Малов Сергей Васильевич,
Шевченко Андрей Константинович,
О'Брайен Стефан Джеймс

УДК 57.087.1, 519.2

ПОИСК ГЕНЕТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ. ЧАСТЬ 1. СТАТИСТИЧЕСКИЕ МЕТОДЫ*

Аннотация

В работе всесторонне изучается методология полногеномного поиска связей фенотипа с одним или несколькими генетическими маркерами. В данной части работы рассмотрены различные типы данных, наиболее часто встречающиеся при проведении генетических исследований, и методы их анализа.

Ключевые слова: полногеномный поиск закономерностей, GWAS, категориальные данные, данные типа времени жизни, модель Кокса, короткие временные ряды, обобщенные линейные модели.

ВВЕДЕНИЕ

В середине XX века было установлено, что вся информация о структуре рибонуклеиновых кислот (РНК) и белков, определяющих свойства всех известных организмов, закодирована в молекуле дезоксирибонуклеиновой кислоты (ДНК) [23], которая находится в каждой клетке. На самом деле в каждой клетке имеется пара комплементарных цепочек ДНК, образующих двойную спираль [40]. За это открытие в 1962 году работа Френсиса Крика и Джеймса Уотсона, впервые предложивших модель двойной спирали, была отмечена Нобелевской премией по физиологии и медицине. Генетическая информация в молекуле ДНК кодируется последовательностью четырех видов нуклеотидов, обозначаемых буквами (символами) «A», «G», «T», «C» по первым буквам названий четырех видов азотистых оснований, входящих в состав соответствующих нуклеотидов.

Длина последовательности ДНК (генома) зависит от типа организма. Скажем, у вирусов длина последовательности ДНК обычно имеет порядок 10^3 пар оснований, у бактерии E. coli – примерно $4{,}6 \cdot 10^6$ пар оснований, у человека – около $3{,}2 \cdot 10^9$ пар оснований, у определенных видов деревьев она исчисляется $3 \cdot 10^{10}$ парами оснований, а у некоторых видов амеб она достигает $6{,}7 \cdot 10^{11}$ пар оснований (Amoeba dubia). Следует заметить, что размер генома не всегда коррелирует со «сложностью» организма и его эволюционным положением.

В клетках, молекулы ДНК особым образом уложены в структуры, называемые хромосомами. Разным видам свойственно различное количество хромосом. Для человека характер-

* Работа поддержана мегагрантом правительства Российской Федерации № 11.G34.31.0068.

но наличие 22-х пар соматических хромосом (аутосом) и двух половых хромосом. У женского организма половые хромосомы являются парными (XX), а у мужского – одна из половых хромосом короче (XY).

Технологии «прочтения» (секвенирования) последовательности ДНК постоянно развиваются, но на сегодняшний день не существует эффективной технологии расшифровки всей последовательности одной молекулы, она собирается с помощью различных алгоритмов из десятков, сотен миллионов, а иногда и миллиардов коротких последовательностей (ридов), полученных с той или иной секвенирующей платформы. Кроме того, в процессе «прочтения» и сборки генома сложно исключить возможность появления ошибок в конечной последовательности. С другой стороны, сейчас уже существуют технологии, позволяющие собирать геном одной единственной исследуемой клетки (single-cell sequencing) многоклеточного или одноклеточного организма.

Последовательность ДНК может меняться под действием физических, химических или биологических факторов. Такие изменения последовательности ДНК принято называть мутациями. Различают три типа простых мутаций: замена одного нуклеотида на другой, вставка и удаление одного или нескольких нуклеотидов. Мутации, происходящие в клетках многоклеточного организма, приводят к различиям в генетическом материале разных клеток. Мутации происходят постоянно, однако в случае многоклеточных организмов, большинство мутаций не оказывает влияния на эволюцию, так как касается только одного организма. Только мутации, происходящие в половых клетках, могут закрепиться в потомстве и влиять на ход эволюции. В результате в популяции могут появляться два или более альтернативных варианта рассматриваемого участка ДНК, называемые аллелями. Вариации в одной позиции исследуемого участка ДНК (локуса), закрепившиеся в популяции, получили название однонуклеотидных полиморфизмов[1] (ОНП). Вероятность закрепления в популяции более двух вариантов ОНП в одной позиции крайне мала, поэтому все известные на текущий момент ОНП имеют по два варианта. Природа возникновения вставок и удалений более сложная, но иногда все-таки удается выявить бинарный маркер, который также приравнивают к ОНП. На текущий момент в геноме человека выявлено около 15 миллионов ОНП.

С учетом постоянной изменчивости абсолютно точная расшифровка последовательности ДНК не является практически актуальной задачей. Имеет смысл говорить о классах последовательностей ДНК, характерных для определенной группы организмов. В частности, можно говорить о последовательности ДНК млекопитающих, ДНК человека (геноме) или ДНК индивида (генотипе). Последовательности ДНК однотипных организмов (генотипы) обычно кодируются с помощью маркеров, поэтому генотип естественно идентифицировать с соответствующей последовательностью маркеров. Простейшие маркеры – ОНП и (или) приравненные к ним бинарные варианты мутаций типа удаления или вставки.

Классически, ге́ном называется последовательность нуклеотидов (участок ДНК), кодирующая белок или функциональную РНК [35]. Мутации внутри гена, закрепленные в процессе эволюции, приводят к появлению различных его вариантов (аллелей этого гена).

