情報科学の基礎理論

東京科学大学　ILA国府台
中林　潤

JH人材育成課　バイオインフォマティシャン育成講座　⑤

バイオインフォマティクスの基礎理論

- バイオインフォマティクス
  “バイオ”＋“インフォマティクス”
  生物学　　情報科学

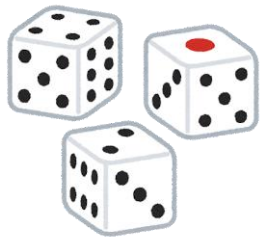
- 情報科学：情報を研究の対象とした科学の分野。情報の処理、操作、統合などを扱う。

- 情報：ある物事に関する知識→研究の対象として扱うための定義が必要

試行と事象

- 事象：試行の結果として得られる事柄

サイコロの例



試行：確率的に起こる出来事（サイコロを振る）
事象：{1の目が出る}、・・・、{6の目が出る}
根源事象：これ以上分解できない事象
{1}、{2}、・・・、{6}
複合事象：根源事象の組み合わせ
偶数が出る {2,4,6}
標本空間：全事象を表す空間
{1,2,3,4,5,6}

確率

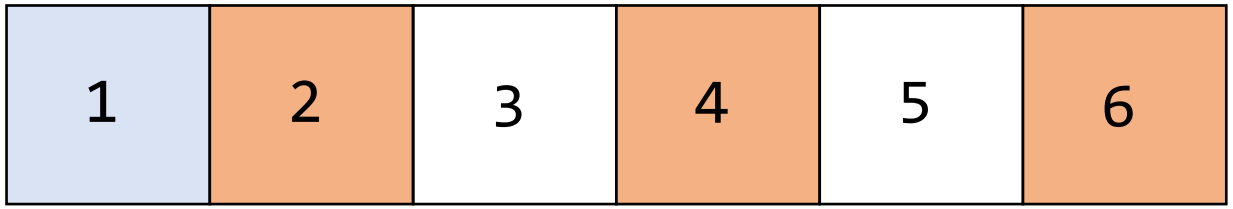
- 確率
  事象の起こりやすさを表す値
  事象Aに対して負でない実数値$P(A)$が対応している
  $0 \leq P(A) \leq 1$

- 確率変数
  事象に割り当てられた数値
  確率的に決まる値

確率は面積の問題として定式化できる

サイコロ

標本空間　面積=1



1の目の出る確率

$P(\{1\}) = \frac{1}{6}$

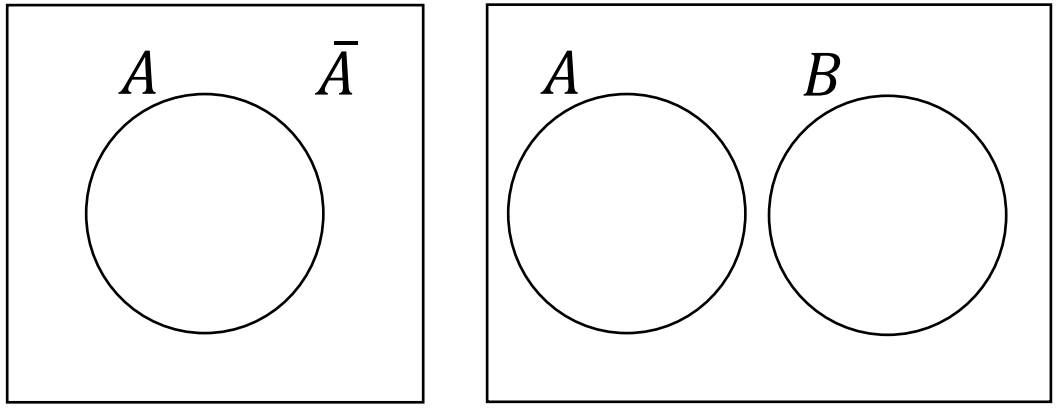
偶数の出る確率

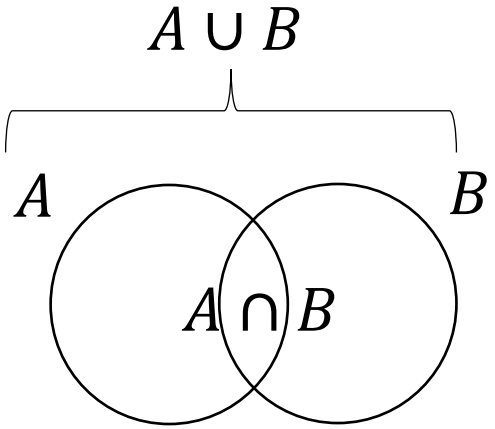
$P(\{2,4,6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$

