統計解析特論 講義メモ $\boxed{\mathbf{1}}$ 基礎事項

講義内容：数理統計学，機械学習について講義する.

- 講義の進めかた
  - 主に板書で基礎事項の解説．時間があれば問題演習も行う.
- 成績
  - レポートと期末試験の結果を総合して評価する.
- 参考文献：適宜，資料を配布する.
  - 機械学習の理論に関して：
    - 金森敬文，“統計的学習理論”，講談社，**2015.**
    - **Mohri, M., et al., “Foundations of Machine Learning”, The MIT Press, 2012.**
  - 確率・統計に関して：
    - 杉山将，“機械学習のための確率と統計”，講談社，**2015.**

主なトピック

- 回帰分析
  - 線形回帰，正則化，交差検証法
  - 高次元回帰，カーネル回帰分析
- 判別分析
  - 学習誤差，予測誤差，ベイズ規則
  - 予測誤差の評価
- カーネル法
  - 再生核ヒルベルト空間
  - サポートベクトルマシン：２値判別，多値判別

機械学習の枠組

観測されたデータ $\Longrightarrow$ 有用な情報を取り出す

- 講義では主に回帰分析・判別分析を扱う：
  データ $(x_1, y_1), \ldots, (x_n, y_n)$ から x と y の間の関係を推定する.
- その他の問題設定
  - 次元削減：高次元データを，情報量を保ちつつ低次元に圧縮.
  - クラスタリング：データをいくつかのグループに分ける.

統計的データ解析と確率論

- 観測データは複雑（ノイズの影響など）
- 確率的なモデリングが有効

$$\text{モデリング} = [\text{確定した構造}] + [\text{ランダムな構造}]$$

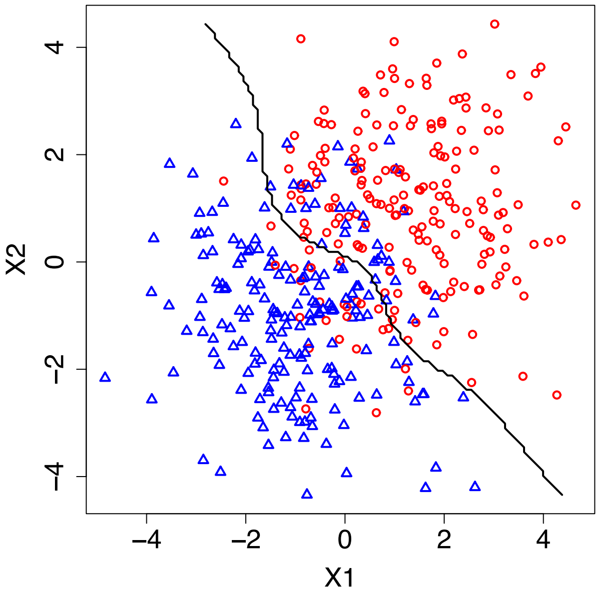
- 確率論を基礎にして，データを解析する
  - ただし本講義では，厳密な測度論的な取り扱いはしない

判別分析

x を入力すると y が出力される状況：

$$\text{例：}\quad \text{メール}\, x \longrightarrow \boxed{\textbf{??}} \longrightarrow y \in \{\text{普通のメール, 迷惑メール}\}$$

- データ $(x_1, y_1), \ldots, (x_n, y_n)$ が観測されている.
- 新たな入力 x に対する出力 y を予測



- 典型的な予測法
  - $x_i \in \mathbb{R}^2,\ \ y_i \in \{+1, -1\}$
  - データから $+1$ と -1 の境界を推定.
  - 境界にもとづいて，新たな入力 x に対する y の値を予測

判別分析の例

- スパム（**spam:**迷惑メール）フィルター
  - x：メール（テキストデータ），$y \in \{\text{spam, non-spam}\}$
  - サンプル：沢山の（過去の）スパムメールと普通のメール
  - 将来のメールがスパムメールかどうかを判定して，仕分けする
- 医療診断
  - x：診察結果，y：病気かどうか.
- 音声認識，顔画像認識
  - x：音声データ **or** 画像データ，y：音声，文字，画像のラベル
  - 郵便番号の自動認識，デジカメの顔検出,
  - 脳波パターン $\rightarrow$ 考えていること **(Brain Computer Interface)**
  - その他，ロボティクスなどへの応用

— 確率論の復習 —

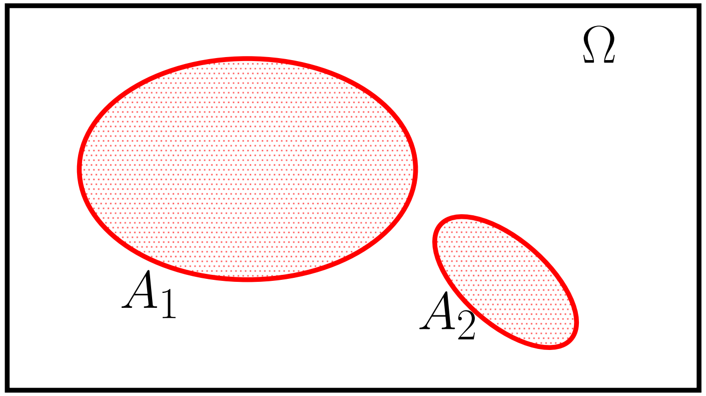
確率の公理

- 確率変数：ランダムな値をとる変数 X（通常は大文字で書く）
- 標本空間 Ω に値をとる確率変数 X に関する確率 $\Pr(\cdot)$ の定義：

1. 集合 $A \subset \Omega$ に対して $0 \leq \Pr(X \in A) \leq 1$.
2. 全集合 Ω の確率は 1. $\Pr(X \in \Omega) = 1$
3. 互いに排反な集合 $A_i,\ i = 1, 2, 3, \ldots$ に対して

$$\Pr(X \in \cup_i A_i) = \sum_i \Pr(X \in A_i).$$

（互いに排反： $i \neq j$ に対して $A_i \cap A_j = \phi$）



（ 簡単のため $\Pr(X \in A)$ を $\Pr(A)$ と表すこともある ）

例 **1** (サイコロの例)**.**

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$
$$X = \text{サイコロの目（確率的に値をとる変数）}$$

$A = \{2, 4, 6\}$ とすると $\Pr(X \in A)$ はサイコロを振って偶数の目がでる確率．公平なサイコロなら $\Pr(X \in A) = 1/2$.