Специфический аллельный состав, характеризующий конкретную особь, называют генотипом. В диплоидном организме генотип представлен двумя наборами идентичных или различающихся аллелей каждого гена или иного участка ДНК, находящимися в соответствующих локусах парных (гомологичных) хромосом. Также можно говорить о генотипе по конкретному гену, участку или позиции (ОНП) в последовательности ДНК индивида. Таким образом, генотип по каждому ОНП в организме человека представлен одним из трех возможных вариантов по данной позиции (+/+; +/–; –/–). Если аллели совпадают (+/+; –/–), то организм называют гомозиготным, если нет (+/–) – гетерозиготным по генотипу в соответствующей позиции.

[1] Single Nucleotide Polymorphism (SNP).

Гаплотип (гаплоидный генотип) – это комбинация аллелей на локусах одной из двух гомологичных хромосом, обычно наследуемых вместе. Иногда (в данной работе) гаплотип определяют как набор ОНП, содержащихся на локусах одной из двух гомологичных хромосом. Генотип диплоидной особи состоит из двух родительских гаплотипов, расположенных на гомологичных хромосомах, полученных от матери и отца соответственно.

Фенотипом называется совокупность наблюдаемых свойств, присущих данному организму в данный момент времени. Элементарные единицы фенотипа (признаки) принято называть фенами [2] (термин фенотип также используется и для отдельных признаков). Главным образом, фенотип – физиологическое и (или) морфологическое следствие генотипа, но на его проявление оказывают влияние и внешние факторы. Наличие одного из трех генотипов (+/+; +/–; –/–) в локусе часто проявляется в виде определенного фена (фенотипа), поэтому выявление влияния определенных маркеров и их комбинаций на фенотип имеет важнейшее практическое значение. Задачи полногеномного поиска закономерностей (GWAS[1]) становятся все более актуальными при современных темпах развития технологий полногеномного секвенирования и накопления генетических данных.

Данная работа состоит из двух частей. В первой части рассмотрены различные типы данных, возникающие при проведении генетических исследований, а также методы их анализа. Во второй части будет обсуждаться проблема интерпретации результатов множества тестов, тесно связанная с задачами распознавания и выявления сигнала, а также реализация статистических методов поиска генетических закономерностей на языке программирования R и их применение в исследовании ВИЧ инфекции и развития СПИД.

2. ТИПЫ БИОСТАТИСТИЧЕСКИХ ДАННЫХ

Разработка дизайна статистического исследования – сложный процесс, включающий в себя постановку эксперимента, сбор данных, их анализ и интерпретацию результатов. На этапе подготовки формируется план исследования, ибо бессистемный сбор данных обычно не позволяет эффективно использовать вложенные ресурсы и получать информативные результаты. Исследование, связанное с человеком, обычно включает в себя организацию когорты пациентов с учетом этических стандартов и сопровождается массой документации. Простейший план эксперимента – разовый скрининг, в рамках которого случайно выбранные согласно запланированным критериям пациенты изучаются на предмет наличия того или иного заболевания (или характерного свойства). Для постановки такого эксперимента обычно требуется существенно меньше ресурсов по сравнению с более сложными экспериментальными планами, но возможности интерпретации результатов анализа в этом случае ограничены. Более сложный, с точки зрения постановки, эксперимент – наблюдение за когортой пациентов в течение определенного времени. Данный план эксперимента гораздо более затратен, и для его реализации требуется много времени. В связи с этим, при постановке такого эксперимента стараются получить максимальное число различного типа данных. При правильной постановке эксперимента результаты могут быть интерпретированы для описания гораздо более широкого набора явлений, по сравнению с разовым скринингом. Основные типы данных, получаемых в результате исследований, связанных с поиском генетических закономерностей, будут описаны далее.

Данные, пригодные для поиска генетических закономерностей, состоят из клинической и генотипической частей. Клинические данные содержат информацию о фенотипах, тогда как генотипические данные представляют собой наборы ОНП вариантов, выявленных у каждого из индивидов. Клинические данные обычно бывают категориальными, типа времени жизни, или короткими временными рядами (лонгитюдными).

[1] Genome wide association study (*англ.*)

Данные, получаемые в результате проведения разового скрининга, часто имеют категориальный тип. Наиболее часто в эпидемиологических исследованиях наблюдается бинарная переменная, принимающая значения 0 и 1 и характеризующая наличие инфекции или болезни у пациента. Иногда наблюдаемая категориальная переменная допускает и большее число значений (уровней). Наконец, даже если наблюдаемая переменная непрерывного типа, то индивидов можно классифицировать в определенное число групп по значениям наблюдаемой переменной, что позволяет свести задачу к категориальной схеме. Поскольку генотип также задается переменной, принимающей три значения, для такого исследования применяют методы категориального анализа (см. [6] или [5]).

Эксперимент может быть как пассивным (свободным), предполагающим случайный выбор индивидов из генеральной совокупности, так и активным (контролируемым),[1] предполагающим контроль исследователем численности групп индивидов с фиксированными значениями наблюдаемой переменной, или ковариат. Вся информация, пригодная для получения статистических выводов, может быть записана в виде таблицы сопряженности, ячейки которой заполняются числами наблюдений с фиксированными значениями наблюдаемой переменной и ковариат. Характерной особенностью полногеномного анализа закономерностей является наличие огромного числа таблиц сопряженности, каждому ОНП соответствует своя таблица сопряженности. Обычно статистические тесты проводят для каждой таблицы по отдельности, однако при интерпретации приходится учитывать результаты всех тестов. Для исследования совместного влияния нескольких генотипов на фенотип необходимо создание общей таблицы на единицу большей размерности, где в качестве дополнительного фактора используют номер ОНП.

