確率統計（４）

概要

1. 統計学の基礎
   - 1次元のデータ
   - 2次元のデータ
   - 相関
2. 確率の基礎
   - 確率変数と確率分布
   - 中心極限定理
   - 独立性
3. 仮説検定
   - t検定
   - カイ二乗検定
4. 推定
   - 点推定・区間推定
   - 最尤推定
   - ベイズ推定

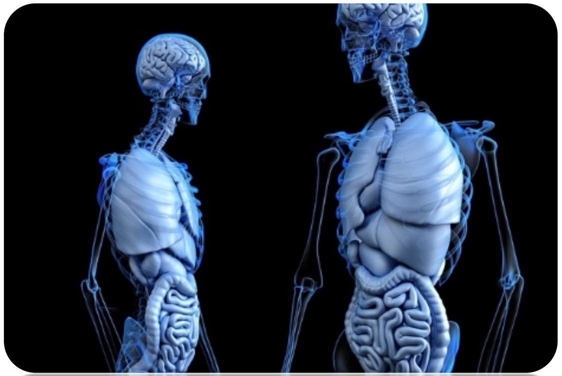
補助資料：http://small-island.work/trial/
ユーザ名：trial
パスワード：trial

推定

世の中の確率的な事象の多くには、実際には観測できない事象や未知のパラメータが多く存在する。統計的な推定の手法は、これらの未知のパラメータ等をサンプルから推し量る方法論である

医療データなどのデータ

患者の様態



自然科学・社会科学

直接測定・観測できない事象

センシング

雑音下での目的とする信号

推定量

母集団のパラメータθ を知りたいときに、それを観測データをもとに推し量る方法

推定の種類

- 点推定

  例：平均μが$\hat{\mu}$であると推定する

- 区間推定

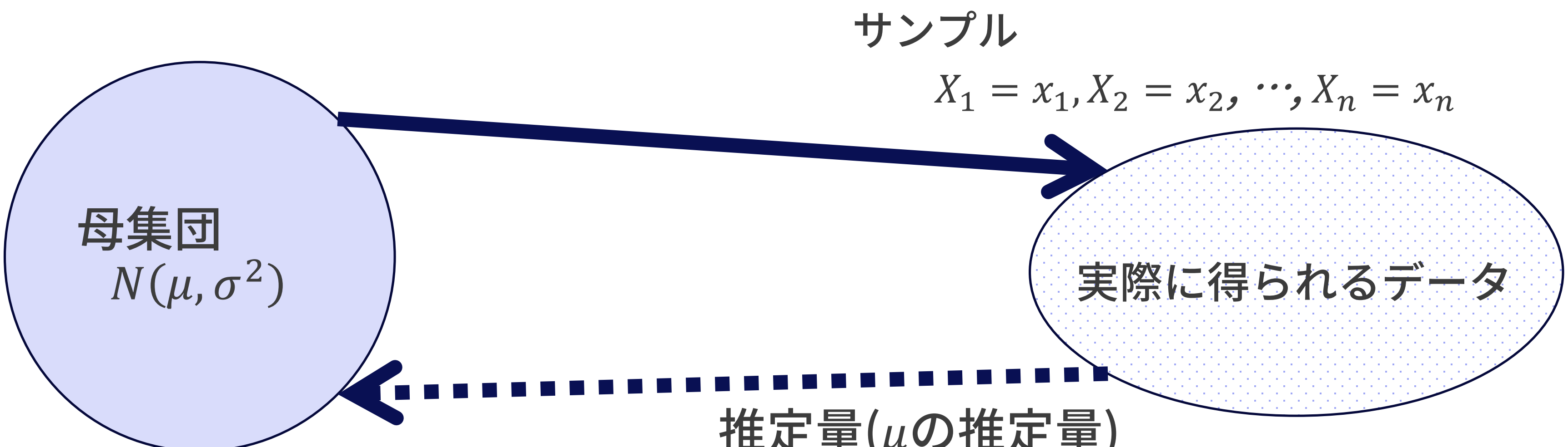
  例：平均μが$\hat{\mu}_{low} \sim \hat{\mu}_{high}$であると推定する

- ベイズ推定

  例：平均μの事後分布が$p(\mu \mid \ldots)$の分布に従うと推定する

推定では観測できないものを推定するという性質上、いくつかの基準が存在し、異なる推定結果になる場合もある。（場合によっては使い分けることも必要）

点推定とサンプルの関係



推定量(μの推定量)

