データ解析

第８回：統計的推測（１）

渡辺澄夫

統計的推測とは

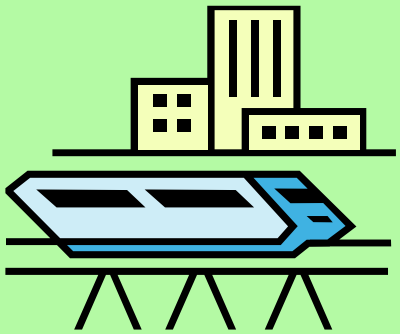
観測データ

回帰・判別分析

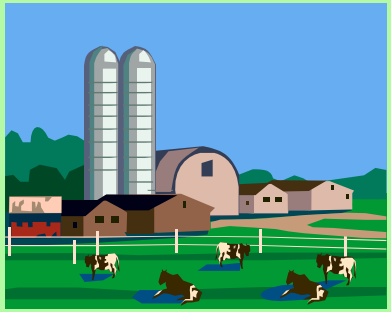
推定と検定

解析方法

ベイズ法
・階層ベイズ法



時系列予測

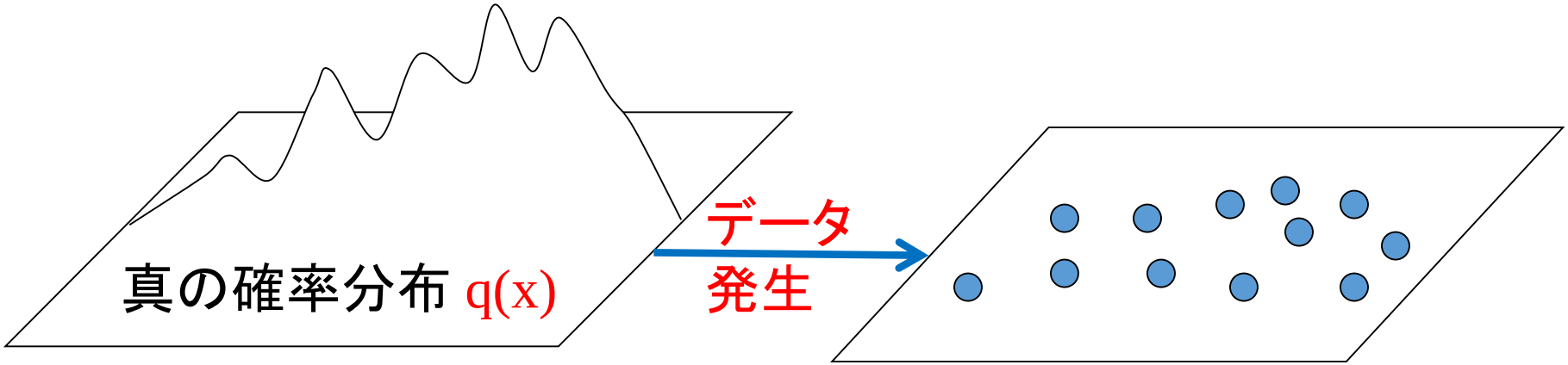


主成分・因子
・クラスタ分析

初等解析から統計モデリングへ

これまでに、初等的なデータ解析の方法である回帰・判別・主成分・因子・クラスタ・時系列分析について習ってきた。これらの方法は、本格的な分析を行なう前にデータのおおよその性質を調べるための予備的な方法である。今日のデータ解析では、データのおおよその性質がわかったら、データの生成過程を確率分布を用いてモデリングすることで真の分布を推測することが行なわれる。まずモデリングの基礎となる概念を学ぼう。

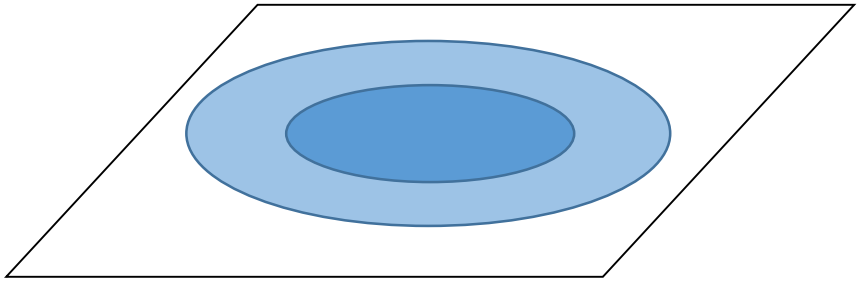
前半のまとめ



データ $X^n = (X_1, X_2, \ldots, X_n)$

データが与えられれば定められた手順に従ってデータ解析の計算を行なうことができる。しかしながら，<u>計算できればそれでいいのか？</u>　実世界では真の分布は不明であるから「定められた手順の計算結果が適切かどうか」を知ることができないのに何を信じることができるのだろうか？。データ解析の最大の弱点はデータ解析者が自身の解析の妥当性を知ることがないまま実用してしまう点にある。

データ解析



回帰・判別・主成分・因子・クラスタ・時系列解析の結果

「データ解析」を数学の対象とする

データ解析を「定められた手順を適用するだけのもの」から「未知の分布を数理的に推測するもの」にしていくためには「データ解析」を数学の対象にしなくてはならない。

データ解析とは関数「データ→解析結果」のことである。この関数の性質を数学的に調べるためには確率論の上にこの関数を乗せて性質を調べることが必要である。

(1) データは確率変数である。未知の確率分布 $q(x)$ に従う。

(2) 「データ→解析結果」の適切さは確率的に変動するから、確率的に変動するものを測ることが必要である。

(3) できるだけ適切さが良好になるようにデータ解析法を作る。

統計的推測

統計的推測とは

データ　$X^n = (X_1, X_2, \ldots, X_n)$ が得られたとき、このデータを発生した確率密度関数 $q(x)$ を知りたい。しかし確率分布の集合は一般には無限次元でありデータは有限であるから、$q(x)$ を完全に知ることはできない。そこでパラメータ w をもつ x の確率密度関数 $p(x|w)$ を使って推測を行なう。