При наблюдении за когортой пациентов в течение определенного промежутка времени в эпидемиологических исследованиях обычно следят за появлением симптомов болезни (инфекции) или за изменением того или иного признака путем проведения анализов. В первом случае получаем данные типа времени жизни[2] (см. [27, 16] или [3]), во втором – короткие временные ряды[3] (см. [22, 11] или [1, 5]).

Анализ данных типа времени жизни обычно направлен на изучение распределения времени перехода T объекта исследования из одного состояния в другое (отказа) или зависимости этого распределения от свойств изучаемого объекта (ковариаты). В эпидемиологии обычно речь идет о времени заражения некоторой инфекцией. В медицине часто изучают время возникновения рецидива некоторой болезни или время жизни с момента проведения хирургической операции или иного вмешательства (лечения). Наиболее информативным является наблюдение времени отказа у каждого из пациентов, однако наблюдение времени отказа всех пациентов практически неосуществимо. Часть пациентов остается в исходном состоянии до конца исследования. Исключение таких пациентов из анализа недопустимо, поскольку это приведет к систематической ошибке оценивания. Полезную информацию о распределении T несет время цензурирования справа U – время ухода пациента из под наблюдения. Если $T \le U$, то наблюдается время отказа, в противном случае наблюдается время цензурирования. Таким образом, для каждого пациента наблюдается пара (X, δ), где $X = T \wedge U$ и $\delta = \mathbb{1}_{\{T \le U\}}$,

$$\mathbb{1}_A = \begin{cases} 1, \text{ если } A \text{ выполнено,} \\ 0, \text{ в противном случае.} \end{cases}$$

Такого рода данные называются цензурированными справа.

[1] Например, Case-control study (*англ.*).
[2] Survival data (*англ.*).
[3] Longitudinal data (*англ.*).

Фактически при проведении когортного исследования время отказа наблюдается с некоторой погрешностью, определяемой интервалами времени между соседними обследованиями. Иными словами, информация об отказе (заболевании) представляется в виде времени последнего обследования, когда пациент еще был здоров L, и времени обследования, когда у него выявили симптомы заболевания W. В случае если пациент остается здоровым до конца эксперимента, наблюдается только $L = U$, а W можно считать равным бесконечности. Такая модель интервального цензурирования была предложена в работе [38], наряду с другими моделями анализа данных типа времени жизни.

Для изучения развития некоторого процесса (заболевания) весьма важным выглядит наблюдение за рядом сопутствующих характеристик, изменяющихся с течением времени. Наборы значений характеристик, измеренных в ряде временных точек, называются временными рядами. Наблюдения могут быть представлены в виде векторов $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{i\,n_i})'$, где i – номер индивида, а n_i – число наблюдений, соответствующих i-му индивиду. В зависимости от плана эксперимента, числа n_i могут быть случайными или фиксированными. Наряду с $\boldsymbol{Y}_i$, обычно наблюдаются времена последовательных обследований t_{ij}, $i = 1, ..., n$, $j = 1, ..., n_i$. Времена t_{ij} также могут быть случайными или контролироваться исследователем.

3. АНАЛИЗ КАТЕГОРИАЛЬНЫХ ДАННЫХ

Если наблюдаемая переменная имеет категориальный тип, то для поиска генотипов, зависимых с наблюдаемой переменной, обычно используют как классические методы категориального анализа, так и обобщенные линейные модели. Методы выявления зависимости в основном применимы как для пассивного, так и для активного экспериментов. В первом случае можно говорить о совместном распределении наблюдаемой переменной и ковариаты, во втором – интерпретация возможна только в терминах условных распределений наблюдаемой переменной при различных значениях ковариаты. Генетические исследования обычно проходят в условиях пассивного плана, однако статистический анализ совместного распределения фенотипа и всего генома не представляется возможным, поскольку число генов в геноме в разы превышает число индивидов. Таким образом, результаты анализа удобнее интерпретировать в терминах условных распределений наблюдаемой переменной при различных значениях отдельных ковариат или небольших наборов ковариат.

3.1. КЛАССИЧЕСКИЕ МЕТОДЫ АНАЛИЗА

Классические методы анализа категориальных данных используют представление исходных данных в виде таблицы сопряженности. В условиях свободного эксперимента распределение значений в таблице сопряженности является мультиномиальным. Отметим, что условные распределения значений в таблице сопряженности, соответствующие различным значениям ковариаты, также являются мультиномиальными и независимость наблюдаемой переменной с ковариатой в условиях свободного эксперимента проверяется теми же тестами, что и однородность условных распределений при всех значениях ковариаты. При наличии дополнительных факторов возможно построение таблиц более высокой размерности (отдельных таблиц для каждого набора значений дополнительных факторов).

При анализе таблиц сопряженности наиболее часто используются критерий хи-квадрат и асимптотически эквивалентный ему критерий отношения правдоподобия. Следует отметить, что эти критерии являются асимптотическими и при наличии малого числа наблюдений в каких-либо ячейках таблицы сопряженности не рекомендованы к применению. В этом случае можно использовать критерии случайных перестановок на базе соответствующих статистик. Идея построения таких критериев состоит в том, что имеющемуся набору

значений наблюдаемой переменной случайным образом сопоставляются имеющиеся значения ковариат, и данная операция проводится K-раз. В каждом случае вычисляется значение статистики и соответствующее P-значение p_k^*, $k = 1, ..., K$. Тогда P-значение критерия случайных перестановок выбирается как R/K, где R – ранг наибольшего из p_k^*, не превышающего исходное P-значение. Использование рандомизации повышает требуемый вычислительный ресурс более чем в K раз, что весьма нежелательно для полногеномного анализа закономерностей. С другой стороны, полученное P-значение не превышает K^{-1}, что с учетом необходимости выбора только очень малых P-значений, далеко не способствует использованию данного метода.

