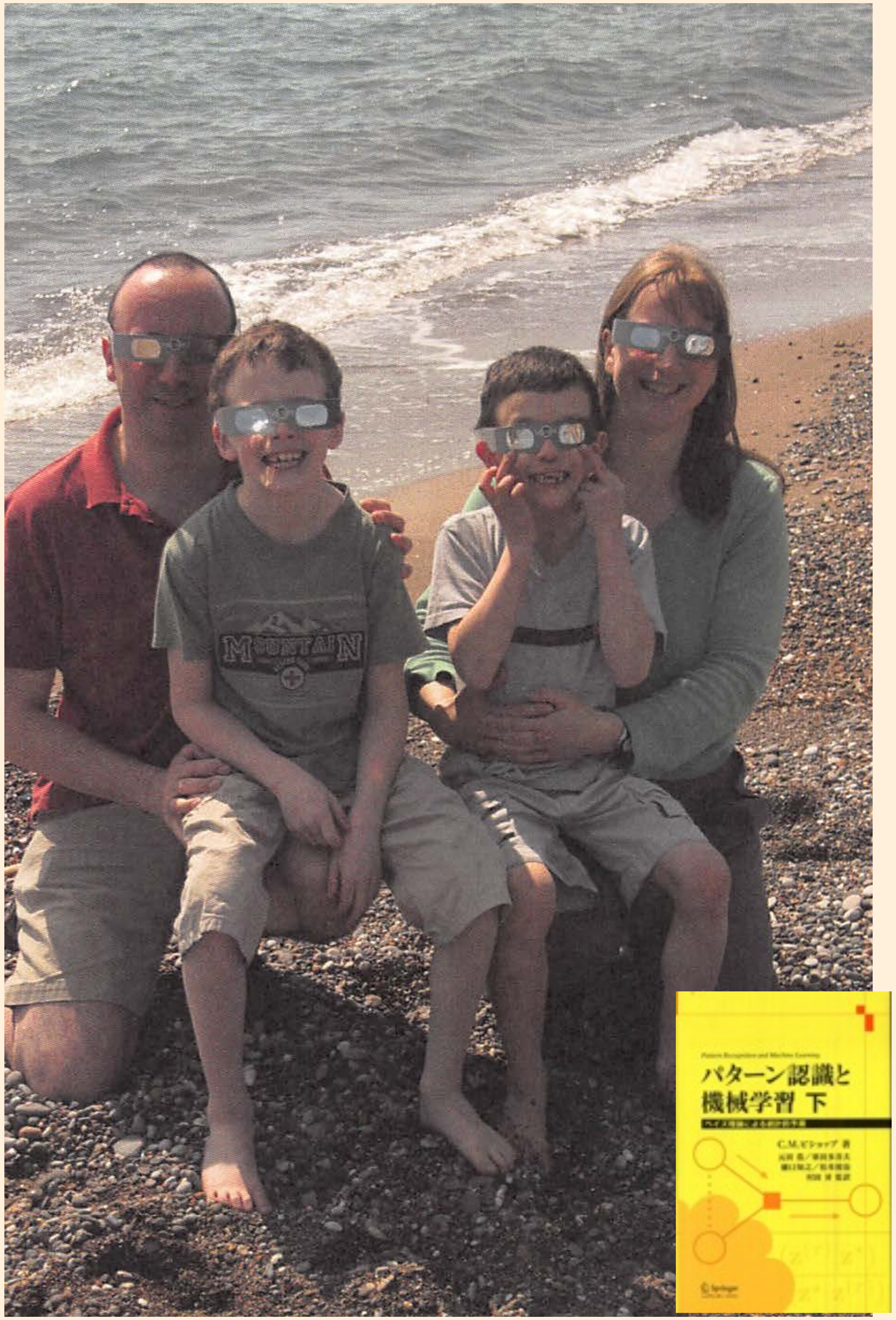
情報理論２

藤田一寿
津山工業高等専門学校情報工学科講師
電気通信大学先進理工学科協力研究員



優れた書籍執筆し、それに使われている
良質な図を公開してくれたBishopに感謝

ver. 20150415

統計、確率

情報理論で確率がなぜ必要？

情報の量を確率で定義しているため、確率を知らないと情報理論がさっぱりわからなくなる。

集合

- なにかの集まり
  - 1, 2, 3, 4, 5という5個の数の集まりをAという集合で表すには
    - A = {1, 2, 3, 4, 5}

  - 集合名 = {要素, 要素, …}
- 空の集合を空集合という。

包含関係

- xが集合Sに含まれる。

$$x \in S$$

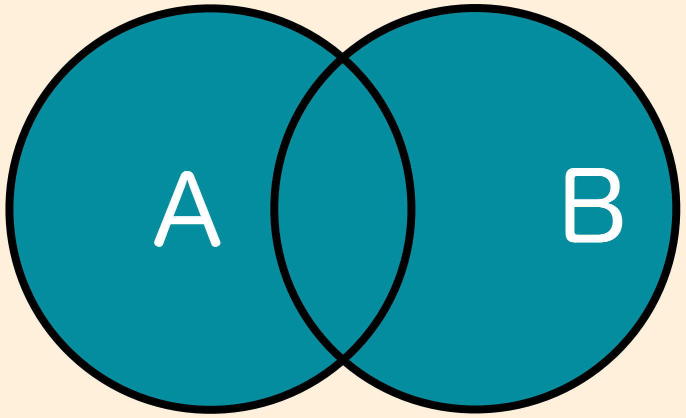
- 集合BはAの部分集合である(含まれる)。

$$B \subset A$$

集合の和、積

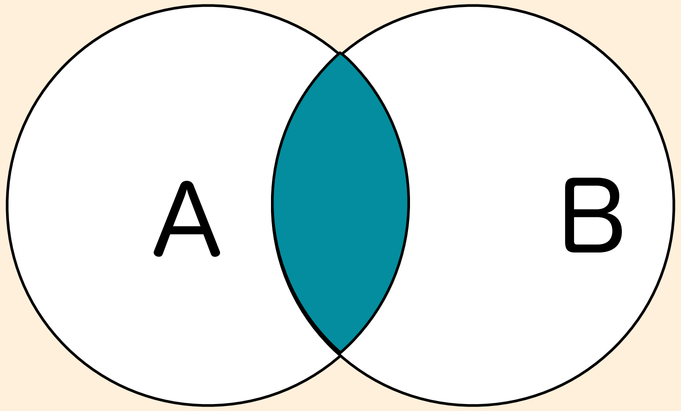
- 集合AとBの和(両方の要素からなる集合)を和集合と呼ぶ。

$A \cup B$



- AとBに属する要素からなる集合を積集合と呼ぶ

$A \cap B$



試行と事象

- 試行
  - 結果が確率的である行為
- 標本点
  - 可能な結果
    - サイコロを1回投げた場合、１から6までの数字が出る。
- 標本空間
  - 標本点の全体の集合
    - サイコロを1回投げた場合、標本空間Ω={1, 2, 3, 4, 5, 6}