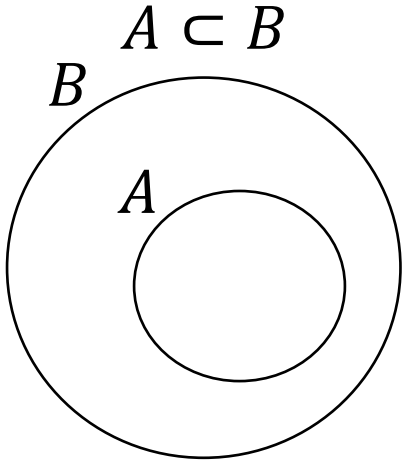
確率：事象の起こりやすさを表す値
　　標本空間中で事象の占める面積

記号による表記

根源事象：$A_1 = \{s_1\}$、・・・、$A_6 = \{s_6\}$
複合事象：$A_1 \cup A_2 \cup A_3 = \{s_1, s_2, s_3\}$
空事象：起こりえない事象、1と2の目が同時に出る　$A_1 \cap A_2 = \emptyset$
お互いに排反な事象：同時に生起しない事象
補事象：ある事象の標本点以外のすべての事象　$P(\bar{A}) = 1 - P(A)$
和事象：事象AまたはBが起こる　$A \cup B$
積事象（結合事象）：事象AとBが同時に起こる　$A \cap B$







期待値と分散

期待値：確率変数を確率で重みづけして足し合わせた値

$E[X] = \sum_{k=1}^{N} x_k P(x_k)$

$1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$

分散：確率変数のばらつきの尺度

$V[X] = \sum_{k=1}^{N} (x_k - E[X])^2 P(x_k)$

$= \sum_{k=1}^{N} \left({x_k}^2 P(x_k) - 2x_k E[X] P(x_k) + (E[X])^2 P(x_k) \right)$

$= \sum_{k=1}^{N} {x_k}^2 P(x_k) - 2E[X] \sum_{k=1}^{N} x_k P(x_k) + (E[X])^2 \sum_{k=1}^{N} P(x_k)$

$= E\left[X^2\right] - (E[X])^2$

＝二乗の期待値－期待値の二乗

$(1 - 3.5)^2 \times \frac{1}{6} + (2 - 3.5)^2 \times \frac{1}{6} + (3 - 3.5)^2 \times \frac{1}{6} + (4 - 3.5)^2 \times \frac{1}{6} + (5 - 3.5)^2 \times \frac{1}{6} + (6 - 3.5)^2 \times \frac{1}{6} = 2.91666 \cdots$

$1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} - 3.5^2 = 2.91666 \cdots$

連続型の確率変数

- ある試行によって得られる変数

  離散型

  連続型

$$f(x) = \frac{1}{\sqrt{2 \times \pi \times \sigma^2}} e^{\left(\frac{-(x-E)^2}{\sigma^2}\right)}$$

連続変数　身体測定の例

身長165-170㎝

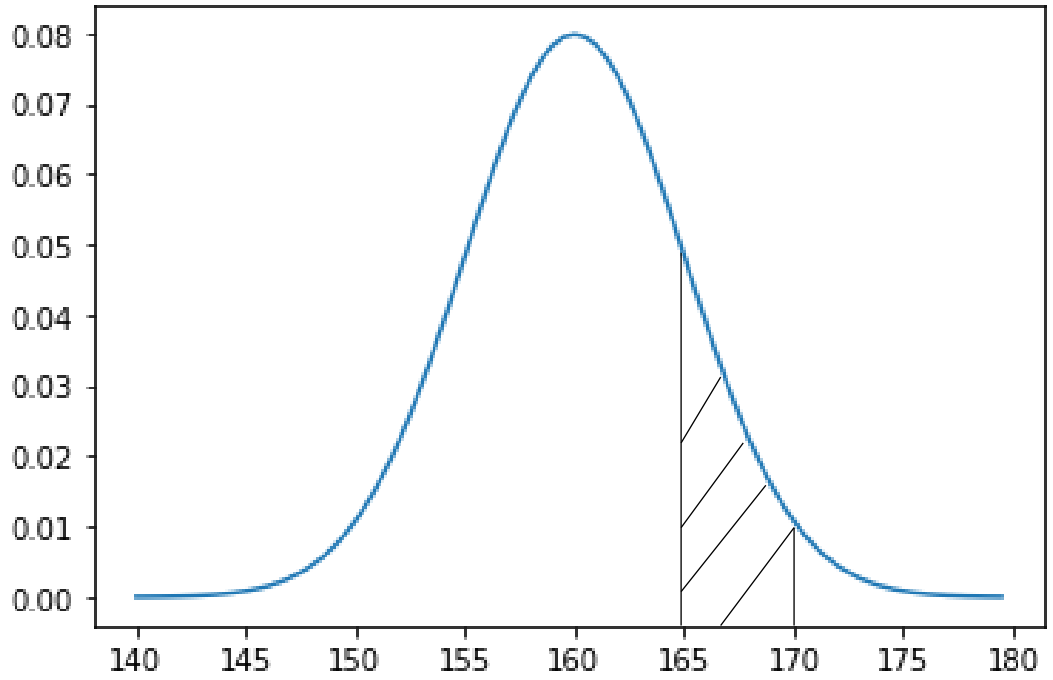
$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

