臨床シーケンス解析をする上で理解しておきたい初歩からの遺伝統計学

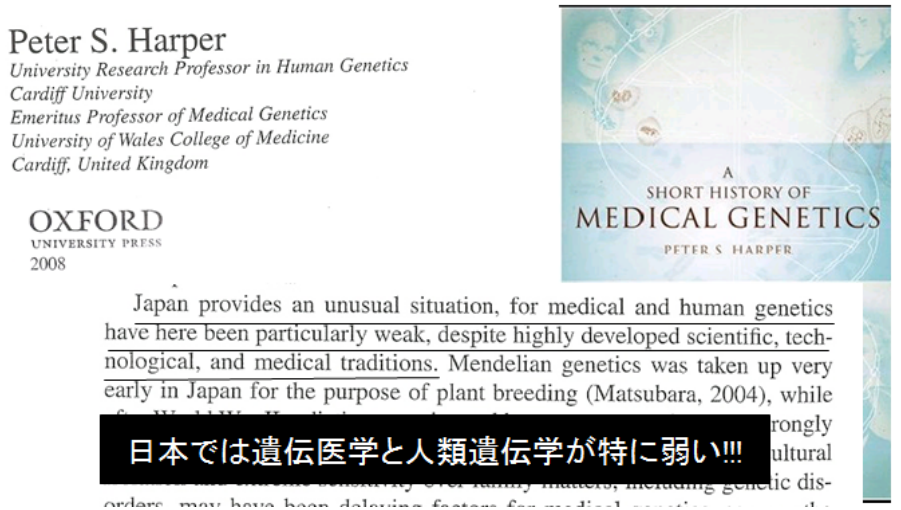
株式会社スタージェン会長
東京女子医科大学膠原病リウマチ痛風センター客員教授
痛風財団理事長
つくば国際臨床薬理クリニック院長

鎌谷直之

日本は遺伝学と統計学が弱い（どこに問題があるのだろう？）

1. 日本では遺伝医学と人類遺伝学が特に弱い（Harper PS）
2. 日本の大学には統計学部、統計学科がほとんど無い
3. 日本では人類遺伝学会、統計学会が極めて小さい
4. 日本からの学術論文には統計に問題がある

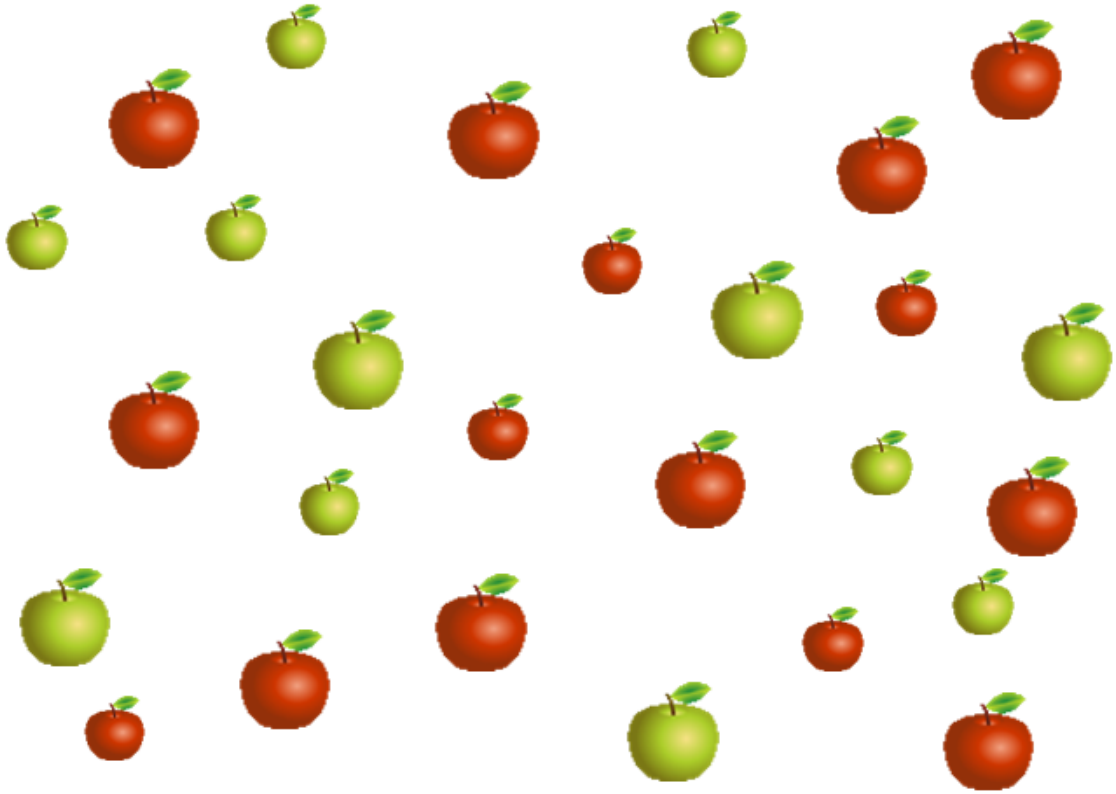
日本人は数学が弱いわけではない、数理統計もわかる。しかし、現実の対象物を数学的に捉える力が不足（特に、不確実性と多様性がある場合）している



情報を抽象的な物として捉えず、
具体的、視覚的に捉える

優秀な英米の遺伝学者には、我々に見えない物が見えている

赤いリンゴと、青いリンゴのどっちがいい？

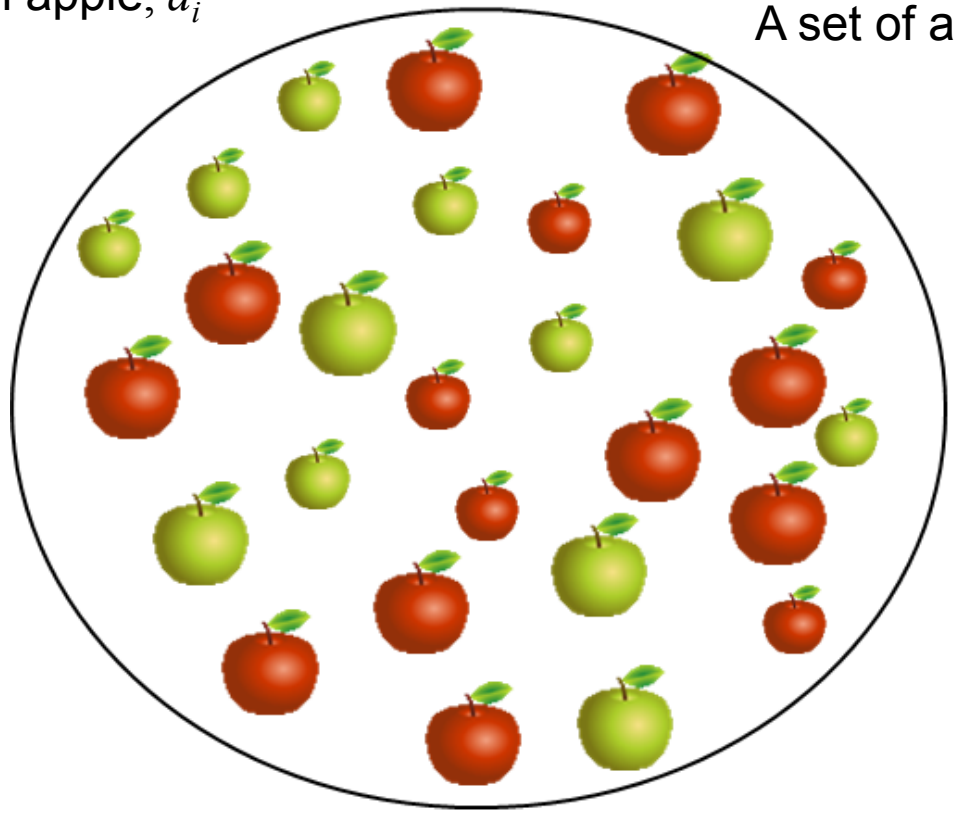


「色々ありますね」では何も進まない

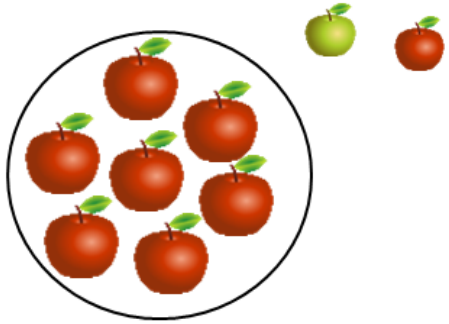
まず、集合を定義する（単数と複数の違いが無い日本語では集合と要素の実感がわきにくい）