- $\hat{\mu} = \frac{X_1+X_2+ \qquad + X_n}{n}$

推定量(σ^2の推定量)

- $\widehat{\sigma^2} = \frac{(X_1-\hat{\mu})^2+(X_2-\hat{\mu})^2+\cdots+(X_n-\hat{\mu})^2}{n-1}$

点推定をする方法：

- モーメントをパラメータにする推定量を利用する
- 最小二乗推定量：観測との最小二乗誤差を最小にする
- 観測に関する確率を最大化する（最尤推定）

一致性

サンプル数を増やしたときに、推定量が母集団の分布のパラメータに近づくという性質を持っていること

任意の $\varepsilon > 0$ に対して

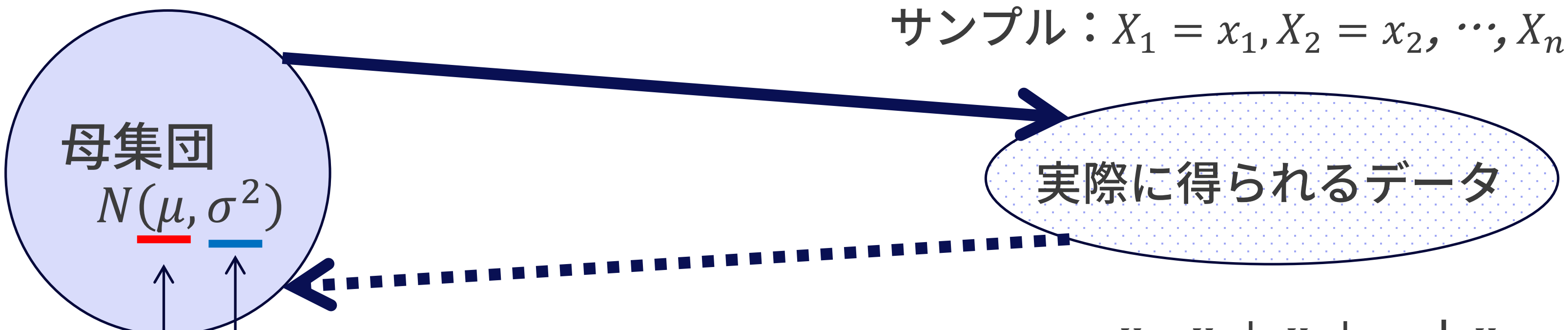
$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| < \varepsilon\right) = 1$$

（確率収束）

サンプル数 n の場合の推定量を $\hat{\theta}_n$

真の値をθとして

サンプル：$X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$



推定量(μの推定量)：$\hat{\mu} = \dfrac{X_1+X_2+X_3+\quad+X_n}{n}$

推定量(σ^2の推定量)$\widehat{\sigma^2} = \dfrac{(X_1-\hat{\mu})^2+(X_2-\hat{\mu})^2+\cdots+(X_n-\hat{\mu})^2}{n-1}$

サンプル数を増やせばほぼ等しい

- 世の中で見かける推定量の多くはこの性質を持つが、適当に決めた推定量などはこの性質を満たすかの確認が必要である。

不偏推定量

平均的に正しくなるような推定

サンプル数 n の場合の推定量を $\hat{\theta}_n$

真の値をθとして

$$E[\hat{\theta}] = \theta$$

不偏推定量では平均しか見ていないが、同じ不偏推定量であるならばより分散が小さい方が効率的であるという（効率性）

サンプル：$X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$

母集団

$N(\mu, \sigma^2)$

実際に得られるデータ

推定量(μの推定量)：$\hat{\mu} = \dfrac{X_1+X_2+X_3+\quad+X_n}{n}$

推定量(σ^2の推定量)$\widehat{\sigma^2} = \dfrac{(X_1-\hat{\mu})^2+(X_2-\hat{\mu})^2+\cdots+(X_n-\hat{\mu})^2}{n-1}$

