

# つくって始める「ベイズ統計入門」補助テキスト

# 目次

- ケーススタディ
- 確率分布の復習クイズ
- 確率モデルの作り方
- 条件付確率
- 最尤推定
- ベイズ統計基礎理論
- MCMC

## このテキストの位置付け

ケーススタディと確率分布の復習クイズは講義内で扱いますが、それ以外の項については想定した講義のゴールよりも少し難しい内容なので、補足として扱うことはあってもこの資料に基づいて講義を進めることはありません。さらに勉強したい方向けの参考資料として、他の書籍を読みながらめくっていただけると理解が捗ると思います。分からないところがあっても気軽に読み飛ばして大丈夫です。

Introduction

# ケーススタディ

# ケーススタディ: 従来のアプローチの課題

ベイズ統計に対して皆さんはどのような印象をお持ちでしょうか？ 難しそう？ 結局何ができるの？ など様々だと思います。しかしながら、ベイズ統計といえども、従来の機械学習や統計的な方法論同様にデータから推定される背後の構造を仮定し、検証するアプローチであることは変わりません。では一体何ができるのでしょうか？

このチャプターではまず、従来の方法が抱えていた課題について以下3つのケースを用いて説明を試みます。

## ケース①: 5段階のアンケート

## ケース②: Afterコロナはいつから？

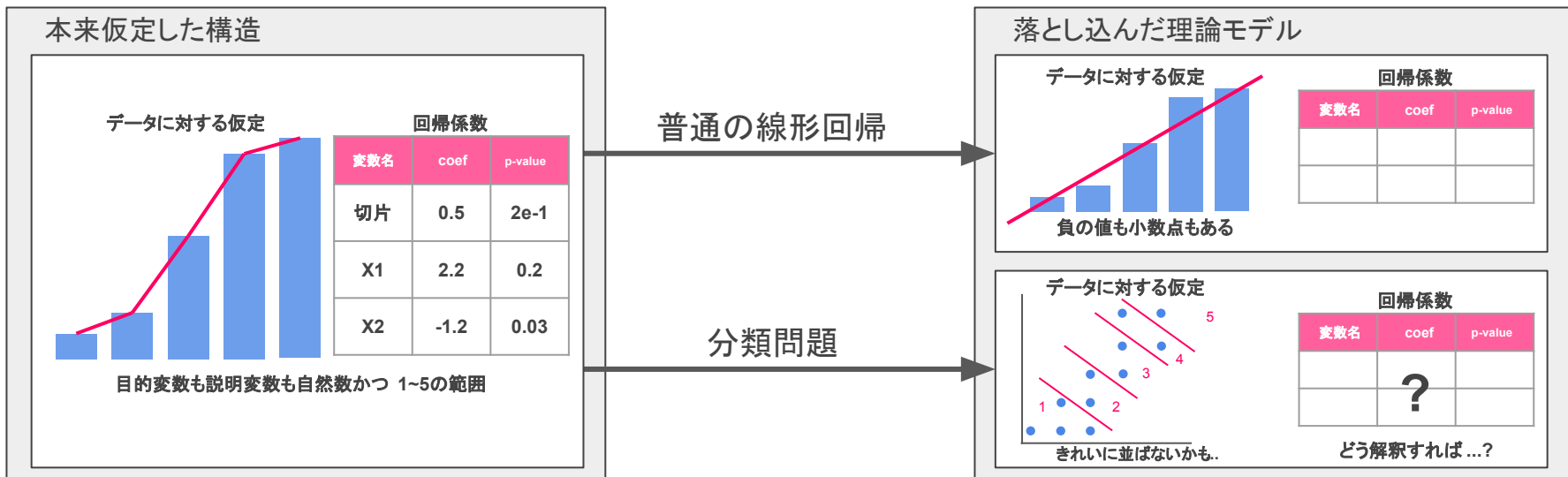
## ケース③: 共通因子と固有因子

勿論それぞれの課題はベイズ統計を用いると対処できるテーマです。似たような課題が身近にないか、ぜひ考えながら読み進めてみてください。

# ケーススタディ: ①5段階のアンケート

まず最初のケースはアンケートの分析です。例えば、社員や顧客への満足度調査などです。

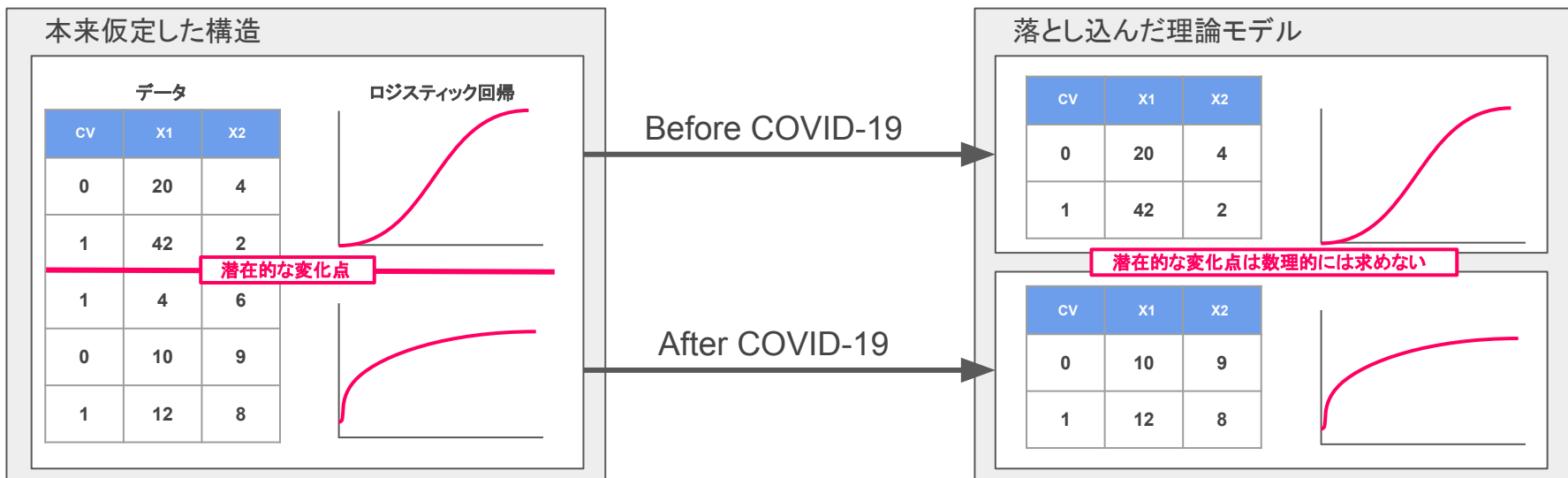
特定の項目を目的変数とした線形回帰分析を行い、説明変数とした項目の回帰係数を確認して施策に結び付けたいとします。では目的変数としたい項目が1~5の5段階評価である場合、回帰問題と捉えるべきでしょうか？それとも分類問題と捉えるべきでしょうか？普通の回帰問題として捉えてしまうと、モデルの仮定としては1~5の範囲にはおさまっていませんし、小数点の値も取り得ます。かと言って分類問題にしたら1~5の大小関係を考慮することができません。



# ケーススタディ: ②Afterコロナはいつからか

次にWebサイトのコンバージョン分析を考えてみます。マーケチームの課題はCOVID-19の前後でコンバージョンに寄与する要素が変化しているため、施策が打ちにくくなっていることです。

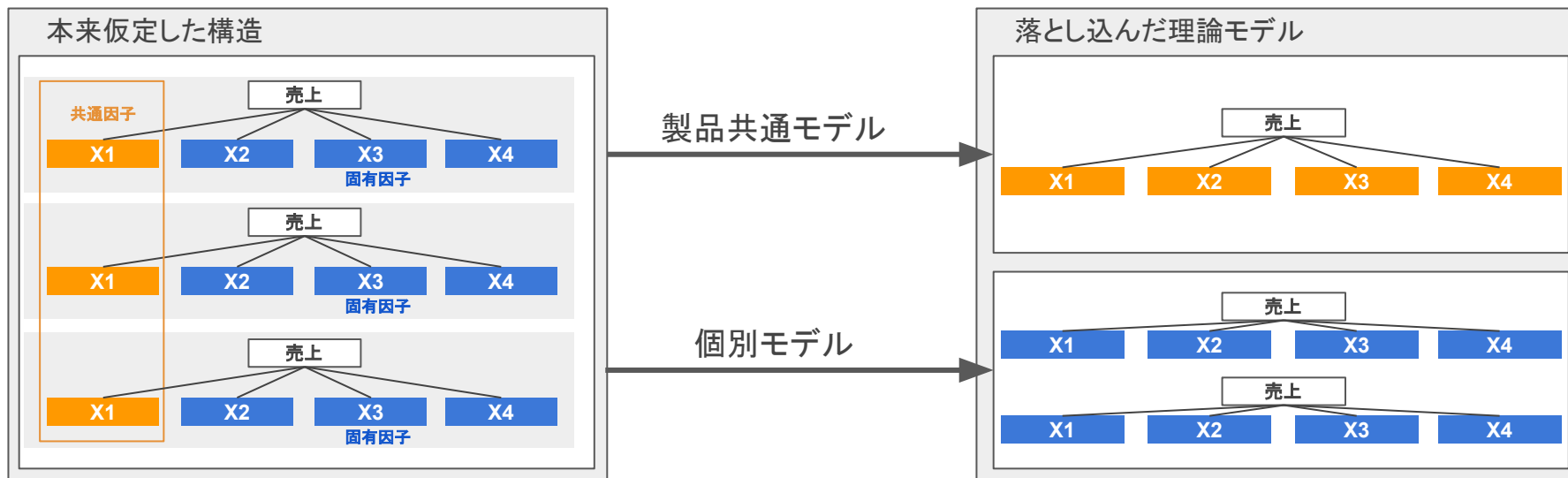
COVID-19以前はコンバージョンを目的変数としたロジスティック回帰を用いて回帰係数(オッズ比)を出力し、施策に活かしていたのですがやり直す必要がありそうです。では、いつからのデータをCOVID-19後のデータとみなせば良いのでしょうか。潜在的には回帰係数の「変化点」があるはずですが、従来の方法では数理的に求めることはできません。



# ケーススタディ:③共通因子と個別因子

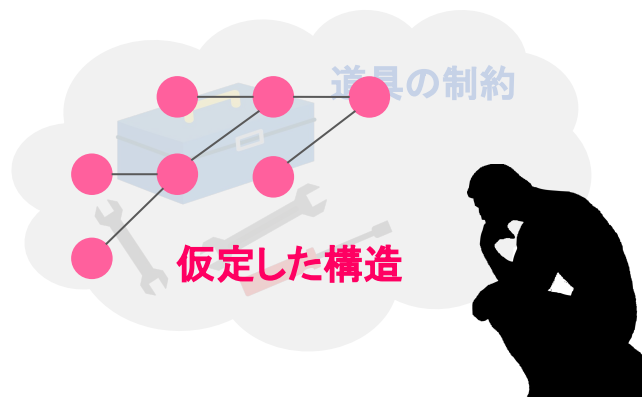
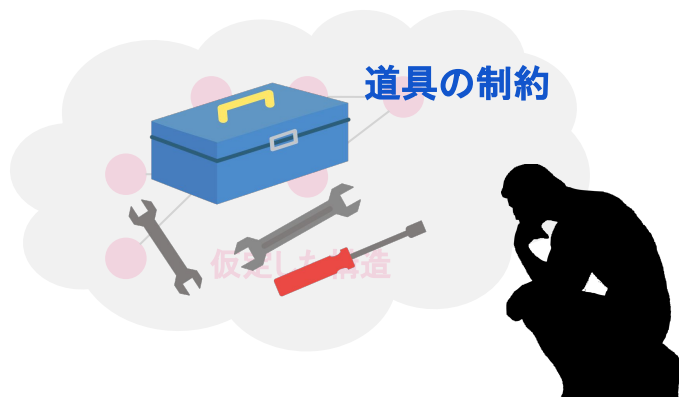
在庫調整のために需要予測を行い調達計画を立てたいとしましょう。各製品ごとに売り上げの傾向が異なることがわかっており、どうやら「固有因子」があること考えた方が良さそうです。その一方で販促施策は商品カテゴリ別に行われているため、各製品間での「共通因子」も考慮する必要があるようです。

しかしながら従来の方法を用いるなら、全ての製品を一緒にモデリングするか、別々に分けるかの2択しかありません。



# ケーススタディ: 様々な応用

ここで紹介したケースは全て「本来仮定した構造」をベイズ統計を用いて検証できるテーマになっています。ベイズ統計を用いることで「道具」を主体としたモデリングから「仮定した構造」を主体としたモデリングがやりやすくなったわけです。何やら回帰モデルっぽいものばかりだな。と感じたかもしれませんが、他にも主成分分析、行列分解やクラスタリングなどの教師なし学習、さらに深層学習もベイズ的にモデリングが可能です。



これだけ柔軟なモデリングが可能であることから、構造化されたデータだけでなく、自然言語、画像処理、信号データ処理など幅広い領域で活用されており、さらに既存の方法論についても基礎理論を与えています。

難しそうだと感じたでしょうか？もちろん深く掘り下げれば難易度は上がりますが、最初の導入はそれほど難しくありません。これから一緒に勉強していきましょう。



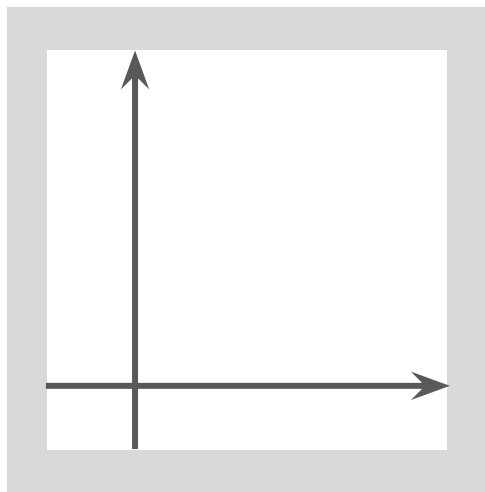
Warming UP!!

# 確率分布の復習クイズ

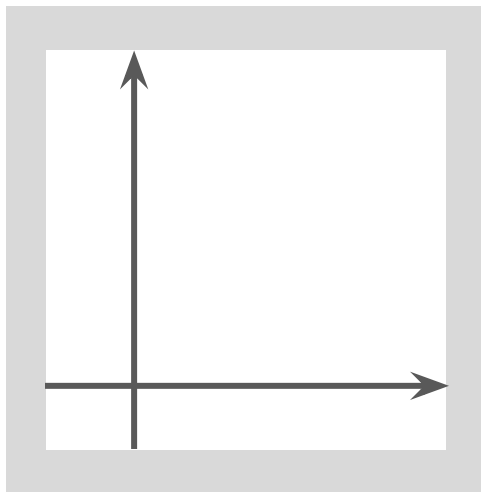
# 確率分布 クイズ ①

それぞれの確率分布の「確率密度(質量)関数」のグラフを書いてみよう

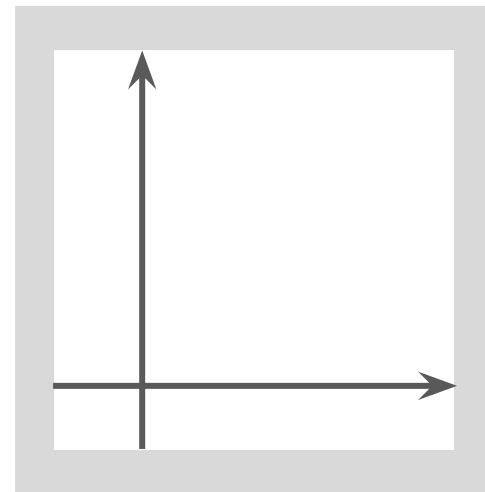
正規分布



二項分布



ポアソン分布



## 確率分布 クイズ ②

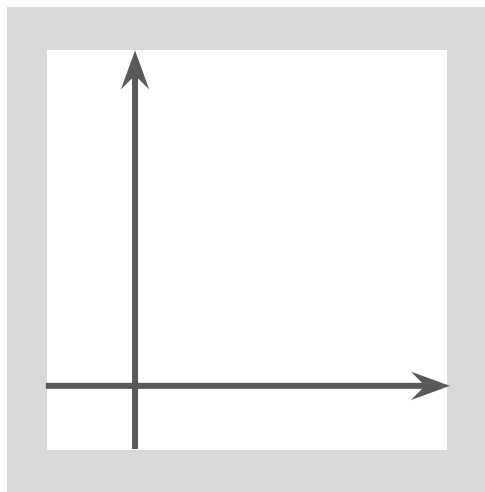
それぞれの確率分布の性質について「正規分布」の例を参考に空欄を埋めてみよう！

| 名称     | 確率変数の種類<br>(離散値 or 連続値) | 確率変数を取りうる値の範囲 |       | パラメーター   | パラメーターを取りうる値の範囲     |                        |
|--------|-------------------------|---------------|-------|----------|---------------------|------------------------|
|        |                         | 最小値           | 最大値   |          | 最小値                 | 最大値                    |
| 正規分布   | 連続値                     | マイナス無限        | プラス無限 | 平均<br>分散 | 平均: マイナス無限<br>分散: 0 | 平均: プラス無限<br>分散: プラス無限 |
| 二項分布   |                         |               |       |          |                     |                        |
| ポアソン分布 |                         |               |       |          |                     |                        |

