ГЕНЕТИКА, 2022, том 58, № 6, с. 713–727

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ

УДК 303.725.23:575.177:303.725.23

ОПИСАНИЕ ДИВЕРГЕНЦИИ СУБПОПУЛЯЦИЙ В ИЕРАРХИЧЕСКОЙ СИСТЕМЕ ПРИ АНАЛИЗЕ ИЗОНИМИИ. I. ДИСПЕРСИЯ КАК ПОКАЗАТЕЛЬ ДИВЕРГЕНЦИИ

© 2022 г. В. П. Пасеков*

Вычислительный центр им. А.А. Дородницына Федерального исследовательского центра "Информатика и управление" Российской академии наук, Москва, 119991 Россия

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

Поступила в редакцию 29.11.2021 г.
После доработки 27.12.2021 г.
Принята к публикации 28.12.2021 г.

Рассматриваются типичные для популяций человека метапопуляции с иерархической подразделенностью на части (субпопуляции), соответствующие классификации субпопуляций на основе административно-территориального деления (скажем, село, сельсовет, район, область и так далее); на основе генеалогического подхода, базирующегося на этногенезе; а также на других принципах биологической классификации. Каждый уровень иерархии представляет собой разбиение метапопуляции на непересекающиеся субпопуляции, суммарно составляющие всю ее и обладающие, в свою очередь, иерархической структурой подразделенности. Изучаются свойства изменчивости количественного признака субпопуляций при иерархической структуре на примере такого признака, как концентрация отдельной фамилии. Анализируется распределение концентрации фамилии по субпопуляциям, характеризуемое на каждом уровне своими средним значением и дисперсией, которая отражает фамильную дивергенцию субпопуляций на соответствующем уровне. Изучение фамильной дивергенции важно, так как она отражает при соответствующих предположениях генетическую дивергенцию и генетическую структуру метапопуляции. Показано, что каждому отдельному уровню иерархии соответствует неотрицательный вклад в полную (общую) дисперсию, равный среднему значению внутригрупповой дисперсии распределения концентрации фамилии по его субпопуляциям. Получено разложение общей дисперсии концентрации фамилии в метапопуляции по вкладам отдельных уровней, обобщающее правило сложения дисперсий. Найдена величина занижения общей дисперсии, когда вместо неподразделенных субпопуляций первого уровня иерархии (допустим, сел) в качестве единиц наблюдения служат субпопуляции более высокого уровня (скажем, районов). Это позволяет судить о степени занижения оценки генетической дивергенции в метапопуляции в результате игнорирования фамильной изменчивости на каком-либо из уровней иерархии. Все население разбивается на два компонента с иерархической структурой подразделенности: сельские и городские жители. Результаты данной работы в равной степени приложимы к каждому из них.

Ключевые слова: иерархическая структура популяций, метапопуляции, концентрации фамилии в субпопуляциях человека, разложение дисперсии концентрации по уровням иерархии.

DOI: 10.31857/S0016675822060054

Существование популяционной структуры (отличий от предположений модели элементарной идеальной популяции без каких-либо подразделений при панмиксии) накладывает свой отпечаток на фамильную структуру популяций. Группировка реальных данных с целью максимального приближения к идеальным конструкциям или из других соображений довольно условна, так как обычно не существует четких естественных границ у групп. Популяционная структура многообразна и не ограничивается случаем подразделения популяции на элементарные непересекающиеся группы. Например, в свою очередь, последние могут состоять из субпопуляций и т.д., образуя иерархическую структуру подразделенности.

Анализ межгрупповой и внутригрупповой изменчивости является интересной биологической проблемой и может пролить свет на особенности микроэволюционного процесса дивергенции популяций. Полное описание метапопуляции с иерархической структурой подразделенности включает данные по каждому уровню иерархии, скажем, средние значения признаков в субпопуляциях и дисперсии, отражая детально характер межгрупповой и внутригрупповой изменчивости с учетом

всех уровней. Чтобы выделить особенности дивергенции субпопуляций необходимо хорошо представлять себе базовые черты, свойственные самой по себе подразделенности метапопуляций, отвлекаясь от воздействия на формирование популяционной структуры прочих факторов. Начать такой анализ можно с изучения свойств произвольного разбиения абстрактной совокупности на части и свойств характерных разбиений природных популяций. Для последних типична иерархическая группировка данных на основе территориального расположения, генеалогической классификации и пр. в соответствии с правилами принятой биологической или иной иерархической классификации популяций.

В случае изучения популяций человека классификация и объединение данных часто производятся на основе административно-территориального деления, имеющего иерархический характер (скажем, село, сельсовет, район, область и др.), генеалогического подхода на основе этногенеза и пр. Получаемая группировка субпопуляций приближенно будет иерархической. Иерархическая структура метапопуляции отражается на ее свойствах, в частности на распределении фамилий в популяциях человека, где типична опора на официальные данные иерархического характера, их сбор и обработку. Настоящая статья мотивирована анализом фамильных данных с ориентацией на популяционную генетику. Использование фамилий для получения выводов о генетической структуре популяций основывается на существующих параллелях в передаче потомкам фамилий и аутосомных аллелей (см., например, [1, 2]). Плодотворность такого использования продемонстрирована в ряде работ [3] (изонимные браки), [4] (фундаментальная монография), в том числе в исследованиях популяций России [5, 6] (медико-генетические аспекты), [7, 8] (антропогенетическое изучение Центральной России), [9] (обширная библиография), см. также критические замечания в [10].

Очевидно, что отдельные или даже все достаточно крупные группы состоят из элементарных популяций, динамические и генетические процессы в которых исследованы теоретически более глубоко, и на полученных для них выводах базируется стандартная обработка материалов. Под элементарной популяцией мы понимаем такую, где более всего удовлетворяются предпосылки, закладываемые в классическую популяционно-генетическую модель Райта–Фишера (см., например, на русском [11]). Среди них основными являются требования панмиксии и равноценности потенциального вклада индивидуумов в следующее поколение. Скорее всего модели элементарной популяции приближенно соответствует субпопуляция самого нижнего первого уровня – село. Выбор элементарной популяции в качестве единицы наблюдения является крайне желательным, и выяснение роли отклонений от него представляет собой важную задачу.

Обоснование выбора единицы наблюдения может основываться на соображениях относительно уровня эндогамии [8]. Такой подход не лишен дискуссионных моментов. Попытаемся обсудить эту проблему. При иерархической структуре подразделенности популяции степень эндогамии может зависеть от уровня иерархии. Чем выше этот уровень, тем более может быть эндогамна популяция при прочих равных условиях. Такая картина увеличения степени эндогамии при движении от уровня сельсовета к району и области наблюдается в Центральной России [7]. Приведем абстрактную модельную иллюстрацию, когда такой характер зависимости имеет место.

Рассмотрим гипотетическую популяцию, равномерно и непрерывно распределенную на плоскости, и некоторый ограниченный ареал внутри данной области. Известно, что вероятность брака уменьшается при увеличении расстояния между местами рождения супругов (изоляция расстоянием). Для анализа того, как влияет на эндогамию этот фактор, предположим, что вероятностью брака при превышении некоторого расстояния между местами рождения можно пренебречь. Понятно, что при таком требовании относительно ограниченности расстояния между местами рождения супругов экзогенные браки могут быть лишь у лиц, родившихся достаточно близко к границе. В результате количество таких браков пропорционально длине границы. Соответственно эндогамные браки заключаются лицами внутри оставшейся части ареала и при достаточной его величине их количество примерно пропорционально площади ареала.

Увеличение ареала при его росте вдоль каждой из осей координат в k раз приводит к такому же удлинению границы в k раз, а площадь увеличится в k^2 раз. Это очевидно в случае прямоугольного и кругового ареалов, но верно и в общем случае. Тем самым отношение количества экзогамных браков к эндогамным приближенно пропорционально отношению $k/k^2 = 1/k$ и уменьшается с увеличением ареала. Таким образом уровень эндогамии при этом растет. При иерархической подразделенности популяции (например, территориальной) у субпопуляции более высокого уровня иерархии размер будет крупнее, чем у более низких по уровню, так как она включает в себя их численности и территории. Следовательно, уровень эндогамии должен повышаться с уровнем иерархии субпопуляций. Хотя предположения данного вывода далеки от реальности, повторим, что качественный характер связи эндогамности с уровнем иерархии подтверждается фактическими наблюдениями [7].

В результате увеличения ареала обитания наступает момент, когда найдутся пары, места рождения членов которых отделены таким расстоянием, что браки между ними практически невозможны. Тем самым кроме роста эндогамии происходит нарушение панмиксии с ее равноценностью образования любых родительских пар. Такая популяция не удовлетворяет предпосылкам модели элементарной популяции. Итак, повышение уровня иерархии ведет не только к увеличению уровня эндогамии, но и к нарушению панмиксии (в конце концов, все население земного шара эндогамно и панмиксия отсутствует).

Цель настоящей работы – анализ общих свойств распределения концентрации фамилии по субпопуляциям при их иерархической организации. Такие свойства являются чисто статистическими характеристиками иерархической структуры, а не особенностью конкретной популяционной системы. Анализируемые свойства относятся к любой иерархически подразделенной популяции и не выводятся из закономерностей той или иной модели микроэволюции популяций. Теоретически это может позволить выделить специфические свойства исследуемого материала.

Применительно к фамильным данным роль единицы наблюдения в иерархической структуре может играть популяция села или субпопуляции более высокого уровня иерархии. Каждая субпопуляция независимо от уровня иерархии характеризуется своими значениями признаков, одним из которых является концентрация фамилии (x). Отдельной субпопуляции соответствует единственное (среднее) значение признака, а его изменчивость (скажем, дисперсия) между субпопуляциями, отражающая их дивергенцию, изучается с разной степенью детализации в зависимости от выбора единицы наблюдения. Так, если такой единицей является индивидуум, а вместо индивидуальных данных используются села, характеризуемые средним весом жителей, то индивидуальная изменчивость остается как бы “за кадром”, и изучение изменчивости веса детализируется до дисперсии среднего веса. Когда структура подразделенности иерархическая, то чем меньше единица наблюдения и ее уровень иерархии, тем полнее охват изменчивости.

Тут возникает вопрос о том, насколько выводы теоретического анализа элементарных популяций приложимы к материалам, представленным в сгруппированном виде. Например, в данных о концентрациях фамилий в районах остается скрытой изменчивость фамильной структуры сел и малых городов, входящих в районы. Использование для анализа изонимии списков избирателей и данных телефонных справочников часто дают сведения о фамилиях в сгруппированном виде. Например, единицей наблюдения может быть район. Исследование того, в каком направлении и в какой степени скрытая изменчивость качественно и количественно сказывается на стандартных оценках коэффициента инбридинга и пр. (см., например, [1, 2]) по фамильным данным, представляет интерес.

Получаемые далее результаты относятся к свойствам иерархически подразделенных совокупностей, они не зависят от их природы и не основываются на предположениях о выборочном характере рассматриваемой иерархической совокупности с требованиями об идентичном и независимом распределении каких-либо данных или на использовании модели процесса формирования иерархической структуры.

Настоящая статья построена следующим образом. Сначала описывается система идентификации субпопуляций в метапопуляции с иерархической структурой подразделенности на субпопуляции. Затем в рамках этой системы рассматриваются соотношения между уровнями иерархии и математическими ожиданиями (средними значениями) и дисперсиями распределения концентрации фамилии как показателями дивергенции субпопуляций. Далее подробно рассматривается изменчивость субпопуляций в случае многоуровневой иерархической структуры метапопуляции и получено разложение общей (полной) дисперсии концентрации фамилий в субпопуляциях, выбранных в качестве единиц наблюдения, на соответствующие отдельным уровням компоненты. В результате получены выражения для степени занижения оценки дивергенции субпопуляций, когда не учитывается неоднородность (подразделенность) субпопуляций, служащих единицами наблюдения. В последнем разделе обсуждаются полученные результаты и указано, что в совместных данных по городскому и сельскому населению нарушается иерархическая структура подразделенности. Это вызвано тем, что город не является объединением непересекающихся сельских субпопуляций. В то же время подразделенность субпопуляций как сельского, так и городского компонентов разбиения всего населения имеет иерархический характер. Каждый из компонентов обладает найденными свойствами иерархических систем. В дальнейшей публикации предполагается проанализировать не дисперсию концентрации фамилии, а вероятность случайной встречи однофамильцев (ср. соответствующие подходы в [1, 2]) и ее связь с коэффициентом инбридинга в популяционной генетике.

Кратко коснемся обозначений и терминологии. Под концентрациями фамилий в популяции подразумеваются концентрации однофамильцев. Векторы набраны полужирным шрифтом, *к обозначениям фамильных аналогов популяционно-генетических характеристик добавлено окончание s.* Символ ◄ обозначет конец доказательства.

