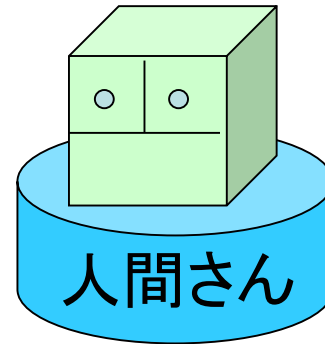
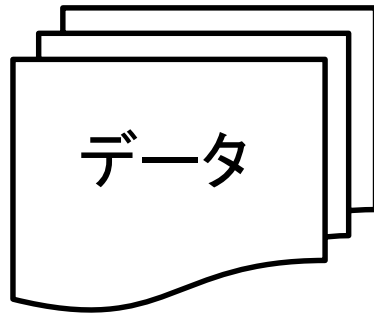


データ解析

第1回 モデルは真ではない

渡辺澄夫

データと人間

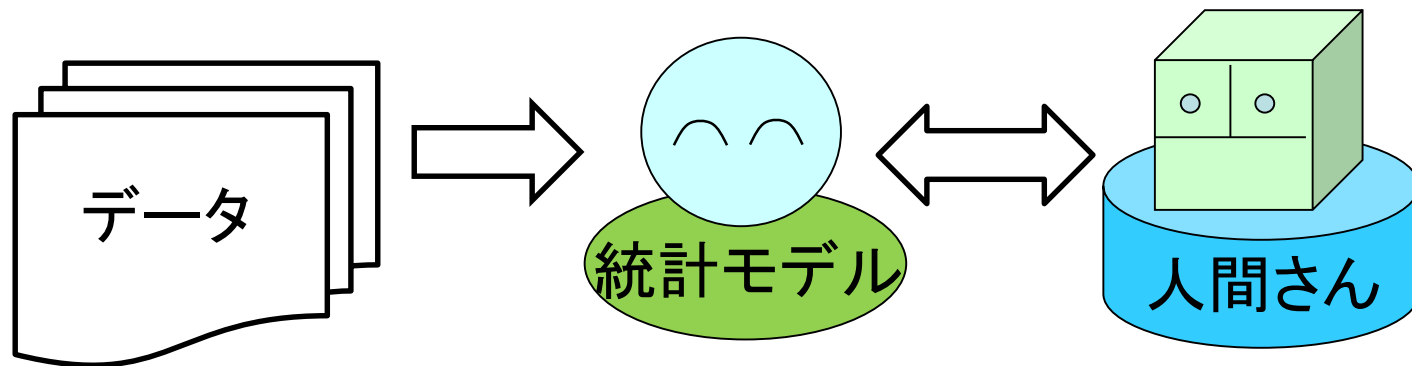


データと人間がある。

コンピュータやネットワークの発展のおかげで膨大なデータが得られるようになった。

人間は、何を目標にして何を行なうのだろうか。

データ モデル 人間



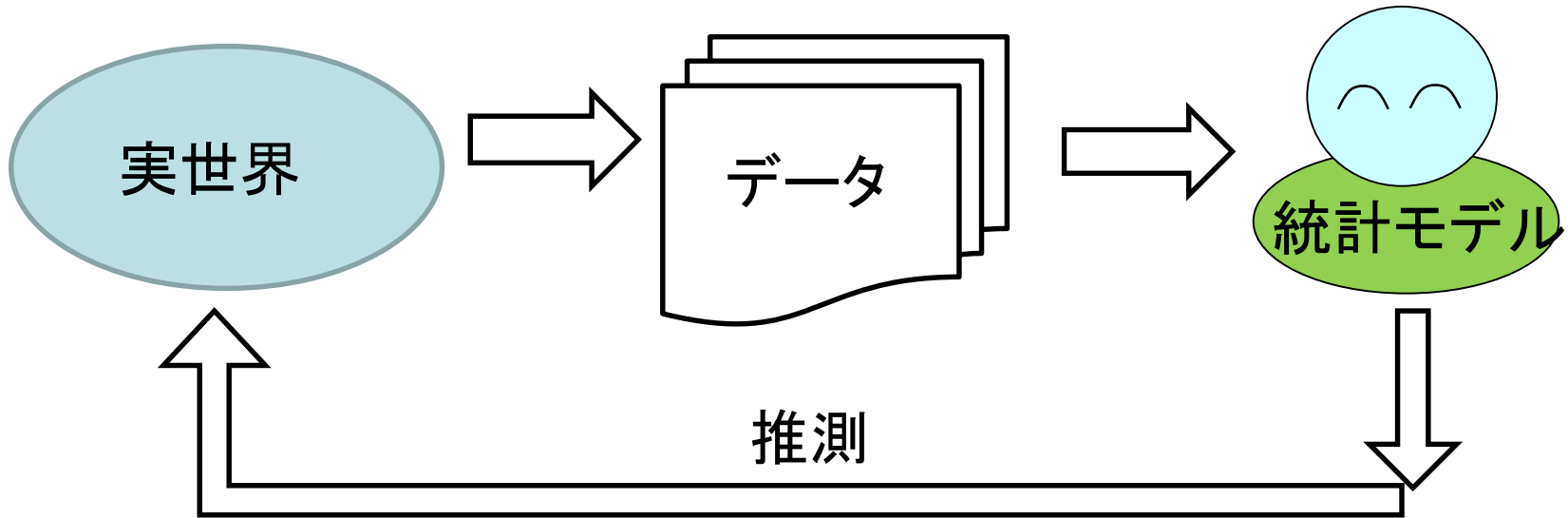
データそのものをよく見ることは大切だ。

しかしデータを見ているだけではわからないことや
気づかないこともある(注)。

そこで人間は統計モデルを作る。 統計モデルが必要だ。

(注) この講義では学びませんが、G.Chaitin が次のことを示しています。「例えばハードディスク上にある固定されたメモリが確率分布から生成されていないとしても、チューリングマシンを用いて情報圧縮するならば符号長を最小とする方法は確率分布を用いたものと漸近的に等価になる」。