$E[\hat{\mu}]$

$E\left[\widehat{\sigma^2}\right]$

平均が等しい

標本分散 vs 不偏分散

標本分散

$$\widehat{s^2} = \frac{(X_1-\hat{\mu})^2+(X_2-\hat{\mu})^2+\cdots+(X_n-\hat{\mu})^2}{n}$$ は不変推定量ではない

$$E\left[\widehat{s^2}\right] = E\left[\frac{(X_1-\hat{\mu})^2+(X_2-\hat{\mu})^2+\cdots+(X_n-\hat{\mu})^2}{n}\right]$$
$$= \frac{1}{n}E\left[\sum_{i=1}^{n}(X_i-\mu)^2+2(X_i-\mu)(\mu-\hat{\mu})+(\mu-\hat{\mu})^2\right]$$
$$= \sigma^2 + 2E[(\hat{\mu}-\mu)(\mu-\hat{\mu})] + E\left[(\mu-\hat{\mu})^2\right]$$
$$= \sigma^2 - E\left[(\mu-\hat{\mu})^2\right]$$

$$\sigma^2 = E[\,(X-\mu)^2]$$

$\widehat{s^2}$ は常に母集団の分散σ^2より少し小さい値になる

$$E\left[\widehat{s^2}\right] = \frac{n-1}{n}\sigma^2$$

正規分布の基本的な性質

平均μ、分散σ^2の正規分布からn　のサンプルを無作為に抽出する時、サンプル平均$\bar{X}$は、$N(\mu, \frac{\sigma^2}{n})$に従う

→$E[(\mu-\hat{\mu})^2] = \sigma^2/n$

不偏分散

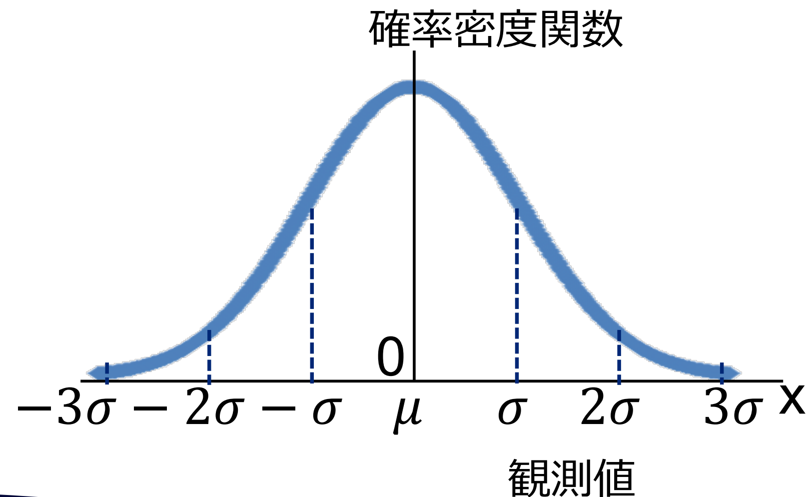
$$\hat{\sigma}^2 = \frac{(X_1-\hat{\mu})^2+(X_2-\hat{\mu})^2+\cdots+(X_n-\hat{\mu})^2}{n-1}$$ は上記のずれの分が修正され、不偏推定量となる

最尤推定

データの観測値が出てくる確率が最大になるようにパラメータを推定する手法

尤度

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(\, x \mid \theta)$$



ガウス分布の例

$$p(\, x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$p(\, x_1, x_2, \dots, x_n \mid \mu, \sigma)$$
$$= p(x_1|\mu,\sigma)p(x_2|\mu,\sigma) \dots p(x_n|\mu,\sigma)$$
$$= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} exp\left\{-\frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \dots + (x_n-\mu)^2}{2\sigma^2}\right\}$$

これをμに関して最大化する

この部分をμに関して最大化する

微分して０とおくと

$$\hat{\mu} = \frac{x_1 + x_2 + x_3 + \quad + x_n}{n}$$

最尤推定によるパラメータの推定の例：ポアソン分布

最尤推定を使うとガウス分布以外のパラメータも同様の手続きで推定量（最尤推定量）を計算することができる

ポアソン分布のパラメータλを推定するための最尤推定量 $\hat{\lambda}$の計算方法を導出する

サンプルに関する尤度は以下のように計算できる

$$P_\lambda(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$
$$= P_\lambda(X_1 = x_1)P_\lambda(X_2 = x_2)\ldots P_\lambda(X_n = x_n)$$
$$= \frac{\lambda^{x_1+x_2+\cdots+x_n} e^{-n\lambda}}{x_1! x_2! \ldots x_n!}$$