<u>問題１．</u>　統計モデル $p(x|w)$ を用いた推測は真の密度関数 $q(x)$ からどのくらいずれているか。

<u>問題２．</u>　真の密度関数 $q(x)$ が不明であるにも関わらず、推測された結果が適切であるかどうかを知るにはどうしたらよいか。

統計モデルの例(1)

ベルヌーイ分布. データ $x \in \{0,1\}$, パラメータ $0<a<1$ について

$$p(x|a) = a^x(1-a)^{1-x}$$

と定義する。すなわち $p(1|a)=a$, $p(0|a) = 1-a$ である。

データ $X^n =(X_1,X_2,\ldots,X_n)$ が0または1に値をとる確率変数の列であるときに $q(x)$ の推測に用いられる。

（注意）$a=0, 1$ もパラメータの集合に含める場合もある。

（例） コインを振って表が出る確率を推測するときなどに現れる。

推測の問題を考える時 $a=0, 1$ は特別な点である。

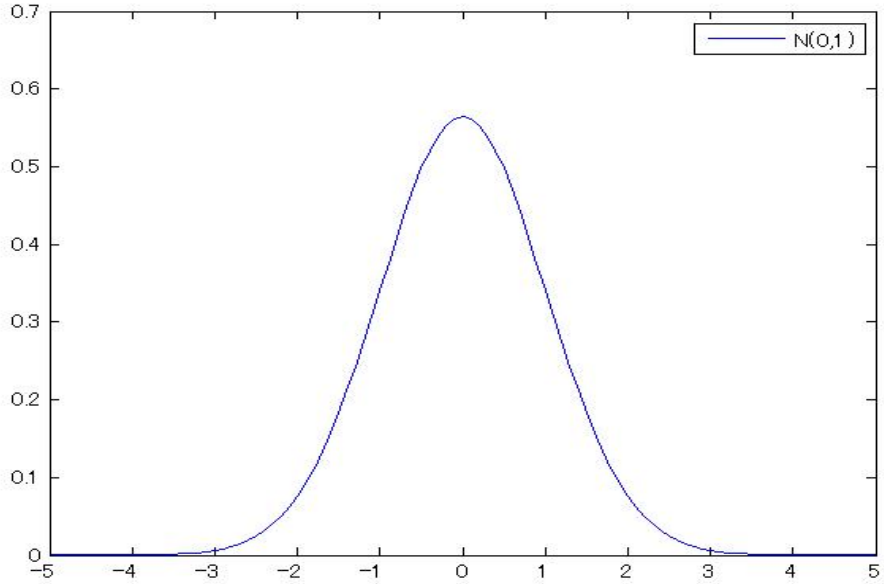
「1000回コインを振って全部表だった。このコインで表が出る確率は？」

統計モデルの例(2)

正規分布. データ x ∈**R** , パラメータ a ∈**R** , s>0 について

$$p(x|a,s) = 1/(2\pi s^2)^{1/2} \exp(-(x-a)^2/(2s^2)).$$

データ $X^n = (X_1, X_2, \ldots, X_n)$ が実数に値をとる確率変数の列であり、真の分布が正規分布であると考えられるときに用いられる。



p(x|0,1) : 標準正規分布

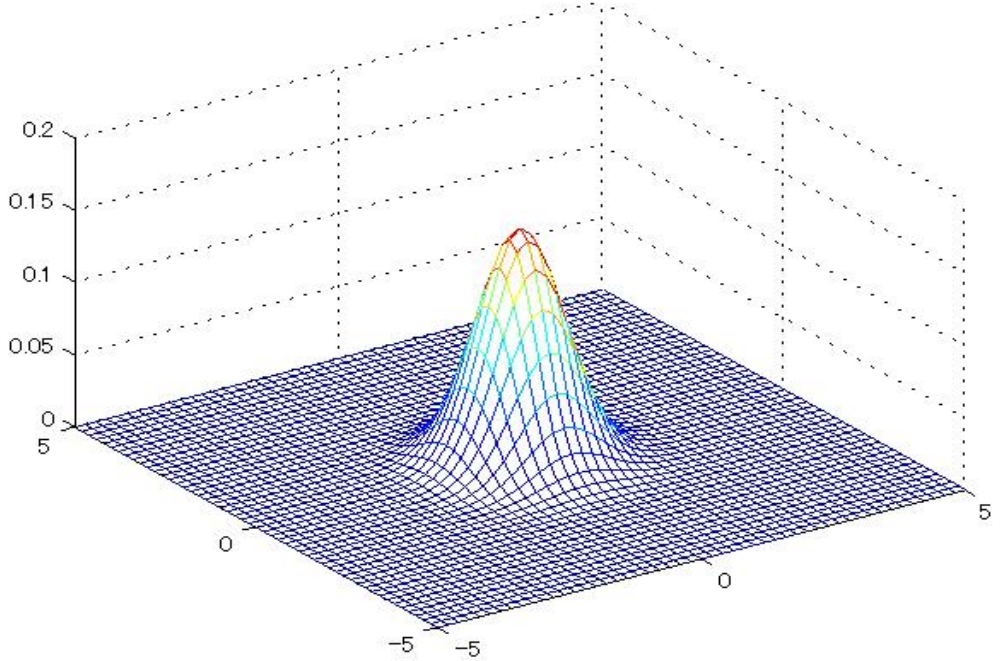
統計モデルの例(3)

<u>高次元正規分布</u>. $x, a \in \mathbf{R}^N$, $\Sigma>0$ (正定値行列)について

$$p(x|a, \Sigma) = 1 / ((2\pi)^N \det(\Sigma))1/2 \exp(- (x-a)\cdot\Sigma^{-1}(x-a)/2).$$

