Wright-Fisher Model （short ver.）

岡田崇

1 数学的準備：二項分布の一般論

この講義で使う確率の道具は，ほぼ二項分布だけである．まず，その意味と平均・分散を確認しておく．

確率 p で成功する試行を N 回独立に行い，成功回数を K とする．このとき

$$K \sim \mathrm{Binomial}(N, p)$$

と書く．この記号は，K が次の確率分布に従うことを表す：

$$\mathbb{P}(K = k) = \binom{N}{k} p^k (1-p)^{N-k}, \qquad k = 0, 1, \ldots, N. \tag{1}$$

ここで

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

は，N 回の試行のうち成功する k 回を選ぶ組合せの数である．式 (1) の三つの因子は，それぞれ

$$\underbrace{\binom{N}{k}}_{\text{成功する試行の選び方}} \quad \underbrace{p^k}_{k\text{ 回の成功}} \quad \underbrace{(1-p)^{N-k}}_{N-k\text{ 回の失敗}}$$

に対応する．

二項分布の平均と分散は

$$\mathbb{E}[K] = Np, \qquad \mathrm{Var}(K) = Np(1-p) \tag{2}$$

である．[1)]

成功割合を

$$X = \frac{K}{N}$$

と書くと，

$$\mathbb{E}[X] = \mathbb{E}\left[\frac{K}{N}\right] = \frac{1}{N}\mathbb{E}[K] = p,$$
$$\mathrm{Var}(X) = \mathrm{Var}\left(\frac{K}{N}\right) = \frac{1}{N^2}\mathrm{Var}(K) = \frac{p(1-p)}{N}.$$

---

1) 簡単に理由を確認する．各試行の成功を表す変数を Y_j と書くと，Y_j は成功なら 1，失敗なら 0 をとる．このとき

$$K = Y_1 + Y_2 + \cdots + Y_N$$

であり，

$$\mathbb{E}[Y_j] = p, \qquad \mathrm{Var}(Y_j) = p(1-p)$$

である．独立な変数の和なので，平均と分散を足し合わせれば式 (2) が得られる．

二項サンプリングによる割合のばらつき

$K \sim \mathrm{Binomial}(N, p)$, $X = K/N$ のとき,

$$\mathbb{E}[X] = p, \qquad \mathrm{Var}(X) = \frac{p(1-p)}{N}$$

である.

正規近似 （Gaussian approximation）

N が十分大きく，p が 0 や 1 に近すぎないとき，二項分布は正規分布で近似できる.

$$X \simeq p + \sqrt{\frac{p(1-p)}{N}}\,\xi, \qquad \xi \sim \mathrm{Normal}(0,1)$$

と書ける．$\mathrm{Normal}(0,1)$ は、平均 0, 分散 1 の正規分布 $p(Z=\xi) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{\xi^2}{2})$.

2 導入：頻度を記述する

本ノートでは，一倍体の無性集団を考える．進化を考えるとき，個々の変異や変異株が「ある」か「ない」かだけでなく，**集団中でどれくらいの割合を占めるか**が最も基本的な量である．この割合を**頻度**と呼ぶ．

たとえば，$x(t)$ をある変異株の頻度とすれば，

$$x(t) = 0 \quad \text{はその変異株がいないこと，}$$

$$x(t) = 1 \quad \text{は全員がその変異株に置き換わったこと}$$

を意味する．

3 Wright–Fisher model: 次世代はサンプリングにより作られる

以下では，集団遺伝学の標準モデルである Wright–Fisher model (WF model) を説明し、上記の確率微分方程式を導く。詳細に入る前に、WF model に基づく頻度ダイナミクスの例を見せる。

頻度軌道の例

図 1　中立な遺伝的浮動と，selection がある場合の頻度ダイナミクスの例．同じ初期頻度から出発しても，1 本 1 本の軌道は大きく異なる．

3.1 中立な場合の平均と分散

集団サイズを一定の N とし，集団には 2 種類の型 A と B がいるとする．Wright–Fisher model では，次世代の N 個体は，親世代からランダムにサンプリングされると考える．