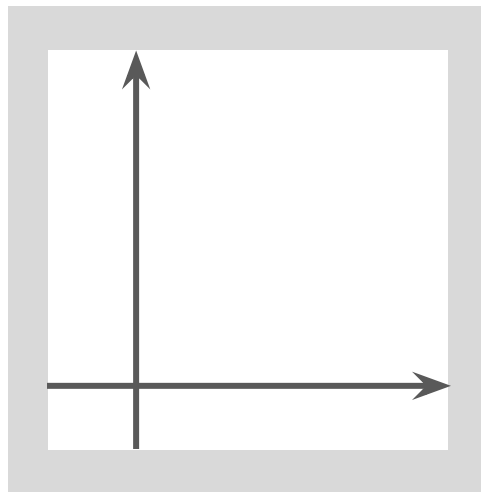
# 確率分布 発展クイズ ① 調べてみよう

それぞれの確率分布の「確率密度(質量)関数」のグラフを書いてみよう

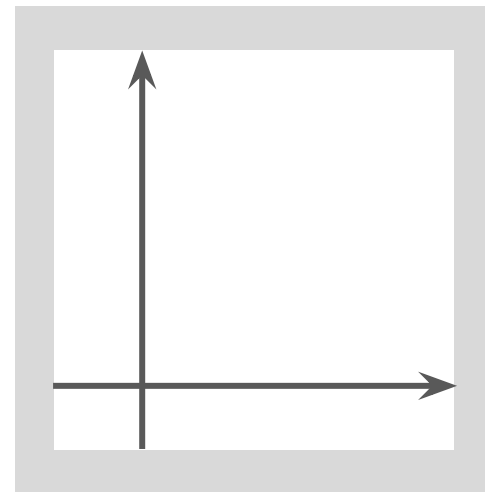
連続一様分布



ベータ分布



ガンマ分布



# 確率分布 発展クイズ ② 調べてみよう

それぞれの確率分布の性質について調べてみよう！

| 名称     | 確率変数の種類<br>(離散値 or 連続値) | 確率変数を取りうる値の範囲 |     | パラメーター | パラメーターを取りうる値の範囲 |     |
|--------|-------------------------|---------------|-----|--------|-----------------|-----|
|        |                         | 最小値           | 最大値 |        | 最小値             | 最大値 |
| 連続一様分布 |                         |               |     |        |                 |     |
| ベータ分布  |                         |               |     |        |                 |     |
| ガンマ分布  |                         |               |     |        |                 |     |

Appendix

# 確率モデルの作り方

# 確率モデルの作り方: 線形回帰モデル

まずは線形回帰モデルから考えてみましょう。線形回帰モデルはデータ間の関係性を以下の式で仮定するものでした。

$$y = \alpha + \beta_1 \times X_1 + \beta_2 \times X_2 + \dots \beta_n \times X_n$$

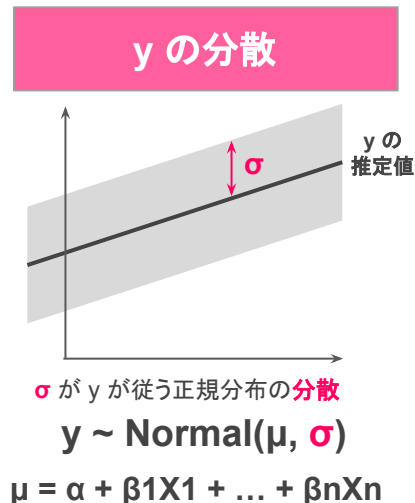
手元のデータに当てはめることで、回帰係数と  $y$  の推定値に対する区間=分散を推定できます。

使用するデータのイメージ

| y  | X1 | X2 | ... | Xn |
|----|----|----|-----|----|
| 10 | 1  | 1  | ... | 9  |
| 7  | 2  | 5  | ... | 2  |
| .  | .  | .  | .   | .  |
| .  | .  | .  | .   | .  |
| .  | .  | .  | .   | .  |
| 5  | 7  | 5  | ... | 0  |

回帰モデルを使ってわかること

| 変数名           | coef     | p-value  |
|---------------|----------|----------|
| $\alpha$ (切片) | -4.0     | 0.201    |
| $\beta_1$     | 3.2      | 2e-10    |
| $\beta_2$     | -5.5     | 0.014    |
| $\vdots$      | $\vdots$ | $\vdots$ |
| $\beta_n$     | 10.2     | 0.023    |



# 確率モデルの作り方: 各変数が従う分布

ではこのシンプルな線形回帰モデルをベイズ的にデザインしてみましょう。

ベイズ統計では推定したい変数を全て「分布」で推定します。「分布」で推定するためには仮定が必要ですが、普通の線形回帰モデルでは、切片と回帰係数に対して  $t$  分布を仮定していました。

ただし、 $\sigma$  ( $y$  が従う分布の分散) の分布はありませんでしたね。

ではベイズ的に推定すると、切片や回帰係数、 $\sigma$  はどのような分布になるのでしょうか？

| 推定したい変数   | 変数の意味    |
|-----------|----------|
| $\alpha$  | 切片       |
| $\beta_1$ | 回帰係数     |
| $\beta_2$ | 回帰係数     |
| $\vdots$  | $\vdots$ |
| $\beta_n$ | 回帰係数     |
| $\sigma$  | $y$ の分散  |

| 普通の線形回帰モデル          |          |
|---------------------|----------|
| $\alpha \sim t(v)$  | $t$ 分布   |
| $\beta_1 \sim t(v)$ | $t$ 分布   |
| $\beta_2 \sim t(v)$ | $t$ 分布   |
| $\vdots$            | $\vdots$ |
| $\beta_n \sim t(v)$ | $t$ 分布   |
| $\sigma \sim ?$     | $?$      |

| ベイズ線形回帰モデル |
|------------|
| ?          |



# 確率モデルの作り方: 事前準備

確率モデルをデザインするために準備しましょう。手順はとても簡単で、①推定したい変数を取りうる値と

②その値を確率変数とする確率分布を考えるだけです。

なお、今後簡単のために  $X_1 \sim X_n$  は全て  $X$ 、 $\beta_1 \sim \beta_n$  は全て  $\beta$  と表記します。

## 手順

①推定したい変数を取りうる値を考える

②その値を確率変数とする確率分布を考える

### ① 推定したい変数を取りうる値

| 未観測変数    | 値の種類 | 最小値     | 最大値 |
|----------|------|---------|-----|
| $\alpha$ | 連続値  | -無限     | +無限 |
| $\beta$  | 連続値  | -無限     | +無限 |
| $\sigma$ | 連続値  | 0 より大きい | +無限 |

### ② 対応する確率分布 の例

| 確率分布の名称                              |
|--------------------------------------|
| Normal, T分布 など                       |
| Normal, T分布 など                       |
| Half-Normal, Half-Cauchy, InvGammaなど |

対応する確率分布はどれを選んで大丈夫なのですが、ここではそれぞれ以下の分布を選択したいと思います。

$\alpha \sim \text{Normal}$ 、 $\beta \sim \text{Normal}$ 、 $\sigma \sim \text{InvGamma}$  (逆ガンマ分布)

# 確率モデルの作り方: 定式化

準備が整いましたので、一旦定式化しておきましょう。

なお、赤い文字は推定したい変数、青い文字は観測されているデータ、緑の文字は事前に準備した確率分布です。

今後同じ意味のものを同じ色で表記するので注意してみてください。

$\alpha \sim \text{Normal}$

$\beta \sim \text{Normal}$

$\sigma \sim \text{InvGamma}$

$\mu = \alpha + \beta \times X$

$y \sim \text{Normal}(\mu, \sigma)$

これで準備完了...と言いたところですが、事前に準備した分布にもパラメーターがあるはずですね？

どうすれば良いかは後述しますので、一旦以下のように表記することになります。

$\alpha \sim \text{Normal}(\alpha\mu, \alpha\sigma)$

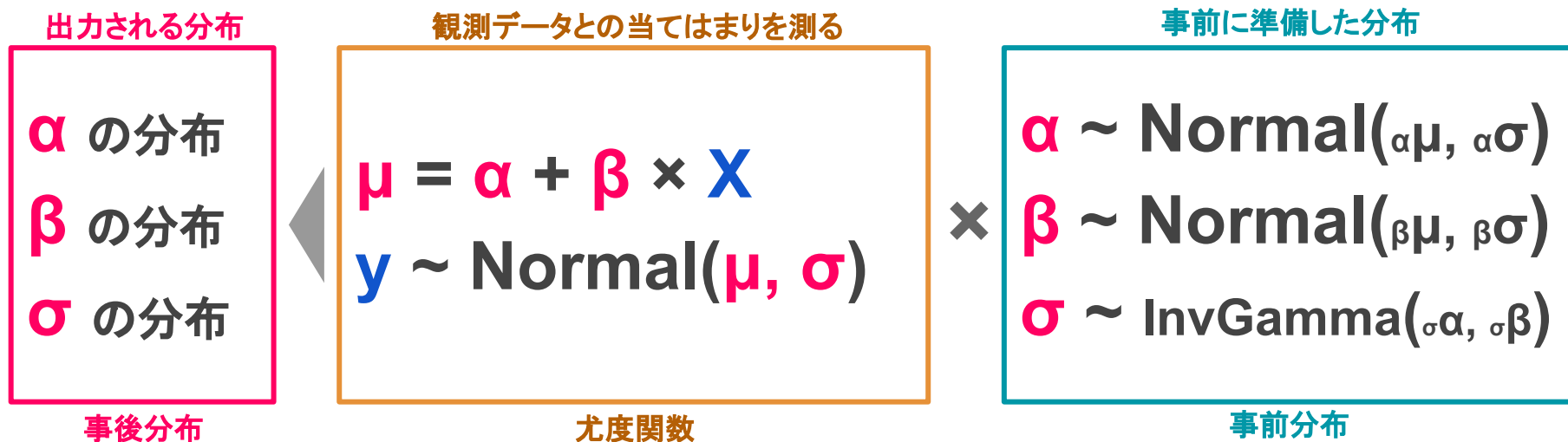
$\beta \sim \text{Normal}(\beta\mu, \beta\sigma)$

$\sigma \sim \text{InvGamma}(\sigma\alpha, \sigma\beta)$

# 確率モデルの作り方: 推定する

では推定してみます。実は  $y$  が従う分布だけは**観測データ**との**当てはまり=尤度**を測る役割を持っています。

この**尤度関数**と**事前に準備した分布**を掛け算すると**推定したい変数の分布**が出力されるようになっています。ええ！なんで！？と思った方は、このテキストの「条件付確率」と「ベイズ統計基礎理論」をぜひ読んでみてください。また「尤度ってなんだっけ...」という方向けには「最尤推定」の章も用意しているので、忘れてしまった方はぜひ読んでみてください。



また各「役割」ごとに名前がついているので覚えておきましょう。**事前に準備した分布**を**事前分布**、**観測変数**との**当てはまりを測る尤度関数**、**出力される分布**を**事後分布**といいます。

# 確率モデルの作り方:出力される分布

事後分布がどのように出力されるかみていきましょう。「解析的」に推定する場合と「サンプリングアルゴリズム」を用いて推定する場合との違いを見てみます。

## 解析的に推定する場合

### 事後分布

確率密度関数

$$\alpha \sim \alpha f(X)$$

$$\beta \sim \beta f(X)$$

$$\sigma \sim \sigma f(X)$$

※密度関数は正規分布、  
t分布など既に名前がついて  
いるものになることが多い

### 解釈の方法

密度関数をプロット



要約統計量や区間を確認

|          | 平均 | 分散 | 区間    |
|----------|----|----|-------|
| $\alpha$ | xx | xx | xx~xx |
| $\beta$  | xx | xx | xx~xx |
| $\sigma$ | xx | xx | xx~xx |

## サンプリングする場合

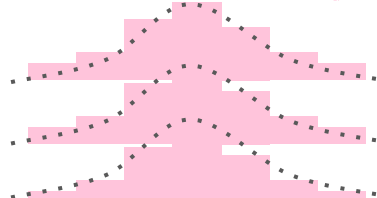
### 事後分布

サンプリングされたデータ

| $\alpha$ | $\beta$ | $\sigma$ |
|----------|---------|----------|
| 2.4      | 3.3     | 1.2      |
| .        | .       | .        |
| .        | .       | .        |
| .        | .       | .        |
| 0.9      | 4.4     | 3.3      |

### 解釈の方法

データをヒストグラムで可視化



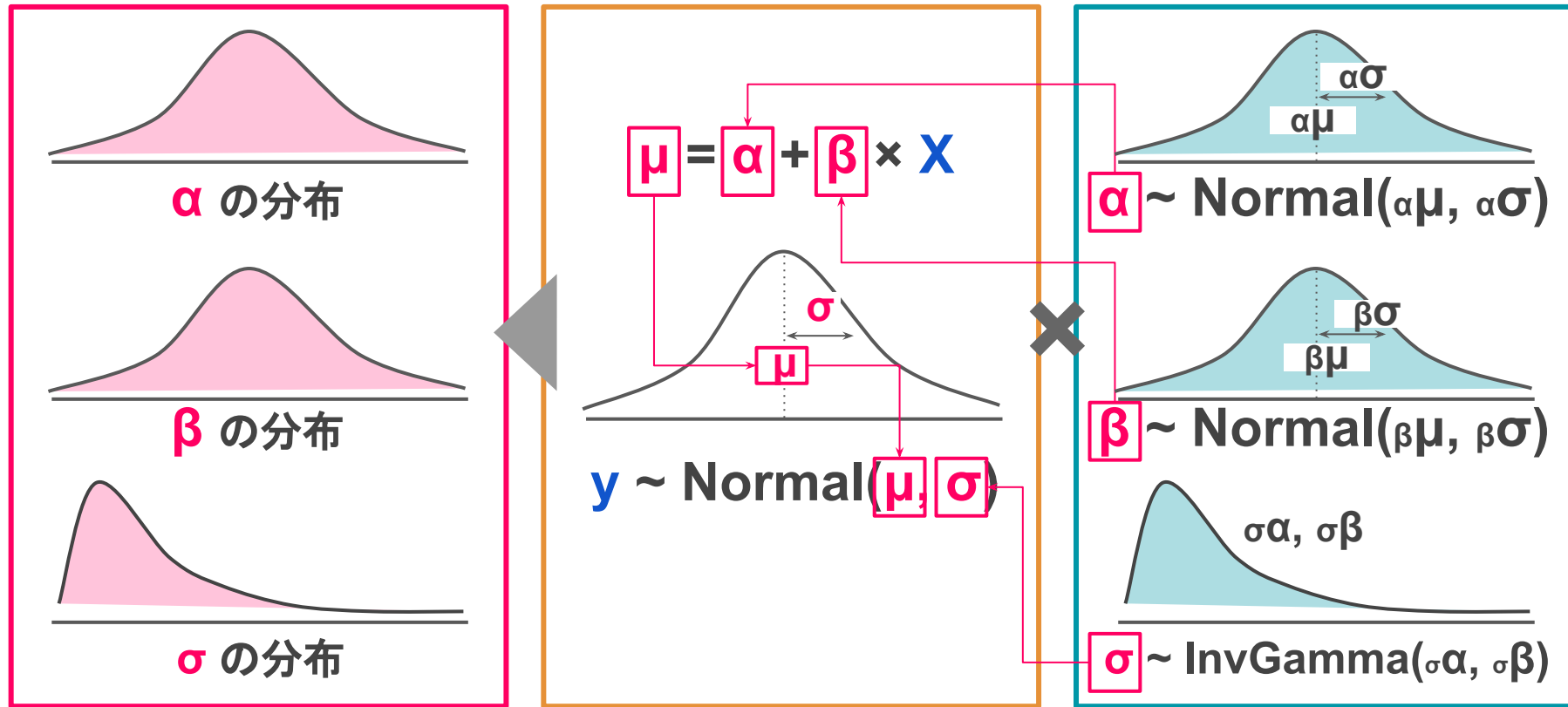
要約統計量や区間を確認

|          | 平均 | 分散 | 区間    |
|----------|----|----|-------|
| $\alpha$ | xx | xx | xx~xx |
| $\beta$  | xx | xx | xx~xx |
| $\sigma$ | xx | xx | xx~xx |

解析的に推定する場合は関数そのものが出力されるのに対して、サンプリングアルゴリズムを用いる場合は、1件1件のデータとして出力される点異なりますが、解釈の方法は概ね同じです。

# 確率モデルの作り方: モデルのデザイン図解①

改めてモデルのデザインと出力結果を図解しながら詳しく解説します。



# 確率モデルの作り方:モデルのデザイン図解②

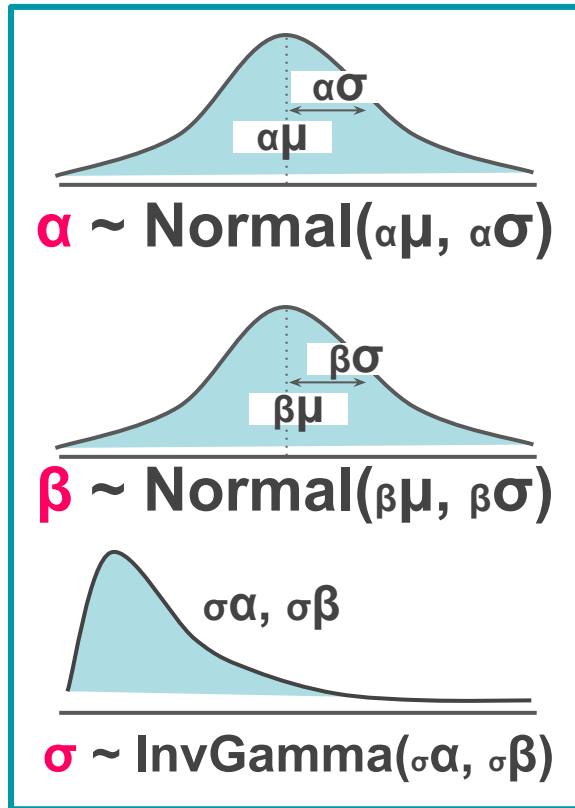
では一つ一つ見てみましょう。まずは**事前分布**です。

今回の値の定義を振り返ると以下の通りでした。

| 推定したい変数  | 意味    | 値の種類 | 最小値    | 最大値 |
|----------|-------|------|--------|-----|
| $\alpha$ | 切片    | 連続値  | -無限    | +無限 |
| $\beta$  | 回帰係数  | 連続値  | -無限    | +無限 |
| $\sigma$ | y の分散 | 連続値  | 0より大きい | +無限 |

推定したい変数が取りうる値の種類 最小, 最大値を考え、該当する確率分布を選べば良いのだね。今回は正規分布と逆ガンマ分布を用いています。

また、ここで考えた確率分布にもパラメーターがありますが、この変数は人間が与えるものなので、**ハイパーパラメーター**と呼びます。実はハイパーパラメーターを推論するための事前分布を与える事もできます。例えば、右の図の $\alpha\mu$ を推定するための事前分布を作ることができるわけです。つまり、事前分布のパラメーターを探すための事前分布を作って、さらにそのパラメーターを探すための事前分布を作って..と言うことができちゃうわけです。