НУМЕРАЦИЯ СУБПОПУЛЯЦИЙ В ИЕРАРХИЧЕСКОЙ СИСТЕМЕ

Как уже говорилось, при исследовании популяций человека часто собирают данные, которые организуют в соответствии с соподчинением субпопуляций. У человека иерархическая структура присуща, напомним, территориальной, этнической (генеалогической) и лингвистической классификациям популяций. Чтобы легче ориентироваться в получаемой при этом картине, рассмотрим абстрактную совокупность любых объектов одной природы с иерархической группировкой по нескольким уровням. При иерархической классификации все исходное множество объектов составляет высший уровень иерархии и разбивается в зависимости от выбранного классификационного принципа на классы (группы), которые образуют предыдущий уровень; каждый класс этого уровня делится на подклассы, которые образуют более низкий уровень, у которого каждый подкласс аналогично разбивается на группы, соответствующие нижеследующему уровню и т.д. *Любой уровень иерархии состоит из групп, представляющих собой разбиение всего множества объектов, т.е. группы одного уровня составляют все исходное множество*. Будем называть объекты самого низкого уровня иерархии в используемом материале *единицами наблюдения*. Когда информация о реально существующих объектах низшего уровня отсутствует (или игнорируется), в их качестве можно выбрать объекты на одном и том же из более высоких уровней иерархии, “забывая” о существовании подразделенности таких “единиц”.

Каждой из полученных подобным образом групп объектов присвоим цифровой идентификатор, соответствующий ее положению в иерархии. Он может быть построен, например, следующим образом. Идентификатор отдельной группы начинается с ее номера внутри множества групп данного уровня и продолжается последовательностью номеров вышестоящих групп, которым “подчиняются” предыдущие. В результате любая из групп однозначно определяется *мультиномером* (*идентификатором*) в виде последовательности из номеров групп все более высокого уровня, подчиняющих все предшествующие. Логика построения этой последовательности напоминает написание почтового адреса (указывающего населенный пункт, район и область), а также принцип генеалогической систематики биологических видов и нумерацию в библиотечном систематическом каталоге. Графически иерархическая классификация отображается древовидной структурой.

У нас объектами являются субпопуляции, скажем, села, группирующиеся в сельсоветы, районы и т.д. с соответствующими уровнями иерархии 1, 2, 3 ... и составляющие всю метапопуляцию. Подчинение одной субпопуляции другой означает вхождение первой в качестве составной части во вторую с более высоким уровнем иерархии. Обозначим номер конкретного села (первый уровень) как s_1; номер сельсовета (второй уровень), куда входит село, как s_2; номер района (третий уровень), включающего указанные сельсовет и село, как s_3; и т.д. Тогда мультиномер $s_1.s_2.s_3.$ ... однозначно определяет рассматриваемое село среди прочих сел первого уровня, $s_2.s_3.s_4.$ ... идентификатор сельсовета, $\mathbf{s}_i \equiv s_i.s_{i+1}.$ ... идентифицирует субпопуляцию i-го уровня среди прочих таких же субпопуляций внутри соответствующей группы следующего уровня $i + 1$. *Таким образом, индекс i у $\mathbf{s}_i$ (s_i) дает уровень иерархии данной субпопуляции*.

В результате субпопуляция $\mathbf{s}_1$ входит в $\mathbf{s}_2$, ..., $\mathbf{s}_i$ входит в $\mathbf{s}_{i+1}$ и т.д., т.е. *субпопуляция некоторого уровня иерархии включает в себя в качестве составной части соответствующие субпопуляции более низкого уровня*. Между объектами и их идентификаторами имеется взаимно однозначное соответствие, и мы иногда будем писать идентификатор вместо названия объекта (села, сельсовета и т.д.). Кроме того, повторим, что множество субпопуляций на каждом отдельно выбранном уровне иерархии представляют собой разбиение всей метапопуляции, т.е. составляют ее целиком.

Данный способ нумерации, например, приложим к концентрации x интересующей фамилии в селе, которую будем обозначать как $x(s_1.s_2.s_3.\ldots) \equiv x(\mathbf{s}_1)$, а концентрацию фамилии в группе i-го уровня как $x(s_i.s_{i+1}.\ldots) \equiv x(\mathbf{s}_i)$, где мультиномер $\mathbf{s}_i \equiv s_i.s_{i+1}.\ldots$ содержит последовательность номеров групп объектов, каждая из которых будет на единицу более высокого уровня и содержит предыдущую. Например, $x(\mathbf{s}_3) = x(s_3.s_4.s_5.\ldots)$ дает концентрацию фамилии в районе с номером s_3, входящем в область (следующий уровень) с номером s_4, и т.д.

При этом на практике реальную подразделенную метапопуляцию можно рассматривать как теоретическую совокупность, а случайный выбор из нее субпопуляций позволяет использовать вероятностный подход, в частности говорить о математических ожиданиях (проще говоря, о средних значениях), дисперсиях и пр. В соответствующем контексте некоторые из номеров $\{s_i\}$ будут рассматриваться как случайные величины, а некоторые как фиксированные. Для наглядности будем писать в мультиномере фиксированные величины после вертикальной черты. Тогда s_1 в $x(s_1|s_2.s_3...)$ является случайной величиной, значениями которой будут номера сел при условии их выбора из фиксированного сельсовета с номером s_2 (который находится внутри своего района с номером s_3 и т.д.); $x(\mathbf{s}_{i-1}|\mathbf{s}_i)$ будет случайной величиной, значениями которой являются концентрации фамилии в субпопуляциях $(i - 1)$-го уровня внутри фиксированной группы $\mathbf{s}_i$.

Первый аргумент $\mathbf{s}_{i-1}$у $x(\mathbf{s}_{i-1}|\mathbf{s}_i)$ указывает на случайно выбираемую субпопуляцию, а второй $\mathbf{s}_i$ на содержащую ее фиксированную группу следующего уровня. Таким образом, $x(\mathbf{s}_i)$ – концентрация фамилии в фиксированной субпопуляции $\mathbf{s}_i$, а $x(\mathbf{s}_{i-1}|\mathbf{s}_i)$ – слу-

чайная величина, принимающая значения концентраций фамилии в субпопуляциях уровня $i-1$, входящих в $\mathbf{s}_i$. Иногда удобней использовать запись, принятую для условных математических ожиданий, и рассматривать фиксированные номера как условие, которое будем отделять вертикальной чертой после обозначения случайной величины, т.е. $x(\mathbf{s}_{i-1}|\mathbf{s}_i)$ и $x(\mathbf{s}_{i-1})|\mathbf{s}_i$ обозначают одну и ту же случайную величину.

СРЕДНИЕ ЗНАЧЕНИЯ КОНЦЕНТРАЦИИ ФАМИЛИИ И ДРУГИХ ХАРАКТЕРИСТИК НА ОТДЕЛЬНЫХ УРОВНЯХ ИЕРАРХИИ

Субпопуляцию каждого уровня характеризует концентрация рассматриваемой фамилии в ней. Так, $x(\mathbf{s}_2)$ обозначает концентрацию фамилии в сельсовете $\mathbf{s}_2$. Она является математическим ожиданием (средним значением) концентрации фамилии в совокупности сел (со случайными номерами $\{s_1\}$) при условии принадлежности каждого из них этому сельсовету с фиксированным номером s_2. Термины "математическое ожидание" и "среднее значение" являются взаимозаменяемыми.

Определение 1. *Математическим ожиданием $E\{x\}$ распределения случайной величины x, которая может принимать конечное число значений $\{x_i\}$ с вероятностями $\{Pr(x_i)\}$, называется константа, определяемая как среднее взвешенное значение для $\{x_i\}$ вида*

$$E\{x\} \equiv x_1 Pr(x_1) + x_2 Pr(x_2) + \ldots = \sum x_i Pr(x_i). \quad (1)$$

Напомним следующие свойства математического ожидания

$$E\{x - E\{x\}\} = 0,$$
$$E\{c\} = c\,(\text{откуда } E\{E\{x\}\} = E\{x\}),$$
$$E\{cx\} = cE\{x\},$$
$$E\{x + y\} = E\{x\} + E\{y\}, \quad E\{xy\} = E\{x\}E\{y\}$$

для любой константы c и любых случайных величин x и y (для произведения xy требуется независимость сомножителей).

Далее для вычисления математических ожиданий мы будем широко использовать следующую *формулу полного математического ожидания* случайной величины x:

$$E\{x\} = E\{E_x\{x|A\}\} = \sum_i E_x\{x|A_i\}Pr(A_i). \quad (2)$$

Здесь A обозначает событие в $\{A_i\}$, в *полной системе* несовместимых случайных событий, реализующихся с вероятностями $\{Pr(A_i)\}$ и таких, что обязательно происходит одно из них; $E_x\{x|A_i\}$ означает условное (условие пишем после вертикальной черты) математическое ожидание для случайной величины x (нижний индекс у E указывает на переменную, которая является случайной и по которой производится усреднение) при условии реализации соответствующего случайного события A_i.

Например, в искусственной ситуации, когда полная система состоит из городских и сельских жителей, средний вес жителя $E\{вес\}$ равен $E\{вес|житель\ города\} \cdot Pr(житель\ города) + E\{вес|житель\ села\} \cdot Pr(житель\ села)$. У нас при случайном выборе субпопуляций (групп) в качестве полной системы обычно рассматриваются (непересекающиеся) субпопуляции $\{\mathbf{s}_i\}$ с одним и тем же уровнем иерархии, вместе составляющие всю метапопуляцию $\mathbf{s}$.

Разбиением какого-либо множества называется его представление в виде объединения произвольного количества попарно непересекающихся непустых подмножеств. Ясно, что все части разбиения совокупности $\mathbf{s}$ на любом уровне иерархии i образуют полную систему случайных событий, реализующихся при выборе наугад $\mathbf{s}_i$ из $\mathbf{s}$. Каждую из таких систем можно использовать в формуле полного математического ожидания (2) для случайной величины.

Очевидно, что в любой совокупности, состоящей из объектов с числовой характеристикой x, среднее значение x в произвольной части совокупности выражается как среднее значение x для объектов, входящих в эту часть. В подразделенной популяции $\mathbf{s}$ концентрация фамилии $x(\mathbf{s}_i)$ в группе $\mathbf{s}_i$ выражается через концентрации входящих в $\mathbf{s}_i$ субпопуляций $\{\mathbf{s}_1\}$, единиц наблюдения, по формуле математического ожидания как

$$x(\mathbf{s}_i) \equiv E_{\mathbf{s}_1}\{x(\mathbf{s}_1)|\mathbf{s}_i\} = \sum_{\mathbf{s}_1} x(\mathbf{s}_1)\, Pr(\mathbf{s}_1|\mathbf{s}_i),$$

где $E_{\mathbf{s}_1}\{x(\mathbf{s}_1)|\mathbf{s}_i\}$ обозначает математическое ожидание случайной величины $x(\mathbf{s}_1|\mathbf{s}_i)$, принимающей значения концентраций фамилии в субпопуляциях $\{\mathbf{s}_1\}$ при выборе наугад $\mathbf{s}_1$ из $\mathbf{s}_i$; нижний индекс у $E_{\mathbf{s}_1}$ служит для облегчения ориентации в уровне иерархии субпопуляций, выбор которых случаен и по которым происходит усреднение; $Pr(\mathbf{s}_1|\mathbf{s}_i)$ – вероятность случайного выбора субпопуляции $\mathbf{s}_1$ при условии, что выбор производится из группы $\mathbf{s}_i$; суммирование осуществляется по всем таким субпопуляциям $\{\mathbf{s}_1\}$ в группе $\mathbf{s}_i$. Для фамильных данных $x(\mathbf{s}_i)$ – это математическое ожидание концентрации фамилии в случайно выбранном селе $\mathbf{s}_1$ внутри $\mathbf{s}_i$ (это среднее взвешенное значение $x(\mathbf{s}_1)$ с весами $\{Pr(\mathbf{s}_1|\mathbf{s}_i)\}$ для концентраций $\{x(\mathbf{s}_1|\mathbf{s}_i)\}$ в селах $\{\mathbf{s}_1\}$).