例 **2.** 確率変数を大文字，実現値を小文字で書くことが多い．

$$\text{確率変数 } X \text{ が 値} z \text{ をとる確率：} \Pr(X = z)$$

確率の計算

公理（だけ）から確認できる

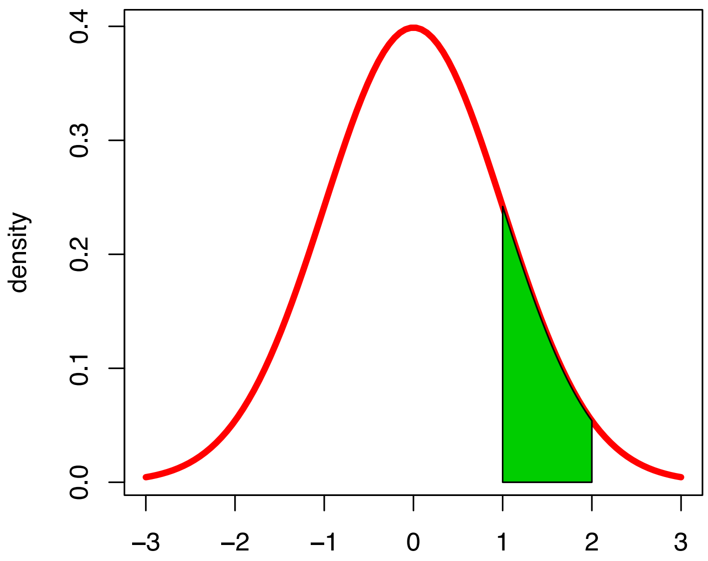
- $\Pr(A) + \Pr(A^c) = 1$,　$A^c : A$ の補集合
- 単調性： $A \subset B \subset \Omega \implies \Pr(A) \leq \Pr(B)$.
  $A \subset B$ のとき $B = A \cup (B \cap A^c)$ （互いに排反）.
  $\therefore \quad \Pr(B) = \Pr(A) + \Pr(B \cap A^c) \geq \Pr(A)$.
- 加法定理： $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

連続値をとる確率変数

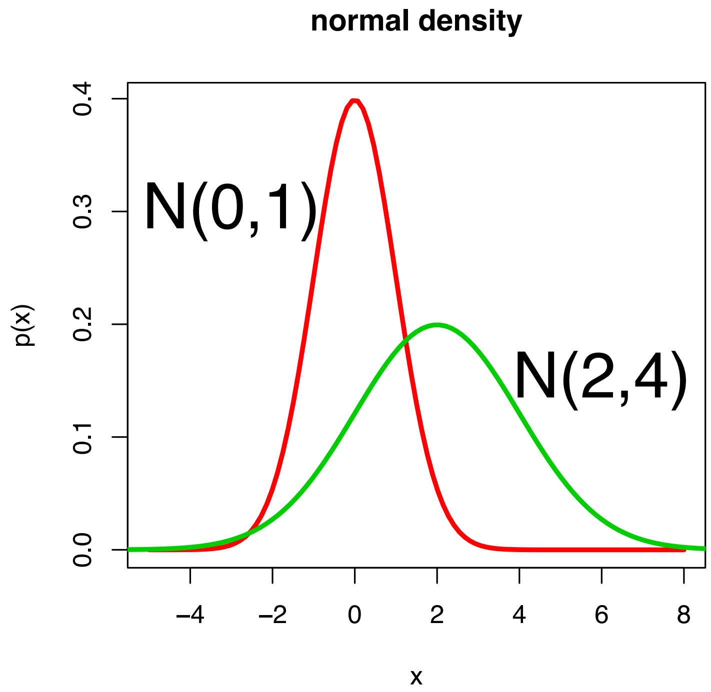
例 3 (正規分布)**.** 確率変数 X が 1次元正規分布にしたがう：

$$\Pr(a \le X \le b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, \qquad (\Omega = \mathbb{R})$$

このとき $X \sim N(\mu, \sigma^2)$ と表す.



$X \sim N(0,1)$, ■ の面積 $= \Pr(1 \le X \le 2)$



確率密度関数

- $\mathbb{R}$ に値をとる確率変数 X： $\Omega = \mathbb{R}$.

$$X \text{ が確率密度 } p(x) \text{ の分布にしたがう}$$

$$\overset{\text{定義}}{\Longleftrightarrow} \quad \Pr(X \in A) = \int_A p(x)dx, \quad A \subset \Omega.$$

- 確率密度関数 $p(x)$ の性質：

$$\text{(i)} \quad \forall x \in \Omega, \quad p(x) \geq 0 \qquad \text{(ii)} \quad \int_\Omega p(x)dx = 1$$

（確率密度関数を確率密度，密度関数，密度と言うこともある）

- サイコロのような離散値をとる確率変数の場合：
積分を和にする．$p(x) = 1/6 \ (x = 1, \ldots, 6)$，$\Pr(X \in A) = \sum_{x \in A} p(x)$.

（離散のとき $p(x)$ を確率関数とよぶこともある）

期待値・分散

X の確率密度を $p(x)$ とする.

- $\Omega = \mathbb{R}$ のとき X の期待値： X がとり得る値の真ん中.

$$E[X] \overset{\text{定義}}{=} \int_{\Omega} x\, p(x) dx \qquad \textbf{(} E(X) \text{ と書くこともある} \textbf{)}$$

離散確率変数のときには $E[X] = \sum_{x \in \Omega} x\, p(x)$ **(**$p(x)$ は確率関数**)**

- 関数 g に対して $g(X)$ の期待値は $E[g(X)] = \int_{\Omega} g(x) p(x) dx$

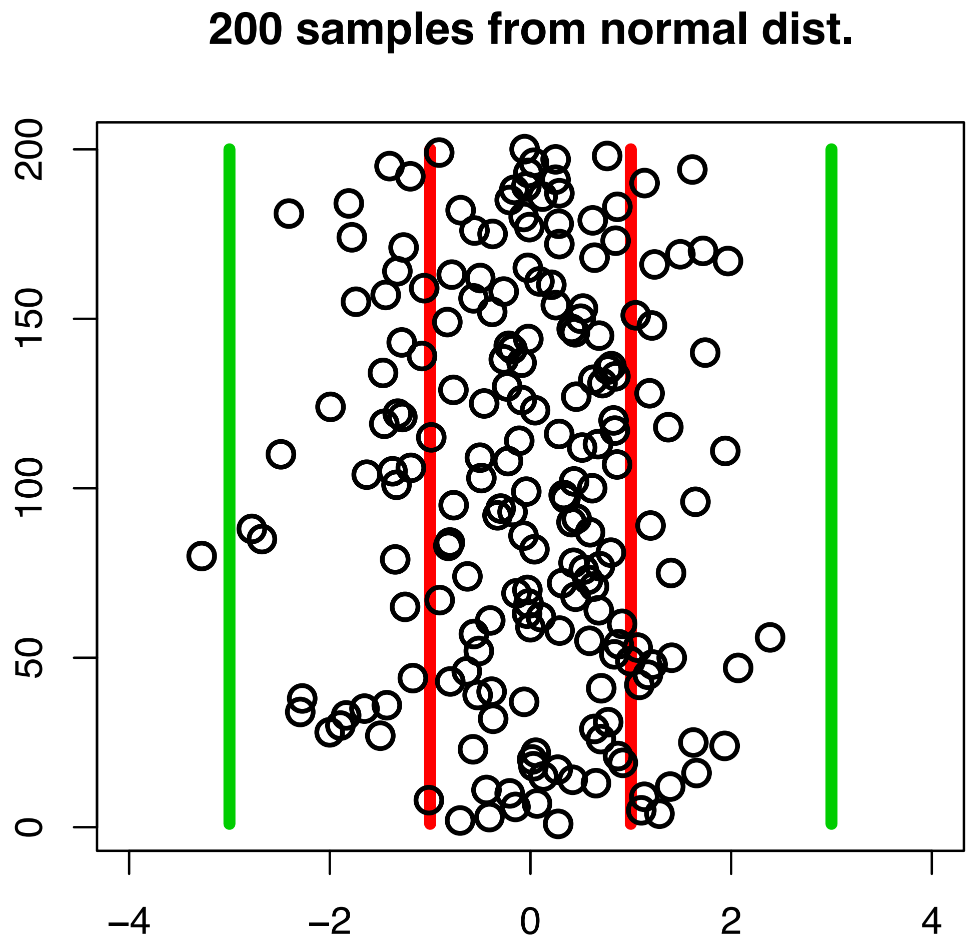
- $\Omega = \mathbb{R}$ のとき X の分散： X のバラツキの大きさ