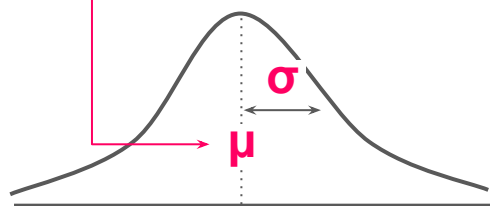


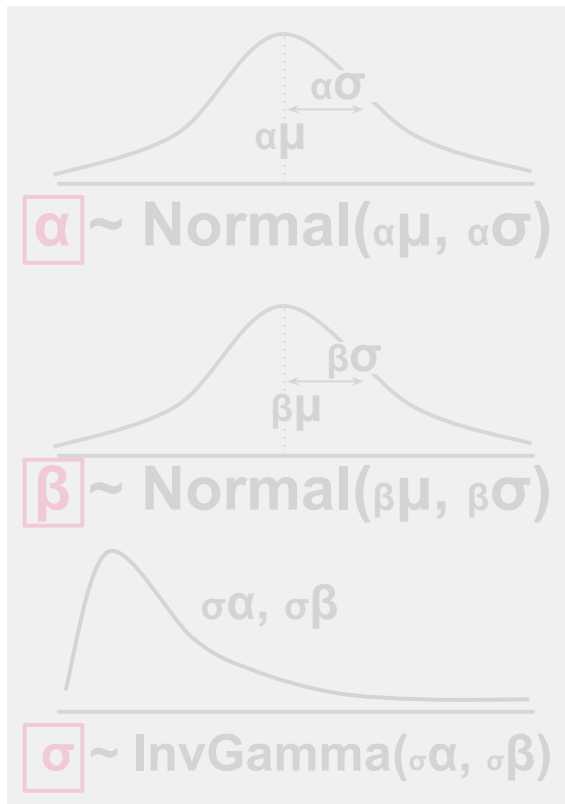
# 確率モデルの作り方:モデルのデザイン図解③

次に**尤度関数**です。

事前分布から生成された各変数が  
 $y$  が従う分布に組み込まれています。  
事前分布と同様にここでも確率分布を仮定して  
います。

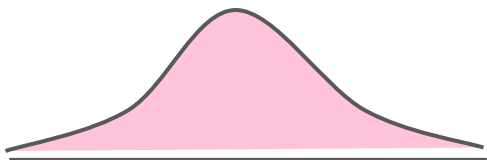
ただし、  
事前分布は**未観測変数が従う分布**でしたが、こ  
こでは**観測変数が従う分布**である点が異なります。  
この分布は**尤度**を測る役割を持っています。

$$\mu = \alpha + \beta \times X$$

$$y \sim \text{Normal}(\mu, \sigma)$$

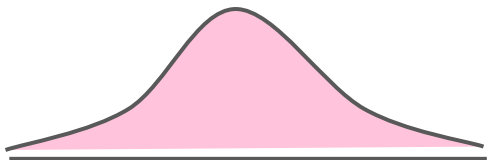


# 確率モデルの作り方:モデルのデザイン図解④

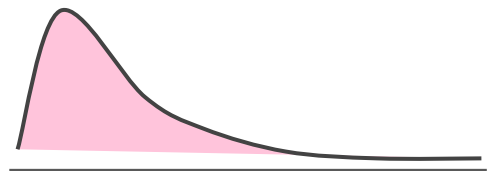
最後に**事後分布**です。



$\alpha \sim ?$



$\beta \sim ?$



$\sigma \sim ?$

**事前分布**と**尤度関数**のデザインが終わったら**未観測変数**を**事後分布**として推論できます。分布で出力できるので、例えば以下のような使い方ができるでしょう。

## 回帰係数と切片

- ・区間推定を用いて、大きくてもこれくらい、小さくてもこれくらい、と言及する。
- ・ある一定の値(例えば0)よりも大きい(もしくは小さい)確率は〇〇%と言及する。

## 分散

- ・y の値は大きくてもこれくらい、小さくてもこれくらい、と言及する。

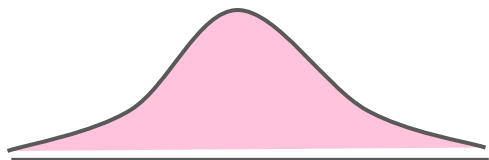
## $\mu(\alpha \text{の平均} + \beta \text{の平均} \times X)$

- ・X だけ明らかなでy が不明のデータを入手した際のy の期待値を求める。つまり予測。

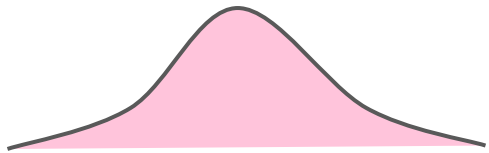


# 確率モデルの作り方:モデルのデザイン図解⑤

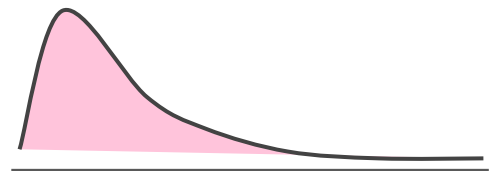
線形回帰モデルにおける**事後分布**は解析的に求めることができます。



$$\alpha \sim t(v)$$



$$\beta \sim t(v)$$



$$\sigma \sim \text{InvGamma}(\sigma\alpha, \sigma\beta)$$

実は、**事前分布**をそれぞれ、 $\alpha, \beta$ を正規分布、 $\sigma$ を逆ガンマ分布とすると解析的に**事後分布**を求めることができ、**事後分布**は $\alpha, \beta$ がt分布、 $\sigma$ がまた逆ガンマ分布に従います。

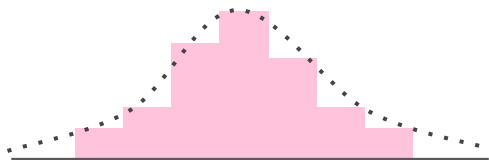
回帰係数のt検定では特段ベイズ統計を意識することはないかもしれませんが、実はこの検定**事後分布**を用いて行っています。

しかしながら、解析的に求める場合は数理的に厳密にデザインする必要があります。また、どのような確率モデルをデザインしても、いつも解析的に求めることができるわけではありません。

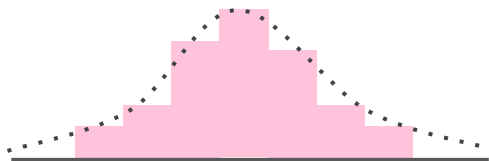
解析的に求められない場合は、事後分布を近似的に求めます。方法はいくつかありますが、サンプリングアルゴリズムを使用した場合の事後分布の出力結果についてみてみましょう

# 確率モデルの作り方:モデルのデザイン図解⑥

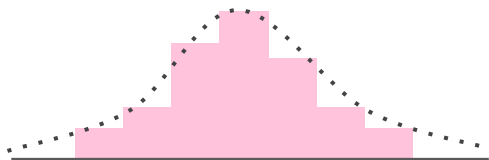
サンプリング アルゴリズムを使用した場合の**事後分布**



$\alpha$



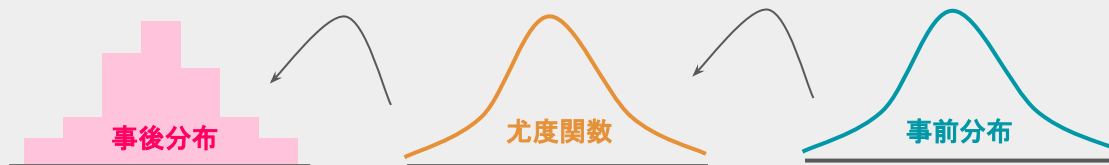
$\beta$



$\sigma$

サンプリングアルゴリズムは、その名の通り**データ**をサンプリングします。  
得られた**データ集合**が実は**推定した事後分布**であり、ヒストグラムにしたり要約統計量を確認することで解釈します。

理屈としては**事前分布**から確率的に生成された**パラメーター**を**尤度関数**に入れてあげると、**事後分布**が出てくるようになっています。詳細な動作原理については「MCMC」の章で解説しています。

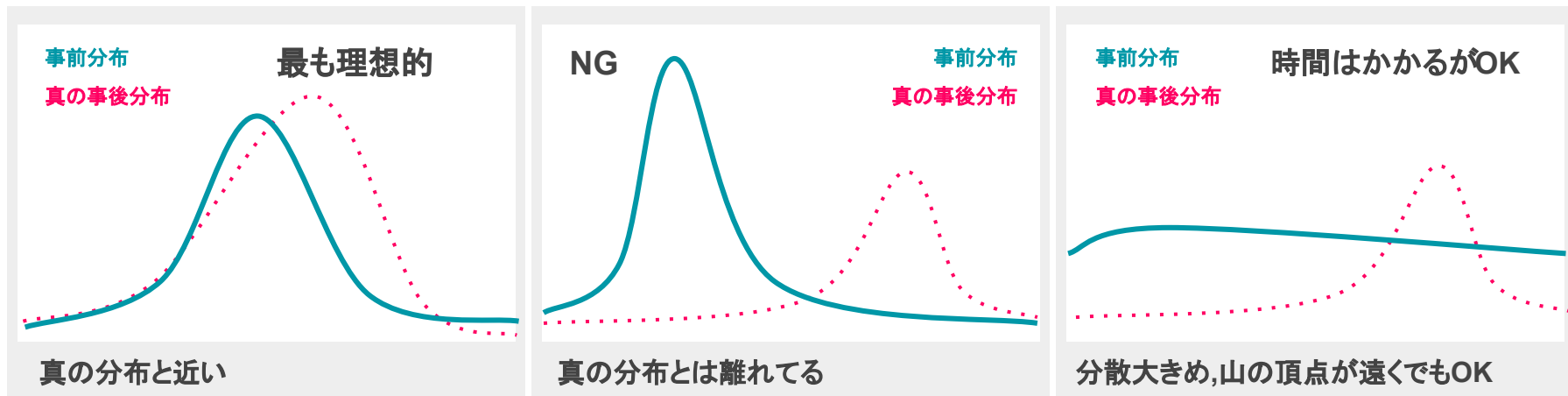


このアルゴリズムはデータを生成しているだけなので**事後分布**を解析的に求められるかどうかを考慮して、**事前分布**を設計する必要がありません。  
また、より当てはまりの良い**事後分布**となるよう生成しています。

ただし、この**事後分布**はアルゴリズムを動かすまでどのような分布になるかはわからない点には注意しましょう。わからないものを推定するので当たり前かもしれませんが。

# 確率モデルの作り方:事前分布のパラメーター

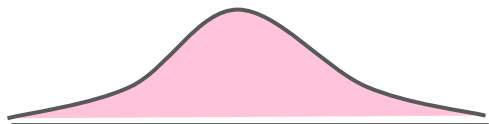
サンプリングアルゴリズムを用いる場合は、**事前分布**のパラメーターは**事後分布**に似るように設計すると、より早く計算が終わり、かつより正確な**事後分布**を推定できますが「そんなの事前にわからない」場合は、分散を大きめ、つまり裾が大きくなだらかなようになるよう設計するのがコツです。**推定したい変数**の範囲が 0 ~ 1 とわかっている場合など、予め区間が明らかな場合は一様分布を用いると良いでしょう。



裾を大きめになるように**事前分布**を設計すると、あとは「いい感じ」に**事後分布**を出力してくれるようアルゴリズムが働いてくれます。

# 確率モデルの作り方: 基底関数モデル①

ここからは「レシピ」をいくつか紹介していきます。ケース⑮段階のアンケートです。  
ここでの仮定は「目的変数が1 ~ 5の自然数」とした線形回帰モデルです。

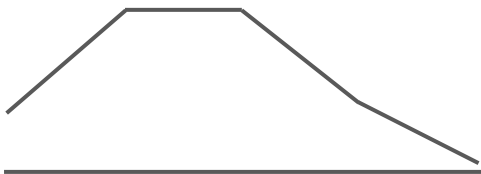


$\alpha$  の分布

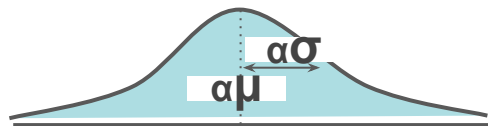


$\beta$  の分布

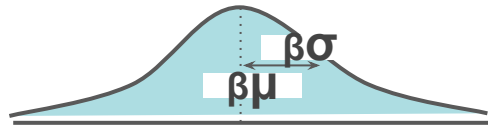
$$\mu = \alpha + \beta \times X$$
$$p = \text{sigmoid}(\mu)$$



$$(y-1) \sim \text{Binom}(p, n=4)$$



$$\alpha \sim \text{Normal}(\alpha\mu, \alpha\sigma)$$



$$\beta \sim \text{Normal}(\beta\mu, \beta\sigma)$$

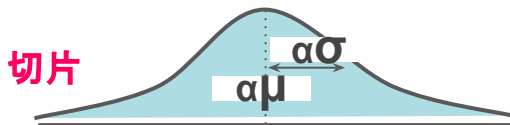
それぞれが何を意味しているのか詳しくみていきましょう。

## 確率モデルの作り方:基底関数モデル②

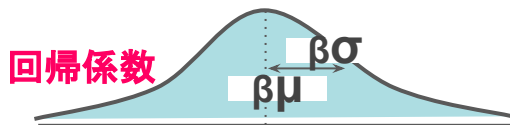
**事前分布**を見ていきましょう。切片の  $\alpha$  回帰係数の  $\beta$  は線形回帰モデル同様ですが、今回は**尤度関数**に二項分布を用いているので、分散の  $\sigma$  はありません。

まずは線形回帰モデルと同じ考え方をします。

そのために、事前分布は切片の  $\alpha$ 、回帰係数の  $\beta$  をそれぞれ正規分布とします。



$$\alpha \sim \text{Normal}(\alpha\mu, \alpha\sigma)$$



$$\beta \sim \text{Normal}(\beta\mu, \beta\sigma)$$

# 確率モデルの作り方:基底関数モデル③

続いて**尤度関数**です。観測データ  $y$  との当てはまりを確認できる分布を考えましょう。ここでは**二項分布**を選択します。 $\alpha + \beta \times X$  を  $\mu$  とするのは線形回帰と同じですが、 $\mu$  をシグモイド関数で変換している点が異なります。

二項分布は  $0 \sim n$  までの値を確率変数とする確率分布でした。

$y$  は  $1 \sim 5$  なので  $0 \sim n$  に合うように  $y - 1$  とし  $n$  を  $4$  とした二項分布とします。

あとは  $p$  の値を求めることができれば良さそうです。

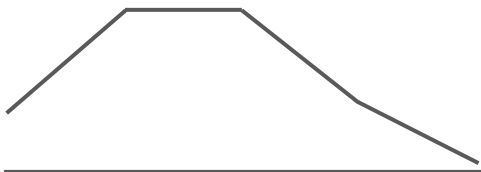
$p$  は確率なので  $0 \sim 1$  の範囲でなければなりません。が、 $\mu = \alpha + \beta \times X$  では  $0 \sim 1$  になりませんので、工夫する必要があります。

ここではsigmoid関数を用いてみました。

ロジスティック回帰に登場した関数ですね。

このような、通常の回帰モデルに、関数を適用したモデルを「基底関数モデル」といいます。ここの基底関数はシグモイド関数です。

$$\mu = \alpha + \beta \times X$$
$$p = \text{sigmoid}(\mu)$$

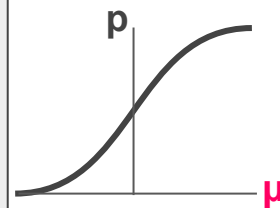


$$(y-1) \sim \text{Binom}(p, n=4)$$

二項分布は  $0 \sim n$  までの値に対して  
 $y$  は  $1 \sim 5$  なので  $-1$  で補正

sigmoid 関数

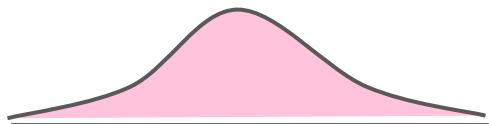
$$p = \text{sigmoid}(\mu)$$
$$p = \frac{1}{1 + \exp(-\mu)}$$



横軸が  $\mu$   
縦軸が  $p$   
S字カーブが  
sigmoid関数

# 確率モデルの作り方:基底関数モデル④

最後の**事後分布**です。 $\alpha$  や  $\beta$  の解釈はロジスティック回帰と同じです。



$\alpha$  の分布



$\beta$  の分布

$\alpha$ と $\beta$ の解釈にヒストグラムや要約統計量を用いる点は線形回帰モデルと同様です。値の解釈はロジスティック回帰と全く同じです。すなわち $\alpha$ ,  $\beta$  の値が大きければ大きいほど、 $p$  の値も1に近づきます。 $p$  の値が1に近づくほど二項分布の確率変数も値が大きくなるという関係です。

$$p = \frac{1}{1 + \exp(-\mu)}$$

$$\mu = \alpha + \beta \times X$$

以下  $y$  を社員満足度としたアンケートを回帰モデルとした事後分布の解釈例です。  
ここではオッズ比は用いていませんが、必要に応じて用いてもいいでしょう。

| 変数            | 意味(例) | 事後分布の平均 | 解釈               |
|---------------|-------|---------|------------------|
| $\alpha$ (切片) | 切片    | -1.6    | -                |
| $\beta_1$     | 年収    | 0.54    | 年収は高い方が満足度が高い？   |
| $\beta_2$     | 残業時間  | -2.2    | 残業時間が多いと満足度が下がる？ |

# 確率モデルの作り方: 潜在変数モデル

ケース② After コロナはいつから？です。目的変数が CV or Not の2値のロジスティック回帰を用いて、コロナ前後での回帰係数の変化と変化点(日付)を推定します。