データ $X^n = (X_1, X_2, \ldots, X_n)$ が$\mathbf{R}^N$に値をとる確率変数の列であり

真の分布が正規分布であると考えられるときに用いられる。



p(x|0,I) : 正規分布

(Iは単位行列)

KL情報量

統計的推測の適切さを測る量c

カルバック・ライブラ情報量

定義. R^N 上の確率密度関数 $q(x), p(x) > 0$ が与えられたとき $K(q||p) = \int q(x) \log (q(x)/p(x))\, dx$ を $q(x)$ と $p(x)$ のカルバックライブラ情報量あるいは相対エントロピーという。

(注意) 以下では $x \in R^N$ の場合を想定する。上記の積分が有限でないときにはカルバックライブラ情報量は無限大であると定義する。xが連続集合でなく離散集合に含まれるときは積分を和に置き換える。実数より濃度の大きな集合上の確率分布に対しても（互いに絶対連続な確率分布に対しては）カルバックライブラ情報量は定義できるが、この講義ではその場合については考察しない。

（注意）この講義では量子力学における統計的推測の問題は考察しない。量子推定においては $q(x)$, $p(x)$ は状態密度関数であり互いに可換でない。そのような場合に統計的推測をどのように行なうか、相対エントロピーをどのように定義するのが適切かなどについて研究が行なわれている。

カルバック・ライブラ情報量の性質

定理. 関数 $q(x)$, $p(x)$ が連続であるとする。次が成り立つ。
(1) 任意の $q(x)$, $p(x)>0$ について $K(q||p) \geqq 0$.
(2) $K(q||p)=0 \Leftrightarrow q(x)=p(x)\ (\forall x)$

証明。$t>0$ の関数を $F(t) = \log t + 1/t - 1$ とおき、微分して増減表を書くとで $F(t) \geqq 0$ であり、$F(t)=0 \Leftrightarrow t=1$ がわかる。また$q(x)$, $p(x)$ の積分値が1になることから $K(q||p)=\int q(x)\, F(q(x)/p(x))\, dx$.
これより(1),(2) が得られた。(証明終)

○ 真の密度関数が $q(x)$ で、推測された密度関数が $p^*(x)$ のとき、$K(q||p^*)$ が小さいほど推測の精度が良いという指標とする。

なぜKL情報量で測るのか

(1) 真の分布 $q(x)$ と学習の結果 $p^*(x)$ の違いを測る量としてKL情報量でないものも提案されている。例: ヘリンガー距離（の２乗）

$$H(q||p) = (1/2) \int (q(x)^{1/2} - p(x)^{1/2})^2 \ dx$$

(2) 応用上で統計的推測によって小さくしたい量が定まっているときは、その量を用いて測るのが適切である。

(3) KL情報量は情報源 $q(x)$ と符号化装置 $p(x)$ における符号長（の増分）であるので、情報理論的な意味を持っている。

(4) 尤度関数 $\Pi\, p(X_i|w)$ を $\Pi\, q(X_i)$ で割り算すると $n \to \infty$ で概収束

$$[\ \Pi\, p(X_i|w) \ / \ \Pi\, q(X_i)\ ]^{1/n} \to \exp(\ -\ K(q(x)||p(x|w))\ )$$

が成り立つので、KL情報量は統計的推測では基礎的な概念である。

汎化誤差と汎化損失

真の分布を q(x) とし学習の結果を p*(x) とするとき汎化誤差を

$$G = \int q(x) \log (q(x) / p^*(x)) \, dx$$

によって定義する。

$$G = \int q(x) \log \; q(x) \, dx \quad - \int q(x) \log p^*(x) \; dx$$

なので、右辺第２項が小さいことと汎化誤差が小さいことは等価。

統計的推測の結果 p*(x) の汎化損失を

$$- \int q(x) \log p^*(x) \; dx$$

によって定義する。(注: $-\int q(x) \log \; q(x) \, dx$ は真の密度関数のエントロピー)。

KL情報量の例(1)

ベルヌーイ分布. $p(x|a) = a^x(1-a)^{1-x}$ についてのKL情報量は

$$K(p(x|a) \,||\, p(x|b)) = \sum_{x=0,1} p(x|a) \log \{ a^x(1-a)^{1-x} / b^x(1-b)^{1-x} \}$$