時刻 t における A の頻度を x_t とすれば、中立な場合 (A と B でサンプルされやすさが同じ場合)、次世代の各個体が A をサンプルする確率　p_t は、

$$p_t = x_t$$

で与えられる。

二項分布の一般論における p を x_t に置き換えて，中立な場合の頻度の統計性として以下を得る。

中立 Wright–Fisher model

$$\mathbb{E}[x_{t+1} \mid x_t] = x_t,$$
$$\mathrm{Var}(x_{t+1} \mid x_t) = \frac{x_t(1-x_t)}{N}.$$

中立な変異の場合、頻度は平均的には変化しないが，サンプリングによって毎世代ばらつく．このランダムな頻度変化を**遺伝的浮動 (genetic drift)** という．[2)]

3.2 Gauss 近似 (拡散近似)

中立過程の拡散近似

$$x_{t+1} = x_t + \sqrt{\frac{x_t(1-x_t)}{N}}\,\xi, \qquad \xi \sim \mathrm{Normal}(0,1) \tag{3}$$

と書ける．等価な表現として、

$$\mathrm{d}x = \sqrt{\frac{x(1-x)}{N}}\,\mathrm{d}W$$

とも書く (W を Wiener process と呼ぶ)。

2) わかりやすい例として、双子の兄弟が次世代に残せる子供の数は同じとは限らない．

4 選択がある場合の Wright-Fisher Model: 次世代に選ばれる確率が変わる

次に，A が B よりも増えやすい場合を考える．ここでは，A の相対的な増えやすさを $1+s$，B の相対的な増えやすさを 1 とする．$s>0$ なら A が有利である．s を、選択係数 (selection coefficient) や相対適応度 (relative fitness) と呼ぶ。応用上の重要性から、以下では $|s| \ll 1$ を仮定する。

4.1 選択がある場合の平均と分散

ある世代 t において A の頻度が x_t であるとする。選択の効果を考慮した A の重み付き頻度は $(1+s)x_t$，B の重み付き頻度は $1-x_t$ であるから，次世代の各個体が A をサンプルする確率 p_t は、

$$
\begin{aligned}
p_t &= \frac{(1+s)x_t}{(1+s)x_t + (1-x_t)} \\
&= \frac{(1+s)x_t}{1+sx_t} \\
&\approx x_t + sx_t(1-x_t). \quad (\text{when } s \ll 1)
\end{aligned} \tag{4}
$$

二項分布の一般論における p を $x_t + sx_t(1-x_t)$ に置き換えば、選択圧のある場合における頻度の統計性として以下を得る。

選択のある Wright–Fisher model

$$
\mathbb{E}[x_{t+1} \mid x_t] = x_t + sx_t(1-x_t), \tag{5}
$$

$$
\mathrm{Var}(x_{t+1} \mid x_t) = \frac{x_t(1-x_t)}{N}. \tag{6}
$$

3)

4.2 Gauss 近似 (拡散近似): 選択と遺伝的浮動を合わせる

選択がある場合における頻度変化の時間発展

$$
x_{t+1} = x_t + sx_t(1-x_t) + \sqrt{\frac{x_t(1-x_t)}{N}}\xi_t, \quad \xi_t \sim \mathrm{Normal}(0,1)
$$

第 1,2 項は バイアス (selection)、第 3 項はランダムな揺らぎ (genetic drift) を表す．等価な表

3) 厳密には分散には s に依存する項もあるが、$|s| \ll 1$ のときには無視できる。

現として、

$$\mathrm{d}x = sx(1-x)\,\mathrm{d}t + \sqrt{\frac{x(1-x)}{N}}\,\mathrm{d}W_t$$

とも書く。[a]

a　物理学の文脈では、Langevin 方程式 $\frac{dx}{dt} = sx(1-x) + \sqrt{\frac{x(1-x)}{N}}\xi(t)$, $\langle \xi(t)\xi(t')\rangle = \delta(t-t')$ と書くことが多い。これらは単なる表現方法の違いであり、シミュレーション実装の際は正規分布を用いた式を用いる。