$\alpha_1$  の分布

$\beta_1$  の分布

$\alpha_2$  の分布

$\beta_2$  の分布

$\tau$  の分布

$$\mu_1 = \alpha_1 + \beta_1 \times X[\tau:]$$

$$\mu_2 = \alpha_2 + \beta_2 \times X[:\tau]$$

$$\mu = [\mu_1, \mu_2]$$

$$p = \text{sigmoid}(\mu)$$

$$y \sim \text{Bernoulli}(p)$$

$$\alpha_1 \sim \text{Normal}(\alpha_1\mu, \alpha_1\sigma)$$

$$\beta_1 \sim \text{Normal}(\beta_1\mu, \beta_1\sigma)$$

$$\alpha_2 \sim \text{Normal}(\alpha_2\mu, \alpha_2\sigma)$$

$$\beta_2 \sim \text{Normal}(\beta_2\mu, \beta_2\sigma)$$

$$\tau \sim \text{Uniform}(a, b)$$



# 確率モデルの作り方: 潜在変数モデル

**事前分布**は before 用の  $\alpha_1$ ,  $\beta_1$ 、after 用の  $\alpha_2$ ,  $\beta_2$ 、変化点検出用の  $\tau$  (タウ)とします。

before と after にそれぞれ独立した切片と回帰係数を推定するのが今回の目的です。

また同時に 変化点も「分布」で推定します。

変化点は「日付」とします。事前分布は  $X$  の日付の最大値から最小値を区間とする一様分布とします。

なお、今回のような「観測できていない未知のデータ」を用いるモデルを「潜在変数モデル」といいます。

before 用

$$\alpha_1 \sim \text{Normal}(\alpha_1\mu, \alpha_1\sigma)$$

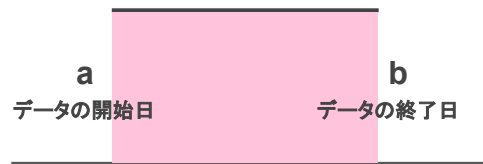
$$\beta_1 \sim \text{Normal}(\beta_1\mu, \beta_1\sigma)$$

after 用

$$\alpha_2 \sim \text{Normal}(\alpha_2\mu, \alpha_2\sigma)$$

$$\beta_2 \sim \text{Normal}(\beta_2\mu, \beta_2\sigma)$$

変化点検出用



$$\tau \sim \text{Uniform}(a, b)$$

# 確率モデルの作り方: 潜在変数モデル

$X$  は日付順に並んでいるとして、前後の  $\mu$  を結合しロジスティック回帰を行います。

$\mu_1$  と  $\mu_2$  を別々に求めるために、一旦  $X$  を  $\tau$  の前後で分割しているイメージです。

一旦分割したものを  $\mu$  として結合しています。その  $\mu$  をシグモイド関数を用いて、0 ~ 1 の範囲に変換します。

$y$  は 0, 1 どちらかの値をとりますので、尤度関数はベルヌーイ分布を用いています。

実はロジスティック回帰と呼んでいるものは尤度関数をベルヌーイ分布とした回帰モデルと同じものです。

$$\mu_1 = \alpha_1 + \beta_1 \times X[\tau:]$$

$$\mu_2 = \alpha_2 + \beta_2 \times X[:\tau]$$

$$\mu = [\mu_1, \mu_2]$$

$$p = \text{sigmoid}(\mu)$$

$$y \sim \text{Bernoulli}(p)$$

$X[\tau:]$  が  $\tau$  以降の  $X$

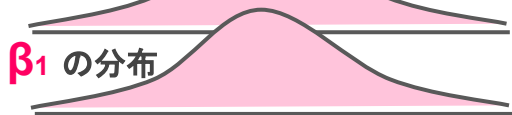
$X[:\tau]$  が  $\tau$  以前の  $X$

便宜的にここでは  $X[\tau:]$  や  $[\mu, \mu]$  のように表記していますが、実装ではライブラリに準拠する必要があるので注意してください。例えば PyMC3 の場合は以下のように実装します。

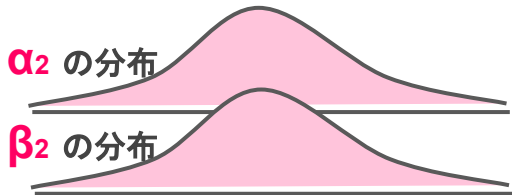
```
mu1 = alpha1 + beta1 * X
mu2 = alpha2 + beta2 * X
mu = pm.math.switch(
    tau > timestamp, mu1, mu2)
p = sigmoid(mu)
obs = pm.Bernoulli(
    'y', p=p, n=4,
    observed=y)
```

# 確率モデルの作り方: 潜在変数モデル

$\alpha_1$  の分布



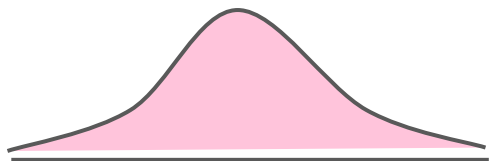
$\beta_1$  の分布



$\alpha_2$  の分布

$\beta_2$  の分布

$T$  の分布

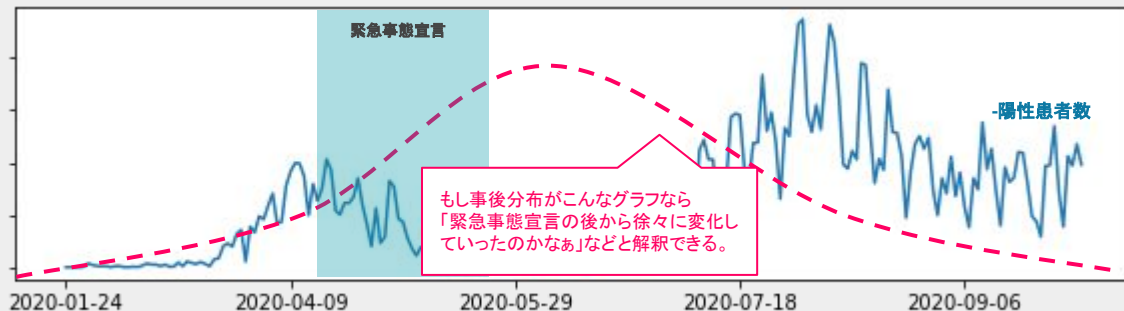


では**事後分布**を解釈していきます。例えば以下のように整理するとわかりやすいでしょう。

| 係数       | before(事後分布の平均) | after(事後分布の平均) |
|----------|-----------------|----------------|
| $\alpha$ | 10.2            | 2.3            |
| $\beta$  | 21.0            | -4.3           |

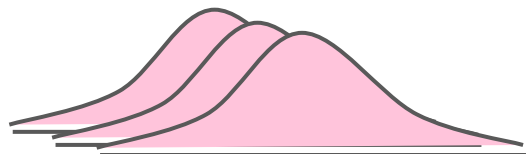
さらに分布で出力しているわけですから、区間を用いて比較してもいいかもしれません。

次に  $T$  を見ていきましょう。 $T$  の山が高い箇所が潜在的な変化点であると解釈します。これは横軸が日付のグラフですので、陽性患者数の推移や緊急事態宣言期間をグラフに重ねるとわかりやすいでしょう。

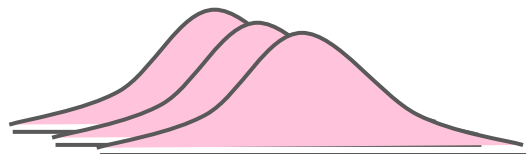


# 確率モデルの作り方: 階層モデル

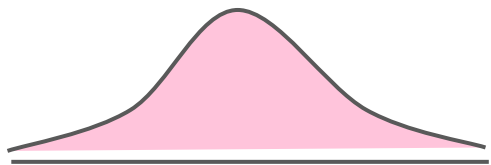
「ケース③共通因子と固有因子」です。需要予測を行う際に、製品別の固有因子と製品間の共通因子を考慮してモデリングを行うシチュエーションです。



$\alpha_1, \beta_1, \sigma_1$  の分布



$\alpha_2, \beta_2, \sigma_2$  の分布



$\beta_{\text{meta}}$  の分布

$$\mu_1 = \beta_{\text{meta}} + \alpha_1 + \beta_1 \times X_1$$

$$y_1 \sim \text{Normal}(\mu_1, \sigma_1)$$

$$\mu_2 = \beta_{\text{meta}} + \alpha_2 + \beta_2 \times X_2$$

$$y_2 \sim \text{Normal}(\mu_2, \sigma_2)$$

$$\alpha_1 \sim \text{Normal}(\alpha_1 \mu, \alpha_1 \sigma)$$

$$\beta_1 \sim \text{Normal}(\beta_1 \mu, \beta_1 \sigma)$$

$$\sigma_1 \sim \text{InvGamma}(\sigma_1 \alpha, \sigma_1 \beta)$$

$$\alpha_2 \sim \text{Normal}(\alpha_2 \mu, \alpha_2 \sigma)$$

$$\beta_2 \sim \text{Normal}(\beta_2 \mu, \beta_2 \sigma)$$

$$\sigma_2 \sim \text{InvGamma}(\sigma_2 \alpha, \sigma_2 \beta)$$

$$\beta_{\text{meta}} \sim \text{Normal}(\beta_{\text{meta}} \mu, \beta_{\text{meta}} \sigma)$$

# 確率モデルの作り方: 階層モデル

事前分布を見えます。ここでは個別の製品1と種類としましたが、もっと種類が多くても同じようにモデリングできます。

ここはとてもシンプルです。それぞれ個別の回帰モデルを考えながら、共通因子用の変数を一つ作っておきます。

共通変数は  $\beta_{\text{meta}}$  としておきました。

製品1の回帰モデル用

$$\alpha_1 \sim \text{Normal}(\alpha_1 \mu, \alpha_1 \sigma)$$

$$\beta_1 \sim \text{Normal}(\beta_1 \mu, \beta_1 \sigma)$$

$$\sigma_1 \sim \text{InvGamma}(\sigma_1 \alpha, \sigma_1 \beta)$$

製品2の回帰モデル用

$$\alpha_2 \sim \text{Normal}(\alpha_2 \mu, \alpha_2 \sigma)$$

$$\beta_2 \sim \text{Normal}(\beta_2 \mu, \beta_2 \sigma)$$

$$\sigma_2 \sim \text{InvGamma}(\sigma_2 \alpha, \sigma_2 \beta)$$

$$\beta_{\text{meta}} \sim \text{Normal}(\beta_{\text{meta}} \mu, \beta_{\text{meta}} \sigma) \quad \text{共通因子}$$

# 確率モデルの作り方: 階層モデル

続いて**尤度関数**です

ではそれぞれの回帰モデルを作しましょう。  
基本的には製品別の回帰モデルをバラバラに作ることにほとんど同じなのですが、 $\beta_{\text{meta}}$  が共通の値として両方の回帰モデルに出現しています。

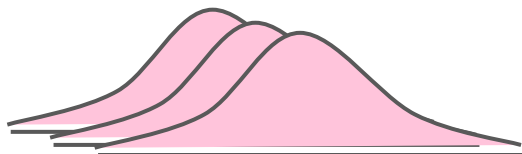
なお、このような個別の因子と共通因子を考慮したモデルを「階層モデル」とか「階層ベイズ」といいます。

$$\begin{array}{c} \text{共通の値} \quad \quad \text{個別の値} \\ \mu_1 = \boxed{\beta_{\text{meta}}} + \boxed{\alpha_1 + \beta_1} \times X_1 \\ y_1 \sim \text{Normal}(\mu_1, \sigma_1) \end{array}$$

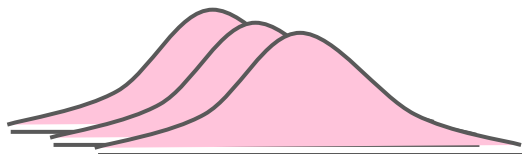
$$\begin{array}{c} \text{共通の値} \quad \quad \text{個別の値} \\ \mu_2 = \boxed{\beta_{\text{meta}}} + \boxed{\alpha_2 + \beta_2} \times X_2 \\ y_2 \sim \text{Normal}(\mu_2, \sigma_2) \end{array}$$

# 確率モデルの作り方: 階層モデル

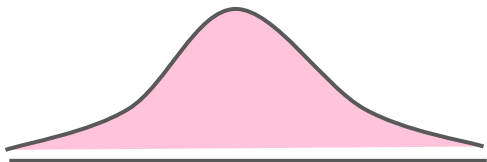
最後に**事後分布**を解釈しましょう。



$\alpha, \beta_1, \sigma_1$  の分布



$\alpha, \beta_2, \sigma_2$  の分布



$\beta_{\text{meta}}$  の分布

では**事後分布**の解釈は簡単ですね。例えば以下のように表記できます。

| 製品区分          | 製品1(事後分布の平均) | 製品2(事後分布の平均) |
|---------------|--------------|--------------|
| $\alpha$      | 10.2         | 2.3          |
| $\beta$       | 21.0         | -4.3         |
| $\text{meta}$ | 5.5          |              |

未来の予測を行う場合は、事後分布の平均値を用いて回帰式にあてはめればOKです。

$$\mu_1 = \beta_{\text{meta}}(5.5) + \alpha_1(10.2) + \beta_1(21.0) \times X_1$$

$$\mu_2 = \beta_{\text{meta}}(5.5) + \alpha_2(2.3) + \beta_2(-4.3) \times X_2$$

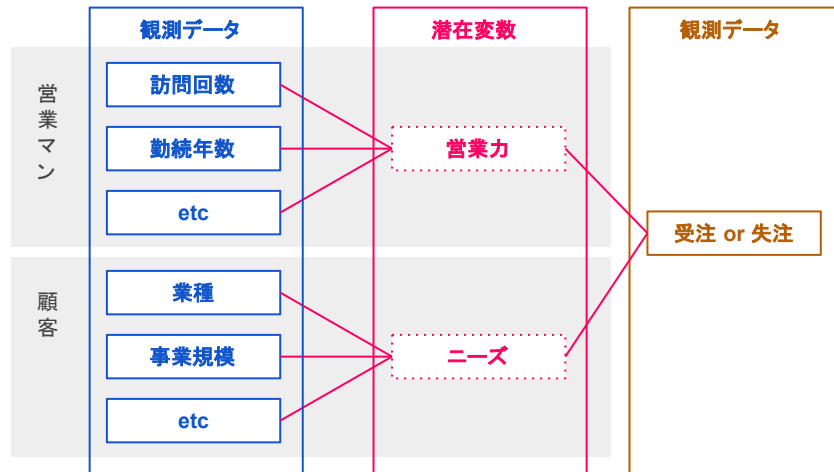
# 確率モデルの作り方:「レシピ」はアイデア次第

ベイズ統計はある程度定型化されたデザインはあるものの、自分の仮定を盛り込んで自由にモデリングできます。今回のアプローチ以外にも様々な応用可能です。例えば以下のようなケースもあるでしょう。

営業会社で「受注するかしないか」をモデル化し背後の構造を検証したい。  
構造として「営業マンの営業力」「顧客のニーズ」によって「受注可否」を説明できると仮定できるとする。しかし「営業力」や「顧客のニーズ」は潜在的な概念であって仮定できてもデータがあるわけではない。  
ただし「営業力」や「顧客のニーズ」を説明する変数はデータから仮定できそう。



モデル化



従来の回帰モデルと比較しても得られる情報の粒度がだいぶ変わってきているのがお分かりでしょうか。  
今回のケースは説明を簡易にするために、線形回帰モデルをベースに説明しましたが、非線形な関係性を考慮する場合は、ガウス過程やベイズ深層学習などの方法もあります。  
また、教師なし学習的な方法もちろんあり、例えば主成分分析のベイズ版である確率的主成分分析や、k-meansの一般化としてGMM(ガウス混合モデル)といったアプローチもあります。  
どのようにしてモデル化するかはみなさん次第です。ぜひ自分なりのアイデアで「レシピ」を作ってみてください。



Appendix

# 条件付確率

# 条件付確率: 問題

## 問題.

袋A、袋Bにはそれぞれ赤玉と白玉が入っています。

それぞれの袋を選ぶ確率、および玉を選ぶ確率が以下に従うとします。

- 袋を選ぶ確率
  - 袋A:  $1/3$
  - 袋B:  $2/3$
- 玉を選ぶ確率(AとBで玉が選ばれる確率が異なる)
  - 袋A
    - 白玉:  $3/4$
    - 赤玉:  $1/4$
  - 袋B
    - 白玉:  $1/2$
    - 赤玉:  $1/2$

では、手元に赤玉がある場合、袋Aが選ばれた確率は？



# 条件付確率：図形で捉える

「袋」「赤玉」「白玉」の選ばれる確率を以下の表で整理します

| 袋の種類 | 選ばれる確率 |
|------|--------|
| A    | 1/3    |
| B    | 2/3    |

袋A

| 玉の種類 | 選ばれる確率 |
|------|--------|
| 赤    | 1/4    |
| 白    | 3/4    |

袋B

| 玉の種類 | 選ばれる確率 |
|------|--------|
| 赤    | 1/2    |
| 白    | 1/2    |

|        |  |         |   |   |   |
|--------|--|---------|---|---|---|
| 2:1で選択 |  | 袋Aは 1:3 |   |   |   |
|        |  | 袋A      | 白 | 白 | 白 |
|        |  | 袋B      | 白 | 白 | 赤 |
|        |  |         | 白 | 白 | 赤 |
|        |  | 袋Bは 1:1 |   |   |   |

Q.手元に「赤玉」がある場合「袋 A」が選ばれた確率は？

答え  $1/5$

|   |   |
|---|---|
|   | A |
| B | B |
| B | B |

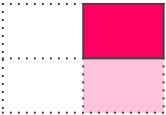
# 条件付確率: 同時確率と条件付確率

ここで改めて数式と各用語を整理するために、各事象の確率を以下のように整理します。

|    | 白    | 赤    |
|----|------|------|
| 袋A | 2/12 | 1/12 |
| 袋B | 4/12 | 4/12 |



| 袋A | 白 | 白 | 白 | 赤 |
|----|---|---|---|---|
| 袋B | 白 | 白 | 赤 | 赤 |
|    | 白 | 白 | 赤 | 赤 |

| 具体例  | 概要                    | 名称    | 数式   |
|--|-----------------------|-------|--|
| 袋Aかつ赤 = $1/12$<br>  | 条件を同時に満たす確率<br>分母が全事象 | 同時確率  | $p(x, y)$                                  |
| 赤の元での袋A = $1/5$<br> | 条件のもとでの確率<br>分母が条件    | 条件付確率 | $p(y x)$<br>xのもとでのy<br>$p(x y)$<br>yのもとでのx |