- 事象
  - 起こりうる事柄(結局事象は部分集合となる)
  - サイコロで奇数の目が出るという事象は、A={1,3,5}である。
- 根源事象
  - ただ一つの標本点からなり、分解ができな事象
    - サイコロの場合はA={1}は根源事象である。

確率とは

ラプラスの定義

*N*個の根源事象があって、それぞれ同等に確からしいとする。事象Aが起こる事象が*R*個ならば、事象Aが起こる確率は

$$P(A) = R/N$$

である。場合の数の数え上げでの確率の計算は、この定義に基いている。

しかし、コイントス、サイコロ、トランプの場合、各根源事象は同等に確からしいと考えられるが、そのように考えられる場合は少ない。

頻度による定義

事象Aが生じるかも知れない試行を*n*回繰り返した時、Aが実際に起こった回数をn_Aとすると、Aが起こる割合*P'*は

$$P'(A) = n_A/n$$

である。試行回数を無限に増やしていったとき、事象Aが生じる頻度確率は

$$P(A) = \lim_{n \to \infty} n_A/n$$

となる。ただ、試行回数をどれだけ増やせば確率が求まるのかわからない。

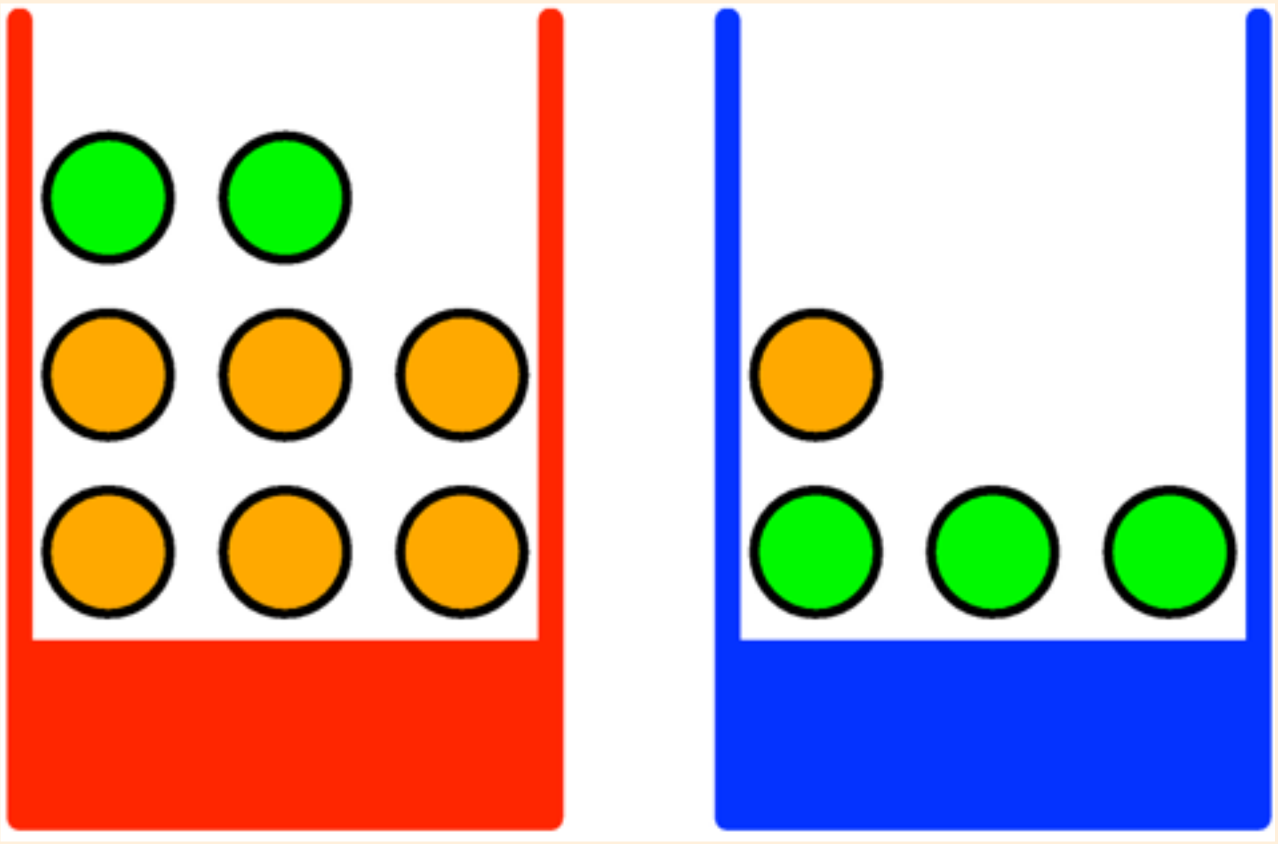
公理主義的な定義

- すべての事象Aに対し $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$ (要するに確率を全部足せば１になる)
- 互いに排他的な事象A1, A2, …,に対し