An apple, a_i

A set of all apples Ω, $\Omega = \{a_i\}$



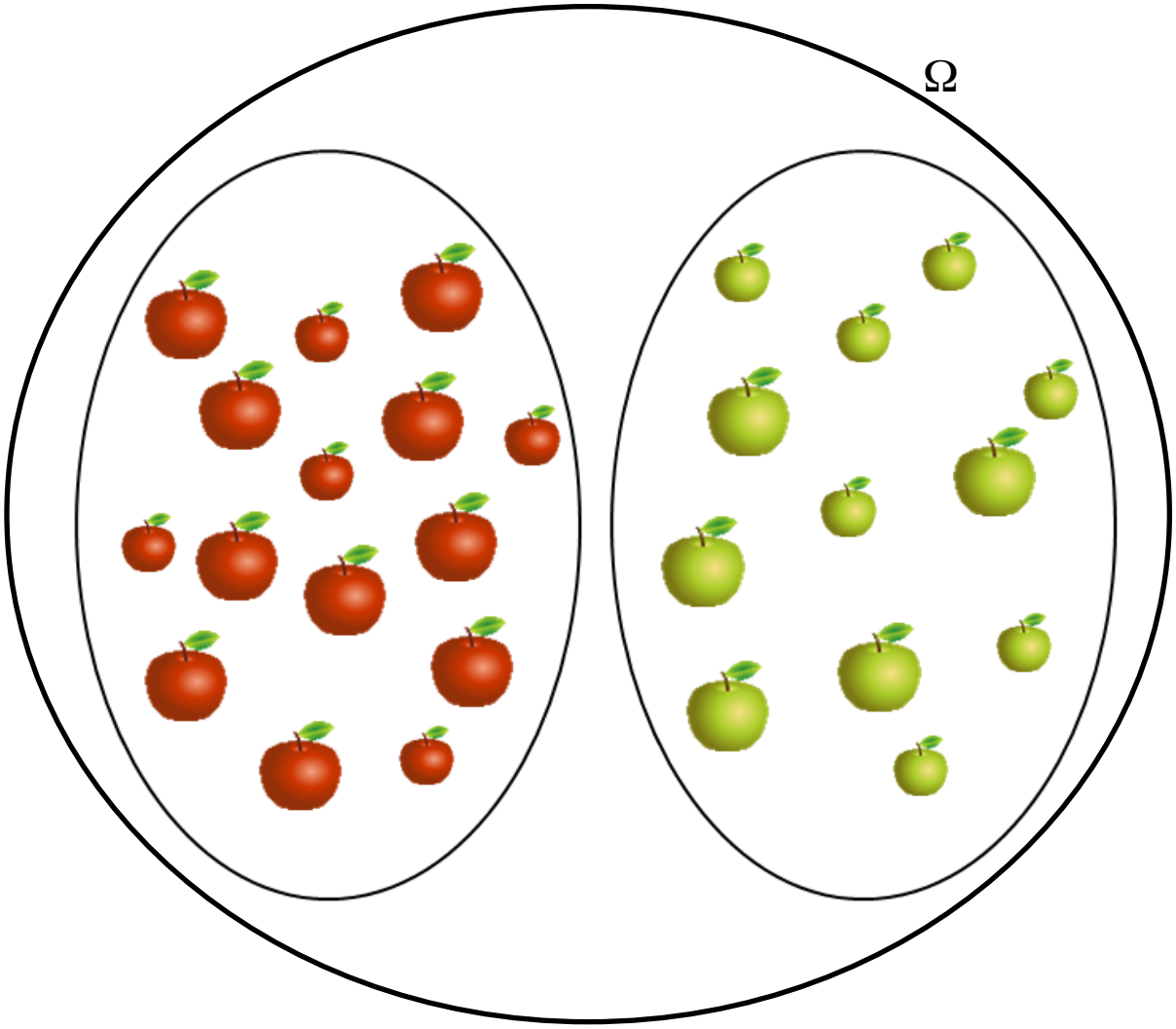
集合の要素は複数で多様
日本人は均一性を重視する傾向
異なる要素は除外する傾向



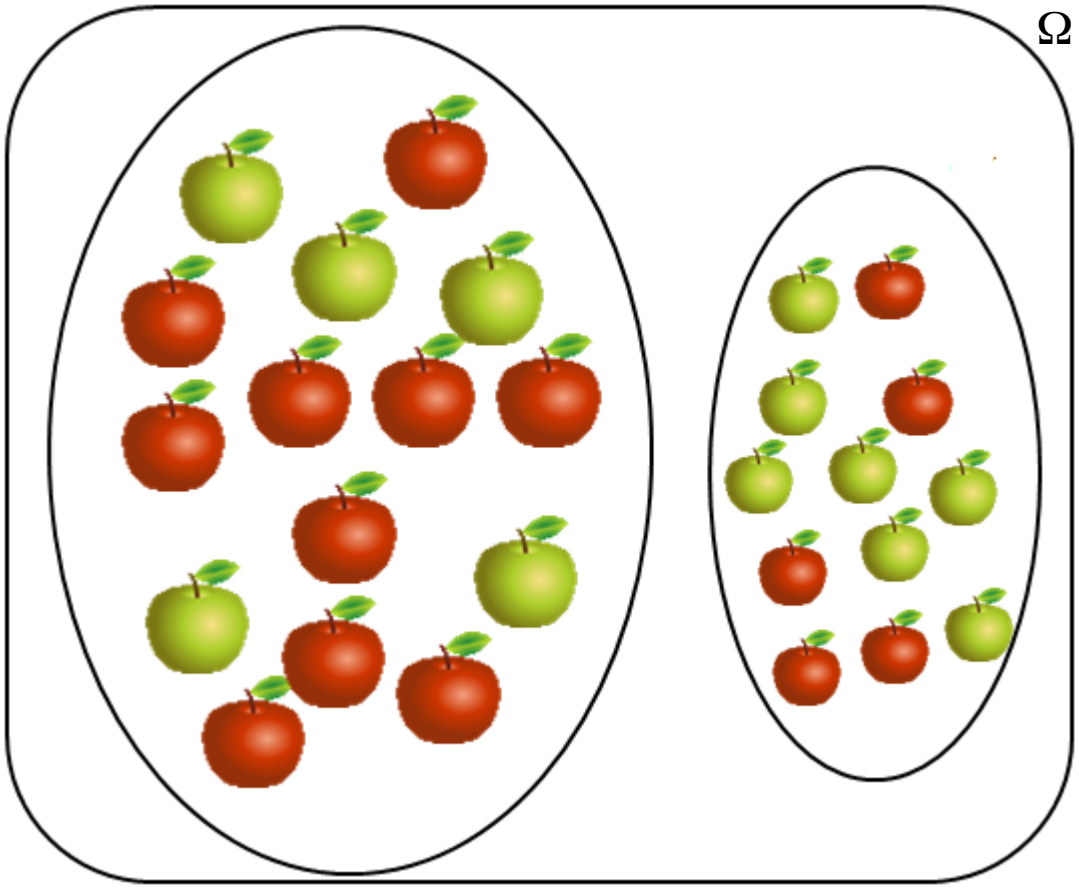
すべて均一なら集合の概念は不要
異質な物は除外して別途考える

「赤いリンゴ」は、特定の要素でも、
任意の要素でも集合でもかまわない

日本の傾向

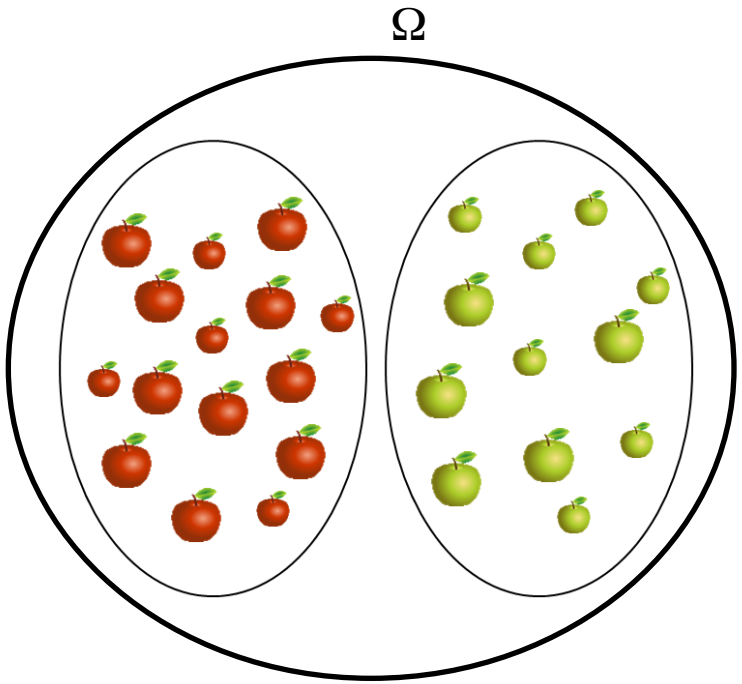


赤いリンゴと、青いリンゴの「部分集合（subsets）」を作る



大きいリンゴと、小さいリンゴの「部分集合」を作る

多くの統計的解析は集合間の関係を取り扱う



Ωと部分集合の関係

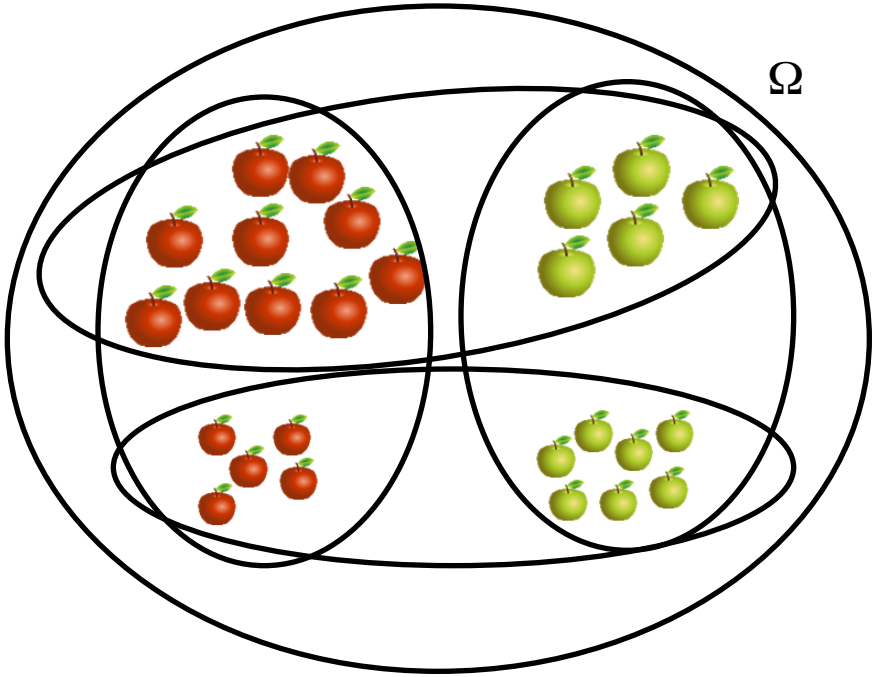
赤いリンゴの割合は = 15/27 = 56%

オッズは 15/12

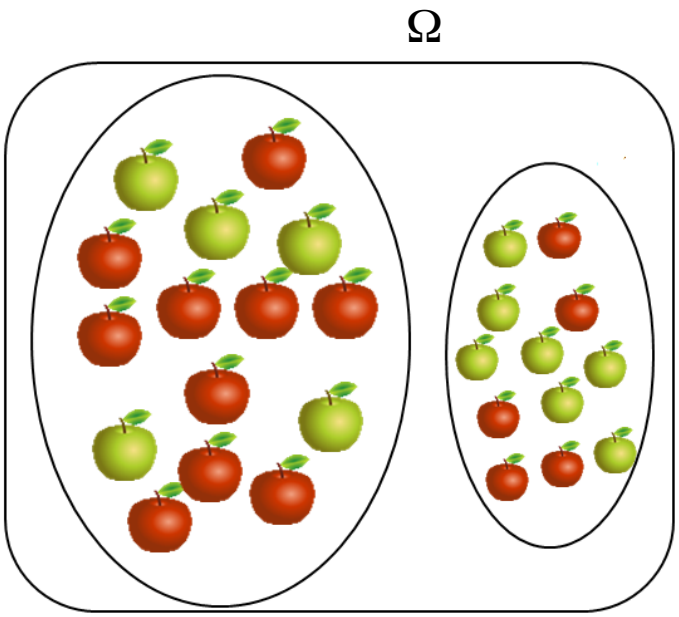
二種類の部分集合（subsets）の関係

赤いリンゴの中で大きいリンゴの割合は10/15 = 67%

青いリンゴの中で大きいリンゴの割合は5/12 = 42%



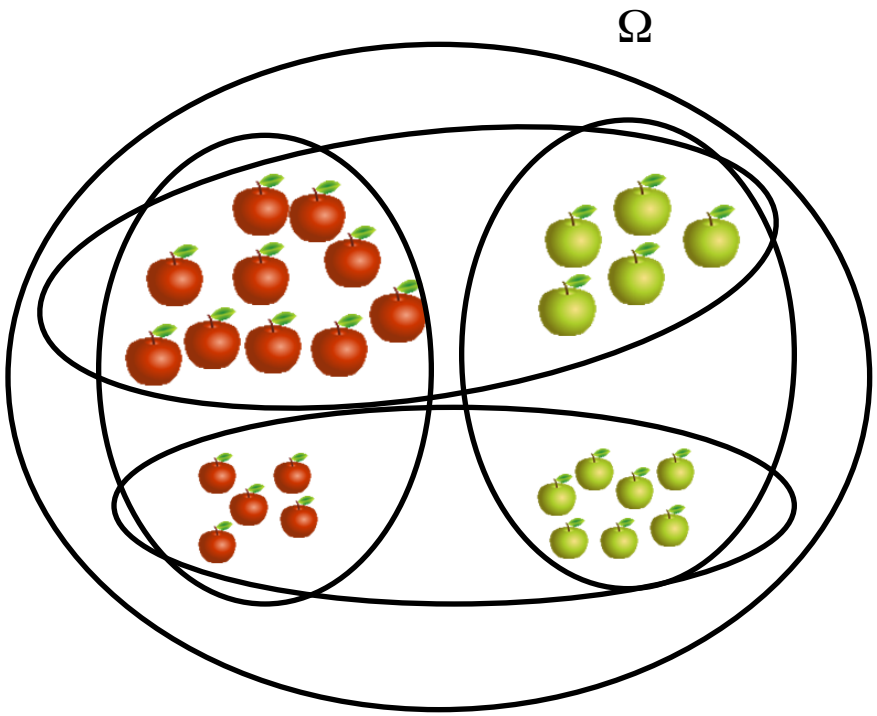
多くの統計的解析は集合間の関係を取り扱う



Ωと部分集合の関係

大きいリンゴの割合は15/27 = 56%

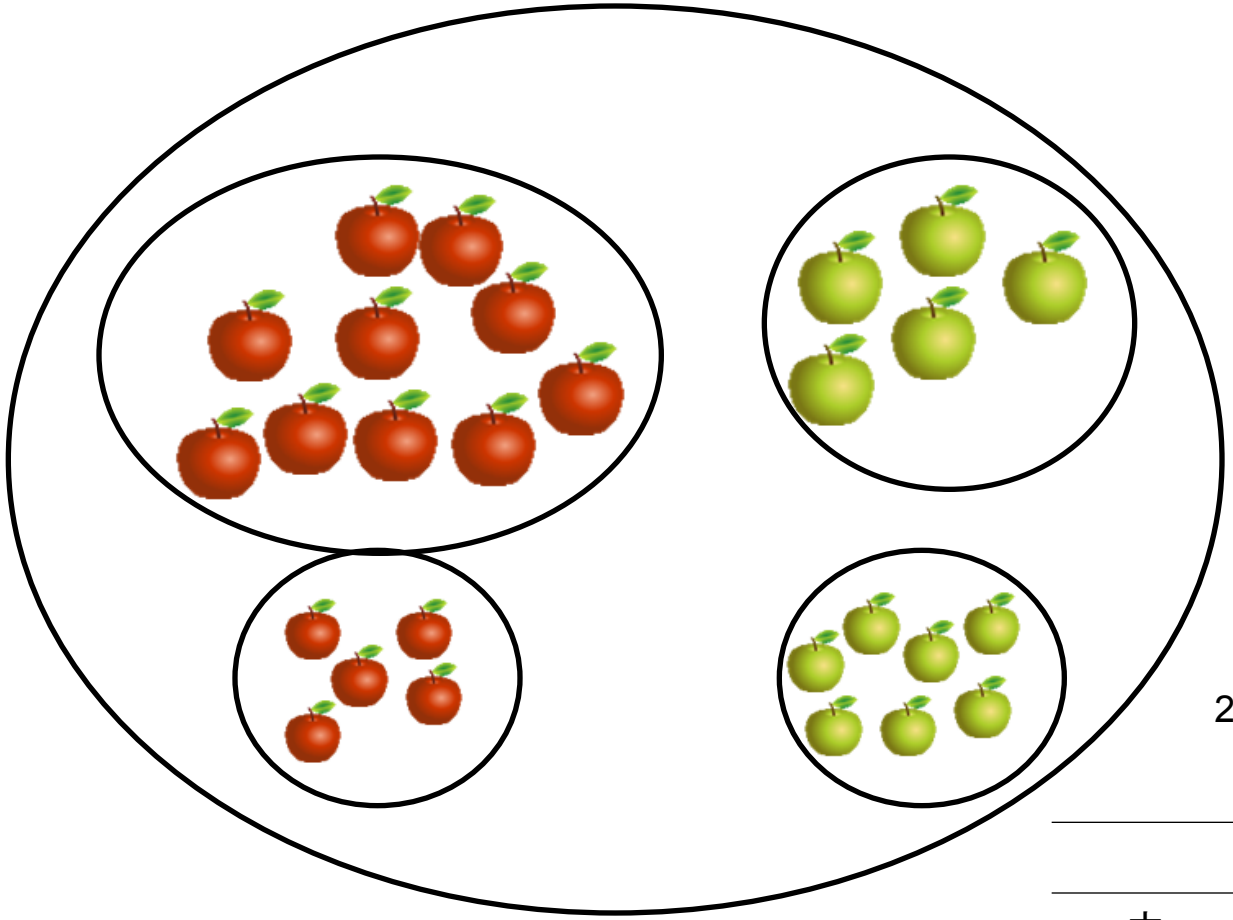
オッズは 15/12,



大きいリンゴの中で赤いリンゴの割合は10/15 = 67%

小さいリンゴの中で赤いリンゴの割合は5/12 = 42%

Ωは4つの互いに排他で、併合するとΩになる集合に分けられる



2 x 2のcontingency tableができる

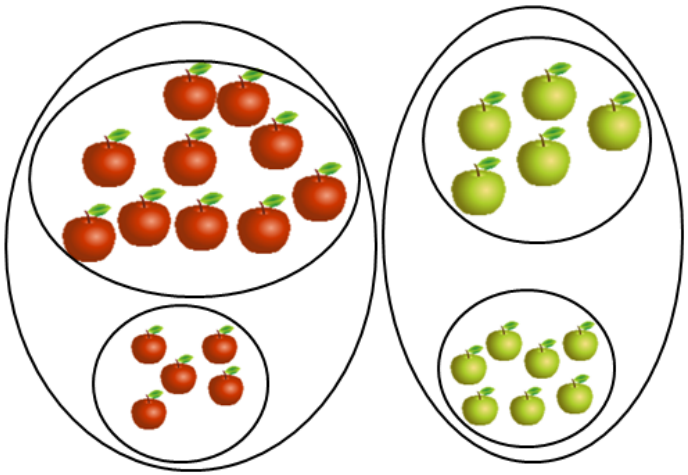
	赤	みどり
大	10	5
小	5	7

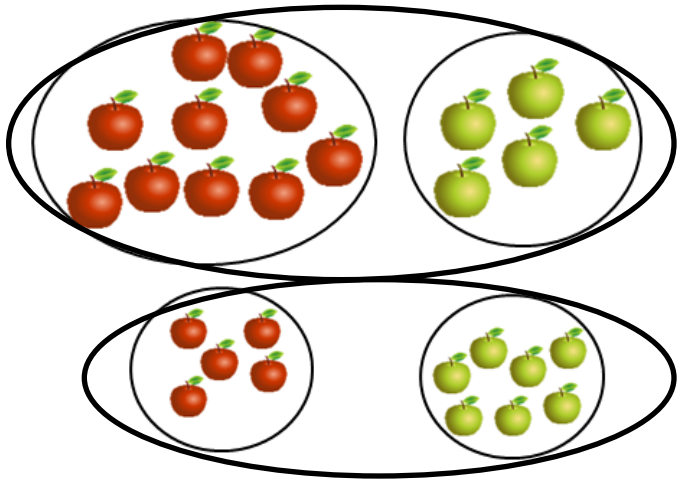
Ω is partitioned into 4 subsets that are all mutually exclusive.

4つの集合を2つずつ併合すると、併合の仕方で、

1. 赤いリンゴの集合と、青いリンゴの集合
2. 大きいリンゴの集合と、小さいリンゴの集合

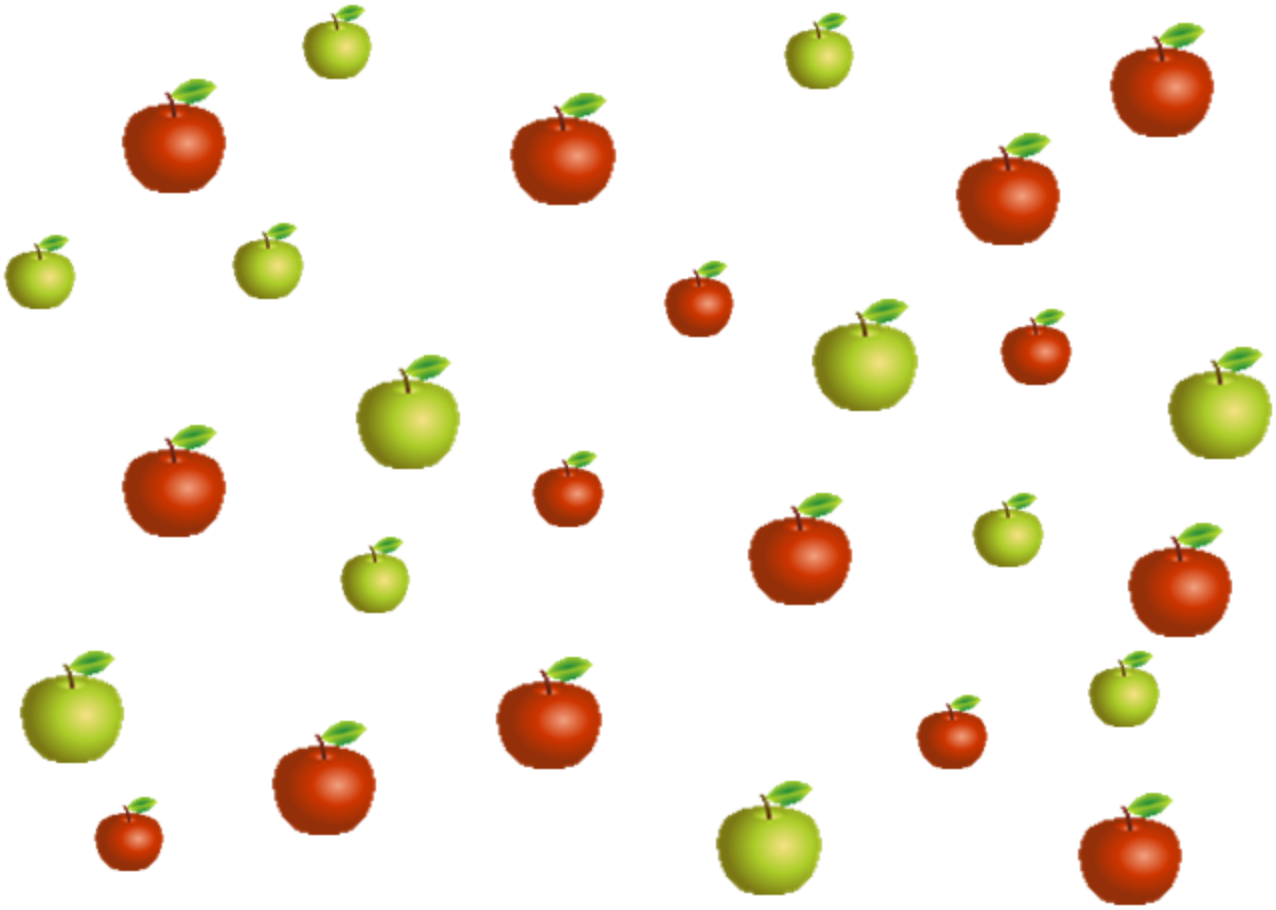
にわけられる。



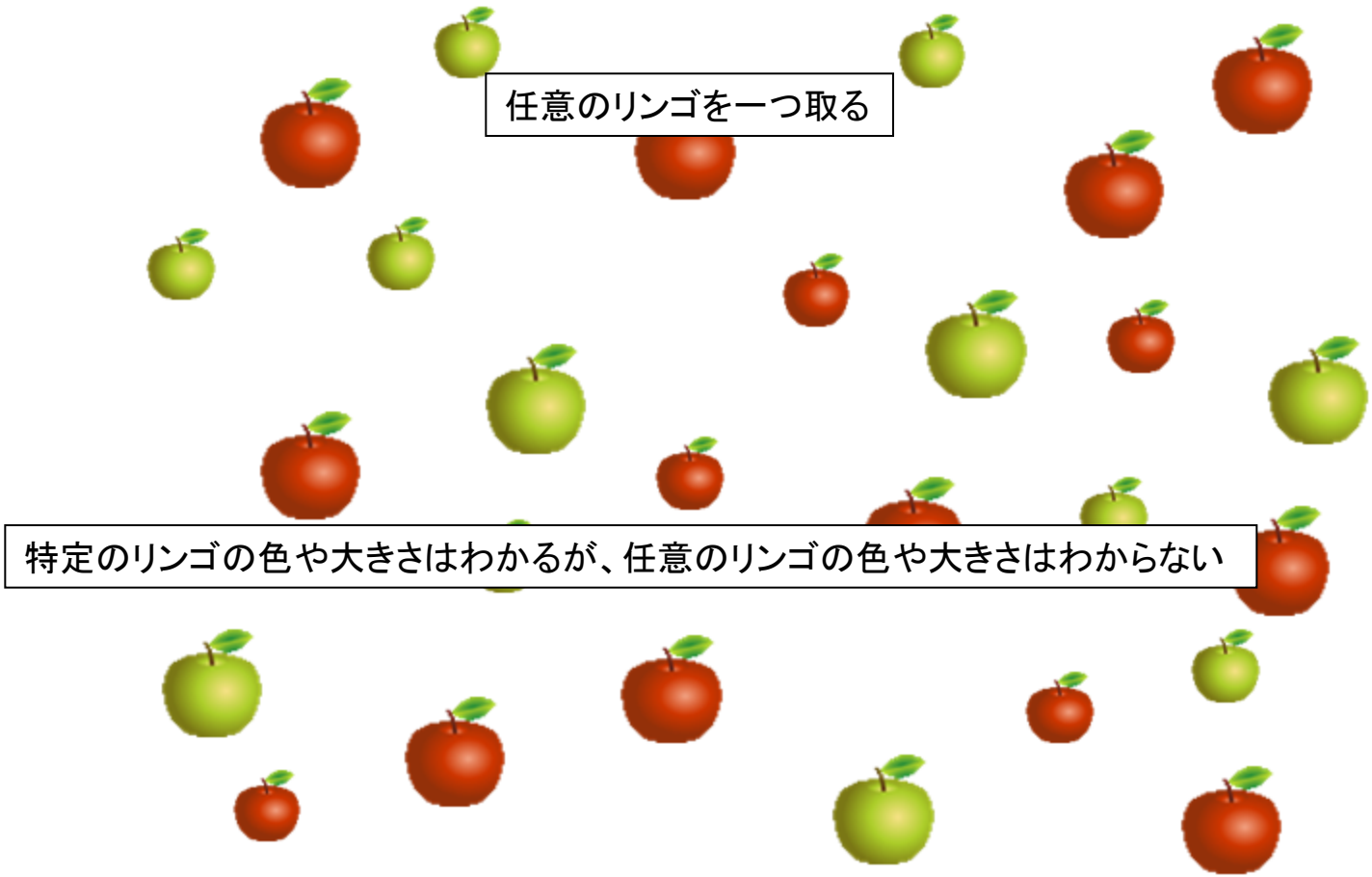


Ω is partitioned into 4 subsets that are all mutually exclusive.