右が  
条件

# 条件付確率：図形と数式と結びつける

「袋」「赤玉」「白玉」の選ばれる確率

| x = 袋の種類 | 選ばれる確率         |
|----------|----------------|
| A        | $p(x=A) = 1/3$ |
| B        | $p(x=B) = 2/3$ |

袋A

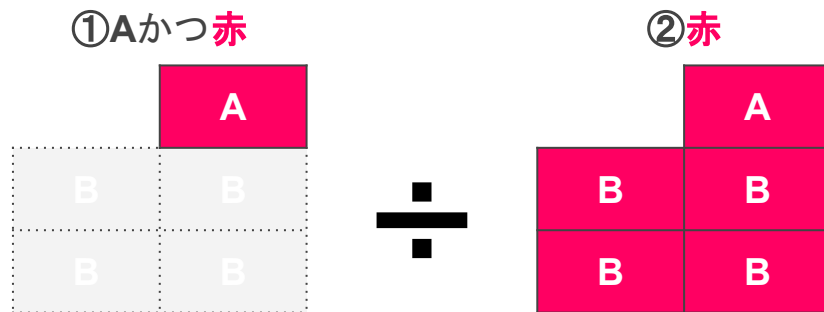
| y = 玉の種類 | 選ばれる確率                      |
|----------|-----------------------------|
| 赤        | $p(y=\text{赤}   x=A) = 1/4$ |
| 白        | $p(y=\text{白}   x=A) = 3/4$ |

袋B

| y = 玉の種類 | 選ばれる確率                      |
|----------|-----------------------------|
| 赤        | $p(y=\text{赤}   x=B) = 1/4$ |
| 白        | $p(y=\text{白}   x=B) = 3/4$ |

Q.手元に「赤玉」がある場合「袋 A」が選ばれた確率は？

→読み替えると $p(x = A | y = \text{赤})$ を求めよという意味



①Aかつ赤

$$p(y = \text{赤} | x = A) \times p(x = A) = \frac{1}{4} \times \frac{1}{3} = \frac{1}{12}$$

②赤

「Aかつ赤」+「Bかつ赤」=「赤」なので

$$\frac{1}{12} + p(y = \text{赤} | x = B) \times p(x = B) = \frac{1}{12} + \frac{1}{2} \times \frac{2}{3} = \frac{5}{12}$$

① ÷ ②なので

$$\frac{1}{12} \div \frac{5}{12} =$$

$$\underline{\underline{1/5}}$$

# 条件付確率: 事前確率, 事後確率, 周辺確率

今回の袋と玉の例をもとに、各用語と数式を整理してみます。

事後確率・事後分布

赤のもとでのAの確率  
 $p(x = A \mid y = \text{赤})$

=

Aかつ赤の確率      事前確率・事前分布

$$p(x = A, y = \text{赤}) = p(y = \text{赤} \mid x = A) \times p(x = A)$$

赤の確率

$$\begin{aligned} p(y = \text{赤}) &= \sum_x p(x, y = \text{赤}) \quad \text{※}\sum_x \text{は} x \text{についての和} \\ &= p(x = A, y = \text{赤}) + p(x = B, y = \text{赤}) \\ &= p(y = \text{赤} \mid x = B) \times p(x = B) + p(y = \text{赤} \mid x = A) \times p(x = A) \end{aligned}$$

周辺確率・周辺分布

ということで式を一般化すると、以下のようにとても単純な式になりました。

$p(x \mid y)$

=

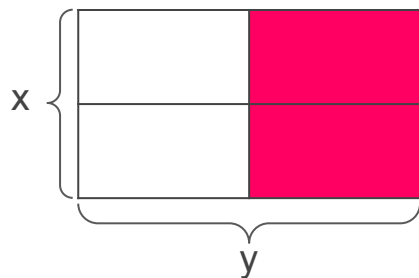
$p(x, y)$

$p(y)$

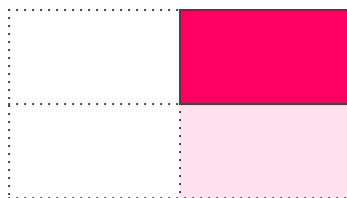
# 条件付確率:連続値の場合

連続値の場合は「袋」や「玉の色」のような「離散値」ではないため、イメージが沸きにくいかもしれませんが、捉え方は同じです。

全事象



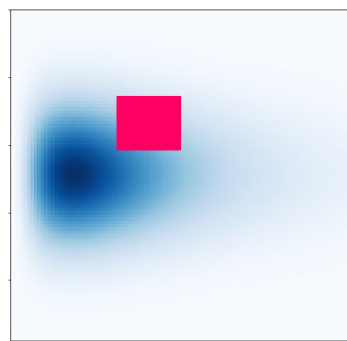
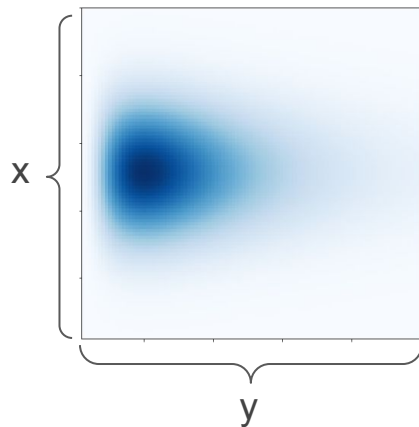
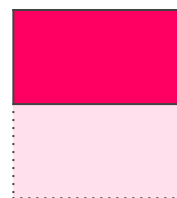
同時確率



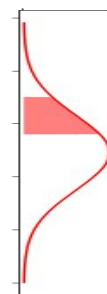
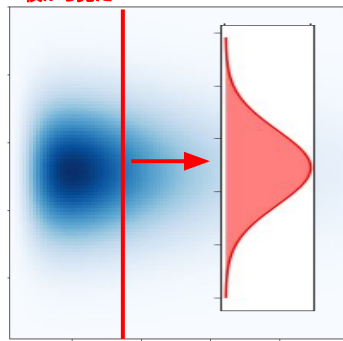
周辺確率



条件付確率



条件を固定して切り出したものを横から見た



# コラム：用語の意味

私(立川)はベイズ統計を勉強し始めた頃に事前分布、事後分布の言葉の意味がわからず「何に対して前で何に対して後なんだろう。」と考え込んでしまい、理解するのに時間がかかってしまいました。ベイズ統計を駆使した確率モデルのデザインに入ると直感的な意味合いも掴めてくる(ように頑張って講義します！)と思いますが、最初のうちはあまり深く考えず、数式に当てはめた「用語」ぐらいに捉えていても大丈夫だと思います。なぜならここまで扱った範囲における条件付き確率はただの「計算」であり、言葉そのものに深い意味がないからです。

今回の例で言えば「手元に赤玉がある場合袋が選ばれた確率」が未知だったので、計算して求めたに過ぎません。ベイズの定理と言う用語も同様です。これも初学者を混乱させる(私がした)要因の一つだと思うのですが、ベイズ統計=ベイズの定理ではありません。

ベイズの定理とは条件付き確率の公式を指しているだけで、ここまでの条件付確率の解説にすでに含んでいます。

もちろん、条件付き確率自体はベイズ統計でも使いますが、ただの条件付き確率の公式なので、どこでも使えるものなんです。

他にもベイズ統計というと頻度主義vsベイズ主義とか主観確率などの用語を目にしたことがある方もいらっしゃると思います。私はこの辺りの用語に「語感から理論を掴める」ような気がして、時間を無駄に浪費してしまいました。今は言葉の意味が分からなくても、少しずつ数式を解いたり、実装して動かしたり、図形を書いたりしながら自分なりに理解できるようになってきましたが、語感で捉えようとしたのはあまり筋が良くなかったなあと反省しています。

まあ誰かに強制されて勉強しているわけではありませんから、そのような反省も楽しみつつ自由に自分のペースでやるに越したことはないとは思っています。

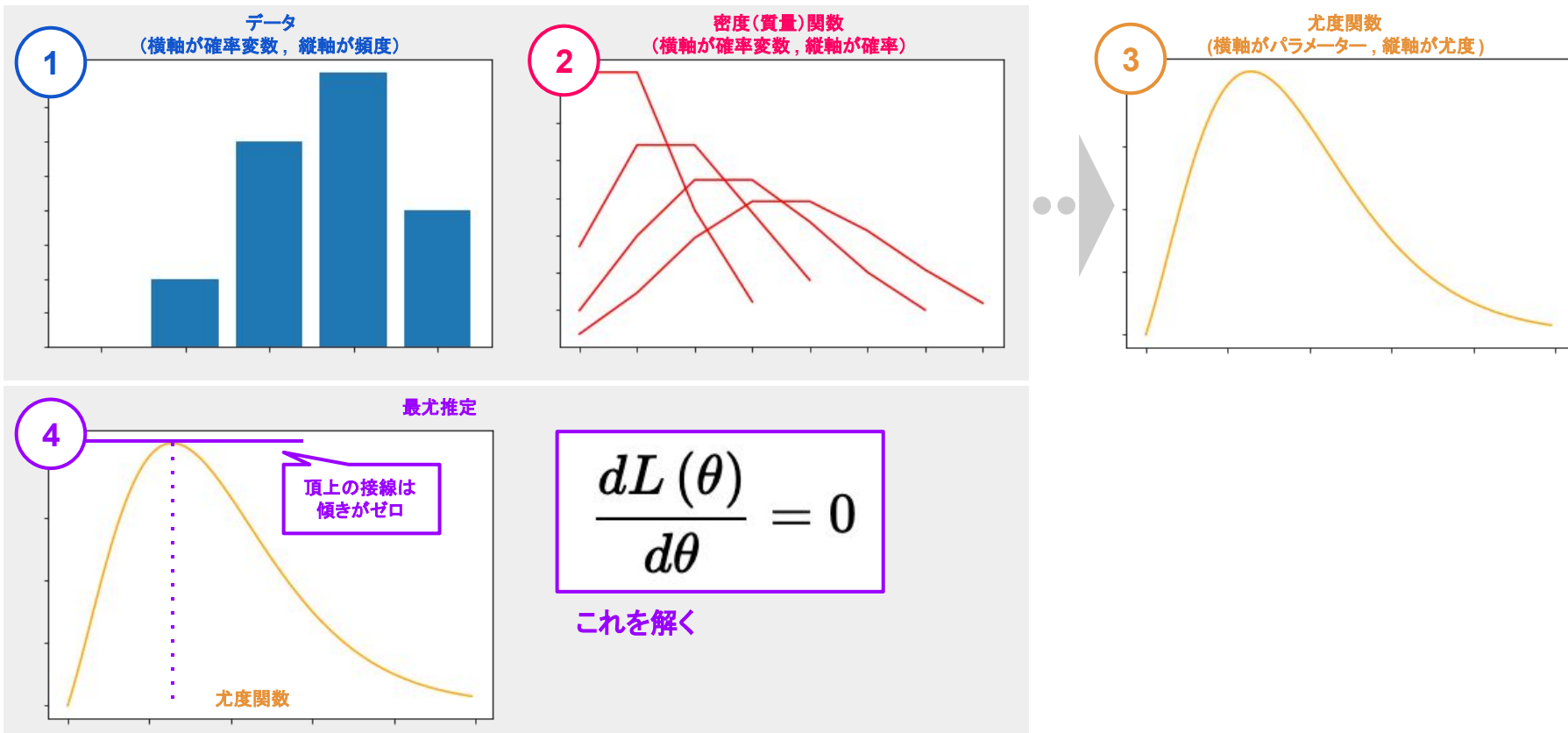


Appendix

# 最尤推定

# 最尤推定：最尤法を用いたパラメーターの推定

## 全体の流れ

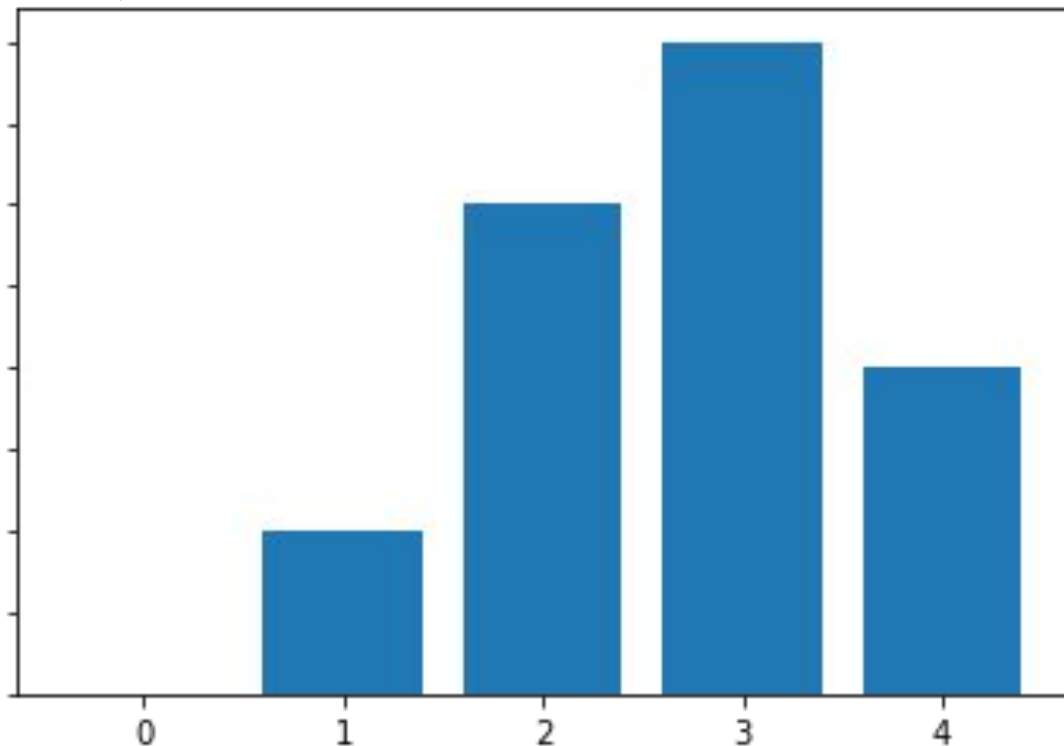


# 最尤推定：確率分布の当てはめ

データ X に当てはまるように確率分布を当てはめてパラメーターを推定したいとします

| id  | X |
|-----|---|
| x1  | 1 |
| x2  | 3 |
| x3  | 4 |
| x4  | 3 |
| x5  | 2 |
| x6  | 4 |
| x7  | 3 |
| x8  | 3 |
| x9  | 2 |
| x10 | 2 |

横軸がX,縦軸が件数



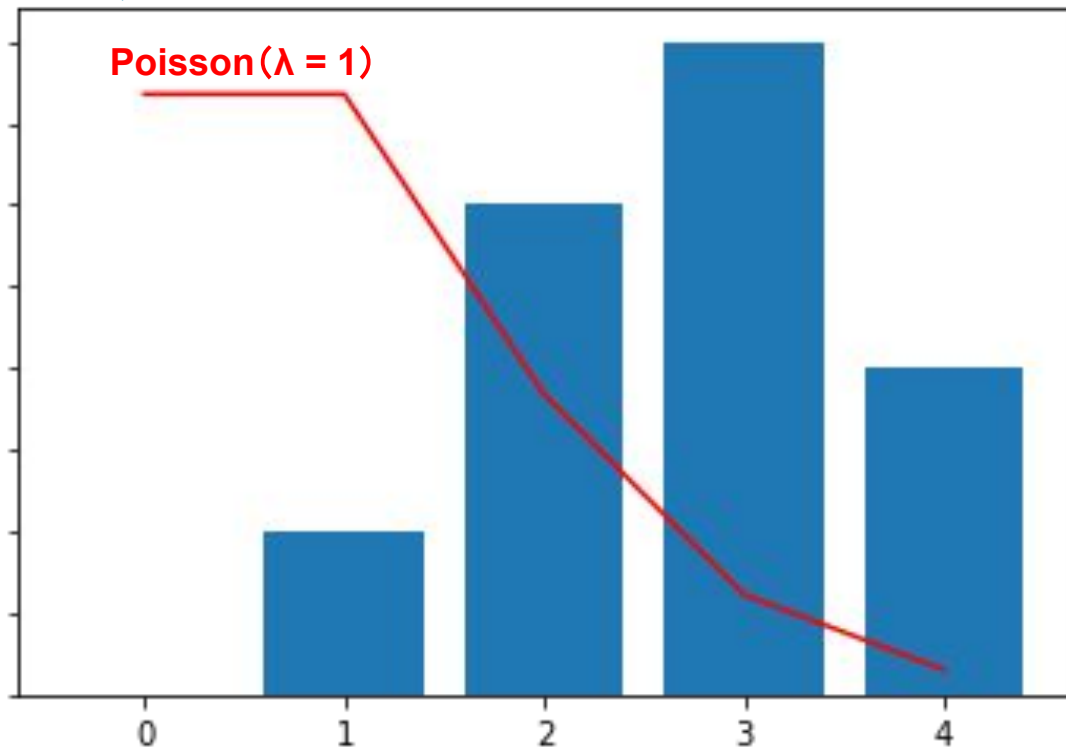
# 最尤推定：確率分布の当てはめ

Poisson分布を用いて  $\lambda = 1$  で当てはめてみます

| id  | X |
|-----|---|
| x1  | 1 |
| x2  | 3 |
| x3  | 4 |
| x4  | 3 |
| x5  | 2 |
| x6  | 4 |
| x7  | 3 |
| x8  | 3 |
| x9  | 2 |
| x10 | 2 |