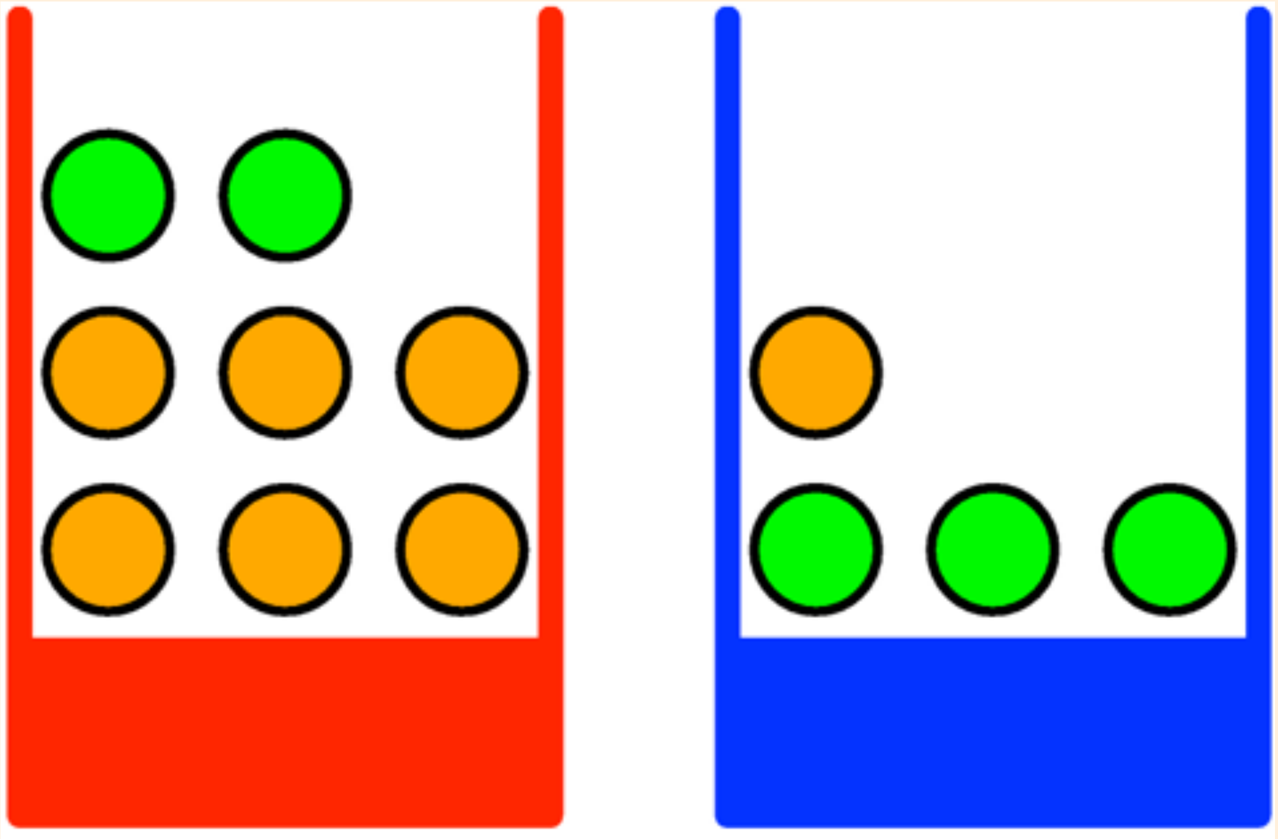
$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