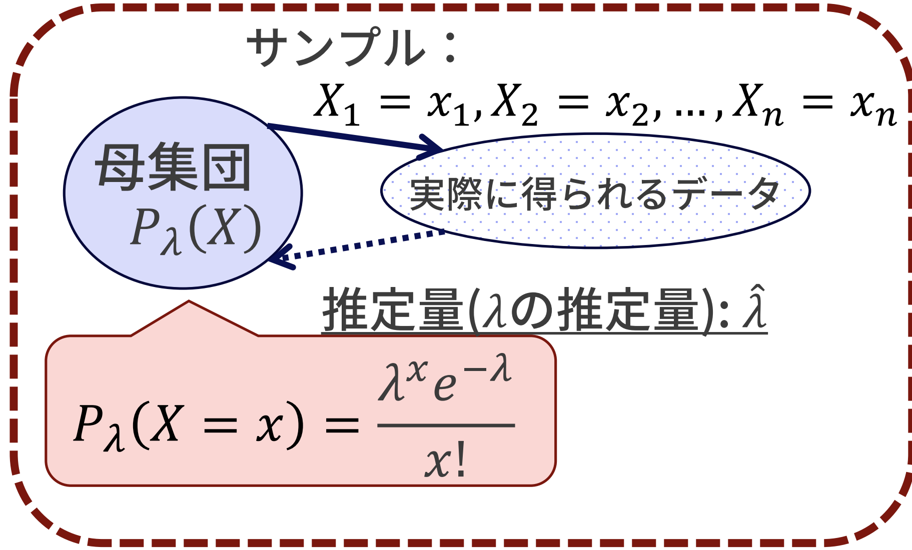
$$\hat{\lambda} = \mathrm{argmax}_\lambda P_\lambda(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

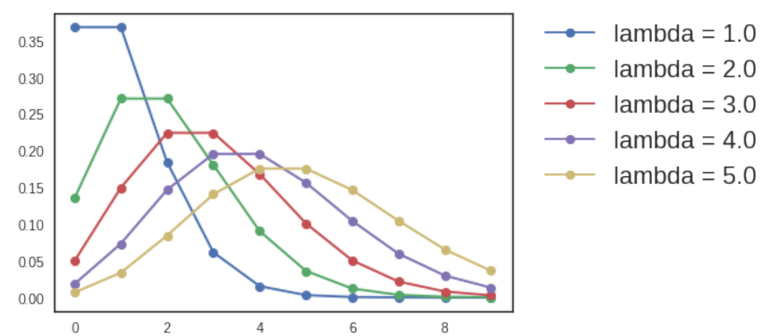
微分してゼロとおいて最大値を計算する

$$(x_1 + x_2 + \cdots + x_n) - n\hat{\lambda} = 0$$

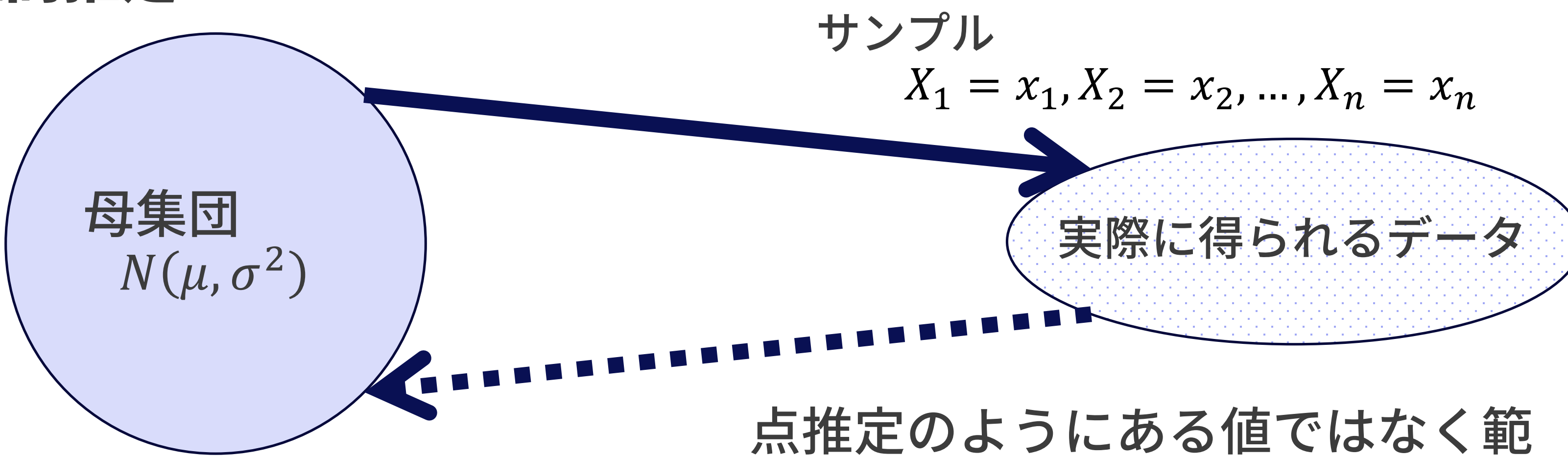
$$\hat{\lambda} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