Альтернативно, можно игнорировать ячейки с малым количеством наблюдений, что снизит эффективность теста, но существенно не отразится на скорости вычислений. Наконец, в рамках предварительного скрининга, можно вычислять значение статистики «как есть», а в случае выбора соответствующей закономерности, провести анализ более точно.

Для таблиц сопряженности 2×2 можно использовать так называемый точный критерий Фишера, основанный на гипергеометрическом распределении значения в выбранной ячейке (например n_{11}) таблицы сопряженности при фиксированных суммарных значениях таблицы по столбцам и по строкам. Точный критерий Фишера достаточно популярен при проведении анализа генетических закономерностей с бинарной классификацией по генотипу, поскольку ситуации с недостаточным числом наблюдений в отдельных ячейках встречаются достаточно часто. Тем не менее, при наличии достаточно больших значений во всех ячейках таблицы рекомендуется использовать асимптотические критерии.

3.2. ИСПОЛЬЗОВАНИЕ ОБОБЩЕННЫХ ЛИНЕЙНЫХ МОДЕЛЕЙ

В основе обобщенной линейной модели однофакторного дисперсионного анализа (простой группировки) лежит регрессионное соотношение

$$g(\mathbb{E}(Y|z = i)) = \mu + \beta_i, \qquad (1)$$

где Y – наблюдаемая переменная (фенотип), μ – базовый уровень (взвешенное среднее), β_i – главный эффект i-го уровня рассматриваемого фактора (генотипа), g – функция связи[1]. Распределение бинарной наблюдаемой переменной $Y \in \{0,1\}$ определяется параметром $p_z = \mathbb{P}(Y = 1|z) = \mathbb{E}(Y|z)$. В этом случае обычно используют модель (1) с функцией связи $g(u) = \log(u/(1-u)) = \text{logit}(u)$ – логистическая регрессия. При наличии у фенотипа трех или более уровней имеет смысл использовать пуассоновскую модель с функцией связи $g(u) = \log(u)$, где в качестве наблюдаемых переменных используются значения из таблицы сопряженности. В условиях пуассоновской модели предполагается, что все значения в таблице сопряженности являются независимыми и имеют распределения Пуассона с параметрами $\mu_{ij} = \lambda p_{ij}$, $\sum_{ij} p_{ij} = 1$, а суммарное число значений в таблице сопряженности имеет распределение Пуассона с параметром λ. Следует отметить, что условное распределение значений в таблице сопряженности при фиксированной сумме в пуассоновской модели является мультиномиальным.

4. АНАЛИЗ ДАННЫХ ТИПА ВРЕМЕНИ ЖИЗНИ

В теории вероятностей для задания распределения случайной величины обычно используют функцию распределения $F(x) = \mathbb{P}(T \le x)$. При работе с данными типа времени жизни с правым цензурированием для задания распределения T: $\mathbb{P}(T \ge 0) = 1$, удобно использовать так называемую функцию отказа $S(x) = 1 - F(x) = \mathbb{P}(T > x)$ или накопленную интенсивность

[1] Link (*англ.*).

$$\Lambda(x) = \int_0^x \frac{dF(x)}{1 - F(x\_)}.$$

В случае абсолютно непрерывного распределения T

$$\Lambda(x) = \int_0^x \lambda(x)dx = \int_0^x p(x)/S(x\_)dx,$$

где $p(x)$ – плотность распределения T, а $S(x\_)$ – предел слева функции S в точке x. Если распределение T является дискретным, то

$$\Lambda(x) = \sum_{u \le x} \lambda(u) = \sum_{u \le x} \mathbb{P}(T = u) / \mathbb{P}(T \ge u).$$

Функцию λ называют интенсивностью отказа. Отметим также, что

$$S(x) = \exp(-\Lambda^c(x)) \prod_{u \le x} (1 - \Delta\Lambda^d(u)),$$

где Λ^c и Λ^d – накопленные интенсивности, соответствующие непрерывной и дискретной компонентам распределения T соответственно.

В случае цензурирования справа каждое наблюдение представляет собой пару (X, δ), где $X = T \wedge U$, $\delta = \mathbb{1}_{\{T \le U\}}$ и U – время цензурирования справа.

Наиболее часто делают предположение о независимости цензурирования. Фактически, условие независимости цензурирования сводится к совпадению условного распределения T при условии $X \ge u$ с усеченным слева распределением T на уровне u при каждом фиксированном u. В частности, если T и U независимые величины, то условие независимости цензурирования выполнено.