$f(x)$：確率密度関数

積分

確率密度 ⇄ 確率分布

微分



確率密度関数

全確率

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

期待値

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

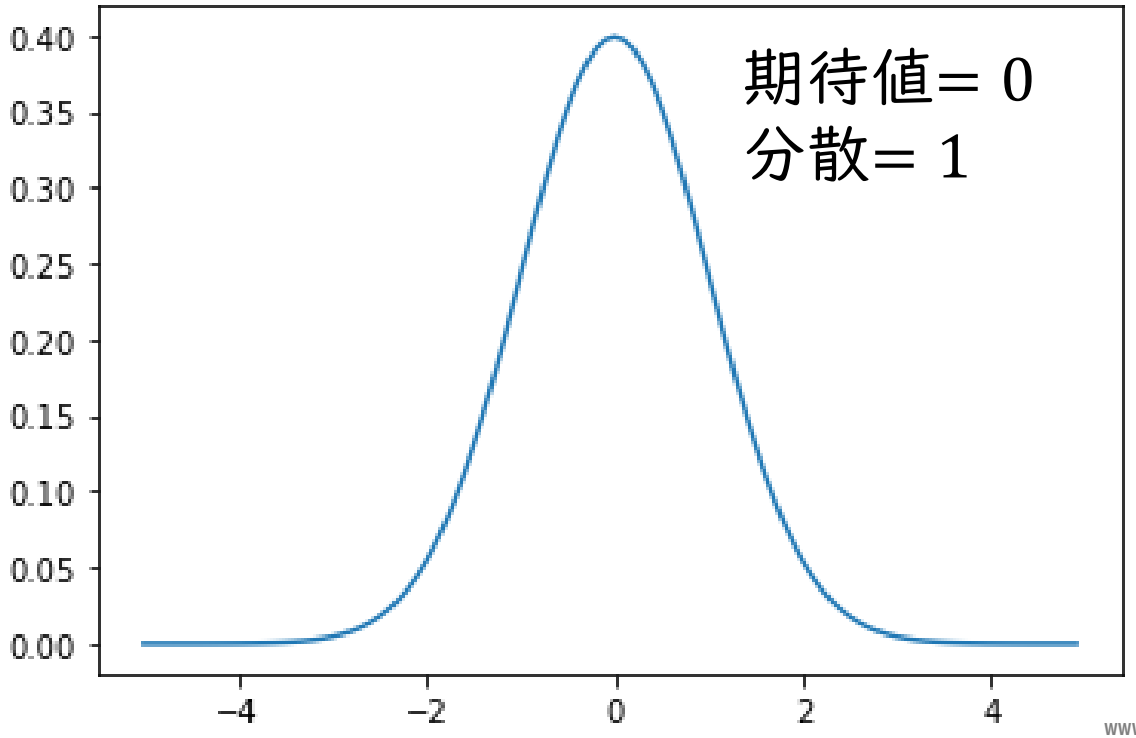
分散

$$V(X) = \int_{-\infty}^{\infty} \big(x - E(X)\big)^2 f(x)dx$$
$$= E\big(X^2\big) - (E[X])^2$$

正規分布の確率密度

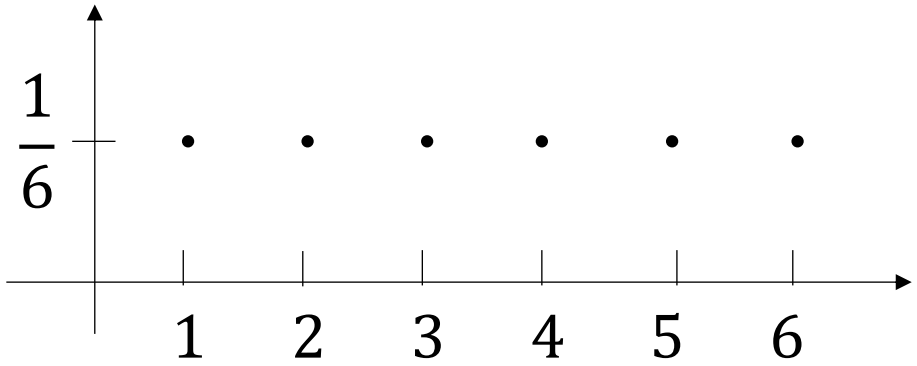
$$f(x) = \frac{1}{\sqrt{2 \times \pi \times \sigma^2}} e^{\left(\frac{-(x-E)^2}{\sigma^2}\right)}$$