Аналогично можно определить значение произвольной функции от концентрации фамилии $g(x(\mathbf{s}_2))$ в субпопуляции второго уровня $\mathbf{s}_2$ как математическое ожидание значений g в составляющих $\mathbf{s}_2$ субпопуляциях первого уровня $\{\mathbf{s}_1\}$. Для уровня i функция $g(x(\mathbf{s}_i))$ также определяется значениями g в $\{\mathbf{s}_{i-1}\}$, составляющих $\mathbf{s}_i$, и $g(x(\mathbf{s}_i))$ дает среднее значение как для набора $\{g(x(\mathbf{s}_{i-1}))\}$, так и для соответствующего набора $\{g(x(\mathbf{s}_1))\}$ из $\{\mathbf{s}_1\}$, содержащихся в $\mathbf{s}_i$. Например, когда роль g играет x,

то получим приведенную ранее формулу для $x(\mathbf{s}_i)$. Выше мы применили одно и то же обозначение $\mathbf{s}_1$ к разным субпопуляциям $\{\mathbf{s}_1\}$ первого уровня, случайно выбираемым из $\mathbf{s}_i$, так как $x(\mathbf{s}_1|\mathbf{s}_i)$ – случайная переменная величина, принимающая значения $\{x(\mathbf{s}_1)\}$ с вероятностями $\{Pr(\mathbf{s}_1|\mathbf{s}_i)\}$. Один из способов фиксировать конкретную субпопуляцию состоит в присваивании ей номера. Тогда $\{s_1\}$ представимо как $\{s_{k1}\}$, где k нумерует субпопуляции данного первого уровня.

Замечание 1. *Пусть рассматривается совокупность* $\mathbf{s}$ *объектов* $\{\mathbf{s}_1\}$*, разбитая на произвольные* (*непересекающиеся*) *части* $\{\mathbf{s}_2\}$*, и для каждого из объектов* $\{\mathbf{s}_1\}$ *определена числовая характеристика* $x = x(\mathbf{s}_1)$.

Тогда математическое ожидание $E\{x(\mathbf{s}_1)\}$ *для значений* x *во всей совокупности* $\mathbf{s}$ *равно среднему значению для математических ожиданий* $E_{\mathbf{s}_1}\{x(\mathbf{s}_1)|\mathbf{s}_{j2}\} \equiv x(\mathbf{s}_{j2})$ *значений* x *в отдельных ее частях*:

$$E\{x(\mathbf{s}_1)\} = E_{j2}\{x(\mathbf{s}_{j2})|\mathbf{s}\} = \sum_j x(\mathbf{s}_{j2})Pr(\mathbf{s}_{j2}|\mathbf{s}), \quad (3)$$

где j *нумерует части* $\{\mathbf{s}_2\} = \{\mathbf{s}_{j2}\}$ *рассматриваемого разбиения*, $Pr(\mathbf{s}_{j2}|\mathbf{s})$ *обозначает вероятность случайного выбора* $\mathbf{s}_{j2}$ *из* $\mathbf{s}$.

Доказательство. Напомним, что если дана совокупность $\mathbf{s}$ объектов $\{\mathbf{s}_1\}$ с числовым признаком $x(\mathbf{s}_1)$, то средним значением x в ней по определению (1) будет $E\{x(\mathbf{s}_1)\} \equiv E_{\mathbf{s}_{k1}}\{x(\mathbf{s}_{k1})|\mathbf{s}\}$. Покажем, что

$$E\{x(\mathbf{s}_1)\} \equiv E_{\mathbf{s}_{k1}}\{x(\mathbf{s}_{k1})|\mathbf{s}\} = E_{j2}\{x(\mathbf{s}_{j2})|\mathbf{s}\},$$
$$\text{т.е. } \sum_{\mathbf{s}_1} x(\mathbf{s}_1)Pr(\mathbf{s}_1|\mathbf{s}) = \sum_j x(\mathbf{s}_{j2})Pr(\mathbf{s}_{j2}|\mathbf{s}).$$

Например, когда объектами являются индивидуумы, характеризуемые своим весом $x(\mathbf{s}_1)$, то эта формула для $E\{x\}$ дает средний вес индивидуума. В рамках фамильных исследований объектами (единицами наблюдения) в совокупности (метапопуляции) являются субпопуляции $\{\mathbf{s}_1\}$, а числовым признаком объекта (субпопуляции) $\mathbf{s}_1$ служит концентрация $x(\mathbf{s}_1)$ рассматриваемой фамилии в $\mathbf{s}_1$. При выборе наугад субпопуляции $\mathbf{s}_1$ из $\mathbf{s}$ значение $x(\mathbf{s}_1)$ будет случайной величиной со значениями $\{x(\mathbf{s}_{k1})\}$, наблюдаемыми в субпопуляциях первого уровня, занумерованными индексом k.

Так как математическое ожидание $E\{x(\mathbf{s}_1)\}$ является взвешенной суммой значений $\{x(\mathbf{s}_1)\}$, то оно *не зависит от порядка слагаемых*. Поэтому можно расположить составляющие математическое ожидание $E\{x(\mathbf{s}_1)\}$ слагаемые блоками, соответствующими частям разбиения совокупности $\mathbf{s}$ на $\{\mathbf{s}_{j2}\}$. Здесь $\mathbf{s}_{j2}$ обозначает j-ю часть совокупности $\mathbf{s}$, состоящую из надлежащих $\{\mathbf{s}_1\} = \{\mathbf{s}_{k1}\}$, где k нумерует объекты внутри $\{\mathbf{s}_{j2}\}$. Другими словами, в соответствующем контексте $\mathbf{s}_1$ ($\mathbf{s}_2$) является случайной переменной со значениями $\{\mathbf{s}_{k1}\}$ ($\{\mathbf{s}_{j2}\}$).

При суммировании по переменной $\mathbf{s}_1$ в случае упорядочивания по j расположения блоков сначала получим взвешенную сумму слагаемых, соответствующих первой части $\mathbf{s}_{12}$ разбиения, потом сумму для второй части $\mathbf{s}_{22}$ и т.д. В итоге математическое ожидание x для всей подразделенной совокупности, разбитой на части $\{\mathbf{s}_{j2}\}$, находится суммированием внутри очередного блока (по $\mathbf{s}_1$ в соответствующей части) и по самим блокам (т.е. по частям совокупности, пока они не будут исчерпаны). В результате получаем сумму по всем $\mathbf{s}_1$ из $\mathbf{s}$, иначе говоря, по всем k:

$$x(\mathbf{s}) \equiv E_{\mathbf{s}_1}\{x(\mathbf{s}_1)|\mathbf{s}\} \equiv \sum_{\mathbf{s}_1} x(\mathbf{s}_1)Pr(\mathbf{s}_1|\mathbf{s}) =$$
$$= \sum_k x(\mathbf{s}_{k1})Pr(\mathbf{s}_{k1}|\mathbf{s}).$$

Здесь $Pr(\mathbf{s}_{k1}|\mathbf{s})$ обозначает вероятность случайного выбора $\mathbf{s}_{k1}$ из $\mathbf{s}$, т.е. выбора соответствующего объекта $\mathbf{s}_1$. Учтем, что согласно очевидному варианту формулы (2) вероятность выбора наугад $\mathbf{s}_{k1}$ из $\mathbf{s}$ выражается через вероятности выборов $\mathbf{s}_{k1}$ из $\mathbf{s}_{j2}$ и $\mathbf{s}_{j2}$ из $\mathbf{s}$ как

$$Pr(\mathbf{s}_{k1}|\mathbf{s}) = \sum_j Pr(\mathbf{s}_{k1}|\mathbf{s}_{j2})Pr(\mathbf{s}_{j2}|\mathbf{s}).$$

В результате подстановки $\sum_j Pr(\mathbf{s}_{k1}|\mathbf{s}_{j2})Pr(\mathbf{s}_{j2}|\mathbf{s})$ вместо $Pr(\mathbf{s}_{k1}|\mathbf{s})$ в $x(\mathbf{s})$ и изменения порядка суммирования получаем

$$x(\mathbf{s}) = \sum_k x(\mathbf{s}_{k1})Pr(\mathbf{s}_{k1}|\mathbf{s}) =$$
$$= \sum_k x(\mathbf{s}_{k1})\sum_j Pr(\mathbf{s}_{k1}|\mathbf{s}_{j2})\,Pr(\mathbf{s}_{j2}|\mathbf{s}) =$$
$$= \sum_j \Big(\sum_k x(\mathbf{s}_{k1})Pr(\mathbf{s}_{k1}|\mathbf{s}_{j2})\Big)Pr(\mathbf{s}_{j2}|\mathbf{s}) =$$
$$= \sum_j x(\mathbf{s}_{j2})Pr(\mathbf{s}_{j2}|\mathbf{s}) = E_{j2}\{x(\mathbf{s}_{j2})|\mathbf{s}\}.$$

Здесь при каждом j суммирование по k (суммирование по $\mathbf{s}_{k1}$) идет не по всей совокупности, а внутри ее j-й части (внутри блока j). Согласно определению (1) внутреннее суммирование дает среднее значение x в субпопуляции $\mathbf{s}_{j2}$ $\left(\sum_k x(\mathbf{s}_{k1})Pr(\mathbf{s}_{k1}|\mathbf{s}_{j2})\right) \equiv x(\mathbf{s}_{j2})$), а внешнее суммирование по j (по $\mathbf{s}_{j2}$) дает сумму по блокам $\left(\sum_j x(\mathbf{s}_{j2})Pr(\mathbf{s}_{j2}|\mathbf{s})\right)$. Таким образом получаем сумму по всем $\{\mathbf{s}_1\}$ в $\mathbf{s}$. Тем самым среднее значение $x(\mathbf{s}_1)$ во всей подразделенной совокупности выражается через средние величины $\{x(\mathbf{s}_{j2})\}$ в составляющих ее частях $\{\mathbf{s}_{j2}\}$ как $x(\mathbf{s})\} \equiv E\{x\} = \sum_j x(\mathbf{s}_{j2})Pr(\mathbf{s}_{j2}|\mathbf{s})$. ◄

Теперь обратимся к иерархически подразделенным совокупностям. У нас иерархическая структура разбиений означает, что любая часть $\mathbf{s}_{ji}$ разбиения состоит, в свою очередь, из частей следующего более низкого уровня иерархии, разби-

тых иерархически вплоть до неподразделенных единиц наблюдения $\{\mathbf{s}_1\}$.

Следствие 2. *Пусть разбиение совокупности* $\mathbf{s}$ *на части* $\{\mathbf{s}_{ji}, j = 1, 2, ...\}$ *является иерархическим, индекс* $i = 1, 2, ...$ *у* $\mathbf{s}_{ji}$ *обозначает уровень иерархии, а* j *– номер части разбиения на этом уровне. Положим, что каждая часть* $\mathbf{s}_{ji}$ *характеризуется соответствующим средним значением* $x(\mathbf{s}_{ji})$ *числового признака* x.

Тогда $x(\mathbf{s}_{ji})$ *выражается через* $\{x(\mathbf{s}_{km}), k = 1, 2, ...\}$, *т.е. через средние значения* x *в содержащихся в* $\mathbf{s}_{ji}$ *частях* $\{\mathbf{s}_{km}\}$ *одного и того же (более низкого) уровня* $m < i$, *как*

$$x(\mathbf{s}_{ji}) \equiv E_{\mathbf{s}_1}\{x(\mathbf{s}_1)|\mathbf{s}_{ji}\} = \sum_k x(\mathbf{s}_{km}) Pr(\mathbf{s}_{km}|\mathbf{s}_{ji}) = E_{km}\{x(\mathbf{s}_{km})|\mathbf{s}_{ji}\}, \quad m < i,$$

и при $m = 2, 3, ...$

$$x(\mathbf{s}_{ji}) \equiv E_x\{x(\mathbf{s}_1)|\mathbf{s}_{ji}\} = E_{k1}\{x(\mathbf{s}_{k1})|\mathbf{s}_{ji}\} = E_{k2}\{x(\mathbf{s}_{k2})|\mathbf{s}_{ji}\} = \ldots = E_{k\,i-1}\{x(\mathbf{s}_{k\,i-1})|\mathbf{s}_{ji}\},$$

$$x(\mathbf{s}_{ji}) \equiv E_{k\,i-1}\{x(\mathbf{s}_{k\,i-1})|\mathbf{s}_{ji}\} = E_{k\,i-1}\{E_{n\,i-2}\{x(\mathbf{s}_{n\,i-2})|\mathbf{s}_{k\,i-1}\}|\mathbf{s}_{ji}\}, \quad (4)$$

где индекс типа km *у* $E_{km}\{x(\mathbf{s}_{km})|\mathbf{s}_{ji}\}$ *указывает на операцию усреднения* $\{x(\mathbf{s}_{km})\}$ *по* k, *когда уровень иерархии* m *фиксирован и на нем случайно выбирается субпопуляция* $\mathbf{s}_{km}$, *входящая в состав* $\mathbf{s}_{ji}$, $k = 1, 2, 3, ...$.

