Pythonで学ぶ
ベイズ推論ハンズオン（前編）
【最尤推定からベイズ統計へ】

本日の予定

7/21 前編
機械学習の手順について

最小二乗法の概要，前提条件について

最尤推定
・尤度とは何か
・最尤推定の考え方
・最尤推定と最小二乗法の関係
・最尤推定とMAP推定の違い
・Pythonによる最尤法の実装

ベイズ推論の基礎
・最尤推定の欠点とベイズ推論による克服
・ベイズ推論の大まかな流れ
・Python（PyMC3）によるベイズ推論の実装，とりあえず動かす

お前誰やねん

- 氏名：大久保　亮介
- 現在，薬学部4年（漢方薬専攻）
- 過去の担当講義：基礎統計→ML，高校数学など

「ベイズ〜」は
結局何がしたいのか？

ベイズ推論

ベイズ推定（ベイズすいてい、英: Bayesian inference）とは、ベイズ確率の考え方に基づき、観測事象（観測された事実）から、推定したい事柄（それの起因である原因事象）を、確率的な意味で推論することを指す。

本講義の大まかな流れ

モデルの種類

学習と推論の方法

	二項分布	正規分布	正規線形モデル	…
最小二乗法	-	-	1	
最尤推定	2	3	4	
ベイズ推論	5	6	7	8

機械学習の手順について

機械学習の手順

データをもとに規則を作る（**学習**）

規則から新たなデータの性質を予測する（**推論**）

モデル構築にあたっての３ステップ

<u>STEP 1　モデルの定義</u>

入力データから出力データを得るための式を定義する。

前述の線形回帰の例だと、

・$y = ax + b$

を定義するところに当たる。

モデルの定義においては、説明変数と目的変数をプロットする（散布図）などで、予め特性をつかんでから数式を決めると良い。

<u>STEP 2　誤差関数の定義</u>

モデルを介して得た予測値と実測値の差を誤差として、それをデータセット全体で足し合わせた誤差の関数を定義する。

前述の線形回帰の例にシンプルな二乗誤差を用いて作成すると、

・$E = \Sigma(y\{実測\} - ax\{実測\} - b)^2$

になる。

<u>STEP 3　最適化</u>

誤差関数の微分を元に、モデルのパラメータをアップデートする。

$$w_{i,t+1} = w_{i,t} - \mu \frac{\partial f}{\partial x_i} \ (i \in N)$$

パラメータの更新にあたっては、上記のような勾配降下法の数式がベースになっている。これを理解するにあたって、<u>数列の漸化式</u>と<u>偏微分</u>を理解しておくと良い。

ベイズ推論の流れ

学習

1. データの特徴をつかむ

2. モデルを決める

3. 事後分布を求める

ベイズの定理

$$p(\theta | D) = \frac{p(D | \theta)p(\theta)}{p(D)}$$

推論

4. 予測分布を求める

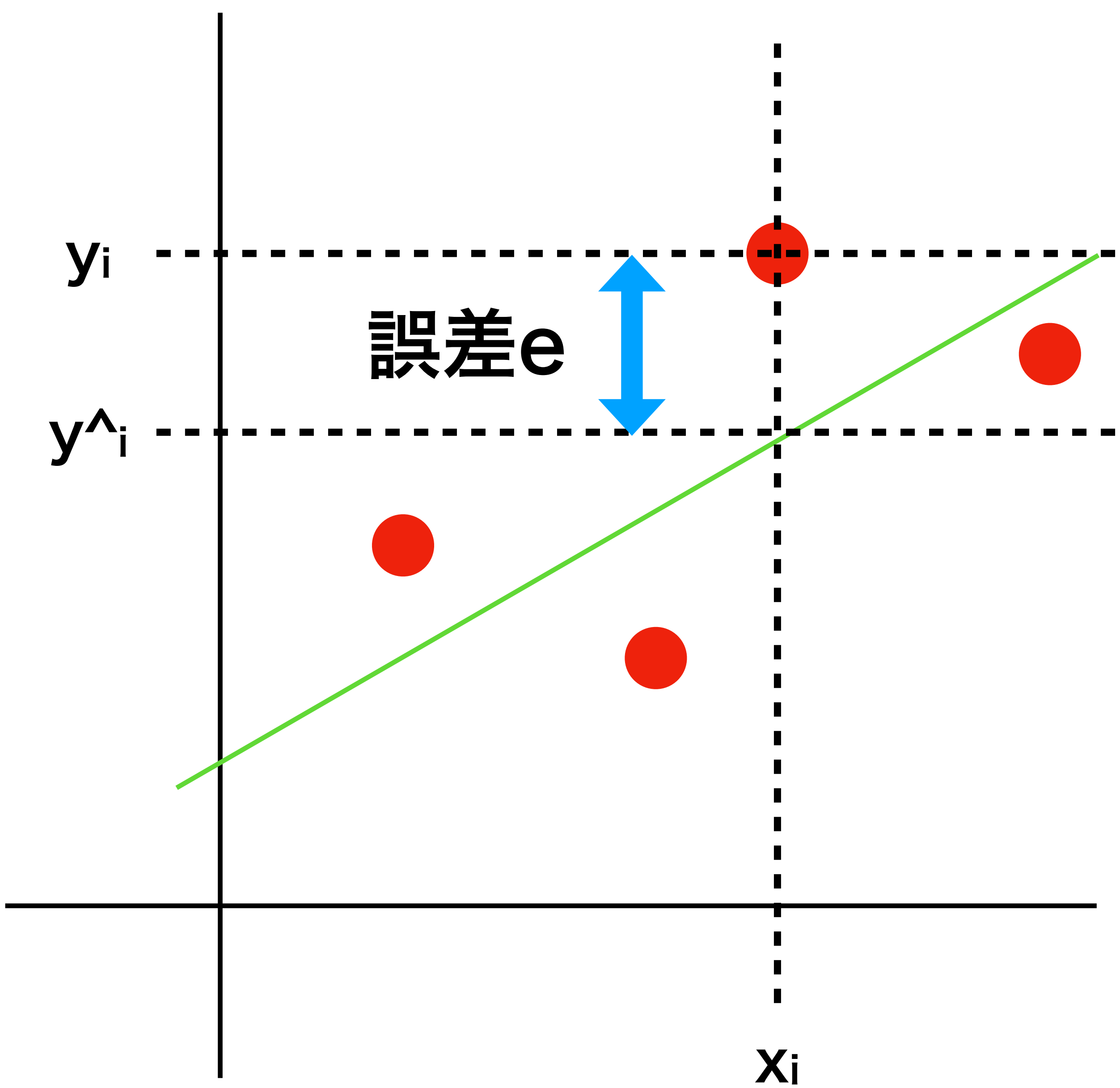
$$p(x_* | D) = \int p(x_* | \theta)p(\theta | D)d\theta$$

（評価）

最小二乗法の概要，前提条件について

最小二乗法とは？

例：単回帰における2乗誤差



$$\hat{y} = \hat{w}x + \hat{b}$$

$$E = \sum (y_i - \hat{y_i})^2$$
$$= \sum (y_i - \hat{w}x_i - \hat{b})^2$$

→Eが最小になる

wとbを推測する

最適解の計算

1. 最小2乗法を用いる，厳密解

$$E = \sum (y_i - \hat{w}x_i - \hat{b})^2$$

$$\Rightarrow \quad \frac{\partial E}{\partial w} = 0 \qquad \frac{\partial E}{\partial b} = 0$$ となるw，bを求める

最適解の計算

2. 勾配降下法を用いる，近似解

$$E = \sum (y_i - \hat{w}x_i - \hat{b})^2$$

$$\rightarrow \quad w_{i,t+1} = w_{i,t} - \mu \frac{\partial f}{\partial x_i} \ (i \in N)$$