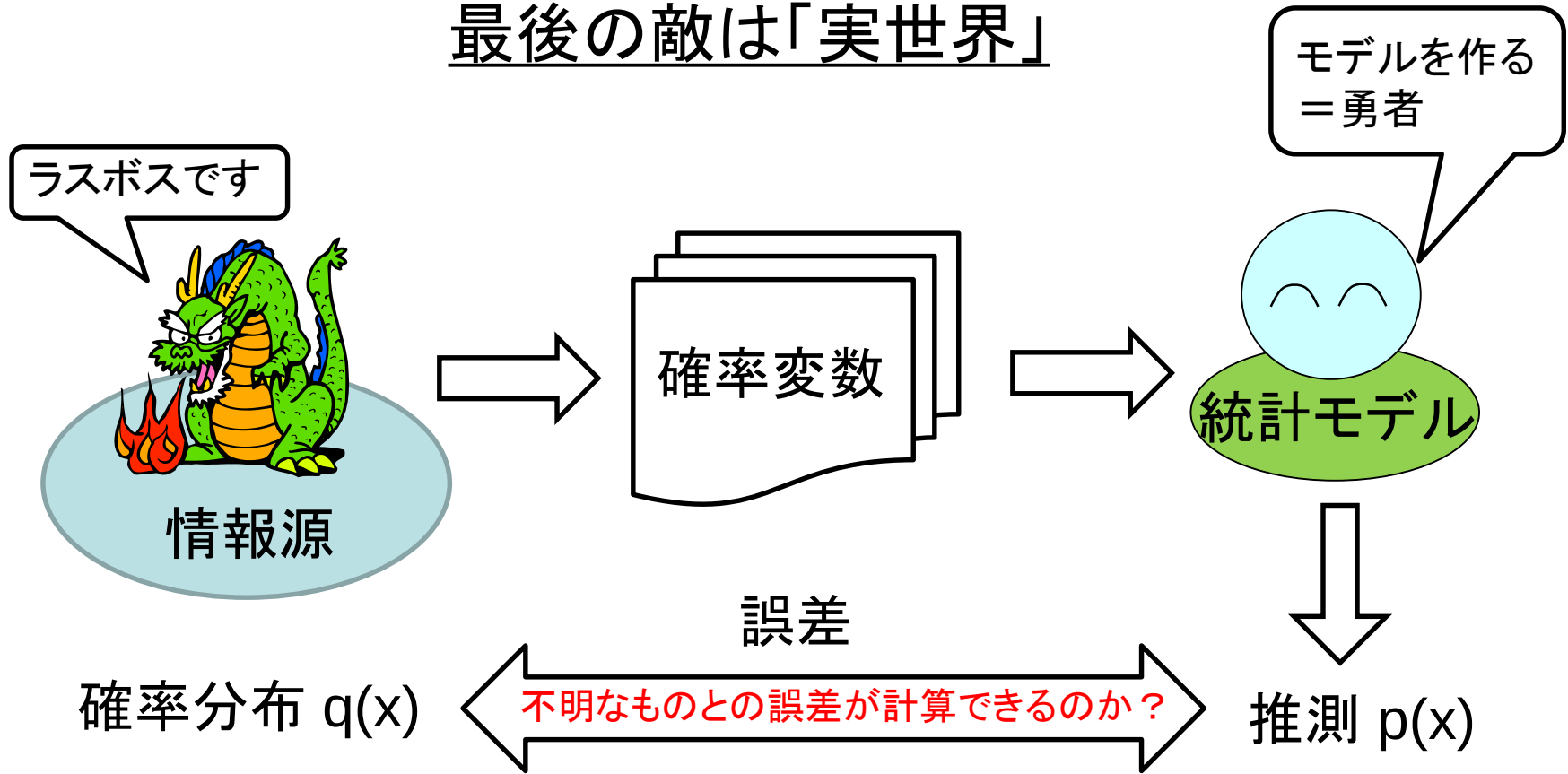
データ解析とは何ですか



モデルは実世界を知るために作られる。人間は実世界を直接に知ることはできないので、データとモデルを通して実世界を推測することになる。その作業のことを「データ解析」という。

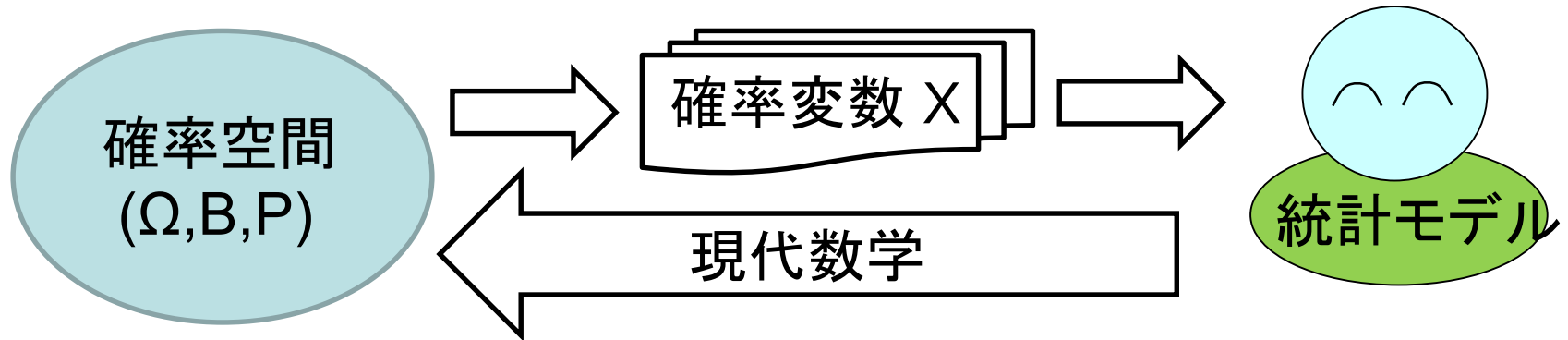
今日、データの膨大化により、データ解析を主務とする業務が増大し「データサイエンティスト」や「機械学習エンジニア」という新しい仕事が生まれている。

最後の敵は「実世界」



実世界 (= 情報源) が未知の確率分布 $q(x)$ である場合を考えよう。人間は統計モデルを設定しデータを経由して情報源を推測する。モデルは仮説であり実世界ではない。推測された結果 $p(x)$ はどの程度に $q(x)$ を正しく推定しているのだろうか。この問題は数学的な普遍性を持っており、あらゆる現代数学が必要になる。

真の目標は何か



データ解析 の真の目標は 未知の確率空間を推測すること である。

注意。たまたま得られたデータについて何かをすることに意味はない。

- ソフトウェアをどうするかは真の目標ではない。
- 計算の複雑さを減らすことは真の目標ではない。
- 最適化問題を解くことは真の目標ではない。

確率論では、確率空間を設定した上で確率変数の挙動を導出する。
統計学では、確率変数の挙動から確率空間そのものを推測する。
完全には達成できない真の目標にどこまで迫ることができるかを問う。

データ解析に関連する教員(数理・計算科学系)

<https://educ.titech.ac.jp/is/> より

三好直人	応用確率論・情報通信システムのモデル化と解析
樺島祥介	統計力学・情報理論・機械学習
高安美佐子	ビッグデータ解析・社会経済物理学
金森敬文	統計的機械学習
福田光浩	連続最適化・凸最適化
山下真	数理最適化・応用数学
中野張	制御理論・確率過程・ファイナンス
小渕智之	統計力学・情報理論・機械学習
金澤輝代士	ビッグデータ解析・社会経済物理学