標準正規分布



一様分布

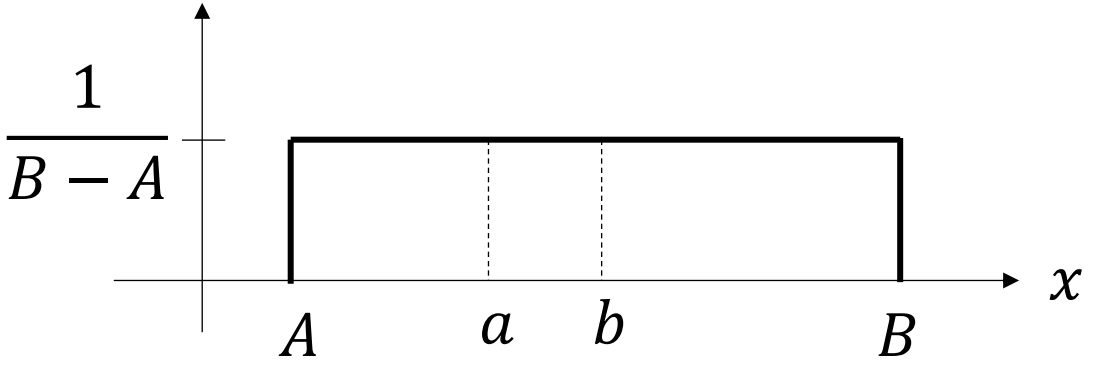
離散型



$P(x = X) = \frac{1}{n}$：確率質量関数

$$\sum_{i=1}^{N} P(X = x_i) = 1$$

連続型



$$P(a \leq x \leq b) = \int_a^b f(x)dx = \frac{1}{B-A} \times (b-a)$$

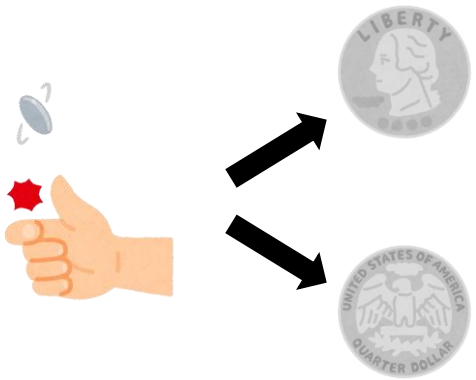
$f(x)$：確率密度関数

$$\int_A^B f(x)dx = 1$$

A：下限値
B：上限値

ベルヌーイ試行

ベルヌーイ試行：二つの結果だけが可能な独立な試行を繰り返すとき、