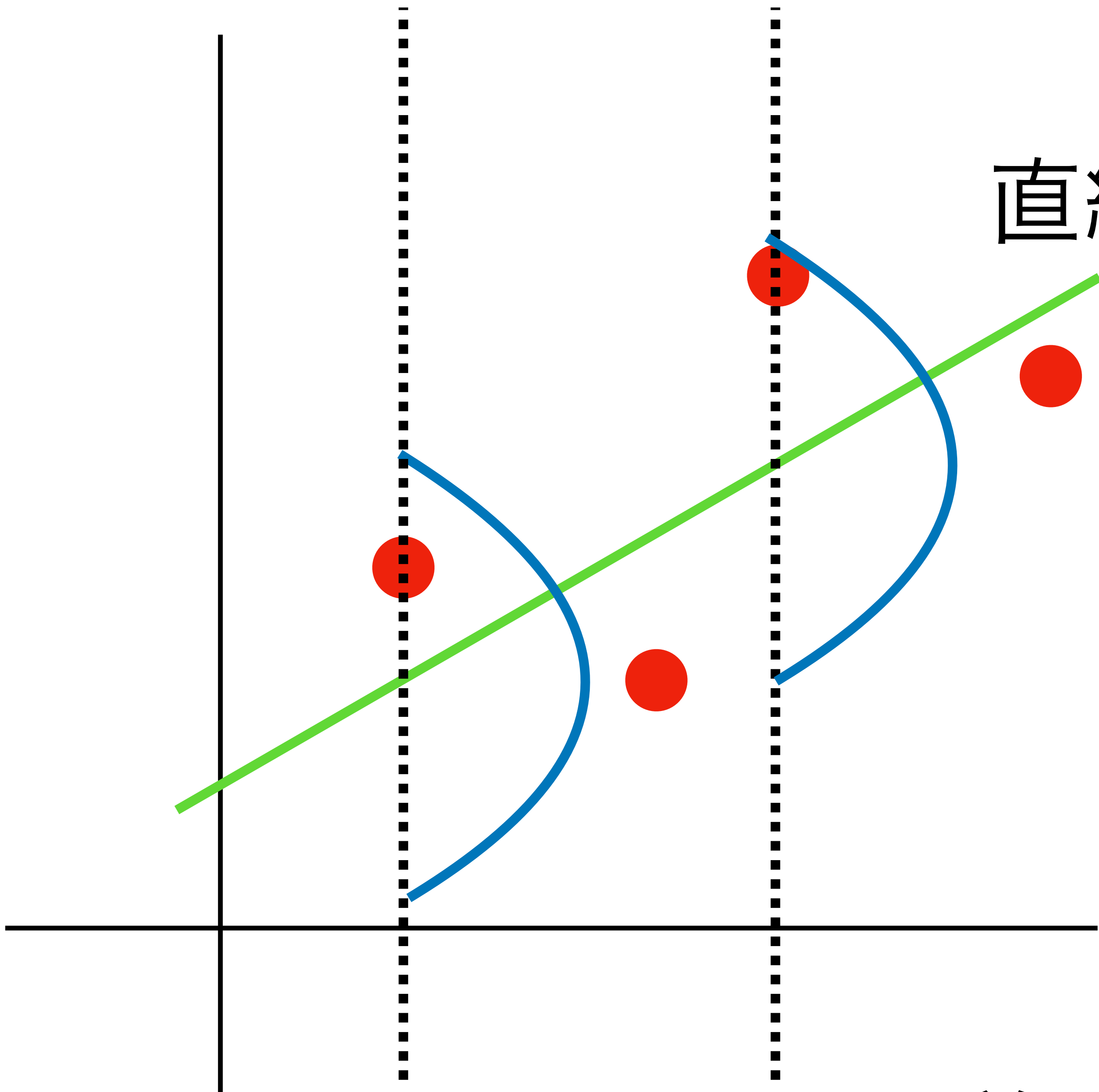
理解しやすい

……が

最小二乗法の前提

1. 系列無相関
2. 分散均一性
3. 説明変数との無相関
4. **正規性**

誤差を正規分布と仮定



$\hat{y} = \hat{w}x + \hat{b}$

直線は**正規分布の平均**を通る

$$N(wx + b, \sigma)$$

→誤差が正規分布じゃなかったら？

最小二乗法のことは
一旦忘れる

最尤推定

最尤推定とは？

最尤推定

出典: フリー百科事典『ウィキペ

最尤推定（さいゆうすいてい、英

最尤法（さいゆうほう、英: met

れたデータからそれが従う確率分

ッシャーが1912年から1922年

基本的理論 [編集]

確率分布関数 f_D と分布の母数 θ のわかっている離散確率分布 D が与えられたとして、そこから n 個の標本 $X_1, X_2, \ldots X_n$ を取り出すことを考えよう。すると分布関数から、観察されたデータが得られる確率を次のように計算することができる：

$$\mathbb{P}(x_1, x_2, \ldots, x_n) = f_D(x_1, \ldots, x_n \mid \theta)$$

しかし、データが分布 D によることはわかっていても、母数 θ の値はわからないかもしれない。どうしたら θ を見積もれるか？ n 個の標本 $X_1, X_2, \ldots X_n$ があれば、この標本から θ の値を見積もることができる。最尤法は母数 θ の一番尤もらしい値を探す（つまり θ のすべての可能な値の中から、観察されたデータセットの尤度を最大にするものを探す）方法である。これは他の推定量を求める方法と対照的である。たとえば θ の不偏推定量は、 θ を過大評価することも過小評価することもないが、必ずしも一番尤もらしい値を与えるとは限らない。 尤度関数を次のように定義する：

$$L(\theta) = f_D(x_1, \ldots, x_n \mid \theta)$$

この関数を母数 θ のすべての可能な値から見て最大になるようにする。そのような値 $\hat{\theta}$ を母数θに対する**最尤推定量**（さいゆうすいていりょう、maximum likelihood estimator、これもMLEと略す）という。最尤推定量は（適当な仮定の下では）しばしば**尤度方程式**（ゆうどほうていしき、likelihood equation）