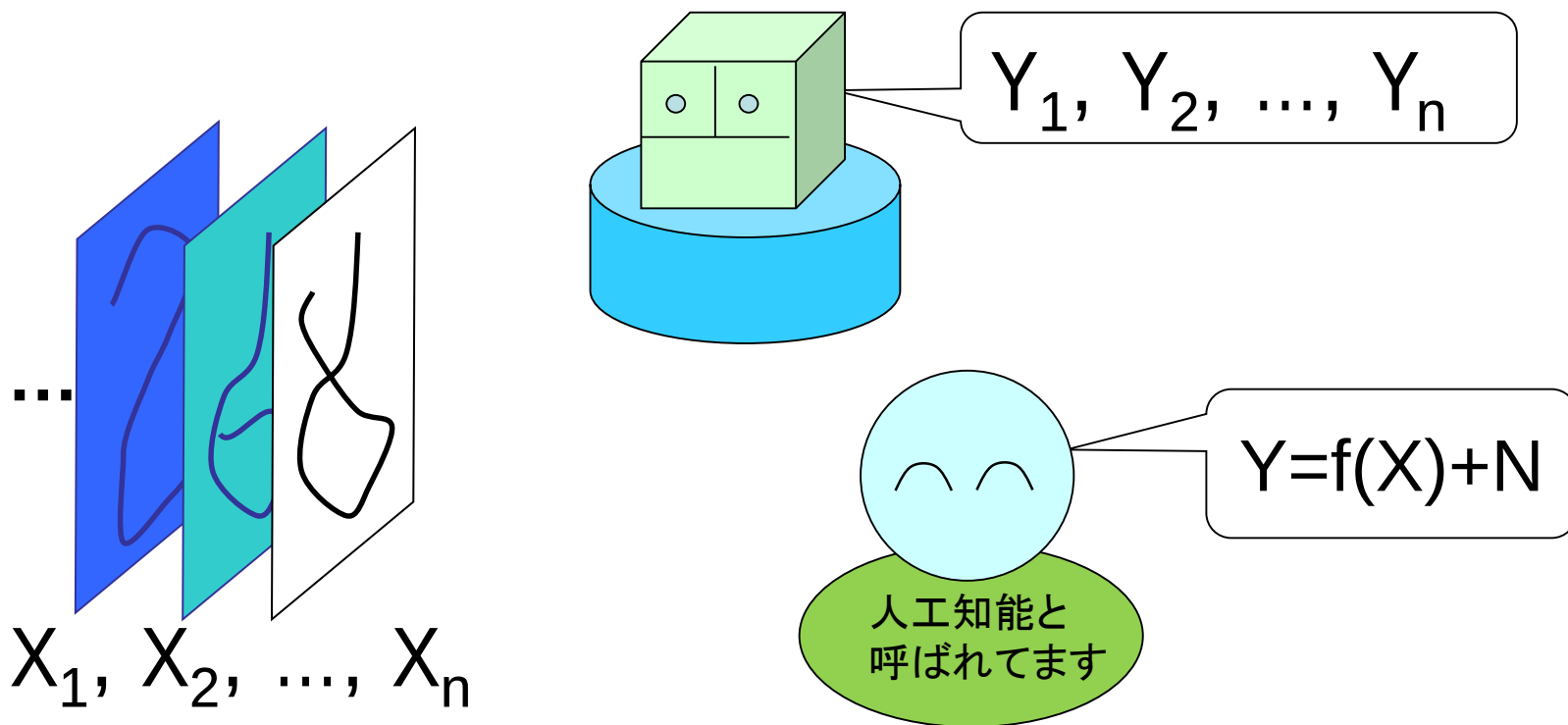
学ぶこと

1. 回帰・判別分析
2. 因子・主成分・クラスタ分析
3. 時系列予測
4. ベイズ法・階層ベイズ法
5. 統計的検定
6. 評価法
7. 実世界

か弱い人間が苛酷な実世界を生き抜くために考え出した方法です。最近ではパワーアップして機械学習に進化し産業や社会のありかたを変えろとまで言われています。役にたつから学ぶのではなく生きるために学ぶ必要があるのです。



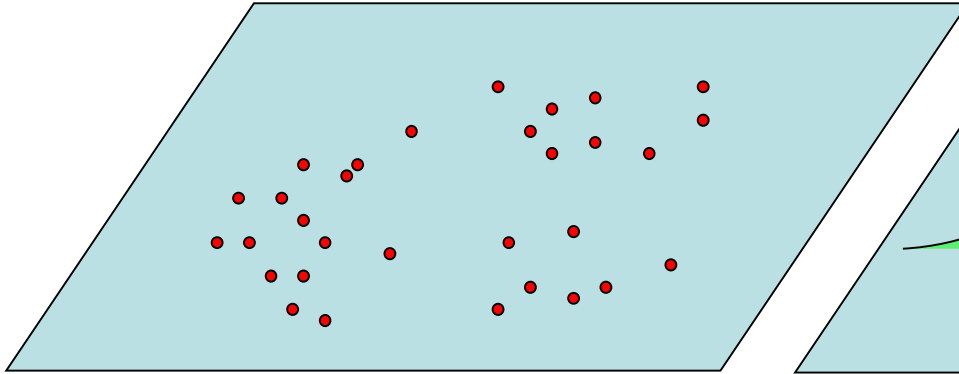
回帰・判別分析 = 教師あり学習



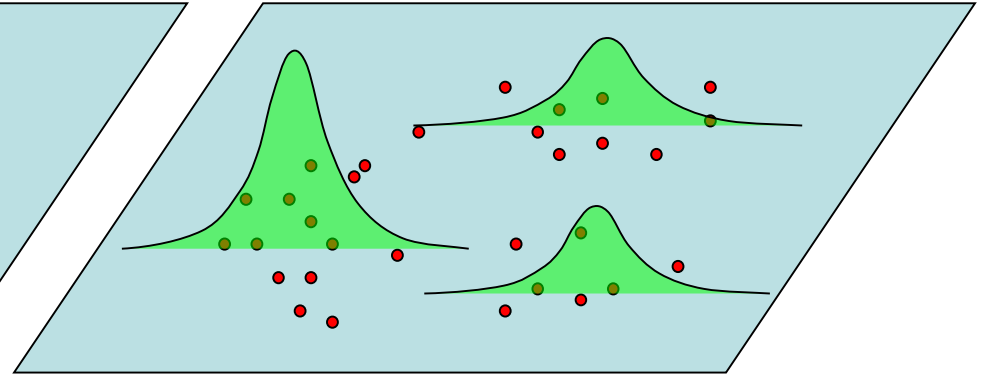
確率変数 (X, Y) のデータが得られたとき、 X から Y への条件つき確率を推測する方法が**回帰分析・判別分析**である。認識・予測など応用が広い。簡単な問題なら線形回帰で解けることも多い。近年、脚光を浴びている深層学習やサポートベクタマシンは、この道の奥にある。