横軸がX,縦軸が件数

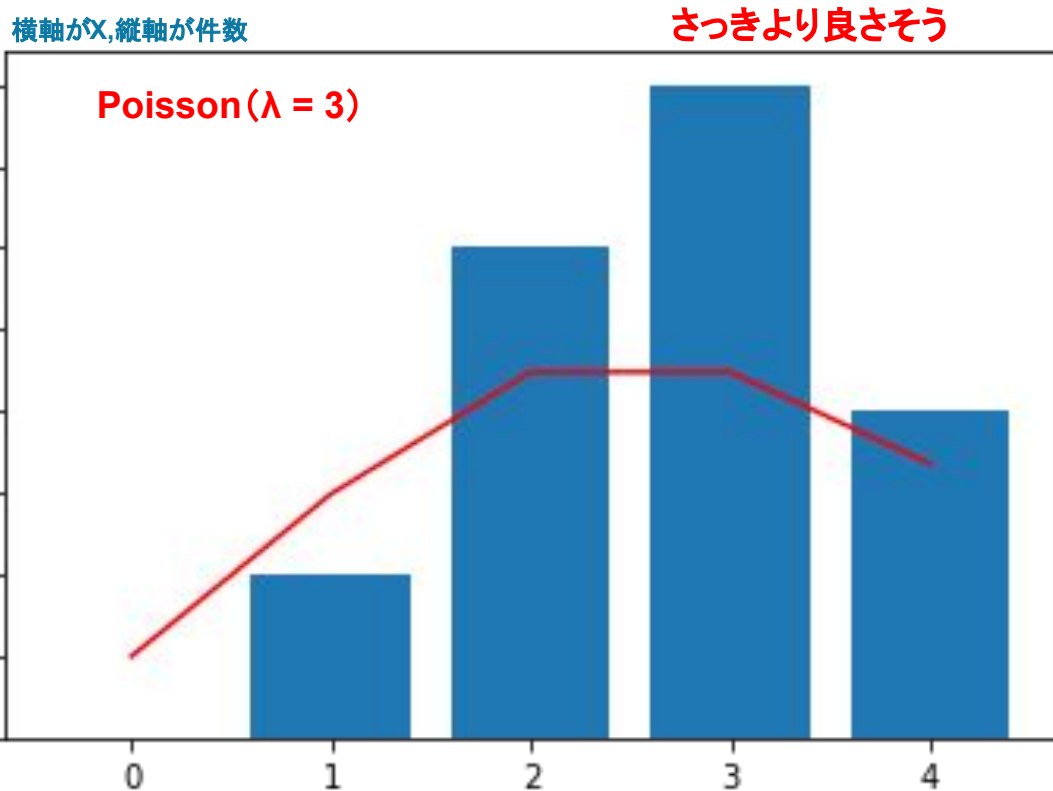
当てはまり悪そう



# 最尤推定：確率分布の当てはめ

Poisson分布を用いて  $\lambda = 3$  で当てはめてみます

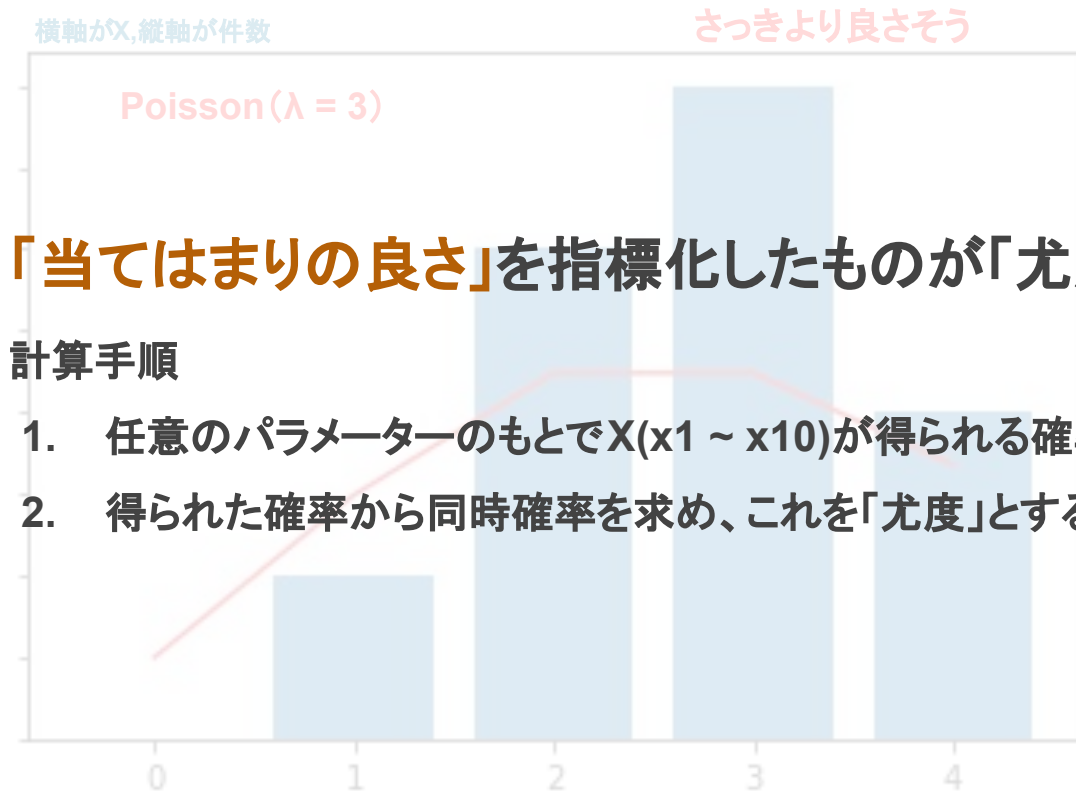
| id  | X |
|-----|---|
| x1  | 1 |
| x2  | 3 |
| x3  | 4 |
| x4  | 3 |
| x5  | 2 |
| x6  | 4 |
| x7  | 3 |
| x8  | 3 |
| x9  | 2 |
| x10 | 2 |



# 最尤推定：確率分布の当てはめ

Poisson分布を用いて  $\lambda = 3$  で当てはめてみます

| id  | X |
|-----|---|
| x1  | 1 |
| x2  | 3 |
| x3  | 4 |
| x4  | 3 |
| x5  | 2 |
| x6  | 4 |
| x7  | 3 |
| x8  | 3 |
| x9  | 2 |
| x10 | 2 |



# 最尤推定：尤度を求める

実際に計算を行い比較してみます

x が得られる確率

| id  | X |
|-----|---|
| x1  | 1 |
| x2  | 3 |
| x3  | 4 |
| x4  | 3 |
| x5  | 2 |
| x6  | 4 |
| x7  | 3 |
| x8  | 3 |
| x9  | 2 |
| x10 | 2 |

| $\lambda = 1$ | $\lambda = 3$ |
|---------------|---------------|
| 36.8%         | 14.9%         |
| 6.1%          | 22.4%         |
| 1.5%          | 16.8%         |
| 6.1%          | 22.4%         |
| 18.4%         | 22.4%         |
| 1.5%          | 16.8%         |
| 6.1%          | 22.4%         |
| 6.1%          | 22.4%         |
| 18.4%         | 22.4%         |
| 18.4%         | 22.4%         |

## $\lambda = 1$ の尤度

$$\begin{aligned} & 36.8\% \times 6.1\% \times 1.5\% \times 6.1\% \times 18.4\% \\ & \quad \times 1.5\% \times 6.1\% \times 6.1\% \times 18.4\% \times 18.4\% \\ & = 7.602e-12 \end{aligned}$$

## $\lambda = 3$ の尤度

$$\begin{aligned} & 14.9\% \times 22.4\% \times 16.8\% \times 22.4\% \times 22.4\% \\ & \quad \times 16.8\% \times 22.4\% \times 22.4\% \times 22.4\% \times 22.4\% \\ & = 1.194e-07 \end{aligned}$$

こちらの方が値が大きい

# 最尤推定：尤度の計算を一般化する

数式で表現してみます。ポアソン分布の確率密度(質量)関数を  $f(x|\lambda)$  とします。

x が得られる確率

| id  | X |
|-----|---|
| x1  | 1 |
| x2  | 3 |
| x3  | 4 |
| x4  | 3 |
| x5  | 2 |
| x6  | 4 |
| x7  | 3 |
| x8  | 3 |
| x9  | 2 |
| x10 | 2 |

| $\lambda = 1$            | $\lambda = 3$            |
|--------------------------|--------------------------|
| $f(x = 1   \lambda = 1)$ | $f(x = 1   \lambda = 3)$ |
| $f(x = 3   \lambda = 1)$ | $f(x = 3   \lambda = 3)$ |
| $f(x = 4   \lambda = 1)$ | $f(x = 4   \lambda = 3)$ |
| $f(x = 3   \lambda = 1)$ | $f(x = 3   \lambda = 3)$ |
| $f(x = 2   \lambda = 1)$ | $f(x = 2   \lambda = 3)$ |
| $f(x = 4   \lambda = 1)$ | $f(x = 4   \lambda = 3)$ |
| $f(x = 3   \lambda = 1)$ | $f(x = 3   \lambda = 3)$ |
| $f(x = 3   \lambda = 1)$ | $f(x = 3   \lambda = 3)$ |
| $f(x = 2   \lambda = 1)$ | $f(x = 2   \lambda = 3)$ |
| $f(x = 2   \lambda = 1)$ | $f(x = 2   \lambda = 3)$ |

## $\lambda = 1$ の尤度

$$f(x = x1 | \lambda = 1) \times f(x = x2 | \lambda = 1) \times \dots \times f(x = x10 | \lambda = 1)$$

$$= \prod_{i=1}^{10} f(x = x_i | \lambda = 1)$$

## $\lambda = 3$ の尤度

$$f(x = x1 | \lambda = 3) \times f(x = x2 | \lambda = 3) \times \dots \times f(x = x10 | \lambda = 3)$$

$$= \prod_{i=1}^{10} f(x = x_i | \lambda = 3)$$

## $\lambda = \theta$ として一般化

$$= \prod_{i=1}^n f(x_i | \theta)$$

$$\prod_{i=1}^n f(x_i | \theta)$$

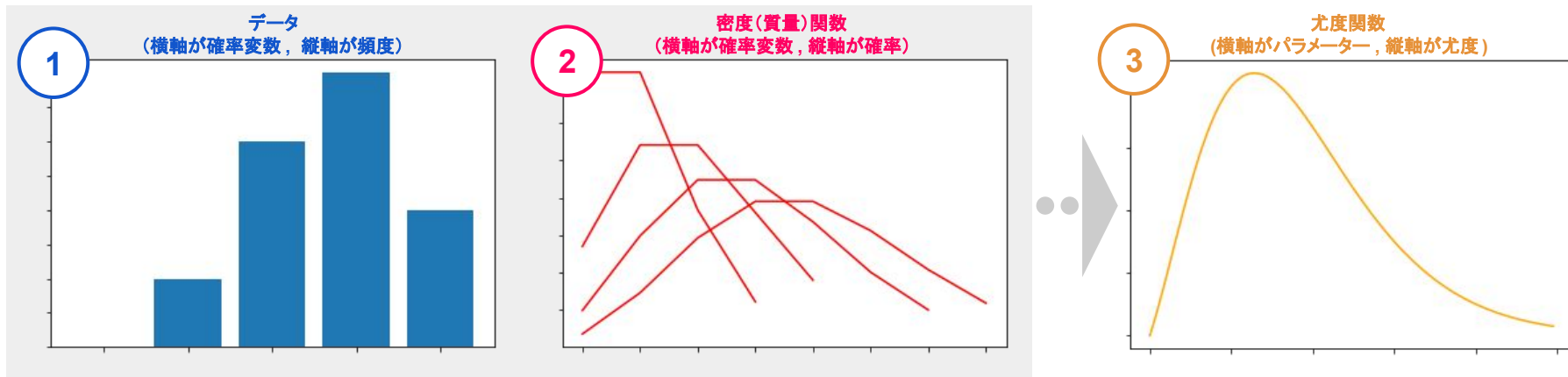
この式はパラメーター入力すると尤度が返ってくる関数になっているので  
**尤度関数**と呼びます。

$\prod$  の記号は「パイ」といいます。円周率  $\pi$  の大文字で数列の掛け算を表しています。 $\Sigma$  の掛け算バージョンですね。

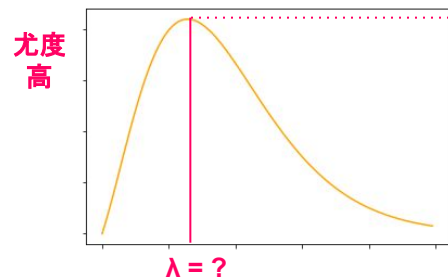
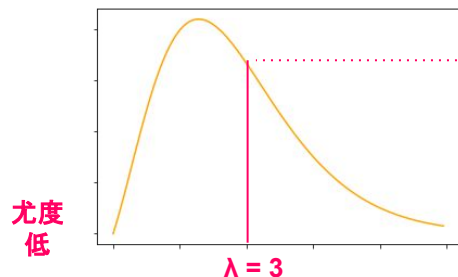
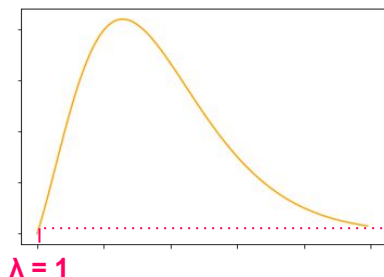


# 最尤推定：尤度関数までのプロセス

尤度を算出する数式を一般化することができました。改めてここまでのプロセスを図示します。



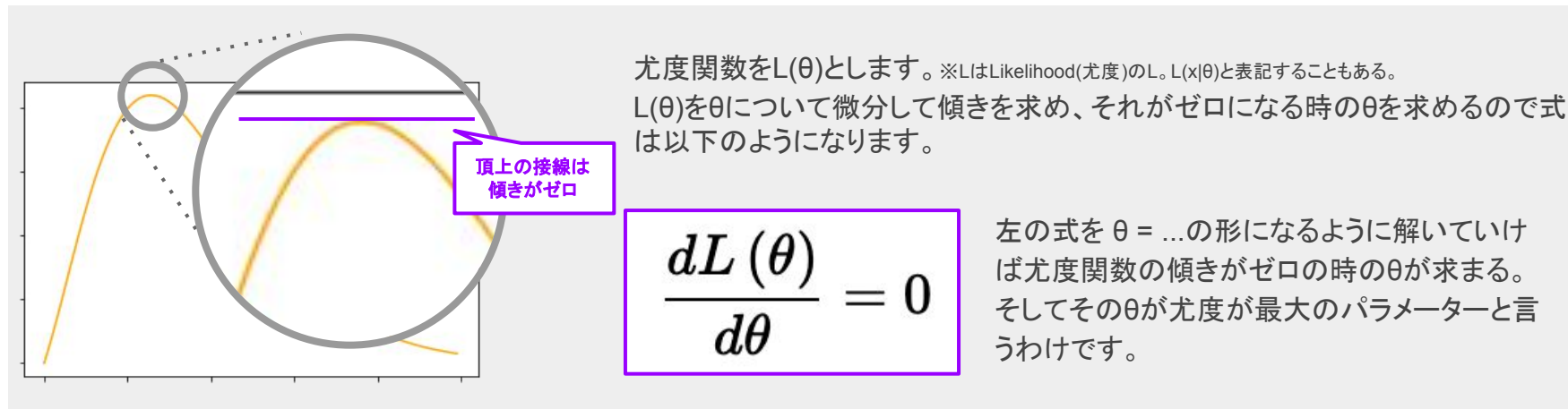
この尤度関数を用いてパラメータを左右に動かしながら「最も尤度が高いパラメーター」を探索します。