Доказательство. Напомним, что в доказанном выше замечании разбиение совокупности на блоки произвольно. Например, при ее иерархической подразделенности такое разбиение $\mathbf{s}_{ji}$ может состоять из частей на $(i-1)$-м уровне и среднее значение x для $\mathbf{s}_{ji}$ является средним для средних значений x (обозначаемых $x(\mathbf{s}_{k\,i-1})$) для частей $\{\mathbf{s}_{k\,i-1}\}$, входящих в $\mathbf{s}_{ji}$ на уровне иерархии $i-1$. В частном случае, когда выбран уровень $m = i-1$, имеем

$$x(\mathbf{s}_{ji}) \equiv E_{k\,i-1}\{x(\mathbf{s}_{k\,i-1})|\mathbf{s}_{ji}\} = \sum_k x(\mathbf{s}_{k\,i-1}) Pr(\mathbf{s}_{k\,i-1}|\mathbf{s}_{ji}), \quad i = 2, 3, \ldots .$$

Точно так же можно выразить среднее значение $x(\mathbf{s}_{k\,i-1})$ для x в части совокупности $\mathbf{s}_{k\,i-1}$ как $E_{n\,i-2}\{x(\mathbf{s}_{n\,i-2})|\mathbf{s}_{k\,i-1}\}$, где n нумерует части разбиения $\mathbf{s}_{k\,i-1}$, откуда

$$x(\mathbf{s}_{ji}) \equiv E_{k\,i-1}\{x(\mathbf{s}_{k\,i-1})|\mathbf{s}_{ji}\} = E_{k\,i-1}\{E_{n\,i-2}\{x(\mathbf{s}_{n\,i-2})|\mathbf{s}_{k\,i-1}\}|\mathbf{s}_{ji}\},$$ т.е.

$$E_{k\,i-1}\{E_{n\,i-2}\{x(\mathbf{s}_{n\,i-2})|\mathbf{s}_{k\,i-1}\}|\mathbf{s}_{ji}\} = E_{k\,i-1}\{x(\mathbf{s}_{k\,i-1})|\mathbf{s}_{ji}\}. \blacktriangleleft$$

Если продолжить эту процедуру, то далее получим, что

$$x(\mathbf{s}_{ji}) \equiv E_{\mathbf{s}_1}\{x(\mathbf{s}_1)|\mathbf{s}_{ji}\} = E_{k\,i-1}\{E_{m\,i-2}\{\ldots\{E_{\mathbf{s}_1}\{x(\mathbf{s}_1)|\mathbf{s}_2\}\ldots|\mathbf{s}_{m\,i-2}\}|\mathbf{s}_{k\,i-1}\}|\mathbf{s}_{ji}\}, \quad i = 2, 3, \ldots,$$

т.е.

$$E_{k\,i-1}\{E_{m\,i-2}\{\ldots\{E_{\mathbf{s}_1}\{x(\mathbf{s}_1)|\mathbf{s}_2\}\ldots|\mathbf{s}_{m\,i-2}\}|\mathbf{s}_{k\,i-1}\}|\mathbf{s}_{ji}\} = E_{\mathbf{s}_1}\{x(\mathbf{s}_1)|\mathbf{s}_{ji}\}.$$

Формула полного математического ожидания (2) верна для любой случайной величины, и выше она использовалась на примере иерархически подразделенной совокупности, где фигурировала случайная величина x. В контексте изучения фамильной структуры мы интерпретируем x как концентрацию фамилии в соответствующей субпопуляции. В дальнейшем будут рассматриваться функции от x, например x^2 или вероятность $Hs(x(\mathbf{s}))$ случайной встречи индивидуумов с разными фамилиями, и случайной величиной может быть произвольная функция $g(x(\mathbf{s}))$.

Очевидно, когда g равняется x, x^2 или Hs, то в субпопуляции уровня иерархии i концентрация $x(\mathbf{s}_i)$ фамилии, ее (средний) квадрат или вероятность Hs случайной встречи двух индивидуумов с разными фамилиями согласно (4) находятся как

$$x(\mathbf{s}_i) \equiv E_{\mathbf{s}_1}\{x(\mathbf{s}_1)|\mathbf{s}_i\} = \sum_{\mathbf{s}_1} x(\mathbf{s}_1) Pr(\mathbf{s}_1|\mathbf{s}_i),$$

$$x^2(\mathbf{s}_i) \equiv E_{\mathbf{s}_1}\{x^2(\mathbf{s}_1)|\mathbf{s}_i\} = \sum_{\mathbf{s}_1} x^2(\mathbf{s}_1) Pr(\mathbf{s}_1|\mathbf{s}_i).$$

$$Hs(x(\mathbf{s}_i)) = E_{\mathbf{s}_1}\{Hs(x(\mathbf{s}_1))|\mathbf{s}_i\} = \sum_{\mathbf{s}_1} Hs(x(\mathbf{s}_1)) Pr(\mathbf{s}_1|\mathbf{s}_i)$$

при $g(x(\mathbf{s}_i)) = Hs(x(\mathbf{s}_i))$. Здесь $Pr(\mathbf{s}_1|\mathbf{s}_i)$ – вероятность случайного выбора субпопуляции $\mathbf{s}_1$ из субпопуляции $\mathbf{s}_i$; суммирование осуществляется по всем $\mathbf{s}_1$ из $\mathbf{s}_i$.

ДИСПЕРСИЯ В ИЕРАРХИЧЕСКИ ПОДРАЗДЕЛЕННОЙ МЕТАПОПУЛЯЦИИ

В метапопуляции с иерархической структурой подразделенности на каждом из уровней существует в общем случае своя фамильная дивергенция субпопуляций. По аналогии с анализом неподразделенной популяции [1, 2] в качестве характеристик дивергенции в совокупности субпопуляций можно использовать дисперсии концентраций фамилий и вероятности случайных встреч индивидуумов с одинаковыми и с разными фамилиями.

Повторим, что когда рассматривается реальная группа субпопуляций с фиксированными состояниями, то при случайном выборе одной из них будем использовать вероятностную технику для вычисления интересующих нас (скажем, средних значений, дисперсий) характеристик. При этом средние значения используются как математические ожидания для описания status quo метапопуляции без

предположения о ее состоянии как случайной выборки из некоторой теоретической совокупности.

Напомним определение дисперсии.

Определение 2. *Дисперсией $V(x)$ случайной величины x называется константа, определяемая формулой*

$$V(x) \equiv E\left\{(x - E\{x\})^2\right\} = \\ = \sum_x (x - E\{x\})^2 \, Pr(x), \text{ или } E\{x^2\} - (E\{x\})^2. \quad (5)$$

Таким образом, дисперсия является средним квадратом расстояния между случайным значением x и математическим ожиданием $E\{x\}$ и, как хорошо известно, равна разности между средним квадратом $E\{x^2\}$ и квадратом среднего $(E\{x\})^2$. У нас дисперсия концентрации фамилии характеризует фамильную дивергенцию субпопуляций и важна также из-за ее связи с генетической дивергенцией и с коэффициентом инбридинга (см., например, [1, 2]). Напомним, что дисперсия случайной величины $x(\mathbf{s}_i|\mathbf{s}_{i+1})$ это у нас дисперсия концентрации x, наблюдаемой в субпопуляции $\mathbf{s}_i$ при выборе наугад $\mathbf{s}_i$ из $\mathbf{s}_{i+1}$.

Мультиномер в качестве идентификатора можно использовать не только вместе с концентрациями фамилий, но и с другими характеристиками состояния популяций, например с дисперсиями. В соответствии с принятыми нами обозначениями $x(\mathbf{s}_i|\mathbf{s}_{i+1})$ является случайной величиной, принимающей значения $\{x(\mathbf{s}_i)\}$, где $x(\mathbf{s}_i)$ – концентрация рассматриваемой фамилии в субпопуляции $\mathbf{s}_i$ внутри $\mathbf{s}_{i+1}$. *Обозначим через $Vs(x(\mathbf{s}_i|\mathbf{s}_{ji+1}))$ дисперсию случайной величины x $(\mathbf{s}_i|\mathbf{s}_{ji+1})$, которая является аргументом у Vs.* Согласно определениям математического ожидания $E_{\mathbf{s}_i}\{x(\mathbf{s}_i)|\mathbf{s}_{ji+1}\}$ и дисперсии $Vs(x(\mathbf{s}_i|\mathbf{s}_{ji+1}))$ для распределения концентрации фамилии по субпопуляциям i-го уровня $\{\mathbf{s}_i\}$, входящим в $\mathbf{s}_{ji+1}$, т.е. для случайной величины $x(\mathbf{s}_i|\mathbf{s}_{ji+1})$, имеем с учетом (5)

$$E_{\mathbf{s}_i}\{x(\mathbf{s}_i)|\mathbf{s}_{ji+1}\} \equiv \sum_{\mathbf{s}_i} x(\mathbf{s}_i)\, Pr(\mathbf{s}_i|\mathbf{s}_{ji+1}) = \\ = \sum_k x(\mathbf{s}_{ki})\, Pr(\mathbf{s}_{ki}|\mathbf{s}_{ji+1}) = x(\mathbf{s}_{ji+1}),$$

$$Vs(x(\mathbf{s}_i|\mathbf{s}_{ji+1})) \equiv E_{\mathbf{s}_i}\left\{\left(x(\mathbf{s}_i) - E_{\mathbf{s}_i}\{x(\mathbf{s}_i)\}\right)^2|\mathbf{s}_{ji+1}\right\} = \\ = E_{\mathbf{s}_i}\left\{(x(\mathbf{s}_i) - x(\mathbf{s}_{i+1}))^2|\mathbf{s}_{ji+1}\right\} = \\ = E_{\mathbf{s}_i}\left\{x^2(\mathbf{s}_i)|\mathbf{s}_{ji+1}\right\} - x^2(\mathbf{s}_{ji+1}),$$

$$Vs(x(\mathbf{s}_i|\mathbf{s}_{jm})) \equiv E_{\mathbf{s}_i}\left\{\left(x(\mathbf{s}_i) - E_{\mathbf{s}_i}\{x(\mathbf{s}_i)\}\right)^2|\mathbf{s}_{jm}\right\} = \\ = E_{\mathbf{s}_i}\left\{x^2(\mathbf{s}_i)|\mathbf{s}_{jm}\right\} - x^2(\mathbf{s}_{jm}), \quad m > i. \quad (6)$$

Здесь к символу дисперсии V (и далее к стандартным в биометрии обозначениям других статистических характеристик) добавляется буква s для напоминания, что речь идет о фамилиях, а не об аллелях в субпопуляциях.

ПОЛНАЯ, МЕЖГРУППОВАЯ И ВНУТРИГРУППОВАЯ ДИСПЕРСИИ

В подразделенной на группы популяции выделяют несколько типов дисперсии: полную (общую) во всей метапопуляции, межгрупповую (дисперсию распределения средних значений признака по группам) и внутригрупповую дисперсии числового признака. Поясним это следующим искусственным примером. Рассмотрим простейшую иерархическую систему из изучаемых в отношении веса жителей сельсовета, состоящего из сел (единицей наблюдения является отдельный житель, характеризуемый значением своего веса). Тогда имеются три типа совокупностей: полная (общая) совокупность значений веса у всех жителей сельсовета, совокупность средних значений веса жителей в отдельных селах сельсовета и совокупности значений веса у жителей внутри отдельных сел сельсовета.

Соответственно будет три типа дисперсий: полная (общая) дисперсия значений веса для всех жителей сельсовета (дисперсия веса у единиц наблюдения при отсутствии у них какой-либо группировки), межгрупповая дисперсия средних значений веса жителей его сел и внутригрупповая дисперсия веса жителей отдельного села (сельсовет в целом характеризует средняя внутригрупповая дисперсия для его сел).

В нашем случае анализа фамилий *на популяционном уровне* единицей наблюдения является отдельное село $\mathbf{s}_1$, характеризуемое теперь не весом, а концентрацией рассматриваемой фамилии $x(\mathbf{s}_1)$ для жителей села $\mathbf{s}_1$ в целом; села группируются в сельсоветы $\{\mathbf{s}_2\}$ с концентрациями фамилии в них $\{x(\mathbf{s}_2)\}$, а сельсоветы образуют некоторый район $\mathbf{s}_3$ с концентрацией фамилии (общим средним значением концентрации) в нем $x(\mathbf{s}_3)$. Очевидно, все вместе села (субпопуляции первого уровня) дают район, точно так же как его дают и все сельсоветы (субпопуляции второго уровня), вообще все субпопуляции любого отдельного уровня образуют метапопуляцию целиком. Для наугад выбранной субпопуляции $\mathbf{s}_1$ из $\mathbf{s}_{j2}$ концентрация фамилии в $\mathbf{s}_1$ будет случайной величиной $x(\mathbf{s}_1|\mathbf{s}_{j2})$ с математическим ожиданием $x(\mathbf{s}_{j2})$, равным среднему значению концентрации фамилии в селах $\{x(\mathbf{s}_1)\}$.