4.3 決定論的なダイナミクス（遺伝的浮動は無視する）

ゆらぎを無視した決定論的近似では,

$$\boxed{\frac{\mathrm{d}x}{\mathrm{d}t} = sx(1-x)} \tag{7}$$

となる．これはロジスティック方程式と呼ばれる。

式 (7) は、変数変換 $u(t) = \log\frac{x(t)}{1-x(t)}$ をすれば容易に解ける[4]．$\frac{du}{dx} = \frac{1}{x(1-x)}$ を用いて、

$$\frac{du}{dt} = s$$
$$u(t) = st + C.$$

したがって、初期頻度を $x(t=0) = x_0$ とすれば,

$$\log\frac{x(t)}{1-x(t)} = st + \log\frac{x_0}{1-x_0}. \tag{8}$$

となる。実用上は Logit 関数による表式が便利なことが多いが、$x(t)$ について解けば、$x(t) = \frac{x_0 e^{st}}{x_0 e^{st} + (1-x_0)}$ となる。

4)　$f(x) = \log\frac{x}{1-x}$ は Logit 関数と呼ばれる．

5 固定確率

ある変異が最終的に集団全体を占めることを**固定**という．初期頻度 x_0 から出発した変異が最終的に固定する確率を

$$P_{\text{fix}}(x_0)$$

と書く．

5.1 中立変異の固定確率

十分長い時間の後には，変異は消滅するか固定することに注意する．したがって、固定確率を $P_{\text{fix}}(x_0)$ と書けば、最終的な事象は

$$x_\infty = \begin{cases} 1 & \text{固定した場合，確率 } P_{\text{fix}}(x_0)\ , \\ 0 & \text{消滅した場合，確率 } 1 - P_{\text{fix}}(x_0) \end{cases}$$

である．したがって、最終的な頻度 x_∞ の期待値は、

$$\mathbb{E}[x_\infty] = 1 \cdot P_{\text{fix}}(x_0) + 0 \cdot \{1 - P_{\text{fix}}(x_0)\} = P_{\text{fix}}(x_0).$$

一方で、中立な場合には、，$\mathbb{E}[x_{t+1} \mid x_t] = x_t$ である。したがって、平均頻度は初期頻度 x_0 のままである:

$$\mathbb{E}[x_\infty] = x_0.$$

したがって，上の 2 つの式を比較すれば、

$$\boxed{P_{\text{fix}}(x_0) = x_0} \tag{9}$$

を得る．

たとえば、新しく 1 コピーだけ現れた中立変異なら $x_0 = 1/N$ なので，

$$P_{\text{fix}}\left(\frac{1}{N}\right) = \frac{1}{N}$$

である．

分子時計

新たな中立変異が 1 世代あたり各個体に確率 μ で生じるとする．集団サイズが N なら，新たな

中立変異は集団内に毎世代、平均

$$\mu N$$

個現れる。

各中立変異の固定確率は

$$\frac{1}{N}$$

であるから，中立変異が固定する速度は

$$\mu N \times \frac{1}{N} = \mu$$

となる．

これは，中立変異の固定速度が集団サイズではなく突然変異率によって決まることを示している．つまり，中立置換はほぼ一定の速度で蓄積すると考えられる．したがって，2 つの系統で中立置換率が同じであるという理想的な状況では，配列の違いの大きさから，それらの系統が分岐してからの時間を推定できる．この考え方が分子時計の基本である．

5.2 有利変異の固定確率

選択係数 $s > 0$ の変異では，中立の場合より固定確率が高くなる．Wright–Fisher 拡散近似では，固定確率は

有益変異の固定確率

$s \ll 1, Ns \gg 1$ とする。頻度 x_0、選択係数 $s > 0$ の変異が最終的に固定する確率は，

$$P_{\text{fix}}(x_0) = \frac{1 - e^{-2Nsx_0}}{1 - e^{-2Ns}}$$

である (導出は Appendix A).