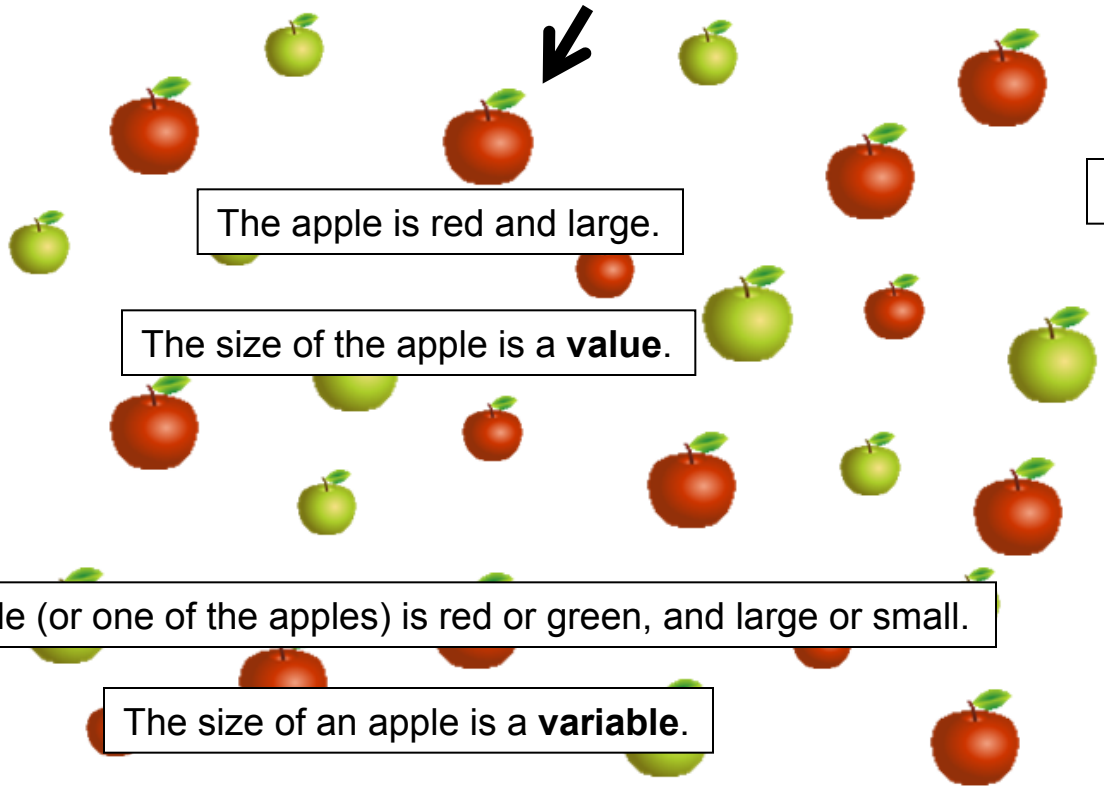
今度は、個々のリンゴを考える



今度は、個々のリンゴを考える



定冠詞、不定冠詞、複数と単数の違いを常に認識



The apple is red and large.

定冠詞 → 値

The size of the apple is a **value**.

An apple (or one of the apples) is red or green, and large or small.

不定冠詞 → 変数

The size of an apple is a **variable**.

日本語では、集合内の任意の対象物（a ….）を考える習慣が弱い

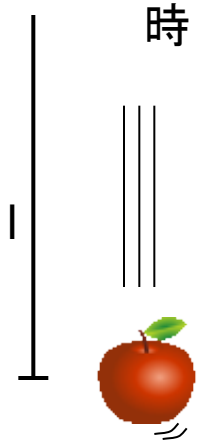
任意の対象物は「variable（変数）」を持つ

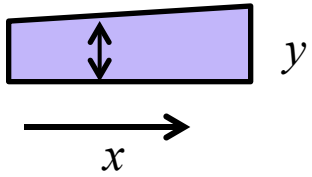
変数は時間的、空間的に動く対象物に限らない

One of, a copy ofは

日本語に無い表現

対象物により異なる対象を「変数」と捉える習慣が必要





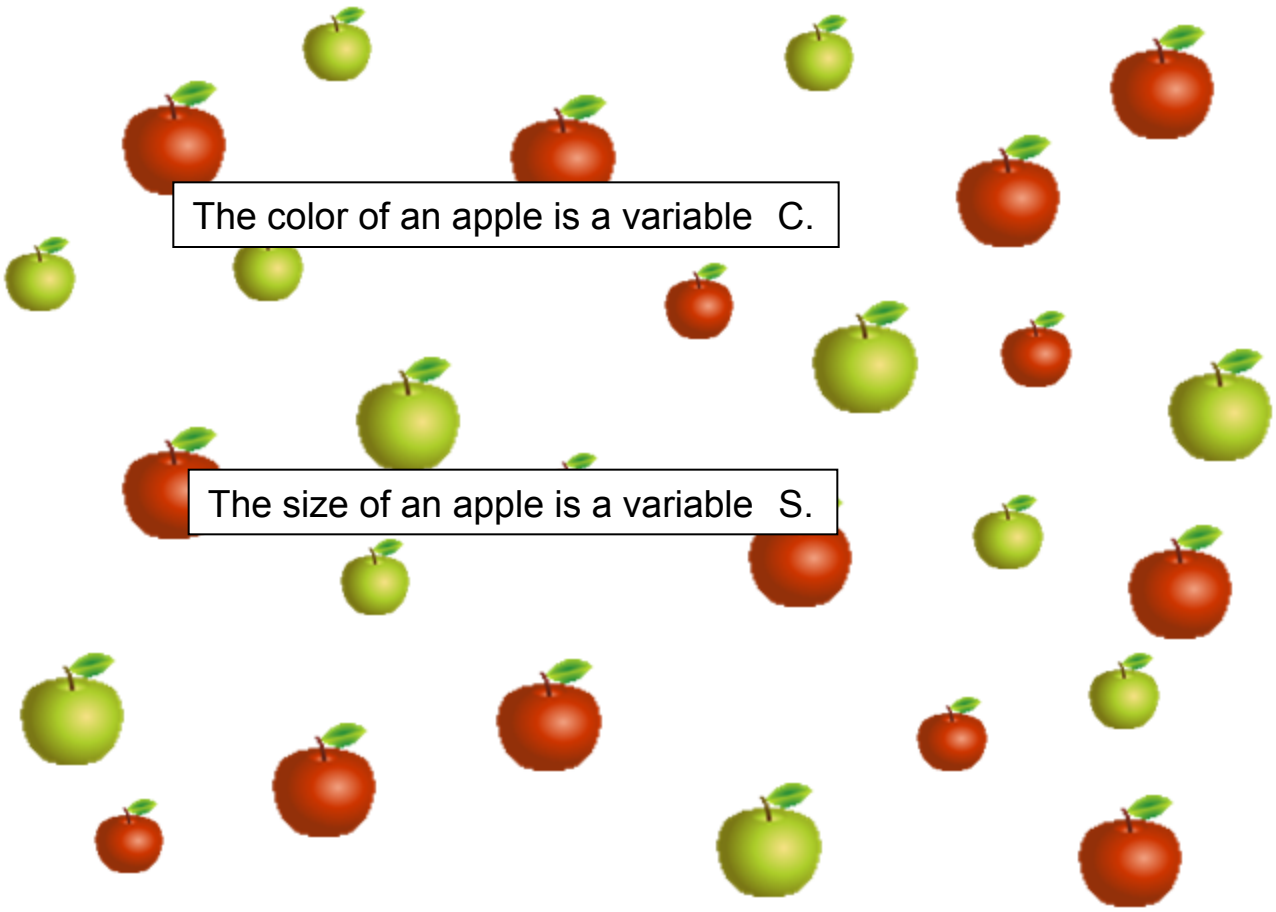
時間により変化する対象も「変数」

空間により変化する対象も「変数」

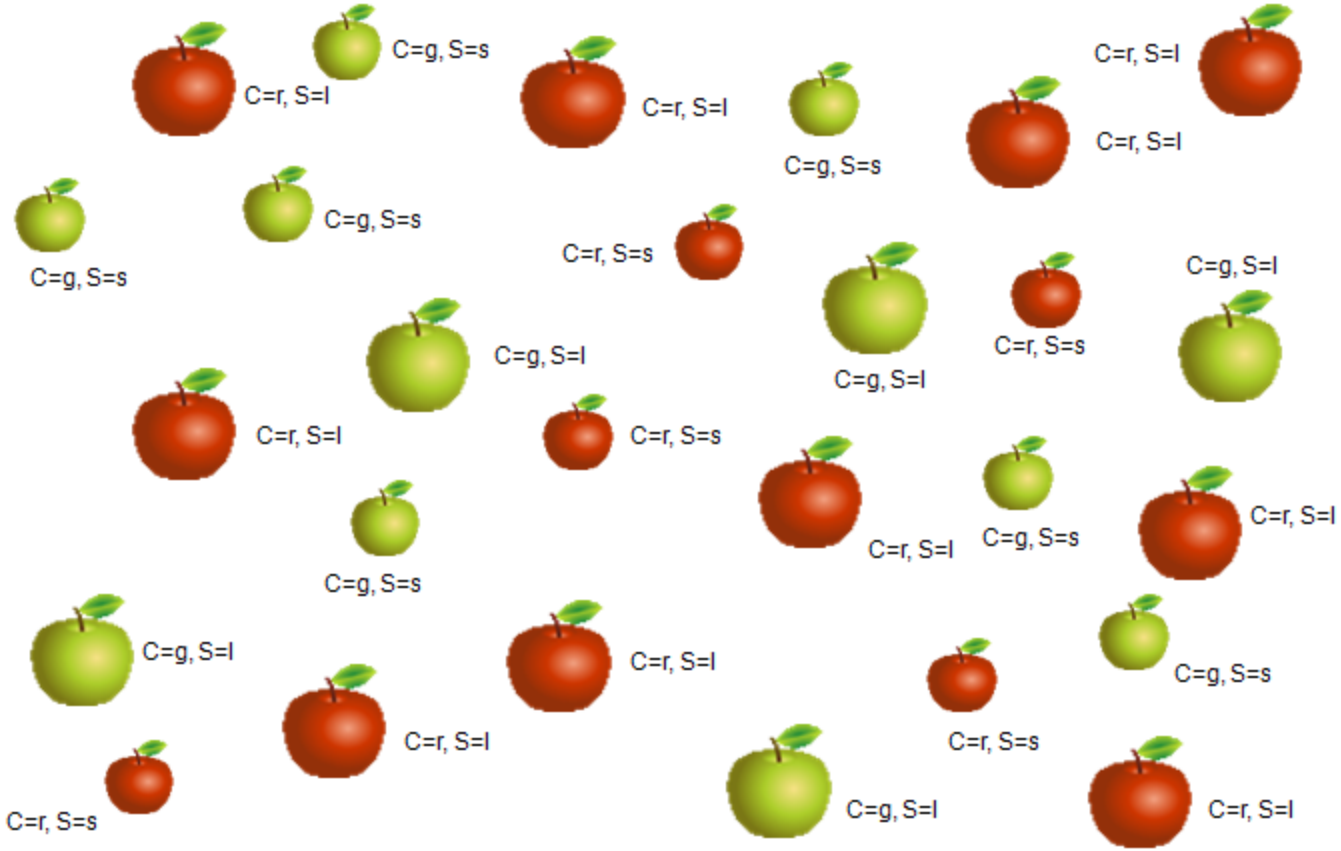
人により異なる対象も「変数」

人々は時や空間のように整然と並んでいないところが捉えにくい

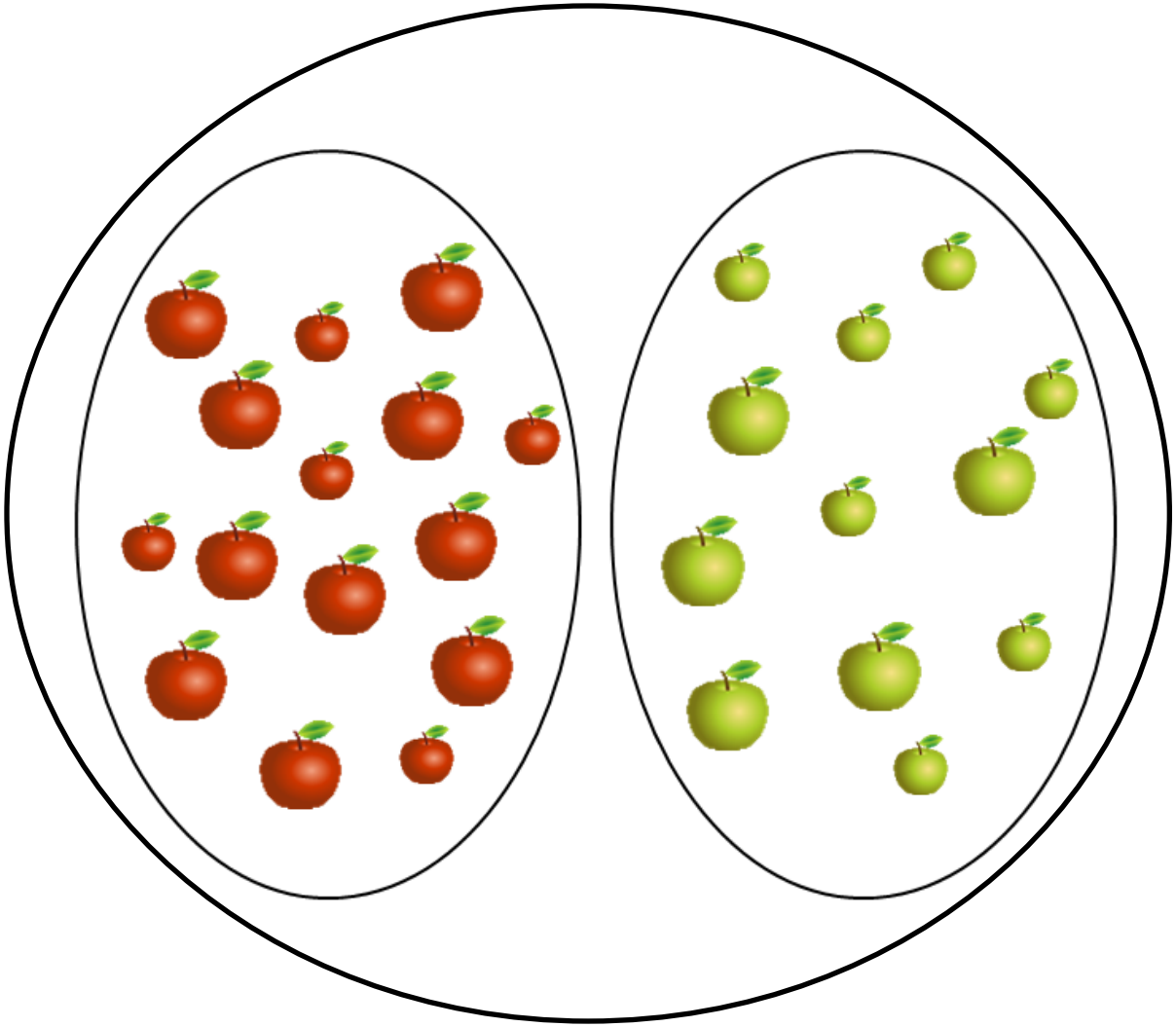
Two variables for a member



The color of an apple is a variable C; The size of an apple is a variable S.



An apple has a value (r or g) for a variable C, and a value (l or s) for a variable S.