$$V[X] \overset{\text{定義}}{=} E[(X - E[X])^2] = \int_{\Omega} (x - E[X])^2 p(x) dx$$

  期待値からのズレ $X - E[X]$ の大きさを２乗で測っている.
  ($V(X)$ と書くこともある)
- $a, b \in \mathbb{R}$ のとき，以下の等式が成り立つ

$$E[aX + b] = aE[X] + b, \qquad V[aX + b] = a^2 V[X]$$

例 **4.** $X \sim N(\mu, \sigma^2)$ のとき $E[X] = \mu$, $V[X] = \sigma^2$ が成立.
X が $E[X] \pm \sqrt{V[X]}$ の範囲に値を取る確率 $\cong 0.682$



期待値 0, 分散 1 の正規分布 $N(0,1)$ からのサンプル

$$\Pr(|X| \leq E[X] + \sqrt{V[X]}) \cong 0.682,$$

$$\Pr(|X| \leq E[X] + 2\sqrt{V[X]}) \cong 0.954$$

多次元の確率分布

- ２つ以上の確率変数の関係を調べることは重要
  - 医療検査の結果と病気にかかっているかどうかの関係
  - **A** 社の株価と **B** 社の株価の関係
- ２つの確率変数 X, Y が密度関数 $p(x, y)$ の 同時確率分布にしたがう：

$$\Pr(a \leq X \leq b,\ c \leq Y \leq d) = \int_a^b \int_c^d p(x, y) dy dx$$

$p(x, y)$：確率密度関数.

$$p(x, y) \geq 0, \qquad \iint_\Omega p(x, y) dy dx = 1$$

- 周辺確率密度関数

$$p_1(x) = \int_{\mathbb{R}} p(x,y)dy, \quad p_2(y) = \int_{\mathbb{R}} p(x,y)dx$$

$p_1(x)$ は (X,Y) の X だけを見たときの密度関数.

$\mathrm{Pr}(a \le X \le b) = \int_a^b dx \int_{\mathbb{R}} dy\, p(x,y) = \int_a^b p_1(x)dx$

- 期待値

$$E[X] = \iint x\, p(x,y)\, dydx = \int x\, p_1(x)\, dx$$

$$E[Y] = \iint y\, p(x,y)\, dydx = \int y\, p_2(y)\, dy$$

- $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ の期待値： $E[Z] = \begin{pmatrix} E[X] \\ E[Y] \end{pmatrix}$

独立性

X, Y の密度関数 $p(x, y)$

- X と Y は 独立：$\boxed{p(x,y) = p_1(x)p_2(y)}$ が成り立つこと

  （同時密度関数が周辺密度関数の積になる）

- 離散確率変数のときも同様.
  $p(x,y) = p_1(x)p_2(y)$ が成立

例 **5.** 2つのサイコロの目を (X, Y) とする．通常 X, Y は独立と仮定する.

$$\Pr(X = 1, Y = 2) = \Pr(X = 1) \times \Pr(Y = 2) = 1/6 \times 1/6 = 1/36.$$

共分散

- ２つの確率変数 X, Y の関連の強さを測る
- X, Y の共分散

$$
\begin{aligned}
\mathrm{Cov}(X, Y) &\overset{\text{定義}}{=} E[(X - E[X])(Y - E[Y])] \\
&= E(XY) - E(X)E(Y)
\end{aligned}
$$

- 分散と共分散の関係

$$
V(X + Y) = V(X) + V(Y) + 2\,\mathrm{Cov}(X, Y)
$$

独立性と共分散

重要：X と Y が独立のときに成り立つ公式

- $\mathrm{Cov}(X, Y) = 0 \iff E[XY] = E[X]E[Y]$
- $V[X+Y] = V[X] + V[Y]$

独立 $\Longrightarrow$ 無相関

逆は一般に成立しない

注意：いつでも（独立でなくても）

$$E[X+Y] = E[X] + E[Y] \quad \text{は成立する.}$$

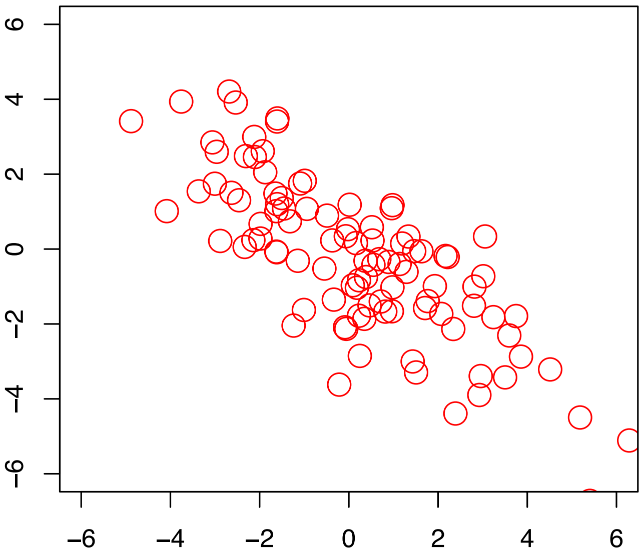
確認の計算：$p(x,y) = p_1(x)p_2(y)$ のとき

$$
\begin{aligned}
E[XY] &= \int xyp(x,y)dxdy \\
&= \int xyp_1(x)p_2(y)dxdy \\
&= \int xp_1(x)dx \int yp_2(y)dy = E(X)E(Y).
\end{aligned}
$$