簡単な例

- 赤と青の2つの箱
- 赤にはりんご2つとオレンジ6つ
- 青にはりんご3つとオレンジ1つ

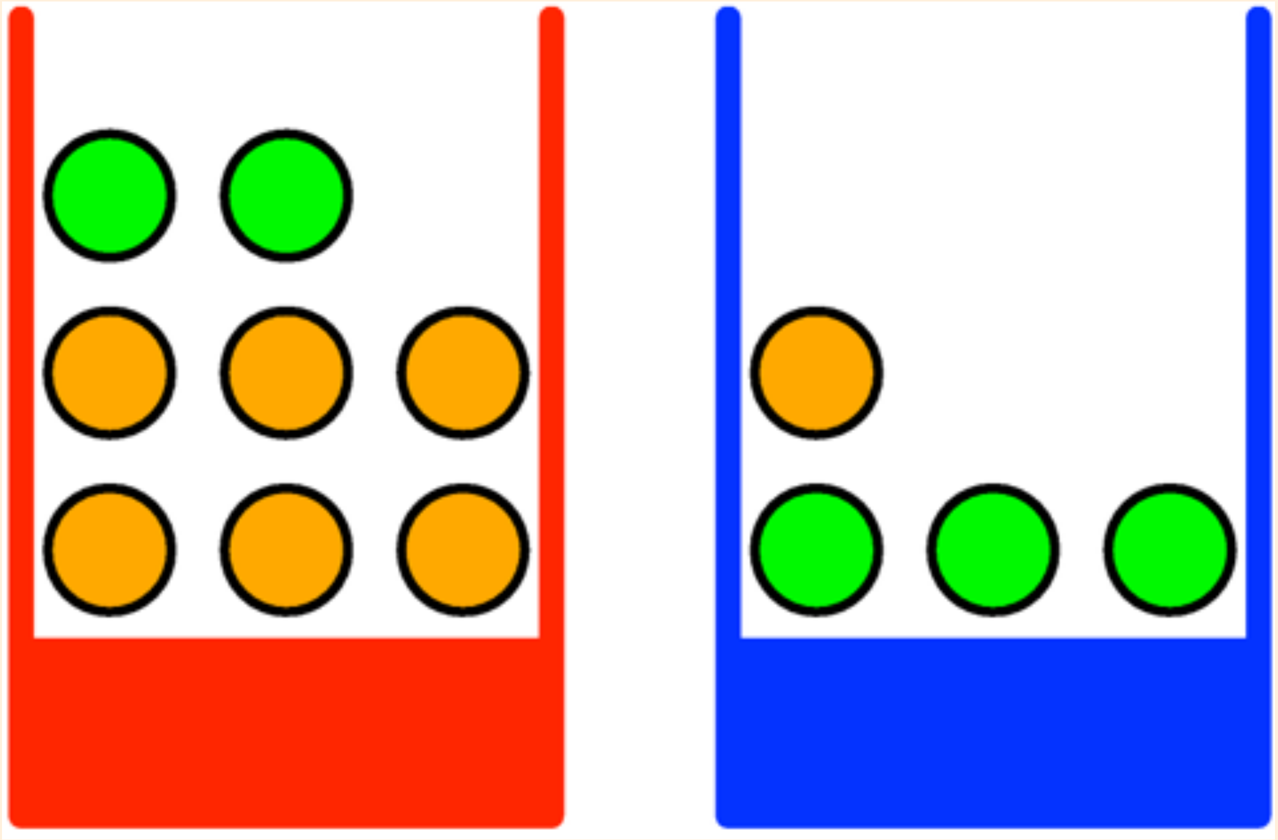


- 箱をある確率に従い、ランダムに選ぶ。
  - 赤い箱は40%の確率で選び、青い箱は60%の確率で選ぶ。
- 箱の中からランダムに果物を選ぶ。



確率変数

- 確率変数は事象を表す変数
- 箱を表す変数B: B = r, B = b
- 果物を表す変数F: F = a, F = o
- 赤い箱が選ばれる確率
  - $P(B = r) = 0.4$
- 青い箱が選ばれる確率
  - $P(B = b) = 0.6$



同時確率

- 赤い箱が選ばれて、りんごが選ばれる
  - P(F = a, B = r)
- 赤い箱が選ばれて、オレンジが選ばれる
  - P(F = o, B = r)
- このように複数の確率変数が同時に決まるときの確率を同時確率という。

