑な無条件分布ではなく, 条件付きまたは階層的に観測値の分布を表現することが多い. たとえば, ランダムに選択された大学生の高さを y とします. 周辺分布 $p(y)$ は, (本質的に) ほぼ 160 と 175cm を中心とする 2 つのほぼ正規分布の混合である. y の分布のより有用な記述は, 身長と性別の共同分布に基づいている. $p(y|\text{female})$ および $p(y|\text{male})$ が平均 160 および 175cm でそれぞれほぼ正常であるという条件付き仕様とともに, $p(\text{male}) \approx p(\text{female}) \approx \frac{1}{2}$ である. 条件付き分散があまり大きくなければ, y の周辺分布は二峰性である. 一般に, 複雑な周辺分布ではなく, 追加の変数を使用して階層構造で複雑さをモデル化することを推奨します. 追加の変数が観測されないか. 22 章で説明したように, このテーマは混合モデルの基礎となります. 私たちは繰り返し, 本の中で条件付きモデリングのテーマに戻ります.

条件付き分布の平均と分散

いくつかの関連量 v を与えられた条件付き平均と分散の観点から確率変数 u の平均と分散を表現することはしばしば有用である. u の平均は, v の周辺分布にわたり条件付き平均を平均することによって得られ,

$$\mathrm{E}(u) = \mathrm{E}(\mathrm{E}(u|v)) \tag{1.8}$$

ここで, 内側の期待値は u での v の条件付き平均で, 外側の期待値は v にわたって取られる. 等式 (1.8) は, u と v の共同分布の観点から期待値を書いて, 次に結合分布を因数分解することによって導出するのが簡単です.

$$\mathrm{E}(u) = \int\int up(u,v)dudv = \int\int up(u|v)dup(v)dv = \int \mathrm{E}(u|v)p(v)dv$$

分散に対する対応する結果は, 条件付き分散の平均および条件付き平均の分散の 2 つの項を含む.

$$\mathrm{var}(u) = \mathrm{E}(\mathrm{var}(u|v)) + \mathrm{var}(\mathrm{E}(u|v)) \tag{1.9}$$

この結果は, (1.9) の右側の項を拡張することによって得られます.

$$
\begin{aligned}
&\mathrm{E}(\mathrm{var}(u|v)) + \mathrm{var}(\mathrm{E}(u|v)) \\
&= \mathrm{E}(\mathrm{E}(u^2|v) - (\mathrm{E}(u|v))^2) + \mathrm{E}((\mathrm{E}(u|v))^2) - (\mathrm{E}(\mathrm{E}(u|v)))^2 \\
&= \mathrm{E}(u^2) - \mathrm{E}((\mathrm{E}(u|v))^2) + \mathrm{E}((\mathrm{E}(u|v))^2) - (\mathrm{E}(u))^2 \\
&= \mathrm{E}(u^2) - (\mathrm{E}(u))^2 \\
&= \mathrm{var}(u)
\end{aligned}
$$

アイデンティティ(1.8) と (1.9) は, u がベクトルの場合にも成り立ちます. この場合, $\mathrm{E}(u)$ はベクトル, $\mathrm{var}(u)$ は行列です.

変数の変換

確率分布をあるパラメータ化から別のパラメータ化に変換することは一般的です. ここでは, 変換された空間上の確率密度についての基本的な結果を検討する. わかりやすくするために, ここでは通常の表記 $p(\cdot)$ の代わりに添字を使用します. $p_u(u)$ がベクトル u の密度であり, $v = f(u)$ に変換するとします. ここで, v は u と同じ数の成分を持ちます.

p_u が離散分布であり, f が 1 対 1 の関数である場合, v の密度は

$$p_v(v) = p_u(f^{-1}(v))$$

によって与えられる. f が多対 1 の関数である場合, 逆関数の各枝に対応する 1 つの項を有する $p_v(v)$ について, この式の右側に項の和が現れる.

p_u が連続分布であり, $v = f(u)$ が 1 対 1 変換である場合, 変換されたベクトルの結合密度は

$$p_v(v) = |J| p_u(f^{-1}(v))$$

である. ただし, $|J|$ は v の関数としての変換 $u = f^{-1}(v)$ のヤコビ行列の行列式である. ヤコビ行列 J は, (i, j) 成分が $\partial u_i / \partial v_j$ に等しい部分導関数の正方行列 (u の成分の数によって与えられる次元を有する) である. ここでもまた, f が多対 1 ならば, $p_v(v)$ は項の和または積分である.

1 次元では, 一般に対数を使ってパラメータ空間を $(0, \infty)$ から $(-\infty, \infty)$ に変換します. 開かれた単位区間 $(0, 1)$ で定義されたパラメータを使って作業するとき, 私たちはしばしばロジスティック変換を使います.

$$\text{logit}(u) = \log\left(\frac{u}{1-u}\right) \tag{1.10}$$

そしてこの逆は,

$$\text{logit}^{-1}(v) = \frac{e^v}{1+e^v}$$

である. 別の一般的な選択肢は, $(0, 1)$ から $(-\infty, \infty)$ に変換する標準的な正規累積分布関数 Φ をプロビット変換 $\Phi^{-1}(u)$ である.