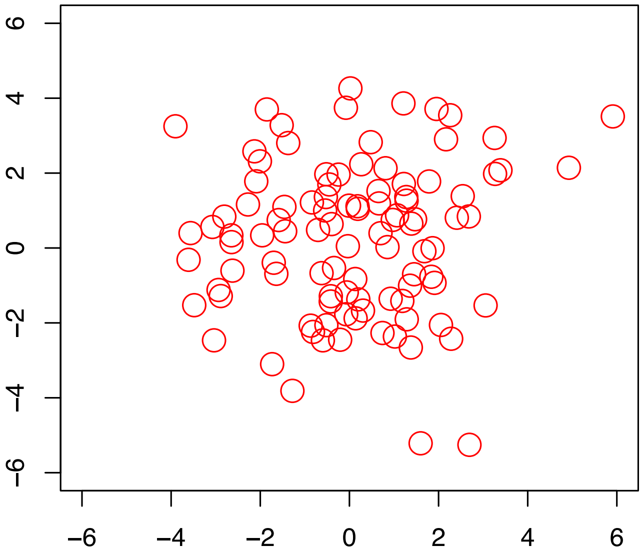
$$
\begin{aligned}
V[X+Y] &= \int (x+y-E(X)-E(Y))^2 p_1(x)p_2(y)dxdy \\
&= \int (x-E(X))^2 p_1(x)p_2(y)dxdy + \int (y-E(Y))^2 p_1(x)p_2(y)dxdy \\
&\quad + 2\int (x-E(X))p_1(x)dx \int (y-E(Y))p_2(y)dy \\
&= V[X]+V[Y].
\end{aligned}
$$

共分散と線形関係

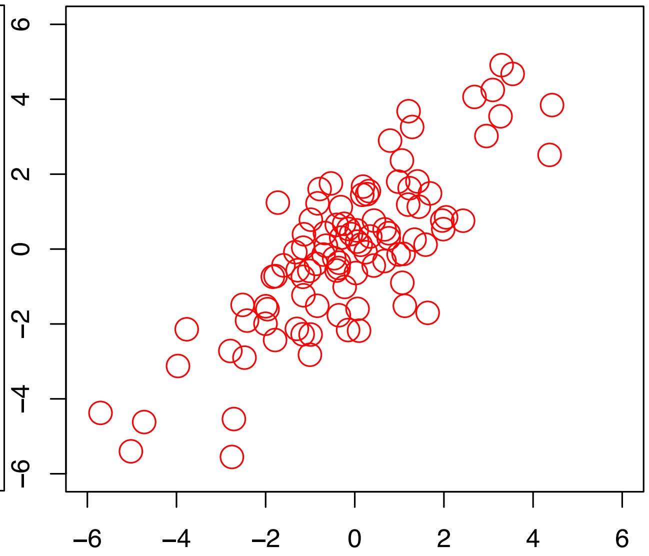
- X と Y に線形な関係はない $\Longrightarrow$ $\mathrm{Cov}(X,Y)=0$ **(無相関)**
- X,Y に線形関係 $\Longrightarrow$ $\mathrm{Cov}(X,Y)>0$ or <0



$\mathrm{Cov}(X,Y)<0$



$\mathrm{Cov}(X,Y)\approx 0$



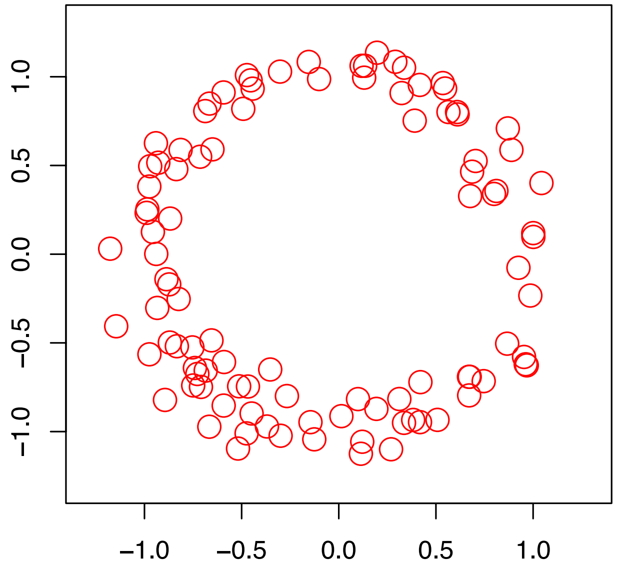
$\mathrm{Cov}(X,Y)>0$

共分散と非線型な関連

共分散は X と Y の線形関係を捉える量.

- $\mathrm{Cov}(X, Y) \approx 0$ だが関連がある場合もある.

(ほぼ) 無相関だが独立でない例



$\mathrm{Cov}(X, Y) \approx 0.03$

多次元確率変数の密度関数

n 個の確率変数：$X_1, X_2, \ldots, X_n$. 集合 $A \subset \mathbb{R}^n$
$(X_1, X_2, \ldots, X_n)$ が集合 A に含まれる確率の計算：

$$P((X_1, \ldots, X_n) \in A) = \int_A p(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

簡単のため多重積分を $\int_A p(x)dx$ と書くこともある

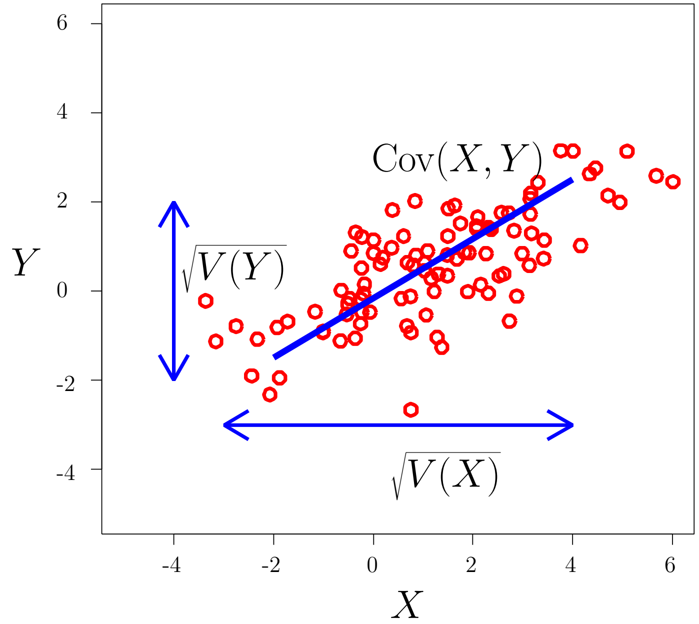
- 密度関数：$p(x_1, \ldots, x_n) \geq 0, \ \int_{\mathbb{R}^n} p(x)dx = 1$.
- 周辺密度：$p_1(x_1) = \int_{\mathbb{R}^{n-1}} p(x_1, \ldots, x_n) dx_2 \cdots dx_n$ など.