どのように探索するのでしょうか？

# 最尤推定：最後のステップ

尤度関数の頂上は傾きがゼロになっているので、尤度関数の微分 = 0 を解いて求めます。

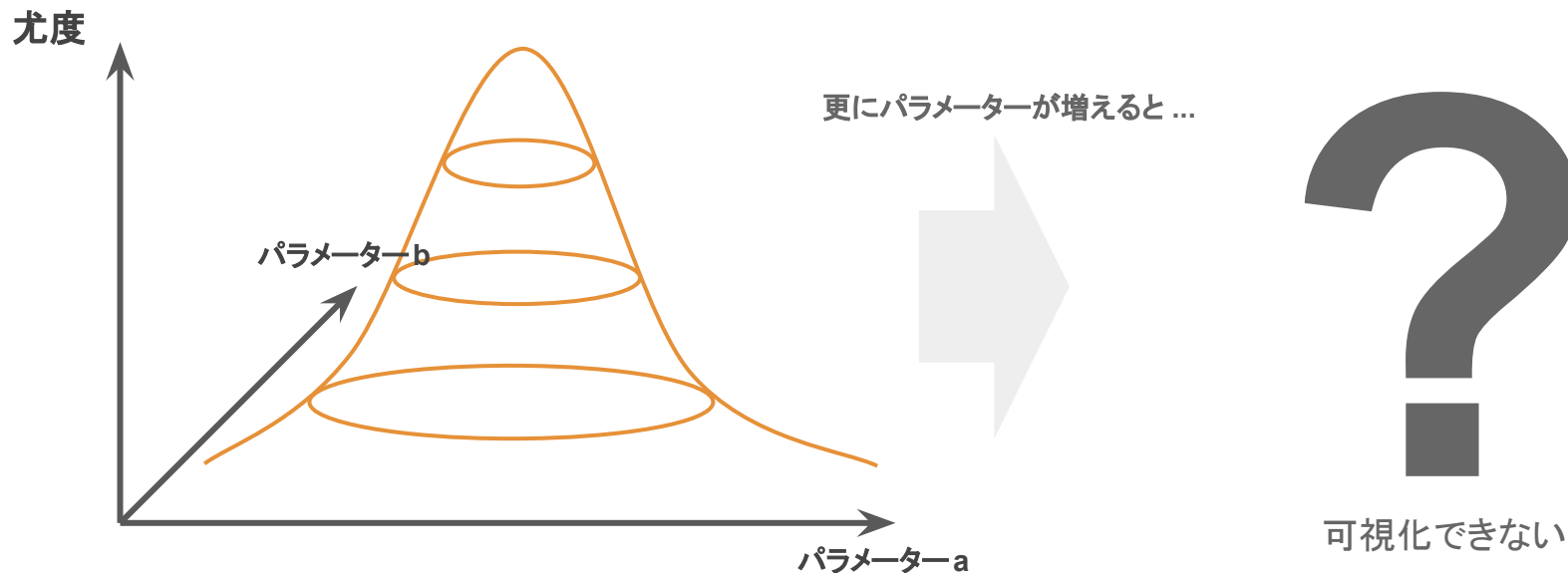


ここでは式の証明は割愛しますが、Poisson分布の最尤推定量(尤度が最大のパラメーター)はデータ $X$ の平均値です。

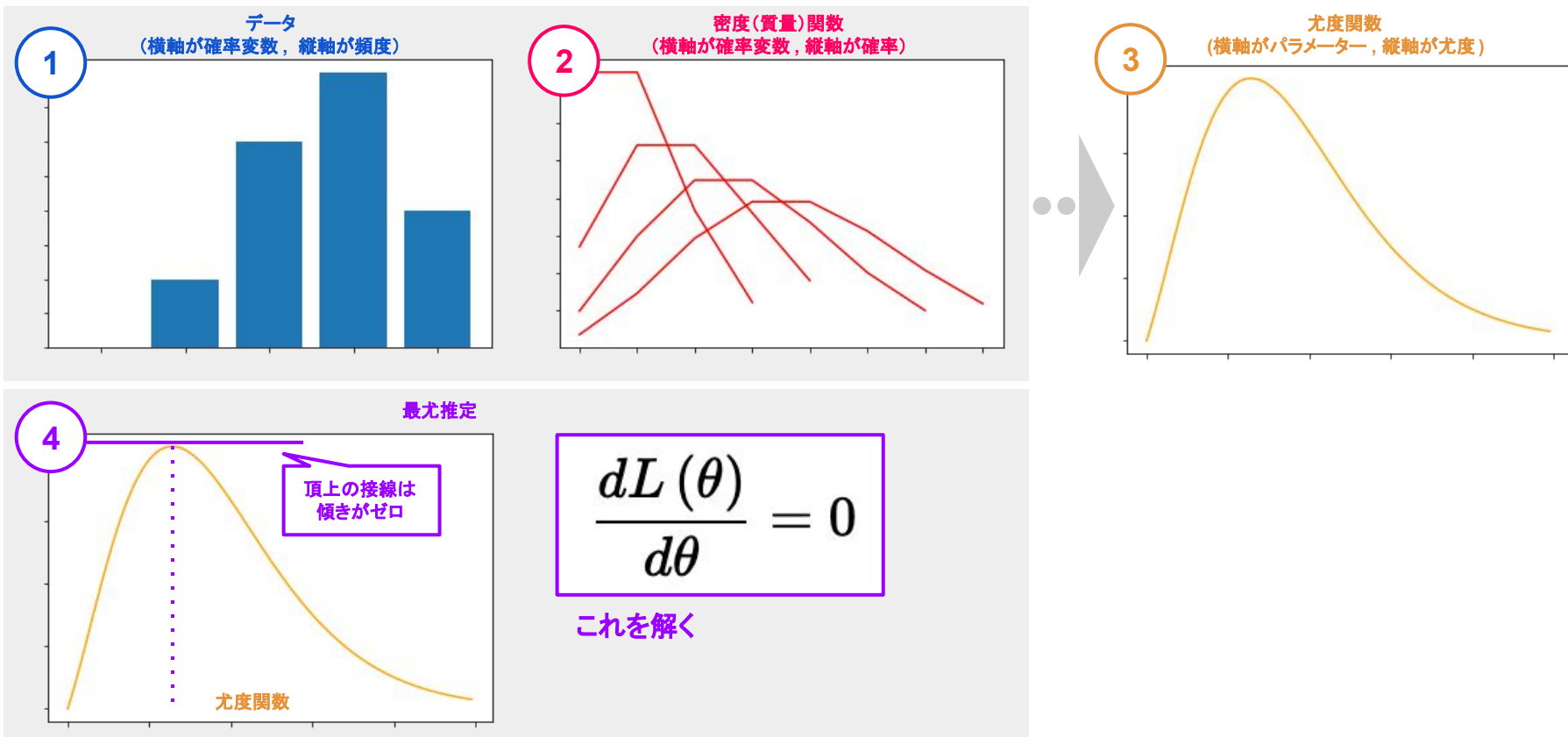
何気なく「パラメーターが平均値だからデータの平均を当てはめてみよう」と考えていたら、それは実は最尤推定を行っていたことになるわけです。

# 最尤推定：パラメーターが複数ある場合

正規分布のようにパラメーターが複数ある場合はどのように最尤推定を行うのでしょうか。  
基本的には同じですが、尤度関数の「次元」が変わります。  
例えばパラメーターが2つある場合の尤度関数は3次元になります。



# 最尤推定：まとめ



# コラム: 機械学習と統計学

最尤推定のプロセスについて触たところで、機械学習との関連について言及しておきたいと思います。

最尤推定は「尤度=当てはまりの良さ」を最大化するパラメーターを求めることでした。機械学習は目的関数を最小化するパラメーターを求めています。つまり、目的関数をマイナス(対数)尤度とすると機械学習と同じことをしているわけです。

一番わかりやすい線形回帰モデルを用いて具体例を示します。

|          | 目的関数 | 学習アルゴリズム |
|----------|------|----------|
| 機械学習ばい表現 | MSE  | 最小二乗法    |
| 統計ばい表現   | 尤度   | 最尤推定     |

実は最尤推定を解いていくと、MSEが現れ、結局最小二乗法と等価になります。他にもK-means法の捉え方を変えたものとしてGMMという手法があり、各クラスターのラベルを確率で出力できるのですがGMMのデザインによつてはK-means法と等価にすることができます。

※機械学習ばい、統計ばいというのは立川の主観

ところで、AICは汎化性能に着目した指標であることを覚えている方もいるかもしれませんが、これは機械学習における交差検証と同じ役割を持っています。ここにも機械学習と共通するコンセプトがあるわけですね。

ちなみにですが、尤度という指標はとても便利で、目的関数と評価関数を分ける必要すらありません。さらに尤度を発展させた指標としてAIC、BIC、WAIC、WBICなどの指標が存在しますが、基礎となる考え方は共通しています。

勉強を始めた頃は、機械学習と統計学を分けて捉えた方が、分野を特定しやすく勉強に取り組みやすいかもしれませんが、本来分けて考えるものではないと私(立川)は考えています。

Appendix

# ベイズ統計基礎理論

# ベイズ統計基礎理論：イカサマコインを見抜く

手元にあるコインが「イカサマコイン」か見抜くために、コインを投げて表裏の回数を数えてみたしましょう。もし2回だけコインを振ってみて、2回連続で表が出たら「イカサマだ！」と考えるのでしょうか？ちょっと無理がありますよね？

表が出る確率が50%ぴったりと仮定しても、2回連続で表が出る確率が25%もあるからです。

| 1回目の結果 | 確率  | 1,2回目の結果 | 確率  | 表が出る回数とその確率 |     |
|--------|-----|----------|-----|-------------|-----|
| 表      | 50% | 表 表      | 25% | 表が出る回数      | 確率  |
|        |     | 表 裏      | 25% | 2回(表表)      | 25% |
| 裏      | 50% | 裏 表      | 25% | 1回(表裏+裏表)   | 50% |
|        |     | 裏 裏      | 25% | 0回(裏裏)      | 25% |

「イカサマコイン」を見抜くためには、何度かコインを振ってみて区間推定を用いることで検証することができます。しかし、そもそも「表が出る確率」の確率分布を求めることはできないのでしょうか？

# ベイズ統計基礎理論：変数と関数を確認

「表が出る確率」の確率分布を求めるにあたって、変数や関数を以下のように定義します。

なお、表が出る回数 が従う分布を二項分布とします。できるだけ同じ意味のものを同じ色で分けるので注意して見てみてください。

## 二項分布の各変数の定義

| 変数  | 意味        | 取りうる最小値 | 取りうる最大値 | 値の種類 |
|-----|-----------|---------|---------|------|
| $X$ | 表が出た回数    | 0       | $n$     | 離散値  |
| $p$ | 表が出る確率    | 0       | 1       | 連続値  |
| $n$ | コインを振った回数 | 0       | 無限      | 離散値  |

## 二項分布の確率質量関数の定義

例. $p=50\%$ と仮定して「2回振って2回とも表」の確率を算出する場合

$$25\% = \text{Binomial}(X=2, n=2, p=0.5)$$

25%も確率、パラメーターの  $p$  も確率でややこしいですが、25%という値はパラメーター  $p=0.5$  のもとで今回の結果が得られる確率つまり尤度を表しています。今回の場合は「試行回数 ( $n$ ) を2回」という1セットを1度だけなので、そのまま尤度関数とします。

尤度関数は以下のように定義しておきます。 ※ $n$ は今回注目する必要がないので省略

$$\text{尤度} = \text{Binomial}(X | p)$$



# ベイズ統計基礎理論：パラメーターが従う分布

では改めて「表が出る確率」の確率分布とはどのような分布なのでしょう。整理すると「表が出た回数  $X$  のもと  $p$  が従う分布」と考えることができます。一旦「表が出る確率」の確率密度関数を  $f()$  とすると...

「表が出る確率」の密度関数  $f(p | X)$  に対して尤度関数は  $\text{Binomial}(X | p)$  でした。

どちらも条件付き確率になっていますね。では「条件付き確率の公式」に当てはめて2つの関数を結び付けてみましょう。

$$\begin{array}{c} \text{事後分布} \\ f(p | X) \\ \text{pが従う分布} \end{array} = \frac{\begin{array}{c} \text{尤度} \\ \text{Binomial}(X | p) \end{array} \times \text{事前分布}(p)}{\int \text{Binomial}(X | p) \times \text{事前分布}(p) dp}$$

※  $p$  についての積分 パラメーターが離散の場合に  $\Delta p$  で足し上げる

周辺尤度・周辺分布 尤度 × 事前分布における全事象を足し上げて1にするための正規化定数

事前分布は  $p$  を確率変数として入力すると確率を出力する確率密度関数になっています。事前分布を決めれば事後分布を求めることができますがその前に式を一般化しておきましょう

# ベイズ統計基礎理論:事後分布の一般公式

一般化するにあたって変数と関数を以下のように定義します。

- データ:  $X$
- パラメーター:  $\theta$
- 事後分布:  $\text{Post}(\theta | X)$
- 尤度:  $L(X | \theta)$
- 事前分布:  $\text{Pre}(\theta)$

$$\text{Post}(\theta | X) = \frac{L(X | \theta) \times \text{Pre}(\theta)}{\int L(X | \theta) \times \text{Pre}(\theta) d\theta}$$

周辺尤度は全事象を足しあげて 1 にするための正規化定数なので、省略して以下のように表記することもあります。  
 $\propto$  は 比例を表す記号で読み方もそのまま「比例」です。

$$\text{Post}(\theta | X) \propto L(X | \theta) \times \text{Pre}(\theta)$$

$\propto$   
比例を表す記号

だいぶスッキリしましたね。ではいよいよ事前分布を定義しましょう。

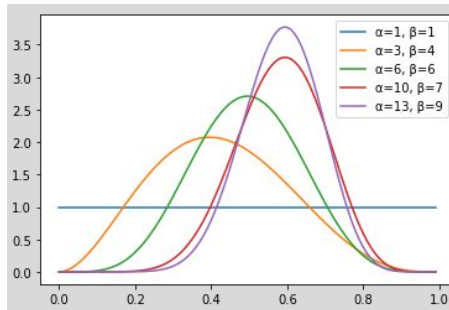
# ベイズ統計基礎理論: 事前分布を定義する

二項分布のパラメーター  $p$  は  $0 \sim 1$  の連続値なので、確率変数が  $0 \sim 1$  の連続値をとる確率分布が良さそうです。そのような分布で代表的なものはBeta分布ですね。Beta分布は以下のような性質を持っています。

Beta分布の確率変数とパラメーター

|         |  |
|---------|--|
| 確率変数の種類 | 連続値  |
| 確率変数の範囲 | $0 \sim 1$   |
| パラメーター  | $\alpha$ (表が出る回数+1)<br>$\beta$ (裏が出る回数+1)<br>表と裏は逆でもOK |

Beta分布の確率密度関数



$\alpha=\beta=1$  (表裏=0) の時は情報が無いので、真横の直線 (一様分布)。それぞれの回数が増えていくとだんだん裾が「狭く」なります。「狭く」ということは二項分布の  $p$  (Beta分布の確率変数、ここでは横軸) に対する「自信」が高まっていると言えます。

実は尤度関数が二項分布の場合、事前分布に Beta分布を用いると事後分布を解析的に求めることができ、事後分布もなんと Beta分布になります。このような関係性を **共役** といい、例えば「**Beta分布は二項分布の共役事前分布**」と表現します。

$$\text{Beta}(p | X) = \frac{\text{Binomial}(X | p) \times \text{Beta}(p | \alpha, \beta)}{\int \text{Binomial}(X | p) \times \text{Beta}(p | \alpha, \beta) dp}$$

また事前分布にパラメーター ( $\alpha$  と  $\beta$ ) が出てきてしまいました。どのように求めるのでしょうか

# ベイズ統計基礎理論：事前分布のパラメーターを考える

事前分布には結果がまだない状態= 無情報、つまり表も裏も0回の状態を与えてみます。  
その上で「表が2回」「裏が0回」という今回の結果を使ってみましょう。

$\text{Beta}(p | X)$   
どんな分布になる？

=

このXには  
表:2回 裏:0回という結果が入っている

$\text{Binomial}(X | p)$

$\alpha = \text{表が出た回数} + 1$   
 $\beta = \text{表が出た回数} + 1$   
なので、以下のようにします。

$\times$

$\text{Beta}(p | \alpha=1, \beta=1)$

/

$\int \text{Binomial}(X | p) \times \text{Beta}(p | \alpha=1, \beta=1) dp$

この計算を解くと**事後分布は表が2回、裏が0回のBeta分布**になります。

**Beta**  
 $\alpha, \beta$  表=2,  $\alpha=3$   
裏=0,  $\beta=1$

=

**Binomial**  
 $X$  表 = 2  
裏 = 0

+

**Beta**  
 $\alpha, \beta$  表=0,  $\alpha=\text{表}+1$   
裏=0,  $\beta=\text{裏}+1$

この性質を用いると、今後新たにコインを振る場合に、表 2回裏0回の結果を事前分布に与えれば、その結果を踏まえた事後分布を導くことができます。更に事前の結果がなくとも「これくらいかな？」という想定を事前分布に与えることもできます。

二項分布とBeta分布の場合は単純な足し算で表現できますが、この性質は分布のペアによって異なります。また、そもそも Beta分布に直接結果を入力すれば事後分布が得られるため、実務的にはこのまま使用することはありませんが、ベイズ統計の基礎理論としてこの性質は重要なので覚えておきましょう。

# ベイズ統計基礎理論：Beta分布と二項分布の関係

ではBeta分布と二項分布の関係について確率密度関数を用いて詳しくみていきます。

## 二項分布

|        |   |
|--------|---|
| 確率質量関数 | $Binom(x) = {}_n C_x p^x (1 - p)^{n-x}$ |
| 確率変数   | x : 表が出る回数                              |
| パラメーター | p : 表が出る確率, n : コインを振る回数                |

## Beta分布

|        |   |
|--------|---|
| 確率質量関数 | $Beta(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1 - p)^{\beta-1}$ |
| 確率変数   | p : 表が出る確率  |
| パラメーター | $\alpha$ : 表が出る回数 + 1, $\beta$ : 裏が出る回数 + 1   |

# ベイズ統計基礎理論：確率密度関数の「基底」

どちらの分布にも似たような構造があります。

## 二項分布

|        |   |
|--------|---|
| 確率質量関数 | $Binom(x) = {}_n C_x p^{\boxed{x}} (1-p)^{\boxed{n-x}}$ <span style="color: red;">表の回数</span> <span style="color: red;">裏の回数</span> |
| 確率変数   | x : 表が出る回数  |
| パラメーター | p : 表が出る確率, n : コインを振る回数  |

## Beta分布

|        |   |
|--------|---|
| 確率質量関数 | $Beta(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\boxed{\alpha-1}} (1-p)^{\boxed{\beta-1}}$ <span style="color: red;">表の回数</span> <span style="color: red;">裏の回数</span> |
| 確率変数   | p : 表が出る回数 <span style="color: red;">実はこの共通部はそれぞれの「分布の形」を定義する役割を持っています。<br/>形を定義する関数を「基底」や「カーネル」といいます。<br/>それぞれの確率変数が異なる点には注意しましょう。</span>  |
| パラメーター | $\alpha$ : 表が出る回数 + 1, $\beta$ : 裏が出る回数 + 1   |

# ベイズ統計基礎理論：確率密度関数の「正規化定数」

では、同じではない構造の部分は何を意味しているか？

## 二項分布

|        |   |
|--------|---|
| 確率質量関数 | $Binom(x) = {}_n C_x p^x (1 - p)^{n-x}$ |
| 確率変数   | x : 表が出る回数                              |
| パラメーター | p : 表が出る確率, n : コインを振る回数                |

## Beta分布

|        |   |
|--------|---|
| 確率質量関数 | $Beta(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1 - p)^{\beta-1}$ |
| 確率変数   | p : 表が出る回数    この「共通しない部分」は確率変数について積分した際に 1になるようにするための定数で「正規化定数」と言います                                  |
| パラメーター | $\alpha$ : 表が出る回数 + 1, $\beta$ : 裏が出る回数 + 1   |

# ベイズ統計基礎理論: 基底と正規化定数の関係

「確率変数について積分し 1 になるように」とはつまり以下の関係です。

正規化定数  $n$                       基底

二項分布

$$A \sum_{i=0}^n p^{x_i} (1 - p)^{n-x_i} = 1$$

x について積分, ただし x は離散なので和を取って 1 にする。  
これを A について解くと  $nCx$  がもとまる

正規化定数                      基底

Beta分布

$$B \int_0^1 p_i^{\alpha-1} (1 - p_i)^{\beta-1} dp = 1$$

p について積分. これを B について解くと  $\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}$  がもとまる