結果の起こる確率が試行を通じて一定であるような試行



表：$\frac{1}{2}$

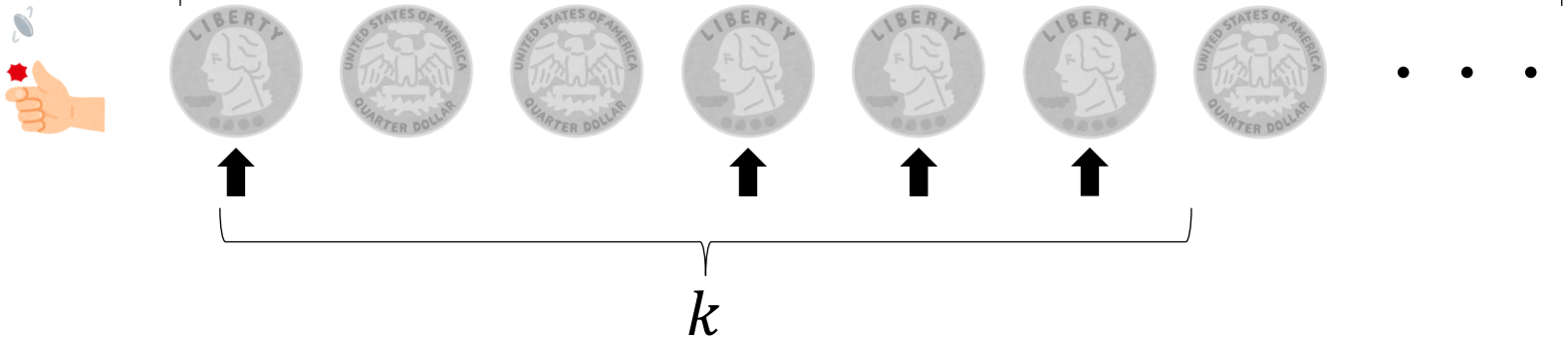
裏：$\frac{1}{2}$

成功：p
失敗：$1-p$

二項分布

二項分布：ベルヌーイ試行をn回独立に行ったとき、k回成功する確率

成功：p
失敗：$1-p$



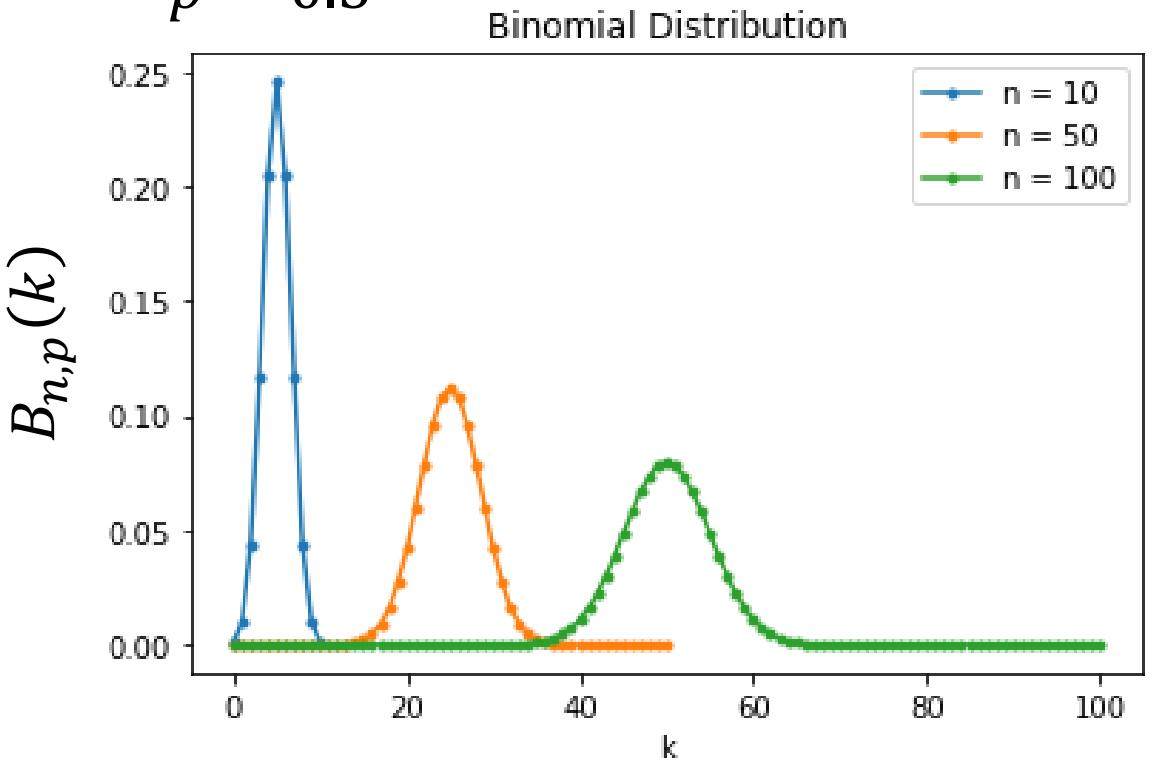
$$B_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \binom{n}{k} = {}_nC_k = \frac{n!}{k!(n-k)!} \text{のこと}$$