サンプル平均で$\hat{\lambda}$が最尤推定できることがわかった





区間推定



点推定のようにある値ではなく範囲を推定する

$$P\left(\hat{\mu}_{\text{low}} \le \mu \le \hat{\mu}_{\text{high}}\right) = 1 - \alpha$$

- 下側信頼限界: $\hat{\mu}_{\text{low}}$
- 上側信頼限界: $\hat{\mu}_{\text{high}}$
- 信頼係数: $1 - \alpha$
- 信頼区間: $[\hat{\mu}_{\text{low}}, \hat{\mu}_{\text{high}}]$

サンプルからμに関する統計量の分布を考える必要があることから、統計的仮説検定と類似した設定である

$\frac{\bar{X}-\mu}{S/\sqrt{n}}$がt分布に従うことからt分布上で、信頼係数を満たす範囲$[t_{\text{low}}, t_{\text{high}}]$を決める

$$t_{\text{low}} \le \frac{\bar{X} - \mu}{S/\sqrt{n}} \le t_{\text{high}}$$

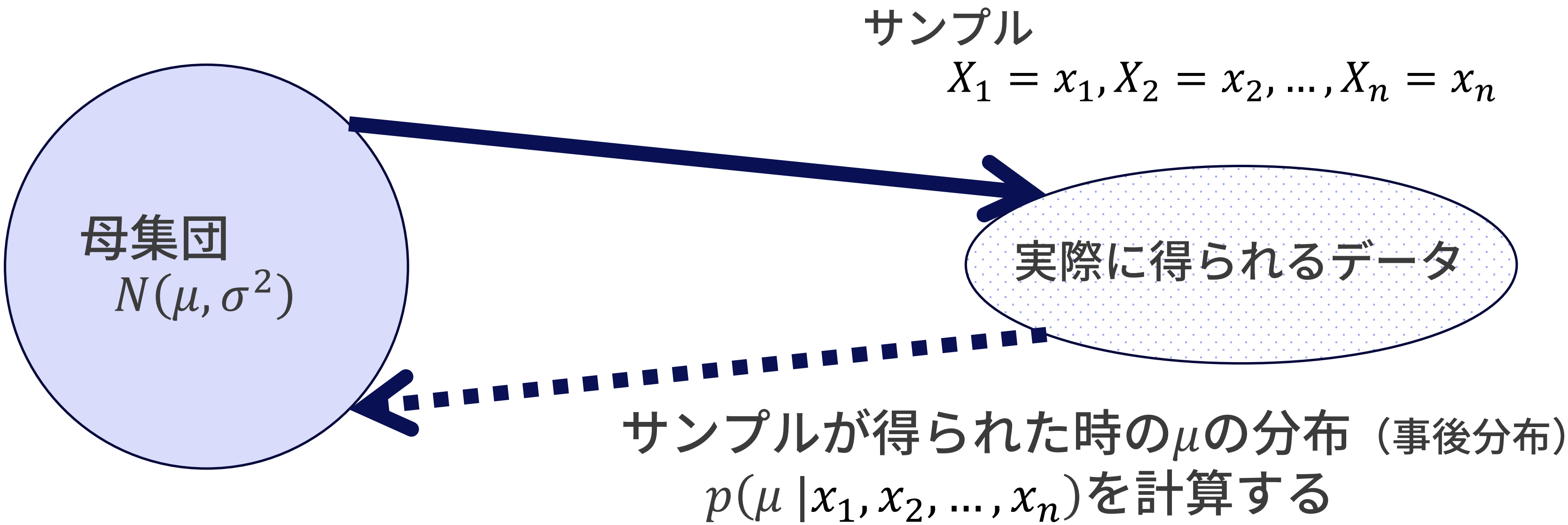
$$\underbrace{\bar{X} - \frac{t_{\text{high}} S}{\sqrt{n}}}_{\hat{\mu}_{\text{low}}} \le \mu \le \underbrace{\bar{X} - \frac{t_{\text{low}} S}{\sqrt{n}}}_{\hat{\mu}_{\text{high}}}$$