多次元確率変数の期待値

- $X = (X_1, \ldots, X_n)^T$ の期待値： $E[X] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{pmatrix} \in \mathbb{R}^n$.
- 分散共分散行列：多次元確率変数のバラツキの傾向をまとめた行列

2 つの確率変数 X と Y の分散共分散行列

$$\overset{\text{定義}}{=} \begin{pmatrix} V(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & V(Y) \end{pmatrix}$$



- 一般に $Z = (X_1, \ldots, X_n)^T$ とすると

$$\text{分散共分散行列 } V[Z] = E[(Z - E[Z])\,(Z - E[Z])^T] \ \in \mathbb{R}^{n \times n}$$

$(V[Z])_{ij}$：X_i, X_j の共分散，対角成分 $(V[Z])_{ii}$ は X_i の分散.

- n 次元確率変数 Z の期待値 $E[Z]$ と分散共分散行列 $V[Z]$．以下が成立.

$$A \in \mathbb{R}^{k \times n},\ \boldsymbol{b} \in \mathbb{R}^k \text{ に対して}$$

$$E[AZ + \boldsymbol{b}] = AE[Z] + \boldsymbol{b}, \quad V[AZ + \boldsymbol{b}] = AV[Z]A^T$$

独立性・独立同一分布

n 個の確率変数：$X_1, X_2, \ldots, X_n$.

- $X_1, \ldots, X_n$ が独立 $\Longleftrightarrow$ 同時密度関数が積に分解

$$p(x_1, \ldots, x_n) = q_1(x_1)q_2(x_2)\cdots q_n(x_n)$$

note: $X = (X_1, \ldots, X_n)$ の分散共分散行列は対角行列.
例：サイコロ X_1 とコイン X_2 を別々に振る.

- $X_1, \ldots, X_n$ が独立に同一の分布にしたがう：

$$p(x_1, \ldots, x_n) = q(x_1)q(x_2)\cdots q(x_n),$$
$$(q = q_1 = \cdots = q_n)$$

note: $X = (X_1, \ldots, X_n)$ の分散共分散行列は単位行列の定数倍.
例：同じサイコロを２回振る. １回目 X_1, ２回目 X_2.

- $X_1, \ldots, X_n$ が独立に同一の分布 P にしたがうとき：

$$X_1, \ldots, X_n \sim_{i.i.d.} P \quad \text{と書く}$$

このとき，$X_1, \ldots, X_n$ の期待値や分散は全て等しい：

$$E[X_1] = \cdots = E[X_n], \qquad V[X_1] = \cdots = V[X_n].$$

例 **6.** $X_1, \ldots, X_n \sim_{i.i.d.} N(0,1)$ のとき,

$$p(x_1, \ldots, x_n) = \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} \times \cdots \times \frac{1}{\sqrt{2\pi}} e^{-x_n^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{i=1}^n x_i^2/2}$$

正規分布と独立性

- a, b $(a \neq 0)$ を定数とすると

$$X \sim N(\mu, \sigma^2) \implies aX + b \sim N(a\mu + b, a^2\sigma^2)$$

- X, Y は独立で $X \sim N(\mu_1, \sigma_1^2),\ Y \sim N(\mu_2, \sigma_2^2)$ のとき

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

したがって $X_1, \ldots, X_n \sim_{i.i.d.} N(\mu, \sigma^2)$ のとき

$$\frac{1}{n}\sum_{i=1}^{n} X_i \ \sim \ N\left(\mu, \frac{\sigma^2}{n}\right)$$

n が大きくなると $\displaystyle\frac{1}{n}\sum_{i=1}^{n} X_i$ の分散が小さくなる

note: 上の関係式は積率母関数を用いて証明できる.

多変量正規分布

$\Omega = \mathbb{R}^d$ に値をとる確率変数の密度関数が

$$\phi(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right\}$$

のとき　$X \sim N_d(\boldsymbol{\mu},\Sigma)$ とかく　$(\mu \in \mathbb{R}^d,\ \ \Sigma \in \mathbb{R}^{d\times d})$

- $X \sim N_d(\boldsymbol{\mu},\Sigma)$ のとき以下が成り立つ：

$$E[X] = \boldsymbol{\mu}, \quad V[X] = \Sigma$$

分散共分散行列の性質

$X \in \mathbb{R}^d$ の分散共分散行列 を $\Sigma \in \mathbb{R}^{d\times d}$ とする

- Σ は対称行列なので，直交行列で対角化可能

$$\Sigma = Q\Lambda Q^T$$

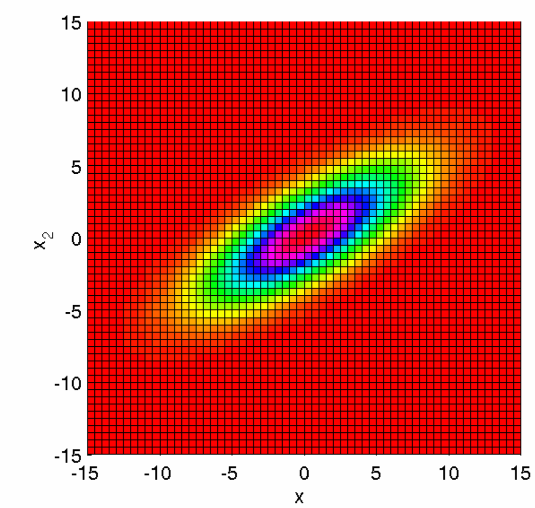
$$Q^TQ = QQ^T = I,\ \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$$

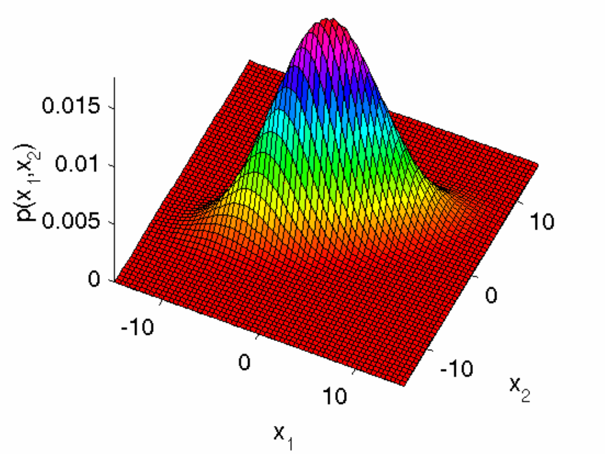
- Σ は非負定値行列： $\boldsymbol{x}^T\Sigma\boldsymbol{x} \geq 0, \quad \forall \boldsymbol{x} \in \mathbb{R}^d$
- Σ の固有値はすべて非負：$\lambda_1, \ldots, \lambda_d \geq 0$.