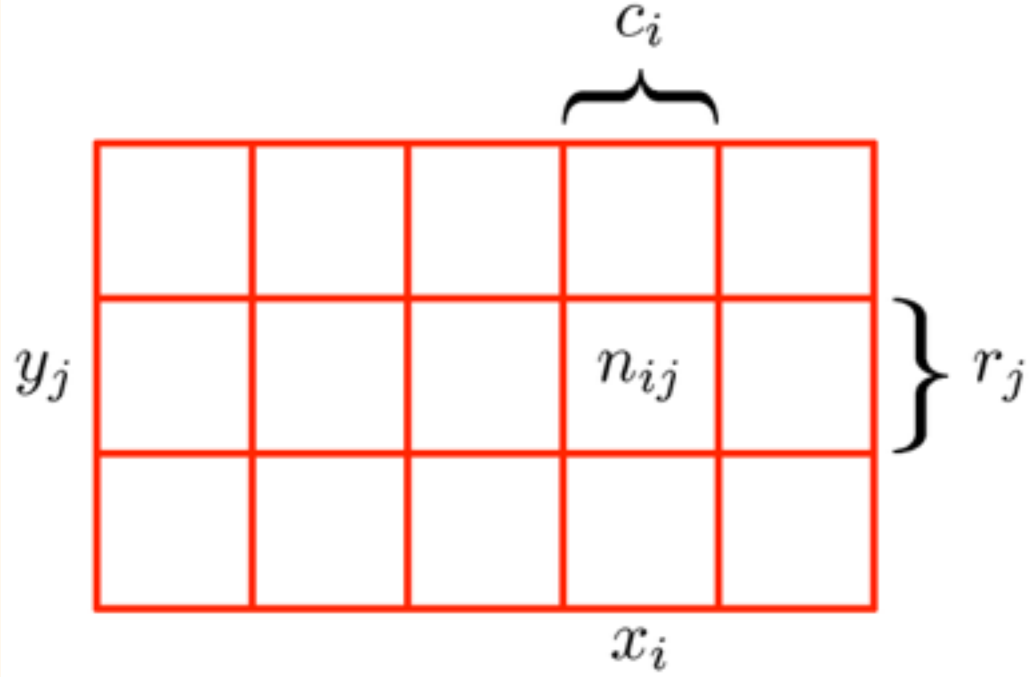
一般的に書いてみる

- 確率変数X, Yがある

$$X = \{x_i | i = 1, ..., M\}$$

$$Y = \{y_j | j = 1, ..., L\}$$

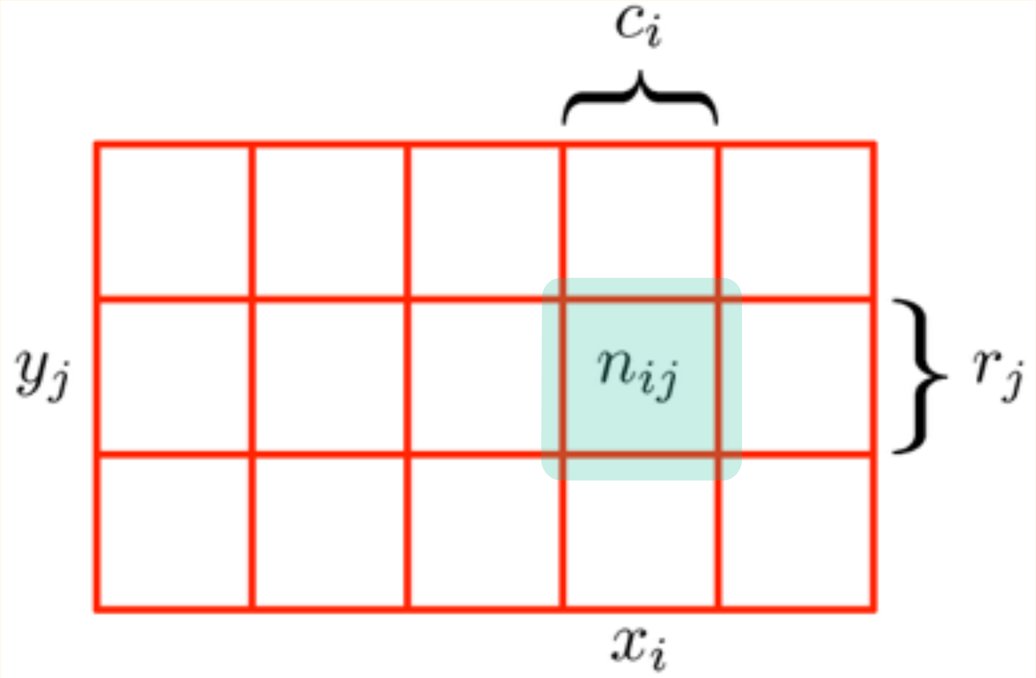
- N回試行した時、X=x_i, Y=y_jとなった回数をn_{ij}とする。



同時確率

- X = x_i, Y = y_jが同時に起こる確率

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$



確率の加法定理

Yを考慮せずXがx_iとなる確率を計算すると

$$P(X = x_i) = \frac{c_i}{N}$$

ここで