1.9 計算とソフトウェア

執筆時点では, グラフや基本シミュレーション, 古典的な単純モデル (回帰, 一般化線形モデル, 局所的重み付き回帰などのノンパラメトリックな方法を含む), 最適化, および単純なプログラミングのためのソフトウェアパッケージ R に主に依存しています. ほとんどのモデルを適合させるためにベイズ推論パッケージ Stan(付録 C を参照) を使用しますが, 本書の教授目的のために, ほとんどの計算を第 1 の原則からどのように実行するかについて説明します. Stan を使用している場合でも, モデルフィッティングの前にデータをプロットして変換し, 後で推論やモデルチェックを表示するために, R 内で作業します.

ベイズデータ分析で生じる特定の計算タスクは次を含む.

・ベクトル操作と行列操作 (表 1.1 を参照)

・確率密度関数の計算 (付録 A を参照)

・確率分布からのシミュレーションの作成 (標準分布については付録 A を, 単純確率過程の例については演習 1.9 を参照)

・構造化プログラミング (ループおよびカスタマイズされた関数を含む)

・線形回帰推定と分散行列を計算する (14 章を参照)

・1 つのページに複数のグラフが表示される, 散布図を含むグラフィックス (例については第 6 章を参照)

一般的な計算方法は, 多くのモデルに適合させ, 徐々に複雑さを増すことです. モデルを作成し, コンピュータを完全に見積もるためにコンピュータを夜間実行させる戦略はお勧めしません. むしろ, モデルを比較的早く適合させることを前提に, より単純なモデルからの推論を開始値として使用し, 推論を表示し, 継続する前にデータと比較することを好む. 本書の 3 部では, ベイズモデリング, 推論, モデル検査の基本的な概念を最初に紹介した後, 計算について詳しく説明します. 付録 C では, R と Stan の計算を 1 つの例でいくつかの異なる方法で実行する方法を示しています.

シミュレーションによる推論の要約

シミュレーションは, 密度関数を明示的に統合できない場合であっても, 確率分布からサンプルを生成することが比較的容易であるため, 多くの適用ベイズ解析の中心的な役割を果たします. シミュレーションを実行する際には, 確率密度関数と分布からのランダム描画の集合のヒストグラムの間の二重性を考慮することが役立ちます. 十分なサンプルがあれば, ヒストグラムは密度に関する実際的な完全な情報を提供することができ, 特に, 様々なサンプルモーメント, パーセンタイル, および他の要約統計量は, 分布のあらゆる側面の推定値を推定できる精度のレベルに提供する. たとえば, θ の分布の 95 次のパーセンタイルを推定するには, $p(\theta)$ からサイズ S のランダムサンプルを引き出し, $0.95S$ 次の統計量を使用します. ほとんどの場合, $S = 1000$ はこのように 95 次パーセンタイルを見積もるのに適しています.

シミュレーションのもう 1 つの利点は, 推定値と確率記述が解析的な形で得られた場合に気付かないかもしれないモデル仕様またはパラメータ化 (例えば図 4.2 を参照) の問題にフラグメンテーションすることがあります. 確率分布から値を生成することは, (擬似) 乱数シーケンスに基づく最新のコンピューティング技術ではしばしば直接的である. うまく設計された擬似乱数発生器は, [0, 1] 上の一様分布からの独立したランダム描画のシーケンスと同じ特性を有すると思われる決定論的シーケンスをもたらす. 付録 A では, 一般的に使用されているいくつかの分布からランダムサンプルを抽出する方法について説明します.

逆累積分布関数を用いるサンプリング

シミュレーションの考え方の紹介として, 逆累積分布関数を使用し離散分布と連続分布からサンプリングする方法について説明する. 1 次元分布 $p(v)$ の累積分布関数 (cumulative distribution function, cdf)F は,

$$F(v_*) = \Pr(v \leq v_*) = \begin{cases} \sum_{v \leq v_*} p(v) & \text{if } p \text{ is discreate} \\ \int_{-\infty}^{v_*} p(v)dv & \text{if } p \text{ is continuous} \end{cases}$$

によって定義される.

逆 cdf を使用して, 以下のように分布 p からランダムサンプルを得ることができる. 最初に, 乱数表, またはおそらくコンピュータ上の乱数関数を使用して $[0, 1]$ の一様分布から乱数値 U を引き出します. さて, $v = F^{-1}(U)$ とする. 関数 F は, 必ずしも 1 対 1 ではなく, 分布が離散的ではないが, $F^{-1}(U)$ は確率 1 で一意である. 値 v は p からの無作為な抽出であり, $F^{-1}(U)$ が単純である限り計算が容易です. 離散分布の場合, F^{-1} は単に表にすることができます.

連続的な例として, v が λ の指数関数分布を持つと仮定する (付録 A を参照). その cdf は $F(v) = 1 - e^{-\lambda v}$ であり, $U = F(v)$ である v の値は $v = -\frac{\log(1-U)}{\lambda}$ である. 次に, $1 - U$ も $[0, 1]$ に一様分布を持つことを認識すると, 私たちは $-\frac{\log U}{\lambda}$ として指数分布からランダム抽出を得ることができることが分かる. 本書の 3 部と付録 A で, 他のシミュレーション方法について説明します.