因子・主成分・クラスタ分析 = 教師なし学習

データ



構造



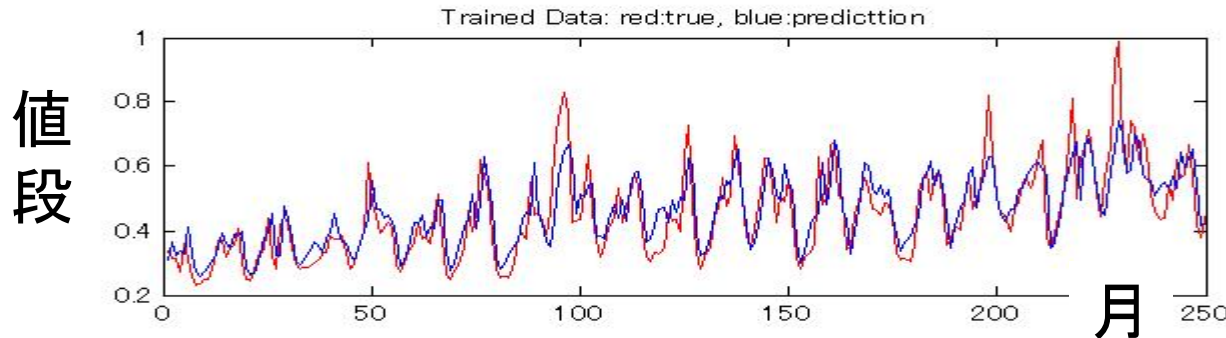
確率変数 X のデータが得られたとき、(X に対するラベル Y が与えられていなくても)、 X の構造や潜在変数を取り出す方法が**因子分析・主成分分析・クラスタ分析**である。

構造や潜在変数を取り出す数理的な技法だけでなく、得られた構造や潜在変数が何を意味するかを考察する力量が必要になる。データサイエンティストへの第一歩。またトピックモデルなどもこの領域に属する。

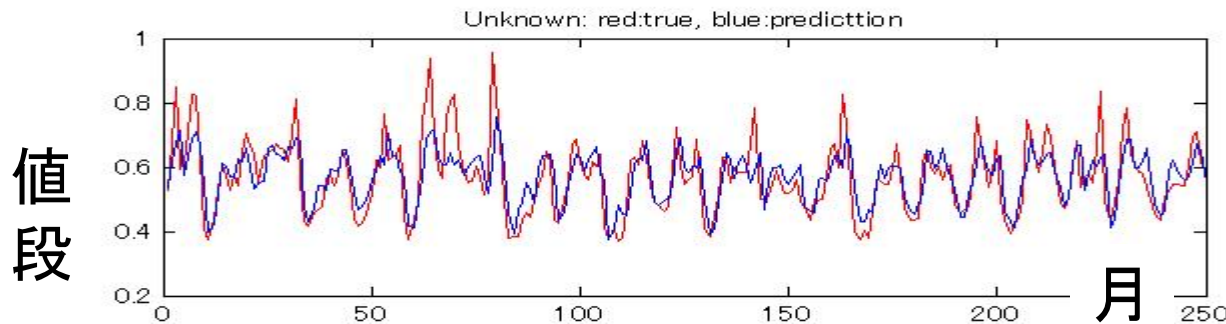
時系列予測



例: 1970年1月から2013年12月までの白菜の値段
「政府統計の総合窓口」のデータを使用しています。
<http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do>