Haldane の公式と Drift barrier

重要な場合として、選択係数 $s > 0$ の変異が集団内に現れたときを考える。初期頻度は $x_0 = \frac{1}{N}$ であるから、固定する確率は、

$$P_{\text{fix}}\left(\frac{1}{N}\right) \simeq 2s \qquad (s \ll 1, Ns \gg 1)$$

である (Haldane の公式). たとえば、5%有利な変異をもつ個体が現れたとしても、その個体が集団に広がる確率は 10%ほどであり、大部分の有益変異については固定せずに消滅することを意味する。

1 コピーが固定する確率が $2s$ という結果から、興味深いことがわかる。有益変異が M コピー

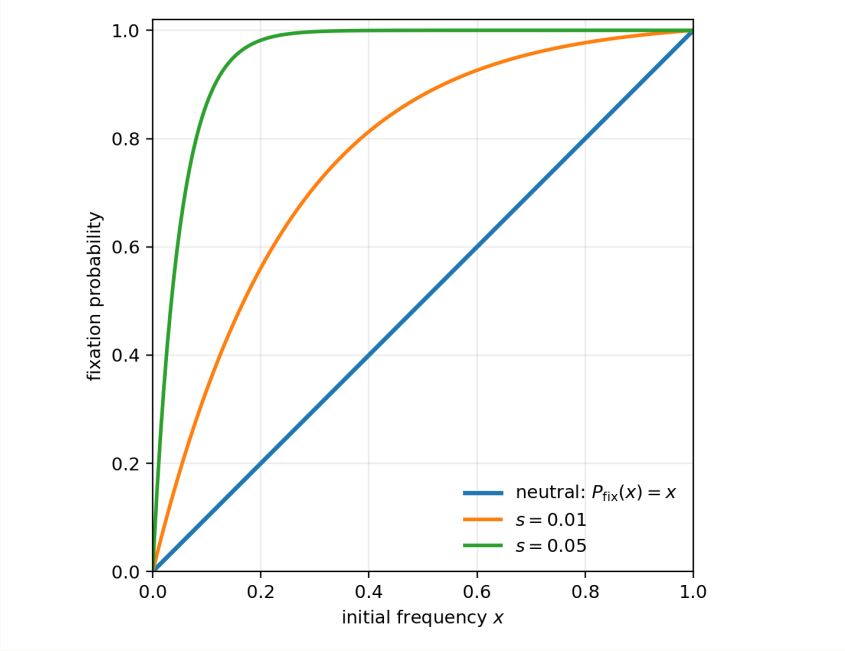

図 2　固定確率。中立では $P_{\mathrm{fix}}(x_0) = x_0$ である．有利な変異では この直線より上にくる．

あるとき，それぞれのコピーが独立に固定へ向かうと粗く近似すれば，固定確率は M 倍で、

$$P_{\mathrm{fix}} \sim 2Ms$$

である．したがって，

$$M \sim \frac{1}{2s}$$

程度のコピー数に達すると，固定確率は $O(1)$ となり，その変異は偶然の消滅をかなり免れやすくなる．

頻度で書けば，

$$x_{\mathrm{est}} \sim \frac{1}{2Ns}$$

である．この頻度より十分低い領域では，遺伝的浮動の影響が大きいために、たとえ有益変異であってもほぼ中立的に振る舞い、偶然に消滅しやすい．一方で，この頻度を十分に超えると，選択による決定論的な増加が強く効く．この境界の頻度 x_{est} は、 **establishment threshold** や **drift barrier** と呼ばれる。