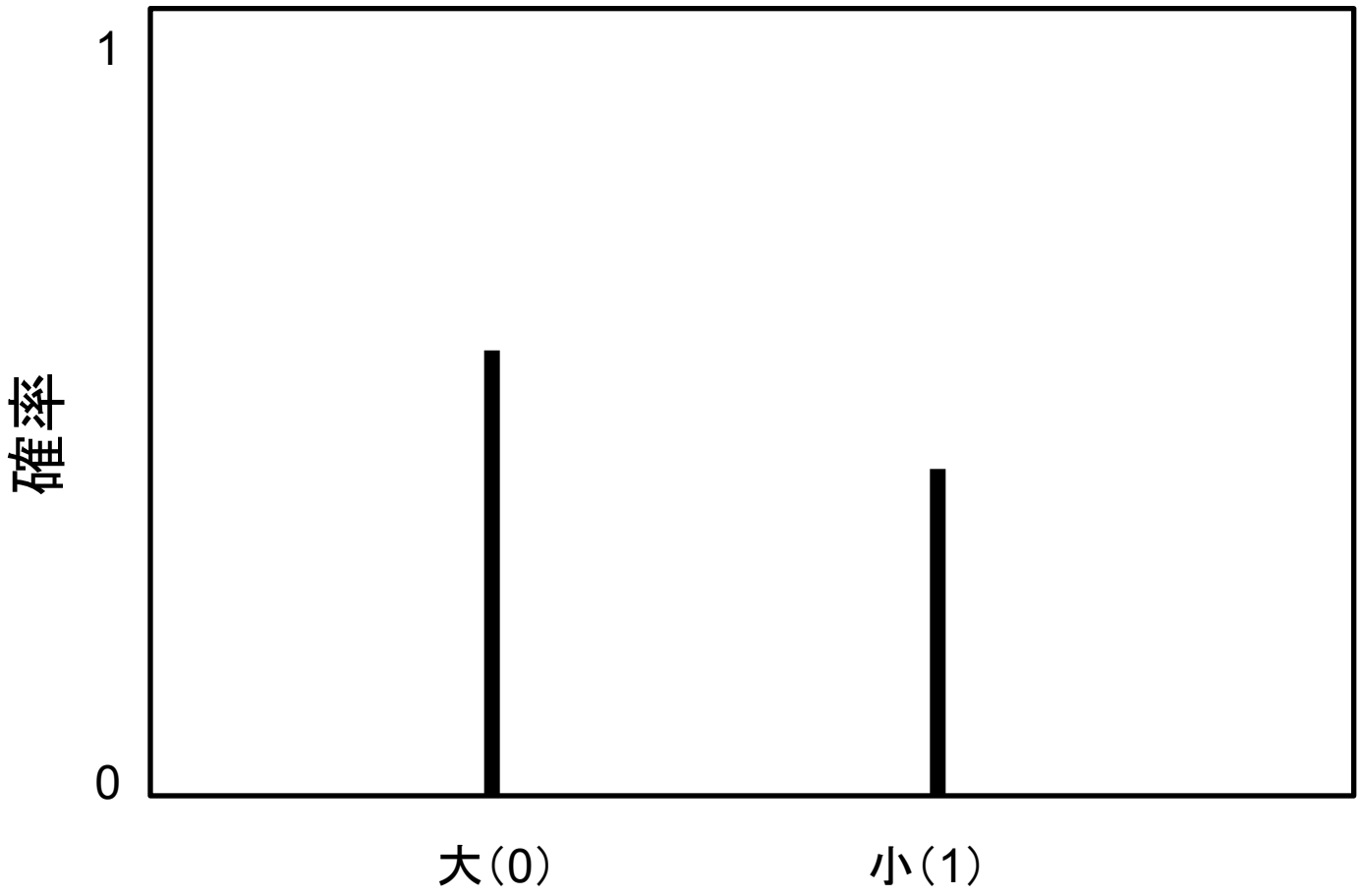
Collect apples with C = r

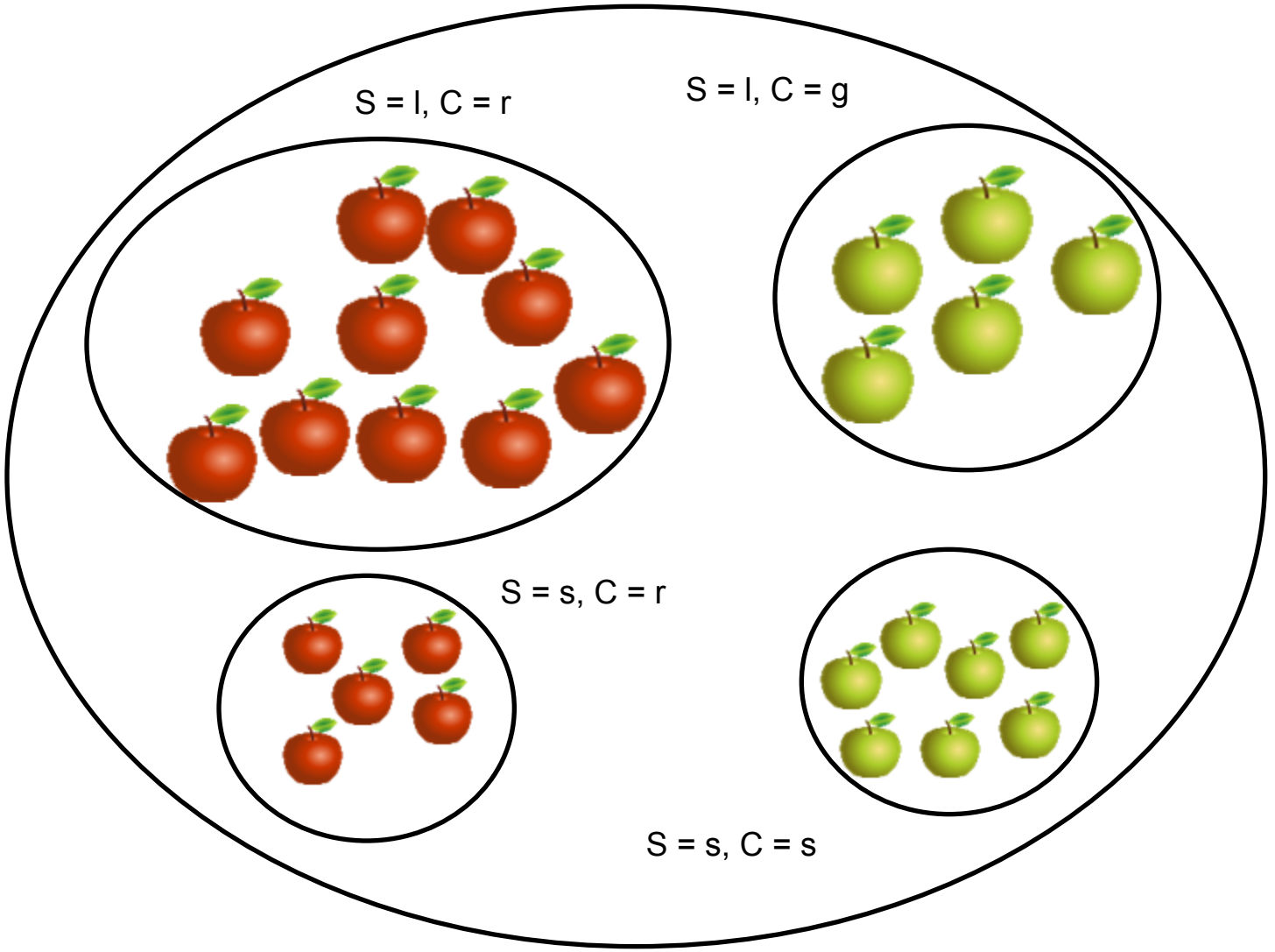
56%

Collect apples with C = g

44%

確率量関数

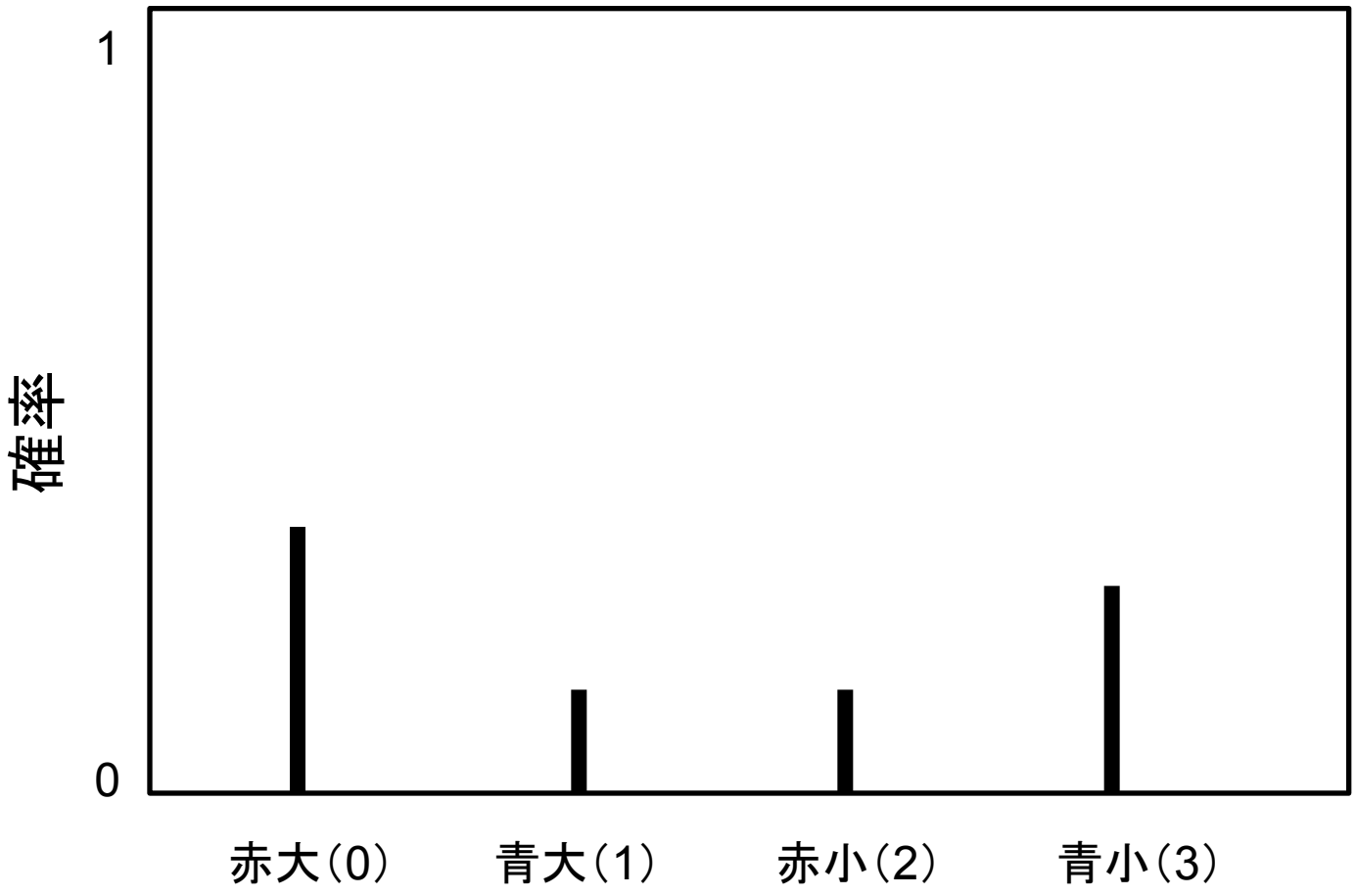


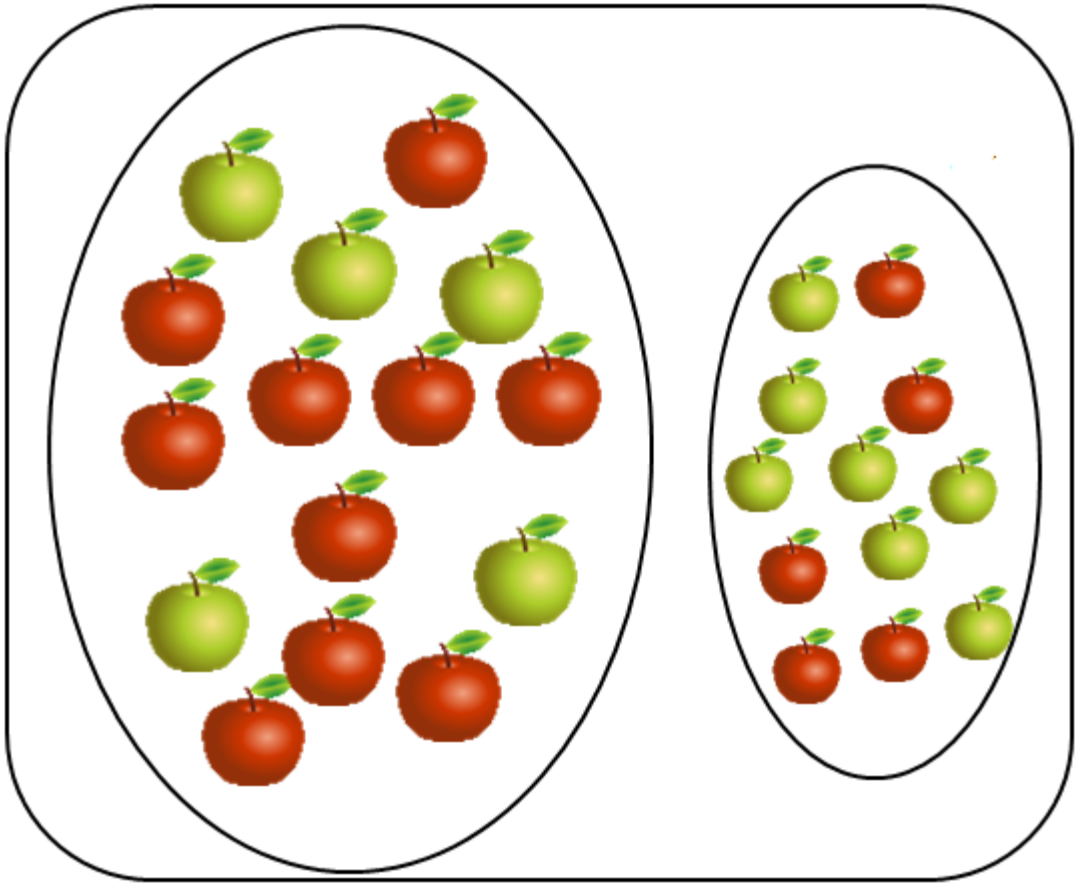


Collect apples using two variables.

A variable or a combination of variables define subsets.

確率量関数





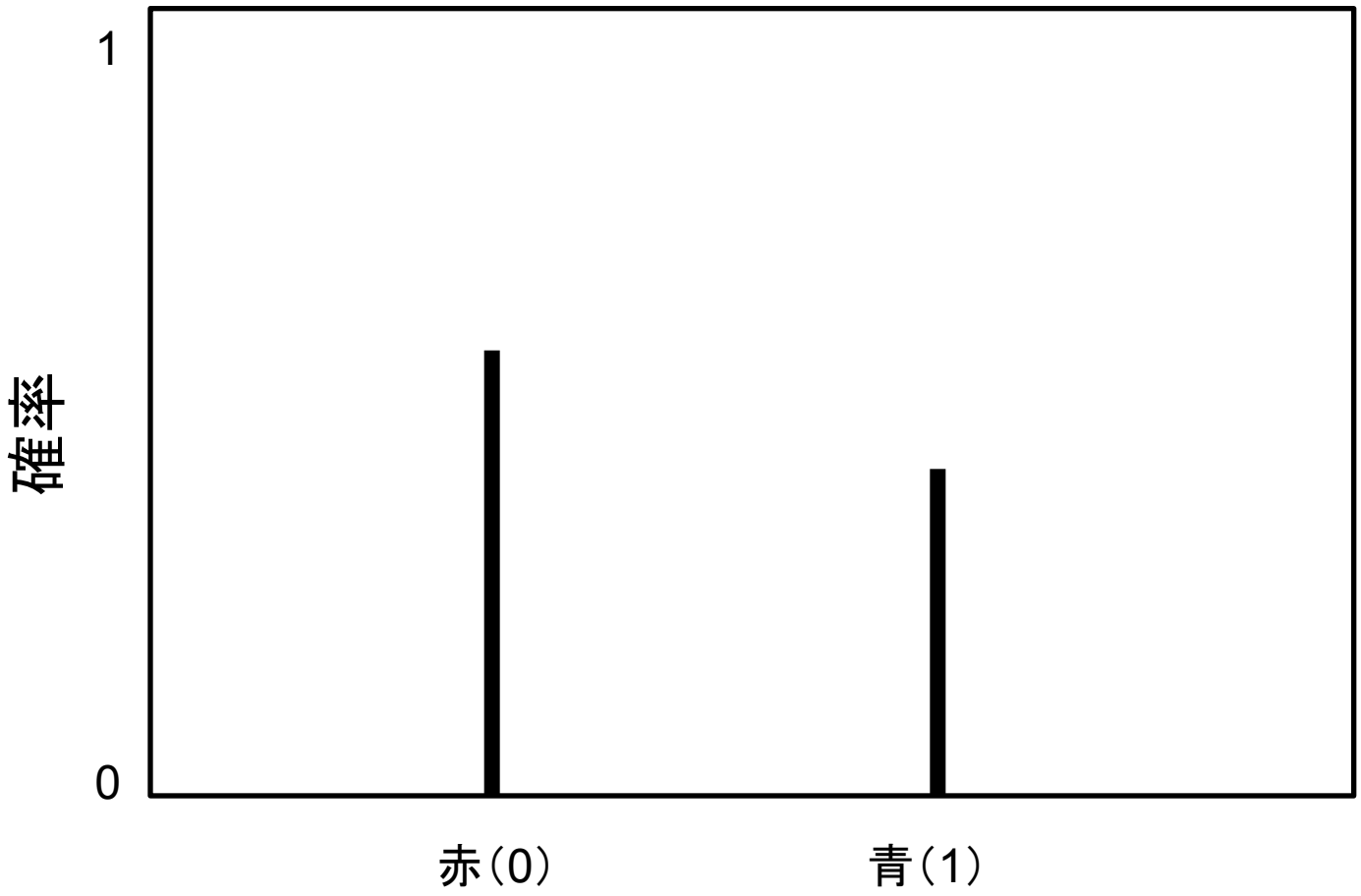
Collect apples with S = I

56%

Collect apples with S = s

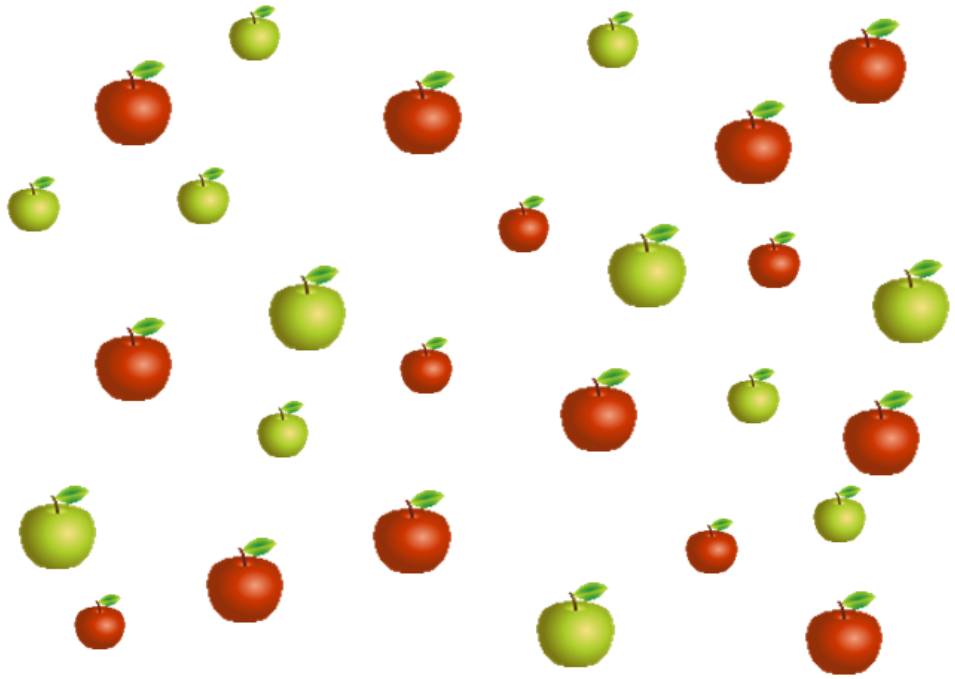
44%

確率量関数



任意のリンゴの色は？大きさは？

（集合から得られた割合を、確率に応用）



Proportion of red apples

Probability that an apple is red.

割合は部分集合の性質

確率は任意の個体の性質

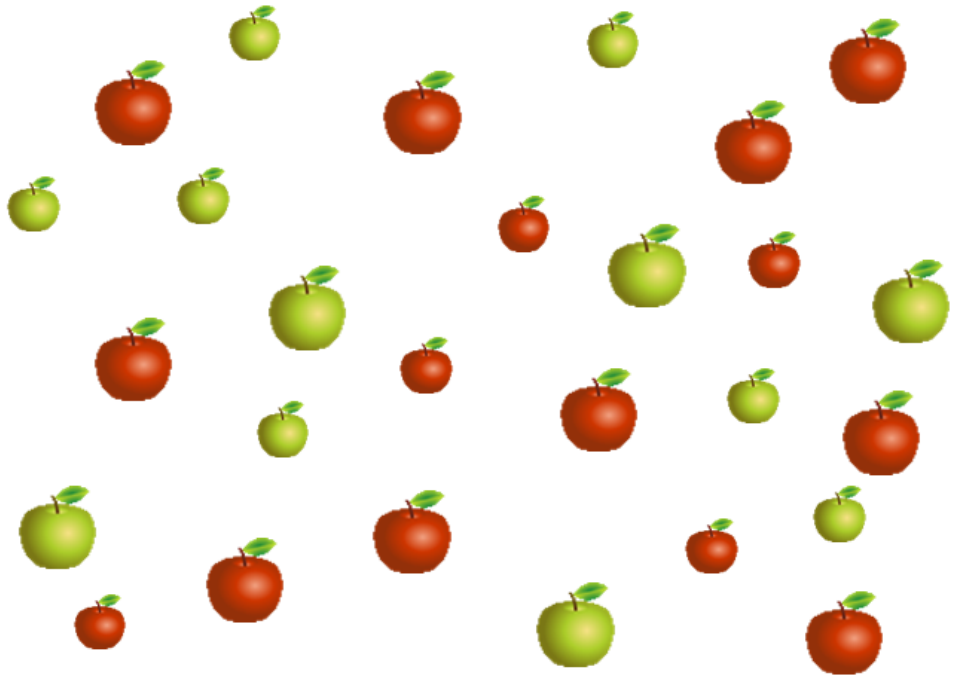
（確率は部分集合の関数）

任意のリンゴの色は？P(C = r) =15/27 = 56%, 任意のリンゴは大きい？ P(S = l) = 15/27 = 56%,

確率は特定のリンゴの性質ではなく、「任意の」リンゴの性質

任意のリンゴの色は？大きさは？

（集合から得られた割合を、確率に応用）



Probability that a large apple is red.