$$F(r) = \sum_{k=0}^{r} \binom{n}{k} p^k (1-p)^{n-k}$$

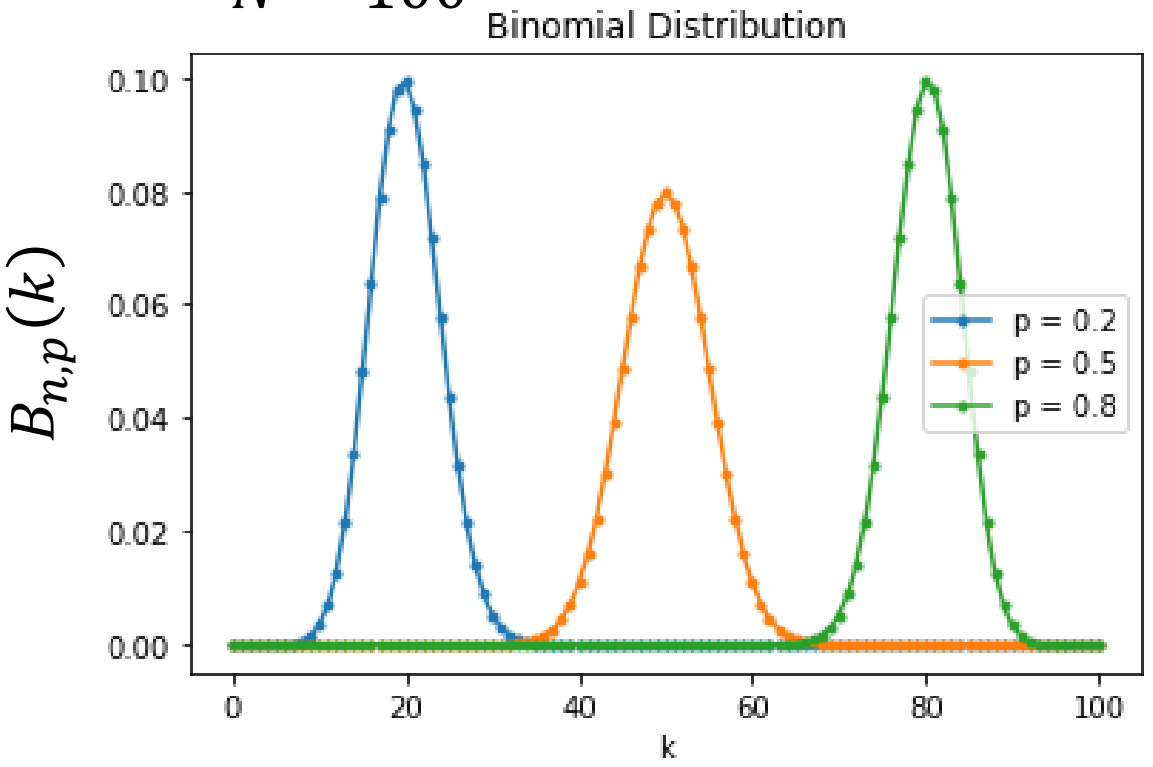
厳密には$B_{n,p}(k)$は二項分布の確率質量関数

二項分布

$p = 0.5$

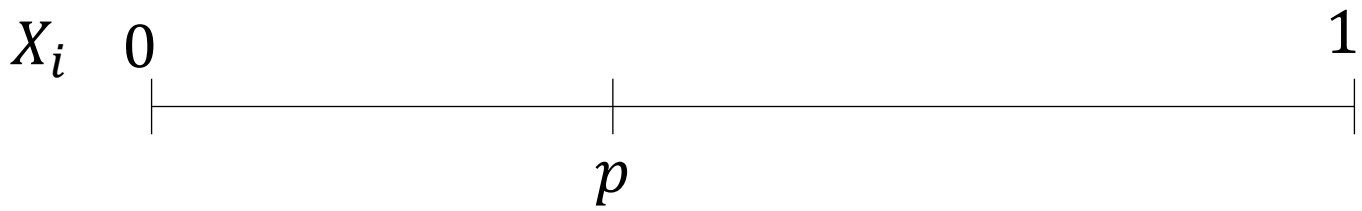


$N = 100$



二項分布の期待値と分散

成功率pの事象で、成功すると1、失敗すると0を取る確率変数X_iを考える



1試行分

X_i の期待値

$0 \times (1-p) + 1 \times p = p$

X_i の分散

$0^2 \times (1-p) + 1^2 \times p - p^2 = p - p^2 = p(1-p)$

2乗の期待値　期待値の2乗