Рассмотрим общее математическое ожидание концентрации рассматриваемой фамилии в распределенной по району $\mathbf{s}_3$ иерархически подразделенной метапопуляции. Она состоит из субпопуляций сельсоветов и сел с мультиномерами $\{\mathbf{s}_2\}$, $\{\mathbf{s}_1\}$, соответствующими уровням иерархии 1 и 2. Концентрации интересующей фамилии в них обозначены как $\{x(\mathbf{s}_2)\}$ и $\{x(\mathbf{s}_1)\}$. Согласно ранее изложенному $x(\mathbf{s}_{j2}) \equiv E_{\mathbf{s}_1}\{x(\mathbf{s}_1)|\mathbf{s}_{j2}\}$, а $x(\mathbf{s}_3) \equiv E_{\mathbf{s}_2}\{x(\mathbf{s}_2)|\mathbf{s}_3\}$. Полная дисперсия (по несгруппированным единицам наблюдения (селам)) – это дисперсия распре-

деления значений концентрации фамилии $\{x(\mathbf{s}_1)|\mathbf{s}_3\}$ по селам всего района.

Распределение концентрации фамилии $\{x(\mathbf{s}_2|\mathbf{s}_3)\}$ по субпопуляциям второго уровня (сельсоветам) характеризуется не только средней концентрацией $x(\mathbf{s}_3)$ (математическим ожиданием), но и межгрупповой дисперсией $Vs_{betw}(x(\mathbf{s}_2|\mathbf{s}_3))$, отражающей фамильную дивергенцию субпопуляций $\{\mathbf{s}_2\}$ друг от друга (по сельсоветам). Эту межгрупповую дисперсию можно представить согласно (6) как разность среднего квадрата и квадрата среднего:

$$Vs_{betw}(x(\mathbf{s}_2|\mathbf{s}_3)) \equiv E_{\mathbf{s}_2}\left\{x^2(\mathbf{s}_2)\middle|\mathbf{s}_3\right\} - \left(E_{\mathbf{s}_2}\left\{x(\mathbf{s}_2)\middle|\mathbf{s}_3\right\}\right)^2 = \\ = E_{\mathbf{s}_2}\left\{x^2(\mathbf{s}_2)\middle|\mathbf{s}_3\right\} - x^2(\mathbf{s}_3).$$

Дивергенция обычно существует между субпопуляциями $\{\mathbf{s}_2\}$ как на данном втором уровне, так и между входящими в отдельные группы $\{\mathbf{s}_2\}$ субпопуляциями $\{\mathbf{s}_1\}$ на единицу меньшего уровня (между селами $\{\mathbf{s}_1\}$ внутри каждого сельсовета). Она характеризуется внутригрупповыми дисперсиями $\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_{j2}))\}$. Для отдельной субпопуляции $\mathbf{s}_{j2}$ внутригрупповая дисперсия концентрации имеет вид

$$Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_{j2})) = E_{\mathbf{s}_1}\left\{x^2(\mathbf{s}_1)\middle|\mathbf{s}_{j2}\right\} - x^2(\mathbf{s}_{j2}) = \\ = \sum_{\mathbf{s}_1} x^2(\mathbf{s}_1))\, Pr(\mathbf{s}_1|\mathbf{s}_{j2}) - x^2(\mathbf{s}_{j2}).$$

Средней внутригрупповой дисперсией (средней дивергенцией концентрации фамилии по селам $\{\mathbf{s}_1\}$ внутри сельсовета $\mathbf{s}_2$ из района $\mathbf{s}_3$) будет

$$E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_3\} = \sum_{\mathbf{s}_2} Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))Pr(\mathbf{s}_2|\mathbf{s}_3) = \\ = E_{\mathbf{s}_2}\left\{E_{\mathbf{s}_1}\left\{x^2(\mathbf{s}_1)\middle|\mathbf{s}_2\right\} - x^2(\mathbf{s}_2)\middle|\mathbf{s}_3\right\} = \\ = E_{\mathbf{s}_2}\left\{E_{\mathbf{s}_1}\left\{x^2(\mathbf{s}_1)\middle|\mathbf{s}_2\right\}\middle|\mathbf{s}_3\right\} - E_{\mathbf{s}_2}\left\{x^2(\mathbf{s}_2)\middle|\mathbf{s}_3\right\} = \\ = E_{\mathbf{s}_1}\left\{x^2(\mathbf{s}_1)\middle|\mathbf{s}_3\right\} - E_{\mathbf{s}_2}\left\{x^2(\mathbf{s}_2)\middle|\mathbf{s}_3\right\},$$

так как $E_{\mathbf{s}_2}\{E_{\mathbf{s}_1}\{x^2(\mathbf{s}_1)|\mathbf{s}_2\}|\mathbf{s}_3\} = E_{\mathbf{s}_1}\{x^2(\mathbf{s}_1)|\mathbf{s}_3\}$ согласно (4).

Данные типы дисперсий для трех уровней иерархии обобщаются на случай метапопуляции $\mathbf{s}_m$ с m уровнями иерархии. На каждом уровне k находятся субпопуляции $\{\mathbf{s}_k\}$, вместе составляющие метапопуляцию $\mathbf{s}_m$ и объединяемые в группы, представляющие субпопуляции $\{\mathbf{s}_{jk+1}, j = 1, 2, ...\}$ следующего $(k + 1)$-го уровня иерархии, где j нумерует субпопуляции с уровнем иерархии $k + 1$. Распределение концентрации $x(\mathbf{s}_k)$ рассматриваемой фамилии внутри этих групп характеризуется средним значением $E_{\mathbf{s}_k}\{x(\mathbf{s}_k)|\mathbf{s}_{jk+1}\} = x(\mathbf{s}_{jk+1})$ и дисперсией $Vs(x(\mathbf{s}_k|\mathbf{s}_{jk+1}))$, т.е. обозначение $Vs_{in}(x(\mathbf{s}_k|\mathbf{s}_{jk+1}))$ относим к дисперсии распределения концентрации $x(\mathbf{s}_k)$ фамилии по субпопуляциям уровня $k < m$ внутри $\mathbf{s}_{jk+1}$. При этом можно использовать те же самые типы дисперсии, что и ранее в простейшем случае.

Под *полной* (*общей*) *дисперсией* распределения концентрации фамилии по всей метапопуляции $\mathbf{s}_m$ понимается дисперсия распределения концентрации по несгруппированным субпопуляциям $\{\mathbf{s}_1\}$ уровня единицы наблюдения (обычно относимым к первому уровню $\{\mathbf{s}_1\}$), т.е. для всей метапопуляции $\mathbf{s}_m$ при отсутствии в ней группировок. Эта дисперсия обозначается как $Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m))$. В качестве условно неподразделенной единицы наблюдения может быть выбрана популяция и более высокого уровня иерархии k, $1 < k < m$ (скажем, при отсутствии данных об уровнях ниже k, либо из научных интересов).

Дисперсия распределения концентрации фамилии по субпопуляциям (группам) $\{\mathbf{s}_k\}$ из $\mathbf{s}_m$

$$Vs_{betw}(x(\mathbf{s}_k|\mathbf{s}_m)) \equiv Vs(x(\mathbf{s}_k|\mathbf{s}_m))$$

называется *межгрупповой дисперсией на уровне* $k < m$. Она характеризует фамильную дивергенцию на данном уровне.

При подразделенности субпопуляций $\{\mathbf{s}_k\}$ j-я из них характеризуется своей *внутригрупповой дисперсией* $Vs(x(\mathbf{s}_{k-1}|\mathbf{s}_{jk})) = Vs_{in}(x(\mathbf{s}_{k-1}|\mathbf{s}_{jk}))$ распределения концентрации фамилии $\{x(\mathbf{s}_{k-1})\}$ по субпопуляциям $\{\mathbf{s}_{k-1}\}$ внутри $\mathbf{s}_{jk}$. Уровень k в целом характеризуется *средней внутригрупповой дисперсией* $Ws(x(\mathbf{s}_{k-1}|\mathbf{s}_{jk}))$:

$$Ws(x(\mathbf{s}_{k-1}|\mathbf{s}_{jk})) \equiv E_{\mathbf{s}_{jk}}\{Vs_{in}(x(\mathbf{s}_{k-1}|\mathbf{s}_{jk}))\} = \\ = E_j\{Vs_{in}(x(\mathbf{s}_{k-1}|\mathbf{s}_{jk}))\}, \quad (7)$$

т.е. средним значением для внутригрупповых дисперсий у субпопуляций уровня k. Можно также рассматривать разные уровни внутригрупповой дисперсии, соответствующие дисперсиям распределения концентраций по соответствующим субпопуляциям на уровне, меньшем k, вплоть до дисперсии концентраций по субпопуляциям первого уровня $\{\mathbf{s}_1\}$.

Отметим неоднозначность при многоуровневой иерархии таких понятий как межгрупповая и внутригрупповая изменчивость без указания уровня, с которым они соотносятся. Одна и та же дисперсия распределения концентрации фамилии, скажем, по сельсоветам является межгрупповой при анализе на уровне района и внутригрупповой на уровне области. В то же время рассматриваемая дисперсия определяется однозначно по случайному аргументу у Vs (стоящему перед вертикальной чертой). Таким образом, индексы *betw*, *in* у дисперсий условны и служат для облегчения ориентации в каком аспекте рассматривается соответствующая дисперсия в данном контексте, а аргумент у Vs универсален. Он указывает на случайную величину, дисперсией которой будет Vs. Статистический смысл дисперсии Vs не зависит от ин-

декса *betw* или *in* и определяется аргументом *Vs*, (напомним, что в обозначении $\mathbf{s}_i|\mathbf{s}_{jk}$ символ $\mathbf{s}_i$ перед вертикальной чертой рассматривается как случайная величина со значениями $\{\mathbf{s}_i\}$ из $\mathbf{s}_{jk}$, $i < k$).

В единицах наблюдения (обычно субпопуляциях первого уровня) по определению отсутствует или игнорируется подразделенность, поэтому у них невозможно определение внутригрупповой дисперсии. Наименьший уровень иерархии субпопуляции, в которой внутригрупповая дисперсия реально существует, равен двум. Субпопуляциями наименьшего уровня иерархии, в которых возможна *средняя* внутригрупповая дисперсия $E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_3\}$, будут трехуровневые метапопуляции $\{\mathbf{s}_3\}$. Повторим, что выражение средней внутригрупповой дисперсии $E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_3\}$ для $\mathbf{s}_3$ интерпретируется следующим образом. Аргумент $x(\mathbf{s}_1|\mathbf{s}_2)$ является случайной величиной, принимающей значения, равные концентрации фамилии в субпопуляции $\mathbf{s}_1$, наугад выбранной из $\mathbf{s}_2$. Эта субпопуляция $\mathbf{s}_2$ сама случайно выбирается из субпопуляции $\mathbf{s}_3$. Математическое ожидание дисперсии данной случайной величины $x(\mathbf{s}_1|\mathbf{s}_2)$ для распределения x по $\mathbf{s}_2$ по определению является средней внутригрупповой дисперсией $E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_3\}$.

Выражение средней внутригрупповой дисперсии $E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_{jk}\}$ для субпопуляции более высокого уровня $k > 3$ интерпретируется сходно. Аргумент $x(\mathbf{s}_1|\mathbf{s}_2)$ является случайной величиной, принимающей значения, равные концентрации фамилии в $\mathbf{s}_1$, наугад выбранной из $\mathbf{s}_2$, а субпопуляция $\mathbf{s}_2$ сама случайно выбирается из субпопуляции $\mathbf{s}_{jk}$. Математическое ожидание дисперсии данной случайной величины $x(\mathbf{s}_1|\mathbf{s}_2)$ по определению называется средней внутригрупповой дисперсией субпопуляции $\mathbf{s}_2$ в $\mathbf{s}_{jk}$.

В более общем случае многоуровневой иерархии вместо $\mathbf{s}_2$ можно взять $\mathbf{s}_i$ и определить среднюю внутригрупповую дисперсию для $\mathbf{s}_i$, на уровне i для субпопуляций внутри $\mathbf{s}_{jk}$, $i < k$. Средняя внутригрупповая дисперсияна уровне i определяется как $E_{\mathbf{s}_i}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_i))|\mathbf{s}_{jk}\}$, $1 < i < k$ для j-ой субпопуляции k-го уровня. Таким образом получаем среднее значение дисперсии распределения концентрации по неподразделенным и несгруппированным субпопуляциям первого уровня внутри субпопуляции $\mathbf{s}_i$ уровня i при случайном выборе $\mathbf{s}_i$ из $\mathbf{s}_{jk}$. Такую дисперсию $E_{\mathbf{s}_i}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_i))|\mathbf{s}_{jk}\} \equiv Ws(x(\mathbf{s}_1|\mathbf{s}_i)|\mathbf{s}_{jk})$ можно интерпретировать как ожидаемое значение дисперсии случайной величины $x(\mathbf{s}_1|\mathbf{s}_i)$, когда $\mathbf{s}_i$ наугад выбирается из $\mathbf{s}_{jk}$.