任意の赤いリンゴは大きい？

$P(S = l \mid C = r) = 10/15 = 67\%$

任意の大きいリンゴは赤い？

$P(C = r \mid S = l) = 10/15 = 67\%$

条件付き確率

1. 任意の大きいリンゴが赤い確率
2. 大きいリンゴの集合の任意のリンゴが赤い確率
3. 大きいリンゴの集合の中の赤いリンゴの割合

$P(C = r \mid S = l) = P(C = r, S = l)/P(S = l)$

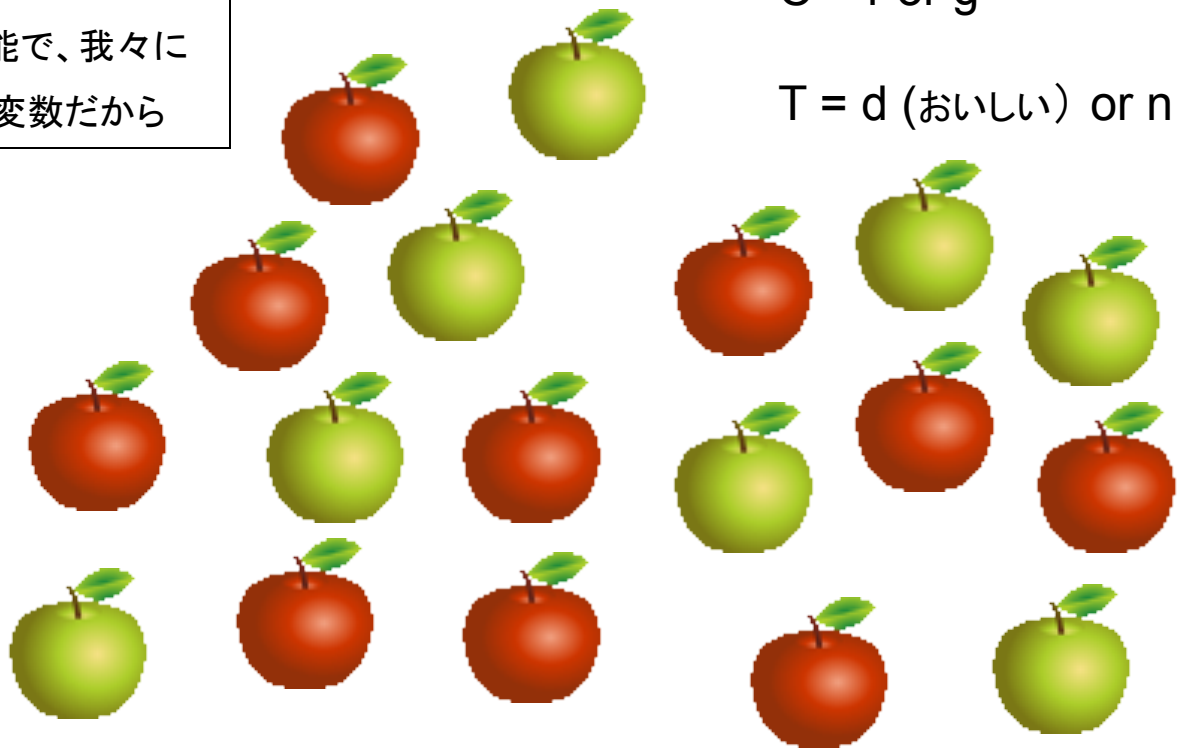
この3つを区別し、しかも同一視できる事が重要

赤いリンゴはおいしいか？（今度は無関心ではいられない）

Tが観察不能で、我々に興味がある変数だから

C = r or g

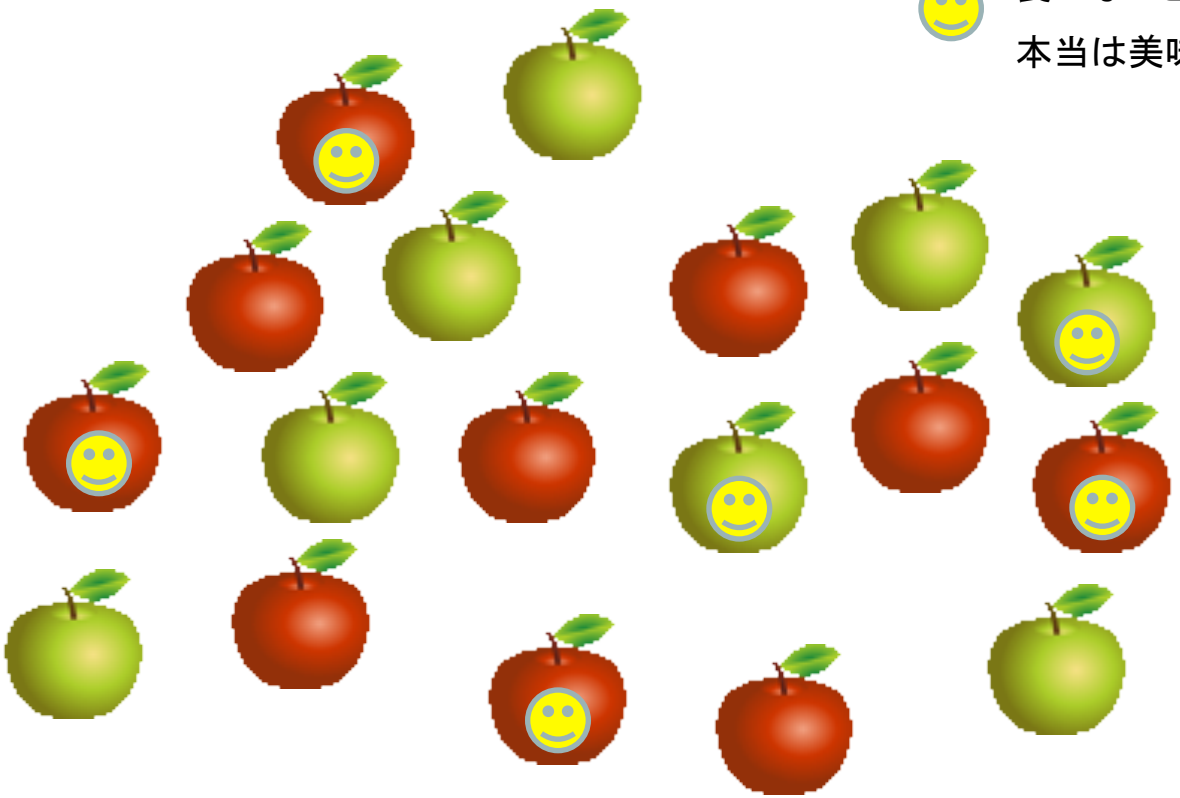
T = d (おいしい) or n (おいしくない)



C: 観察できる変数、T: 観察できない変数

観察できない変数の予測こそ重要

多くの場合、統計、疫学の仕事はこのような物です



食べないとわからないけど、
本当は美味しいリンゴ

赤いリンゴの方がおいしい
確率が高いように見える
（しかしホントかな？）

確率は「部分集合の確率」または「任意の要素の確率」

$P(T = d \mid C = r) = 4/10 = 0.4$

$P(T = d \mid C = g) = 2/8 = 0.25$

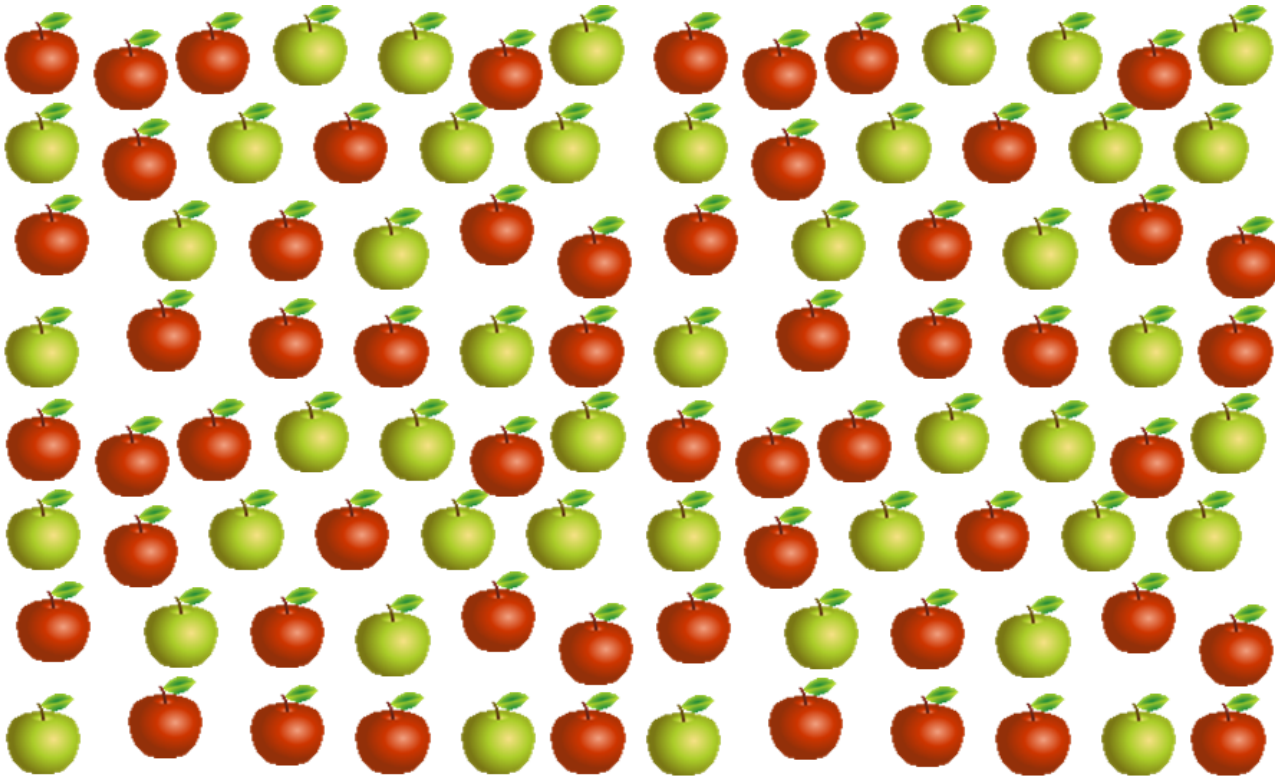
$P(T = d) = 6/18 = 0.33$

あなたは、どう考えますか？

1. 赤いリンゴでもおいしくない事もあるから意味が無い
2. 赤いリンゴの方がおいしい確率が高い

確率により、赤を選びますか、それとも
不確実なら科学はいらないと考えますか？

そんな少ない数じゃ信用できない、数を増やせば結果は変わるんじゃない？

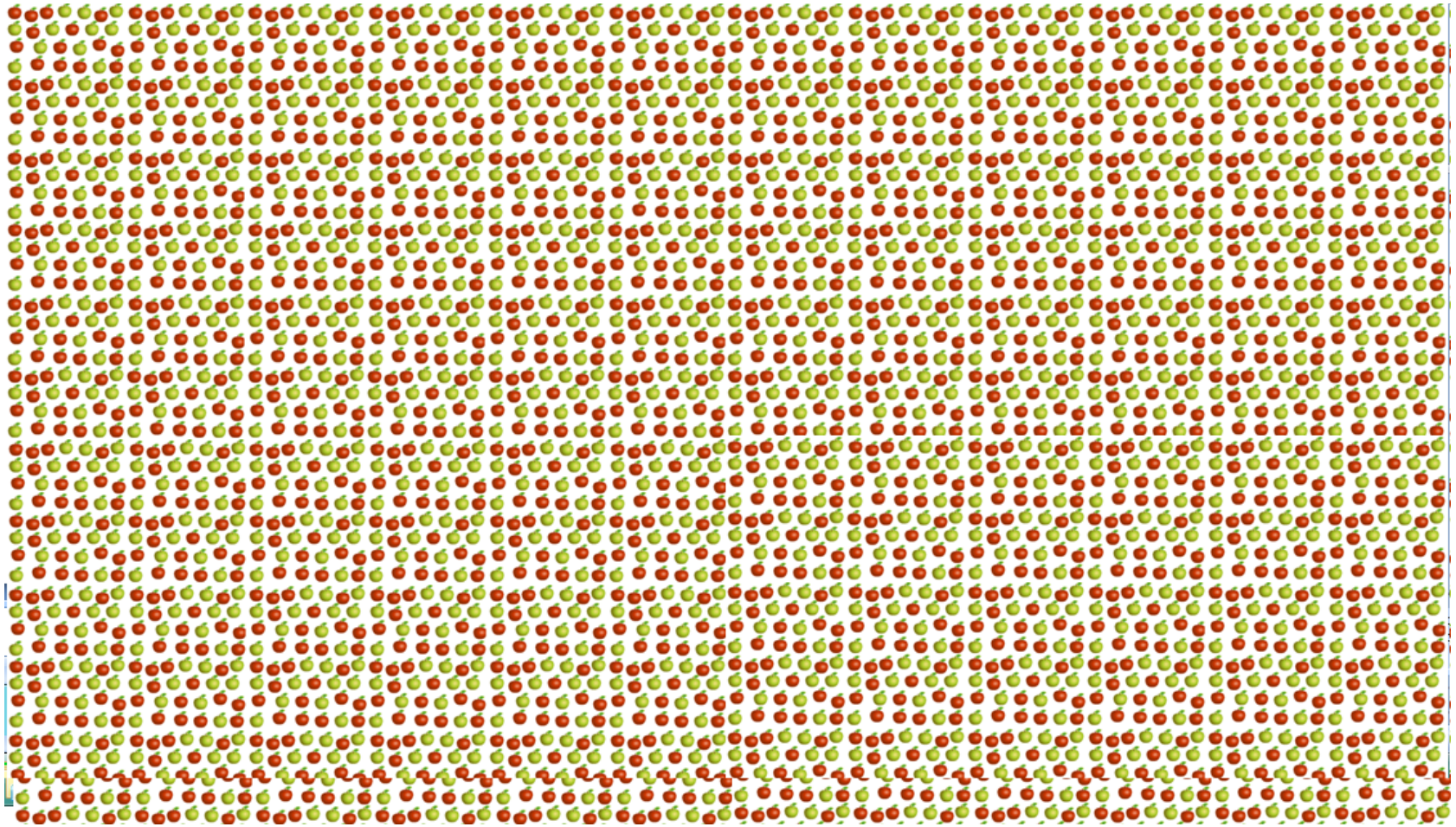


そりゃそうですよね、じゃあ、はい!!!

$P(T = d \mid C = r) = 0.42$

$P(T = d \mid C = g) = 0.28$

それでもだめ、じゃあ、はい!!!