$$= \sum_{x=0,1} p(x|a) \{ x \log(a/b) + (1-x) \log((1-a)/(1-b)) \}$$

$$= a \log(a/b) + (1-a) \log ((1-a) / (1-b)).$$

ただし a=0 または a=1 では

$$K(p(x|0) \,||\, p(x|b)) = \log (1 / (1-b)),$$

$$K(p(x|1) \,||\, p(x|b)) = \log (1 / b).$$

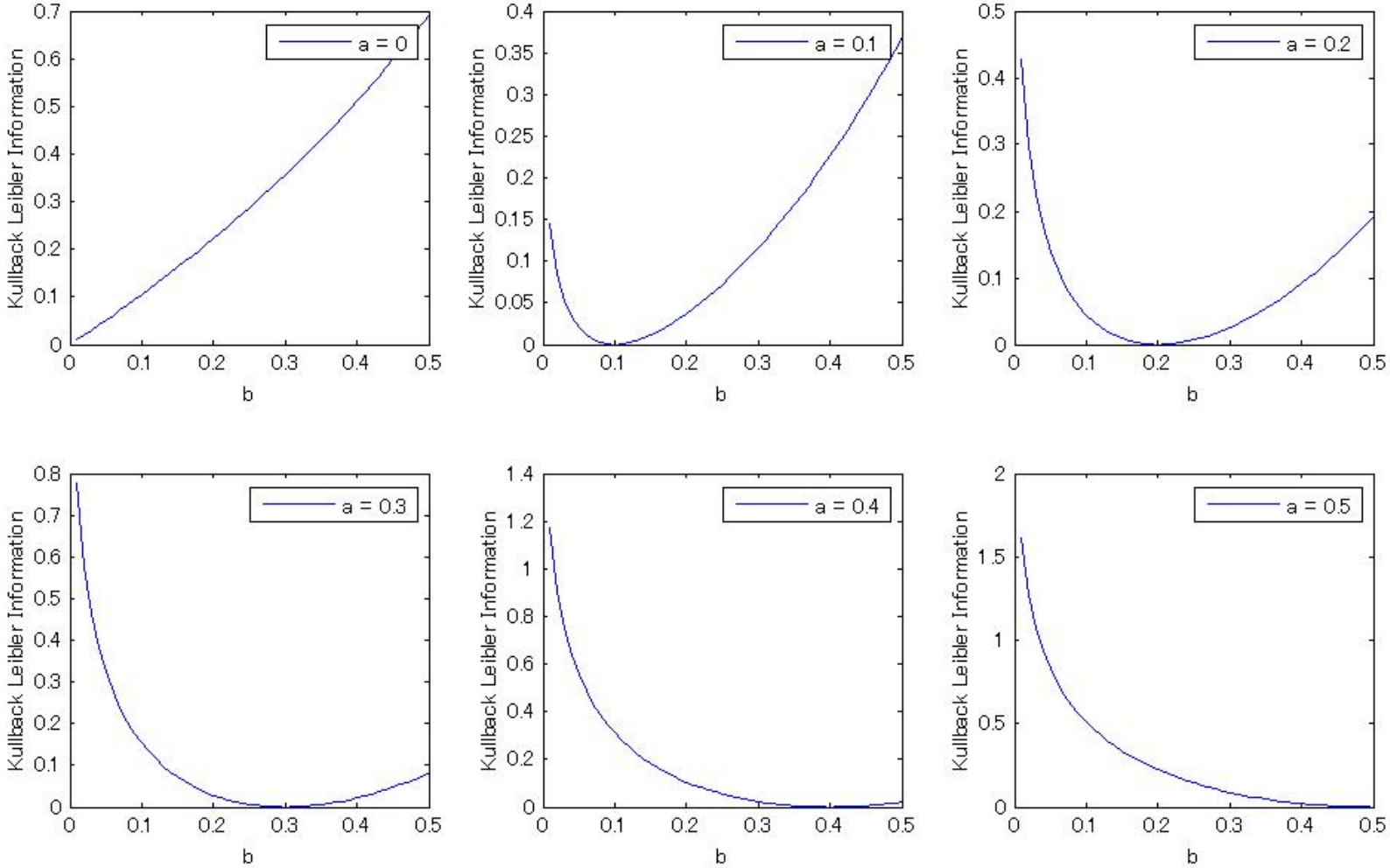
(注意) 0<a<1 で b=0, 1 のときはKL情報量は発散している。

KL情報量の例(1)つづき

ベルヌーイ分布.

K(p(x|a)||p(x|b)) = a log(a/b) + (1-a) log ((1-a) / (1-b)).

a を固定して b を 0.01から0.5 まで変化させたときのKL情報量の様子。



KL情報量の例(2)

正規分布. $p(x|a,s) = 1/(2\pi s^2)^{1/2} \exp(-(x-a)^2/(2s^2))$ についてKL情報量は

$$K(p(x|a,s) \,||\, p(x|b,t))$$

$$= \int p(x|a,s) \log \frac{1/(2\pi s^2)^{1/2} \exp(-(x-a)^2/(2s^2))}{1/(2\pi t^2)^{1/2} \exp(-(x-b)^2/(2t^2))} dx$$

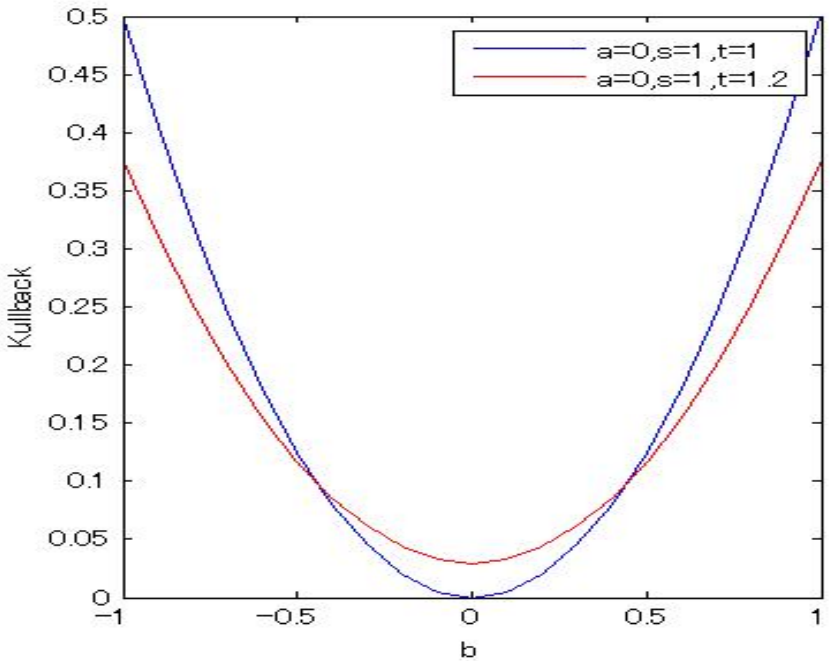
$$= \int p(x|a,s) \{ \log(t/s) - (x-a)^2/(2s^2) + (x-b)^2/(2t^2) \} dx$$