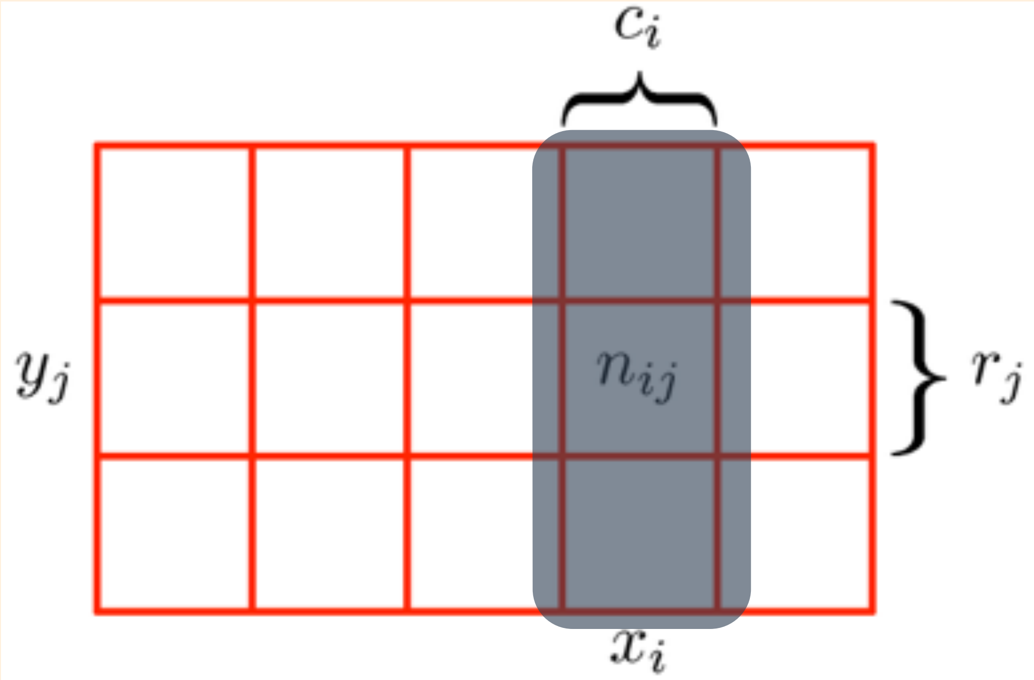
$$c_i = \sum_j n_{ij}$$

よって

$$P(X = x_i) = \frac{n_{i0} + n_{i1} \dots}{N}$$

$$= \frac{n_{i0}}{N} + \frac{n_{i1}}{N} \dots$$

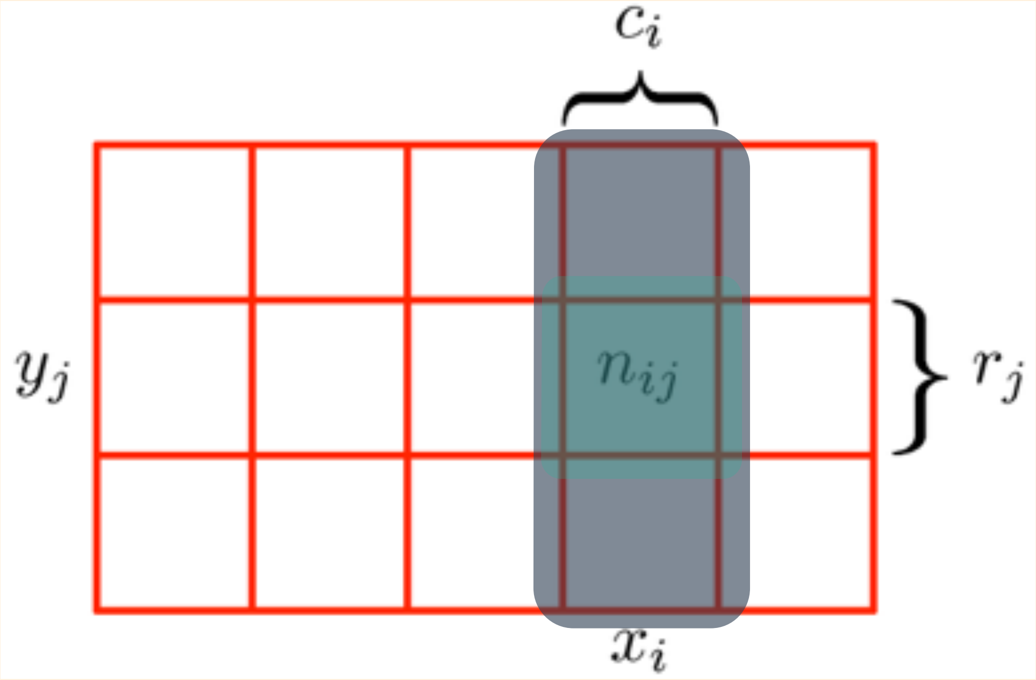
$$= \sum_j p(X = x_i, Y = y_j)$$ 加法定理



Yについて周辺化した。その結果、周辺確率が出てくる。

条件付き確率

- $X=x_i$となることが確定している時、$Y=y_j$が起こる確率は $P(Y=y_j|X=x_i)=\frac{n_{ij}}{c_i}$
  - $P(Y=y_j|X=x_i)$: 条件付き確率