無限大の集団から得られた割合（確率 = パラメータ）

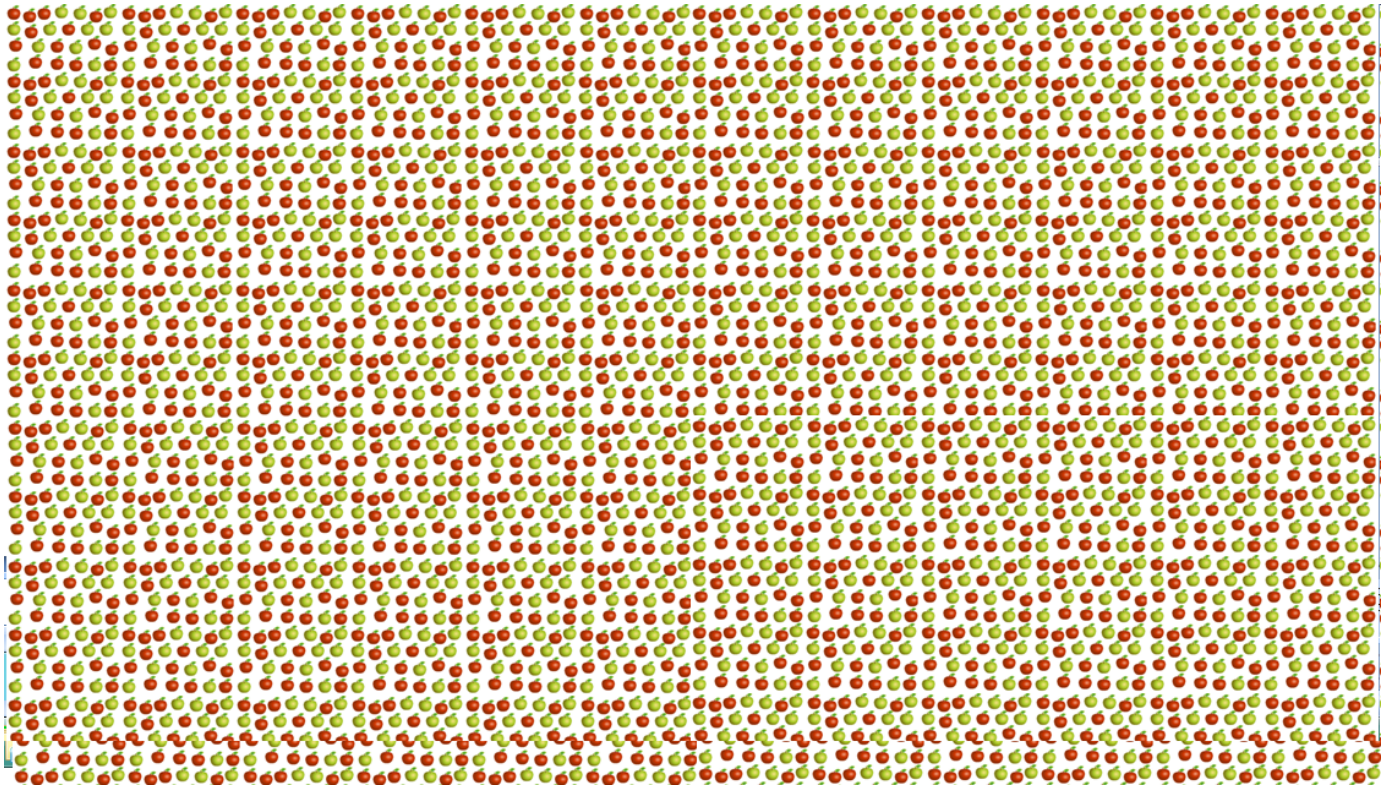
数を増やせば割合は収束する（大数の法則）

$P(T = d \mid C = r) = 0.424$

$P(T = d \mid C = g) = 0.275$

赤が薬を服用した人、緑が服用しなかった人で、おいしいと言うのが「病気が良くなる事」で、おいしくないと言うのが「よくならない事」だとしたら、

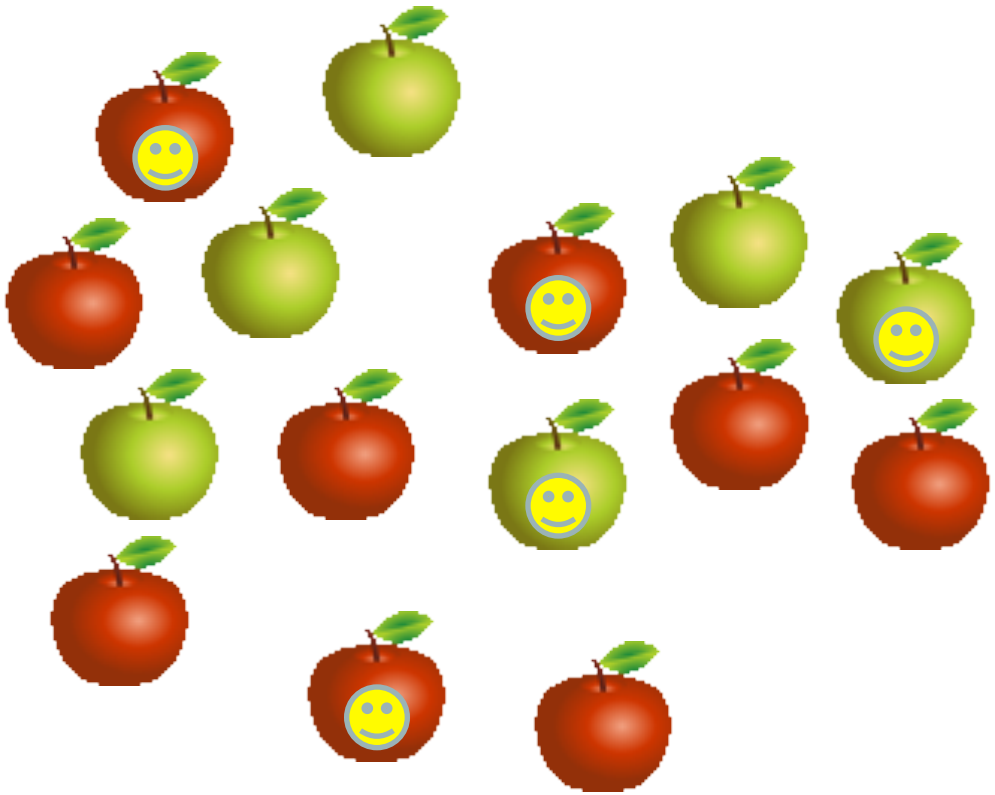
この薬を服用しますか？　薬を服用しても病気が良くならない人もいるから無意味？



$P(T = d \mid C = r) = 0.424$

$P(T = d \mid C = g) = 0.275$

「おいしい、おいしくない」とリンゴの色は独立



任意のリンゴを食べても
任意の赤いリンゴを食べても
任意の青いリンゴを食べても
おいしい確率は同じ

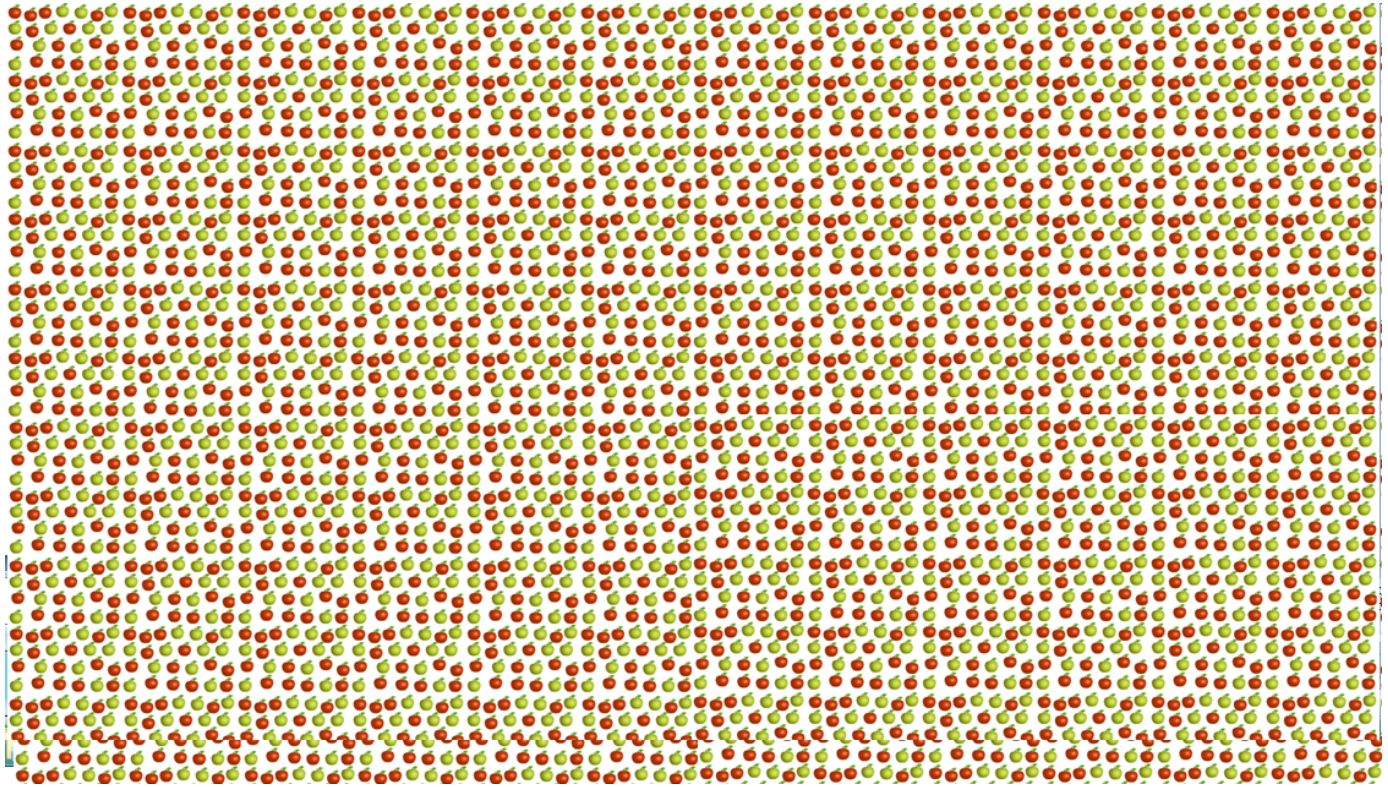
$P(T = d \mid C = r) = 3/9 = 0.33$

$P(T = d \mid C = g) = 2/6 = 0.33$

$P(T = d) = 5/15 = 0.33$

あなたは、どのリンゴを選びますか？

あなたはどうしますか、この薬を承認しますか



任意の服用した人も、任意の服用して
いない人も、よくなる確率は同じ
薬の服用と、よくなることは「独立」

$P(T = d \mid C = r) = 0.33$

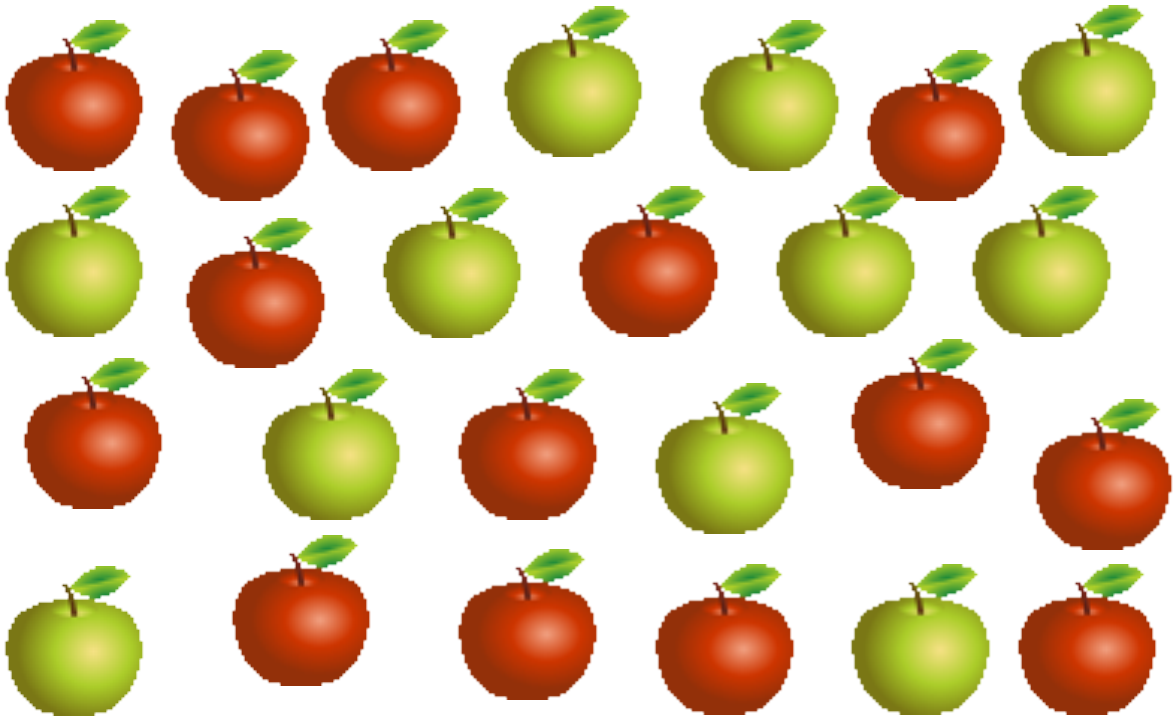
$P(T = d \mid C = g) = 0.33$

$P(T = d) = 0.33$

リンゴにはポリフェノールが含まれています。

ポリフェノールの含量は多いほどいいです。

しかし、ポリフェノールはすべてのリンゴに含まれており、量が問題です



個々のリンゴの含むポリフェノールの量は変数Φで表わす事ができるが、Φは連続の値を取る

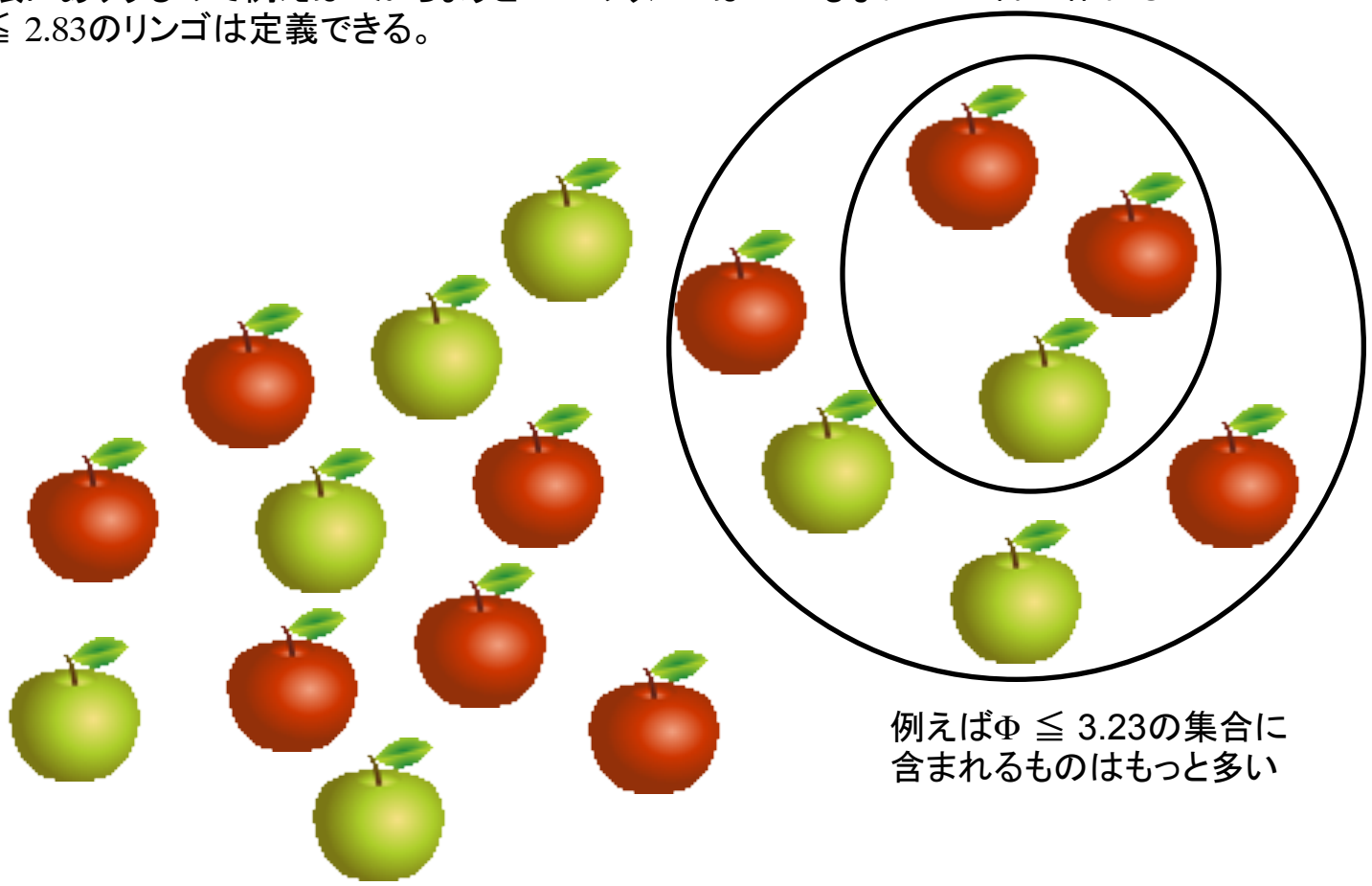
赤青、大小のような離散の変数では無い

可算名詞、非可算名詞の区別が無い日本語で
はこの大きな違いに気付きにくい

問題はΦ（ポリフェノールの量）によって部分集合を定義する事が難しいこと

Φの値は無限にありうるので例えばΦがちょうど2.83のリンゴは一つもない