РАЗЛОЖЕНИЕ ДИСПЕРСИИ РАСПРЕДЕЛЕНИЯ КОНЦЕНТРАЦИИ ФАМИЛИИ ПО УРОВНЯМ ИЕРАРХИИ

Как известно из дисперсионного анализа (см., например, [12]), *сумма квадратов отклонений* значений признака от общепопуляционного среднего значения в популяции, подразделенной на группы, равна сумме межгрупповых и внутригрупповых сумм квадратов отклонений (называемых в [12] вариациями). Соответствующие *дисперсии* практически удовлетворяют тому же соотношению, но при замене “сумма квадратов отклонений” на “дисперсия”, а вместо “внутригрупповая дисперсия” будет “*средняя* внутригрупповая дисперсия”, т.е. добавляется термин “*средняя*”. В биометрии это соотношение известно как *правило сложения дисперсий*. Мне не удалось найти распространенный учебник биометрии, где оно приведено, но по поисковому запросу в Интернете появляется множество ссылок. Согласно этому правилу

в совокупности из нескольких групп произвольных объектов с каким-либо числовым признаком x общая (*полная*) *дисперсия x во всей совокупности равна сумме межгрупповой дисперсии* (*дисперсии распределения средних значений x в группах*) *и средней внутригрупповой дисперсии* (*средней дисперсии x внутри групп*).

Понятно, что согласно приведенной формулировке данное статистическое правило не зависит от природы объектов и принципов их объединения в группы, которые могут быть произвольными. Правило выполняется не только для дисперсий, но и при совместном изучении нескольких признаков для их матриц ковариаций. В любом случае это правило может использоваться для проверки безошибочности вычислений.

В случае анализа фамильной структуры данное правило принимает следующий вид.

Замечание 3 (*правило сложения дисперсий*). *Пусть подразделенная метапопуляция* $\mathbf{s}_3$ *разбита на субпопуляции* $\{\mathbf{s}_2\}$, *каждая из которых включает непересекающиеся неподразделенные группы* $\{\mathbf{s}_1\}$, *являющиеся единицами наблюдения с концентрациями рассматриваемой фамилии в них* $\{x(\mathbf{s}_1|\mathbf{s}_2)\}$.

Тогда полная (*общая*) *дисперсия* $Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_3))$ *распределения концентрации фамилии по несгруппированным единицам наблюдения* $\{\mathbf{s}_1\}$ *во всей метапопуляции* $\mathbf{s}_3$ *равна сумме межгрупповой дисперсии* $Vs_{betw}(x(\mathbf{s}_2|\mathbf{s}_3))$, *характеризующей фамильную дивергенцию средних значений концентраций* $\{x(\mathbf{s}_2)\}$ *у субпопуляций* $\{\mathbf{s}_2\}$, *и средней внутригрупповой дисперсии* $W(x(\mathbf{s}_1|\mathbf{s}_2)|\mathbf{s}_3) \equiv E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_3\}$, *характеризующей среднюю фамильную дивергенциюх* ($\mathbf{s}_1$) *внутри субпопуляций* $\{\mathbf{s}_2\}$:

$$Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_3)) = Vs_{betw}(x(\mathbf{s}_2|\mathbf{s}_3)) + W(x(\mathbf{s}_1|\mathbf{s}_2)|\mathbf{s}_3),$$
$$W(x(\mathbf{s}_1|\mathbf{s}_2)|\mathbf{s}_3) \equiv E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_3\}. \quad (8)$$

Доказательство. Рассмотрим сумму указанных дисперсий:

$$Vs_{betw}(x(\mathbf{s}_2|\mathbf{s}_3)) + E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_3\} =$$
$$= Vs_{betw}(x(\mathbf{s}_2|\mathbf{s}_3)) + E_{\mathbf{s}_2}\left\{E_{\mathbf{s}_1}\left\{x^2(\mathbf{s}_1)|\mathbf{s}_2\right\} - x^2(\mathbf{s}_2)|\mathbf{s}_3\right\} =$$
$$= \left(E_{\mathbf{s}_2}\left\{x^2(\mathbf{s}_2)|\mathbf{s}_3\right\} - x^2(\mathbf{s}_3)\right) +$$
$$+E_{\mathbf{s}_2}\left\{E_{\mathbf{s}_1}\left\{x^2(\mathbf{s}_1)|\mathbf{s}_2\right\} - x^2(\mathbf{s}_2)|\mathbf{s}_3\right\} =$$
$$= E_{\mathbf{s}_2}\left\{x^2(\mathbf{s}_2)|\mathbf{s}_3\right\} -$$
$$- x^2(\mathbf{s}_3) + E_{\mathbf{s}_1}\left\{x^2(\mathbf{s}_1)|\mathbf{s}_3\right\} - E_{\mathbf{s}_2}\left\{x^2(\mathbf{s}_2)|\mathbf{s}_3\right\} =$$
$$E_{\mathbf{s}_1}\left\{x^2(\mathbf{s}_1)|\mathbf{s}_3\right\} - x^2(\mathbf{s}_3) = Vs(x(\mathbf{s}_1|\mathbf{s}_3)),$$

так как $E_{\mathbf{s}_2}\{E_{\mathbf{s}_1}\{x^2(\mathbf{s}_1)|\mathbf{s}_2\}|\mathbf{s}_3\} = E_{\mathbf{s}_1}\{x^2(\mathbf{s}_1)|\mathbf{s}_3\}$ согласно (4), и после сокращения получаем общую дисперсию $Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_3)) = Vs(x(\mathbf{s}_1|\mathbf{s}_3))$. ◄

Данное правило также справедливо по отношению к соответствующим матрицам ковариаций.

Рассмотрим правило сложения дисперсий применительно к свойствам многоуровневых иерархически подразделенных метапопуляций. Пусть дана метапопуляция с m уровнями иерархии, в которой в качестве единицы наблюдения выбрана субпопуляция уровня n и выбран уровень иерархии k: $m > k > n$ (в рассмотренном выше случае $m = 3$, $k = 2$, $n = 1$). Очевидно, все множество субпопуляций $\{\mathbf{s}_k\}$ данного уровня (или любого другого) составляют метапопуляцию целиком. Условно считаем, что единицы наблюдения характеризуются только соответствующими концентрациями фамилии, внутригрупповая дисперсия в них либо отсутствует, либо сведений о ней не имеется, наконец, она может игнорироваться. Таким образом, мы допускаем произвольное количество уровней иерархии и *не требуем, чтобы выбранные уровни k, n и m были соседними*. Рассмотрим, как выглядит правило сложения дисперсий в этом случае.

Следствие 4. *Пусть в иерархически подразделенной метапопуляции $\mathbf{s}_m$ с m уровнями иерархии выбраны в качестве единицы наблюдения субпопуляции $\{\mathbf{s}_n\}$ уровня n и выбран уровень иерархии k, $m > k > n$ с непересекающимися субпопуляциями $\{\mathbf{s}_k\}$, вместе образующими $\mathbf{s}_m$. Тогда в $\mathbf{s}_m$ выполняется следующий вариант правила сложения дисперсий.*

Общая (полная) дисперсия $Vs_{tot}(x(\mathbf{s}_n|\mathbf{s}_m))$, понимаемая как дисперсия распределения по всей метапопуляции $\mathbf{s}_m$ концентрации x интересующей фамилии в субпопуляциях $\{\mathbf{s}_n\}$, рассматриваемых (условно) как единицы наблюдения, разлагается в сумму

1) *межгрупповой дисперсии $Vs_{betw}(x(\mathbf{s}_k|\mathbf{s}_m))$ распределения концентрации по составляющим $\mathbf{s}_m$ субпопуляциям $\{\mathbf{s}_k\}$ на произвольном уровне $k < m$ и*

2) *среднего значения $E_{\mathbf{s}_k}\{Vs_{in}(x(\mathbf{s}_n|\mathbf{s}_k))\}$ внутригрупповой дисперсии $Vs_{in}(x(\mathbf{s}_n|\mathbf{s}_k))$ распределения концентрации фамилии $\{x(\mathbf{s}_n|\mathbf{s}_k)\}$ по субпопуляциям-единицам наблюдения $\{\mathbf{s}_n\}$ внутри субпопуляций $\{\mathbf{s}_k\}$ уровня k ($n < k$):*

$$Vs_{tot}(x(\mathbf{s}_n|\mathbf{s}_m)) = Vs_{betw}(x(\mathbf{s}_k|\mathbf{s}_m)) + E_{\mathbf{s}_k}\{Vs_{in}(x(\mathbf{s}_n|\mathbf{s}_k))|\mathbf{s}_m\}. \quad (9)$$

Согласно данному правилу эти дисперсии зависимы, и по значениям любой пары дисперсий определяется значение третьей.

Доказательство. Распишем межгрупповую и внутригрупповую дисперсии по формуле (6) как разности среднего квадрата и квадрата среднего. Например, для случайной величины $x(\mathbf{s}_n|\mathbf{s}_k)$ среднее значение будет равно $x(\mathbf{s}_k)$ для соответствующей субпопуляции $\mathbf{s}_k$ согласно (3), а (внутригрупповой) дисперсией будет

$$Vs_{in}(x(\mathbf{s}_n|\mathbf{s}_k)) = E_{\mathbf{s}_n}\left\{x^2(\mathbf{s}_n)|\mathbf{s}_k\right\} -$$
$$- (E_{\mathbf{s}_n}\left\{x(\mathbf{s}_n)|\mathbf{s}_k\right\}^2 = E_{\mathbf{s}_n}\left\{x^2(\mathbf{s}_n)|\mathbf{s}_k\right\} - x^2(\mathbf{s}_k).$$

Здесь под $\mathbf{s}_k$ подразумевается какая-либо конкретная субпопуляция уровня k (скажем, $\mathbf{s}_{jk}$ с номером j).

Межгрупповая дисперсия имеет вид $Vs_{betw}(x(\mathbf{s}_k|\mathbf{s}_m)) = E_{\mathbf{s}_k}\{x^2(\mathbf{s}_k)|\mathbf{s}_m\} - x^2(\mathbf{s}_m)$.

Найдем сумму межгрупповой и *средней* внутригрупповой дисперсий:

$$Vs_{betw}(x(\mathbf{s}_k|\mathbf{s}_m)) + E_{\mathbf{s}_k}\{Vs_{in}(x(\mathbf{s}_n|\mathbf{s}_k))|\mathbf{s}_m\} =$$
$$= \left(E_{\mathbf{s}_k}\left\{x^2(\mathbf{s}_k)|\mathbf{s}_m\right\} - x^2(\mathbf{s}_m)\right) +$$
$$+ E_{\mathbf{s}_k}\left\{E_{\mathbf{s}_n}\left\{x^2(\mathbf{s}_n)|\mathbf{s}_k\right\} - x^2(\mathbf{s}_k)|\mathbf{s}_m\right\} =$$
$$= E_{\mathbf{s}_k}\left\{x^2(\mathbf{s}_k)|\mathbf{s}_m\right\} - x^2(\mathbf{s}_m) +$$
$$+ E_{\mathbf{s}_n}\left\{x^2(\mathbf{s}_n)|\mathbf{s}_m\right\} - E_{\mathbf{s}_k}\left\{x^2(\mathbf{s}_k)|\mathbf{s}_m\right\} =$$
$$= E_{\mathbf{s}_n}\left\{x^2(\mathbf{s}_n)|\mathbf{s}_m\right\} - x^2(\mathbf{s}_m) = Vs(x(\mathbf{s}_n|\mathbf{s}_m)),$$

поскольку $E_{\mathbf{s}_k}\{E_{\mathbf{s}_n}x^2(\mathbf{s}_n)|\mathbf{s}_k\}|\mathbf{s}_m\} = E_{\mathbf{s}_n}\{x^2(\mathbf{s}_n)|\mathbf{s}_m\}$ в соответствии с (4), а после сокращения получаем $E_{\mathbf{s}_n}\{x^2(\mathbf{s}_n)|\mathbf{s}_m\} - x^2(\mathbf{s}_m)$, т.е. полную дисперсию $Vs_{tot}(x(\mathbf{s}_n|\mathbf{s}_m))$. ◄

Понятно, что значение полной дисперсии зависит от выбора единицы наблюдения, и дисперсия будет наибольшей при неподразделенной единице (не содержащей субпопуляций). Например, когда единицей наблюдения является сельсовет и соответственно имеется информация только по концентрации фамилии в сельсоветах $\{\mathbf{s}_2\}$, то общая дисперсия концентрации фамилии в $\mathbf{s}_3$ совпадает с межгрупповой дисперсией распределения концентрации по сельсоветам. При этом выпадает неотрицательный вклад фамильной дивергенции сел внутри сельсоветов, т.е. средний внутригрупповой компонент $E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))\}$. Это может привести к существенному уменьшению полной дисперсии, следовательно к занижению показателя дивергенции субпопуляций внутри района и вытекающему отсюда уменьшению

оценки (см., например, [1, 2]) коэффициента инбридинга популяции по фамильным данным, максимальной при единице наблюдения минимального уровня. Когда субпопуляции $\{\mathbf{s}_{m-1}\}$ являются единицами наблюдения, то межгрупповая дисперсия совпадает с полной.