分散共分散行列 Σ の固有値と固有ベクトルを計算
$\Longrightarrow$ データの散らばり方が分かる

2次元正規分布のプロット

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 20 & 10 \\ 10 & 9 \end{pmatrix}$$





等高線：$(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c$

Σ の固有値，固有ベクトル：

$\lambda_1 \cong 26.0,\ \boldsymbol{q}_1 \cong (0.86,\ 0.51), \quad \lambda_2 \cong 3.1,\ \boldsymbol{q}_2 \cong (-0.51,\ 0.86)$

固有ベクトル：楕円形の等高線の軸方向に対応.

多変量正規分布の性質

- $\Omega = \mathbb{R}^d$, $A \in \mathbb{R}^{k\times d}$ $(\mathrm{rank}(A) = k)$, $\boldsymbol{b} \in \mathbb{R}^k$ のとき

$$X \sim N_d(\boldsymbol{\mu}, \Sigma) \implies AX + \boldsymbol{b} \sim N_k(A\boldsymbol{\mu} + \boldsymbol{b},\ A\Sigma A^\top)$$

- $X \sim N_d(\boldsymbol{\mu}_1, \Sigma_1)$, $Y \sim N_d(\boldsymbol{\mu}_2, \Sigma_2)$ とする．X, Y が独立のとき

$$X + Y \sim N_d(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2,\ \Sigma_1 + \Sigma_2)$$

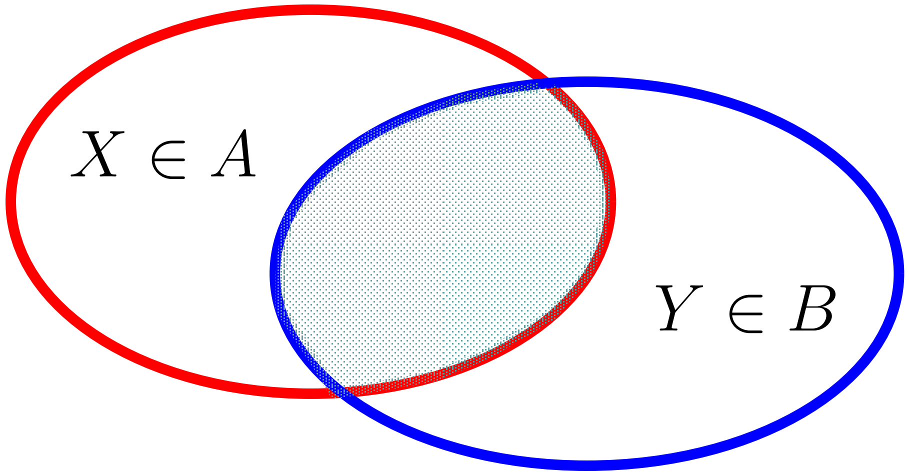
- $X_1, \ldots, X_n \sim_{i.i.d.} N(0, \sigma^2)$ のとき $X = (X_1, \ldots, X_n)^T$ の分布は

$$X \sim N_n(\mathbf{0}, \sigma^2 I_n), \qquad (I_n \text{ は } n \text{ 次元単位行列})$$

条件付き確率・条件付き確率密度

- X, Y に関する確率 $\Pr(X \in A, Y \in B)$ が与えられているとき「$X \in A$ の条件のもとで $Y \in B$」となる確率：

$$\Pr(Y \in B \mid X \in A) = \frac{\Pr(X \in A, Y \in B)}{\Pr(X \in A)}$$



- 密度 $p(x,y)$ のもとで，x が与えられたときの y の条件付き密度

$$p(y|x) = \frac{p(x,y)}{\displaystyle\int_{\mathbb{R}} p(x,y)dy} = \frac{p(x,y)}{p_1(x)}. \quad (p_1(x) : x \text{ の周辺密度})$$

$$^{\forall}x, \quad p(y|x) \geq 0, \quad \int p(y|x)dy = 1.$$

「$X \in [x, x+dx]$ の条件下で $Y \in [y, y+dy]$ となる確率」

$$= \frac{\Pr(X \in [x, x+dx],\ Y \in [y, y+dy])}{\Pr(X \in [x, x+dx])}$$

$$= \frac{p(x,y)dxdy}{p_1(x)dx}$$

$$= p(y|x)dy$$

ベイズの定理

$$\Pr(X \in A | Y \in B) = \frac{\Pr(Y \in B | X \in A)\Pr(X \in A)}{\Pr(Y \in B)}$$

証明： $\Pr(X \in A|Y \in B)\Pr(Y \in B) = \Pr(X \in A, Y \in B)$

$$= \Pr(Y \in B | X \in A)\Pr(X \in A)$$

解釈：X を原因，Y を結果と考えると・・・

- $\Pr(Y|X)$： 原因 X から結果 Y への関係
- $\Pr(X|Y)$： 結果 Y を見て，原因 X について推論

条件付き確率密度：混合分布の例

X は $\mathbb{R}$ 上の確率変数, Y は $\{0, 1\}$ 上の確率変数.

$$\Pr(Y = 0) = q, \quad \Pr(Y = 1) = 1 - q$$

$$X\text{ の条件付き密度：}\quad p(x|Y = 0) = p_0(x), \quad p(x|Y = 1) = p_1(x)$$

このとき X の周辺密度は $p(x) = q \cdot p_0(x) + (1 - q) \cdot p_1(x)$.

note: $p(x) = \int p(x|y)p(y)dy$. この例では Y は離散なので,
積分ではなく和になる.

ベイズの定理より

$$\begin{aligned}\Pr(Y=1|X\in[x,x+dx]) &= \frac{p(X\in[x,x+dx]\,|\,Y=1)\Pr(Y=1)}{p(X\in[x,x+dx])}\\ &= \frac{(1-q)p_1(x)dx}{\{qp_0(x)+(1-q)p_1(x)\}dx}\\ &= \frac{1}{\dfrac{q}{1-q}\cdot\dfrac{p_0(x)}{p_1(x)}+1}\end{aligned}$$

簡単のため $\Pr(Y=1|X\in[x,x+dx])$ を $\Pr(Y=1|x)$ と書く.

