ГЕНЕТИКА, 2024, том 60, № 12, с. 94–109

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ

УДК 303.725.23:575.177

О СТАБИЛИЗАЦИИ ТЕМПА ДИВЕРГЕНЦИИ ИЗОНИМИИ

© 2024 г. В. П. Пасеков[1, *]

[1]*Федеральный исследовательский центр "Информатика и управление" Российской академии наук, Москва, 119991 Россия*

**e-mail: pass40@mail.ru*

Поступила в редакцию 19.06.2024 г.
После доработки 22.07.2024 г.
Принята к публикации 24.07.2024 г.

Проведен теоретический анализ фамильного состояния популяции (вектора концентраций однофамильцев в мужском компоненте популяции) и его динамики в результате случайного фамильного дрейфа. Используется аппроксимация такого процесса моделью Райта – Фишера популяции с неперекрывающимися поколениями, неподверженной давлению отбора, т. е. последовательностью вложенных случайных выборок с возвращением из совокупности фамилий отцов. Размер выборки равен $N/2$ согласно численности мужского компонента в популяции размера N. В одной и той же популяции одновременно протекают процессы случайного дрейфа как фамилий, так и генов. Их кардинальное различие в том, что размер выборки фамилий вчетверо меньше, чем выборки аллелей аутосомного локуса. Анализ случайного дрейфа упрощается при переходе от координат-концентраций к квадратным корням из них. При смене поколений состояние получает выборочное отклонение, измеряемое угловым расстоянием, а его средний квадрат дает темп дивергенции, стабилизирующийся в новых координатах. Дана адаптация (применительно к анализу фамильного дрейфа) известного в популяционной генетике результата о характере дивергенции на этапе относительно малого по сравнению с размером популяции количества поколений. Дивергенция фамилий протекает в 4 раза быстрее дивергенции концентраций аллелей.

Ключевые слова: случайный фамильный дрейф, дивергенция концентраций фамилий и аллелей, изонимия, угловые расстояния, стабилизация темпа дивергенции.

DOI: 10.31857/S0016758241200103 **EDN:** VZVXWJ

Изучение фамильной структуры популяций человека интересно не только само по себе, но и как отражение действующих на уровне популяции процессов, как отражение происхождения популяций и как косвенное свидетельство характера генетической структуры (см. [1], где имеется обширная библиография). Дело не только в том, что фамилии могут наследоваться патрилинейно и передаваться сходно с генами негомологичного участка Y-хромосомы, но и в том, что характер типичных популяционных процессов (миграция, изоляция, популяционные волны численности и др.) близким образом влияет на распределение генов и на распределение фамилий. Сходство в передаче потомкам фамилии и генов позволяет использовать фамильные данные при изучении структуры ДНК Y-хромосомы (см. обзор [2, 3]) и в ряде случаев сузить круг фамилий подозреваемых в криминалистике. К настоящему времени проведены широкие исследования фамильной структуры во многих странах и их внутренних регионах, в том числе в России (см., например, [4 с картографическим анализом, 5]). Количество соответствующих работ перевалило за половину тысячи, и обзор современного состояния данной области заслуживает отдельной публикации, а здесь мы ограничились ссылками преимущественно на монографии, но упомянем посвященную библиографии работу [6], в которой источники сгруппированы по изучаемым странам.

Отметим, что с термином изонимия связаны оставшиеся за рамками настоящей статьи популярные подходы, основанные на использовании данных по частоте браков между однофамильцами для оценивания коэффициента инбридинга в популяции [7–9] (см. критические замечания в [10]). Мы не рассматриваем используемые в публикациях такие характеристики фамильной структуры популяции, как индекс случайной изонимии и показатели разнообразия фамилий. При анализе распределения фамилий наш фокус лежит на других подходах и методах, применяемых в популяционной генетике. Конечно, при этом требуются определенные коррекции в методах исследования и в интерпретации результатов. *Цель настоящей*

работы состоит в адаптации методов популяционно-генетического анализа применительно к изучению фамильной структуры и ее связи с генетической структурой, а также обоснования теоретического фундамента таких методов.

Для достижения указанной цели используем упрощенную модель случайного фамильного дрейфа в популяции с неперекрывающимися поколениями [11–13]. Согласно закономерностям репродукции при оплодотворении зигота получает случайным образом один из двух аллелей аутосомного локуса отца и один от матери, т. е. генотип потомка представляет собой случайную выборку аллеля от отца и аллеля от матери. На популяционном уровне при случайном комбинировании генотипов родителей при неперекрывающихся поколениях *генетический* состав популяции потомков является результатом случайного выбора аллелей из родительской популяции. Аналогично *фамильный* состав потомков формируется как случайная выборка фамилий из мужской составляющей родительской популяции. Данная модель в популяционной генетике известна как модель Райта – Фишера. В ряду неперекрывающихся поколений мы получаем последовательность вложенных выборок. Динамику фамильного состава (изонимии), изменяющегося в результате выборочных ошибок при "копировании" родительского состава, назовем по аналогии с генным дрейфом *процессом фамильного дрейфа*.

Использование предположения о неперекрывающихся поколениях, когда речь идет о популяциях человека, проблематично, так как входит в противоречие с реальным положением вещей. Однако допустить такое использование можно на основе многочисленных результатов изучения с его помощью разнообразных реальных популяций. Достаточно вспомнить проверку закона Харди–Вайнберга, полученного в своей классической форме для популяций с неперекрывающимися поколениями. Отметим также, что многие выводы при изучении разнообразных природных популяций получены с помощью приложения результатов непрерывной аппроксимации для дискретных моделей популяций с неперекрывающимися поколениями.

Другая проблема использования модели случайного дрейфа связана с тем, что реальные популяции подвержены одновременному давлению нескольких факторов микроэволюции. Тем не менее такое использование оправдано, так как при сравнимом по результатам давлении систематических факторов и случайного дрейфа последний доминирует на относительно небольших промежутках времени [14, 15]. В данном контексте фамильное состояние очередной популяции потомков моделируется как результат случайной выборки фамилий из их совокупности в мужском компоненте родительской популяции.

Дальнейшее изложение придерживается следующего плана. Сначала формулируются основные понятия, используемые при изучении фамильной структуры. Затем обсуждается переход от традиционных фамильных состояний популяции в терминах концентраций однофамильцев к состояниям с координатами в виде квадратных корней из концентраций. Далее обосновывается аппроксимация распределения фамильных состояний популяции, описываемых угловым отклонением θ от начального состояния, нормальным распределением. Преимущество новых координат состоит в достижении независимости от состояния популяции эффектов случайного дрейфа.

Кратко коснемся обозначений. Названия векторов и матриц набраны полужирным шрифтом (заглавными буквами для матриц, матрица с элементами a_{ij} обозначается как $[a_{ij}]$). К обозначениям фамильных аналогов популяционно-генетических характеристик добавлено окончание *s* (для дисперсий *Vs* и углов θs соответственно). Символ *E* относится к операции получения среднего значения (*математического ожидания*). Когда у *E* имеется нижний индекс, то подразумевается, что усреднение производится по переменной, обозначаемой этим индексом. Расстояние между точками x и y в Евклидовом пространстве обозначаем как $|x - y|$. Знак тождества "≡" используется в смысле равенства по определению. Символ ◄ отмечает конец доказательства.

ОСОБЕННОСТИ ВЫБОРОЧНОГО ДРЕЙФА ФАМИЛИЙ В ОДНОЙ ИЗОЛИРОВАННОЙ ПОПУЛЯЦИИ

Фамильное состояние популяции определяется как набор (вектор) концентраций групп мужчин-однофамильцев, короче концентраций фамилий в популяции. В модели процесса случайного дрейфа фамилий последовательность фамильных состояний по неперекрывающимся поколениям представляет собой *цепь результатов вложенных случайных выборок с возвращением* из фамилий мужских компонентов соответствующих родительских популяций. Вероятность появления определенной фамилии при извлечении выборочной единицы (у нас сына) равна концентрации этой фамилии среди родителей (среди глав семей). Формально каждая выборка рассматривается как мужская составляющая популяции в очередном поколении, а последовательность выборок в этой схеме определяет динамику изонимии в ряду неперекрывающихся поколений. Распределение состава выборки (распределение возможного фамильного состава популяции в следующем поколении, т. е. концентраций фамилий) является *полиномиальным* (мультиномиальным).

Любое состояние популяции как совокупности, состоящей из групп однотипных объектов (у нас групп однофамильцев), можно геометрически представить в Евклидовом пространстве как точку (вектор из начала координат) **x** на части гиперплоскости над полуосями неотрицательных координат (см. рис. 1), координаты точки **x** равны концентрациям групп $\{x_i\}$. Эта гиперплоскость отсекает единичные отрезки на осях координат и состоит из множества точек **x** таких, что в случае k групп

$$\mathbf{x}=\left(x_1,x_2,\ldots,x_k\right)^T,\quad x_i\geq 0,\ \sum_{i=1}^{k}x_i=(\mathbf{x},\mathbf{e})=1,\quad \mathbf{e}\equiv(1,1,\ldots.1)^T.$$

Здесь T – символ транспонирования, **(x, e)** – скалярное произведение вектора-состояния **x** и вектора нормали **e** к рассматриваемой плоскости $(\mathbf{x}-\mathbf{p},\mathbf{e})=0$, к плоскости отклонений **x** от (*начального*) состояния **p**. На границе множества состояний (фазового пространства) концентрация одной из групп равна нулю. Выборочное отклонение **x** от **p** можно охарактеризовать квадратом Евклидова расстояния $|\mathbf{x}-\mathbf{p}|$ между **x** и **p**.

$$|\mathbf{x}-\mathbf{p}|^2\equiv(\mathbf{x}-\mathbf{p},\mathbf{x}-\mathbf{p})=\sum_{i=1}^{k}\left(x_i-p_i\right)^2.$$

Задача настоящей работы состоит в упрощении анализа динамики фамильного и генетического состояний популяции с неперекрывающимися поколениями, изменяющихся в результате случайного дрейфа. Анализ как бы обращает нас к модели случайного генного дрейфа с дискретным временем, для которой большинство результатов выведены с использованием аппроксимации непрерывными аналогами. Получаемые выводы могут использоваться при оценивании инбридинга.

Повторим, что процесс случайного фамильного дрейфа популяции с неперекрывающимися поколениями является последовательностью выборочных изменений фамильного состояния при смене поколений (последовательностью вложенных случайных выборок с возвращением). Состояние популяции с k вариантами фамилий в следующем (первом) поколении представляет собой результат случайной выборки с возвращением фамилий из множества фамилий мужчин родительской популяции. Размер выборки фамилий равен $N(1)/2$, где $N(1)$ – численность диплоидной популяции в первом поколении, а $N(1)/2$ – численность ее мужского компонента, передающего свои фамилии по поколениям.

Выборка является случайной при независимых выборах фамилии для каждого потомка. Повторим, что ее размер равен $N(1)/2$. *Хотя мы далее интерпретируем N как размер популяции, ключевым является размер выборки мужчин, носителей наследуемых фамилий* (с учетом дополнительных поправок его можно назвать эффективным дисперсионным размером мужского компонента, в нашем случае он взят для простоты равным $N/2$). Интерпретация N как общего размера популяции условна, и N фактически играет роль параметра. Чем меньше размер $N/2$ мужского компонента популяции, тем более интенсивны выборочные отклонения нового фамильного состояния от прежнего, а величина разброса выборочных колебаний определяет “темп” дивергенции фамильных состояний от начального.

В первом поколении вероятность попадания в выборку i-й фамилии равна p_i при каждом из $N(1)/2$ испытаний (при каждом выборе фамилии для потомка). Вероятности $\{p_i\}$ равны концентрациям фамилий в начальном фамильном состоянии популяции **p**. При описанной схеме распределение результатов выборки является полиномиальным, как говорилось выше.