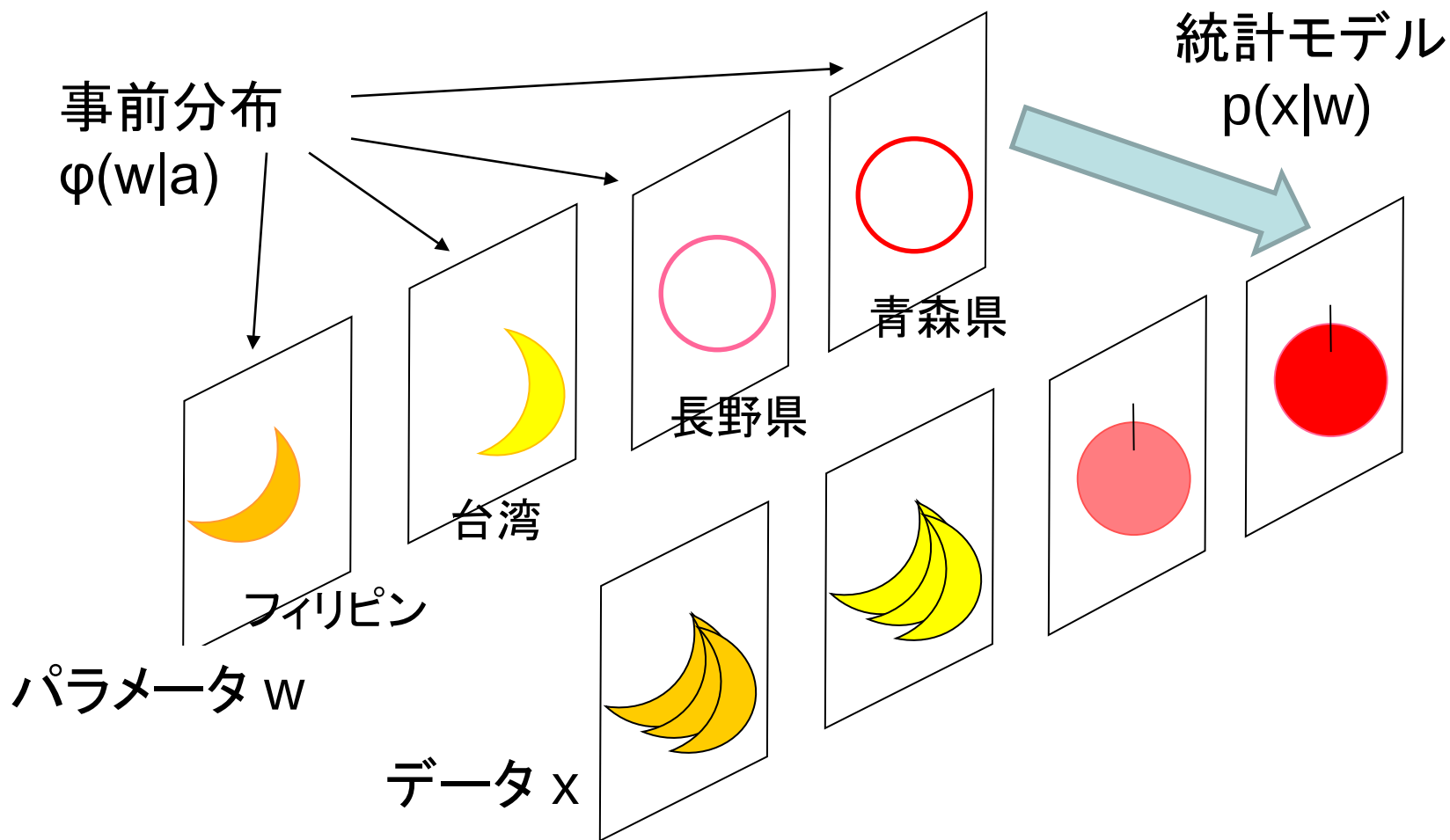
学習データ
赤: 真
青: 学習結果



テストデータ
赤: 真
青: 予測

過去の時系列から未来を予測する。回帰分析とよく似ているが $\times$ が時間とともに変動する点が異なる。経済・金融などへの応用は多すぎるほどある。既に人間トレーダーは人工知能に及ばない時代が来ているそうです。

ベイズ法・階層ベイズ法



果物の特徴 x の確率分布を定めるパラメータ w は、果物が取れる地域により似ている点と異なる点がある。事前分布も含めてモデル化し推測しよう。現実のデータ解析で頻繁に現れる構造である。

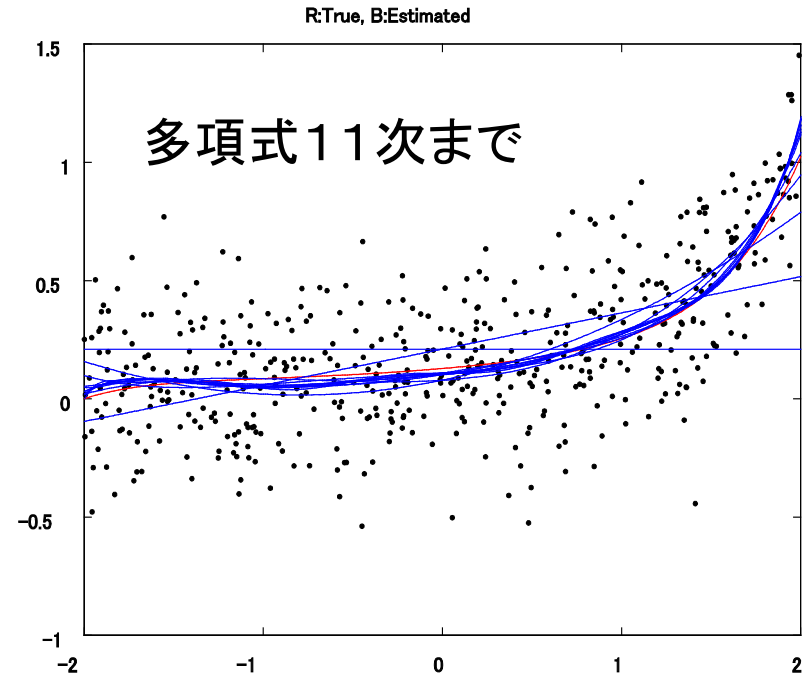
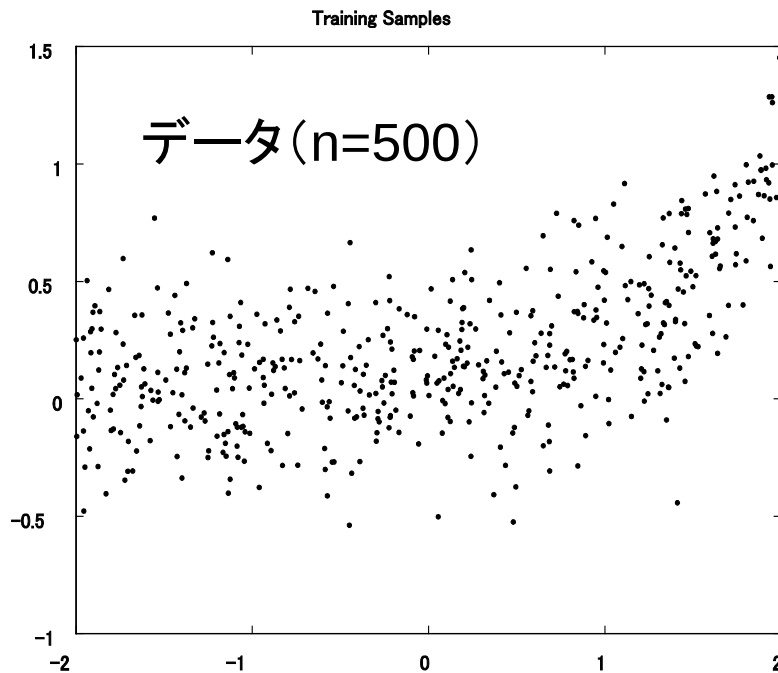
統計的検定

これまでわが社の将棋ソフトは名人との対局で勝率が4割だった。このたび、わが精鋭チームが神経回路網を100億回自己対戦させ強化学習したところ、ついに名人との対戦で3勝2敗になった。わが社のソフトは本当に強くなったのだろうか。



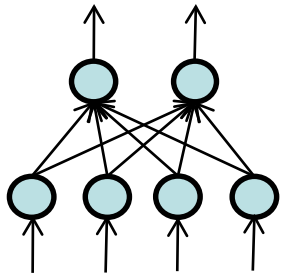
真の確率分布がわかることのない実世界から得られたデータに対して、弱い人間はデータの解析を行なって何らかの推測を行なうがその推測はどの程度に正しいのだろうか。**統計的検定**では、実世界に対して仮説を作りデータの説明可能性を数量的に調べる。

評価法

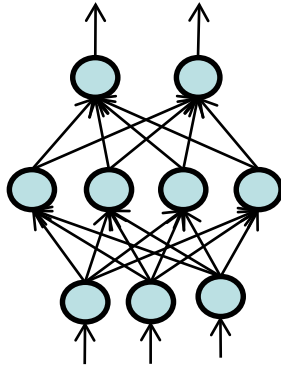


真の確率空間がわかることのない実世界から得られたデータに対して、弱い人間は複数のモデルの候補を作り、複数の推測を行なうことができる。その中に実世界とぴったりと一致するモデルは存在しないが、最も適切なモデルはどれだろうか。そこで「適切さ」とは何に対する適切さだろうか。データ解析が「ただやってみただけ」とは違う客観性を持つにはどうしたらよいのか。数理科学によってもたらされた光である**情報量規準**を学ぼう。