4.1. НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ

Пусть $(X_1, \delta_1), ..., (X_n, \delta_n)$ – исходные данные типа времени жизни с правым цензурированием; $T_{(1)} \le ... \le T_{(L)}$ – последовательные времена наблюдавшихся отказов, $L \le n$. Аналог эмпирического распределения в случае независимого правого цензурирования имеет накопленную интенсивность

$$\hat{\Lambda}(x) = \sum_{i:T_{(i)} \le x} \frac{d_i}{Y_i},$$

где $d_i = \sum_j \mathbb{1}_{\{T_j = T_{(i)}, \delta_j = 1\}}$ – число отказов, наблюдавшихся в момент времени $T_{(i)}$, $Y_i = \sum_j \mathbb{1}_{\{X_j \ge T_{(i)}\}}$ – число объектов с неопределенным исходом (не отказавших и не цензурированных) к моменту времени $T_{(i)}$. Оценка $\hat{\Lambda}$ называется оценкой Нельсона–Аалена. Соответствующая функция отказа называется оценкой Каплана–Мейера

$$\hat{S}(x) = \prod_{i:T_{(i)} \le x} (1 - \Delta\hat{\Lambda}(T_{(i)})) = \prod_{i:T_{(i)} \le x} (1 - d_i / Y_i). \quad (2)$$

Данное распределение является дискретным с атомами в точках $T_{(1)}, ..., T_{(L)}$ и весами

$$q(T_{(k)}) = \hat{S}(T_{(k-1)}) - \hat{S}(T_{(k)}) = \frac{d_k}{Y_k} \hat{S}(T_{(k-1)}) = \frac{d_k}{Y_k} \prod_{i=1}^{k-1} (1 - d_i / Y_i).$$

Рассмотрим еще один алгоритм вычисления данного распределения. Для этого удобнее рассмотреть все последовательные времена наблюдаемых событий (отказов или цензурирований) $X_{(1)} \le ... \le X_{(m)}$, $L \le m \le n$; $\tilde{d}_i^{\,f} = \sum_{j=1}^{i} \mathbb{1}_{\{X_j = X_{(i)}, \delta_j = 1\}}$, $\tilde{d}_i^{\,f} = \sum_{j=1}^{i} \mathbb{1}_{\{X_j = X_{(i)}, \delta_j = 0\}}$ и $\tilde{d}_i = \tilde{d}_i^{\,f} + \tilde{d}_i^{\,c}$ – числа отказов, цензурирований и общее число событий соответственно, наблю-

давшихся в момент времени $X_{(i)}$; $\widetilde{Y}_i = \sum_j \mathbb{I}_{\{X_j \geq X_{(i)}\}} = \sum_{j \geq i} d_j$. Формула (2) в новых обозначениях будет выглядеть следующим образом:

$$\hat{S}(x) = \prod_{i:X_{(i)} \leq x} (1 - \widetilde{d}_i^{\,f} / \widetilde{Y}_i) .$$

С использованием соотношений

$$1 - \frac{\widetilde{d}_i^{\,f}}{\widetilde{Y}_i} = \frac{\widetilde{Y}_i - \widetilde{d}_i^{\,f}}{\widetilde{Y}_i} = \frac{\widetilde{Y}_i - \widetilde{d}_i}{\widetilde{Y}_i} \frac{\widetilde{Y}_i - \widetilde{d}_i^{\,f}}{\widetilde{Y}_i - \widetilde{d}_i} = \frac{\widetilde{Y}_{i+1}}{\widetilde{Y}_i} \left(1 + \frac{\widetilde{d}_i - \widetilde{d}_i^{\,f}}{\widetilde{Y}_{i+1}} \right)$$

получаем, что

$$\hat{S}(x) = \frac{\widetilde{Y}_{k+1}}{n} \prod_{i:X_{(i)} \leq x} (1 + \widetilde{d}_i^{\,c} / \widetilde{Y}_{i+1}) \quad \text{и} \quad q(X_{(k)}) = \frac{\widetilde{d}_k^{\,f}}{n} \prod_{i=1}^{k-1} (1 + \widetilde{d}_i^{\,c} / \widetilde{Y}_{i+1}) .$$

Таким образом, для получения данного распределения можно использовать следующий алгоритм:

1) начинаем с эмпирического распределения на множестве времен наблюдаемых событий $X_1, ..., X_n$ (то есть каждому X_i сопоставляем вес $1/n$);

2) последовательно перебирем все наблюдаемые времена событий $X_1, ..., X_n$ в порядке возрастания значений, если $\delta_i = 1$, оставляем текущий вес, соответствующий выбранному наблюдению X_i, в точке X_i, а если $\delta_i = 0$, то распределяем его равномерно по всем $X_j: X_j > X_i$. Отметим, что построенное распределение является собственным в том и только в том случае, если $T_{(L)} = X_{(m)}$ и все события, наблюдающиеся в момент времени $X_{(m)}$, являются отказами.

Отметим важные асимптотические свойства оценок Нельсона–Аалена и Каплана–Мейера, которые в дальнейшем будут использованы при построении категориальных тестов. Пусть $(T_1, U_1), ..., (T_n, U_n)$ – выборка из двумерного распределения с непрерывной функцией распределения F компоненты T, $\gamma_T = \sup\{x : F(x) < 1\}$. Тогда, при выполнении условия независимости цензурирования

$$\sqrt{n}(\hat{\Lambda}(x) - \Lambda(x)) \Rightarrow W_{\tau^2(x)} \tag{3}$$

в $D([0, t])$, $t < \gamma_T$, где W – стандартный винеровский процесс, $\tau^2(x)$ – функция вариации. В качестве состоятельной оценки $\tau^2(x)$ обычно используют

$$\hat{\tau}^2(x) = n \sum_{i:T_{(i)} \leq x} d_i / Y_i^2 \to_P \tau^2(x) \tag{4}$$

при $n \to \infty$. Аналогичный результат справедлив для оценки Каплана–Мейера

$$\sqrt{n}(\hat{S}(x) / S(x) - 1) \Rightarrow W_{\sigma^2(x)} \tag{5}$$

в $D([0, t])$, $t < \gamma_T$, где W – стандартный винеровский процесс, $\sigma^2(x)$ – функция вариации. Состоятельная оценка для $\sigma^2(x)$ может быть вычислена по формуле Гринвуда

$$\hat{\sigma}^2(x) = n \sum_{i:T_{(i)} \leq x} d_i / (Y_i (Y_i - d_i)) \to_P \sigma^2(x) \tag{6}$$

при $n \to \infty$.