Пусть вектор $\mathbf{x}=\mathbf{x}(1)$ с концентрациями фамилий $\{x_i\}$ обозначает состояние популяции в первом поколении. Если рассматривать только какую-либо одну из координат вектора-состояния **x** (концентрацию отдельной, скажем, i-й фамилии), то вероятность попадания этой фамилии в выборку (“успеха”) в результате одного из $N(1)/2$ испытаний при формировании первого поколения равна p_i. Концентрация i-й фамилии x_i в выборке (в следующем поколении) является результатом деления количества успехов на размер выборки $N(1)/2$ (т. е. деления суммы $N(1)/2$ независимых биномиальных переменных с вероятностью успеха p_i, равной концентрации фамилии среди родоначальников). Соответствующее распределение количества успехов в выборке является биномиальным.

Ожидаемым (средним) значением концентрации x_i для i-й фамилии в новом поколении будет прежнее значение p_i, а дисперсия выборочных отклонений x_i от p_i равна $p_i(1-p_i)/\dfrac{N(1)}{2}$ в соответствии со свойствами биномиальных испытаний. Таким образом, *у случайного дрейфа нет преимущественного направления* (ожидаемое значение концентраций фамилий в следующем поколении совпадает с предыдущим значением).

При отсутствии направления у динамики фамильного состояния в результате случайного дрейфа ее можно характеризовать разбросом возможных отклонений состояний от начального значения, увеличивающимся в силу накопления выборочных ошибок в ряду поколений, т. е. характеризовать степенью и темпом дивергенции от исходного положения. Величина фамильной дивергенции за поколение, “темп” ненаправленной эволюции, измеряемая, скажем, средним абсолютным отклонением или средним квадратическим

отклонением, или средним квадратом отклонения (дисперсией), равным $V(x_i) = p_i(1 - p_i)/\frac{N(1)}{2}$ (а не просто средним отклонением, которое при ненаправленной эволюции равно нулю), зависит, как видим, от значения рассматриваемой концентрации p_i в родительской популяции. При одинаковых прочих условиях выборочная дисперсия x_i как характеристика скорости ненаправленной дивергенции определяется значением p_i (дисперсия пропорциональна $p_i(1 - p_i)$). Тем самым темп дивергенции для концентрации i-го аллеля (фамилии) зависит от текущего значения p_i и со временем меняется вместе с ним, как подчеркивалось многими исследователями.

В один и тот же момент времени в одной и той же популяции темп дивергенции для другой фамилии с другой концентрацией будет в общем случае иным. Выборочные дисперсии характеризуют случайную ненаправленную динамику и могут служить показателем скорости дивергенции. Однако затруднительно определить, из-за чего различия в величине выборочного отклонения разных фамилий достигают наблюдаемого значения — объясняется ли это только темпом дивергенции, зависимым от их концентраций среди родителей, или причиной является, например, давление некоторого фактора. Кроме того, одинаковые значения отклонений характеризуются по-разному при разных значениях p.

При одновременном изучении концентраций множества фамилий разброс выборочных отклонений характеризуется матрицей ковариаций, зависящей от значений концентраций фамилий в родительской популяции. В случае разных концентраций фамилий темп дивергенции от начального состояния отличается как по различным направлениям (по разным осям координат, на которых откладываются концентрации соответствующих фамилий), так и по отдельной оси в зависимости от значения концентрации. Здесь возникает задача оптимального объединения без потери информации данных по отдельным фамилиям для получения единой характеристики дивергенции фамильного состояния от начального значения и между разными популяциями с общим происхождением.

На этом пути желательно использовать такое преобразование координат, когда темп дивергенции стабилен и не зависит ни от направления, ни от текущего состояния. Тогда упрощается построение обобщенной характеристики динамики отклонений состояния популяции от начальной точки для получения единого показателя, облегчающего сравнение популяций. Такими полезными статистическими особенностями дивергенции обладает, например, случай, когда выборочные отклонения имели бы стандартное многомерное нормальное распределение с единичными дисперсиями по каждой из независимо и случайно изменяющихся переменных состояния.

Для получения подобного показателя рассмотрим в следующем разделе обобщение преобразования $\theta = \arccos\sqrt{p}$ (облегчающего изучение динамики отдельной фамилии подобно используемому в [16]) на случай анализа одновременных изменений множества фамилий (см., например, [15]). При рассматриваемом обобщении достигается изотропность темпа отклонения (дивергенции) популяции от начального состояния в результате случайного дрейфа.

ПРОСТРАНСТВО СОСТОЯНИЙ ПОПУЛЯЦИИ С МНОЖЕСТВОМ ФАМИЛИЙ КАК ОБЛАСТЬ ГИПЕРСФЕРЫ

В связи с зависимостью скорости дивергенции от состояния популяции **x** возникает задача добиться, чтобы характер случайных выборочных колебаний был бы одним и тем же для любого вектора концентраций **x** различных фамилий (аллелей), т. е. не зависел от состояния. Начать можно с такого преобразования отдельных концентраций, при котором на дисперсию преобразованной биномиальной переменной не влияет вероятность успеха. Для этой цели Р. Фишером были предложены арксинус-преобразование и преобразование $\cos\theta = 1 - 2p$ [16, 17]. Последнее было использовано им для анализа генного дрейфа по концентрации одного аллеля, его применение стабилизировало выборочную дисперсию, принимающую постоянное значение. У нас подобный подход означает изучение свойств случайной динамики концентраций фамилий по отдельности. Результаты такого изучения остаются корректными в качестве части общей динамики при исследовании всего множества фамилий, так как процесс случайного дрейфа допускает произвольную группировку фамилий, предельным случаем которой будет группа из одной фамилии.

Для стабилизации темпа дивергенции по концентрации *одной* из фамилий можно использовать угловую переменную θ, получаемую преобразованием $\theta = \arccos\sqrt{p}$, при котором выборочная дисперсия θ не зависит от концентрации p. Геометрически углу θ соответствует согласно школьному курсу тригонометрии точка на единичной окружности с центром в начале координат (точнее, у нас точка на части этой окружности в первой четверти) или радиус-вектор данной точки. При этом на оси абсцисс откладывается $\sqrt{p}$, косинус угла между радиус-вектором указанной точки p и осью абсцисс, а по оси ординат $\sqrt{1-p}$ (косинус угла с осью ординат) — см. рис. 1. Угол θ *измеряем*

в радианах. На тригонометрической окружности он совпадает с длиной дуги между осью абсцисс и $y \equiv \sqrt{p}$. При малых отклонениях y_1 от $y = y_0$ длина дуги между этими точками приближенно равна хорде между y_0 и y_1 или расстоянию между y_0 и соответствующей y_1 точкой на касательной прямой в точке y_0 к окружности, что используется в дальнейшем. Каждая точка $y \equiv (\sqrt{p},\ \sqrt{1-p}\ )$ на окружности служит геометрическим образом фамильного состояния популяции при наличии только двух фамилий, а пространством состояний является часть тригонометрической окружности в первой четверти.

При обобщении этой картины на случай k фамилий (групп) в популяции [14] получим, что в пространстве состояний (фазовом пространстве) будет k осей координат, на которых откладываются k значений $\{\sqrt{x_i}\}$, направляющих косинусов $y_i = \sqrt{x_i}$ радиуса-вектора состояния в k-мерном пространстве (косинусов углов θ_i между радиус-вектором и i-й осью координат). На этом пути перейдем к более строгому изучению стабилизации темпа дивергенции в последовательности выборочных отклонений при случайном дрейфе популяции с несколькими группами однофамильцев. При этом через **p** обозначаем *начальное состояние популяции*, а **x** относим к состоянию с учетом выборочного отклонения.

Итак, перейдем к изучению свойств преобразования координат $\{y = \sqrt{x_i}\}$, при котором по i-й оси откладывается корень квадратный из концентрации i-й фамилии. Все множество {y} состояний популяции в новых координатах состоит из точек части поверхности гиперсферы (с радиусом $R = 1$ и с центром в начале координат), которая находится над полуосями неотрицательных координат с границей из состояний, у которых имеется нулевая координата. Например, для трех переменных (концентраций) множество состояний популяции состоит из точек {**y**} вида

$$\mathbf{y} = (y_1, y_2, y_3)^T = \left(\sqrt{x_1}, \sqrt{x_2}, \sqrt{x_3}\right)^T,$$

$$R^2 = {y_1}^2 + {y_2}^2 + {y_3}^2 = x_1 + x_2 + x_3 = 1,$$

что иллюстрируется рис. 1. Оказывается, что дивергенция состояний, кроме примыкающих к границе фазового пространства, обладает желательными свойствами стабилизации (правда, за такое преимущество приходится платить ограничением на величину промежутка времени, когда преимущество существует).

Пусть $\theta(\mathbf{y}, \mathbf{y}_0)$ обозначает угол между двумя векторными состояниями **y** и $\mathbf{y}_0$ (соответствующими **x** и **p**, см. рис. 1). Отклонения **x** от начального состояния **p** (**y** от $\mathbf{y}_0$) можно измерять различными способами, из которых в координатах **y** длина дуги большого круга на гиперсфере привлекательна своими статистическими и геометрическими свойствами. Она аналогична прямой в Евклидовом пространстве в том смысле, что также дает кратчайшее

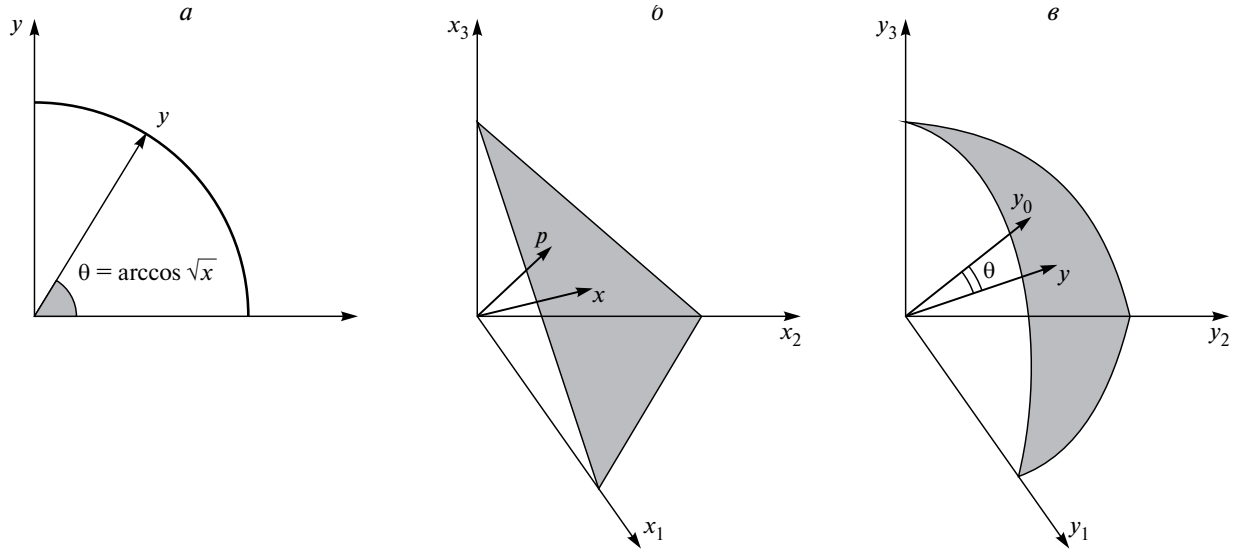

Рис. 1. Пространство состояний популяции в различных системах координат.
а – затененный угол θ между радиус-вектором y_0 и осью абсцисс;
б – затененная часть плоскости как пространство состояний популяции в терминах концентраций групп;
в – затененная часть сферы как пространство состояний популяции в терминах квадратных корней из концентраций.
Объяснения см. в тексте.