事前, 事後の予測量のシミュレーション

実際には, 我々は, モデルパラメータ θ の事後分布, おそらくは未知の観測可能な $\tilde{y}$ の事後予測分布から抽出をシミュレートすることに最も関心が高い. S 個のシミュレーション抽出のセットの結果は, 表 1.1 に示すように, 行列内のコンピュータに格納することができます. シミュレーション抽出のインデックス付けには $s = 1, \cdots, S$ という表記法を使用します. $(\theta^s, \tilde{y}^s)$ は, それらの結合事後分布からのパラメータおよび予測量の対応する結合抽出である.

これらのシミュレートされた値から, 既存の S 個の $(\theta, \tilde{y})$ を使用して表 1.1 の新しい列を計算するだけで, θ_1/θ_3 などの関心のある任意の量の事後分布を推定することができます. $\Pr(\tilde{y}_1 + \tilde{y}_2 > e^{\theta_1})$ のような事象の事後確率を, それが真である S 個のシミュレーションの割合で推定できる. 我々はしばしば事後区間に関心がある. 例えば, $\Pr(\theta_j < a) = 0.025$, $\Pr(\theta_j > b) = 0.025$ のパラメータ θ_j の中央の 95%の事後間隔 $[a, b]$ である. これらの値は θ_j の適切なシミュレートされた値, 例えば $S = 1000$ であれば 25 番目と 976 番目の順位統計値によって直接推定することができる. 我々は, 一般に推論を 50%と 95%の間隔で要約する.

いくつかの簡単な例で事後分布のシミュレーションを使っていくつかの経験を積んだ後, 10.5 節でシミュレーションの推論の精度に戻ります.

1.10 応用統計のベイズ推論

ベイズ手法を使用するための実用的な理論的根拠は, 複数のレベルのランダム性の組み込みと結果的に異なる情報源からの情報を結合することによって導入される固有の柔軟性であり, 推論要約にすべての合理的な不確実性の源を組み込む. このような方法は, 複雑なデータ構造において平滑化された推定値に自然につながり, 結果的に現実のより良い解答を得る能力を有する.

ベイズ手法に焦点を当てるもう一つの理由は, より心理的であり, 統計学者と統計学者の仕事の消費者である主題分野のクライアントまたは専門家との関係に関係します. 多くの実用的な場合では, クライアントは統計学者によって提供された区間推定値をベイズ間隔として解釈します. つまり, データのエビデンスを条件とする未知量の取りうる値についての確率表記として解釈します. このような直接の確率表記は, 未知量 (またはより一般的には, 未知量のベクトルの確率モデル) の事前確率モデルを必要とするため, クライアントが想定する回答の種類は統計学者によって提供されます.

最後に, ベイズ推論は, 複雑な現実世界の関係を表現しようと試みて近似値を常に含む確率モデルに条件付きです. ベイズの回答は, データによっては不可能な, 科学的に合理的な仮定の範囲で劇的に変化します. その結果得られる可能性のある結論を正当なものとして振る舞わなければならず, 統計学者はクライアントにこの事実を知らせる責任があると考えます.

本書では, 複雑なデータ構造を科学的な問題に関連させ, そのモデルの適合性をチェックし, 合理的なモデリングの仮定に対する結論の感度を調べるために, モデル (特に 5 章で議論されるような階層的モデル) の構築に焦点を当てる. この観点から, ベイズ手法の強みは, (1) 複数の情報源からの情報を結合する能力 (最終的には実際にはより大きな客観性を可能にする), (2) 統計的問題における未知数に関する不確実性のより包括的な説明.

その他の重要なテーマは, 多くの現代の適用された統計的実践に共通するものであり, ベイズ的であろうとなかろうと, 次のようなものです.

・多くのパラメータを使用する意欲

・モデルの階層的構造化は, 見積もりの部分的なプーリングを達成するための不可欠なツールであり, 代替的な情報源の間で科学的な方法で妥協する

・モデル検査 - 観察された可能性のあるデータと可能性のある将来のデータに対するモデルの内在的適合度を調べることによってだけでなく, 推定知識と予測の推論を実質的な知識

・単純な点推定ではなく, 分布の形での推論または少なくとも区間推定による強調

・計算の主要な方法としてのシミュレーションの使用. 「共同確率分布」に対する現代的な計算の対応は, 無作為に描かれた値の集合であり, 失われたデータを扱うための鍵となるツールは, 複数の代入の方法です（計算と複数の代入については後の章で詳しく説明します）

・ベイズモデルを明示的に呼び出さない可能性のあるデータ分析手法を理解し, 場合によっては改善するためのツールとしての確率モデルの使用

・モデル内のすべての変数に条件付きで, データがランダムサンプルとして表示されるという目標を近似するために, できるだけ多くの背景情報を分析に含めることの重要性

・推定仮説の推論がモデル仮定に強くなるという性質を持つように研究を設計することの重要性