4.2. ИСПОЛЬЗОВАНИЕ МЕТОДОВ КАТЕГОРИАЛЬНОГО АНАЛИЗА

Методы категориального анализа применимы и в случае, если наблюдаемые переменные непрерывного типа. Наиболее известный пример – хи-квадрат критерий, когда множество значений переменных разбивают на интервалы, получая при этом дискретные распределения. Данный подход применим и для данных типа времени жизни. Применимость клас-

сических категориальных тестов при анализе данных типа времени жизни детально рассмотрена в работе [30]. Основные результаты, связанные с применимостью классических методов категориального анализа при анализе данных типа времени жизни, будут рассмотрены далее.

Каждому наблюдению сопоставим пару (T, z), где T – время отказа (заболевания), $z \in \{1, ..., d\}$ – ковариата категориального типа, значение которой характеризует распределение T. Обозначим F_z, S_z и Λ_z функцию распределения, функцию отказа и накопленную интенсивность распределения T при значении ковариаты z. Пусть $0 < t_1 < \ldots < t_{s-1} < \infty$ разбиение множества допустимых значений T: $0 < F_z(t_1) < \ldots < F_z(t_{s-1}) < 1$ при всех допустимых значениях z. Гипотеза однородности

$$H_0: F_1(x) = \ldots = F_d(x) \text{ при всех } x \geq 0$$

заменяется на более слабую гипотезу

$$H_0^c: F_1(t_j) = \ldots = F_d(t_j) \text{ при всех } j = 1, ..., s-1.$$

Если все отказы наблюдаются, то их можно классифицировать в s групп и получить таблицу сопряженности $s \times d$. В этом случае можно использовать классические методы категориального анализа, основанные на сопоставлении полученных частот и теоретических вероятностей $p_{k|z} = \mathbb{P}(T \in I_k | z)$, $I_1 = [0, t_1]$, $I_k = (t_{k-1}, t_k]$ при $k = 2, ..., s-1$ и $I_s = (t_s, \infty)$. В общем случае наблюдения, цензурированные до момента времени t_{s-1}, не удается классифицировать в ту или иную группу, поскольку точное время отказа неизвестно. Отметим, что любые правила классификации таких наблюдений или удаления их из анализа ведут к смещениям ожидаемых частот из таблицы сопряженности по отношению к теоретическим вероятностям. В то же время категориальная схема накопления статистической информации может быть использована при решении некоторых задач и в этом случае. В качестве примеров рассмотрим три способа классификации наблюдений:

1) все наблюдения, цензурированные до момента t_s, удаляются из анализа;

2) все цензурированные наблюдения удаляются из анализа;

3) все наблюдения классифицируются по наблюдаемому моменту события $X = T \wedge U$.

Пусть $\nu_{i|j} = n_{ij}/n_{\cdot j}$, где n_{ij} – число наблюдений с ковариатой $z = j$, классифицированных в i-ю группу, $n_{\cdot j} = \sum_i n_{ij}$. Очевидно, что в первом случае

$$q_{i|z} = \mathbb{E}\nu_{i|z} = \mathbb{P}(T \in I_i | \{T \leq U\} \cup \{U > t_s\}; z).$$

Предположим дополнительно, что T и U – независимые величины и функция отказа G величины U не зависит от ковариаты. Тогда,

$$q_{i|z} = \int_{I_i} G(x) dF_z(x) \Big/ \left(S_z(t_s) G(t_s) + \int_0^{t_s} G(x) dF_z(x) \right), \; i = 1, ..., s-1$$

и $q_{s|z} = 1 - \sum_{i=1}^{s-1} q_{i|z}$. Аналогично, во втором случае

$$q_{i|z} = \mathbb{P}(T \in I_i | T \leq U; z) = \int_{I_i} G(x) dF_z(x) \Big/ \int_0^{\infty} G(x) dF_z(x)$$

и в третьем случае

$$q_{i|z} = \mathbb{P}(T \wedge U \in I_i | z) = \begin{cases} 1 - G(t_1) S_z(t_1), & i = 1, \\ G(t_{i-1}) S_z(t_{i-1}) - G(t_i) S_z(t_i), & i = 2, ..., s, \\ G(t_s) S_z(t_s), & i = s. \end{cases}$$

Классические категориальные тесты применимы в этом случае для проверки гипотезы

$$H_0^b: q_{i|1} = \ldots = q_{i|d}, \; i = 1, ..., s.$$

Очевидно, что отвержение данной гипотезы влечет отвержение H_0, но не H_0^c. Таким образом, классические категориальные тесты могут быть использованы для проверки гипо-

тезы H_0, но значения статистик по сути не могут быть использованы для обоснования характера и величины отклонений от основной гипотезы, поскольку в выражения для $q_{i|z}$ помимо параметра F входит мешающий параметр G.

4.3. ПРОВЕРКА КАТЕГОРИАЛЬНОЙ ГИПОТЕЗЫ ОДНОРОДНОСТИ ПО ЦЕНЗУРИРОВАННЫМ СПРАВА ДАННЫМ

Для проверки гипотезы однородности по цензурированным данным типа времени жизни наиболее часто применяются так называемые G^ρ-статистики [21]. Менее популярны критерии типа хи-квадрат, разработанные для проверки категориальной гипотезы H_0^c. Мы рассмотрим еще один простой метод построения статистических критериев для проверки H_0^c, обсуждавшийся в работе [30]. Для построения критериев нам потребуются асимптотические свойства (3), (5) и метод сравнений, широко использующийся в дисперсионном анализе.