расстояние между двумя точками (теперь на гиперсфере). Кроме того, длина дуги на единичной гиперсфере совпадает с угловым расстоянием $\theta(\mathbf{x}, \mathbf{p}) = \theta(\mathbf{y}_1, \mathbf{y}_0)$, которое, как говорилось, измеряем в радианах

$$\theta(\mathbf{x}, \mathbf{p}) \equiv \theta(\mathbf{y}_1, \mathbf{y}_0) \equiv \arccos\left(\sum_{i=1}^{k} \sqrt{x_i}\sqrt{p_i}\right). \qquad (1)$$

Рассматриваемое преобразование было предложено в [18] с точки зрения, главным образом, изучения выборочных свойств статистики $\cos\theta$, ее связи с критерием хи-квадрат и др. В [14, 19] доказана изотропность пространства выборочных отклонений на гиперсфере. Если сдвинуть одинаковым образом как точку **x**, так и **p**, то обычное (Евклидово) расстояние между ними в Евклидовом пространстве останется прежним. Аналогично угловое расстояние $\theta(\mathbf{x}, \mathbf{p})$ на гиперсфере не изменится при соответствующем сдвиге **x** и **p**. Описанная картина верна для любой популяции как совокупности, состоящей из непересекающихся групп однотипных объектов.

АНАЛИЗ РАЗБРОСА УГЛОВОГО ОТКЛОНЕНИЯ

При случайном фамильном дрейфе по одной фамилии с начальной концентрацией p в популяции результат добавления в выборку потомка (у нас в выборку размера $N/2$ потомков мужского пола) является случайной величиной со значениями 1 (если у потомка окажется рассматриваемая фамилия, вероятность этого "успеха" равна p) и 0 (в противном случае). Распределение суммы случайных величин (количества успехов во всей выборке), получаемой при этом, является *биномиальным*. Повторим, что согласно известным свойствам биномиального распределения с вероятностью успеха p (см., например, [20]) у полученной концентрации x в случайной выборке размера $N/2$ математическое ожидание (среднее значение) $E\{x\}$ и дисперсия $Vs\{x\}$ равны соответственно

$$E\{x\} = p, \quad Vs\{x\} = p(1-p)\Big/\frac{N}{2}.$$

Согласно центральной предельной теореме сумма достаточно большого количества сравнительно малых случайных величин ведет себя как нормальная случайная величина, т. е. при большом $N/2$, где N – размер популяции с учетом обоих полов, распределение x (суммы "успехов", деленной на $N/2$) является приближенно нормальным с приведенными значениями $E\{x\}$ и $Vs\{x\}$, обозначаемым как $N(p, p(1-p)/\frac{N}{2})$.

При использовании преобразования $\theta(p) = \arccos\sqrt{p}$ у угла $\theta(x)$ также будет приближенно нормальное распределение. Покажем, что у него математическое ожидание приближенно равно $\theta(p)$, а дисперсия не зависит от p. У нас p обозначает исходную концентрацию рассматриваемой фамилии, а в контексте популяционной генетики p имеет смысл концентрации рассматриваемого аллеля аутосомного локуса в родоначальной популяции. Напомним, что когда речь идет об аллелях, для углового отклонения используем обозначение "θ", а когда имеются в виду фамилии, к "θ" добавляем "s", т. е. значение $\theta s(p)$ характеризует фамильное состояние популяции.

Найдем приближенно дисперсию значений $\theta s(x)$ в новом поколении с помощью известного δ-метода (см., например, [20]) следующим образом. Новая концентрация в поколении потомков получается прибавлением к p случайного выборочного отклонения δp с нулевым математическим ожиданием и дисперсией $p(1-p)/\frac{N}{2}$. Для получения примерного значения $\theta s(x) \equiv \theta s(p + \delta p)$ в следующем поколении используем член первого порядка по δp в разложении Тейлора $\theta s(p)$: $\theta s(x) \approx \theta s(p) + (d\theta s(p)/dp)\delta p$. Здесь $\theta s(p)$ – константа, $d\theta s(p)/dp$ – постоянный множитель при случайной переменной $\delta p \equiv x - p$ с нулевой средней величиной и с дисперсией $p(1-p)/\frac{N}{2}$, соответствующей дисперсии концентрации фамилии в следующем поколении.

Как известно, дисперсия произведения константы на случайную величину δp равна произведению дисперсии δp на квадрат константы. Отсюда вычисление приближенной (межпопуляционной) дисперсии $V(\theta s)$ теоретически мыслимых вариантов фамильных состояний θs популяции в следующем поколении дает известное значение

$$V(\theta s) \approx \left(d\theta s(p)/dp\right)^2 p(1-p)\Big/\frac{N}{2} = \frac{2}{N}.$$

Таким образом, при замене p на $\theta s = \theta s(p)$ выборочное отклонение новой переменной θs от фамильного состояния родительской популяции $\theta s(p)$ при достаточной величине N приближенно имеет нормальное распределение с нулевой средней и дисперсией $\frac{2}{N}$, независимой от значения p.

Ремарка 1. Отметим, что при выводе дисперсии $V(\theta s)$ мы в разложении Тейлора ограничились членом с первой производной $d\theta s(p)/dp$ и пренебрегли следующими. Однако уже вторая производная неограниченно растет, когда p стремится к нулю. Поэтому можно пользоваться полученной аппроксимацией дисперсии, когда p превышает надлежащий порог, выбираемый из условия малости эффекта следующего члена в разложении Тейлора для θs.

Обоснуем более строго свойства углового расстояния при случайном дрейфе генов и фамилий.

Результат 2. *Пусть концентрация x аллеля (фамилии) в популяции с неперекрывающимися поколениями определяется при каждой смене поколений случайной выборкой с возвращением из совокупности аллелей рассматриваемого локуса (из совокупности фамилий) родительского поколения. Положим, что размер такой выборки на шаге* τ *равен* $2Ne(\tau)\left(\frac{Ne(\tau)}{2}\right)$, *где* $Ne(\tau)$ *обозначает эффективную численность популяции в поколении* τ (см., например, [11, 12]).

Пусть t поколений тому назад концентрация данного аллеля (фамилии) в популяции была равна p, и расстояние (отклонение) между текущим x и начальным p состояниями в угловых координатах $\theta(x) \equiv \equiv \arccos(\sqrt{x})$ *находится как* $|\theta(\sqrt{x}) - \theta(\sqrt{p})|$ (*соответственно как* $|\theta s(\sqrt{x}) - \theta s(\sqrt{p})|$).

Тогда при $\frac{\widetilde{N}e(t)}{t} \to \infty$ *и при* $\sqrt{p}$, *превышающем надлежащим образом выбранный порог, асимптотические распределения для* $\theta(\sqrt{x(t)})$ *и* $|\theta(\sqrt{x(t)}) - \theta(\sqrt{p})|^2$, *а также для* θs (*переменной фамильного состояния, аналогичной* θ) *и их параметры имеют вид*:

$$
\begin{aligned}
&E\left\{\theta\left(\sqrt{x(t)}\right)\right\} = \theta\left(\sqrt{p}\right), \quad E\left\{\left|\theta\left(\sqrt{x(t)}\right) - \theta\left(\sqrt{p}\right)\right|^2\right\} = \frac{t}{8\widetilde{N}e(t)},\\
&\theta\left(\sqrt{x(t)}\right) = N\left(\sqrt{p}, \frac{t}{8\widetilde{N}e(t)}\right), \frac{8\widetilde{N}e(t)}{t}\left|\theta\left(\sqrt{x(t)}\right) - \theta s\left(\sqrt{p}\right)\right|^2 = \chi^2,\\
&E\left\{\theta s\left(\sqrt{x(t)}\right)\right\} = \theta s\left(\sqrt{p}\right), \quad E\left\{\left|\theta s\left(\sqrt{x(t)}\right) - \theta s\left(\sqrt{p}\right)\right|^2\right\} = \frac{t}{2\widetilde{N}e(t)},\\
&\theta s\left(\sqrt{x(t)}\right) = N\left(\sqrt{p}, \frac{t}{2\widetilde{N}e(t)}\right), \frac{2\widetilde{N}e(t)}{t}\left|\theta s\left(\sqrt{x(t)}\right) - \theta s\left(\sqrt{p}\right)\right|^2 = \chi^2.
\end{aligned}
\tag{2}
$$

Здесь $N(m, V)$ *– символ нормального распределения с математическим ожиданием m и дисперсией V,* $\widetilde{N}e(t) \equiv t \Big/ \sum_{\tau=1}^{t} \frac{1}{Ne(\tau)}$, $\widetilde{N}e(t)$ *и* χ^2*обозначают среднюю гармоническую численность популяции для ряда* $\{Ne(\tau), \tau = 1, 2, ..., t\}$ *и распределение хи-квадрат* (*с одной степенью свободы*) *соответственно.*

Доказательство проведем для конкретности в случае анализа фамильных состояний. Рассмотрим последовательность нескольких поколений, в τ-м из которых фамильное состояние популяции представляет собой выборку $Ne(\tau)/2$ фамилий из предыдущего поколения. Пусть на первом шаге реализовалось состояние $\theta s(1) \equiv \theta s(x(1))$. Случайное отклонение $\delta\theta s(1)$ нового значения θs от начального состояния $\theta s(p)$, как говорилось ранее, приближенно имеет нормальное распределение с нулевой средней и независимой от p дисперсией $\frac{1}{2Ne(1)}$. На втором шаге следующее выборочное отклонение $\delta\theta s(2)$ не коррелирует с предыдущим и будет нормальным с нулевой средней и независимой от $x(1)$ дисперсией $\frac{1}{2Ne(2)}$, где $Ne(2)$ – очередной эффективный размер популяции. Распределение суммы двух нормально распределенных некоррелирующих случайных отклонений с нулевыми средними и дисперсиями $\frac{1}{2Ne(1)}$ и $\frac{1}{2Ne(2)}$ является нормальным распределением с нулевой средней и дисперсией $\frac{1}{2Ne(1)} + \frac{1}{2Ne(2)}$. Продолжая эти рассуждения, мы получим, что в поколении t итоговое суммарное отклонение значения θs от начальной величины приближенно распределено нормально с нулевой средней и дисперсией $\sum_{\tau=1}^{t} \frac{1}{2Ne(t)} = \frac{t}{2\widetilde{N}e(\tau)}$, складывающейся из дисперсий отклонений на отдельных шагах (поколениях). Здесь $\widetilde{N}e(t) \equiv \equiv t/\sum_{\tau=1}^{t} \frac{1}{Ne(\tau)}$, $Ne(\tau)$ обозначает эффективный размер популяции в поколении τ, $\widetilde{N}e(t)$ – средняя гармоническая численность популяции для ряда $\{Ne(\tau)\}$, равная обратной величине к среднему арифметическому $\left(\sum_{\tau=1}^{t} \frac{1}{Ne(\tau)}\right)/t$ для $\left\{\frac{1}{Ne(\tau)}\right\}$. Отсюда вытекает, что у нормированного квадрата углового расстояния $\frac{2\widetilde{N}e(t)}{t}\,\theta s^2$ будет распределение хи-квадрат с одной степенью свободы. ◄

Распределение отклонения $\theta s(t)$ от начального значения в простом частном случае постоянного размера N у популяции приближенно является нормальным $N(0, \frac{t}{2N})$ с нулевым математическим ожиданием и дисперсией $\frac{t}{2N}$, а нормированный квадрат углового расстояния $\frac{2N}{t}\theta s^2$ имеет распределение хи-квадрат с одной степенью свободы. Еще раз напомним, что приведенные результаты корректны, когда значение t мало по сравнению $\widetilde{N}e(t)$.