$\bar{X}$: サンプル平均
S: 不偏分散

ベイズ推定（やや発展的）

点推定や区間推定では母集団におけるμは確率変数ではなく単なる値であり、
それを観測される確率変数や統計量を用いて推定していた

ベイズ推定ではμを確率変数とみなし、観測からこの確率分布を直接推定する



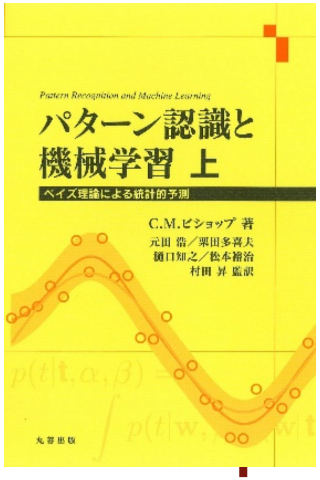
ベイズ推定の利点：

- 点や区間よりもμ ▲振る舞いが詳細にわかる
- データ数が少ない場合でも良く動くことが多い
- より複雑な場合でも同様の原理で計算可能（統計的機械学習へ）

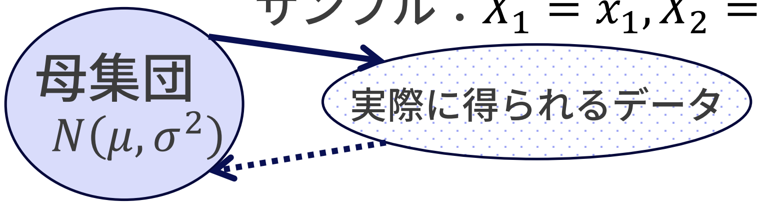
ベイズ推定の例（やや発展的）

ここではベイズ推定のイメージをつかむため、単純な例のみを示す

より詳細は右の教科書および機械学習の講義にて紹介する



サンプル：$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$



ここでは, σ^2は既知として
推定しないこととする

サンプルが得られた時のμの分布$p(\mu\,|x_1, x_2, \ldots, x_n)$を計算する（事後分布と呼ばれる）

尤度と同一の式：
母集団からのサンプルなので、
$N(\mu, \sigma^2)$

ベイズの定理

（導出は第２回確率統計：e-learning の条件付き確率の定義から容易に計算できる）

事後分布

$$p(\mu\,|x_1, x_2, \ldots, x_n) = \frac{p(x_1, x_2, \ldots, x_n|\mu)p(\mu)}{p(x_1, x_2, \ldots, x_n)}$$

事前分布と呼ばれる
$N(0, 1)$とする

周辺尤度と呼ばれる
（μ に依存しない）

計算

$$p(\mu\,|x_1, x_2, \ldots, x_n) \propto \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2+\sigma^2\mu^2}{2\sigma^2}\right)$$

$$\propto \exp\left(-\frac{\sum_{i=1}^{n} -2x_i/(n+\sigma^2)\mu+\mu^2}{2\sigma^2/(n+\sigma^2)}\right)$$

$$\propto N\left(\textstyle\sum_i x_i/(n+\sigma^2), \sigma^2/(n+\sigma^2)\right)$$

この場合、事後分布は正規分布

μに関係のある部分以外は比例
定数となるので、比例記号

μについて二次式なので平方完
成することで正規分布にでき、
μの事後分布を計算で来ている

統計的機械学習へ

- 統計的機械学習では、これまでに学んだガウス分布などのパラメータを推定するのと同様の枠組みで、例えば、以下のような状況に対応していく

  - 実際のデータは多次元のデータであることから、
    - より多次元の分布のパラメータを推定する

  - 実際のデータでは変数間に依存関係があることから、
    - 依存関係があるような多次元の分布のパラメータを推定する

  - 実際のデータではガウス分布のような山なりの単純な分布とは限らないことから
    - より複雑な分布を推定する（例えば山が複数あるような分布など（多峰性））