しかし、Φ ≦ 2.83のリンゴは定義できる。

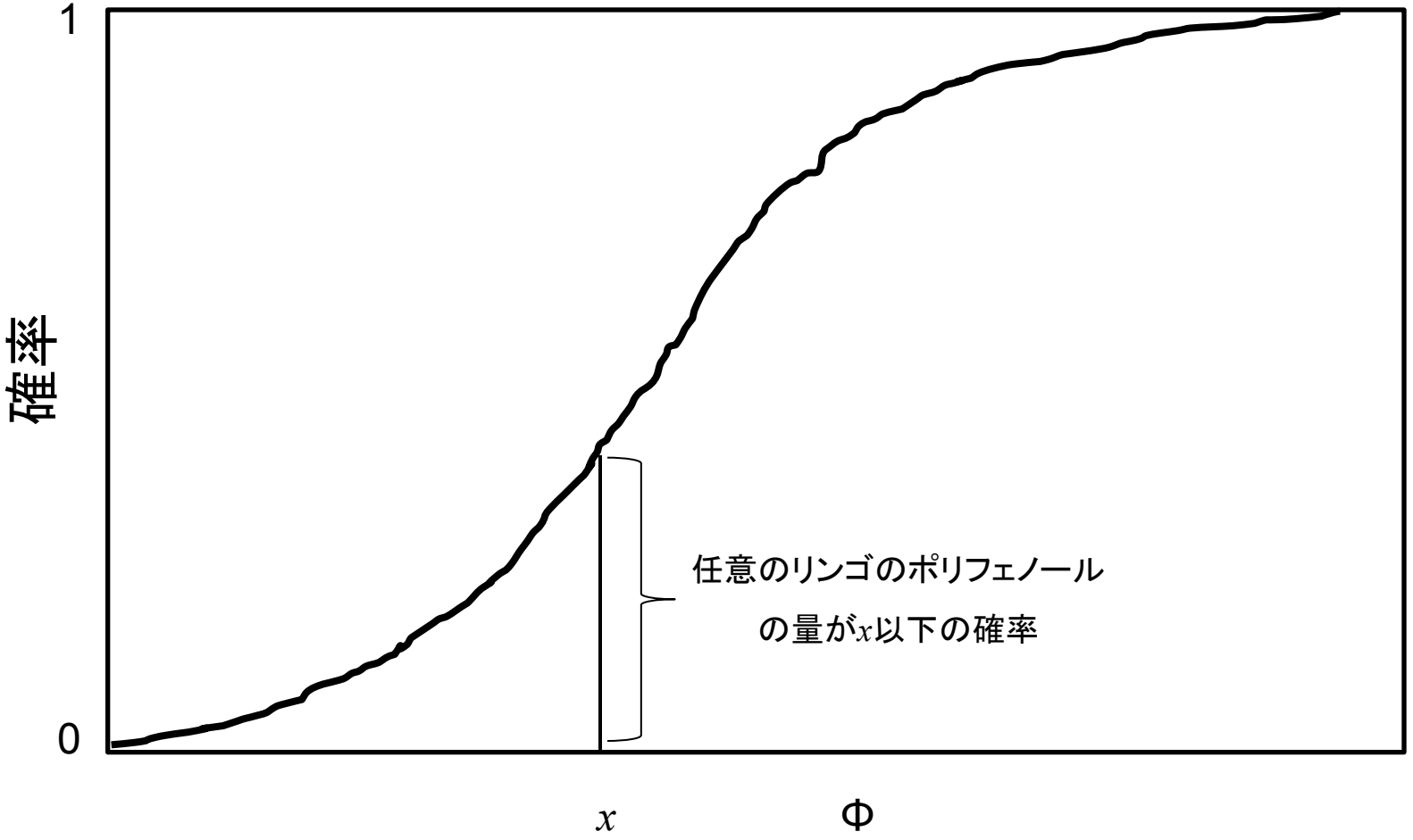
例えばΦ ≦ 2.83の集合は作れる



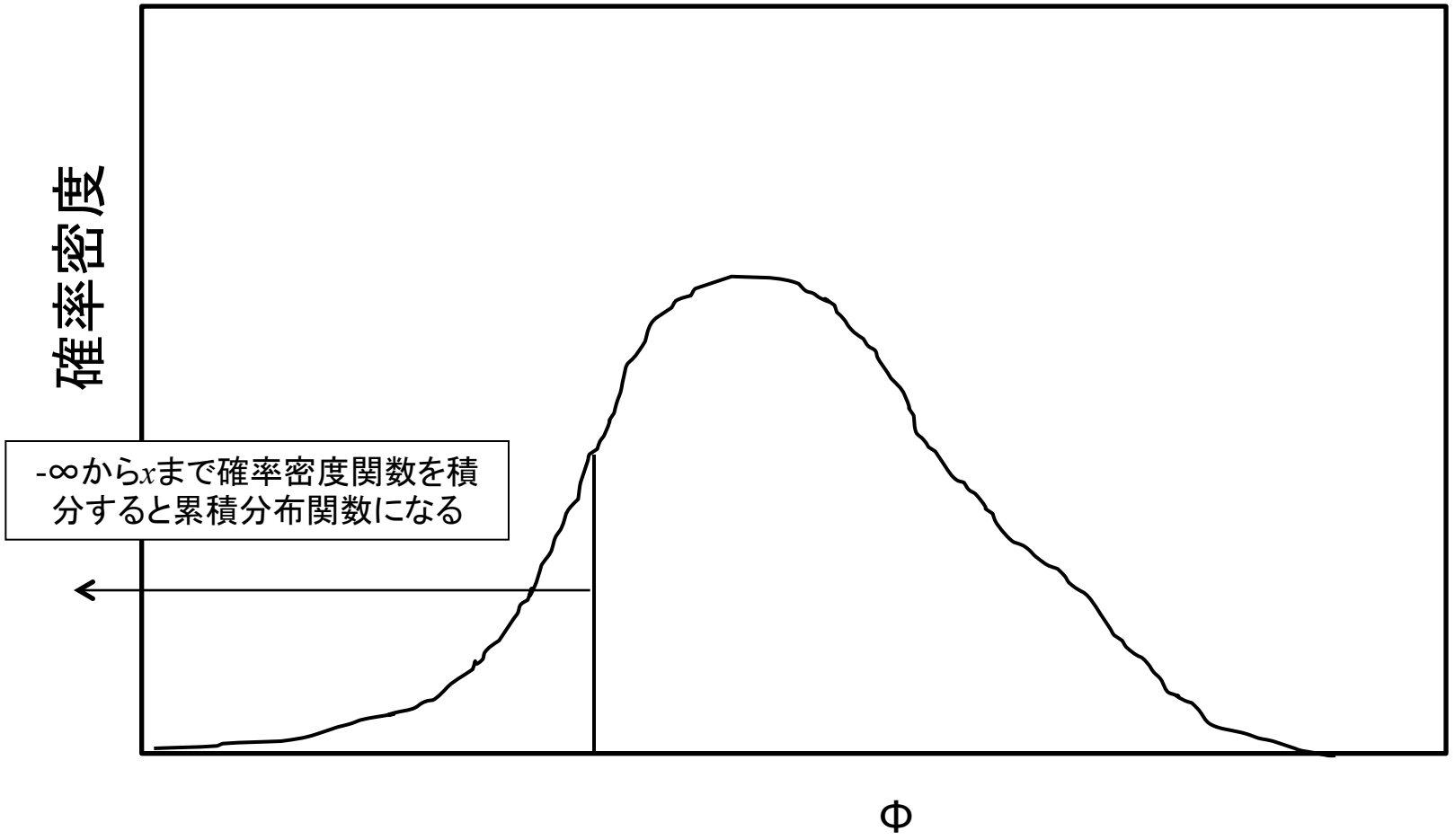
例えばΦ ≦ 3.23の集合に含まれるものはもっと多い

連続変数の場合も、集合が思考の原点

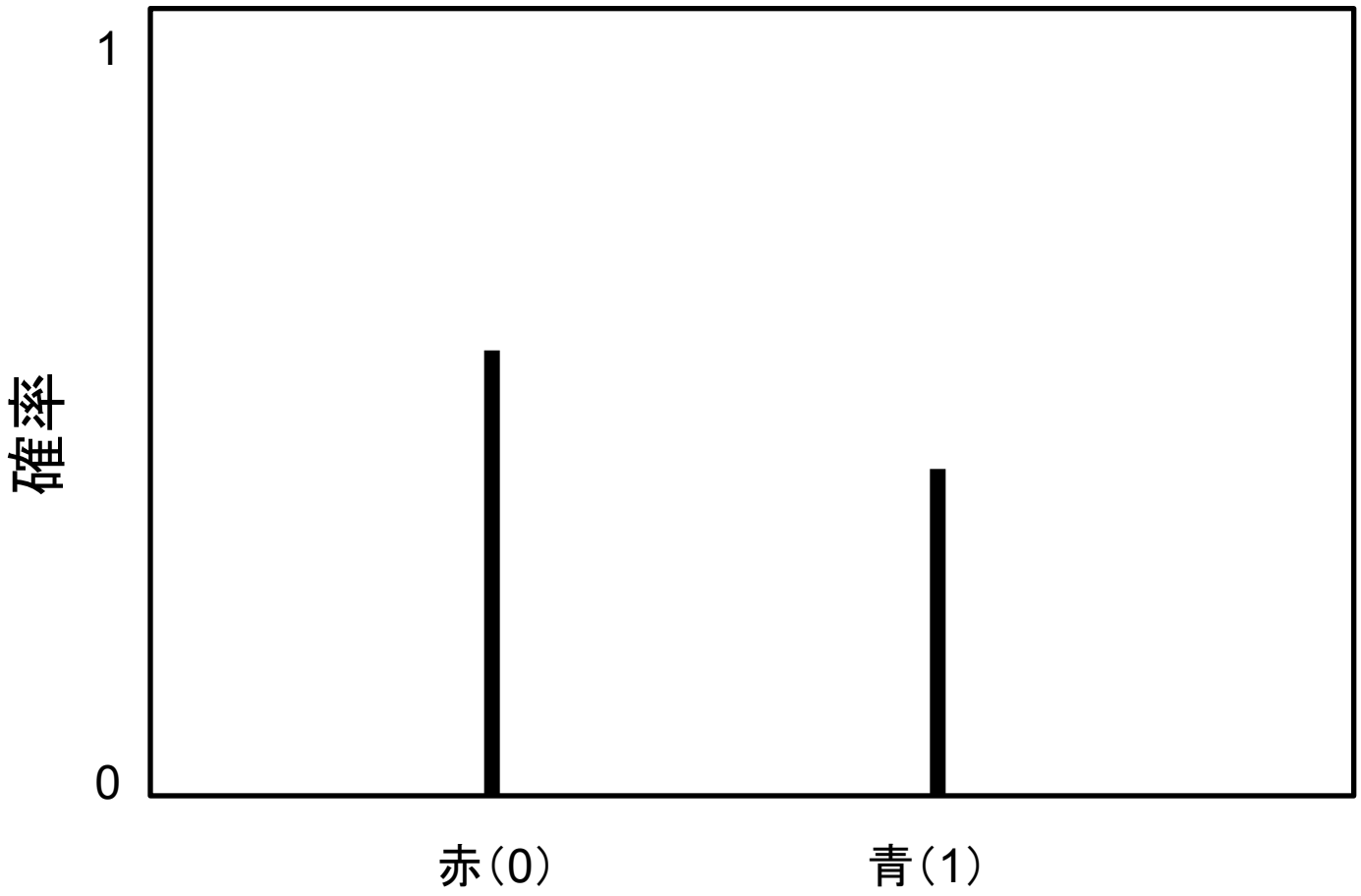
累積分布関数



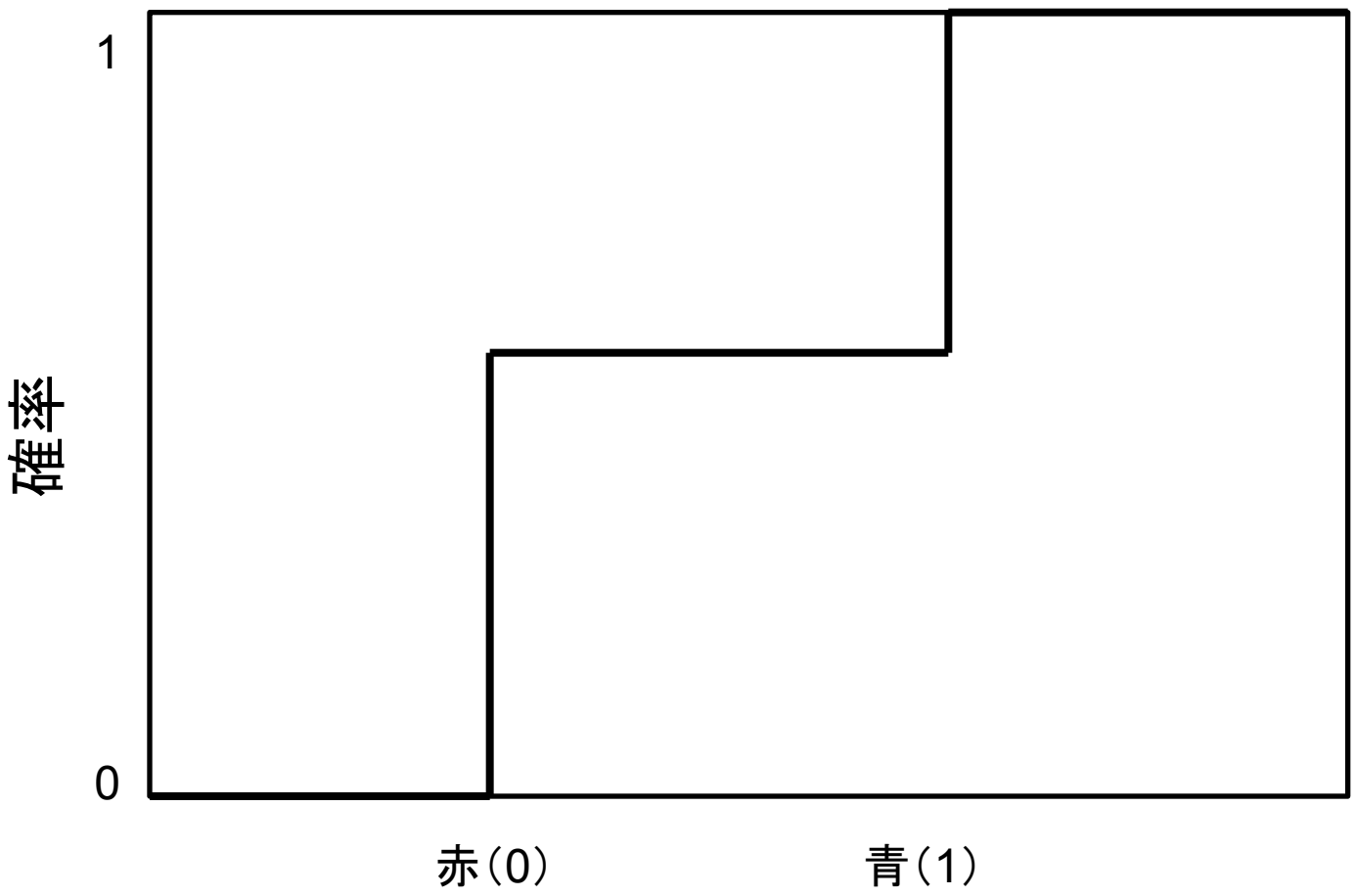
確率密度関数



確率量関数



累積分布関数



離散の変数の場合も、累積分布関数は共通

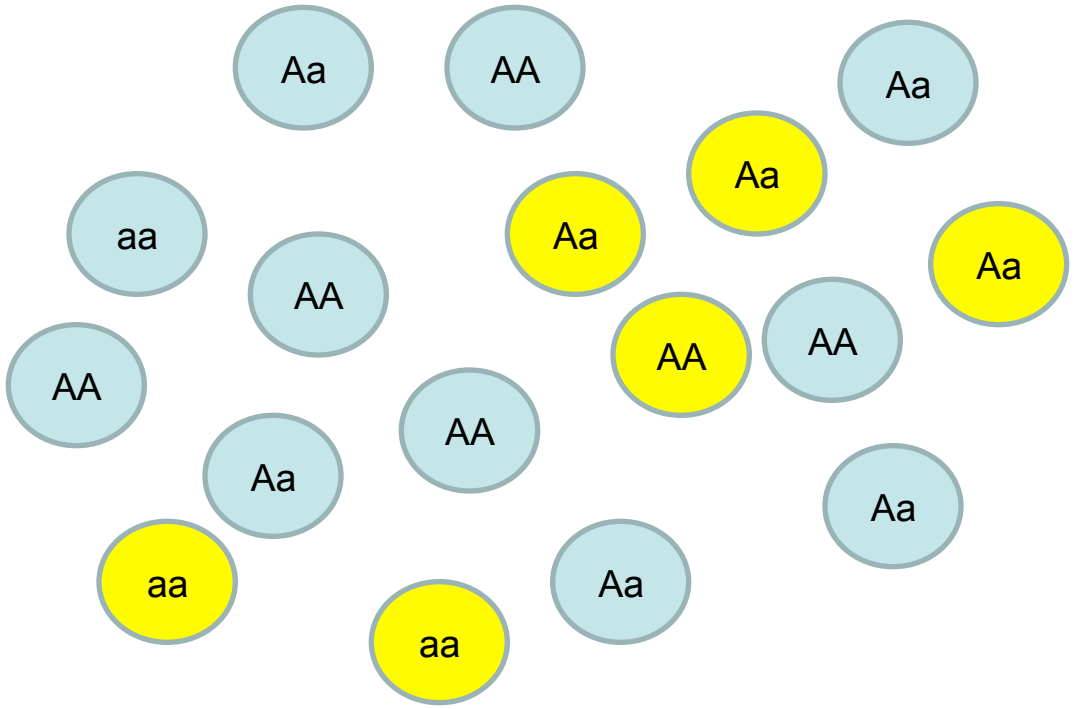
メンデルの法則に出てくる、「座位」「アレル」
「遺伝型」「表現型」も同じこと

これらは「モノ」ではなく「情報」

1. 集合か要素か
2. 任意の要素か特定の要素か
3. 変数か値か
4. 質的情報か量的情報か（表現型）

日本人にメンデルの法則がわかりにくい理由は、概念が「モノ」ではなく「情報」だから

メンデルの法則に出てくる、「座位」「アレル」「遺伝型」「表現型」も同じこと



非病気

病気

AA遺伝型の集合

任意のAAの人

特定のAAの人

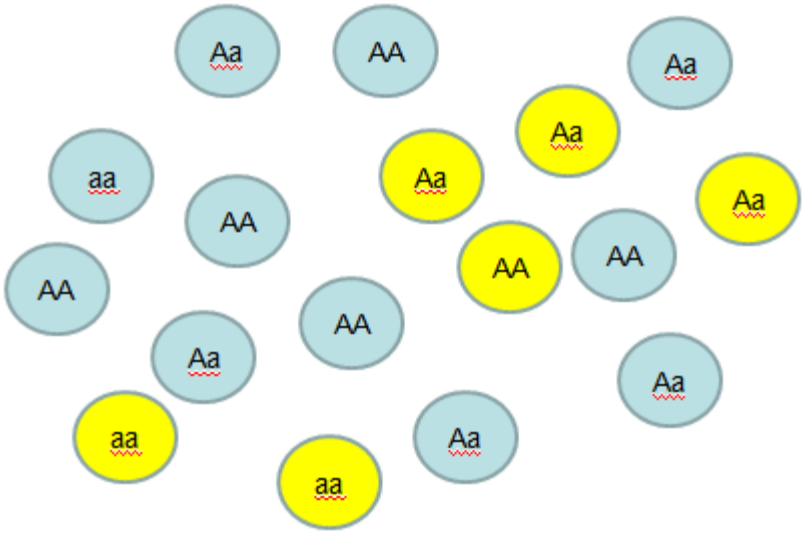
病気の人の集合

任意の病気の人

特定の病気の人

遺伝学用語は、「集合」「要素」「任意の」「特定の」「量か質」の区別をする必要がある

1. P(Aa): 任意の個体がAaの遺伝型の確率
2. P(D): 任意の個体が病気である確率（罹患率）
3. P(D|Aa): 任意のAaの遺伝型の個体が病気である確率（浸透率）
4. P(Φ ≦ 173): 任意の個体の身長が173以下である確率
5. P(Φ ≦ 173 |Aa): 任意のAaの個体の身長が173以下である確率

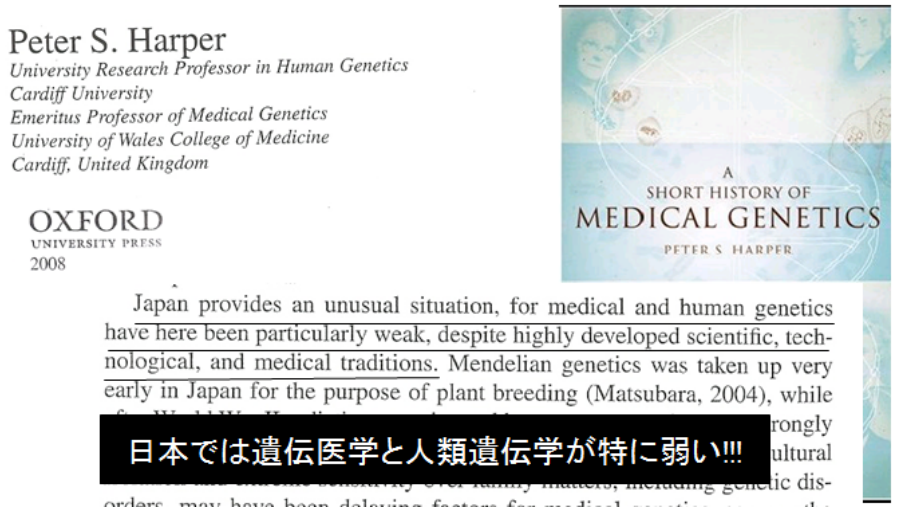


日本は遺伝学と統計学が弱い（どこに問題があるのだろう？）

1. 集合と要素の区別に弱い（集合名詞、複数と単数が無い）
2. 任意の要素と特定の要素の違いに弱い（定冠詞と不定冠詞が無い）
3. 個人の性質を表わす変数と値の違いに弱い（定冠詞と不定冠詞が無い）
4. 質的対象物と量的対象物の違いに弱い（可算名詞と非可算名詞が無い）