Теперь обобщим рассмотренное правило. Покажем, что в случае иерархической подразделенности метапопуляции с произвольным количеством уровней иерархии полная дисперсия распределения концентрации фамилии разлагается не только на межгрупповую и внутригрупповую дисперсии, но последняя еще разлагается на компоненты, соответствующие отдельным уровням.

Результат 5 (*разложение полной дисперсии*). *В иерархически подразделенной метапопуляции* $\mathbf{s}_m$ *с уровнями иерархии* $i = 1, 2, \ldots, m$ *полная* (*общая*) *дисперсия* $Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m))$, *т.е. дисперсия распределения во всей подразделенной метапопуляции* $\mathbf{s}_m$ *концентрации рассматриваемой фамилии* (*по несгруппированным и неподразделенным субпопуляциям* $\{\mathbf{s}_1\}$ *уровня единицы наблюдения*), *разлагается в сумму*

1) *межгрупповой дисперсии* $Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m))$ *распределения концентрации по субпопуляциям* $\{\mathbf{s}_{m-1}\}$ *в* $\mathbf{s}_m$ *и*

2) *средней внутригрупповой дисперсии* $E_{\mathbf{s}_{m-1}}\{Vs_{in}(x(\mathbf{s}_1))|\mathbf{s}_{m-1}\}$ *распределения концентрации по единицам наблюдения* $\{\mathbf{s}_1\}$ *внутри отдельных субпопуляций* $\mathbf{s}_{m-1}$;

3) *эта средняя внутригрупповая дисперсия* $E_{\mathbf{s}_{m-1}}\{Vs_{in}(x(\mathbf{s}_1))|\mathbf{s}_{m-1}\}$, *в свою очередь, разлагается на сумму средних внутригрупповых дисперсий* $E_{\mathbf{s}_{i+1}}\{Vs_{in}(x(s_i)|\mathbf{s}_{i+1})|\mathbf{s}_m\}$, *соответствующих отдельным уровням иерархии* $\{i\}$, *т.е.*

$$\begin{aligned} Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m)) &= Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m)) + \\ &+ E_{\mathbf{s}_{m-1}}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_{m-1}))|\mathbf{s}_m\} = \\ = Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m)) &+ E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_m\} + \\ &+ E_{\mathbf{s}_3}\{Vs_{in}(x(\mathbf{s}_2|\mathbf{s}_3))|\mathbf{s}_m\} + \\ +\ldots &+ E_{\mathbf{s}_{m-1}}\{Vs_{in}(x(\mathbf{s}_{m-2}|\mathbf{s}_{m-1}))|\mathbf{s}_m\} = \\ = Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m)) &+ \sum_{i=1}^{m-2} E_{\mathbf{s}_{i+1}}\{Vs_{in}(x(\mathbf{s}_i|\mathbf{s}_{i+1}))|\mathbf{s}_m\}. \end{aligned} \quad (10)$$

Доказательство. Рассмотрим иерархически подразделенную метапопуляцию $\mathbf{s}_m$ с m уровнями иерархии. Непересекающиеся субпопуляции каждого отдельного уровня i образуют всю метапопуляцию $\mathbf{s}_m$ (дают ее разбиение на группы). Положим, что в качестве единицы наблюдения выбраны субпопуляции первого уровня $\{\mathbf{s}_1\}$, а затем второго $\{\mathbf{s}_2\}$, и рассмотрим субпопуляции $\{\mathbf{s}_{m-1}\}$ уровня $m - 1 > 2$. Полные (общие) дисперсии $Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m))$ и $Vs_{tot}(x(\mathbf{s}_2|\mathbf{s}_m))$ концентрации фамилии в $\mathbf{s}_m$, когда единицей наблюдения выбраны субпопуляции $\mathbf{s}_1$ и $\mathbf{s}_2$ соответственно, выражаются согласно правилу сложения дисперсий в виде (9) при $n = 1$, $i = m - 1$ как

$$\begin{aligned} Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m)) &= Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m)) + \\ &+ E_{\mathbf{s}_{m-1}}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_{m-1}))|\mathbf{s}_m\}, \\ Vs_{tot}(x(\mathbf{s}_2|\mathbf{s}_m)) &= Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m)) + \\ &+ E_{\mathbf{s}_{m-1}}\{Vs_{in}(x(\mathbf{s}_2|\mathbf{s}_{m-1}))|\mathbf{s}_m\}. \end{aligned}$$

Проанализируем как изменилась полная дисперсия концентрации в $\mathbf{s}_m$ в результате изменения единицы наблюдения. Для этого найдем разность Δ_{12} приведенных дисперсий, где сократим члены $Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m))$, распишем $Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_{m-1}))$ согласно (9) при $n = 1$, $i = 2$ и произведем дальнейшие сокращения:

$$\begin{aligned} \Delta_{12} &\equiv Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m)) - Vs_{tot}(x(\mathbf{s}_2|\mathbf{s}_m)) = \\ &= Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m)) + E_{\mathbf{s}_{m-1}}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_{m-1}))|\mathbf{s}_m\} - \\ &- \left(Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m)) + E_{\mathbf{s}_{m-1}}\{Vs_{in}(x(\mathbf{s}_2|\mathbf{s}_{m-1}))|\mathbf{s}_m\}\right) = \\ &= E_{\mathbf{s}_{m-1}}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_{m-1}))|\mathbf{s}_m\} - \\ &- E_{\mathbf{s}_{m-1}}\{Vs_{in}(x(\mathbf{s}_2|\mathbf{s}_{m-1}))|\mathbf{s}_m\}. \end{aligned}$$

Подставим сюда $Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_{m-1})) = Vs(x(\mathbf{s}_2|\mathbf{s}_{m-1}) + E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_m\}$ согласно (9) при $n = 1$, $i = 2$ и учтем, что по (4) $E_{\mathbf{s}_{m-1}}\{E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))\}|\mathbf{s}_m\} = E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_m\}$:

$$\begin{aligned} \Delta_{12} &= E_{\mathbf{s}_{m-1}}\{Vs(x(\mathbf{s}_2|\mathbf{s}_{m-1}))|\mathbf{s}_m\} + \\ &+ E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_m\} - \\ &- E_{\mathbf{s}_{m-1}}\{Vs_{in}(x(\mathbf{s}_2|\mathbf{s}_{m-1}))|\mathbf{s}_m\} = \\ &= E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_m\} \geq 0. \end{aligned}$$

Точно так же находим, что изменением дисперсии концентрации фамилии в $\mathbf{s}_m$ при переходе от уровня единицы наблюдения $\mathbf{s}_2$ к $\mathbf{s}_3$ будет

$$\begin{aligned} \Delta_{23} &\equiv Vs_{tot}(x(\mathbf{s}_2|\mathbf{s}_m)) - \\ &- Vs_{tot}(x(\mathbf{s}_3|\mathbf{s}_m)) = E_{\mathbf{s}_3}\{Vs(x(\mathbf{s}_2|\mathbf{s}_3)) \geq 0. \end{aligned}$$

Отсюда

$$\begin{aligned} \Delta_{12} + \Delta_{23} &= (Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m)) - Vs_{tot}(x(\mathbf{s}_2|\mathbf{s}_m))) + \\ &+ (Vs_{tot}(x(\mathbf{s}_2|\mathbf{s}_m)) - Vs_{tot}(x(\mathbf{s}_3|\mathbf{s}_m))) = \\ &= Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m)) - Vs_{tot}(x(\mathbf{s}_3 \mid \mathbf{s}_m)). \end{aligned}$$

Следовательно,

$$\begin{aligned} Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m)) &= Vs_{tot}(x(\mathbf{s}_3|\mathbf{s}_m)) + \Delta_{12} + \Delta_{23} = \\ &= Vs_{tot}(x(\mathbf{s}_3|\mathbf{s}_m)) + E_{\mathbf{s}_2}\{Vs(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_m\} + \\ &+ E_{\mathbf{s}_3}\{Vs(x(\mathbf{s}_2|\mathbf{s}_3)). \end{aligned}$$

Аналогично при переходе от единицы наблюдения уровня $\mathbf{s}_i$ к уровню $i + 1$ изменение дисперсии будет равно $E_{\mathbf{s}_{i+1}}\{Vs(x(\mathbf{s}_i|\mathbf{s}_{i+1}))|\mathbf{s}_m\}$. В результате последовательности $j - 1$ таких шагов общая дисперсия при единице наблюдения $\mathbf{s}_1$ выражается через общую дисперсию при единице наблюдения $\mathbf{s}_j$ как

$$Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m)) = Vs_{tot}(x(\mathbf{s}_j|\mathbf{s}_m)) + \\ + \sum_{i=1}^{j-1} E_{\mathbf{s}_{i+1}}\{Vs(x(\mathbf{s}_i|\mathbf{s}_{i+1}))\mathbf{s}_m\}.$$

Когда на последнем шагу j равно $m-1$, получаем полную дисперсию в виде следующего разложения:

$$Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m)) = Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m)) + \\ + \sum_{i=1}^{m-2} E_{\mathbf{s}_{i+1}}\{Vs_{in}(x(\mathbf{s}_i|\mathbf{s}_{i+1}))\},$$

где $Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m)) = Vs_{tot}(x(\mathbf{s}_{m-1}|\mathbf{s}_m)) = Vs(x(\mathbf{s}_{m-1}|\mathbf{s}_m))$, так как значение дисперсии не зависит от индексов *betw* или *tot*. ◄

Таким образом, разложение полной дисперсии концентрации фамилии в случае иерархически подразделенной трехуровневой метапопуляции равно сумме межгрупповой и внутригрупповой дисперсий и совпадает с правилом сложения дисперсий (8). В случае метапопуляции с четырьмя уровнями иерархии разложение полной дисперсии имеет вид

$$Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_4)) = Vs_{betw}(x(\mathbf{s}_3|\mathbf{s}_4)) + \\ + E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_4\} + \\ + E_{\mathbf{s}_3}\{Vs_{in}(x(\mathbf{s}_2|\mathbf{s}_3))|\mathbf{s}_4\},$$

а при пяти уровнях

$$Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_5)) = Vs_{betw}(x(\mathbf{s}_4|\mathbf{s}_5)) + \\ + E_{\mathbf{s}_2}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_2))|\mathbf{s}_5\} + E_{\mathbf{s}_3}\{Vs_{in}(x(\mathbf{s}_2|\mathbf{s}_3))|\mathbf{s}_5\} + \\ + E_{\mathbf{s}_4}\{Vs_{in}(x(\mathbf{s}_3|\mathbf{s}_4))|\mathbf{s}_5\}.$$

Здесь видно, что каждому уровню иерархии i соответствует вклад в общую дисперсию, равный $E_{\mathbf{s}_{i+1}}\{Vs_{in}(x(\mathbf{s}_i|\mathbf{s}_{i+1}))\}$.

Напомним об условности термина межгрупповая дисперсия. Так, например, межгрупповая дисперсия $V(x(\mathbf{s}_{m-1}|\mathbf{s}_m)$ является внутригрупповой для всей метапопуляции $\mathbf{s}_m$. Поэтому можно сформулировать доказанный результат следующим образом. *Дисперсия распределения во всей подразделенной метапопуляции* $\mathbf{s}_m$ *концентрации рассматриваемого аллеля по несгруппированным и неподразделенным субпопуляциям* $\{\mathbf{s}_1\}$ *уровня единицы наблюдения, разлагается в сумму средних внутригрупповых дисперсий* $E_{\mathbf{s}_{i+1}}\{V_{in}(x(s_i))|\mathbf{s}_{i+1}\}$, *соответствующих отдельным уровням иерархии*.

Ремарка 6. Доказанный результат остается верным, если заменить $m-1$ на любой другой уровень иерархии i $(1 < i < m-1)$.