Хотя далее у нас речь идет о совокупности фамилий, результаты имеют общий характер и приложимы к случайным выборкам из любой совокупности дискретных объектов, сгруппированных согласно их типам с соответствующими вероятностями попадания типов в выборку. Таким образом, если слово “фамилия” заменить на название объекта, то выводы останутся верными для такого случая, например для концентраций аллелей. Доказываемые факты относительно выборочных свойств угла θ являются вариантом результата, полученного [18] в области статистики, адаптированным применительно к дрейфу фамилий в духе анализа

генного дрейфа в [14]. Сформулируем эти факты более строго для многомерного случая.

Результат 3. *Пусть дана случайная выборка с возвращением размера N/2 из популяции с k вариантами фамилий. Положим, вероятность извлечения i-й фамилии равна ее концентрации* $x_i > 0$, $i = 1, 2, \ldots k$, $\sum_{i=1}^{k} x_i = 1$. *Вектор с координатами* $\{x_i\}$ *обозначим как* **x**, *а случайный вектор концентраций фамилий в выборке как* $\mathbf{x}_1 \equiv \mathbf{x}(1)$ *и определим преобразование* **y**(**x**) *как*

$$\mathbf{y}(\mathbf{x}) = \left(y_1, y_2, \ldots y_k\right)^T \equiv \left(\sqrt{x_1}, \sqrt{x_2}, \ldots \sqrt{x_k}\right)^T,$$
$$\sum_{i=1}^{k} y_i^2 = 1, \quad \mathbf{y}(1) \equiv \mathbf{y}\left(\mathbf{x}(1)\right), \mathbf{y}_0 = \mathbf{y}\left(\mathbf{x}_0\right). \tag{3}$$

Тогда асимптотически при $N \to \infty$

$$\mathbf{y}(1) = N\left(\mathbf{y}_0, \frac{1}{2N}\mathbf{W}\left(\mathbf{y}_0\right)\right), \quad \mathbf{W}\left(\mathbf{y}_0\right) \equiv \mathbf{I} - \mathbf{y}_0\mathbf{y}_0^T. \tag{4}$$

Здесь $N(\mathbf{m}, \mathbf{V})$ *– символ многомерного нормального распределения с вектором математического ожидания* ***m*** *и матрицей ковариаций* **V**, **I** *– единичная матрица.*

Доказательство. Распределение фамильного состава выборки, получаемой при сделанных предположениях, является *полиномиальным* (мультиномиальным). Согласно известным свойствам полиномиального распределения с вероятностями $\{x_i\} = \mathbf{x}$ (см., например, [20]) у полученных в случайной выборке размера $\frac{N}{2}$ концентраций $\mathbf{x}_1$ (и их отклонений от **x**) матрица ковариаций **V**(**x**) имеет вид

$$\mathbf{V}(\mathbf{x}) = \frac{2}{N}\left[x_i\left(\delta_{ij} - x_j\right)\right] = \frac{2}{N}\left(\mathbf{D}(\mathbf{x}) - \mathbf{x}\mathbf{x}^T\right) =$$
$$= \frac{2}{N}\left(\mathbf{D}^2(\mathbf{y}) - \left(\mathbf{D}(\mathbf{y})\mathbf{y}\right)\left(\mathbf{y}^T\mathbf{D}(\mathbf{y})\right)\right),$$

где δ_{ij} обозначает символ Кронекера ($\delta_{ij} = 1$ при $i = j$ и нулю в противном случае), **D**(**x**)(**D**(**y**)) – диагональная матрица с координатами вектора **x** (соответственно **y**) на главной диагонали. Дальнейшее доказательство разобьем на пункты.

1. При преобразовании (3) можно приближенно найти математическое ожидание $E\{\mathbf{y}(\mathbf{x})\}$ и матрицу ковариаций **V**(**y**) для координат $\{y_i\}$ вектора $\mathbf{y} = \mathbf{y}(\mathbf{x})$ с помощью δ-метода, использующего члены первого порядка в разложении Тейлора **y**(**x**). Применим его как к дисперсиям (см., скажем, [20]), так и ковариациям. Повторим, что когда $y(x)$ получается преобразованием случайной переменной x с математическим ожиданием x_0, то

$$y(x) = y\left(x_0 + (x - x_0)\right) \equiv y(x_0 + \delta x) \approx y(x_0) + \left(dy(x_0)/dx\right)\delta x,$$
$$\delta y \equiv y(x) - y(x_0) \approx \left(dy(x_0)/dx\right)\delta x,$$

где $dy(x_0)/dx$ – константа, а δx – случайное отклонение x от x_0 с нулевым математическим ожиданием ($E\{\delta x\} = 0$). Отсюда

$$E\left\{y(x)\right\} \approx E\left\{y(x_0)\right\} + \left(dy(x_0)/dx\right)E\left\{\delta x\right\} = y(x_0),$$
$$V(y) \equiv E\left\{\delta y^2\right\} \approx \left(dy(x_0)/dx\right)^2 V(x).$$

2. Когда **y**(**x**) получен преобразованием случайного вектора **x** с математическим ожиданием $\mathbf{x}_0$, то для i-й координаты $y_i(\mathbf{x})$ вектора **y**(**x**) имеем

$$y_i(\mathbf{x}) = y_i\left(\mathbf{x}_0 + (\mathbf{x} - \mathbf{x}_0)\right) \equiv y_i(\mathbf{x}_0 + \delta x) \approx$$
$$\approx y_i(\mathbf{x}_0) + \sum_j \left(\partial y_i(\mathbf{x}_0)/\partial x_j\right)\delta x_j,$$

где $\partial y_i(\mathbf{x}_0)/\partial x_j$ – константы, а δx_j – случайные отклонения координат **x** от координат $\mathbf{x}_0$ с нулевым математическим ожиданием ($E\{\delta x_j\} = 0$).

В векторно-матричном виде это соотношение можно переписать как

$$\mathbf{y}(\mathbf{x}) = \mathbf{y}(\mathbf{x}_0 + \delta\mathbf{x}) \approx \mathbf{y}(\mathbf{x}_0) + [\partial_{ij}\mathbf{y}(\mathbf{x}_0)]\,\delta\mathbf{x},$$
$$\delta\mathbf{y} \equiv \mathbf{y}(\mathbf{x}) - \mathbf{y}(\mathbf{x}_0) \approx \left[\partial_{ij}\mathbf{y}(\mathbf{x}_0)\right]\delta\mathbf{x}, \left[\partial_{ij}\mathbf{y}(\mathbf{x}_0)\right] \equiv \left[\partial y_i(\mathbf{x}_0)/\partial x_j\right].$$

Так как константы $\partial_{ij}\mathbf{y}(\mathbf{x}_0)$ можно выносить за знак математического ожидания E и согласно полученному выше $E\{\delta\mathbf{x}\} = \mathbf{0}$, то

$$E\left\{\mathbf{y}(\mathbf{x})\right\} \approx E\left\{\mathbf{y}(\mathbf{x}_0)\right\} + \left[\partial_{ij}\mathbf{y}(\mathbf{x}_0)\right]E\left\{\delta\mathbf{x}\right\} = \mathbf{y}(\mathbf{x}_0).$$

3. Теперь обратимся к вычислению матрицы ковариаций **V**(**y**) случайного вектора-столбца **y**. По определению $\mathbf{V}(\mathbf{y}) \equiv E\{\delta\mathbf{y} \times \delta\mathbf{y}^T\}$, $\mathbf{V}(\mathbf{x}) \equiv E\{\delta\mathbf{x} \times \delta\mathbf{x}^T\}$, подстановка в **V**(**y**) приведенного выше значения $\delta\mathbf{y}$ дает

$$\mathbf{V}(\mathbf{y}) = \mathbf{V}(\delta\mathbf{y}) \approx E\left\{\left[\partial_{ij}\mathbf{y}(\mathbf{x}_0)\right]\delta\mathbf{x} \times \delta\mathbf{x}^T\left[\partial_{ij}\mathbf{y}(\mathbf{x}_0)\right]^T\right\} =$$
$$= \left[\partial_{ij}\mathbf{y}(\mathbf{x}_0)\right]E\left\{\delta\mathbf{x} \times \delta\mathbf{x}^T\right\}\left[\partial_{ij}\mathbf{y}(\mathbf{x}_0)\right]^T = \tag{5}$$
$$= \left[\partial_{ij}\mathbf{y}(\mathbf{x}_0)\right]\mathbf{V}(\mathbf{x})\left[\partial_{ij}\mathbf{y}(\mathbf{x}_0)\right]^T.$$

Напомним, что при анализе фамильной структуры размер случайной выборки равен $N/2$, $\mathbf{y}(\mathbf{x}) = \{\sqrt{x_i}\}$ и

$$\mathbf{V}(\mathbf{x}) = \frac{2}{N}\left(\mathbf{D}(\mathbf{x}) - \mathbf{x}\mathbf{x}^T\right) = \frac{2}{N}\left(\mathbf{D}^2(\mathbf{y}) - \left(\mathbf{D}(\mathbf{y})\mathbf{y}\right)\left(\mathbf{y}^T\mathbf{D}(\mathbf{y})\right)\right),$$

$$\left[\partial_{ij}\mathbf{y}(\mathbf{x})\right] \equiv \left[\partial y_i \,/\, \partial x_j\right] = \frac{1}{2}\left[\frac{\delta_{ij}}{\sqrt{x_i}}\right] = \frac{1}{2}\mathbf{D}^{-1}(\mathbf{y}).$$

Подстановка этих выражений в формулу (5) для $\mathbf{V}(\mathbf{y})$ дает

$$\mathbf{V}(\mathbf{y}) \approx \frac{1}{2}\mathbf{D}^{-1}(\mathbf{y})\left(\frac{2}{N}\left(\mathbf{D}^2(\mathbf{y}) - \left(\mathbf{D}(\mathbf{y})\mathbf{y}\right)\left(\mathbf{y}^T\mathbf{D}(\mathbf{y})\right)\right)\right)\frac{1}{2}\mathbf{D}^{-1}(\mathbf{y}) =$$

$$= \frac{1}{2N}\left(\mathbf{I} - \mathbf{y}\mathbf{y}^T\right) \equiv \frac{1}{2N}\mathbf{W}(\mathbf{y}),$$

где I – единичная матрица, $\mathbf{W}(\mathbf{y}) \equiv [\delta_{ij} - y_i y_j] = \mathrm{I} - \mathbf{y}\mathbf{y}^T$. Значит, в результате замены (3) приближенно матрица ковариаций $\mathbf{V}(\mathbf{y})$ для новых переменных $\mathbf{y}$ (и для их отклонений $\boldsymbol{\delta}\mathbf{y}$ от математического ожидания y_0) пропорциональна $\mathbf{W}(\mathbf{y}) \equiv [\delta_{ij} - y_i y_j]$ с множителем $\frac{1}{2N}$ при анализе фамильного дрейфа (и $\frac{1}{8N}$ при дрейфе генов).

4. Чем больше размер выборки, тем теснее располагаются выборочные отклонения вблизи нулевого значения и тем лучше аппроксимируется их распределение многомерным нормальным, причем асимптотически

$$\mathbf{x}_1 = N\left(\mathbf{x}, \frac{2}{N}\left(\mathbf{D}(\mathbf{x}) - \mathbf{x}\mathbf{x}^T\right)\right), \quad \mathbf{y} = N\left(\mathbf{y}_0, \frac{1}{2N}\mathbf{W}(\mathbf{y}_0)\right),$$

$$\delta\mathbf{y} = \mathbf{y} - \mathbf{y}_0 = N\left(0, \frac{1}{2N}\mathbf{W}(\mathbf{y}_0)\right).$$

◀

Теперь покажем, что распределение квадратов Евклидова $|\boldsymbol{\delta}\mathbf{y}|$ и углового θs расстояний между $\mathbf{y}_1$ *и* $\mathbf{y}_0$ является широко употребляемым в биометрии распределением хи-квадрат.