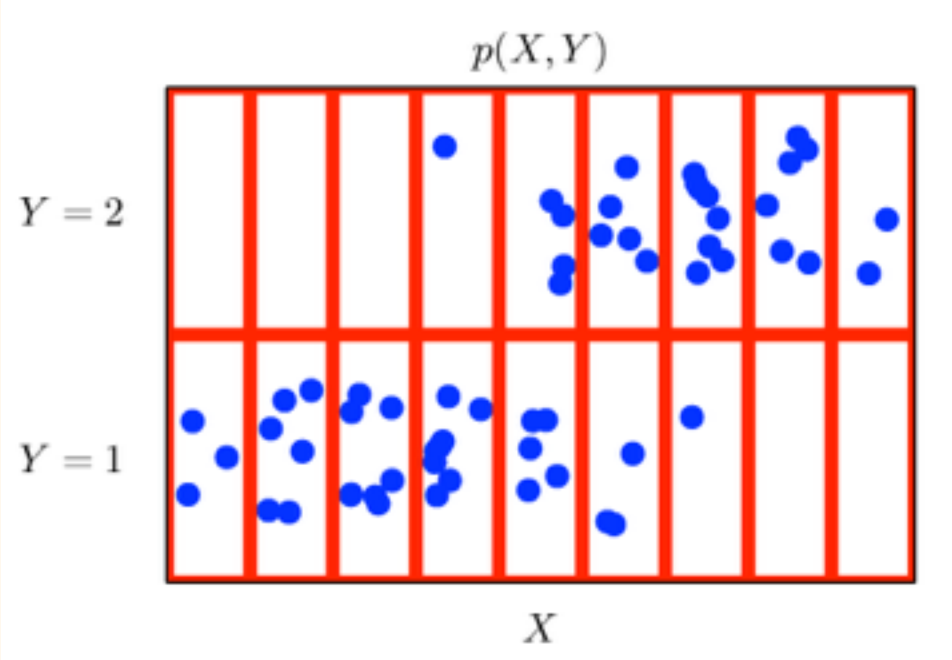


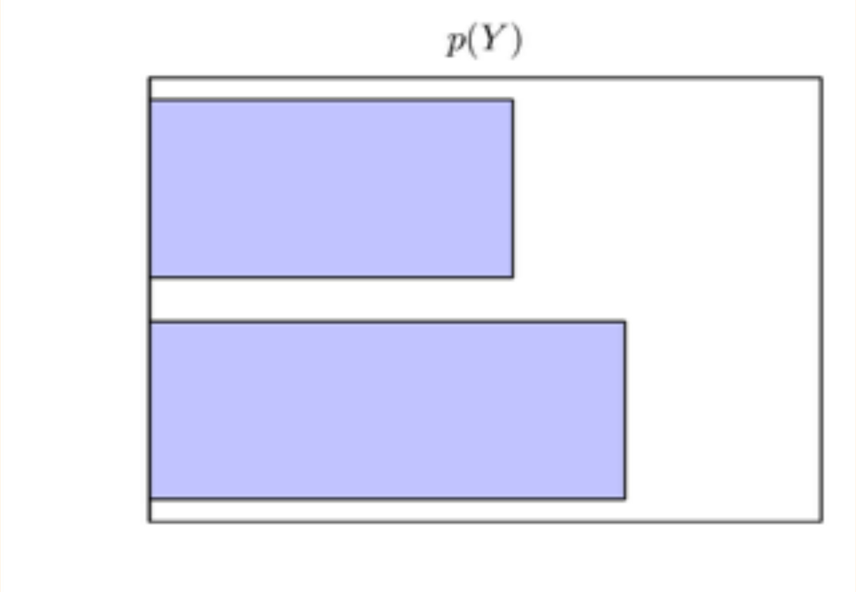
乘法定理

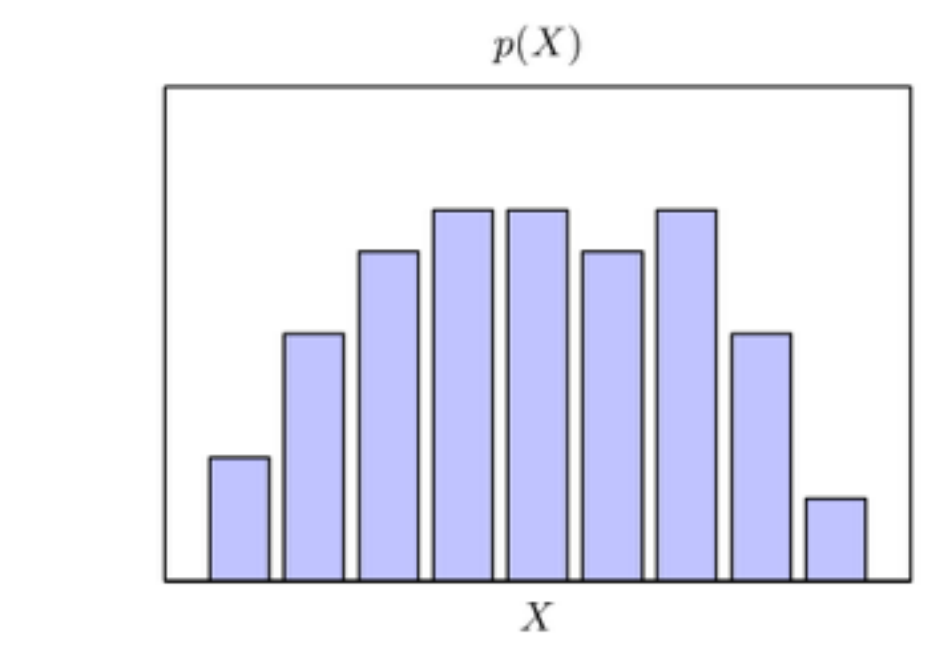
$$\begin{aligned} P(Y = y_j | X = x_i) &= \frac{n_{ij}}{c_i} \\ &= \frac{\frac{n_{ij}}{N}}{\frac{c_i}{N}} \\ &= \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} \end{aligned}$$

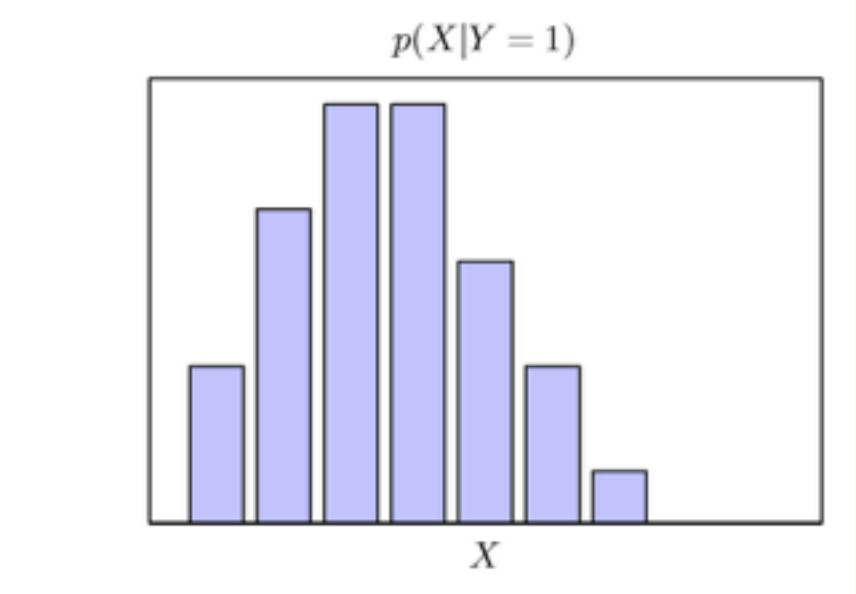
$$P(X = x_i, Y = y_j) = P(Y = y_j | X = x_i) P(X = x_i)$$ 乘法定理

2変数の確率分布









ベイズの定理

- 乗法定理より

$$P(Y|X) = \frac{P(X,Y)}{P(X)}$$

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

$$\Rightarrow$$

$$P(X,Y) = P(Y|X)P(X)$$

$$P(X,Y) = P(X|Y)P(Y)$$

$$P(Y|X)P(X) = P(X|Y)P(Y)$$

$$\boxed{P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}}$$

ベイズの定理

トーマス・ベイズ(1702-1761)

- 1702年イングランド生まれ
- 牧師

この肖像画はベイズではないらしい

ベイズの定理

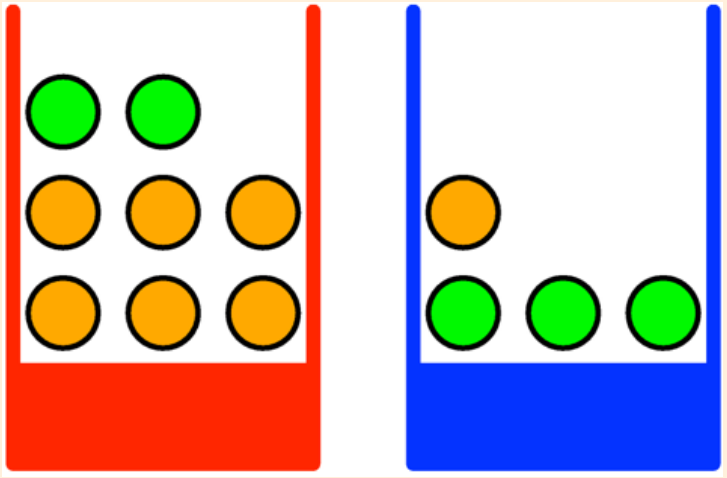
もっともらしさ

$$\underset{\text{事後確率}}{\boxed{P(Y|X)}} = \frac{\overset{\text{尤度}}{\boxed{P(X|Y)}}\overset{\text{事前確率}}{\boxed{P(Y)}}}{P(X)}$$