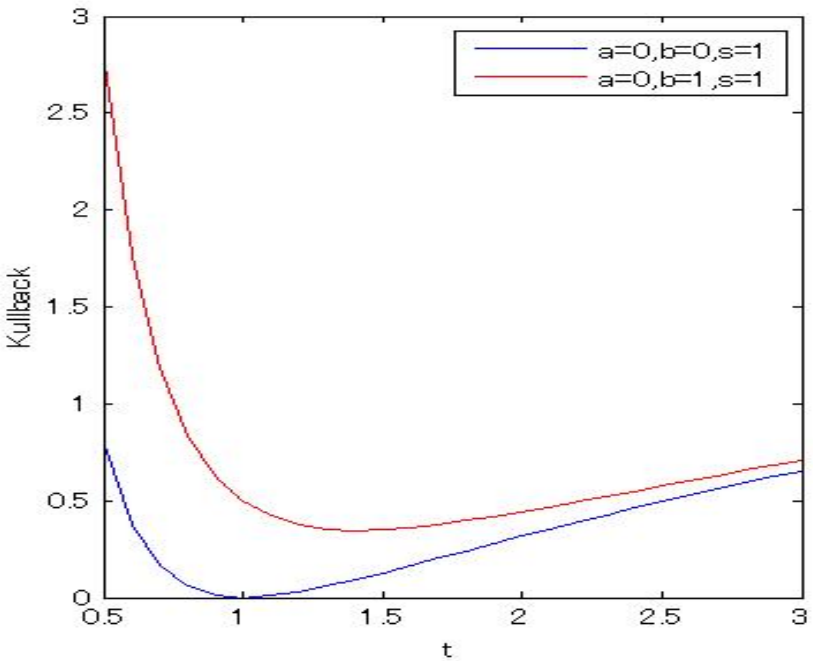
$$= \log(t/s) -1/2 + \{ s^2 + (a-b)^2\} / (2t^2)$$

（注意）$t=s$ のときは $(a-b)^2/(2s^2)$ になる。

KL情報量の例(2)つづき

正規分布. $K(p(x|a,s) \,||\, p(x|b,t)) = \log(t/s) - 1/2 + \{ s^2 + (a-b)^2\} / (2t^2)$





青：標準偏差が同じ(s=t=1)平均bを変化さた。
赤：標準偏差が異なる（s=1,t=1.2)平均bを変化させた。|a-b|>0.5 では標準偏差が異なるほうがKL情報量が小さい。

青：平均が同じ(a=b=1)標準偏差tを変化させた。
赤：平均が異なる(a=1,b=2)標準偏差tを変化させた。KL情報量を最小にする点は標準偏差が等しいところではない。

KL情報量の例(3)

高次元正規分布. $p(x|a,\Sigma) = 1 / ((2\pi)^N \det(\Sigma))^{1/2} \exp(- (x-a) \Sigma^{-1}(x-a)/2)$

のKL情報量は

$$K(p(x|a, \Sigma) \,||\, p(x|b,\Theta))$$

$$= \int p(x|a, \Sigma) \log \frac{1 / ((2\pi)^N \det(\Sigma))^{1/2} \exp(- (x-a) \Sigma^{-1}(x-a)/2)}{1 / ((2\pi)^N \det(\Theta))^{1/2} \exp(- (x-b) \Theta^{-1}(x-b)/2)} dx$$

$$= \int p(x|a, \Sigma) (1/2) \{ \log \det(\Sigma^{-1}\Theta) - \mathrm{tr}(\Sigma^{-1}(x-a)(x-a)^T - \Theta^{-1}(x-b)(x-b)^T) \} dx$$

$$= (1/2) \{ \log \det(\Sigma^{-1}\Theta) - N + \mathrm{tr}(\Sigma\Theta^{-1}) + (a-b)\cdot\Theta^{-1}(a-b)) \}$$

KL情報量の例(3)のつづき

高次元正規分布. で a=b=0 のとき

$$K(p(x|0, S) \,||\, p(x|0,T)) = (1/2) \{ \log \det(S^{-1}T) - N + tr(ST^{-1})) \}$$

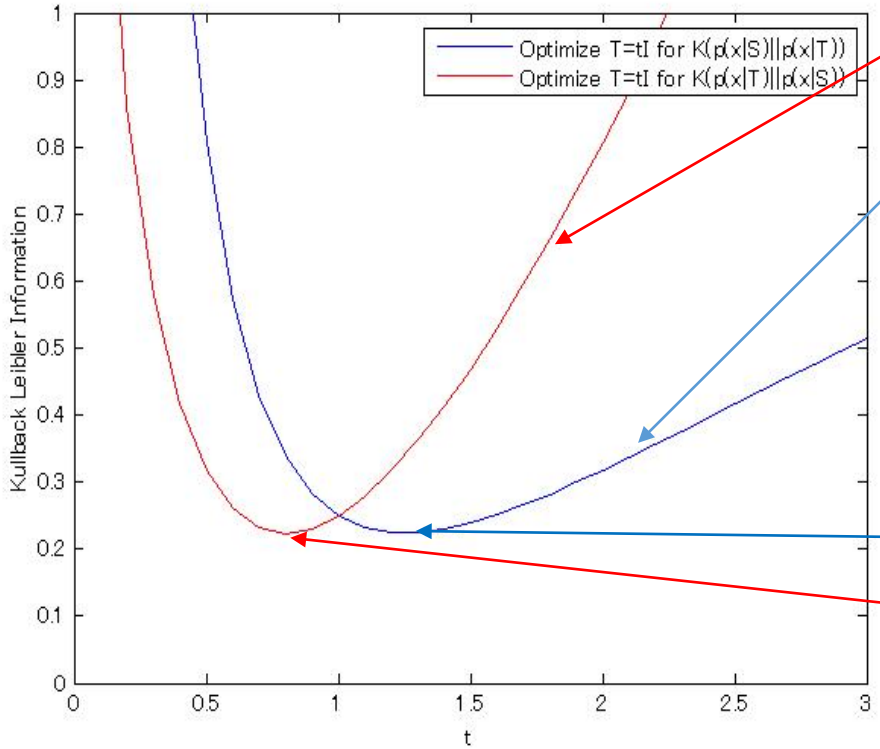
$$S = \begin{pmatrix} 2 & 0 \\ 0 & 0.5 \end{pmatrix} \quad T = \begin{pmatrix} t & 0 \\ 0 & t \end{pmatrix}$$