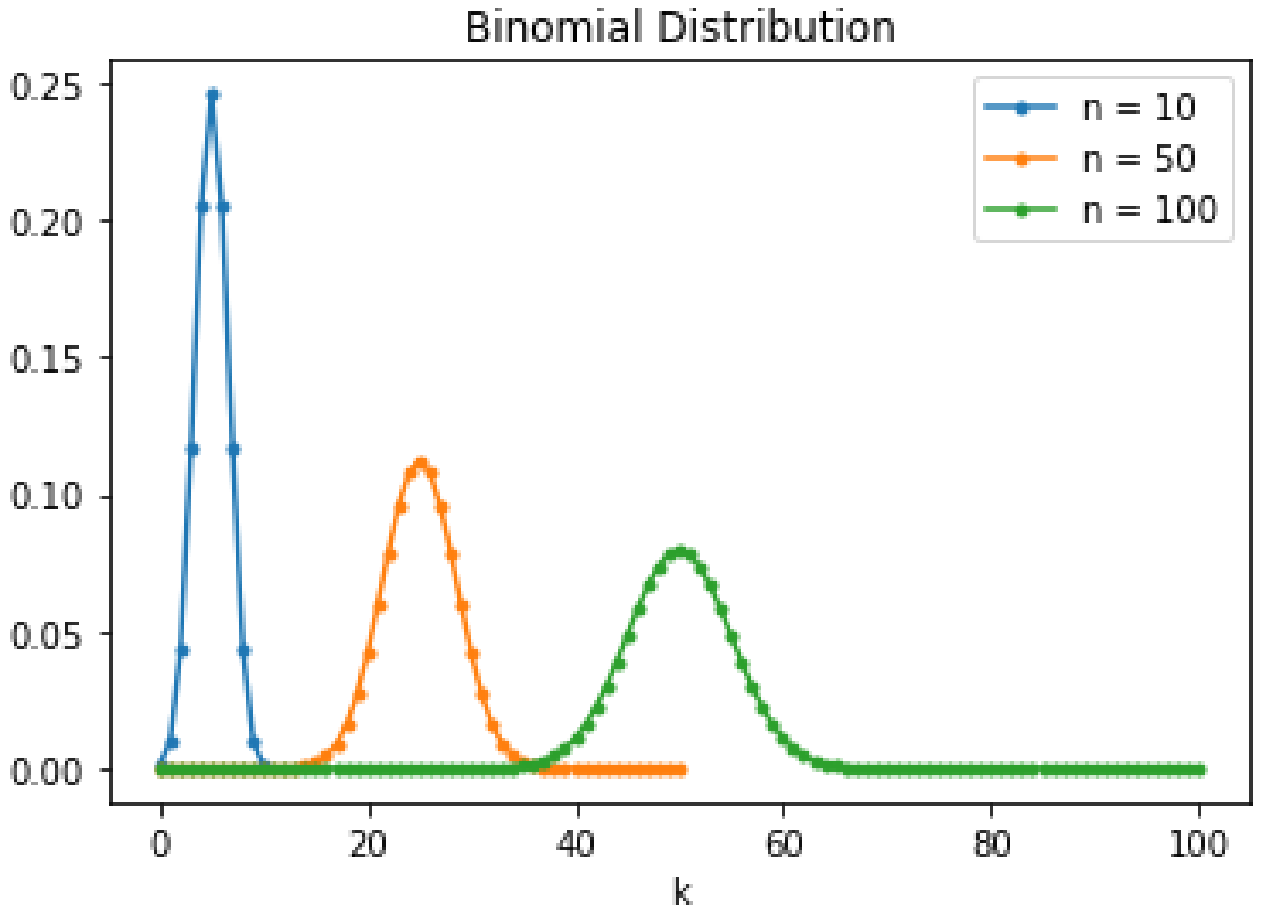
n 回繰り返す

期待値：np

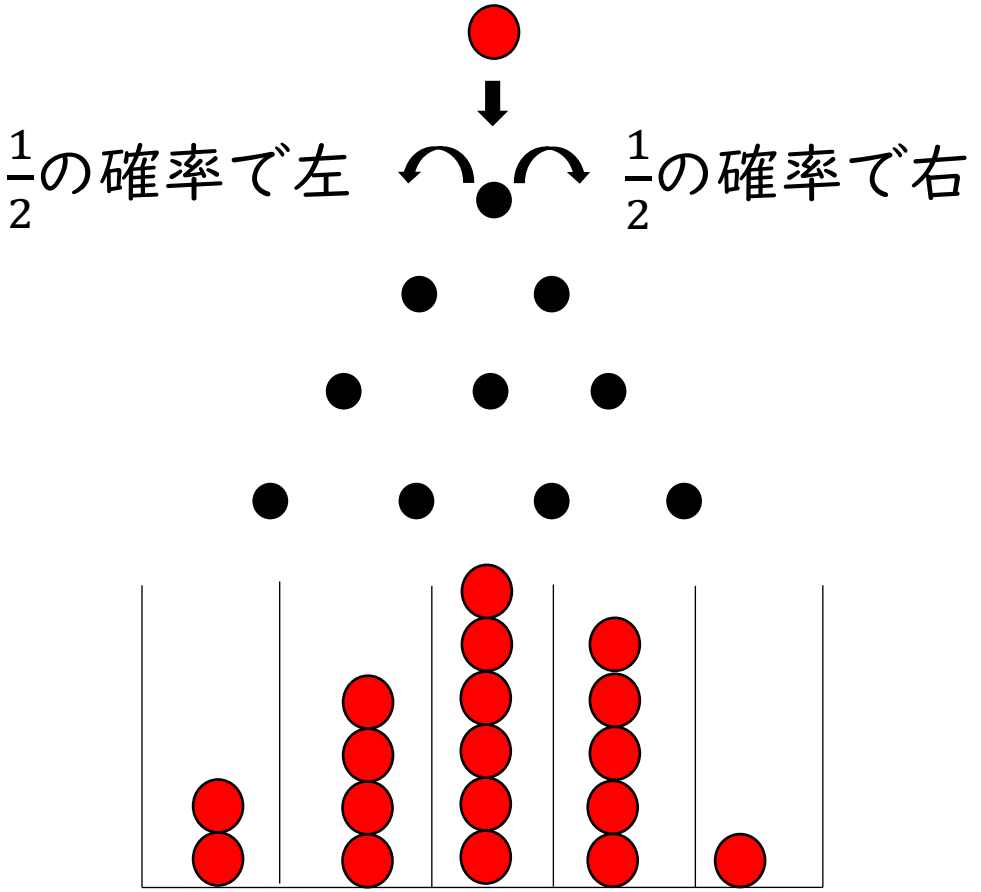
分散：$np(1-p)$

正規分布と二項分布の関係

- 二項分布の期待値npと分散$np(1-p)$が大きいとき、正規分布で近似できる



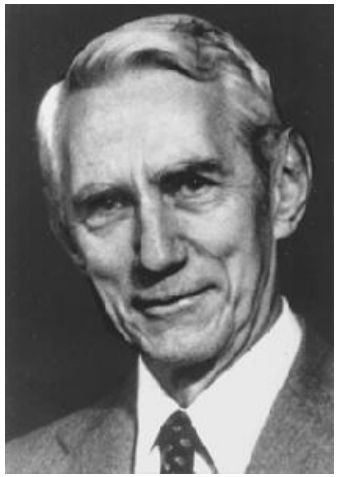
Galton Board



ランダムな選択の繰り返し→正規分布

シャノンの情報量

クロード・エルウッド・シャノン
“A Mathematical Theory of Communication”
Bell System Technical Journal, vol. 27, pp. 379–423, 623–656, 1948