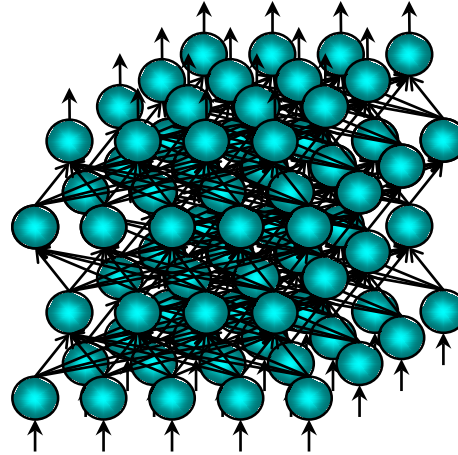
現代から未来へ



1960



1985



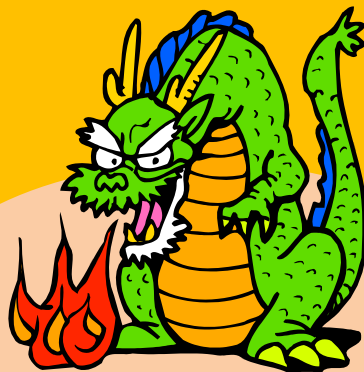
2015



2040

この講義では「データ解析」の初歩なことを学びます。今日ではデータの次元も個数も大きくなり、統計モデルも高度化しています。「機械学習」と呼ばれる分野では極めて大きな複雑さを持つモデルが使われます。また現代数学の方法が必要とされるようになってきています(例: 代数幾何、微分幾何、多変数複素関数論、関数解析、関数空間上の確率変数の収束、超関数、ゼータ関数など)。数理科学が時代を切り拓くときがきています。