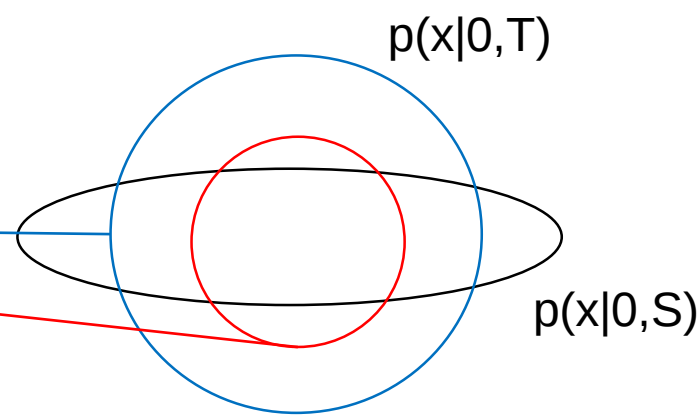
t を変化させたときの

K(p(x|0, T) || p(x|0,S))

K(p(x|0, S) || p(x|0,T))

を描いた。KL情報量を最小にする t は上では縮み下では広がる。

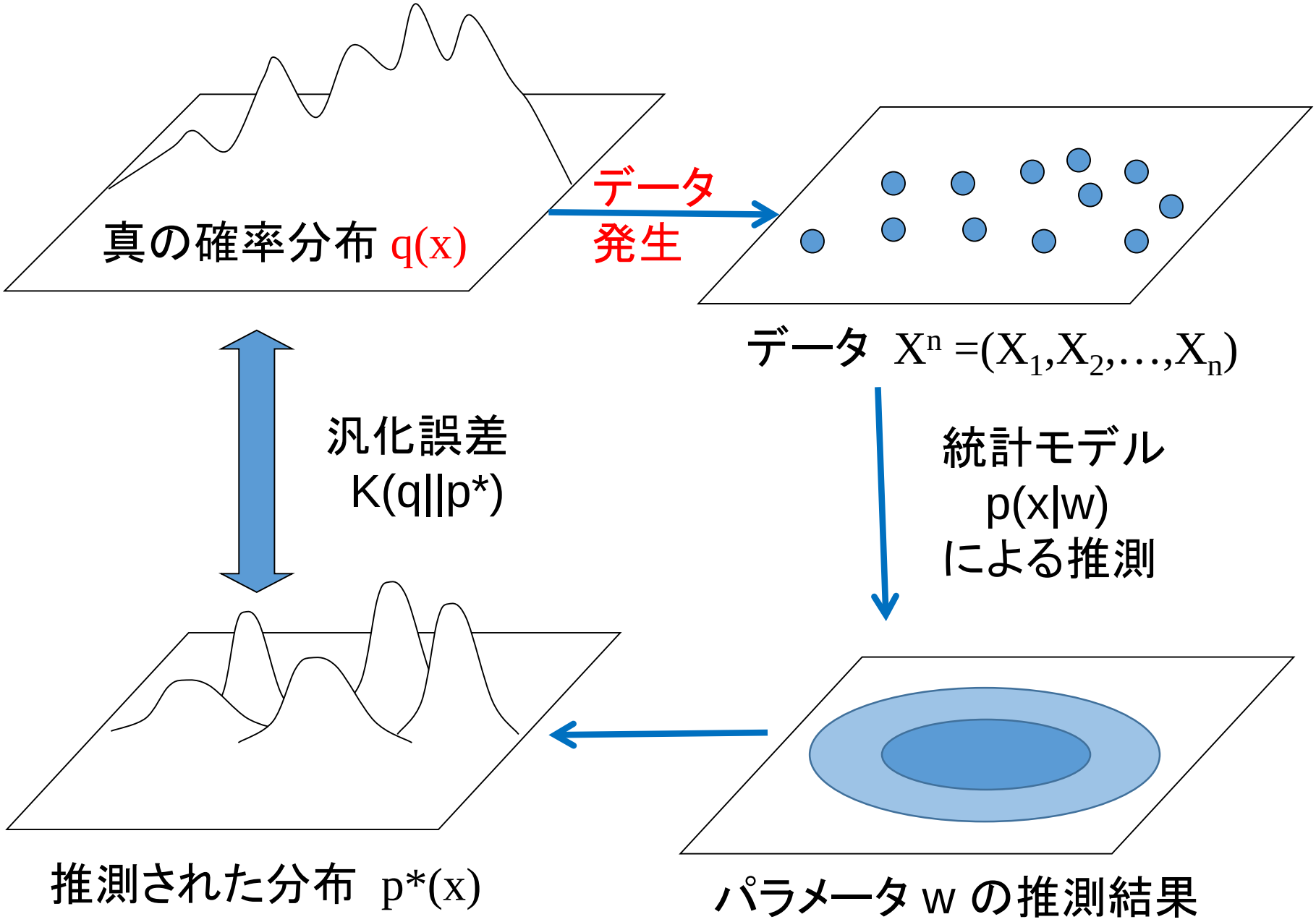




推測とKL情報量

統計的推測の方法の良好さを

KL情報量で比較することができる


真の確率分布 q(x)
データ
発生
データ $X^n = (X_1, X_2, \ldots, X_n)$
統計モデル
p(x|w)
による推測
汎化誤差
K(q||p*)
推測された分布 p*(x)
パラメータ w の推測結果