$$\frac{\partial}{\partial\theta} \log L(\theta) = 0$$

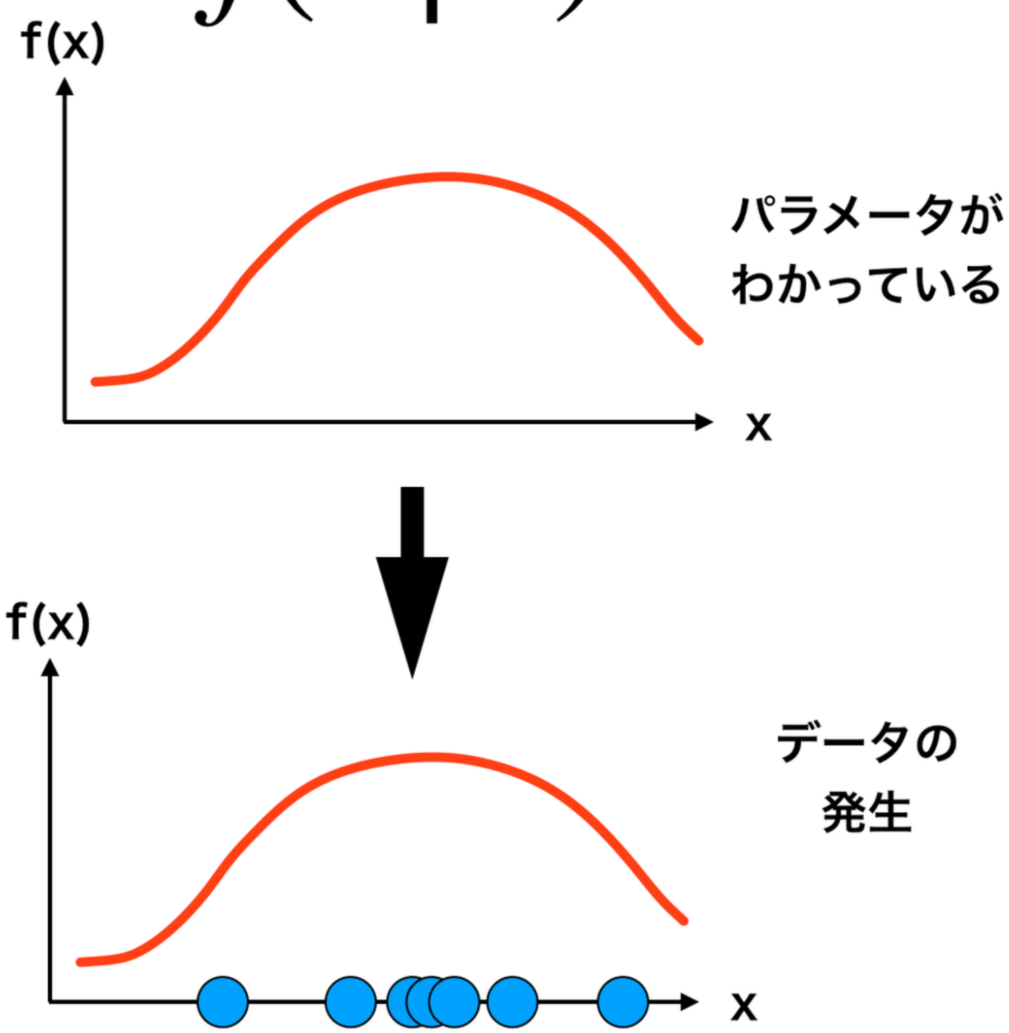
の解として求められる。

？？？

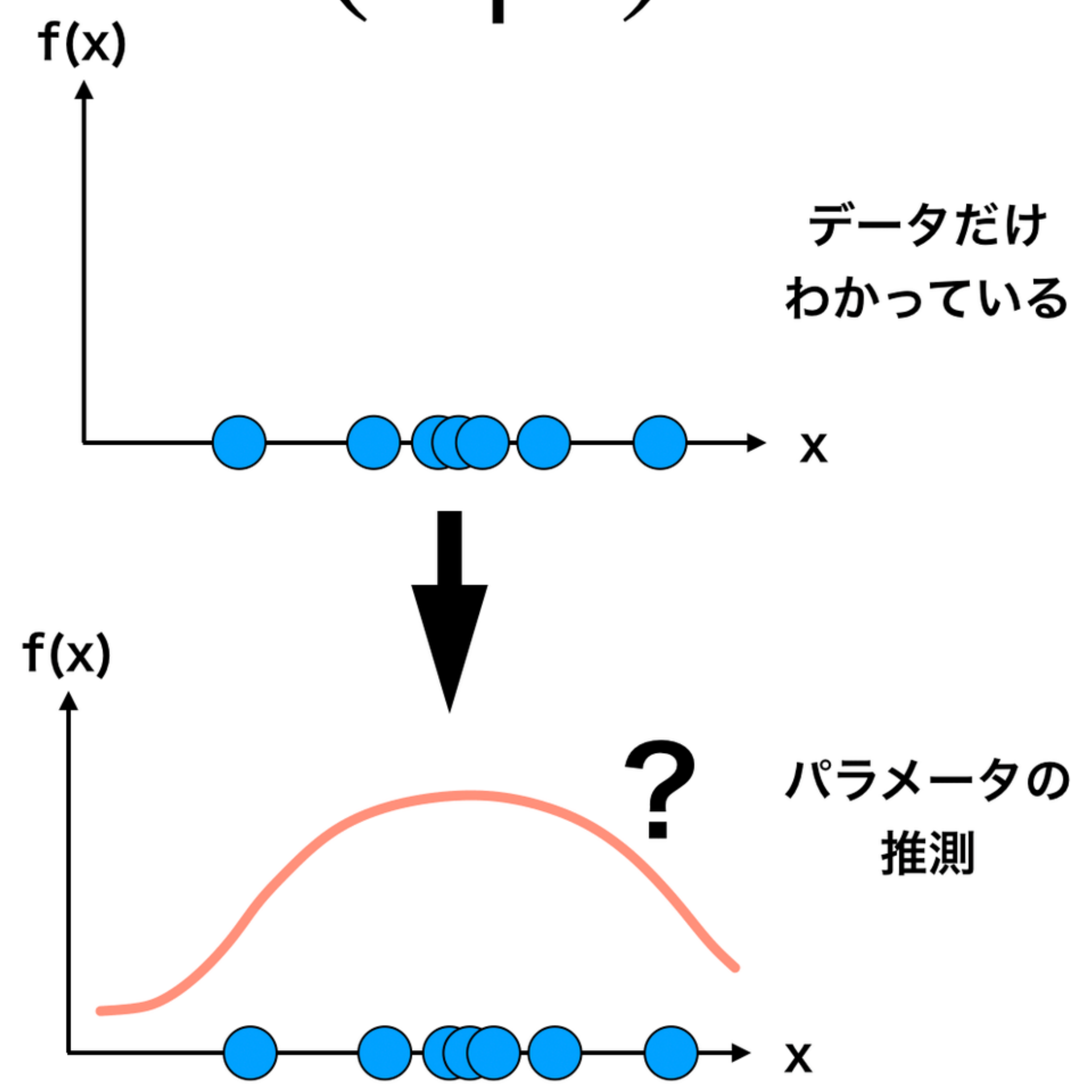
尤度の理解が難しい

確率密度関数と
尤度関数

$$f(x \mid \theta)$$



$$L(\theta \mid x)$$

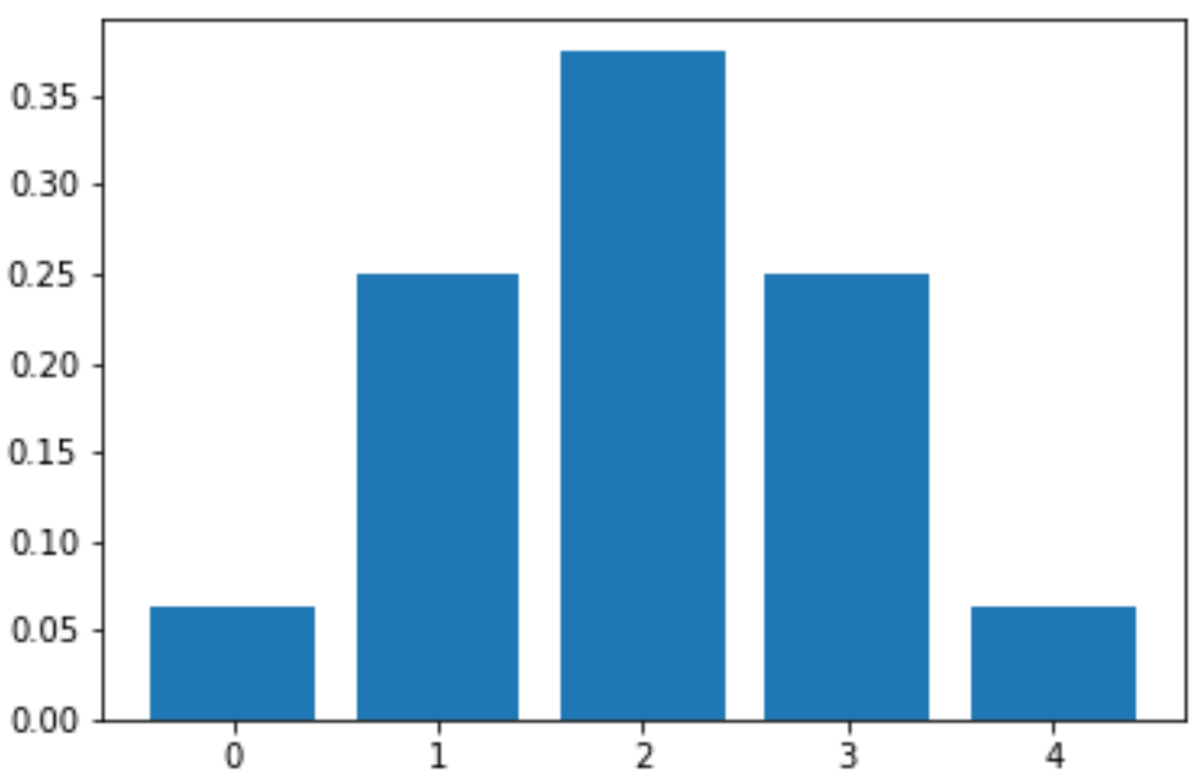


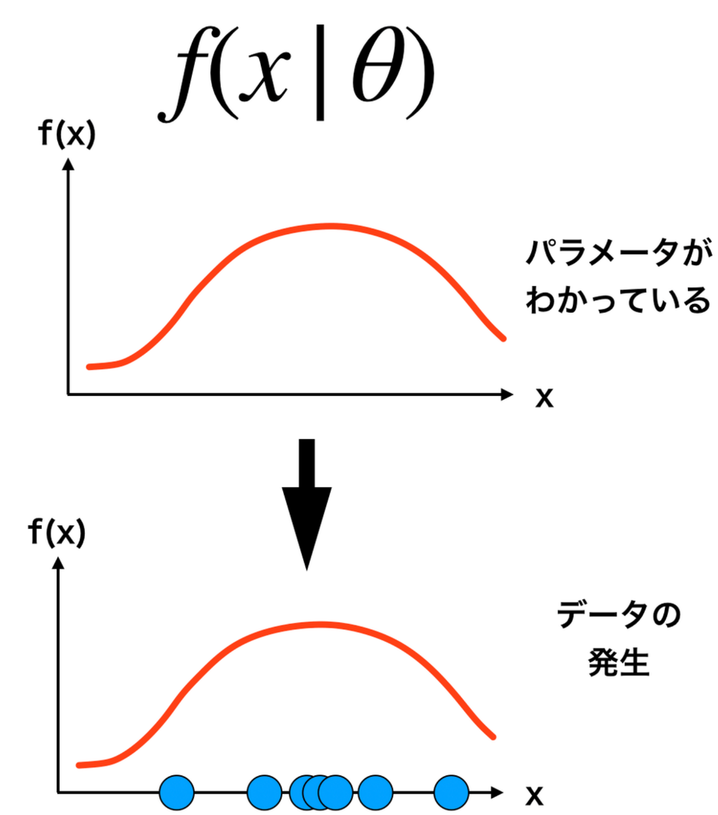
確率密度関数の例1

二項分布（ベルヌーイ分布）

表が出る**確率p=0.5**のコインを4回投げたとき
2回表が出る確率は？

$$P(X = k) = {}_nC_k p^k (1-p)^{n-k}$$





尤度関数の例1

二項分布（ベルヌーイ分布）

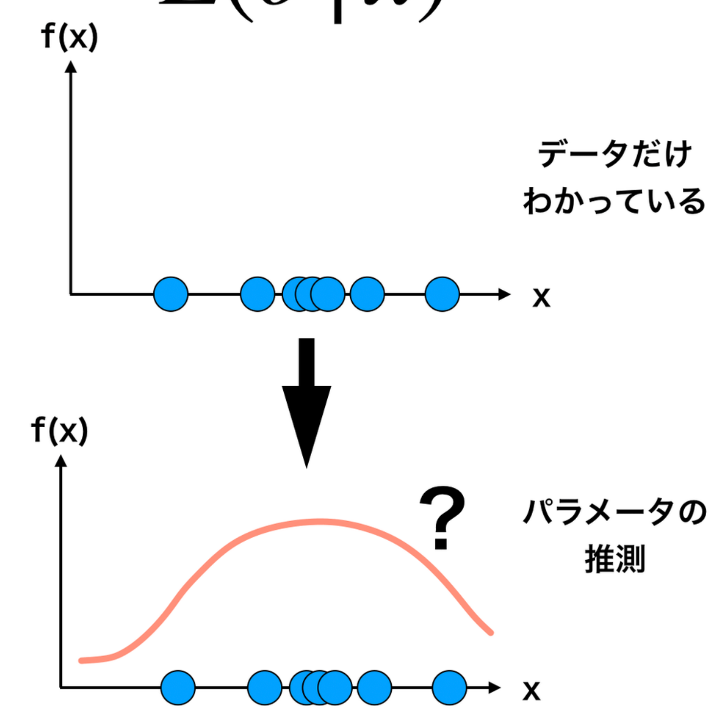