Результат 4. *В рамках предыдущего результата Евклидово* $|\boldsymbol{\delta}\mathbf{y}|$ *и угловое* θs *расстояния между* $\mathbf{y}_1$ *и* $\mathbf{y}_0$ *удовлетворяют*

$$2N|\delta\mathbf{y}|^2 \equiv 2N\left(\mathbf{y}_1 - \mathbf{y}_0, \mathbf{y}_1 - \mathbf{y}_0\right) = 2N\left(\delta\mathbf{y}, \delta\mathbf{y}\right) = \chi^2_{k-1},$$

$$2N\theta s^2(\mathbf{y}_1, \mathbf{y}_0) = \chi^2_{k-1}; \theta s = \arccos\left(\sum_{i=1}^{k}\sqrt{x_i(t)}\sqrt{p_i}\right), \quad (6)$$

где χ^2_{k-1} *обозначает распределение хи-квадрат с* $k-1$ *степенями свободы.*

Доказательство. При сделанных ранее предположениях угловое расстояние θs аппроксимирует Евклидово расстояние на касательной плоскости к гиперсфере. Соответственно распределения квадратов этих расстояний приближенно одинаковы. Матрица ковариаций для $\boldsymbol{\delta}\mathbf{y}$ равна $\frac{1}{2N}\mathbf{W}(\mathbf{y}_0)$, $\mathbf{W}(\mathbf{y}) \equiv \mathbf{I} - \mathbf{y}\mathbf{y}^T$; $\mathbf{y}^T\mathbf{y} = 1$ и для $\sqrt{2N}\,\boldsymbol{\delta}\mathbf{y}$ она равна $\mathbf{W}(\mathbf{y}_0)$. Здесь $\mathbf{I}$ – единичная матрица размера k.

Заметим, что матрица $\mathbf{W}(\mathbf{y})$ является идемпотентной (т. е., как можно легко проверить, $\mathbf{W}^2(\mathbf{y}) = \mathbf{W}(\mathbf{y})$). Кроме того, приближенно $E\{\boldsymbol{\delta}\mathbf{y}\} = 0$. Известно, что для такого случая

$$\left(\sqrt{2N}\delta\mathbf{y}, \sqrt{2N}\delta\mathbf{y},\right) = 2N\left(\delta\mathbf{y}, \delta\mathbf{y}\right) = \chi^2_{\mathbf{tr}\,\mathbf{W}} = \chi^2_{k-1},$$

где $\mathrm{tr}\,\mathbf{W}$ обозначает след матрицы $\mathbf{W}$ (сумму ее диагональных элементов, равную у нас $k-1$). Следовательно, произведение $2N$ на квадрат расстояния (отклонения $|\boldsymbol{\delta}\mathbf{y}|$) между фамильными состояниями $\mathbf{y}_1 \equiv \mathbf{y}(1)$ и $\mathbf{y}_0$ приближенно имеет распределение хи-квадрат с $k-1$ степенями свободы. При больших N отклонения $\boldsymbol{\delta}\mathbf{y}$ с близкой к единице вероятностью малы, значения углового расстояния $\theta s^2(\mathbf{y}_1, \mathbf{y}_0)$ аппроксимируются величинами $|\boldsymbol{\delta}\mathbf{y}|^2$ и асимптотически $2N\theta s^2\left(\mathbf{y}_1, \mathbf{y}_0\right) = \chi^2_{k-1}$. ◀

Обратимся к более наглядной геометрической картине приведенного результата [14]. Напомним, что переход (3) от координат пространства фамильных состояний $\mathbf{x}$,

$$\mathbf{x} = \left(x_1, x_2, \ldots, x_k\right)^T, \quad \sum_{i=1}^{k} x_i = \left(\mathbf{x}, \mathbf{e}\right) = 1, \quad \mathbf{e} \equiv \left(1, 1, \ldots 1\right)^T,$$

как части гиперплоскости над полуосями неотрицательных координат, к координатам $\{y_i \equiv \sqrt{x_i}\}$ геометрически означает преобразование пространства фамильных состояний (симплекс) в часть гиперсферы с единичным радиусом (см. рис. 1 в трехмерном случае). При этом матрица ковариаций для выборочных отклонений $\boldsymbol{\delta}\mathbf{y}$ новых переменных приближенно равна $\mathbf{V}(\mathbf{y})$ со следующими легко проверяемыми свойствами:

$$\mathbf{V}(\mathbf{y}) = \frac{1}{2N}\mathbf{W}(\mathbf{y}), \ \mathbf{W}(\mathbf{y}) \equiv \mathbf{I} - \mathbf{y}\mathbf{y}^T = \left[\delta_{ij} - y_i y_j\right],$$

$$\mathbf{y}^T\mathbf{y} = 1; \quad \mathbf{W}(\mathbf{y})\mathbf{y} = 0, \quad \mathbf{W}(\mathbf{y})\mathbf{v} = \mathbf{v}, \ \mathbf{v}:(\mathbf{y}, \mathbf{v}) = 0.$$

Таким образом, вектор $\mathbf{y}$ является собственным вектором матриц $\mathbf{V}(\mathbf{y})$ и $\mathbf{W}(\mathbf{y})$ с собственным числом $\lambda = 0$, остальные собственные векторы $\mathbf{W}(\mathbf{y})$ ортогональны $\mathbf{y}$ с равными единице собственными числами. Очевидно, $\mathbf{y}$ является вектором единичной нормали к гиперсфере (3).

Перейдем к новой системе координат, в которой на одной из осей лежит вектор нормали, а остальные оси располагаются в касательной плоскости в точке $\mathbf{y}$, образуя ортонормированную систему. Данное преобразование с ортонормированной матрицей перехода не изменяет распределение выборочных отклонений и собственных чисел матрицы ковариаций. В новой системе координат она является диагональной, ее главная диагональ состоит из дисперсий по новым координатам (из собственных чисел λ). Дисперсия (собственное число) по нормали равна нулю, т. е. выборочные отклонения по

направлению вектора нормали невозможны. Это означает, что вся выборочная изменчивость сконцентрирована на гиперсфере. Будем аппроксимировать ее в касательной плоскости. Повторим, что равенство $\mathbf{Wv} = \mathbf{v}$ означает, что любой вектор $\mathbf{v}$, ортогональный $\mathbf{y}$ (лежащий в касательной плоскости к гиперсфере в точке $\mathbf{y}$), будет собственным для матрицы $\mathbf{W}$ с $\lambda = 1$ [14]. Поэтому на касательной плоскости дисперсии выборочных отклонений одинаковы по любому направлению (изотропность). Займемся изучением свойств θs в ряду поколений.

АППРОКСИМАЦИЯ УГЛОВОГО ОТКЛОНЕНИЯ ПО МНОЖЕСТВУ ФАМИЛИЙ ПРИ ОТНОСИТЕЛЬНО МАЛОМ КОЛИЧЕСТВЕ ПОКОЛЕНИЙ

Большинство из описанных свойств углового расстояния по множеству фамилий соответствуют свойствам выборки, которая с точки зрения случайного дрейфа рассматривается как характеристика только одного шага в цепи изменений популяции в поколениях. Иная точка зрения фокусируется по рекомендации Р. Фишера на свойствах динамики углового расстояния в ряду поколений в результате случайных выборочных колебаний состояния популяции. Она была широко популяризирована Л. Кавалли-Сфорца с соавт. [21, 12] в отношении генного дрейфа на относительно малом промежутке времени, не приводившими, однако, теоретических обоснований. В дальнейшем анализ данной ситуации был представлен в [14], изложенный также в [22].

Геометрические свойства процесса случайного дрейфа, аппроксимируемого диффузионным процессом на гиперсфере в римановом пространстве, и его асимптотика на небольших промежутках времени рассматривались в [23, 24]. Общий случай асимптотики диффузионных процессов в римановом пространстве на небольших временах рассмотрен в предположении невырожденной матрицы диффузии внутри и на границе фазового пространства в [25]. В монографии [26] проанализированы разносторонние информационно-геометрические свойства модели Райта – Фишера, одним из примеров которой является рассматриваемый случай фамильного дрейфа.

Приведенные выше результаты о свойствах углового расстояния для множества фамилий в контексте динамики фамильного состояния соответствуют однократной смене поколений, сопровождаемой случайным выборочным отклонением фамильного состава популяции от $\mathbf{y}_0$ до $\mathbf{y}(1)$. Конечно, наибольший интерес представляет динамика в течение не одного, а ряда поколений. Эта динамика описывается результатами последовательности вложенных выборок, соответствующими последовательности смены поколений популяции.

Ремарка 5. Проанализируем второе поколение под другим углом зрения. Рассмотрим его как гипотетический *ансамбль* популяций, состоящий из возможных вариантов популяций-потомков популяций первого поколения. Этот ансамбль можно интерпретировать как иерархически подразделенную метапопуляцию. Она разбивается на группы, происходящие от разных популяций первого поколения. В силу ненаправленного характера случайного дрейфа ожидаемые концентрации фамилий в каждой отдельной группе популяций совпадают с концентрациями у породившей группу популяции первого поколения. Поэтому межгрупповая дисперсия распределения концентрации отдельной фамилии *по группам* второго поколения такая же, как дисперсия распределения ее концентраций *по популяциям* первого поколения. Аналогично *межгрупповая матрица ковариаций* $\mathbf{Vs}_{betw}(\mathbf{y}(2)|\mathbf{y}(0))$ совпадает с матрицей ковариаций распределения концентраций по популяциям первого поколения, равной согласно (4) значению $\frac{1}{2Ne(1)}\mathbf{W}(\mathbf{y}_0)$:

$$\mathbf{Vs}_{betw}\left(\mathbf{y}(2)\mid\mathbf{y}(0)\right)=\mathbf{Vs}\left(\mathbf{y}(1)\mid\mathbf{y}(0)\right)=\frac{1}{2Ne(1)}\mathbf{W}\left(\mathbf{y}_0\right).$$

В каждой отдельной группе популяций, происходящих от некоторой популяции первого поколения с концентрациями фамилий $\mathbf{y}(1)$, матрица ковариаций согласно (4) равна $\frac{1}{2Ne(2)}\mathbf{W}(\mathbf{y}_1))$ и рассматривается как *внутригрупповая матрица ковариаций*.

Итак, все популяции второго поколения образуют метапопуляцию, подразделенную на группы (см. рис. 2). Дивергенция групп между собой по концентрации отдельной фамилии характеризуется *межгрупповой дисперсией* распределения ее концентраций по группам. Дивергенция популяций внутри групп метапопуляции характеризуется *средней внутригрупповой дисперсией* распределения концентраций фамилии по популяциям внутри групп.

Дивергенция популяций всего ансамбля характеризуется *общей (полной) дисперсией* распределения концентраций фамилий по всем популяциям второго уровня. По *правилу сложения дисперсий* (см., например, его применение к фамильной структуре в [27]) полная дисперсия равна сумме межгрупповой и средней внутригрупповой дисперсий. Это правило остается верным в случае не одной, а множества фамилий, если слово “дисперсия” заменить на “матрица ковариаций”.

Рассмотрим динамику матрицы ковариаций $\mathbf{Vs}(\mathbf{y}(t))$. Начнем изучение со случая двух поколений случайного дрейфа.

