Шаг нулевой. Постановка задачи.
Рассмотрим набор $\boldsymbol{A}$, представляющий множество различных генов для хромосомы в заданном конкретном локусе. Во время генерации особи ген из набора выбирается случайно и с равной вероятностью. Поставим вопрос: какой должна быть численность популяции, чтобы поддерживалось разнообразие генов в популяции? Полагаем, что мутации отсутствуют.

Для простоты положим, что хромосома состоит из одного гена. Набор доступных генов, из которых может состоять хромосома обозначим за $\boldsymbol{A} = (\boldsymbol{a_1}, \boldsymbol{a_2}, .., \boldsymbol{a_M})$. Саму популяцию будет представлять в виде вектора $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, .., \boldsymbol{x_N})$**.** Особь будем представлять в виде последовательности генов. Например $\boldsymbol{x_1} = \boldsymbol{a_1}$

Шаг первый. Определения.
Введем определение полной популяции.
Полная популяций – это такая популяций, особей которых представляют ВСЕ гены из набора доступных. Например, пусть $\boldsymbol{A} = (\boldsymbol{a_1}, \boldsymbol{a_2}, .., \boldsymbol{a_M})$ и $\boldsymbol{X} = (\boldsymbol{a_1}, \boldsymbol{a_2}, .., \boldsymbol{a_M})$**,** тогда $\boldsymbol{X}$ – полная популяций.
Неполная популяций – это такая популяция, которая не является полной. Например пусть $\boldsymbol{A} = (\boldsymbol{a_1}, \boldsymbol{a_2}, .., \boldsymbol{a_M})$ **и** $\boldsymbol{X} = (\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_1} .., \boldsymbol{a_2})$**,** тогда **X –** неполная популяций.
Замечу, что N заведомо больше или равен M. В противном случае из популяции невозможно создать полную.

Шаг первый. Формулировка идеи.
Сколько различных популяций из N особей можно составить при наборе из M генов? Очевидно, что $\boldsymbol{M^N}$**.** Далеко не все эти популяции будут содержать такой набор хромосом, который обеспечивает полноту популяции. Например популяция $\boldsymbol{X} = (\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_1} .., \boldsymbol{a_2})$. Вероятность сгенерировать такую популяцию, конечно, мала, но существует похожие по типу популяции, и вместе они дают немалую вероятность генерации неполной популяции. Как выше было указано, мутации в данной модели исключены. В таком случае мы никогда не получим популяцию, в которой бы была особь например с геном $\boldsymbol{a_3}$.

Обозначим за **Y** множество всех неполных популяций. Тогда вероятность сгенерировать полную популяцию есть $\boldsymbol{P} = \mathbf{1} - \frac{\boldsymbol{CardY}}{\boldsymbol{M^N}}$. Очевидно, что $\boldsymbol{CardY}$ зависит от размера популяции N. Поэтому, определяя желаемую вероятность получить полную популяцию P, мы сможем регулировать параметр N.

Шаг второй. Реализация идеи.
Итак, пусть имеется набор $\boldsymbol{A} = (\boldsymbol{a_1}, \boldsymbol{a_2}, .., \boldsymbol{a_M})$ и популяция $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, .., \boldsymbol{x_N})$. Число различных популяций, которые мы можем составить равняется $\boldsymbol{M^N}$. Однако это количество включается в себя такие популяции, которые не являются полными. Одна из таких популяций есть такая популяция, в которой нет особи с геном $\boldsymbol{a_M}$ $\boldsymbol{X'} = (\boldsymbol{a_1}, \boldsymbol{a_2}, .., \boldsymbol{a_{M-1}}, \boldsymbol{a_1}, ..)$**.** Однако заметим, что эта популяция является полной, если $\boldsymbol{A'} = (\boldsymbol{a_1}, \boldsymbol{a_2}, .., \boldsymbol{a_{M-1}})$**.** Другими словами, она является полной при некотором наборе $\boldsymbol{A'}$**,** являющимся подмножеством множества $\boldsymbol{A}$**!**