コインを4回投げて2回表が出たとき,

表が出る**確率p**は？

p=0.1よりも

p=0.5のほうがもっともらしい

$$L(\theta\,|\,x)$$



2回表が出たとき，p=0.1よりもp=0.5のほうがもっともらしい

p=0.1よりも

p=0.5のほうがもっともらしい？

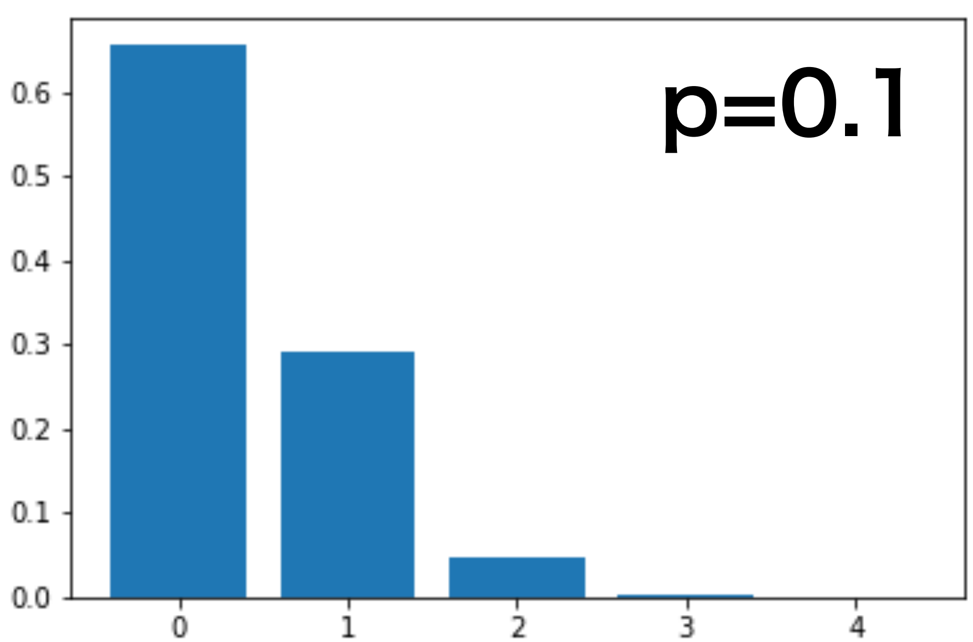
$$P(X=k) = {}_nC_k p^k (1-p)^{n-k}$$

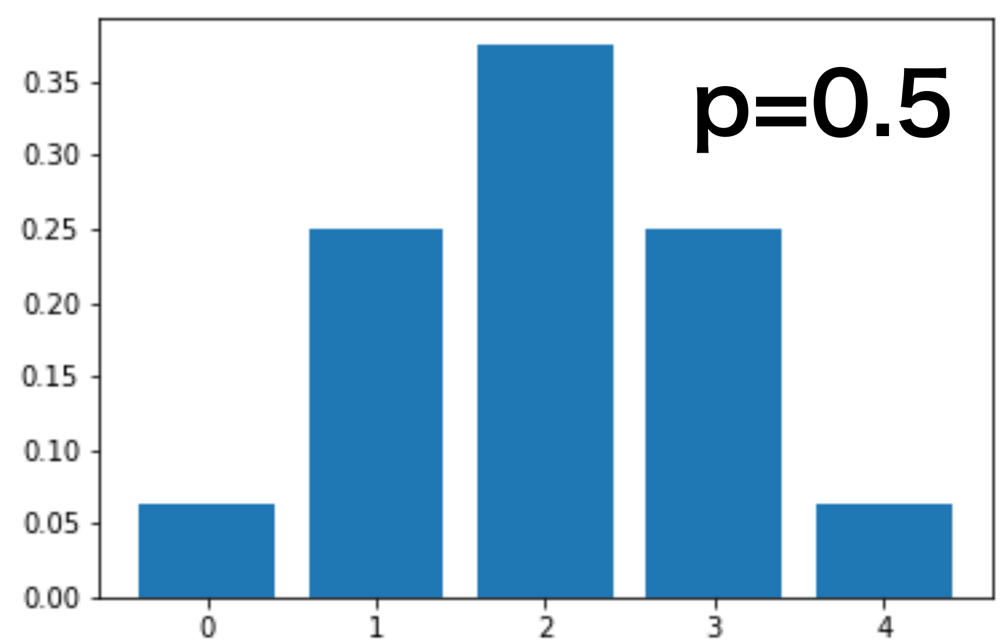
p=0.1のとき，表が2回でる確率

$${}_4C_2(0.1)^2(0.9)^2 = 0.0486$$

p=0.5のとき，表が2回でる確率

$$\boxed{{}_4C_2(0.5)^2(0.5)^2 = 0.375}$$



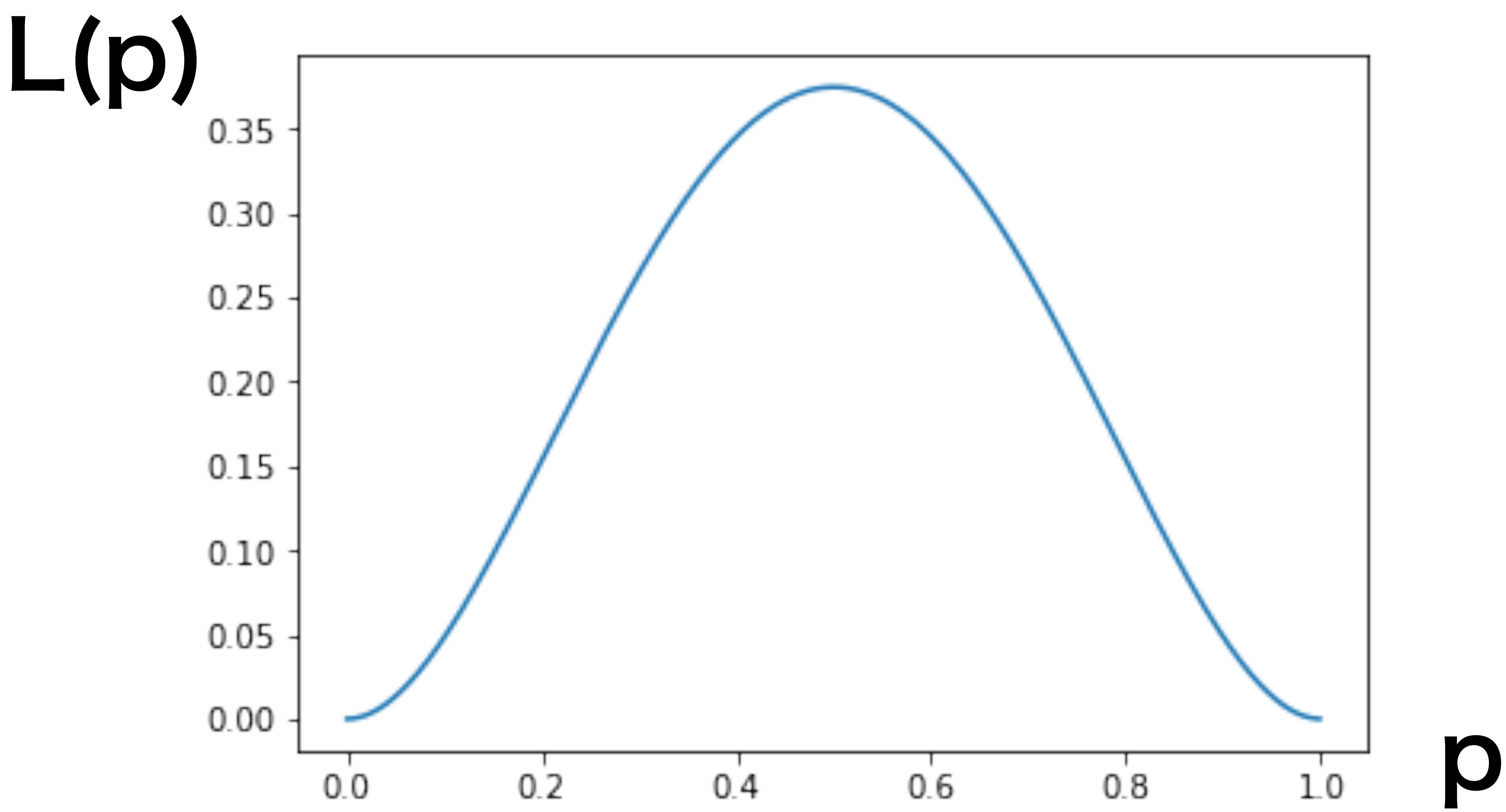


尤度関数の形1

コインを4回投げて2回表が出たとき,

表が出る**確率p**は？