Результат 6. *Пусть рассматривается последовательность двух независимых вложенных случайных*

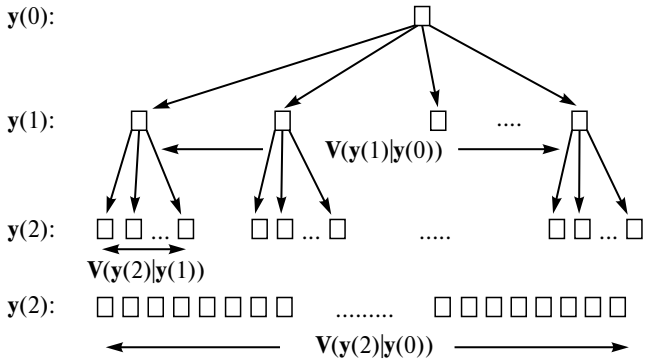

Рис. 2. Межгрупповая, внутригрупповые и полная матрицы ковариаций концентраций состояния популяций:
□ – обозначение популяции; **y**(0) – состояние родоначальной популяции; **y**(1) – случайные состояния ее потомков, популяций первого поколения, разброс возможных состояний **y**(1) характеризуется матрицей ковариаций **V**(**y**(1)|**y**(0)), служащей межгрупповой матрицей ковариаций для популяций следующего поколения, где **y**(0) фиксировано; **y**(2) – случайные состояния популяций, потомков родоначальной популяции во втором поколении. Они образуют метапопуляцию, состоящую из групп с происхождением от отдельных популяций первого поколения и с внутригрупповыми матрицами ковариаций **V**(**y**(2)|**y**(1)). Здесь **y**(1) случайно варьирует между группами; следующая строка относится ко второму поколению без разбиения на группы его популяций, разброс которых характеризуется полной матрицей ковариаций **Vs**(**y**(2)). Стрелки, направленные сверху вниз, соединяют родительскую популяцию с популяцией потомков.

выборок с возвращением из популяции с k вариантами фамилий, где их концентрации превышают надлежащий порог. Положим, что размер такой выборки (эффективный размер популяции) в поколении $\tau = 1, 2$ *равен* $\frac{Ne(\tau)}{2}$, *а вероятность попадания в выборку фамилии i-го типа равна* $x_i(\tau) > 0, \sum_{i=1}^{k} x_i(\tau) = 1,$ *где* x_i *– ее концентрация в родительской популяции.*

Тогда во втором поколении матрица ковариаций **Vs**(**y**(2)) *случайного вектора* $\mathbf{y}(2) \equiv \left\{\mathbf{y}_i(2) \equiv \sqrt{x_i(2)}\right\}$ *приближенно выражается как*

$$\mathbf{Vs}\left(\mathbf{y}\left(2\right)\right) \approx \left(\frac{1}{2Ne(1)} + \frac{1}{2Ne(2)}\right)\left(\mathbf{I} - \mathbf{y}\left(0\right)\mathbf{y}^T\left(0\right)\right) \equiv$$
$$\equiv \left(\frac{1}{2Ne(1)} + \frac{1}{2Ne(2)}\right)\mathbf{W}\left(\mathbf{y}_0\right).$$

Доказательство. Учтем, что при смене поколений к фамильному состоянию популяции добавляется случайное выборочное отклонение **δ**, причем $E\{\boldsymbol{\delta}\} = \mathbf{0}$ независимо от номера поколения. Для первого поколения имеем

$$E\left\{\mathbf{y}\left(1\right)\middle|\mathbf{y}\left(0\right)\right\} = E\left\{\mathbf{y}\left(0\right) + \boldsymbol{\delta}\left(1\right)\middle|\mathbf{y}\left(0\right)\right\} = \mathbf{y}\left(0\right),$$
$$E\left\{\mathbf{y}\left(2\right)\middle|\mathbf{y}\left(1\right)\right\} = E\left\{\mathbf{y}\left(1\right) + \boldsymbol{\delta}\left(2\right)\middle|\mathbf{y}\left(1\right)\right\} = \mathbf{y}\left(1\right),$$
$$\mathbf{Vs}\left(\mathbf{y}\left(1\right) \mid \mathbf{y}\left(0\right)\right) = \mathbf{Vs}\left(\boldsymbol{\delta}\left(1\right)\right) \mid \mathbf{y}\left(0\right)\right) =$$
$$= \frac{1}{2Ne(1)}\left(\mathbf{I} - \mathbf{y}\left(0\right)\mathbf{y}^T\left(0\right)\right) \equiv \frac{1}{2Ne(1)}\mathbf{W}\left(\mathbf{y}_0\right).$$

Отсюда очевидно, что матрица ковариаций во втором поколении при условии **y**(1) равна $\frac{1}{2Ne(2)}\mathbf{W}(\mathbf{y}_1)$.

Рассмотрим ансамбль популяций второго поколения, интерпретируемый как иерархически подразделенная метапопуляция, состоящая из групп. Каждая группа порождается соответствующей популяцией первого поколения и содержит возможные варианты популяций-потомков с разными фамильными состояниями. В нашем случае межгрупповая матрица ковариаций согласно предыдущей ремарке имеет вид

$$\mathbf{Vs}_{betw}(\mathbf{y}(2)|\mathbf{y}(0)) = \mathbf{Vs}(\mathbf{y}(1)|\mathbf{y}(0)) = \frac{1}{2Ne(1)}\mathbf{W}(\mathbf{y}_0).$$

Дисперсия распределения концентрации отдельной фамилии по популяциям внутри какой-либо группы, порождаемой популяцией первого поколения с фамильным состоянием **y**(1) (при условии **y**(1)), является внутригрупповой дисперсией (Vs_{in}) для этой группы, а при рассмотрении множества фамилий вместо дисперсии имеем матрицу ковариаций $\mathbf{Vs}_{in}(\mathbf{y}(2))|\mathbf{y}(1))$ концентраций фамилий, являющуюся внутригрупповой матрицей ковариаций. Согласно *правилу сложения дисперсий (ковариаций)* полные (для всей метапопуляции-ансамбля) значения данных характеристик изменчивости равны сумме межгруппового и *среднего* внутригруппового значений этих показателей, т. е. полная матрица ковариации $\mathbf{Vs}_{tot}(\mathbf{y}(\mathbf{2}))$ концентраций фамилий популяций второго поколения имеет вид

$$\mathbf{Vs}_{tot}\left(\mathbf{y}\left(2\right) \mid \mathbf{y}\left(0\right)\right) = \mathbf{Vs}_{betw}\left(\mathbf{y}\left(2\right) \mid \mathbf{y}\left(0\right)\right) +$$
$$+ E_{\mathbf{y}(1)}\{\mathbf{Vs}_{in}\left(\mathbf{y}\left(2\right) \mid \mathbf{y}\left(1\right)\right)\} =$$
$$= \mathbf{Vs}\left(\mathbf{y}\left(1\right) \mid \mathbf{y}\left(0\right)\right) + E_{\mathbf{y}(1)}\left\{\mathbf{Vs}\left(\mathbf{y}\left(2\right) \mid \mathbf{y}\left(1\right)\right)\right\} =$$
$$= \frac{1}{2Ne(1)}\mathbf{W}\left(\mathbf{y}_0\right) + E_{\mathbf{y}(1)}\left\{\frac{1}{2Ne(2)}\mathbf{W}\left(\mathbf{y}_1\right) \mid \mathbf{y}\left(0\right)\right)\right\}.$$

Рассмотрим второе слагаемое. Вспомним, что для любой случайной величины *x*

$$V(x) \equiv E\left\{\left(x - E\{x\}\right)^2\right\} = E\left\{x^2\right\} - \left(E\{x\}\right)^2,$$

$$E\left\{x^2\right\} = \left(E\{x\}\right)^2 + V(x).$$

Аналогично для любого случайного вектора **y** имеем

$$\mathbf{V}(\mathbf{y}) \equiv E\left\{\left((\mathbf{y} - E\{\mathbf{y}\})(\mathbf{y} - E\{\mathbf{y}\})^T\right)\right\} = E\{\mathbf{y}\mathbf{y}^T\} - E\{\mathbf{y}\}E\{\mathbf{y}^T\},$$

$$E\{\mathbf{y}\mathbf{y}^T\} = \mathbf{V}(\mathbf{y}) + E\{\mathbf{y}\}E\{\mathbf{y}^T\}, \quad (7)$$

где **V**(**y**) – матрица ковариаций вектора **y**.

Таким образом, у нас средняя внутригрупповая матрица ковариации имеет вид

$$E_{\mathbf{y}(1)}\left\{\mathbf{Vs}_{in}\left(\mathbf{y}(2)\right) \mid \mathbf{y}(1)\right\} = E_{\mathbf{y}(1)}\left\{\frac{1}{2Ne(2)}\mathbf{W}\left(\mathbf{y}(1)\right) \mid \mathbf{y}(0)\right\} \equiv$$

$$\equiv E_{\mathbf{y}(1)}\left\{\frac{1}{2Ne(2)}\left(\mathbf{I} - \mathbf{y}(1)\mathbf{y}^T(1)\right) \mid \mathbf{y}(0)\right\} =$$

$$= \frac{1}{2Ne(2)}\mathbf{I} - \frac{1}{2Ne(2)}E_{\mathbf{y}(1)}\left\{\mathbf{y}(1)\mathbf{y}^T(1)\right) \mid \mathbf{y}(0))\right\}$$

Подставим сюда $E_{\mathbf{y}(1)}\{\mathbf{y}(1)\mathbf{y}^T(1)\}$ согласно (7) и продолжим равенства

$$= \frac{1}{2Ne(2)}\mathbf{I} - \frac{1}{2Ne(2)}\mathbf{V}\left(\mathbf{y}(1) \mid \mathbf{y}(0)\right) + E\left\{\mathbf{y}(1)\right\}E\left\{\mathbf{y}^T(1)\right\} =$$

$$= \frac{1}{2Ne(2)}\mathbf{I} - \frac{1}{2Ne(2)}\left(\frac{1}{2Ne(1)}\mathbf{W}\left(\mathbf{y}(0)\right) + \mathbf{y}(0)\mathbf{y}^T(0)\right) \approx$$

$$\approx \frac{1}{2Ne(2)}\left(\mathbf{I} - \mathbf{y}(0)\mathbf{y}^T(0)\right) \equiv \frac{1}{2Ne(2)}\mathbf{W}\left(\mathbf{y}_0\right).$$

Здесь мы пренебрегли членом малой величины, содержащим произведение $\frac{1}{Ne(1)Ne(2)}$.

В итоге получаем

$$\mathrm{Vs}\left(\mathbf{y}(2)|\mathbf{y}(0)\right) = \frac{1}{2Ne(1)}\mathbf{W}(\mathbf{y}_0) + \mathrm{E}_{\mathrm{y}(1)}\left\{\frac{1}{2Ne(2)}\mathbf{W}(\mathbf{y}_1)|\mathbf{y}(0))\right\}$$

$$\approx \left(\frac{1}{2Ne(1)} + \frac{1}{2Ne(2)}\right)\mathrm{W}(\mathrm{y}_0).$$

◄

Напомним, что при относительно небольшой по сравнению с размером популяции длине *t* для последовательности поколений процесс динамики под влиянием многих недоминирующих по давлению факторов хорошо аппроксимируется процессом дрейфа, поскольку дивергенция из-за фактора случайного дрейфа будет порядка $\sqrt{t}$, а из-за возможного давления систематических факторов порядка *t* и $\sqrt{t} >> t$ при малых *t* согласно [14, 15]. Поэтому дальше мы ограничимся именно такой аппроксимацией. Так как речь идет о произвольной совокупности дискретных объектов (в частности фамилий), то нижеследующий результат сформулируем для произвольной совокупности (в случае фамилий в нем под **V** подразумевается **Vs**). Дадим модификацию обоснования в [14] динамики матрицы ковариаций в этом случае.