旅の地図

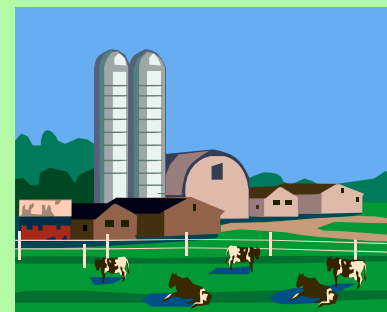


実世界

観測データ

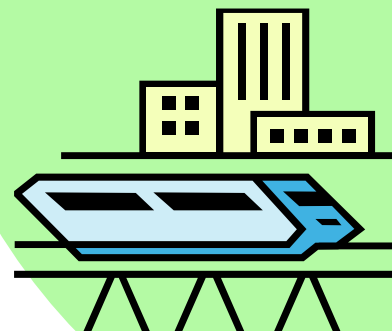


回帰・判別分析



因子・主成分
・クラスタ分析

解析方法



時系列予測

推定と検定



ベイズ法
・階層ベイズ法

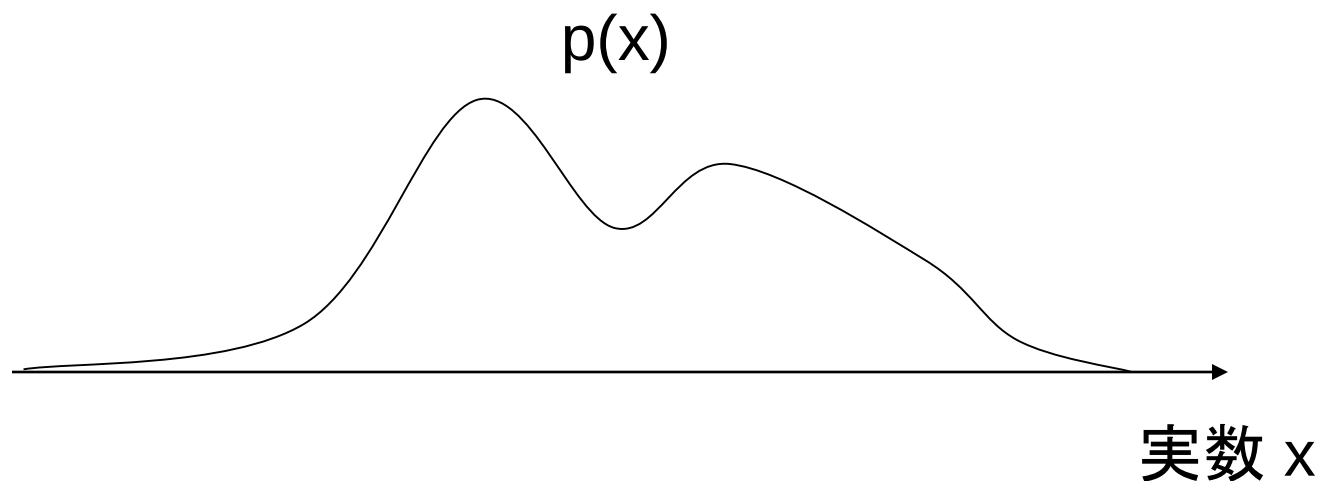


評価法



統計的検定

準備 実数上の確率密度関数

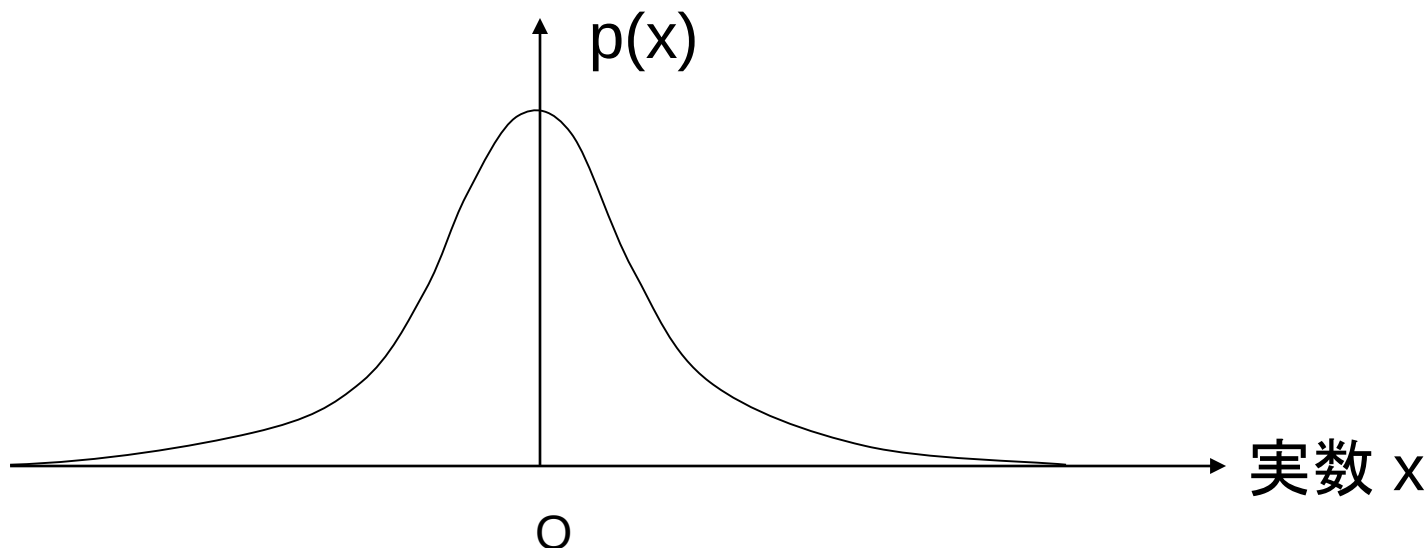


確率密度関数の性質

$$p(x) \geq 0$$

$$\int p(x) dx = 1$$

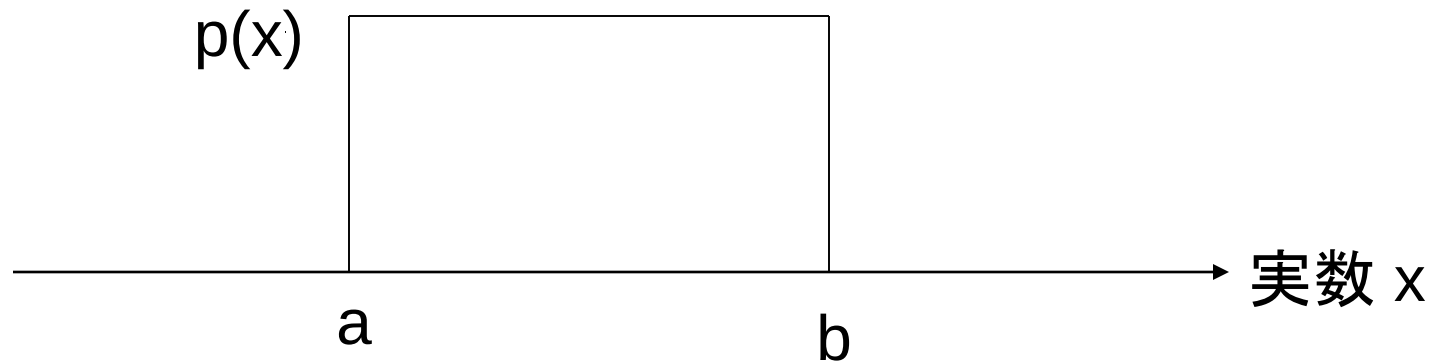
確率密度関数の例(1)



正規分布

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

確率密度関数の例(2)

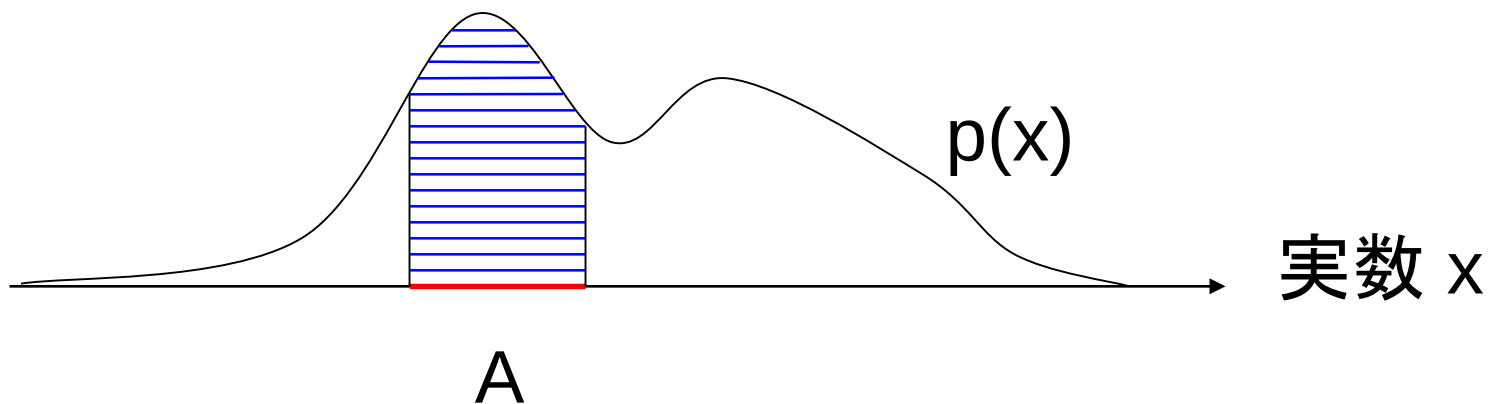


$$p(x) = \begin{cases} 1/(b-a) & (a \leq x \leq b) \\ 0 & (\text{それ以外}) \end{cases}$$