$$L(p \mid n, k) = {}_nC_k p^k (1-p)^{n-k}$$

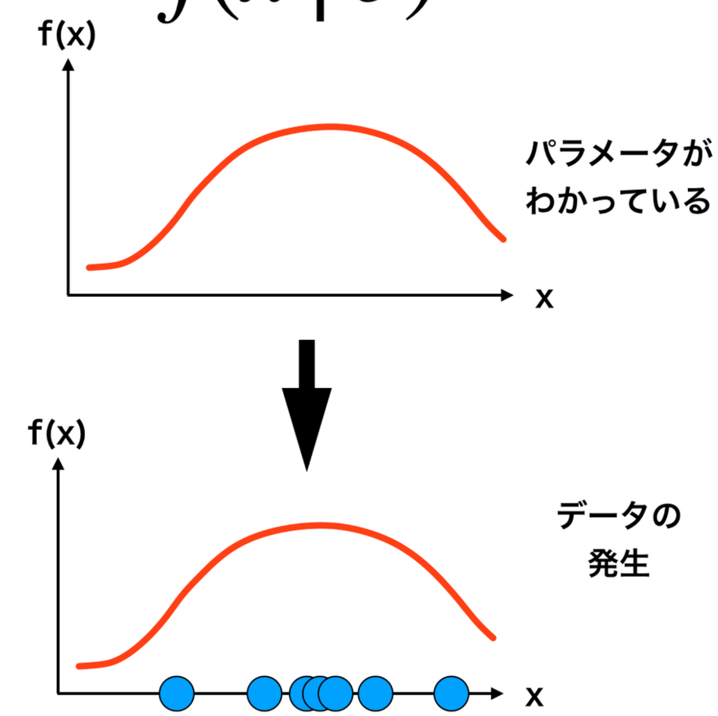


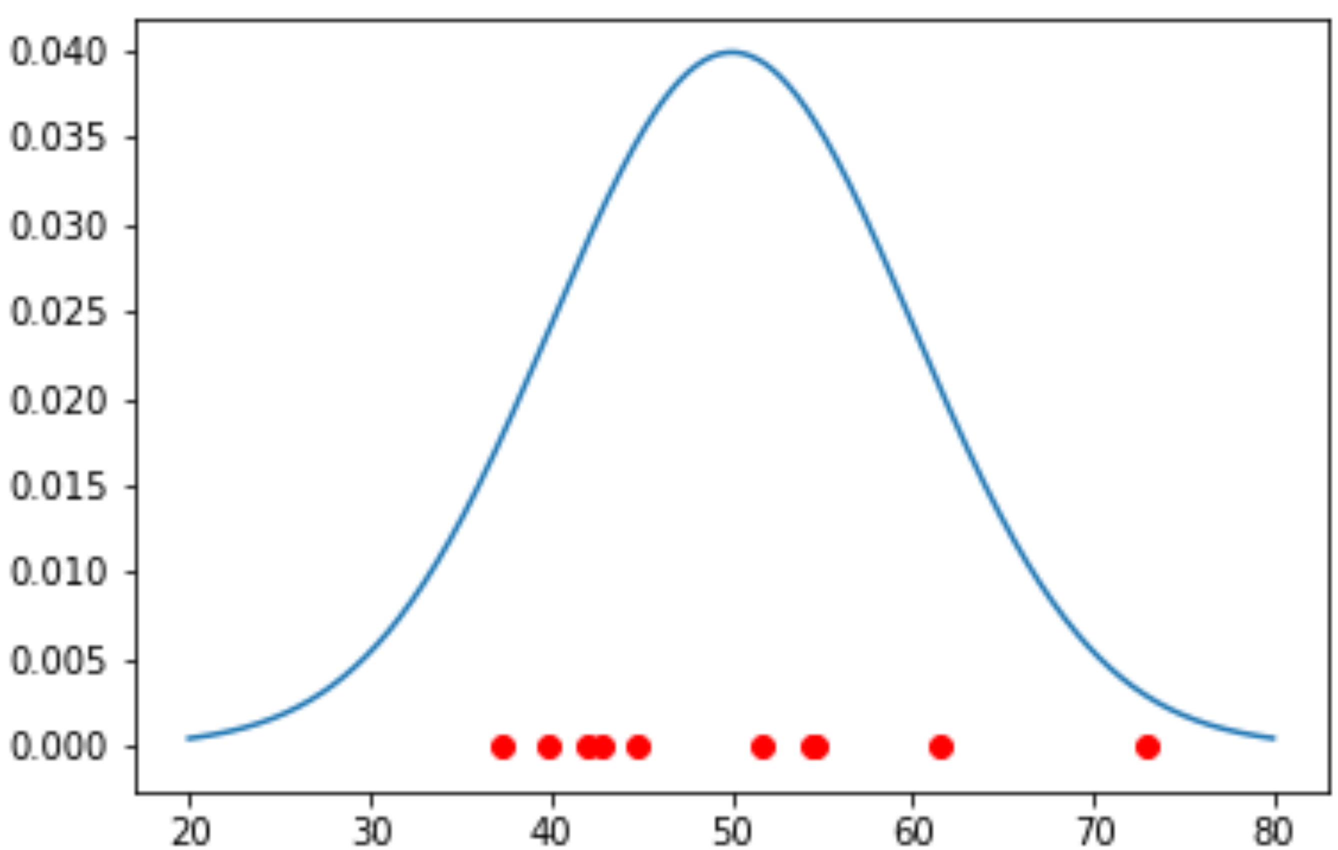
確率密度関数の例2

正規分布

正規分布N(50, 10)にしたがう
データを20個発生させる

$$f(x \mid \theta)$$





尤度関数の例2

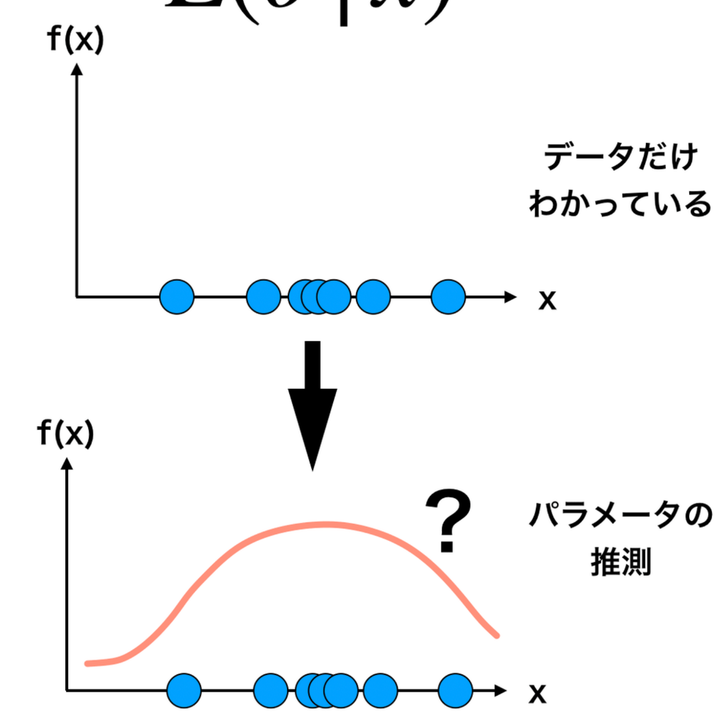
正規分布

正規分布N(μ, σ)にしたがう
データを20個発生させる
→最もよいμ, σは？

μ=30よりも
μ=50のほうがもっともらしい

σ=5よりも
σ=10のほうがもっともらしい

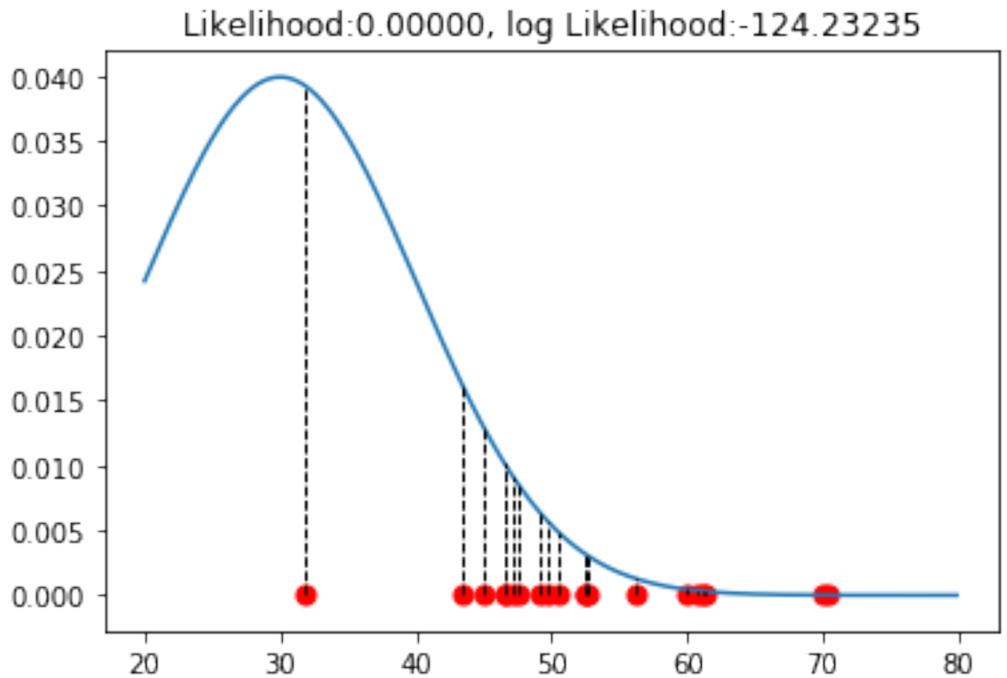
$$L(\theta | x)$$

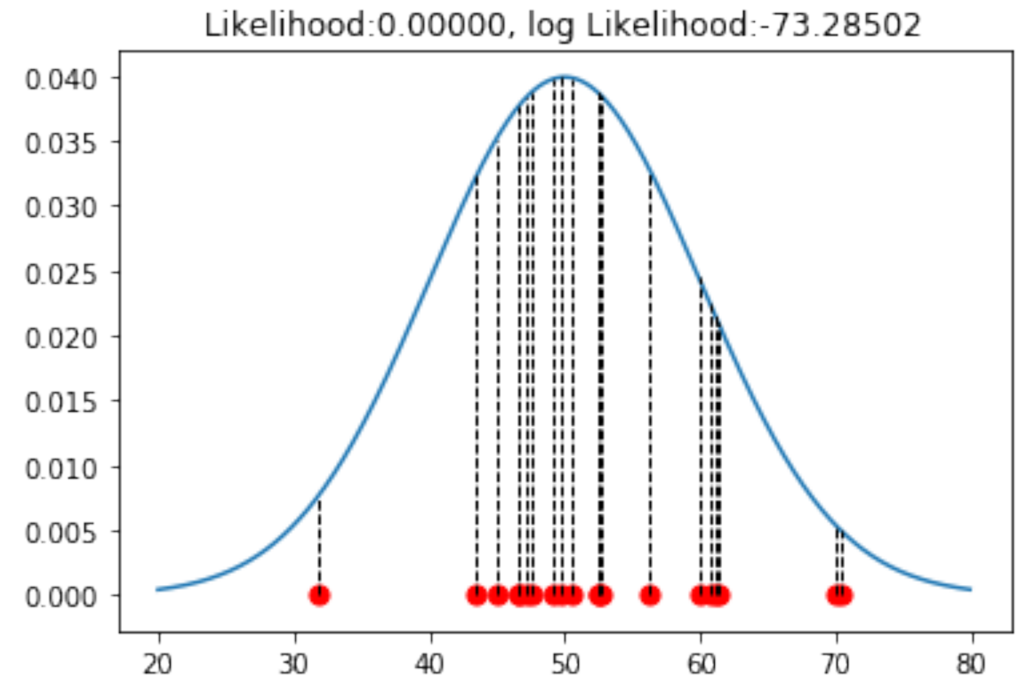


μ=30よりも

μ=50のほうがもっともらしい?