- 情報：あいまいさを減らすもの
- 情報量 $= -\log_2$(事象の起こる確率)
- “surprise”の度合いを表す

情報量の具体例

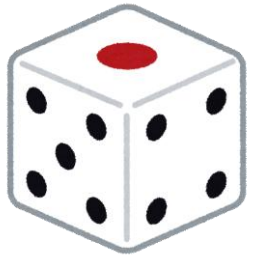
- コイン投げ

  情報量$= -\log_2$（表が出る確率）

  $= -\log_2 \frac{1}{2} = 1$ (bit)

- サイコロ

  情報量$= -\log_2 \frac{1}{6} = 2.58496$



- トランプ

  情報量$= -\log_2 \frac{1}{53} = 5.72792$

エントロピー（平均情報量）

エントロピー（平均情報量）：情報源の平均情報量，情報全体としての情報量を表す。
エントロピーが大きいほど情報源があいまいであることを意味する。

$$H(p) = -\sum_{x_i=1}^{M} p(x_i) \log_2 p(x_i)$$

$$H(p) = -\int p(x) \log p(x)\, dx$$

コインの重心の傾きを知らない場合（表$\frac{1}{2}$, 裏$\frac{1}{2}$）

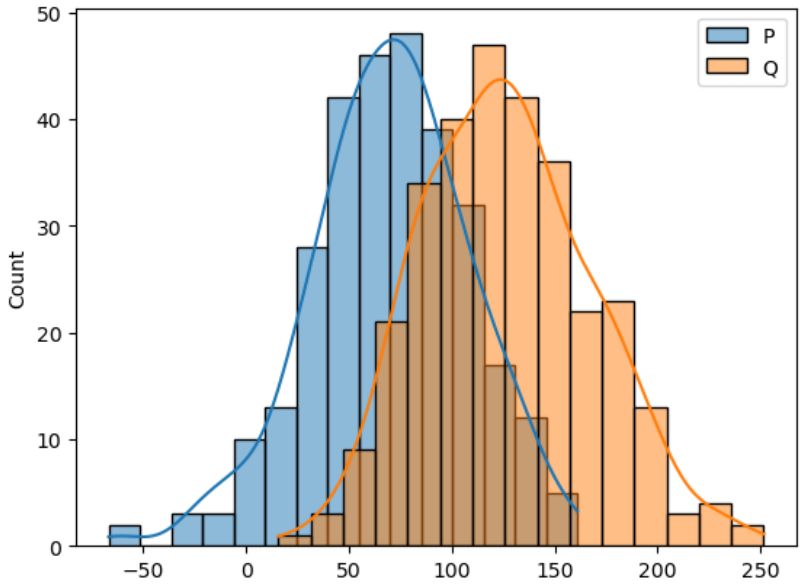
$$H(p) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

コインの重心の傾きを知っている場合（表$\frac{4}{5}$, 裏$\frac{1}{5}$）

$$H(p) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 0.7219$$

Kullback-Leibler Divergence

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$
$$= \int p(x) \log p(x) dx - \int p(x) \log q(x)\, dx$$



確率密度$p(x)$と$q(x)$がどれくらい似ているかを表す指標
$p(x)$と$q(x)$が完全に一致していると0になる

Cross Entropy（交差エントロピー）

離散型

$$H(p,q) = -\sum_x p(x) \log q(x)$$

連続型

$$H(p,q) = -\int_x p(x) \log q(x)\, dx$$

KL DivergenceとCross Entropyの関係

$$H(p,q) = -\sum_x p(x)\log q(x)$$
$$= -\sum_x p(x)\log q(x) + \sum_x p(x)\log p(x) - \sum_x p(x)\log p(x)$$
$$= -\sum_x p(x)\log p(x) + \sum_x p(x)\log p(x) - \sum_x p(x)\log q(x)$$
$$= H(p) + \sum_x p(x)\log\frac{p(x)}{q(x)}$$
$$= H(p) + D_{KL}(P||Q)$$

Cross Entropyを最小化することは、KL Divergenceを最小化することに等しい。

$p(x)$：真の分布

$q(x)$：予測分布

Cross Entropyを最小化すると、予測分布が真の分布に近づく。

まとめ

- 情報：あいまいさを減らすもの
- 情報量

$$I(x) = -\log_2(p(x))$$

- エントロピー

$$H(p) = -\sum_i p(x_i) \log p(x_i)$$

- 交差エントロピー

$$H(p, q) = -\sum_i p(x_i) \log q(x_i)$$

- KLダイバージェンス

$$D_{KL}(P||Q) = \int p(x) \log p(x) dx - \int p(x) \log q(x)\, dx$$