良い推定法を作りたい

(1) 与えられたデータ $X^n=(X_1,X_2,\ldots,X_n)$ に対して推測された分布 $p^*(x)$ を定める方法「 $X^n \Rightarrow p^*(x)$」には様々なものがある（次週に説明する）。

(2) 真の分布 $q(x)$ が同じでも、そこから発生するデータ $X^n=(X_1,X_2,\ldots,X_n)$ は実験するたびにばらつく（確率変数）。

(3) 従って推測された分布 $p^*(x)$ もばらつく。汎化誤差 $K(q||p^*)$ もばらつく。

(4) まず推定法を固定したときの汎化誤差の挙動を解明する。

(5) 次に汎化誤差ができるだけ小さくなるように推定法を決める。

(6) 真の分布が不明であっても(5)ができるようになることが目標。

例で考える:コインを投げる

コインを投げる例で考える。

(1) 真の分布を q(x) とする（表確率が q(1) 、裏確率が q(0)) 。

(2) 独立なデータ $X^n = (X_1, X_2, \ldots, X_n)$ は、0または1に値をとる確率変数の列。

(3) 統計モデルとしてベルヌーイ分布 $p(x|a) = a^x(1-a)^{1-x}$ を考える。

(4) 推測法「$X^n = (X_1, X_2, \ldots, X_n) \rightarrow p^*(1), p^*(0)$」を決める。

(5) 汎化誤差を求める。

$$K(q||p^*) = q(1) \log(q(1) / p^*(1)) + q(0) \log (q(0) / p^*(0))$$

上記が1回の推測実験である。データの出方によるばらつきや、データの数 n の影響などを調べたい場合には、それぞれについて繰り返しや変更を行って実験する。

推定する方法

ここでは、次の推定法を考察する。Small≧0 を固定して

p*(1) =（表が出た回数+Small）/（投げた回数+2*Small）

p*(0) = 1-p*(1)