μ=30 μ=50



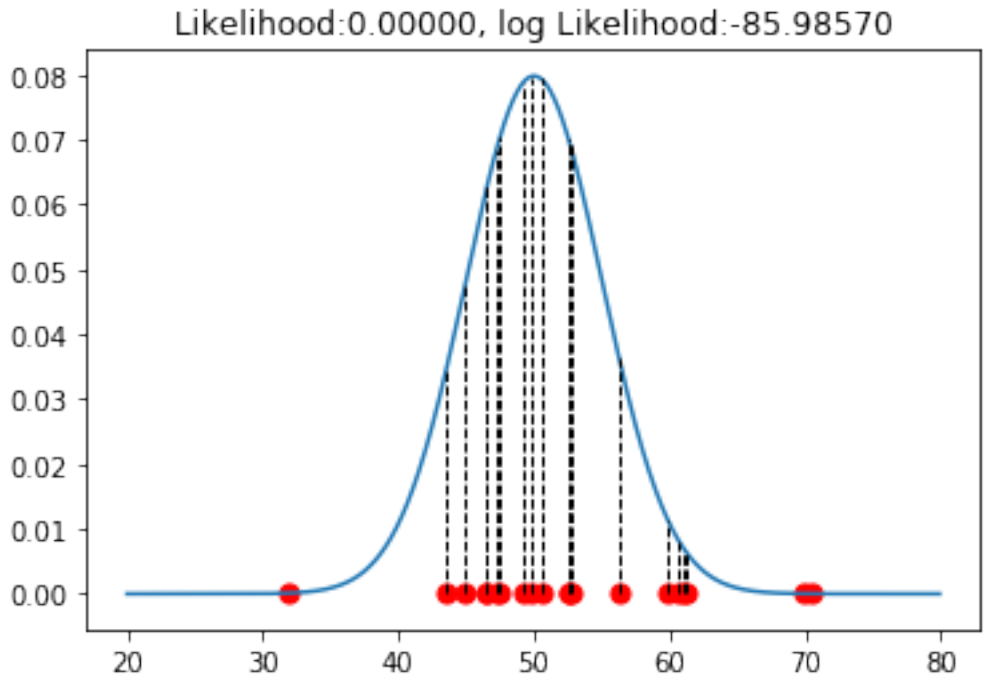


両方ともσ=10

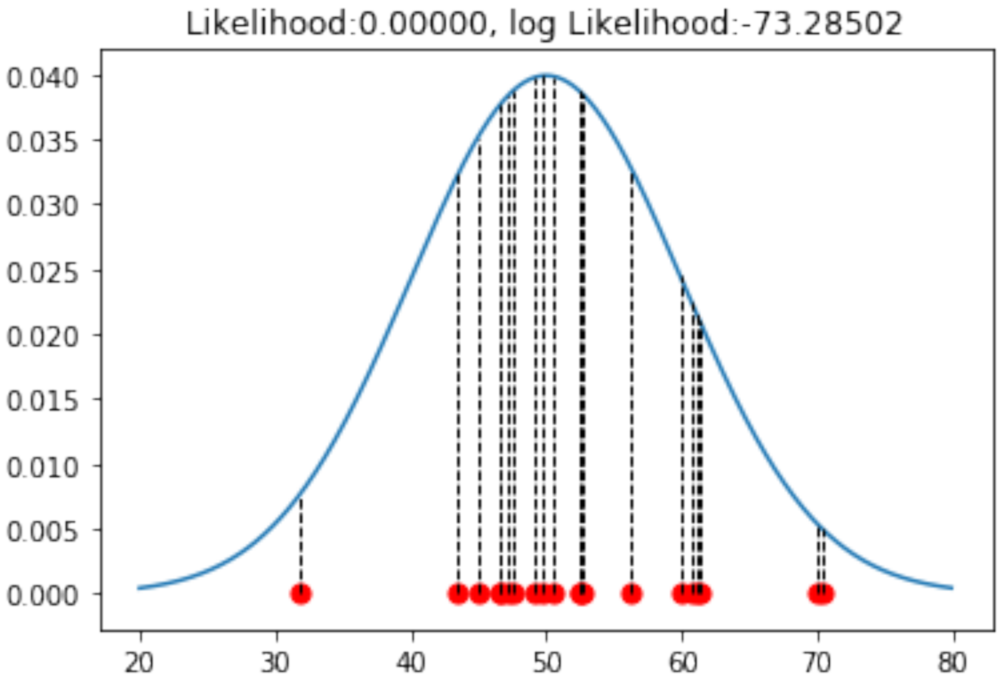
σ=5よりも

σ=10のほうがもっともらしい?