事後確率は尤度X事前確率に比例する。

箱の例では

- 箱を選ぶ確率: P(B)
  - 事前確率(Prior probability)

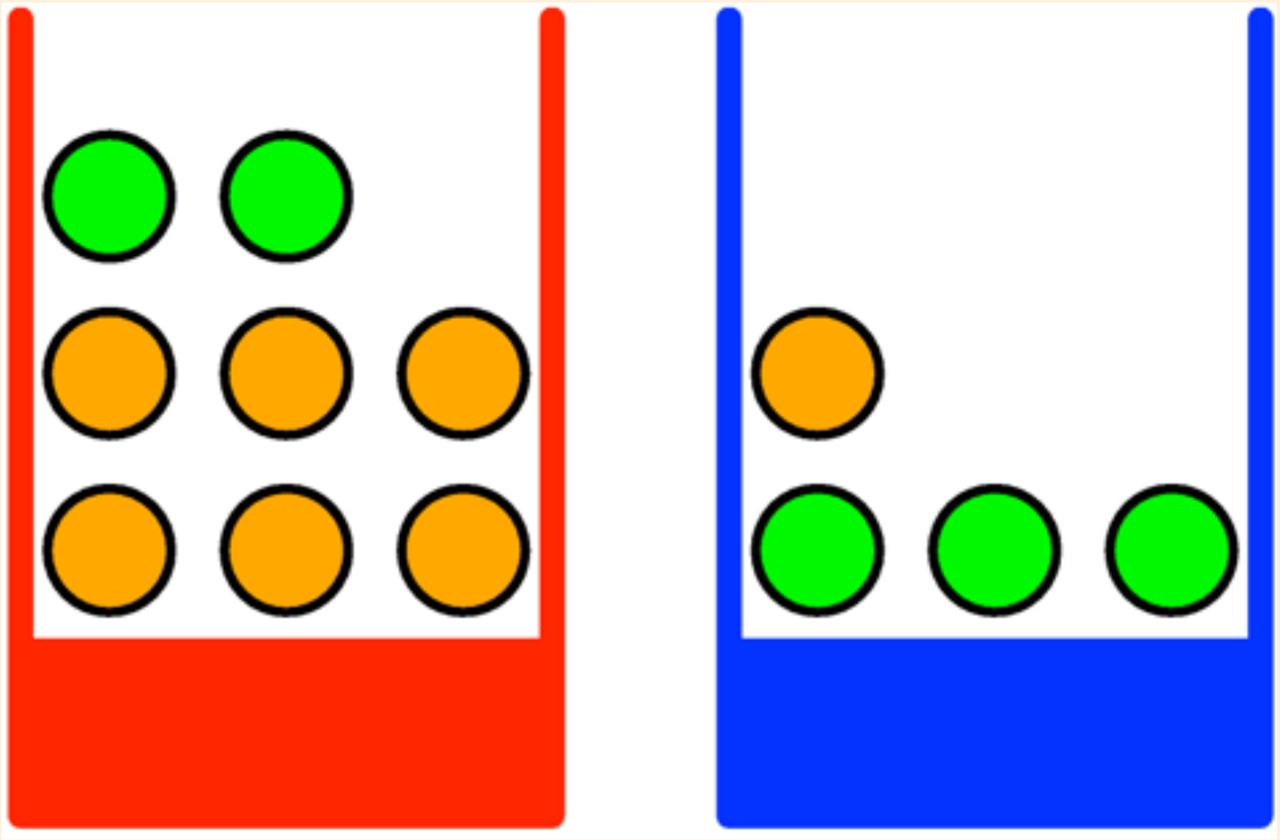


- 選んだ果物から箱を選んだ確率が分かる: P(B|F)
  - 事後確率(Posterior probability)
- P(F|B)は尤度(Likelihood)

$$P(B|F) = \frac{P(F|B)P(B)}{P(F)}$$

箱の例

- 箱の選択確率
  - P(B=r) = 0.4
  - P(B=b) = 0.6
- 箱ごとの果物の選択確率
  - P(F = a| B=r) = 1/4
  - P(F = o| B=r) = 3/4
  - P(F = a| B=b) = 3/4
  - P(F = o| B=b) = 1/4



問題

- 果物がりんごである確率P(F = a)を求めよ。

- 果物がオレンジであるという条件のもので、箱が赤である確率P(B = r| F = o)を求めよ。

問題

ビデオが映らなくなった事象をB、その原因として、

　A1: ビデオの電子回路の故障(発生確率65%)

　A2: モータの故障(発生確率25%)

　A3: データの破損(発生確率10%)

とする。このとき、条件付き確率を

　$P(B|A1) = 30\%$

　$P(B|A2) = 60\%$

　$P(B|A3) = 10\%$

とする。ベイズの定理を用いて３つの事後確率を求めよ。

塩野著「デジタル情報理論」より

次回

- 確率分布