# ベイズ統計基礎理論:事後分布の導出

では「正規化定数」と「基底」うまく使って事後分布を求めます。事後分布は尤度関数と事前分布を掛け算すれば良いのですが、計算を簡単にするためにまず「基底」どうしで掛けます。

$$\begin{aligned} & p^x(1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^x p^{\alpha-1} (1-p)^{n-x} (1-p)^{\beta-1} \\ &= p^{\alpha+x-1} (1-p)^{\beta+n-x-1} \end{aligned}$$

ここで  $\alpha + x = \alpha_{post}$ ,  $\beta + n - x = \beta_{post}$  とすると

$$= p^{\alpha_{post}-1} (1-p)^{\beta_{post}-1}$$

となりこれはBeta分布の「基底」です。この「基底」に「正規化定数」をかけてあげれば確率分布となります。

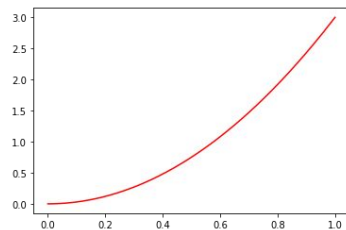
$$Posterior(p) = \frac{\Gamma(\alpha_{post} + \beta_{post})}{\Gamma(\alpha_{post})\Gamma(\beta_{post})} p^{\alpha_{post}-1} (1-p)^{\beta_{post}-1}$$

この計算を覚える必要はありませんが、確率密度関数は規定と正規化定数で構成されていることを知っておくと、後々自分で勉強する際にわかりやすくなるので頭の片隅に入れておくといいでしょう。

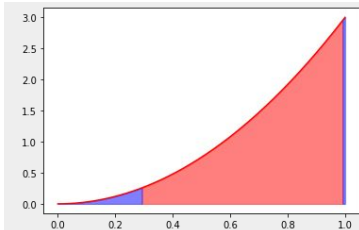
# ベイズ統計基礎理論：区間推定は両端だけ見ればいい？

事後分布が求まりましたので、区間推定を行ってみましょう。しかし、従来の区間推定の考え方通り、95%区間の上端と下端を求めると最頻値が区間外に位置してしまいます。また事後分布は多峰性のある分布になってしまうことしばしばあるため、両端に着目した区間推定の場合は、分布の「谷」が区間内になってしまう事もあります。そのため、事後分布の区間推定では頻度に着目した**HPD(最高事後密度)区間**を用いることがあります。

事後分布(Beta分布)

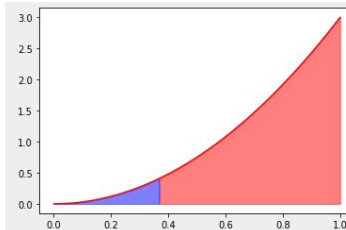


両端から求める区間95%



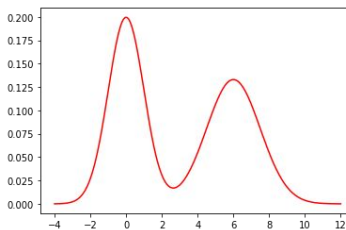
下端: 29%  
上端: 99%

HPD区間(95%)

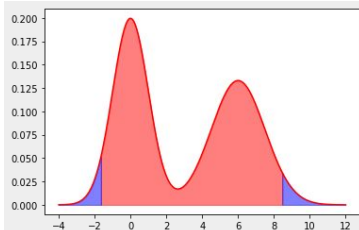


下端: 37%  
上端: 100%

事後分布(多峰分布)

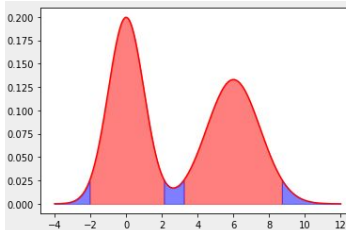


両端から求める区間95%



下端: -1.6  
上端: 8.5

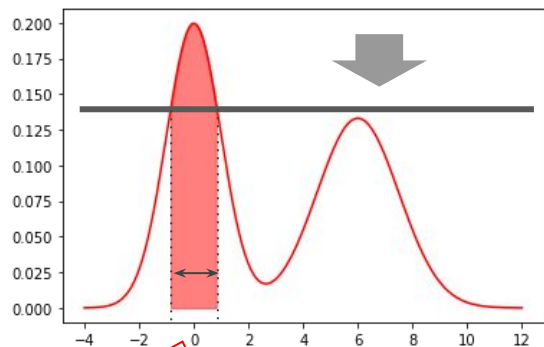
HPD区間(95%)



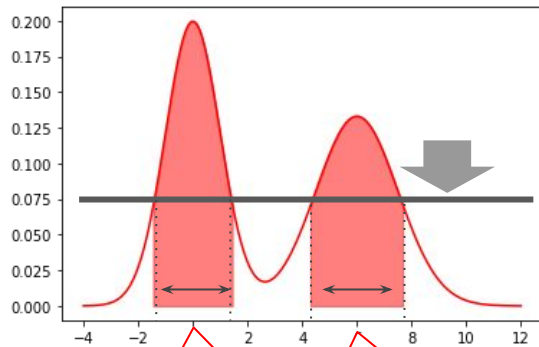
左の山  
・下端: -2.0  
・上端: 2.1  
右の山  
・下端: 3.2  
・上端: 8.7

# ベイズ統計基礎理論:HPD区間算出の考え方

HPD(最高事後密度)区間の算出は頻度に着目して上から積分区間を定義していきます。積分した値が95%に満たした時点でそれ以上積分しないようにすれば終了というわけです。もちろん95%以外の閾値を設けることも可能です。

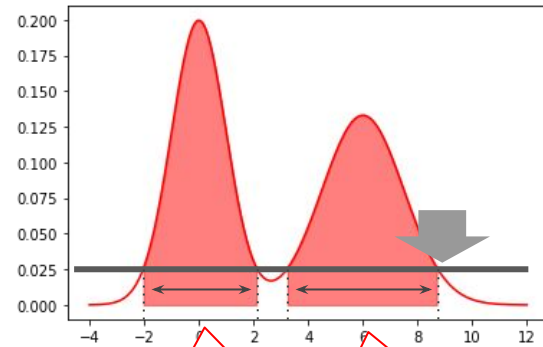


この区間で積分



この区間で積分

この区間で積分



この区間で積分

この区間で積分

この方法なら必ず最頻値を含むことができますし、多峰性のある分布の「谷」を区間外とすることができます。

# ベイズ統計基礎理論：イカサマコインだった？

では、当初のテーマに戻りましょう。

2回コインを振って2回連続で表が出ました。

この結果をもとに、二項分布とBeta分布を使って事後分布を求め、HPD95%区間を用いると、

「表が出る確率  $p$  は 少なくとも約37%以上である」

という結果になりました。

この結果を持って「イカサマ」とするかどうかは人間の判断によります。

# Appendix

# MCMC

# MCMC: 改めて事後分布の算出式

サンプリングアルゴリズムにもいくつか種類がありますが、ここでは代表的な「MCMC」について解説します。  
まずは**事後分布**と**尤度関数**、**事前分布**の関係を振り返ります。

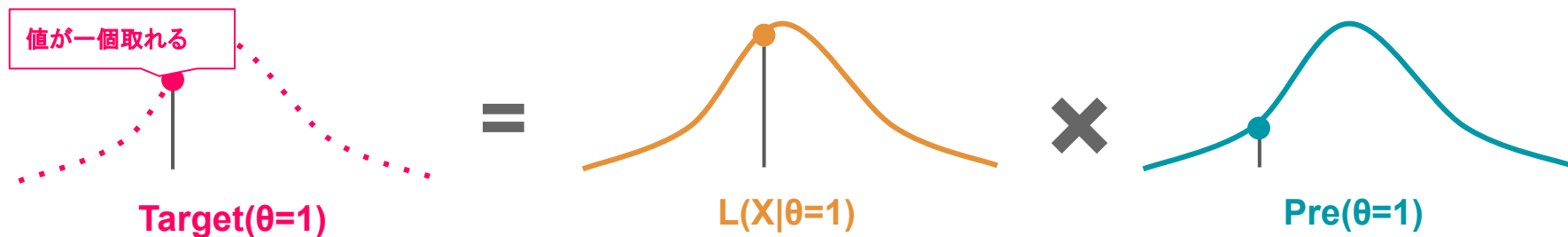
A diagram illustrating the relationship between three probability distributions. On the left is a pink bell-shaped curve labeled **事後分布** (Posterior) and  $\text{Post}(\theta)$ . In the middle is an orange bell-shaped curve labeled **尤度関数** (Likelihood) and  $L(X|\theta)$ . On the right is a teal bell-shaped curve labeled **事前分布** (Prior) and  $\text{Pre}(\theta)$ . Between the pink and orange curves is a large infinity symbol ( $\infty$ ). Between the orange and teal curves is a large multiplication symbol ( $\times$ ).

事後分布を「確率分布」とするためには正規化定数である「周辺尤度」の計算が必要ですが、複雑な確率モデルでは周辺尤度がもとまらないこともあります。MCMCではそのボトルネックを解消するために、周辺尤度を用いません。  
したがって以後、MCMC法における「事後分布」と「確率分布の性質を満たす事後分布」を区別するために「**目標分布**」と表記します。

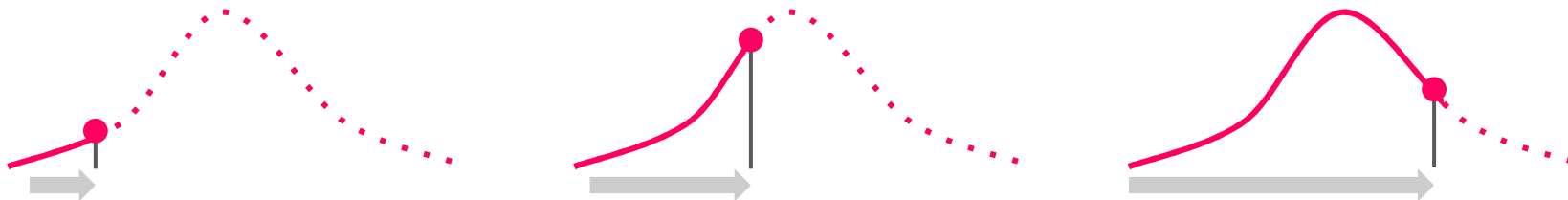
A diagram illustrating the target distribution. On the left is a pink bell-shaped curve labeled **目標分布** (Target) and  $\text{Target}(\theta)$ . Below it, in pink text, is the note **積分して1にならない** (Does not integrate to 1). In the middle is an orange bell-shaped curve labeled **尤度関数** (Likelihood) and  $L(X|\theta)$ . On the right is a teal bell-shaped curve labeled **事前分布** (Prior) and  $\text{Pre}(\theta)$ . Between the pink and orange curves is an equals sign ( $=$ ). Between the orange and teal curves is a large multiplication symbol ( $\times$ ).

# MCMC: Monte Carlo 法

MCMCとは「Monte Carlo」「Markov Chain」の略語で、それぞれのアプローチを組み合わせたものを意味しています。まず「Monte Carlo」法のアプローチから考えましょう。目標分布の形がわからないとはいえ、 $\theta$  に数字を代入してあげれば、目標分布の値を取り出すことができます。そうですね？事前分布の $\theta$ が一つ決まれば尤度も計算できますから。



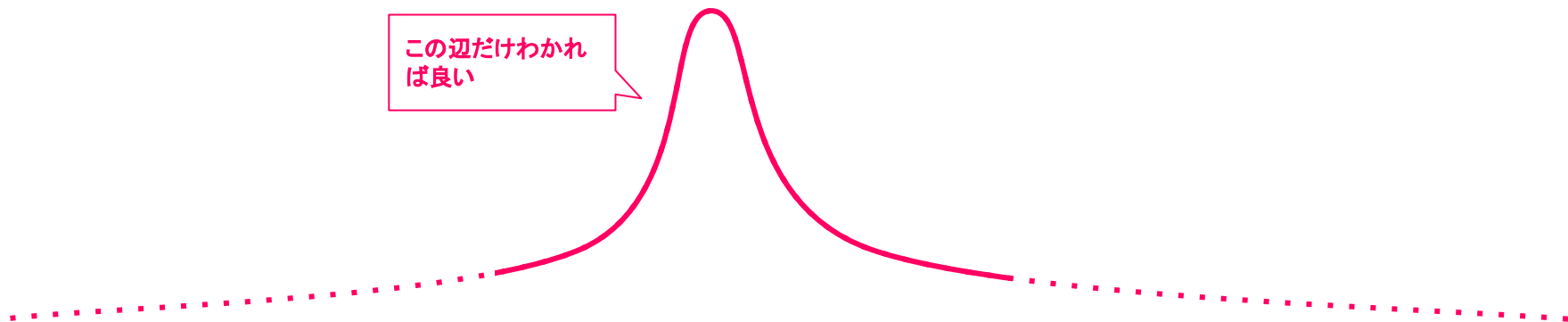
ということは  $\theta$  の値を物凄く小さい値から物凄く大きい値まで 0.0001 などの小さい値刻みで実験すれば提案分布の「形」はほぼわかるはずです。



これが「Monte Carlo」法です。

# MCMC: Monte Carlo 法の問題点

Monte Carlo 法はシンプルでわかりやすいのですが、非常に無駄が多い方法です。  
実際は山の高いところだけわかれば分布の形は解釈できるはずですし、そもそもマイナス無限からプラス無限まで計算することは不可能です。



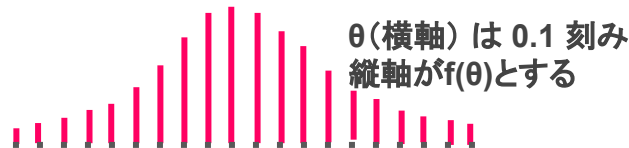
とはいえ、「山の形」が事前にわかるなら計算する必要がありません。そもそも「山の形」が分からないから求めようとしているわけですから。

ということで次からここに「Markov Chain」を組み込んだアルゴリズムを説明します。



# MCMC: MCMC法のアルゴリズム

MCMC法も「 $\theta$  の値を動かしながら分布の形を推定する」という  
コンセプトはMC法と共通しています。  
一旦簡単のために  $\theta$  は 0.1 刻みで動かすことにします。

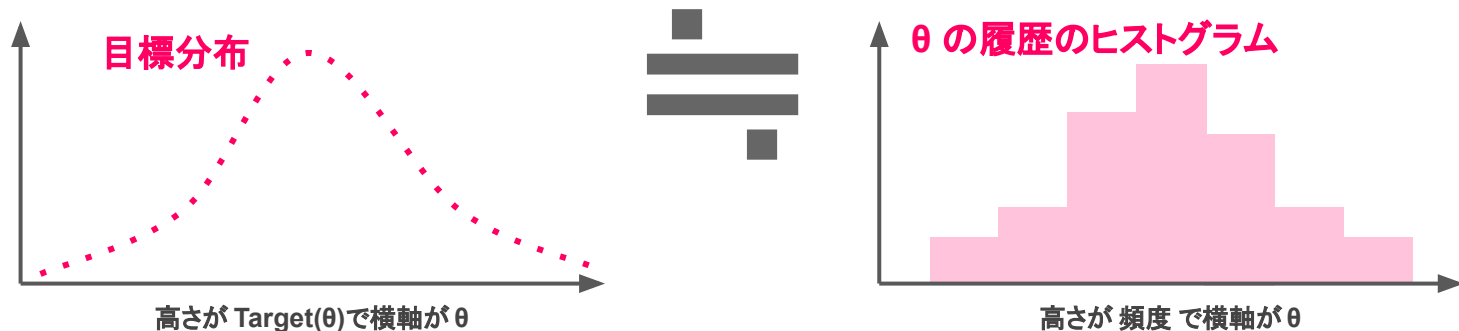


## MCMC法のアルゴリズム

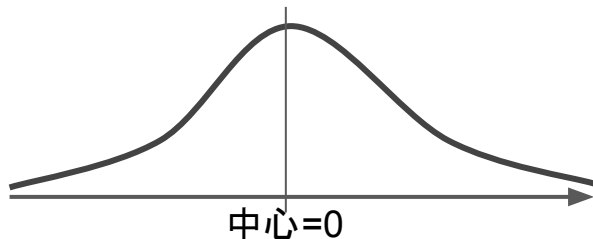
- ①  $\theta$  の初期値をランダムに決めます。  
なお  $\theta$  の右下の数字はステップ番号です。
  - ② 1歩進めましょう。左右どちらに進むかは50%のランダムです。
  - ③  $f(\theta_{(0)}) < f(\theta_{(1)})$  の場合は  $\theta_{(1)}$  に進みます。
  - ④ 次の1歩はまた左右どちらか 50% の確率でランダムに進みます。
  - ⑤ 次のステップで左が候補になったとします。
  - ⑥  $f(\theta_{(1)}) > f(\theta_{(2)})$  の場合は、 $f(\theta_{(2)}) / f(\theta_{(1)})$  の確率で  $\theta_{(2)}$  に進み  $1 - f(\theta_{(2)}) / f(\theta_{(1)})$  の確率で留まります
- ②に戻る

# MCMC: 目標分布と $\theta$ の履歴との関係

前述の「ステップ」を何度も何度も繰り返すと $\theta$ の「履歴」が集まってきます。その $\theta$ の「履歴」が目標分布と近似的に一致することが知られています。そしてこの $\theta$ の「履歴」を事後分布と見なそう。と考えているわけです。



なお今回は簡単のために「ステップ幅を 0.1」とし、「50%の確率で左右どちらかに進む」としていましたが、中心が 0 で左右対称の分布 (例えば正規分布など。これを「提案分布」という) を用いる方法もあります。



この分布からランダムにサンプリングした値を歩幅とします。マイナスの場合は左に進むのと同じ意味になります。また、0付近の値がサンプリングされやすく大きい値はサンプリングされにくくなります。真ん中が 0 なので戻るか進むかは50%ずつ、かつ、歩幅は0 付近が多いけど時々遠くに進む。ということになります。

MCMC法もいくつか種類があるのですが、基本的な考え方は共通しています。