σ=5

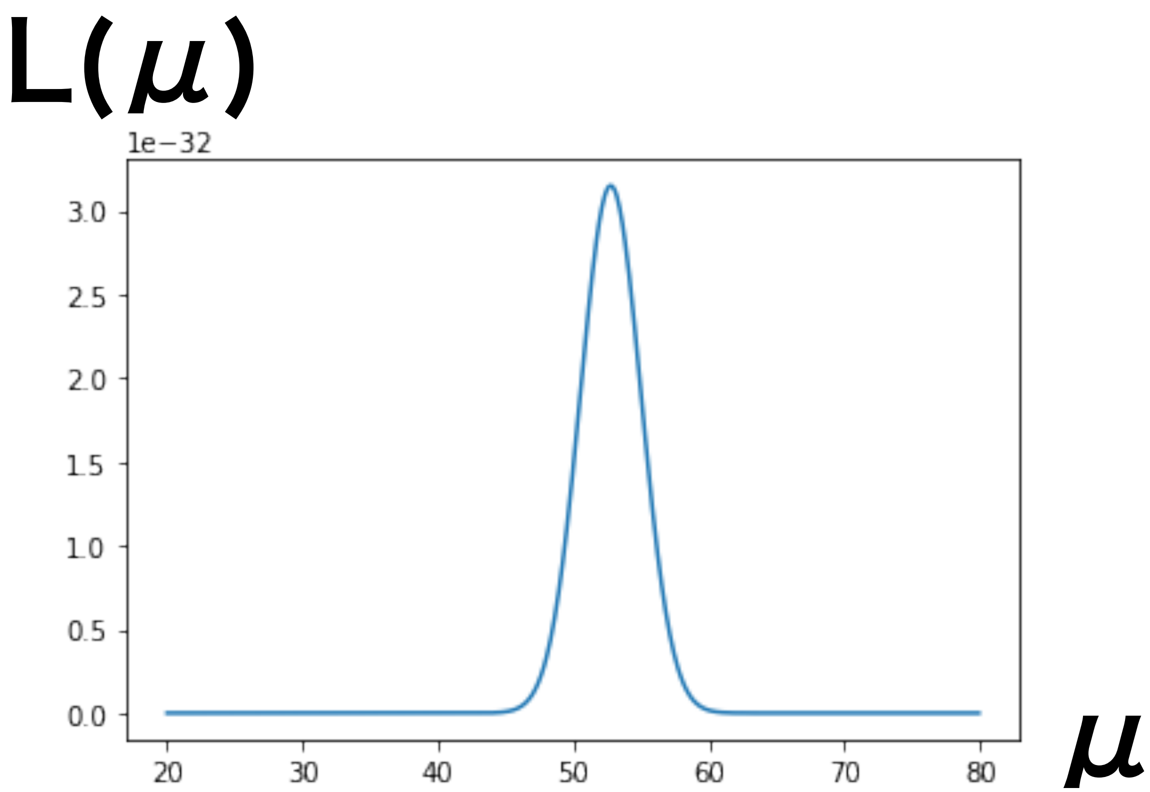


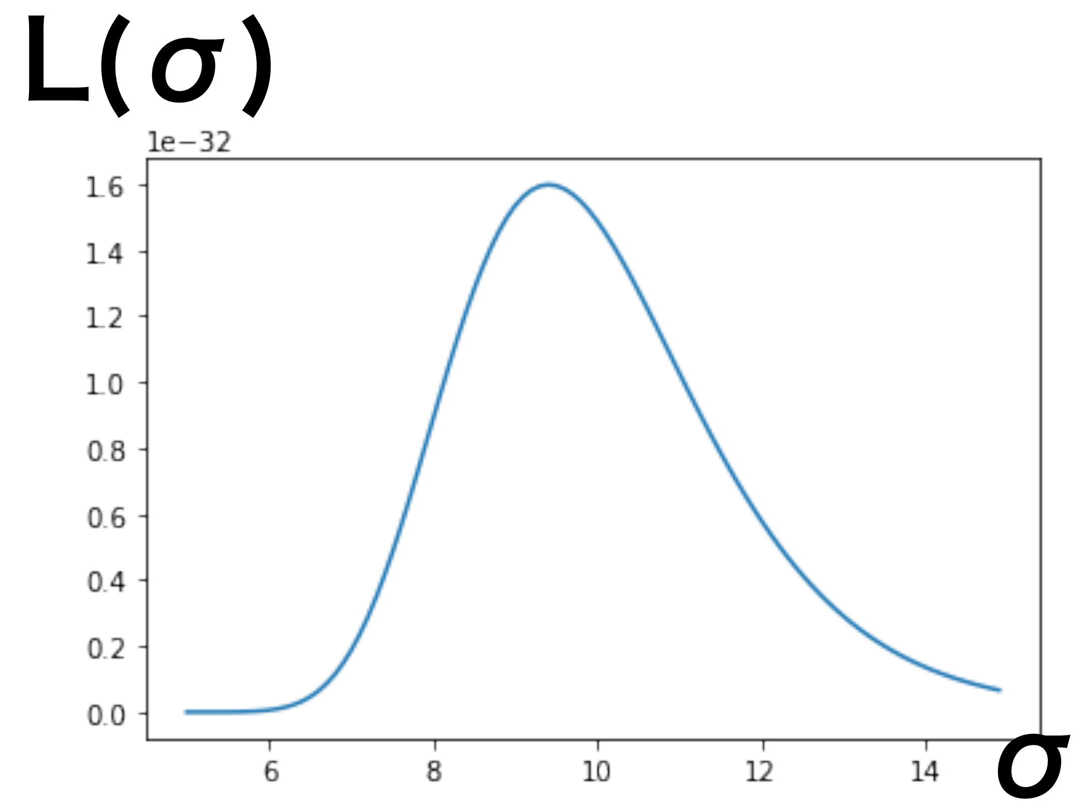
σ=10



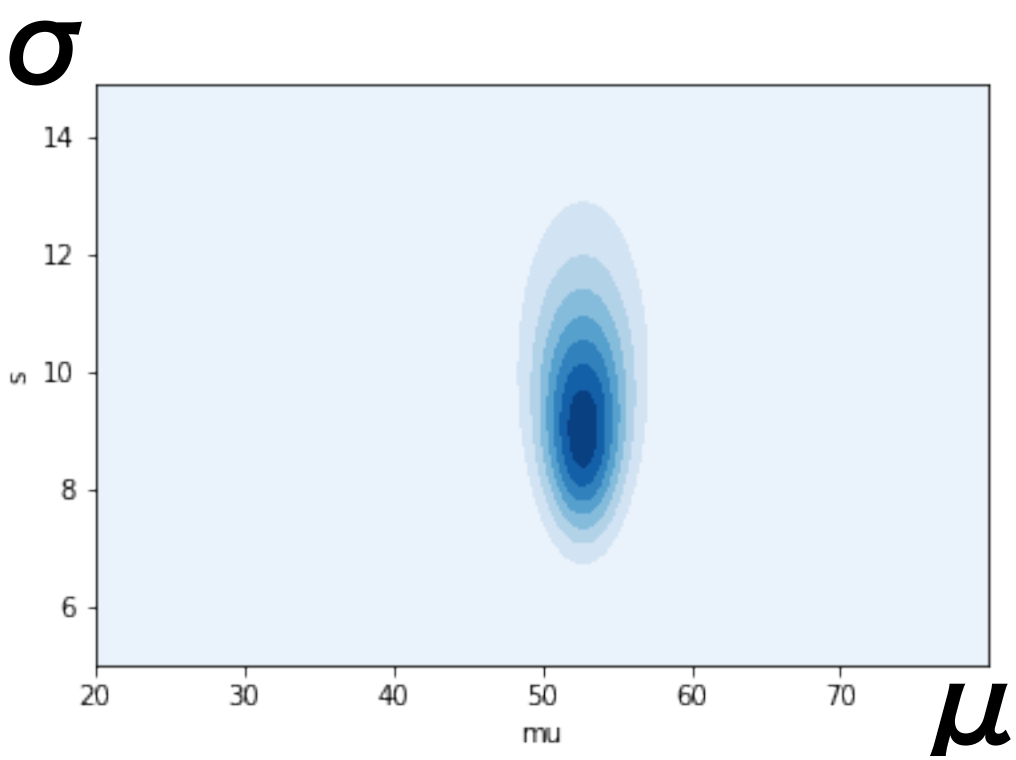
両方ともμ=50

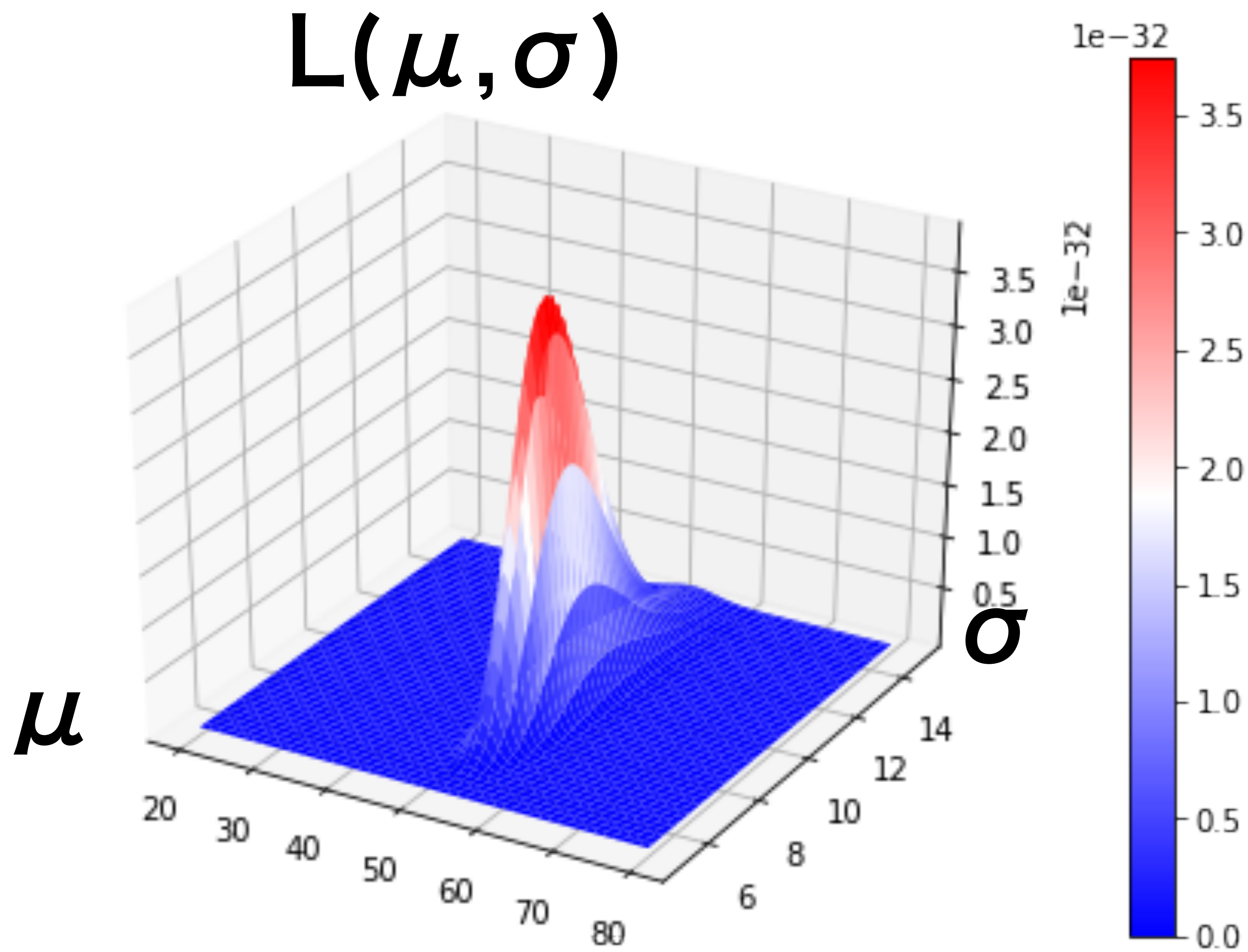
尤度関数の形2


L(μ)
1e−32
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
30
40
50
60
70
80
μ


L(σ)
1e−32
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
6
8
10
12
14
σ

尤度関数の形2


σ
14
12
s 10
8
6
20
30
40
50
60
70
mu
μ


L(μ,σ)
1e−32
3.5
3.0
2.5
2.0
1.5
1.0
0.5
σ
μ
20
30
40
50
60
70
80
6
8
10
12
14
1e−32
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