$r(x)=\dfrac{q}{1-q}\cdot\dfrac{p_0(x)}{p_1(x)}$ とおくと

$$\Pr(Y=1\,|\,x)=\frac{1}{r(x)+1},\qquad \Pr(Y=0\,|\,x)=\frac{r(x)}{r(x)+1}$$

確率不等式 (必要に応じて追加資料を配布する)

- **Jensen's ineqality :** X を $\mathbb{R}^k$ に値をとる確率変数, $g : \mathbb{R}^k \longrightarrow \mathbb{R}$ を凸関数とすると

$$E[g(X)] \;\geq\; g(E[X])$$

- **Chebyshev's inequality :** $\mathbb{R}$ に値をとる確率変数 X に対して

$$P(|X - E[X]| \geq \varepsilon) \leq \frac{V[X]}{\varepsilon^2} \quad \text{(大数の法則の証明に用いる)}$$

- **Hoeffding's inequality :** $X_1, \ldots, X_n \sim_{i.i.d.} P$.
$\mu = E[X_i],\ \Pr(a \leq X_i \leq b) = 1$ のとき

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \varepsilon\right) \;\leq\; 2e^{-2n\varepsilon^2/(b-a)^2}$$

— 復習：統計の基礎 —

統計モデルを用いた推定

- データ X を観測．その確率法則について推論．
- $X \sim p(x)$ とする．統計モデルを設定：

$$\text{統計モデル：}\ \mathcal{P} = \{p(x;\theta) \mid \theta \in \Theta \subset \mathbb{R}^d\},$$

$$\text{仮定：}\ p(x) = p(x;\theta^*) \in \mathcal{P}.$$

- データ X からパラメータ θ^* を推定

尤度原理

統計モデル：$\mathcal{P} = \{p(x;\theta) \mid \theta \in \Theta \subset \mathbb{R}^k\}$
データ $X \sim p(x;\theta)$．X からパラメータ θ を推定

- 尤度原理：「典型的なデータ」が観測された，と考える．
- 最尤推定量：X の実現値を x とするとき

$$p(x;\widehat{\theta}) \;=\; \max_{\theta\in\Theta}\, p(x;\theta)$$

となる $\widehat{\theta} \in \Theta$ を最尤推定量という．
– 観測データを最も出現させやすい確率を推定量とする

独立同一分布にしたがうデータ

- データ： $X_1, \dots, X_n \sim_{i.i.d.} p(x;\theta^*)$.
- 真のパラメータ θ^* の最尤推定量：

$$\max_{\theta} \underbrace{\prod_{i=1}^{n} p(X_i;\theta)}_{\text{尤度関数}} \longrightarrow \widehat{\theta}_n$$

$$\iff \max_{\theta} \underbrace{\sum_{i=1}^{n} \log p(X_i;\theta)}_{\text{対数尤度関数}} \longrightarrow \widehat{\theta}_n$$

多変量正規分布パラメータの最尤推定量

$$X_1, \ldots, X_n \sim_{i.i.d.} N_d(\boldsymbol{\mu}, \Sigma)$$

- $N_d(\boldsymbol{\mu}, \Sigma)$ の確率密度：

$$\phi(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

$(\boldsymbol{\mu}, \Sigma)$ の代わりに $(\boldsymbol{\mu}, \Sigma^{-1})$ をパラメータと考える：

対数尤度 $$\ell(\boldsymbol{\mu}, \Sigma^{-1}) = -\frac{1}{2}\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^T \Sigma^{-1}(X_i - \boldsymbol{\mu}) + \frac{n}{2}\log|\Sigma^{-1}| - \frac{d}{2}\log 2\pi$$

最尤推定量の計算 (極値を求める)

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}} = \Sigma^{-1} \sum_{i=1}^{n} (X_i - \boldsymbol{\mu})$$

$$\begin{aligned}
\frac{\partial \ell}{\partial (\Sigma^{-1})_{k\ell}} &= -\frac{1}{2} \sum_{i=1}^{n} (X_i - \boldsymbol{\mu}_k)(X_i - \boldsymbol{\mu}_\ell) + \frac{n}{2} \frac{\partial}{\partial (\Sigma^{-1})_{k\ell}} \log |\Sigma^{-1}| \\
&= -\frac{1}{2} \sum_{i=1}^{n} (X_i - \boldsymbol{\mu}_k)(X_i - \boldsymbol{\mu}_\ell) + \frac{n}{2} \frac{1}{|\Sigma^{-1}|} \frac{\partial |\Sigma^{-1}|}{\partial (\Sigma^{-1})_{k\ell}} \\
&= -\frac{1}{2} \sum_{i=1}^{n} (X_i - \boldsymbol{\mu}_k)(X_i - \boldsymbol{\mu}_\ell) + \frac{n}{2} \frac{\widetilde{\Sigma^{-1}}_{k\ell}}{|\Sigma^{-1}|} \qquad (\widetilde{\Sigma^{-1}}_{k\ell}\text{：小行列式}) \\
&= -\frac{1}{2} \sum_{i=1}^{n} (X_i - \boldsymbol{\mu}_k)(X_i - \boldsymbol{\mu}_\ell) + \frac{n}{2} \Sigma_{\ell k}
\end{aligned}$$

極値問題を解いて

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\boldsymbol{\mu}})(X_i - \widehat{\boldsymbol{\mu}})^T$$

- 期待値は不偏推定量，分散共分散行列は不偏推定量でない.
- $\ell(\boldsymbol{\mu}, \Sigma^{-1})$ の代わりに $\ell(\boldsymbol{\mu}, \Sigma)$ について計算しても同じ結果（Σ の変換に関する **Jacobian** 因子が出てくるだけ）
  $\longrightarrow$ 最尤推定量の共変性

例 7 (血液型：表現型の人数から遺伝子型の確率を推定).

- 血液型： **A, B, AB, O.**
- 対立遺伝子： **a, b, o.**

  a,b,o の確率： $\theta_a,\ \theta_b,\ \theta_o$

$$\theta_a+\theta_b+\theta_o=1,\quad \theta_a,\theta_b,\theta_o>0.$$

表現型	遺伝子型	人数
A	**aa，ao，oa**	n_A
B	**bb，bo，ob**	n_B
AB	**ab, ba**	n_{AB}
O	**oo**	n_O

$$\Pr(A)=\theta_a^2+2\theta_a\theta_o,\quad \Pr(B)=\theta_b^2+2\theta_b\theta_o$$
$$\Pr(AB)=2\theta_a\theta_b,\qquad \Pr(O)=\theta_o^2$$

対数尤度：
$$\begin{aligned}&\log\Pr(A)^{n_A}\Pr(B)^{n_B}\Pr(AB)^{n_{AB}}\Pr(O)^{n_O}\\ &=n_A\log(\theta_a^2+2\theta_a\theta_o)+n_B\log(\theta_b^2+2\theta_b\theta_o)\\ &\quad+n_{AB}\log(2\theta_a\theta_b)+n_O\log(\theta_o^2)\quad\longrightarrow\quad\max_\theta\end{aligned}$$

最尤推定量の統計的性質

最尤推定量は良い性質をもっている.

$$X_1, \ldots, X_n \sim_{i.i.d.} p(x;\theta^*) \quad \longrightarrow \quad \text{最尤推定量} \quad \widehat{\theta}_n$$

適当な正則条件のもとで以下が成立；

- **(統計的一致性)** $\widehat{\theta}_n \xrightarrow{p} \theta^* \quad (n \to \infty)$：
  データが十分多ければ，ほぼ正しい推定が可能.
- **(有効推定量)** 他の推定量と比べて誤差（分散）が小さい

— 補足事項 —

- 大数の法則，中心極限定理

漸近理論１：大数の法則

$X_1, \ldots, X_n \sim_{i.i.d.} P$ として $E(X_i) = \mu$ とおく.