区間 $[a, b]$ の一様分布という

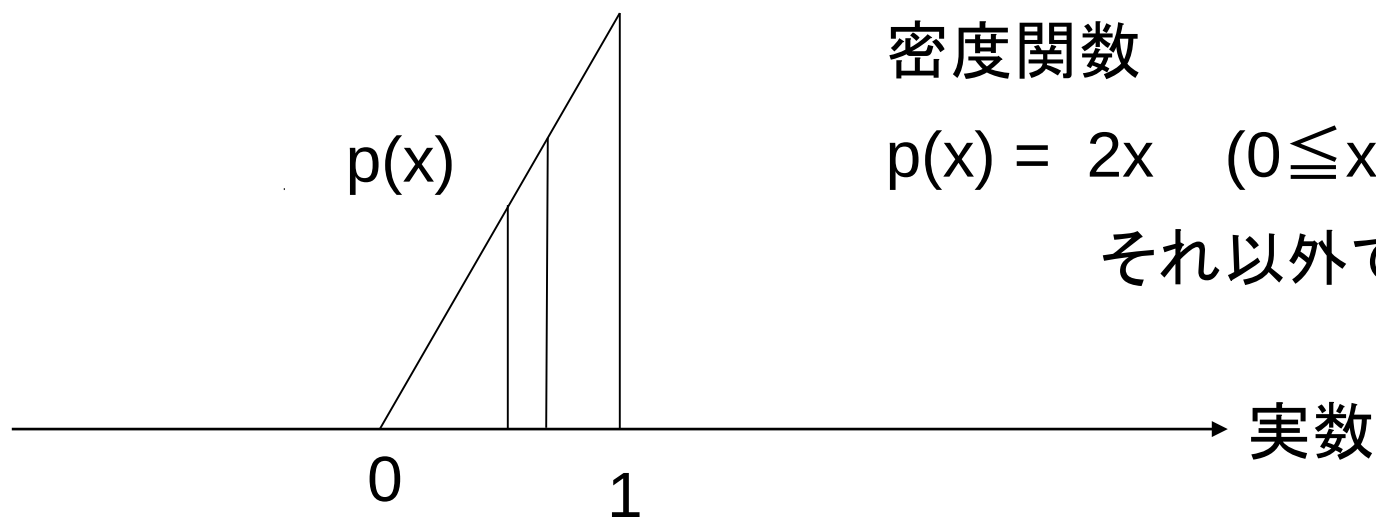
準備 集合の確率

集合Aの確率



$$\text{「A の確率」} = \int_A p(x) dx$$

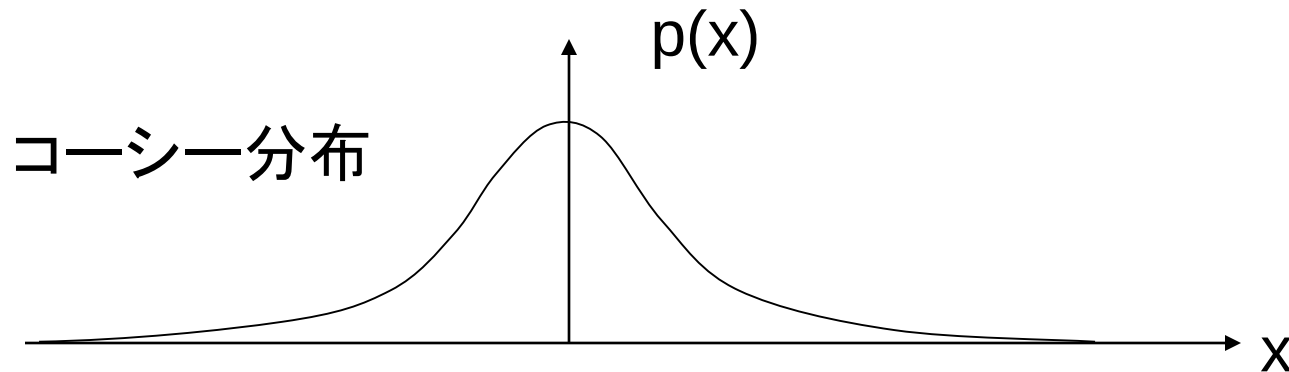
例



集合 $\{ 0.5 \leq x \leq 0.7 \}$ の確率 P は

$$P = \int_{0.5}^{0.7} 2x \, dx = 0.24$$

問題1

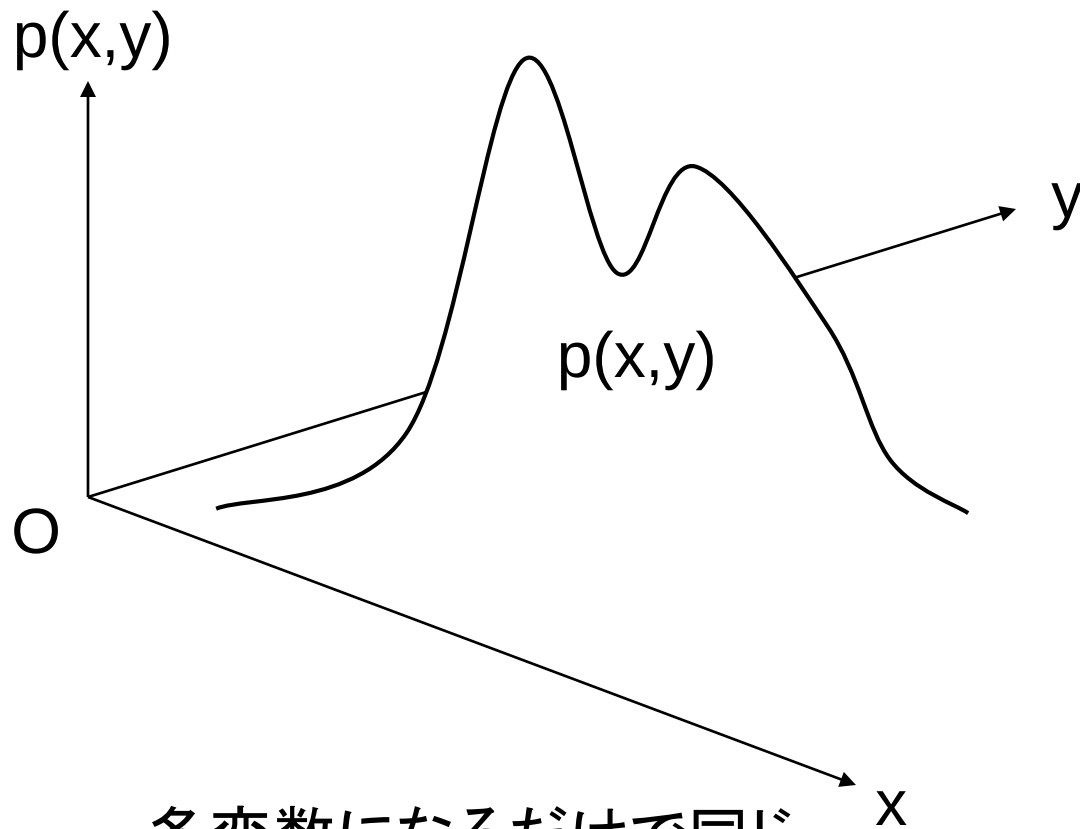


確率変数 X が次の密度関数を持つとする。

$$p(x) = \frac{A}{1+x^2}$$

- (1) 定数 A を求めよ。
- (2) $0 \leq X \leq \sqrt{3}$ となる確率を求めよ。

ユークリッド空間上の確率変数



多次元の場合

R^N 上の中の集合の確率

$$\text{「}A\text{ の確率」} = \int_A p(x_1, \dots, x_N) dx_1 \cdots dx_N$$

$p(x,y)$ は (x,y) の組の起こりやすさを表す

⇒ 同時確率密度関数 という

問題2

$\mathbb{R}^2$ 上の確率変数 X が次の密度関数を持つとする。

$$p(x,y) = \frac{1}{Z} \exp\left(-\frac{x^2 + y^2}{2}\right)$$

- (1) 極座標表示 (r, θ) を用いて定数 Z を求めよ。
- (2) 原点を中心とし半径 a の円に X が入る確率を a で表せ。