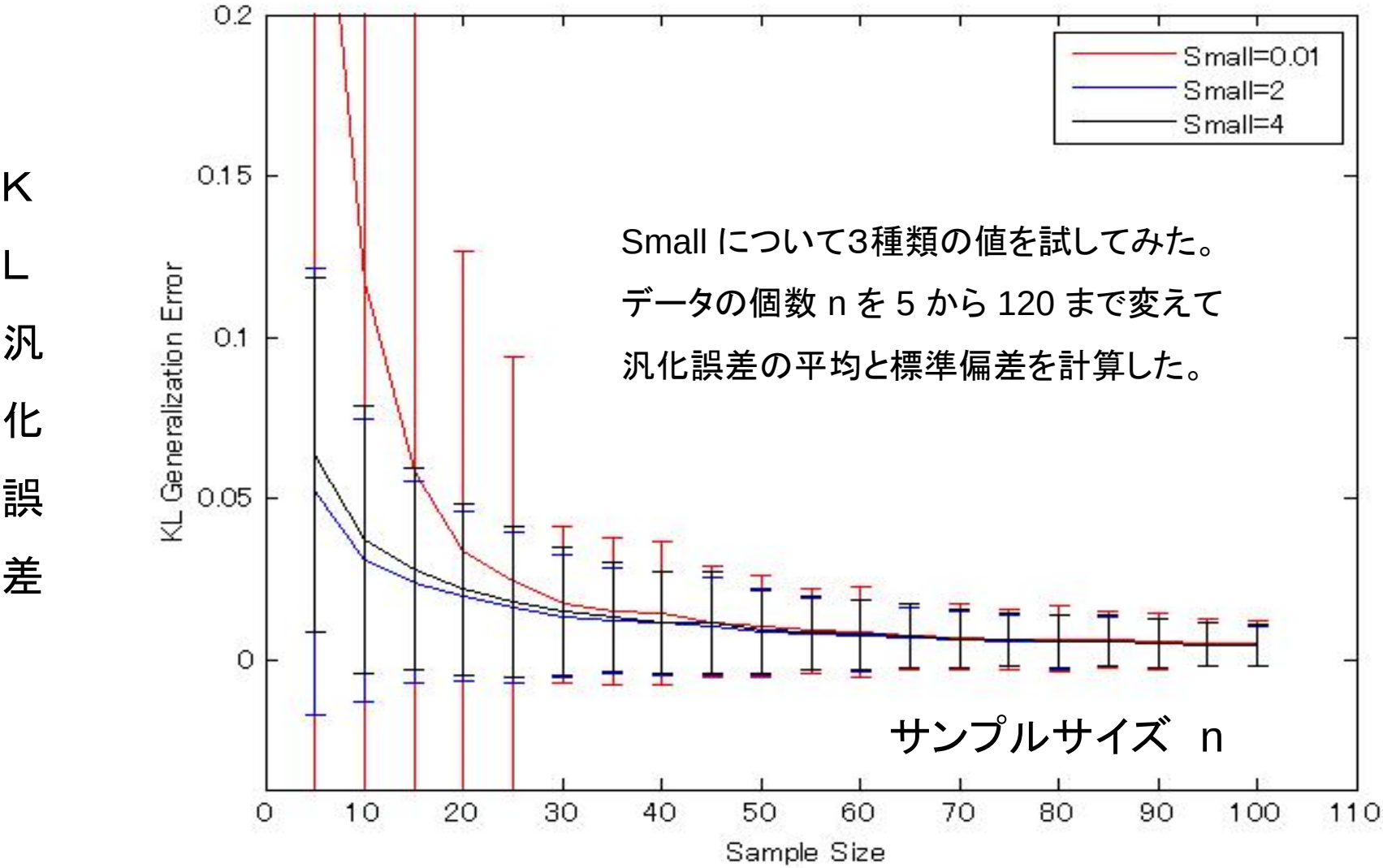
(1) 日常感覚である Small=0 は最も良い推測だろうか。

(2) Small を変化させるとKL情報量はどうかわるだろうか。

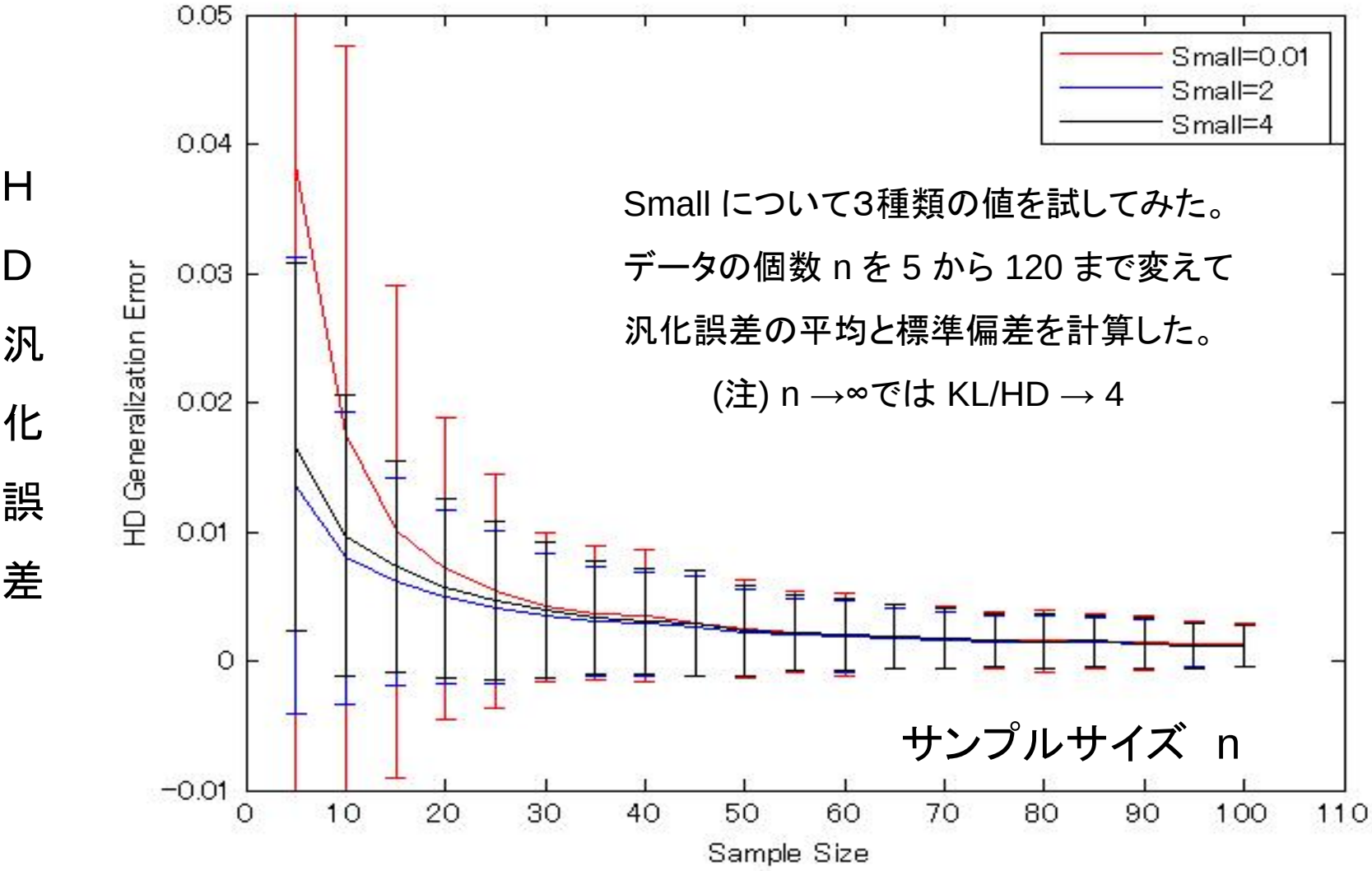
真の分布を q(1)=1/4, q(0)=3/4 に固定する。

同じ条件で 1000 回の独立な試行を行って平均と分散を求める。

ベルヌーイ分布のKL汎化誤差の挙動



ベルヌーイ分布のHellinger汎化誤差の挙動



わかったこと

(1) コインを投げて表が出る確率を推測する場合、データ解析手法「p*(1) = 表が出た回数 / 投げた回数」はKL情報量から見て精度の良い結果を与えるわけではない (n<25ではかなり悪い推測である)。

(2) 今回の比較では「p*(1) = (表が出た回数+2) / (投げた回数+4)」が最も精度がよかった。

(3) データ解析法の候補としては無限の可能性がある。例えば、「p*(1)=[(表が出た回数)/(投げた回数)]$^{0.9}$」なども推定法の候補になるかもしれない。無限の可能性を全て比較することは(一般には)できない。候補としてはどのようなものを考えるのがよいのだろうか。

データ解析について再考する

(1) 方法 「n 回投げてk回表のとき (k/n) を推定確率とする」は、定められた手順に従ってコインの表が出る確率を推定しているのでデータ解析のひとつである。

(2) データ解析を実問題に適用するだけなら、(k/n) はすぐに計算できて終了である。「答えは (k/n) です」。しかし、それは正しいのだろうか。
最も良い推測だろうか。

(3) データ解析で大切なのは、ここからである。<u>考察している問題に手続きを適用するだけなら何でもよいことになる</u>。「結果はどの程度に信じてもよいか」「もっと適切な方法があるのではないか」を知るにはどうしたらよいだろうか。