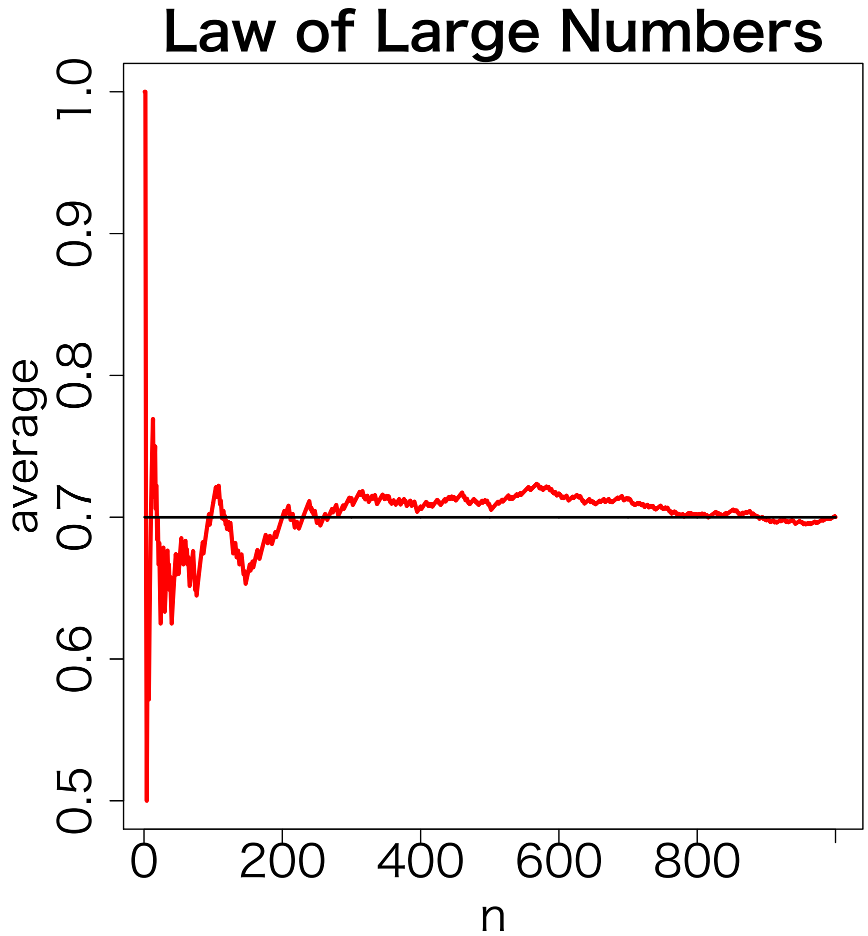
- 大数の法則：$\bar{X}_n \overset{\text{定義}}{=} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ とすると

$$\forall \varepsilon > 0, \lim_{n\to\infty} \Pr(|\bar{X}_n - \mu| > \varepsilon) = 0.$$

  - 意味：n が十分大きいと，高い確率で $\bar{X}_n$ は μ にほぼ等しい.
  - これを $\bar{X}_n \xrightarrow{p} \mu$ と書き「$\bar{X}_n$ が μ に確率収束する」という.
- $f(x)$ を連続関数とするとき，普通の極限と類似の関係：

$$\bar{X}_n \xrightarrow{p} \mu \quad \text{ならば} \quad f(\bar{X}_n) \xrightarrow{p} f(\mu) \quad \text{が成り立つ}$$

例 **8** (大数の法則の例)**.** 表の確率が**0.7**のコイン.
n回振って表が出た割合をプロット



例 **9.** コインをn回振る．k回目が表なら $X_k = 1$，裏なら $X_k = 0$とするとn 回のうち $\sum_{k=1}^{n} X_k$ 回表が出ることになる．$X_1, \dots, X_n$ は独立とする．表が出る確率を p とすると $E(X_i) = p$ となり，

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{p} p$$

が成り立つ．また $f(x) = x(1-x)$ とすると

$$f\Big(\frac{1}{n}\sum_{k=1}^{n} X_k\Big) \xrightarrow{p} p(1-p)$$

となる．

したがって，$E(X_i) = p$ を $\bar{X}_n$ で推定でき，$V(X_i) = p(1-p)$ を $f(\bar{X}_n)$ で推定できる．

漸近理論２：中心極限定理

$X_1, \dots, X_n \sim_{i.i.d.} P$ とする

- 中心極限定理：$E(X_i) = \mu,\ V(X_i) = \sigma^2$ のとき

$$Z_n \overset{\text{定義}}{=} \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - \mu}{\sigma} = \sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma}$$ とすると,

$$\lim_{n \to \infty} \Pr(Z_n \le z) = \int_{-\infty}^{z} \phi(x; 0, 1) dx$$ が成り立つ.

- 大数の法則：$\bar{X}_n - \mu \xrightarrow{p} 0$.
  中心極限定理では $\bar{X}_n - \mu$ を $\sqrt{n}(\bar{X}_n - \mu)$ に拡大して，極限の分布を詳しく見ている.

例 10. コインを n 回振る．k 回目が表なら $X_k = 1$，裏なら $X_k = 0$ として，$X_1, \ldots, X_n$ は独立とする．表が出る確率を p とすると，$E(X_i) = p$, $V(X_i) = p(1-p)$ となる．このとき

$$Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_i - p}{\sqrt{p(1-p)}} = \sqrt{n} \cdot \frac{\bar{X}_n - p}{\sqrt{p(1-p)}}$$

とすると

$$\lim_{n\to\infty} \Pr(Z_n \le z) = \int_{-\infty}^{z} \phi(x;0,1)dx \quad \text{が成り立つ．} \tag{1}$$

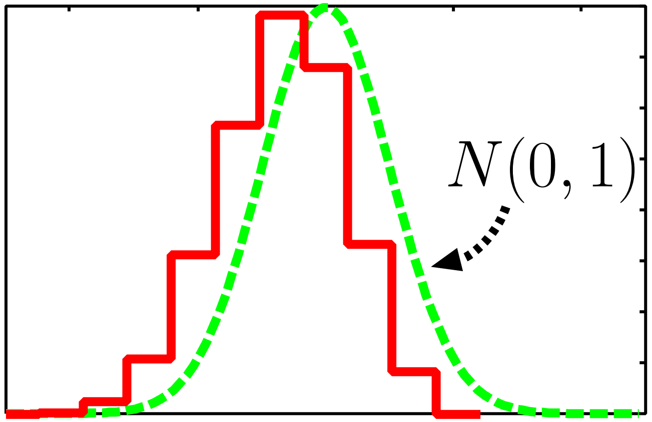
式(1)のように分布関数が収束することを

$$Z_n \xrightarrow{d} N(0,1)$$

と書くこともある.

- 表の確率が**0.3**のコイン.
- n回振るときのZ_nの密度関数をプロット

$n = 10$



$n = 100$