Исходя из того, что особей в популяции всё также **N** штук, а набор теперь состоит из генов на единицу меньше, то количество возможных популяций, которые мы можем составить на данном наборе генов есть $(\boldsymbol{M-1})^{\boldsymbol{N}}$**.** Заметим, что мы точно также могли бы рассматривать популяцию, в которой отсутствует особь с геном $\boldsymbol{a_1}$**:**

$\boldsymbol{X' = (a_2, a_3, .., a_{M-1}, a_2, ..)}$ и набор $\boldsymbol{A' = (a_2, a_3, .., a_M)}$. Поэтому на самом деле число различных популяций, которые мы можем составить есть $\boldsymbol{C_M^{M-1}(M-1)^N}$. Однако это число включает лишние популяции, о которых я скажу позже.

Итак, теперь $\boldsymbol{X' = (a_1, a_2, .., a_{M-1}, a_1, ..)}$ – полный набор при $\boldsymbol{A' = (a_1, a_2, .., a_{M-1})}$. Число различных популяций, которые мы можем составить равняется $\boldsymbol{(M-1)^N}$. Однако это количество включается в себя такие популяции, которые не являются полными, например $\boldsymbol{X'' = (a_1, a_2, .., a_{M-2}, a_1, ..)}$. Однако популяция $\boldsymbol{X'' = (a_1, a_2, .., a_{M-2}, a_1, ..)}$ является полной при наборе $\boldsymbol{A'' = (a_1, a_2, .., a_{M-2})}$. Число различных популяций, которые могут быть составлены из этого набора $\boldsymbol{C_{M-1}^{M-2}(M-2)^N}$

Дальше следуют повторяющиеся рассуждения до тех пор, пока мы не упремся на набор из одного гена, например $\boldsymbol{A''' = (a_1)}$Из данного набора можно образовать единственную популяцию, которая будет являться полной $\boldsymbol{X''' = (a_1, \dots, a_1,)}$. Таким образом искомая вероятность:

$$P = \frac{M^N - C_M^{M-1}[(M-1)^N - C_{M-1}^{M-2}[(M-2)^N - ...]]}{M^N} =$$

$$1 - \frac{C_M^{M-1}[(M-1)^N - C_{M-1}^{M-2}[(M-2)^N - ...]]}{M^N}$$

Но это не итоговый вид формулы. Каждый раз, когда мы образуем подмножество **A** возникают возможные повторяющие друг друга популяции. Например, рассмотрим набор генов из 4-ёх генов: **{a, b, c, d}.**

1. На первом шаге мы образуем подмножества

{a, b, c}, {a, b, d}, {a, c, d}, {b, c, d}

Видим, что получается по 1! = 1 одинаковому подмножества (тождественному самому себе)

2. На втором шаге мы образуем подмножества

{a, b}, {a, c}, {b, c} {a, b}, {a, d}, {b, d} {a, c}, {a, d}, {c, d} {b, c}, {b, d}, {c, d}

Видим, что получается по 2! = 2 одинаковых подмножества

3. На третьем шаге мы образуем подмножества

{a}, {b}, {a}, {c} {b}, {c}, {a}, {b}, {a}, {d}, {b}, {d}, {a}, {c}, {a}, {d}, {c}, {d}, {b}, {c}, {b}, {d}, {c}, {d}

Видим, что получается по 3! = 6 одинаковых подмножеств.

Если положить первоначально набор из 5-ти генов **{a, b, c, d, e},** то в последней итерации получится 4! = 24 одинаковых подмножеств. Интуиция подсказывает, что на 6-ти одинаковых генах будет уже 5! = 120 одинаковых подмножеств. Этот факт вносит корректировку в формулу:

$$P = 1 - \frac{\frac{C_M^{M-1}}{1!}[(M-1)^N - \frac{C_{M-1}^{M-2}}{2!}[(M-2)^N - ...[\frac{C_{M-(M-1)}^{M-(M-1)}}{(M-1)!}(M-(\mathrm{M}-1))^N]]]}{M^N}$$