Следствие 7. *Если в иерархически подразделенной метапопуляции* $\mathbf{s}_m$ *с m уровнями иерархии в качестве единицы наблюдения выбраны субпопуляции уровня* $1 < n < m-1$, *то*

1) *из полной (общей) дисперсии* $Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m))$ *распределения концентрации фамилии* (10) *по неподразделенным субпопуляциям уровня единицы наблюдения* $\{\mathbf{s}_1\}$ *выпадают неотрицательные вклады в дивергенцию субпопуляций, соответствующие уровням ниже n* (*внутригрупповые дисперсии* $E_{\mathbf{s}_{i+1}}\{Vs_{in}(x(\mathbf{s}_i|\mathbf{s}_{i+1}))|\mathbf{s}_m\}$), *суммарно равные* $\sum_{i=1}^{n-1} E_{\mathbf{s}_{i+1}}\{Vs_{in}(x(\mathbf{s}_i|\mathbf{s}_{i+1}))\}$. *В итоге роль общей дисперсии при единице наблюдения* $\{\mathbf{s}_n\}$ *играет*

$$Vs_{tot}(x(\mathbf{s}_n|\mathbf{s}_m)) = Vs_{betw}(x(\mathbf{s}_{m-1}|\mathbf{s}_m)) + \\ + \sum_{i=n}^{m-2} E_{\mathbf{s}_{i+1}}\{Vs_{in}(x(\mathbf{s}_i|\mathbf{s}_{i+1}))|\mathbf{s}_m\}).$$

При увеличении уровня иерархии у единицы измерения от n до $n+1 < i$ *полная дисперсия уменьшается на* $E_{\mathbf{s}_{n+1}}\{Vs_{in}(x(\mathbf{s}_n|\mathbf{s}_{n+1}))|\mathbf{s}_m\}$;

2) *полная дисперсия* $Vs_{tot}(x(\mathbf{s}_n|\mathbf{s}_m))$, *межгрупповая* $Vs(x(\mathbf{s}_i|\mathbf{s}_{i+1}))$ *и внутригрупповая* $E_{\mathbf{s}_i}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_i))|\mathbf{s}_m\}$ *дисперсии концентрации рассматриваемой фамилии, соответствующие уровню* $i < m$, *связаны соотношением*

$$Vs(x(\mathbf{s}_i|\mathbf{s}_{i+1})) = Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m)) - \\ - E_{\mathbf{s}_i}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_i))|\mathbf{s}_m\},$$

откуда по значениям любых двух дисперсий можно найти величину третьей;

3) *при случайном формировании субпопуляций следующего уровня* $i+1$ *из групп на предыдущем i межгрупповая дисперсия не возрастает* (*уменьшается*):

$$Vs(x(\mathbf{s}_{i-1}|\mathbf{s}_i)) \geq Vs(x(\mathbf{s}_i|\mathbf{s}_{i+1})).$$

Доказательство опирается на полученные ранее результаты.

1. Обоснование п. 1 совпадает с приведенным выше, просто уровнем единицы наблюдения будет не первый, а n-й. Изменение полной дисперсии при увеличении уровня иерархии у единицы измерения следует из формулы (10) разложения Vs_{tot}.

2. Напомним, что множество субпопуляций на каждом отдельном уровне i дает всю иерархически подразделенную метапопуляцию $\mathbf{s}_m$, и каждая из этих субпопуляций состоит из единиц наблюдения $\{\mathbf{s}_1\}$. Полная дисперсия согласно (9) представима при $n=1$, $k=i$ как

$$Vs(x(\mathbf{s}_1|\mathbf{s}_m)) = Vs(x(\mathbf{s}_i|\mathbf{s}_m)) + E_{\mathbf{s}_i}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_i))\},$$

т.е. равна сумме межгрупповой и средней внутригрупповой дисперсий. Отсюда по величине любых двух дисперсий можно найти значение третьей.

При заданной полной дисперсии увеличение (уменьшение) межгрупповой $Vs(x(\mathbf{s}_i|\mathbf{s}_m))$ или внутригрупповой $E_{\mathbf{s}_i}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_i))|\mathbf{s}_m\}$ дисперсий концентрации рассматриваемой фамилии, соответствующие уровню $i < m$, связаны соотношением

$$Vs(x(\mathbf{s}_i|\mathbf{s}_m)) = Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m)) - \\ - E_{\mathbf{s}_i}\{Vs_{in}(x(\mathbf{s}_1|\mathbf{s}_i))|\mathbf{s}_m\},$$

где $Vs_{tot}(x(\mathbf{s}_1|\mathbf{s}_m))$ – полная дисперсия распределения концентрации фамилии по неподразделенным субпопуляциям уровня единицы наблюдения $\{\mathbf{s}_1\}$. Поэтому увеличение одной из них связано с уменьшением в той же степени величины другой.

3. При переходе к более высокому уровню иерархии межгрупповая дисперсия не возрастает (уменьшается) при случайном формировании групп. Чтобы избежать доказательства с громоздкими выкладками при гипергеометрическом распределении в выборках без возвращения, просто укажем на интуитивное ожидание данного свойства. Оно опирается на очевидное уменьшение размаха изменчивости при усреднении в группе, что приводит к устранению крайних вариантов. ◄

Правило сложения дисперсий верно не только для совокупности субпопуляций, но и для совокупности **s** произвольных объектов (единиц наблюдения) при ее разбиении на любые непересекающиеся группы $\{\mathbf{s}_{ji}\}$. Пусть каждый объект характеризуется значением x некоторого числового признака. В частности, объектом может быть индивидуум, а признаком его вес, разные группы, скажем, состоят из индивидуумов с разными типами питания. Напомним, что когда единицей наблюдения является индивидуум, а вместо индивидуальных данных используются села, характеризуемые средним весом жителей, то изучение изменчивости веса детализируется до дисперсии среднего веса, а индивидуальная изменчивость остается как бы “за кадром”. Чем мельче единица наблюдения и ее уровень иерархии, тем полнее охват изменчивости. При фамильных исследованиях выбор в качестве единицы наблюдения административного района означает уменьшение общей фамильной дивергенции на дивергенцию на уровнях сел и сельсоветов.

ОБСУЖДЕНИЕ

Для (мета)популяций человека типична иерархическая подразделенность на части (субпопуляции), соответствующие классификации на базе административно-территориального деления, скажем, село, сельсовет, район, область и т.д.; на основе генеалогического подхода, базирующегося на этногенезе; или на использовании других принципов биологической классификации. Каждый уровень иерархии представляет собой разбиение метапопуляции на непересекающиеся субпопуляции, суммарно составляющие всю ее и обладающие, в свою очередь, иерархической структурой подразделенности. Данной структуре как таковой присущи специфические свойства изменчивости количественных признаков ее частей, независимые от природы иерархической системы (например, эти свойства будут и у системы из неживых объектов) и от факторов ее формирования (скажем, от миграций).

Здесь важно исследование роли единицы наблюдения на дивергенцию количественных признаков частей произвольной иерархической системы не обязательно биологической природы. Сама единица допускается любой, лишь бы была на одном из уровней иерархии рассматриваемой системы. В частности, системой может быть метапопуляция из субпопуляций село, сельсовет и так далее, а полученные результаты приложимы к такому признаку, как концентрация фамилии в популяциях человека, когда предметом изучения служит распределение концентрации фамилии по субпопуляциям системы.

Особый интерес представляет собой дисперсия распределения концентрации фамилии внутри и между субпопуляциями как характеристика фамильной дивергенции в метапопуляции. При иерархической подразделенности на каждом уровне иерархии будут в общем случае свои среднее значение и дисперсия концентрации, которая отражает фамильную дивергенцию субпопуляций на соответствующем уровне. Изучение фамильной дивергенции важно, так как при соответствующих предположениях она отражает генетическую дивергенцию и генетическую структуру метапопуляции.

Проведенное исследование показывает, что каждому отдельному уровню иерархии соответствует неотрицательный вклад в полную (общую) дисперсию концентрации в системе, равный среднему значению внутригрупповой дисперсии распределения концентрации фамилии по его субпопуляциям. Получено разложение общей дисперсии концентрации фамилии в метапопуляции по вкладам отдельных уровней, обобщающее правило сложения дисперсий. Отсюда находится величина занижения общей дисперсии, когда вместо неподразделенных субпопуляций первого уровня иерархии (допустим, сел) в качестве единиц наблюдения служат субпопуляции более высокого уровня (скажем, районов). Это позволяет судить о степени занижения оценки генетической дивергенции в метапопуляции в результате игнорирования фамильной изменчивости на низких уровнях иерархии.

Приложение найденных результатов к реальным данным наталкивается на определенные трудности. Опишем одну из них и соответствующее направление дальнейших исследований. Например, рассмотрим проблемы, связанные с нарушением иерархической структуры. Так, город не является объединением непересекающихся сельских субпопуляций, скажем, уровня сел или сельсоветов, т.е. использование объединенных данных по городам и сельским субпопуляциям нарушает иерархический характер подразделенности метапопуляции. Однако при отдельном изучении фамильной структуры городов (сел) наблюдается их иерархическая группировка согласно административным образованиям. В настоящее время городское на-

селение России составляет порядка 74.56%, а сельское только 25.44%. Требуются дополнительные исследования в отношении способов объединения получаемых таким образом фамильных данных с учетом разного вклада указанных компонентов. Соответственно необходим обоснованный метод использования получаемых в итоге результатов для выводов относительно генетической структуры метапопуляции.

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. *Пасеков В.П.* К анализу случайных процессов изонимии. I. Структура изонимии // Генетика. 2021. Т. 57. № 10. С. 1194–1204. https://doi.org/10.31857/S001667582110009X
2. *Пасеков В.П.* К анализу случайных процессов изонимии. II. Динамика дивергенции популяций // Генетика. 2021. Т 57. № 11. С. 1318–1329. https://doi.org/10.31857/S0016675821110114
3. *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.
4. *Lasker W.G.* Surnames and Genetic Structure. Cambridge: Cambr. Univ. Press, 1985. 2005. 148 p.
5. *Ревазов А.А., Парадеева Г.М., Русакова Г.И.* Пригодность русских фамилий в качестве квазигенетического маркера // Генетика. 1986. Т. 22. № 4. С. 699–703.
6. *Гинтер Е.К., Зинченко Р.А., Ельчинова Г.И. и др.* Роль факторов популяционной динамики в распространении наследственной патологии в российских популяциях // Мед. генетика. 2004. Т. 3. № 12. С. 548–555.
7. *Балановская Е.В., Сорокина И.Н., Чурносов М.И.* Описание “генетического ландшафта” районных популяций Центральной России // Вестник новых медицинских технологий. 2007. Т. 10. № 1.
8. *Сорокина И.Н., Чурносов М.И., Балтуцкая И.В. и др.* Антропогенетическое изучение населения Центральной России. М.: Изд-во РАМН, 2014. 336 с.
9. *Сорокина И.Н., Рудых Н.А., Крикун Е.Н., Сокорев С.Н.* Применение фамилий в популяционно-генетических исследованиях (на примере зарубежных популяций) // Науч. ведомости БелГУ. Сер. Медицина. Фармация. 2016. № 19(240). Вып. 35. С. 5–10.
10. *Rogers A.R.* Doubts about isonymy // Human Biology. 1991. V. 63. № 5. P. 663–668.
11. *Свирежев Ю.М., Пасеков В.П.* Основы математической генетики. М.: Наука, 1982. 511 с.
12. *Гланц С.* Медико-биологическая статистика. М.: Практика, 1998. 459 с.

Description of Divergence of Subpopulations in the Hierarchical System under the Analysis of Isonymy. I. Variance as an Indicator of Divergence

V. P. Passekov*

Dorodnicyn Computing Centre, Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia

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

We consider (typical for human) metapopulations with a hierarchical subdivision into parts (subpopulations) corresponding to the classification of subpopulations on the basis of administrative-territorial division (for example, a village, a village council, a district, a region, and so on); on the basis of genealogical approach grounded in ethnogenesis; and also on other principles of biological classification. Each level of the hierarchy is a partition of the metapopulation into nonintersecting subpopulations, which in total make up all of it and, in turn, have a hierarchical structure of partion. The properties of variability of the quantitative trait of subpopulations under a hierarchical structure are studied using the example of such a trait as the concentration of some surname. The distribution of the concentration of the surname over subpopulations is analyzed, it is characterized at each level by its mean value and variance, which reflects the surname divergence of subpopulations at the corresponding level. The study of surname divergence is important, since, under appropriate assumptions, it reflects the genetic divergence and the genetic structure of the metapopulation. It is shown that each separate level of the hierarchy corresponds to a non-negative contribution to the total variance, equal to the average value of the intragroup variance of the distribution of the concentration of the surname by its subpopulations. The decomposition of the total variance of the concentration of the surname in the metapopulation by the contributions of individual levels is obtained, generalizing the rule for addition of variances. The value of the underestimation of the total variance is found, when, instead of unsubdivided subpopulations of the first level of the hierarchy (say, villages), subpopulations of a higher level (say, districts) serve as observation units. This makes it possible to judge the degree of underestimation of the assessment of genetic divergence in the metapopulation as a result of ignoring the surname variability at the low levels of the hierarchy. The entire population is divided into two components with a hierarchical structure of subdivision: rural and urban residents. The results of this work are equally applicable to each of them.

Keywords: hierarchical structure of populations, metapopulations, concentration of surname in human subpopulations, decomposition of surname concentration variance by hierarchy levels.