Переформулируем гипотезу H_0^c в следующем виде:

$$H_0^c:\ S_1(t_j) = \ldots = S_d(t_j),\ j = 1, \ldots, s-1.$$

Введем $\theta_{ij} = S_i(t_j)$, $i = 1, \ldots, d$, $j = 1, \ldots, s-1$. Отметим, что оценки Каплана–Мейера $S_1, \ldots, S_d$ обладают свойством независимости при условии, что значения ковариат изначально фиксированы. С использованием (5) получаем сходимость

$$\sqrt{n_i}\,(\hat{\theta}_{i1} - \theta_{i1}, \ldots, \hat{\theta}_{i\,s-1} - \theta_{i\,s-1})' \Rightarrow N(0, \Sigma_i) \text{ при } n_i \to \infty,$$

где $\Sigma_i = \|\sigma_{i:qr}\|: \sigma_{i:qr} = \theta_{iq}\theta_{ir}\sigma_i^2(t_q \wedge t_r)$, $q, r = 1, \ldots, s-1$, n_i – число наблюдений с ковариатой $z = i$. Для оценки матрицы ковариации используем $\hat{\Sigma}_i$ с элементами $\hat{\sigma}_{i:qr} = \hat{\theta}_{iq}\hat{\theta}_{ir}\hat{\sigma}_i(t_q \wedge t_r)$, а $\hat{\sigma}_i(t)$ вычисляется по формуле Гринвуда (6). Рассмотрим $\theta = (\theta_{11}, \ldots, \theta_{1\,s-1}, \ldots, \theta_{d1}, \ldots, \theta_{ds-1})'$ и $\hat{\theta} = (\hat{\theta}_{11}, \ldots, \hat{\theta}_{1\,s-1}, \ldots, \hat{\theta}_{d1}, \ldots, \hat{\theta}_{ds-1})'$. Тогда,

$$\sqrt{n}\,(\hat{\theta} - \theta) \Rightarrow N(0, \Sigma_i). \qquad (7)$$

где $\Sigma = \mathrm{diag}(l_1\Sigma_1, \ldots, l_n\Sigma_d)$ – блочно-диагональная матрица, $l_i = n/n_i$ при $i = 1, \ldots, d$.

Сравнением параметра $\theta_j = (\theta_{1j}, \ldots, \theta_{dj})$ называется линейная комбинация $\sum_i a_i\theta_{ij}$: $\sum_i a_i = 0$. Пусть $\mathbf{A} = \|a_{ik}\|$ – $d \times (d-1)$-матрица ранга $d-1$: $\sum_i a_{ik} = 0$ при $k = 1, \ldots, d-1$. Известно, что $\theta_{1j} = \ldots = \theta_{dj}$ равносильно обращению в нуль всех $d-1$ линейно независимых сравнений: $\mathbf{A}'\theta_j = 0$.

Сопоставим каждому a_{ij} диагональную матрицу $\mathbf{A}_{ij} = a_{ij}\,\mathbf{I}_{s-1}$, где $\mathbf{I}_{s-1}$ – единичная матрица размерности $s-1$, и введем матрицу сравнений $\mathbf{B}$ размера $(d-1)(s-1) \times d(s-1)$, составленную из блоков $\mathbf{A}_{ij}$ в соответствующем порядке. Матрица $\mathbf{B}$ является матрицей сравнений параметра θ, и основная гипотеза H_0 может быть переформулирована в следующем виде:

$$H_0: \mathbf{B}\theta = 0.$$

С использованием (7) и свойств нормального распределения получаем сходимость

$$n\hat{\theta}'\,\hat{\boldsymbol{Q}}^{-1}\hat{\theta} \Rightarrow \chi^2_{(d-1)(s-1)},$$

где $\hat{\boldsymbol{Q}} = \mathbf{B}'(\mathbf{B}\hat{\Sigma}\,\mathbf{B}')^{-1}\mathbf{B}$.

Отметим, что гипотеза H_0^c может быть переформулирована в виде

$$H_0^c : \Lambda_1(t_j) = ... = \Lambda_d(t_j),\ j = 1, ..., s-1,$$

что позволяет использовать асимптотическую нормальность (3) для построения критерия, который строится аналогично с использованием сравнений.

4.4. ПАРАМЕТРИЧЕСКИЕ И СЕМИПАРАМЕТРИЧЕСКИЕ МОДЕЛИ

В параметрических моделях делается предположение о принадлежности распределения времени отказа к некоторому параметрическому семейству распределений, сконцентрированных на положительной полуоси при каждом значении ковариаты z. Методы статистического анализа в условиях параметрической модели в основном базируются на функции правдоподобия. В случае правого цензурирования функция правдоподобия допускает разложение

$$L(\boldsymbol{X},\delta;\theta) = L^f(\boldsymbol{X},\delta;\theta)L^c(\boldsymbol{X},\delta;\theta),$$

где L^f выражается в терминах условных распределений T при условии X и δ, а L^c – в терминах условных распределений U при условии X и δ. Если L^f и L^c выражаются через независимые параметры θ_f и θ_c $(\theta = (\theta_f, \theta_c))$ соответственно, а интерес представляет оценивание θ_f при мешающем параметре θ_c, то цензурироавние называется неинформативным.