A 固定確率の導出

アレル頻度の確率過程を X_t と書き，$X_0 = x$ から出発するとする．0 または 1 に初めて到達する時刻を

$$\tau = \inf\{t \geq 0 : X_t \in \{0, 1\}\}$$

とし，固定確率を

$$u(x) = \mathbb{P}_x(X_\tau = 1) = \mathbb{E}_x\left[\mathbf{1}_{\{X_\tau = 1\}}\right]$$

と定義する．下付きの x は，初期値が $X_0 = x$ であることを表す．

なぜ $u(x)$ に対する backward equation を解けばよいのかを説明する．短い時間 Δt 後の状態 $X_{\Delta t}$ で条件づけると，反復期待値の法則より

$$\begin{aligned} u(x) &= \mathbb{E}_x\left[\mathbf{1}_{\{X_\tau = 1\}}\right] && (10)\\ &= \mathbb{E}_x\left[\mathbb{E}_x\left[\mathbf{1}_{\{X_\tau = 1\}} \mid X_{\Delta t}\right]\right]. && (11) \end{aligned}$$

Wright–Fisher 拡散は Markov 過程なので，$X_{\Delta t} = y$ が分かれば，それ以前の履歴は将来の固定確率に影響しない．したがって内側の条件付き期待値は，状態 y から出発した固定確率 $u(y)$ に等しい．式 (11) は

$$u(x) = \mathbb{E}_x[u(X_{\Delta t})] \tag{12}$$

となる．これは「最初の短い時間だけ進め，その後の固定確率を平均する」という first-step decomposition である．確率密度の時間発展を求める forward equation とは異なり，ここでは**初期状態 x の関数** $u(x)$ を求めるため，backward equation が現れる．

$X_{\Delta t} = x + \Delta X$ と書く．拡散近似では

$$\begin{aligned} \mathbb{E}[\Delta X \mid X_0 = x] &= sx(1-x)\Delta t + o(\Delta t),\\ \mathbb{E}[(\Delta X)^2 \mid X_0 = x] &= \frac{x(1-x)}{N}\Delta t + o(\Delta t). \end{aligned}$$

また，Taylor 展開により

$$u(x + \Delta X) = u(x) + u'(x)\Delta X + \frac{1}{2}u''(x)(\Delta X)^2 + o(\Delta t)$$

である．これを式 (12) に代入して条件付き期待値をとると，

$$u(x) = u(x) + \Delta t\left\{sx(1-x)u'(x) + \frac{x(1-x)}{2N}u''(x)\right\} + o(\Delta t).$$

両辺から $u(x)$ を引き，Δt で割って $\Delta t \to 0$ とすると，

$$0 = sx(1-x)u'(x) + \frac{x(1-x)}{2N}u''(x). \tag{13}$$

これが固定確率に対する backward equation である．吸収境界から境界条件は

$$u(0) = 0, \qquad u(1) = 1$$

となる．

$0 < x < 1$ では $x(1-x)$ で割ることができるので，

$$0 = su'(x) + \frac{1}{2N}u''(x)$$

すなわち

$$u''(x) + 2Nsu'(x) = 0$$

である．$v(x) = u'(x)$ とおくと，

$$v'(x) + 2Nsv(x) = 0$$

であり，

$$v(x) = Ce^{-2Nsx}$$

となる．したがって，

$$u(x) = A + B\{1 - e^{-2Nsx}\}$$

と書ける．

境界条件 $u(0) = 0$ より $A = 0$．また $u(1) = 1$ より $B = \frac{1}{1-e^{-2Ns}}$ である．よって，

$$u(x) = \frac{1 - e^{-2Nsx}}{1 - e^{-2Ns}}$$

を得る．

$s \to 0$ の極限では，ロピタルの定理または $e^{-y} \simeq 1 - y$ を使って

$$u(x) \to x$$

となる．これは中立変異の固定確率と一致する．また、$x = \frac{1}{N}$ を代入すると、

$$u(\frac{1}{N}) = \frac{1 - e^{-2s}}{1 - e^{-2Ns}} \approx 2s$$

となる (Haldane's formula).