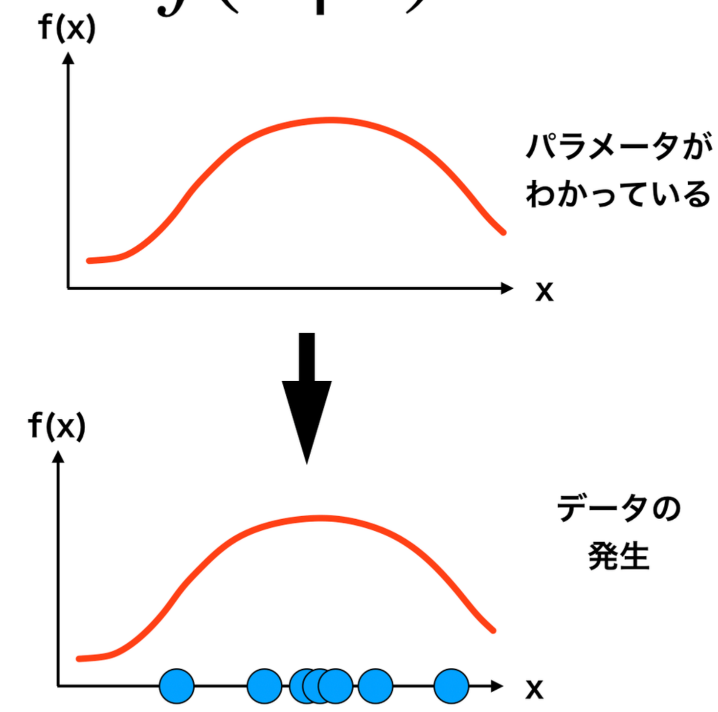
確率密度関数の例3

線形回帰モデル

y = 2x+10, 誤差N(0, 1)にしたがうデータ

正規分布N(2x + 10, 1)にしたがう
データを100個発生させる

$$f(x \mid \theta)$$



尤度関数の例3

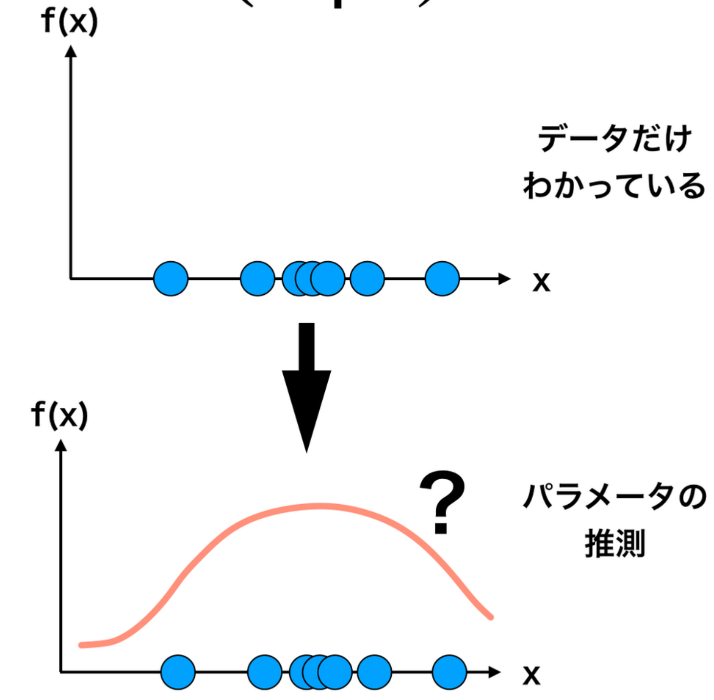
線形回帰モデル

正規分布N(ax + b, σ^2)にしたがう

データを100個発生させる

→最もよいa, b, σ^2は？

$$L(\theta|x)$$



尤度関数の式

データが複数ある場合，尤度関数は

$$L(\theta|x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n|\theta)$$

$$= p(x_1)p(x_2)...p(x_n) = \prod_{i=1}^{N} p_i$$

となる。

尤度は同時確率と等しい

ただし尤度は確率ではない

（確率の定義を思い出す）

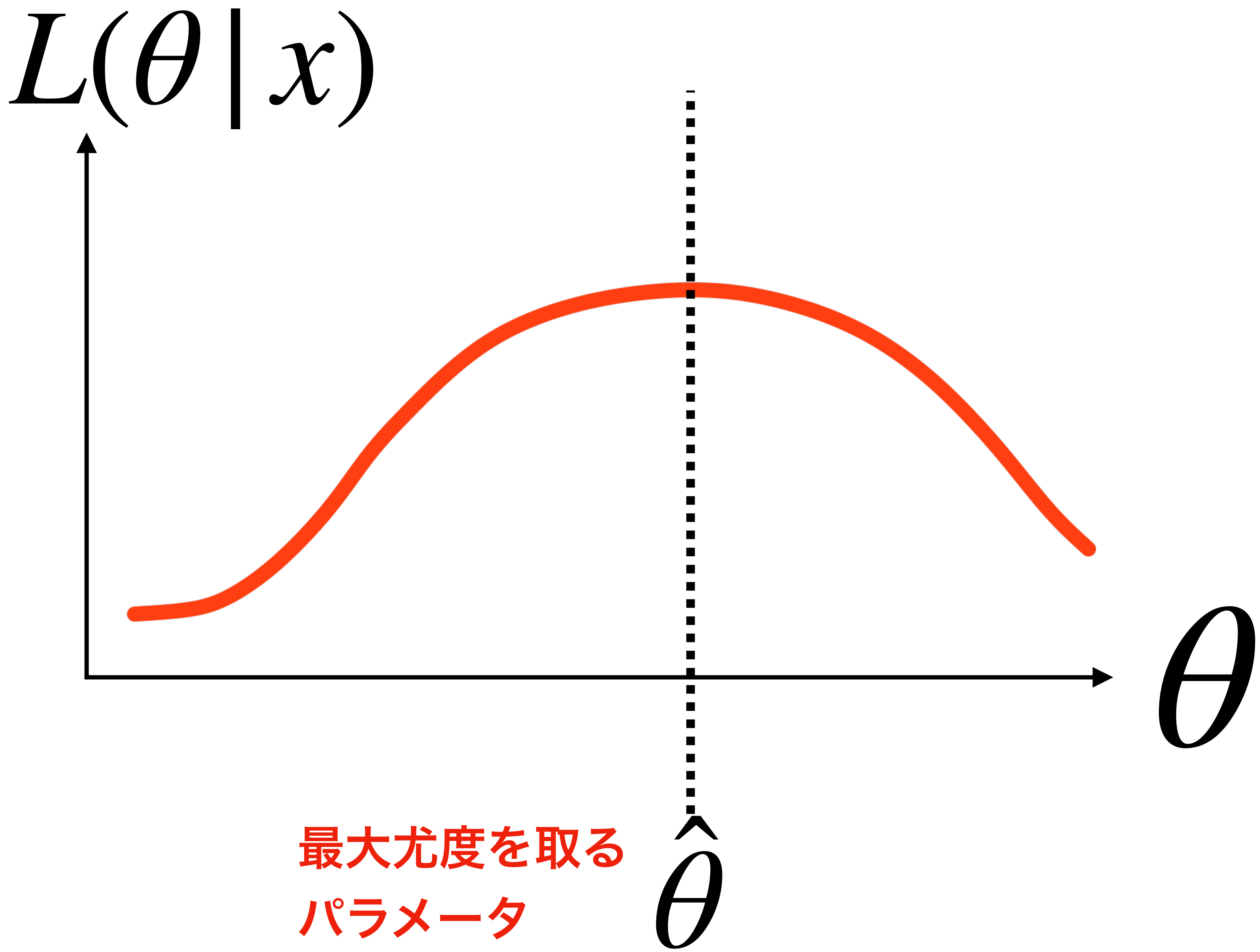
で，

一番いいパラメータは？

最尤推定

尤度関数についての最大化

→微分して0



最尤推定の計算

尤度関数： $L(\theta) = {}_nC_k \theta^k (1-\theta)^{n-k}$

尤度関数を最大にする θ を求める

→対数をとって，それを微分して0とおく

$$\log L(\theta) = \log {}_nC_k + k \log \theta + (n-k) \log(1-\theta)$$

$$\frac{d}{d\theta} \log L(\theta) = \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

$$\frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

のときの θ が
最尤推定量となる

最尤推定から
最小二乗法の導出

最尤推定とは，尤度関数$L(\theta|x)$（以後xを省略）が最大になるパラメータθを求めることである。最大化するにおいて，**対数尤度関数**

$$\ln L(\theta) = \ln \prod p(x_i) = \sum \ln p(x_i)$$

で計算すると楽である。ここで**最尤推定量**$\hat{\theta}$は

$$\frac{\partial}{\partial \theta} \ln L(\theta) = 0$$

となるθである。

よって正規分布では

対数尤度関数：$\ln L(\mu, \sigma^2) = \frac{1}{2\sigma^2} \sum (x_i - \mu) + \frac{N}{2} \ln 2\pi\sigma^2$

それぞれ偏微分して0と置くと，

$$\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2) = -\frac{1}{\sigma^2} \sum (x_i - \mu) = 0$$

$$\frac{\partial}{\partial \sigma^2} \ln L(\mu, \sigma^2) = -\frac{1}{2\sigma^4} \sum (x_i - \mu)^2 + \frac{N}{2\sigma^2} = 0$$

これらの式を連立させて解いたものをそれぞれ$\hat{\mu}$，$\hat{\sigma^2}$と置くと，

$$\hat{\mu} = \frac{1}{N} \sum x_i$$

$$\hat{\sigma^2} = \frac{1}{N} \sum (x_i - \hat{\mu})^2$$

となる。

応用として，回帰分析における最尤推定が重要である。

$y_i = ax_i + b + \epsilon_i$の誤差$\epsilon_i$について

- 各iについて独立
- 正規分布$N(0, \sigma^2)$に従う

と仮定したときの最尤推定量は，**最小二乗法**における正規方程式の解と一致する。

$\rightarrow$

最尤推定では
誤差が正規分布じゃなくても
考えることができる

MAP推定とは？

最尤推定

$$\theta_{ML} = \arg \max_{\theta} p(D \mid \theta)$$

最大事後確率（maximum a posteriori, **MAP**）**推定**

$$\theta_{MAP} = \arg \max_{\theta} p(D \mid \theta) p(\theta)$$

注意：
最尤推定と同じく
点推定

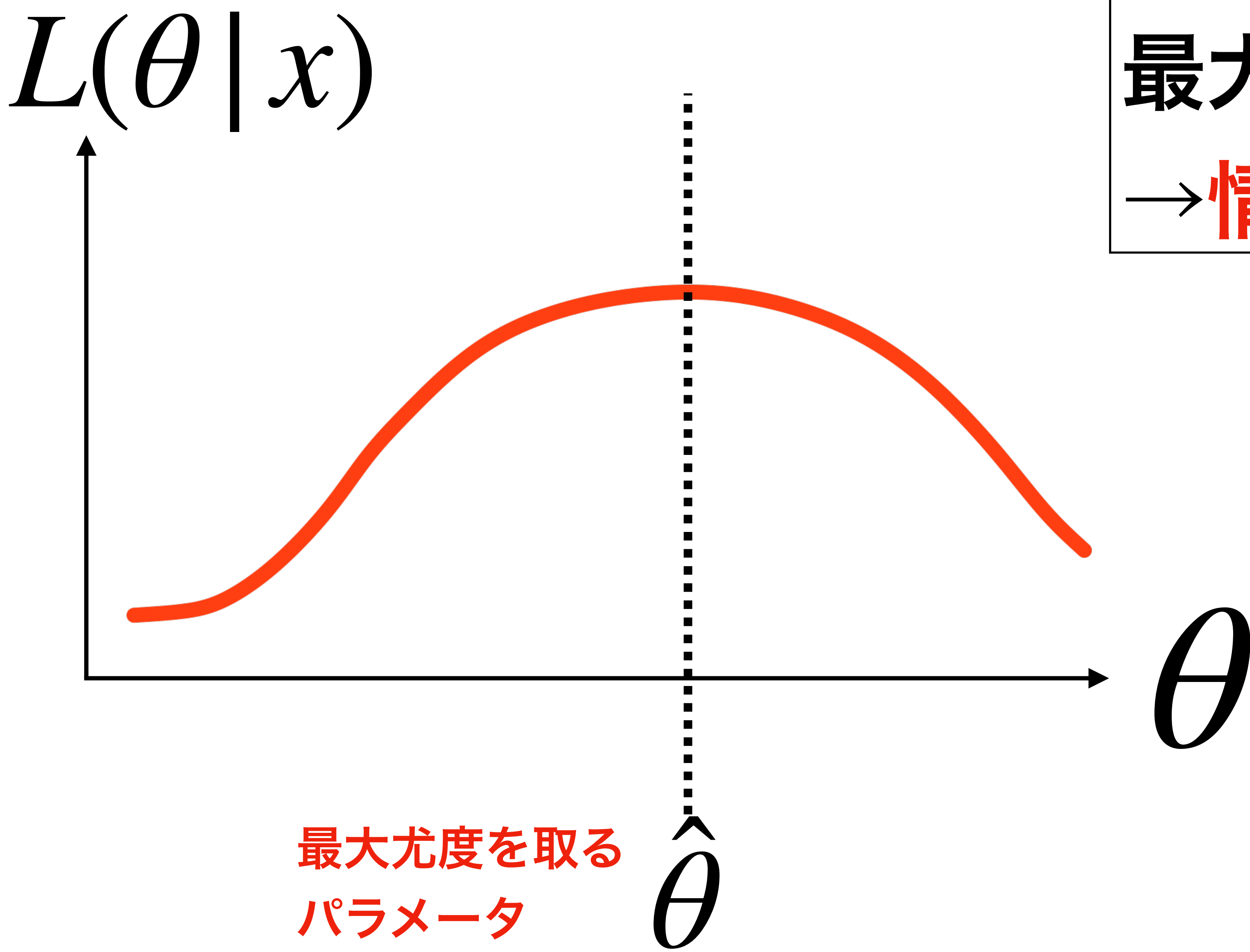
事前分布

→いわゆる正則化として
はたらく

Pythonによる
最尤推定の実装

ベイズ推論の基礎

最尤推定の欠点



尤度関数のうち
最大値しか使っていない
→**情報が失われている**

特定の値（点）で
出力
→**過学習のおそれ**
（予測能はない）

ベイズ推論による克服

ベイズの定理

尤度　事前分布

事後分布

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

周辺分布

尤度関数を
余すところなく使える

確率分布（区間）で
出力
→予測能あり

$$p(x_* \mid D) = \int p(x_* \mid \theta)p(\theta \mid D)d\theta$$

未知の
データ

詳しい話は

赤池 弘次「エントロピーとモデルの尤度」

https://www.jstage.jst.go.jp/article/butsuri1946/35/7/35_7_608/_article/-char/ja/

前編では
ほんの流れだけ

ベイズ推論の流れ（再掲）

学習

1. データの特徴をつかむ

2. モデルを決める

3. 事後分布を求める

ベイズの定理

$$p(\theta | D) = \frac{p(D | \theta)p(\theta)}{p(D)}$$

推論

4. 予測分布を求める

$$p(x_* | D) = \int p(x_* | \theta)p(\theta | D)d\theta$$

1. データの特徴をつかむ

例：コインを投げて，5回中2回表が出た
→表が出る確率は？

表1，裏0とすると
0,1,0,0,1
と表せる

2. モデルを決める

ベルヌーイ分布（二項分布）
でモデリングする

尤度

$$L(\theta \mid x) = p(x \mid \theta) = Bern(x \mid \theta)$$

事前分布

$$p(\theta) = Beta(\theta \mid a, b)$$

ベータ分布は
二項分布に対して
共役である

3. 事後分布を求める

条件付き確率より

$$p(D, \theta) = p(D \mid \theta)p(\theta)$$

ベイズの定理

尤度　**事前分布**

事後分布

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

周辺分布

→勝手に計算してくれる

4. 予測分布を求める

$$p(x_* | D) = \int p(x_* | \theta) p(\theta | D) d\theta$$

未知の
データ

→さまざまなθについて
モデルの平均

Python（PyMC3）による
ベイズ推論の実装，
とりあえず動かす