Результат 7. *Пусть рассматриваются последовательности независимых вложенных случайных выборок с возвращением из совокупности с k типами объектов (в нашем случае фамилий из популяции) и с такой группировкой объектов, когда минимальная концентрация среди групп превышает надлежащий порог. Положим, что размер выборки (эффективный размер популяции) на шаге* τ *(в поколении* τ*) равен* $\frac{Ne(\tau)}{2}$*, а вероятность попадания в выборку объекта i-го типа равна* $x_i(\tau) > 0$, $\sum_{i=1}^{k} x_i(\tau) = 1$, *где* x_i *– доля (концентрация) группы объектов i-го типа в "родительской" совокупности. Тогда на шаге t (в поколении t) матрица ковариации* **V**(**y**(*t*)) *случайного вектора* **y**(*t*) ≡ $\{y_i(t) \equiv \sqrt{x_i(t)}\}$ *при относительно малом t (при* $\frac{2\widetilde{N}e(t)}{t} >> 1$, *где* $\widetilde{N}e(t)$ *– среднее гармоническое значение для рассматриваемого ряда эффективных размеров популяции* $\{Ne(\tau), \tau = 1, 2, ..., t\}$*) приближенно выражается как*

$$\mathbf{V}(\mathbf{y}(t)) = \left(\sum_{\tau=1}^{t} c(\tau)\right)\left(\mathbf{I} - \mathbf{y}(0)\mathbf{y}^T(0)\right) =$$

$$\frac{t}{2\widetilde{N}e(t)}\left(\mathbf{I} - \mathbf{y}(0)\mathbf{y}^T(0)\right) = \frac{t}{2\widetilde{N}e(t)}\mathbf{W}\left(\mathbf{y}_0\right), \quad (8)$$

$$c(\tau) \equiv \frac{1}{2Ne(\tau)},\ \widetilde{N}e(t) \equiv t \Big/ \sum_{\tau=1}^{t}\frac{1}{Ne(\tau)}.$$

Доказательство проведем по индукции. Мы подметили закономерность динамики **V**(**y**(*t*)) в двух первых поколениях, согласно которой существует *t* (*t* = 2), когда (8) выполняется для любой последовательности длиной не больше *t*. Покажем, что то же самое верно и при *t* + 1. Для этого рассмотрим ансамбль популяций в поколении *t* + 1 как метапопуляцию, подразделенную на группы. *Каждая группа порождается соответствующей популяцией первого поколения*, т. е. состоит из возможных вариантов ее популяций-потомков в поколении *t* + 1 с разными фамильными состояниями.

Как и при анализе двух поколений, ожидаемые концентрации фамилий в группах совпадают с фамильным состоянием порождающей популяции, принимающим значения {**y**(1)}, а *дисперсия Vs*(y(1)) *распределения концентраций отдельной фамилии по*

популяциям первого поколения является межгрупповой дисперсией Vs_{betw} для ансамбля на любом другом из последующих поколений. Аналогично матрица ковариаций для распределения по популяциям первого поколения концентраций множества фамилий является межгрупповой матрицей ковариаций, которая в нашем случае имеет согласно (4) вид $\mathbf{Vs}_{betw} = \mathbf{Vs}(\mathbf{y}(1)|\mathbf{y}(0)) = \frac{1}{2Ne(1)}\mathbf{W}(\mathbf{y}_0)$.

Для *отдельной* группы в поколении $t+1$, происходящей от популяции с фамильным состоянием $\mathbf{y}(1)$, матрица ковариации $\mathbf{Vs}(\mathbf{y}(t+1)|\mathbf{y}(1))$ является *полной*. По предположению индукции доказываемая формула верна в случае до t поколений, отделяющих ансамбль от порождающей популяции. Она верна как для $\mathbf{Vs}_{tot}(\mathbf{y}(t)|\mathbf{y}(0))$, так и для $\mathbf{Vs}_{tot}(\mathbf{y}(t+1)|\mathbf{y}(1))$, поскольку для них число поколений, отделяющих $\mathbf{y}(t)$ от $\mathbf{y}(0)$ и $\mathbf{y}(t+1)$ от $\mathbf{y}(1)$, одно и то же (равно t), т. е. матрица $\mathbf{Vs}_{tot}(\mathbf{y}(t+1)|\mathbf{y}(1))$ находится согласно предположению индукции по формуле (8). В метапопуляции она представляет собой одну из *внутригрупповых матриц*. По правилу сложения ковариаций и с учетом предположения индукции применительно к матрице $\mathbf{Vs}_{tot}(\mathbf{y}(t+1)|\mathbf{y}(1))$, рассматриваемой как внутригрупповая, получаем

$$\mathbf{Vs}_{tot}\left(\mathbf{y}(t+1)\mid\mathbf{y}(0)\right)=\mathbf{Vs}_{betw}\left(\mathbf{y}(1)\mid\mathbf{y}(0)\right)+$$

$$+E_{\mathbf{y}(1)}\left\{\mathbf{Vs}_{in}\left(\mathbf{y}(t+1)\right)\mid\mathbf{y}(1)\right\}=$$

$$=\frac{1}{2Ne(1)}\mathbf{W}(\mathbf{y}_0)+\left(\sum_{\tau=2}^{t+1}c(\tau)\right)E_{\mathbf{y}(1)}\left\{\mathbf{W}\left(\mathbf{y}(1)\right)\mid\mathbf{y}(0)\right\},$$

$$E_{\mathbf{y}(1)}\left\{\mathbf{W}\left(\mathbf{y}(1)\right)\mid\mathbf{y}(0)\right\}\equiv E_{\mathbf{y}(1)}\left\{\left(\mathbf{I}-\mathbf{y}(1)\mathbf{y}^T(1)\right)\mid\mathbf{y}(0)\right)\right\}=$$

$$=\mathbf{I}-E_{\mathbf{y}(1)}\left\{\left(\mathbf{y}(1)\mathbf{y}^T(1)\right)\mid\mathbf{y}(0)\right)\right\}.$$

Подставим сюда $E_{\mathbf{y}(1)}\{\mathbf{y}(1)\mathbf{y}^T(1)\}$ согласно (7) и продолжим равенства

$$E_{\mathbf{y}(1)}\left\{\mathbf{W}\left(\mathbf{y}(1)\right)\middle|\mathbf{y}(0)\right\}=\mathbf{I}-\left(\mathbf{V}\left(\mathbf{y}(1)\middle|\mathbf{y}(0)\right)+E\left\{\mathbf{y}(1)\right\}E\left\{\mathbf{y}^T(1)\right\}\right)$$

$$=\mathbf{I}-\left(c(1)\mathbf{W}(\mathbf{y}_0)+\mathbf{y}(0)\mathbf{y}^T(0)\right)=\left(1-c(1)\right)\mathbf{W}(\mathbf{y}_0).$$

Продлим преобразование $\mathbf{Vs}_{tot}(\mathbf{y}(t+1)|\mathbf{y}(0))$ подстановкой найденного значения $E_{\mathbf{y}(1)}\left\{\mathbf{W}\left(\mathbf{y}(1)\right)\mid\mathbf{y}(0)\right\}$:

$$\mathbf{Vs}_{tot}\left(\mathbf{y}(t+1)\mid\mathbf{y}(0)\right)=c(1)\,\mathbf{W}(\mathbf{y}_0))+\left(\sum_{\tau=2}^{t+1}c(\tau)\right)\left(1-c(1)\right)\mathbf{W}(\mathbf{y}_0)\approx$$

$$\approx\left(\sum_{\tau=1}^{t+1}c(\tau))\right)\mathbf{W}(\mathbf{y}_0)=\frac{t}{2\widetilde{N}e(t)}\mathbf{W}(\mathbf{y}_0),\quad c(\tau)\equiv\frac{t}{2Ne(\tau)}.$$

Здесь мы пренебрегли слагаемыми, содержащими произведение $c(\tau)c(1)=\frac{1}{2Ne(\tau)}\,\frac{1}{2Ne(1)}$, и учли, что по определению $\widetilde{N}e(t)\equiv t/\sum_{\tau=1}^{t}\frac{1}{Ne(\tau)}$ как обратная величина к среднему арифметическому для $\left\{\frac{1}{Ne(\tau)}\right\}$.

Таким образом, выполняется переход индукции: если $\mathbf{V}(\mathbf{y}((t))$ зависит от t по предлагаемой формуле, то она верна при $t+1$ и, значит, при любом (относительно малом) t. ◄

Следствие 8. *В условиях предыдущего результата при* $\frac{2\widetilde{N}e(t)}{t}\to\infty$, *где* t *– количество поколений, и при такой группировке фамилий, когда минимальная концентрация среди групп превышает надлежащий порог, асимптотическое распределение фамильного состояния* $\mathbf{y}(t)$ *является многомерным нормальным с математическим ожиданием* $\mathbf{y}(0)$ *и матрицей ковариации, пропорциональной идемпотентной матрице*

$$\mathbf{W}(\mathbf{y}_0)\equiv\mathbf{I}-\mathbf{y}(0)\mathbf{y}^T(0):$$

$$\mathbf{y}(t)=N\left(y(0),\frac{t}{2\widetilde{N}e(t)}\left(\mathbf{I}-\mathbf{y}(0)\mathbf{y}^T(0)\right)\right),\ \left(\mathbf{I}-\mathbf{y}(0)\mathbf{y}^T(0)\right)^2=$$

$$=\left(\mathbf{I}-\mathbf{y}(0)\mathbf{y}^T(0)\right)^2=\mathbf{I}-\mathbf{y}(0)\mathbf{y}^T(0).$$

Распределение квадрата Евклидова расстояния $\Delta(\mathbf{y}(t))$ *между* $\mathbf{y}(t)$ *и* $\mathbf{y}(0)$, *т. е. величины* $\Delta^2(\mathbf{y}(t))\equiv$ $\equiv|\mathbf{y}(t)-\mathbf{y}(0)|^2\equiv\sum_{i=1}^{k}(y_i(t)-y_i(0))^2$, *удовлетворяет*

$$\frac{2\widetilde{N}e(t)}{t}\Delta^2(\mathbf{y}(t))\equiv\frac{2\widetilde{N}e(t)}{t}\left|\mathbf{y}(t)-\mathbf{y}(0)\right|^2=\chi^2_{k1},\frac{2\widetilde{N}e(t)}{t}\gg1.$$

Для углового расстояния $\theta s(\mathbf{y}(t))$ *между текущим фамильным состоянием* $\mathbf{x}(t)$ *и начальным* $\mathbf{p}$ (*между* $\mathbf{y}(t)$ *и* $\mathbf{y}_0$ *на гиперсфере*), *определяемого согласно* (1) *как*:

$$\theta s(t)=\theta s\left(\mathbf{x}(t),\,\mathbf{p}\right)=\theta s\left(\mathbf{y}(t),\mathbf{y}_0\right)\equiv\arccos\left(\sum_{i=1}^{k}\sqrt{x_i(t)}\sqrt{p_i}\right),$$

асимптотически выполняется

$$\frac{2\widetilde{N}e(t)}{t}\theta s^2\left(\mathbf{x}(t),\mathbf{p}\right)=\chi^2_{k-1}$$

Здесь $\widetilde{N}e(t)$ *и* χ^2_{k-1} *обозначают среднюю гармоническую численность популяции для ряда* $\{Ne(\tau)\}$ *длиной* t *поколений и распределение хи-квадрат с* $k-1$ *степенями свободы соответственно.*

Доказательство вытекает из центральной предельной теоремы, приближенного равенства Евклидова и углового расстояний в предположениях

следствия и ненаправленного характера случайного дрейфа. ◄

ОБСУЖДЕНИЕ

Интерес автора к рассматриваемым проблемам мотивирован наблюдениями малых деревень России, где резко преобладала одна или несколько фамилий. Такая ситуация описывалась не раз в художественной литературе упоминаниями типа “у нас в деревне все Смирновы”. Картина, когда большинство жителей деревни оказываются однофамильцами, поражает. Например, при изучении автором популяций европейского севера России [28] встретилась деревня с 91% однофамильцев при общем количестве жителей 126 человек. Для городского жителя такая ситуация парадоксальна, и возникает желание дать теоретическое объяснение наблюдаемым различиям между фамильным и генетическим разнообразием. Поэтому настоящая статья фокусируется на теоретических аспектах анализа фамильной структуры.