弱点を理解し、新しい概念に慣れれば克服できる

情報を直感的、情緒的に理解せず、具体的、視覚的に理解



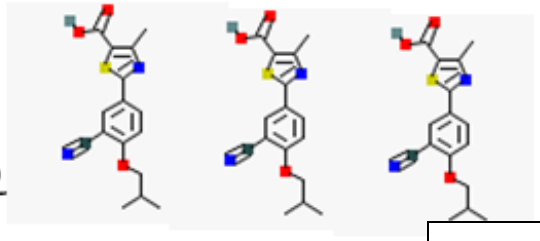
確実性と均一性

VS

不確実性と多様性

「不確実性と多様性」は人間の特徴

「確実な過程」を繰り返すと、「均一」なものが出来る

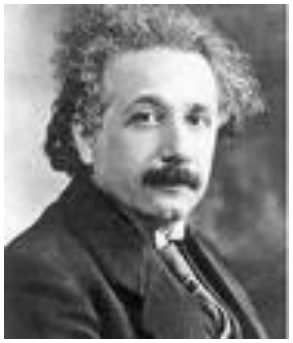


「確実性と均一性」が
製造業の本質

「不確実な過程」を繰り返す
と「多様性」が発生する

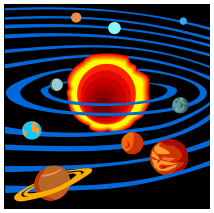
「不確実性と多様性」が
サービス産業の本質

モノと生き物の本質的な違い



不確実性は生物に必然

「神はサイコロを振りたまわず」



根本原理が確実な過程

「生物の世界では絶えずサイコロが振られている」

根本原理が不確実な過程

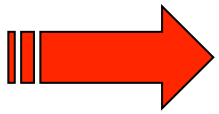
ゲノムの多様性がすべての多様性の起源

確率を本格的に定義しよう

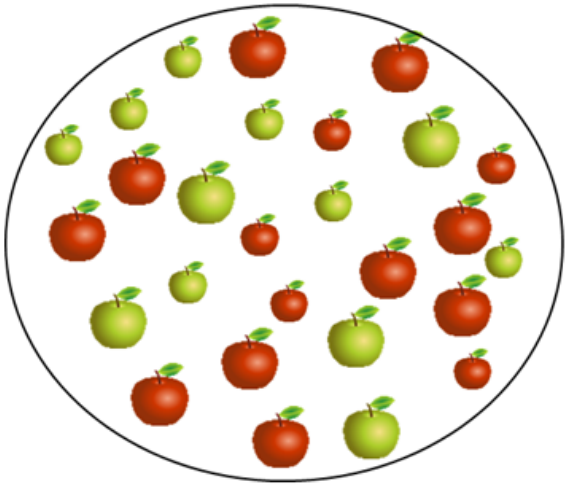
コルモゴロフによる公理的確率論

まず1つの試行（実験）を定義する

試行



結果

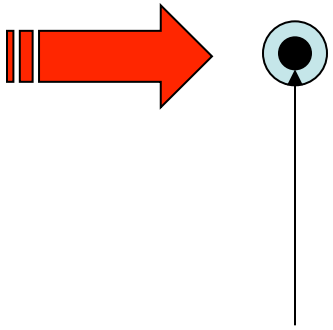


例えば「一つのリンゴを取る」

確実な過程と不確実な過程

確実な過程

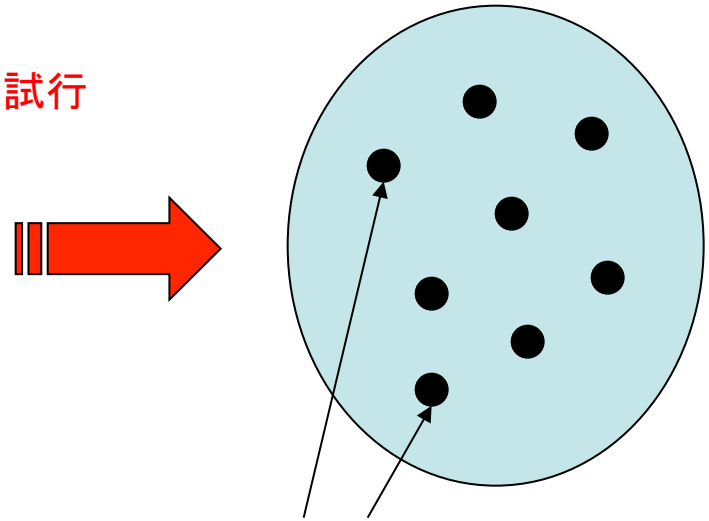
試行



起こりうる結果は1つだけ

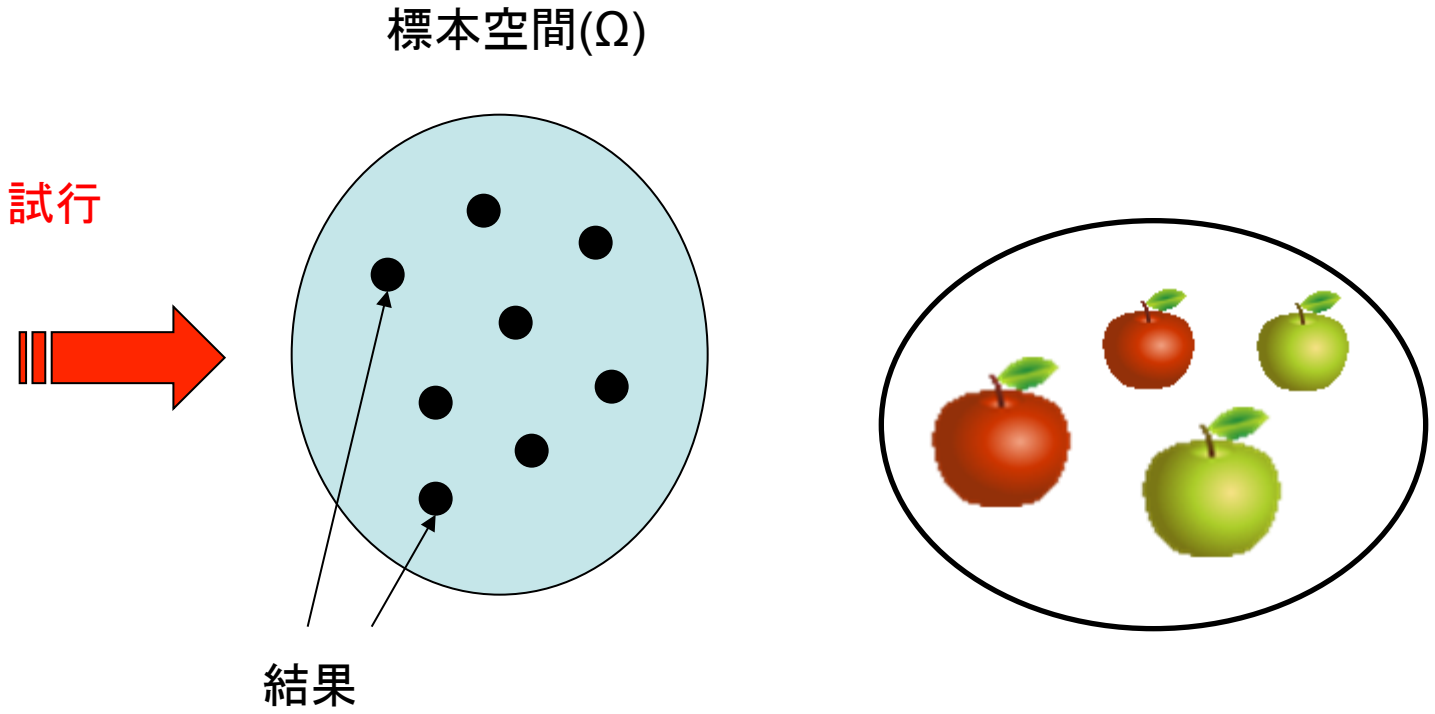
不確実な過程

試行



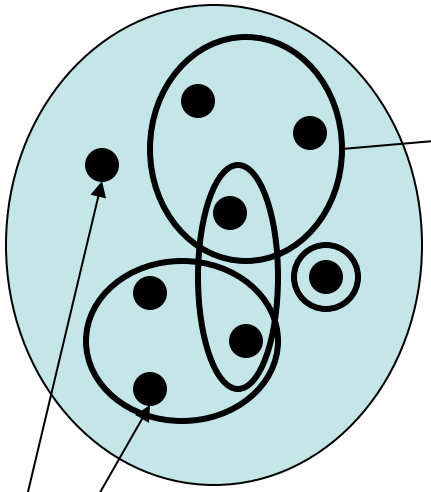
起こりうる結果は2つ以上
(無限かもしれない).

標本空間(Ω)は起こりうる結果の集合である



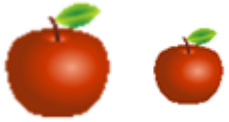
標本空間は実験によって起こりうる
すべての結果の集合である

Ω（標本空間）



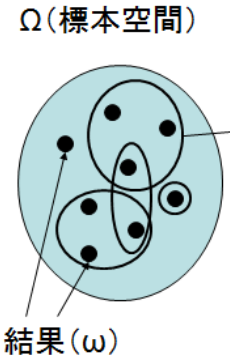
結果（ω）

1つの出来事は、標本空間で定義された結果の集合、つまり標本空間の部分集合である。



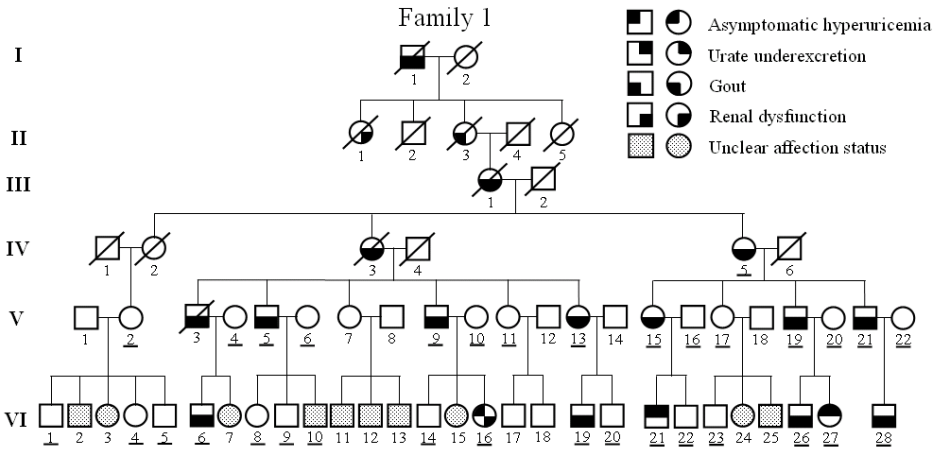
赤い小さなリンゴは結果だが、
赤いリンゴは出来事

確率は「出来事（集合）の関数」である。



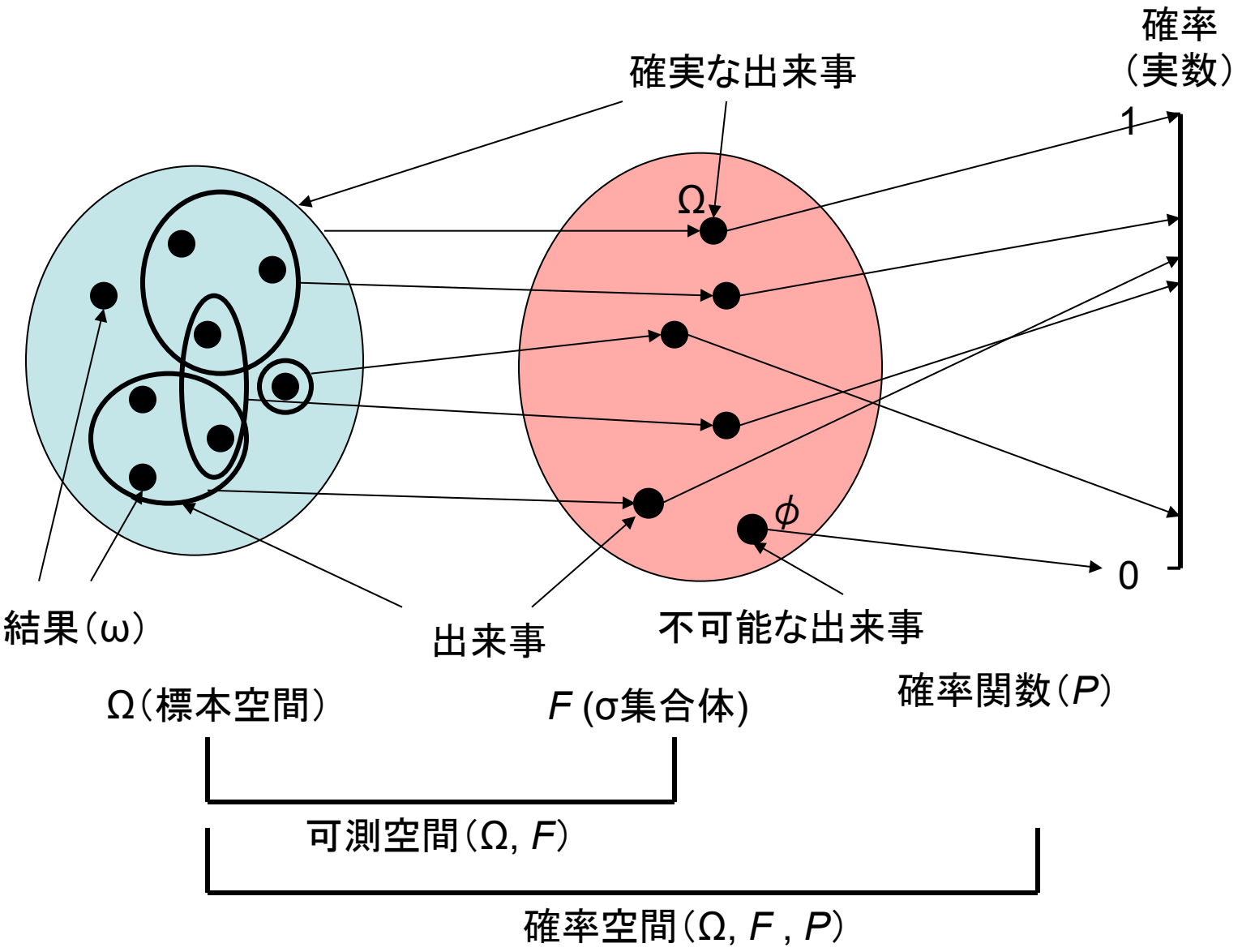
家系のすべてのアレルの動きや表現型の確定した情報を含むものが「結果」で、特定の個人が病気という情報は「出来事」

Familial Juvenile Hyperuricemic Nephropathy (FJHN)の家系



結果の数は膨大（宇宙の星より多い）

確率空間の構造


確率
（実数）
確実な出来事
1
Ω
ϕ
0
結果（ω）
出来事
不可能な出来事
Ω（標本空間）
F (σ集合体)
確率関数（P）
可測空間（Ω, F）
確率空間（Ω, F , P）

膨大なデータ解析では「エラー」は必然

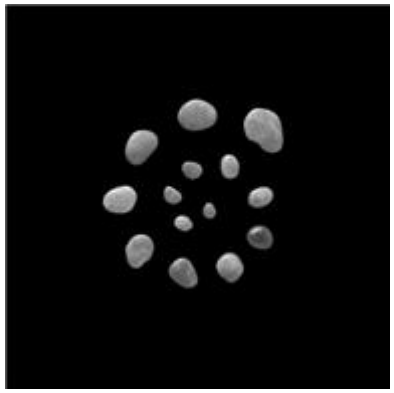
エラーの確率を把握する事が重要

精度が99.999%でも、全ゲノムでは30,000個のエラー

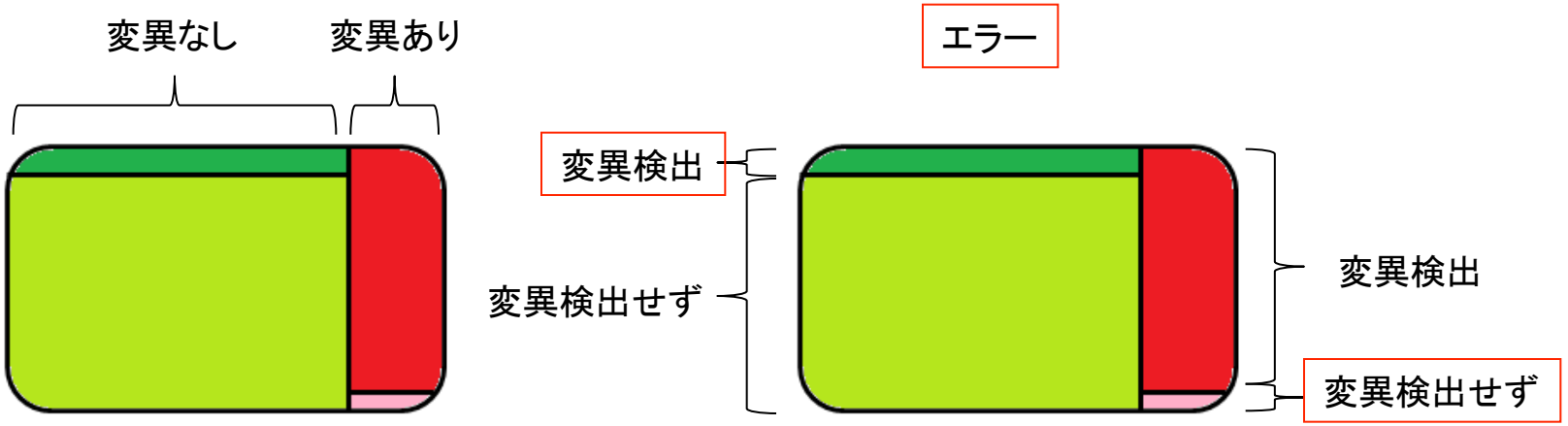
SNPがエラーの可能性より、新規変異がエラーの可能性の方が高い

1. 月で石器に似た石が見つかった場合と、地球の無人島で同じような石器に似た石が見つかった場合とで、どちらが本物である可能性が高い？

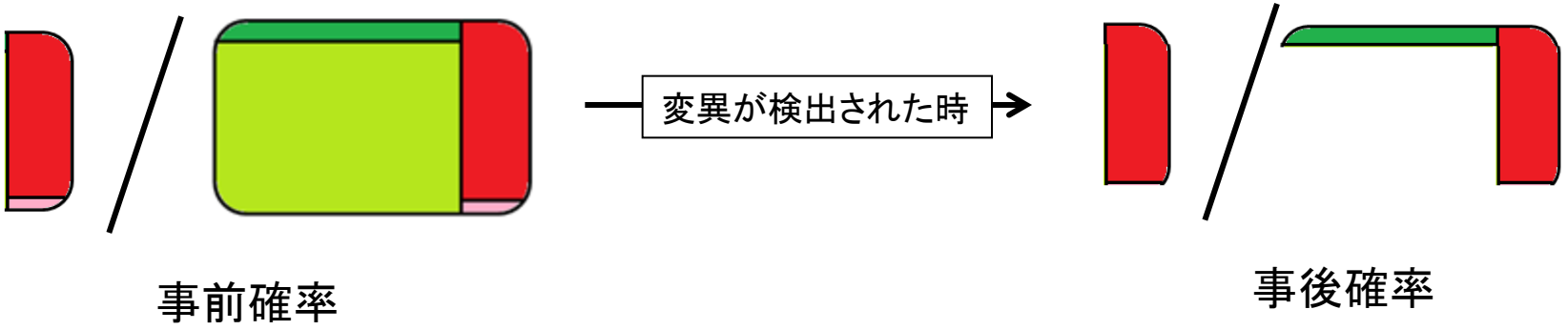
2. 本物の石器が存在する「事前確率」が違うと、石器に似たものが見つかった時、それが本物である「事後確率」が違う



事前確率と事後確率を正しく理解する



本物の変異である確率は？

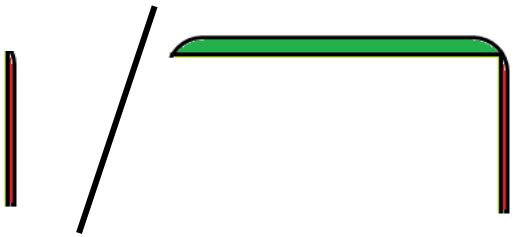


事前確率

事後確率

変異の事前確率が極めて低い時は

エラーの確率が極めて低くても



検出された変異が本物である確率は下がる

（FDRはかなり高い）

シークエンスエラーが極めて低くても、新たな変異のエラー率（偽発見率：FDR）は低くない

0.012%　99.999%

$$Pr(R|S) = \frac{Pr(R)Pr(S|R)}{Pr(R)Pr(S|R) + Pr(N)Pr(S|N)} = 0.923 = 1 - 0.077$$

0.012%　99.999%　99.988%　0.001%

False discovery rate of a rare variant 7.7%

R: rare variant

N: no rare variant

S: positive signal of rare variant

もともと、妊婦の1,000人に1人がダウン症を妊娠していることがわかっている。

妊婦の血液を用いた新しい遺伝子診断で胎児のダウン症の診断が可能である。

ダウン症を妊娠している妊婦の99.9%（感度）はこの検査陽性であり、ダウン症を妊娠していない妊婦の99.9%（特異度）がこの検査で陰性である。

ある妊婦がこの検査で陽性の時、胎児がダウン症である確率はどれくらいか？

$$\frac{\overset{\text{事前確率}}{0.001} \times \overset{\text{感度}}{0.999}}{0.001 \times 0.999 + 0.999 \times 0.001} = 0.5 \quad \text{陽性的中率}$$

事前確率　感度　1 - 事前確率　1 – 特異度

事前確率が1%なら陽性的中率は91%

事前確率が5%なら陽性的中率は98%

検査陽性による疾患/非疾患のオッズ（disease odds)の増加

事前	検査陰性	検査陽性	合計
非疾患	$(1-q_N)(1-\pi)$	$q_N(1-\pi)$	$1-\pi$
疾患	$(1-q_D)\pi$	$q_D\pi$	π
合計	$(1-q_N)(1-\pi)+(1-q_D)\pi$	$q_N(1-\pi)+q_D\pi$	1

全集団における割合、または確率: π: 有病率, q_D: 感度, $(1-q_N)$: 特異度

$$\frac{\pi_+}{1-\pi_+} = \frac{\pi}{1-\pi}\cdot\frac{q_D}{q_N}$$

事後オッズ = 事前オッズ x 尤度比

検査陽性により、事後のオッズ（ $\frac{\pi_+}{1-\pi_+}$ ）は、事前のオッズ（ $\frac{\pi}{1-\pi}$ ）の尤度比（$\frac{q_D}{q_N}$）倍になる

ご静聴、ありがとうございました。