Результаты анализа фамильной структуры популяции важны также своими параллелями с анализом генетической структуры в силу сходства патрилинейной передачи фамилии и генетической информации потомкам. В популяциях конечного размера такая передача сопровождается случайными флуктуациями как концентраций фамилий, так и аллелей. Напомним используемые нами подходы к изучению этой ситуации. При анализе флуктуаций мы используем предположение о неперекрывании поколений, т. е. пренебрегаем существованием возрастной структуры у человека. Это предположение можно рассматривать как аппроксимацию реальной ситуации, часто применяемую в популяционной генетике, например при использовании закона Харди – Вайнберга.

При моделировании флуктуаций частот аллелей аутосомного локуса нередко используется модель Райта – Фишера, которую можно сформулировать в виде процесса вложенных выборок, формирующих состав нового поколения как случайную выборку аллелей поколения родителей. Та же самая картина получается при рассмотрении мужского компонента популяции, фамилии которого наследуются по поколениям при их патрилинейной передаче от отца к сыну. Данный процесс называем процессом случайного дрейфа (генов и фамилий одновременно в одной и той же популяции). Понятно, что качественные свойства случайного генного дрейфа и дрейфа фамилий одинаковы, но количественно различаются. Суть в том, что при генном дрейфе новое поколение с численностью N формируется как случайная выборка с возвращением $2N$ гамет из пула родительских гамет, а состав мужского компонента нового поколения как выборка $N/2$ фамилий отцов. Размер выборки определяет интенсивность флуктуаций, которые больше в 4 раза для фамилий.

Понятно, что сходство процессов генного дрейфа и дрейфа фамилий означает возможность использования для анализа фамильного дрейфа методов, наработанных в популяционной генетике в течение длительного времени. В данной статье внимание концентрируется на стабилизации темпа дивергенции фамильного состояния популяции (вектора концентраций фамилий) от начального положения. Решение этой задачи опирается на известное нелинейное преобразование $y_i = \sqrt{x_i}$ концентраций фамилий x_i (координат фамильного состояния) и анализ углового расстояния θs между состояниями. Приведены модификации соответствующих популяционно-генетических подходов, расширены и углублены обоснования аппроксимации свойств случайного дрейфа на относительно небольших промежутках времени в поколениях.

В одной и той же популяции случайный дрейф приводит к фамильной дивергенции, вчетверо превосходящей дивергенцию генетическую на относительно небольшом промежутке времени в поколениях. Этот вывод характеризует с другой точки зрения результат о четырехкратном различии между стандартным коэффициентом инбридинга и его фамильным аналогом, полученный автором в [15] при анализе фамильного дрейфа в терминах концентраций фамилий $\{x_i\}$. Преимуществом изложенного подхода к анализу в пространстве $\{y_i = \sqrt{x_i}\}$ является независимость дивергенции от начального состояния и постоянный темп ее увеличения за поколение в случае неизменяющегося эффективного размера популяции. Кроме того, *дивергенция, отражаемая средним квадратом углового расстояния от начального положения, обратно пропорциональна среднему гармоническому эффективному размеру популяции и прямо пропорциональна количеству поколений дивергенции.* Данный результат важен для сравнения популяций и их систем и для решения микротаксономических задач.

В заключение напомним основные условия корректности этих результатов на рассматриваемом промежутке времени:

отклонениями от патрилинейной передачи фамилии можно пренебречь;

размеры популяции должны быть достаточно велики, чтобы выборочные изменения фамильного состояния аппроксимировались нормальным распределением;

на рассматриваемом промежутке времени концентрации фамилий не должны быть слишком малыми во избежание нарушения свойств преобразования $y_i = \sqrt{x_i}$;

сам рассматриваемый промежуток времени в поколениях должен быть небольшим по сравнению

со средним гармоническим эффективным размером популяции на нем.

Настоящая статья не содержит каких-либо исследований с использованием в качестве объекта животных.

Настоящая статья не содержит каких-либо исследований с участием в качестве объекта людей.

СПИСОК ЛИТЕРАТУРЫ

1. *Lasker G.W.* Surnames and Genetic Structure. CambridgeUniv. Press, 2005. 148 p.
2. *King T.E., Jobling M.A.* What's in a name Y chromosomes, surnames and the genetic genealogy revolution // Trends in Genetics. 2009. V. 25. Iss. 8. P. 351–360.
3. *Jobling M.A.* In the name of the father- surnames and genetics // Trends in Genetics. 2001. V. 17. № 6. P. 353–357.
4. *Балановская Е.В., Балановский О.П.* Русский генофонд на Русской равнине. М.: Луч, 2007. 415 с.
5. *Сорокина И.Н., Чурносов М.И., Балтуцкая И.В. и др.* Антропогенетическое изучение населения Центральной России. М.: Изд-во РАМН, 2014. 336 с.
6. *Colantonio S.E., Lasker G.W., Kaplan B.A., Fuster V.* Use of surname models in human population biology: A review of recent developments // Human Biology. 2003. V. 75. № 6. P. 785–807.
7. *Crow J.F., Mange A.P.* Measurement of inbreeding from the frequency of marriages between persons of the same surname // Social Biology. 1982. V. 29. № 1/2. P. 101–105.
8. *Crow J.F.* The estimation of inbreeding from isonymy // Human Biology. 1980. V. 52. № 1. P. 1–14.
9. *Crow J.F.* The estimation of inbreeding from isonymy (reprint) with an update // Human Biology. 1989. V. 61. № 5/6. Special issue on foundations of anthropological genetics. P. 935–948.
10. *Rogers A.R.* Doubts about isonymy // Human Biology. 1991. V. 63. № 5. P. 663–668.
11. *Ли Ч.* Введение в популяционную генетику. М.: Мир, 1978. 555 с. (*Li C.C.* First course in population genetics. California: Boxwood Press Pacific Grove, 1976).
12. *Кимура М.* Молекулярная эволюция: теория нейтральности. М.: Мир. 1985. 394 с. (*Kimura M.* The Neutral Theory of Molecular Evolution. Cambridge: Cambr. Univ. Press., 1983)
13. *Хедрик Ф.* Генетика популяций. М.: Техносфера. 2003. 592 с. (*Hedrick P.W.* Genetics of Populations. 2nded. Boston: Jones and Bartlett Publ., 2000. 553 pp.)
14. *Малютов М.Б., Пасеков В.П.* Об одной статистической задаче популяционной генетики // Теория вероятностей и ее применения. 1971. Т. 16. Вып. 3. С. 579–581. (*Ma1yutov M.B., Pasekov V.P.* On one statistical problem of population genetics // Theory of Probability and its Applications. 1971. Iss. V. 16. № 3. P. 559–566)
15. *Пасеков В.П.* К анализу случайных процессов изонимии. I. Структура изонимии // Генетика. 2021. Т. 57. № 10. С. 1194–1204. doi: 10.31857/S001667582110009X (*Passekov V.P.* To the Analysis of Random Processes of Isonymy: I. Isonymic Structure // Rus. J. Genet. 2021. V. 57. № 10, P. 1214–1222. doi: 10.1134/S1022795421100094)
16. *Fisher R.A.* On the dominance ratio // Proc. R. Soc. Edinb. 1922. V. 42. P. 321–341 (Bull. Math. Biol. 1990. V. 52. № 1–2. P. 297–318)
17. *Fisher R.A.* The Genetical Theory of Natural Selection. Oxford: Clarendon Press, 1930. 272 p.
18. *Bhattacharyya A.* On a measure of divergence between two multinomial populations // Sankhya. 1946. V. 7. Part 4. P. 401–406.
19. *Edwards A.W.F.* Distances between populations on the basis of gene frequencies // Biometrics. 1971. V. 27. № 4. P. 873–881.
20. *Вейр Б.* Анализ генетических данных: дискретные генетические признаки. М.: Мир, 1995. 400 с. (*Weir B.S.* Genetic data analysis: Methods for discrete population genetic data. Sunderland: Sinauer, 1990.)
21. *Cavalli-Sforza L.L., Edwards A.W.F.* Phylogenetic analysis. Models and estimation procedures // Am. J. Hum. Genet. 1967. V. 19. P. 233–257 (Evolution. 1967. V. 21. № 3. P. 550–570).
22. *Свирежев Ю.М., Пасеков В.П.* Основы математической генетики. М.: Наука, 1982. 511 с. (*Svirezhev Y.M., Passekov V.P.* Fundamentals of mathematical evolutionary genetics. Kluwer Acad. Publ., Dordrecht et al., 1990. 395 p.)
23. *Antonelli P.L., Strobeck C.* The geometry of random drift. I. Stochastic distance and diffusion // Adv. Appl. Probab. 1977. V. 9. № 2. P. 238–249.
24. *Papangelou F.* The large deviations of a multi-allele Wright–Fisher process mapped on the sphere // Ann. Appl. Prob. 2000. V. 10. № 4. P. 1259–1273.
25. *Молчанов С.А.* Диффузионные процессы и риманова геометрия // УМН. 1975. Т. 30. Вып. 1(181). С. 3–59. (*Molchanov S.A.* Diffusion processes and Riemannian geometry // Russ. Math. Surveys. 1975. V. 30. Iss. 1. P. 1–63)
26. *Hofrichter J., Jost J., Tran T.D.* Information geometry and population genetics: The mathematical structure of the Wright–Fisher model. Springer, 2017. 320 p.
27. *Пасеков В.П.* Описание дивергенции субпопуляций в иерархической системе при анализе изонимии. I. Дисперсия как показатель дивергенции // Генетика. 2022. Т. 58. № 6. С. 713–727 doi: 10.31857/S0016675822060054 (*Passekov V.P.* Description of Divergence of Subpopulations in the

Hierarchical System When Analyzing Isonymy: I. Variance as an Indicator of Divergence // Rus. J. Genet. 2022. V. 58, № 6. P. 736–750. doi: 10.1134/S1022795422060059)

28. *Пасеков В.П., Ревазов А.А.* К популяционной генетике населения европейского севера СССР. Сообщение I. Данные по структуре шести деревень Архангельской области // Генетика, Т. 11. № 7. 1975. С. 145–155.

On Stabilizing the Rate of Isonymy Divergence

V. P. Passekov[1, *]

[1] *Federal Research Center "Computer Science and Control", Russian Academy of Sciences, Moscow 119991 Russia*
**e-mail: pass40@mail.ru*

A theoretical analysis of the surname state of the population (the vector of namesake concentrations in the male component of the population) and its dynamics as a result of random surname drift is presented. An approximation of such a process by the Wright-Fisher model of a population with non-overlapping generations without selection pressure is used, i.e., an approximation by a sequence of nested random samples with the replacement from fathers' surnames in the population. The sample size is N/2 according to the size of the male component in the population of size N. In the same population, processes of random drift of both surnames and genes simultaneously occur. Their cardinal difference is that the sample size of surnames is four times smaller than the sample size of autosomal locus alleles. The analysis of random drift is simplified when moving from concentration coordinates to the square roots of them. As generations change, the state receives a sample deviation, measured by angular distance, and its mean square gives the rate of divergence, stabilizing in the new coordinates. An adaptation (in relation to the analysis of surname drift) of a known in population genetics result about the nature of divergence at a stage of a relatively small number of generations compared to the size of the population is given. The divergence of surnames occurs four times faster than the divergence of allele concentrations.

Keywords: random surname drift, divergence of surname and allele concentrations, isonymy, angular distances, stabilization of the divergence rate.