В случае независимого неинформативного цензурирования L^c не зависит от параметра θ_f, а следовательно, $L(\boldsymbol{X},\delta,z;\theta) \cong L^f(\boldsymbol{X},\delta,z;\theta_f)$, и

$$L^f(\boldsymbol{X},\delta,z;\theta_f) = \prod_{i=1}^n p_{z_i}(X_i;\theta_f)^{\delta_i} S_{z_i}(X_i;\theta_f)^{1-\delta_i} = \prod_{i=1}^n \lambda_{z_i}(X_i;\theta_f)^{\delta_i} S_{z_i}(X_i;\theta_f).$$

При выполнении определенных условий регулярности выбранного параметрического семейства распределений работает стандартный критерий отношения правдоподобия. Также, можно использовать асимптотическую нормальность оценок максимального правдоподобия для построения критериев типа Вальда.

Наряду с параметрическими при анализе данных типа времени жизни используют семипараметрические модели. Наиболее часто в приложениях используется модель пропорциональных интенсивностей Кокса [10]

$$\lambda_z(x) = \exp(\boldsymbol{X}(z)'\beta)\lambda_0(x),\ x \geq 0,$$

где λ_0 – неизвестная базовая интенсивность, соответствующая нулевому значению ковариаты. Данную модель можно задать соотношением

$$\ln(\lambda_z(x)/\lambda_{z_1}(x)) = \left((\boldsymbol{X}(z) - \boldsymbol{X}(z_1))'\beta\right),\ x \geq 0,$$

что выглядит более корректным, поскольку значение $z = 0$ может отсутствовать в множестве допустимых значений ковариат. Модель Кокса для исследования связи генотипа с фенотипом выглядит следующим образом:

$$\ln(\lambda_i(x)/\lambda_j(x)) = \beta_i - \beta_j,$$

где λ_s – интенсивность заболеваемости в группе с s-м генотипом, β_s – соответствующий параметр. Таким образом, гипотеза отсутствия связи фенотипа с выбранным генотипом выполнена в случае равенства нулю всевозможных сравнений параметров β_i.

Существует ряд обобщений данной модели [28, 7], однако модель Кокса по-прежнему используется наиболее часто, в первую очередь, в силу того, что методы статистического анализа для модели Кокса существенно проще, чем для ее обобщений. Последнее обусловлено тем, что в модели Кокса функция правдоподобия L^f не зависит от мешающего параметра λ_0, а это позволяет использовать стандартные методы, базирующиеся на функции правдоподобия.

В заключение отметим, что условие пропорциональности интенсивностей при различных значениях ковариат весьма ограничительно. В частности, большинство стандартных параметрических семейств распределений (кроме экспоненциального) не обладают свойством пропорциональности интенсивностей, а следовательно, соответствующие параметрические модели не укладываются в модель Кокса.

4.5. ИНТЕРВАЛЬНОЕ ЦЕНЗУРИРОВАНИЕ

В общей модели интервального цензурирования [38] наблюдение представляется в виде пары (L,W) времен соседних обследований, между которыми произошел отказ (заболевание), и $W = \infty$ для цензурированных справа наблюдений. Фактически в модели интервального цензурирования присутствуют времена обследований $V_1, ..., V_k$, которые считаются независимыми от T. При $k = 1$ данные называются *current status dat*a, при фиксированном k говорят об интервальном цензурировании с k границами[1] (см. [19]). Общий случай интервального цензурирования со случайным k, как и случай $k > 2$, с вычислительной точки зрения удобно сводить к интервальному цензурированию с 2-мя границами [39], однако известные свойства распределений, вообще говоря, не переносятся на случай общего интервального цензурирования. Точное значение T не наблюдается, но известно, что $T \in [V_s,V_{s+1}] = [L, W]$ при некотором $s \in \{0, ..., k\}$, $V_0 = 0$, тогда как в случае, если пациент остается здоровым до конца обследования, $L < T$, и в этом случае L совпадает с временем цензурирования U, а W считается равным бесконечности. Упрощенная[2] функция правдоподобия представляет собой произведение вероятностей попадания T в наблюдаемый интервал $(L, W]$ и выражается в терминах приращений функций отказа на наблюдаемых интервалах

$$L^s(\boldsymbol{L},\boldsymbol{W},z;\theta_f) = \prod_{i=1}^{n}(S_{z_i}(L_i;\theta_f) - S_{z_i}(W_i;\theta_f)),$$

где (L_i, W_i, z_i) – наблюдаемые границы интервала и значение ковариаты, соответствующие i-му индивиду. Неинформативность цензурирования заключается в эквивалентности упрощенной и полной функций правдоподобия. Условия неинформативности интервального цензурирования получены в [32].

В параметрическом случае упрощенная функция правдоподобия легко выписывается с учетом выбранной модели, и все классические свойства оценок максимального правдоподобия сохраняются. В отличие от модели с правым цензурированием, использование семипараметрической модели Кокса требует достаточно сложных вычислений и не обходится без учета непараметрической части Λ_0. В работе [17] рассматривался случай дискретного распределения границ интервалов с конечным множеством значений. Задачи в данной постановке по сути могут быть сведены к работе с группированными данными. В работе Pan [13] было отмечено, что методы, использовавшиеся в [17], неприменимы для построения тестов типа Вальда и отношения правдоподобия в общем случае в связи с большим числом мешающих параметров. В работе [13] предлагается метод группировки с границами, зависящими от исходных данных, и рассмотрен ряд дополнительных моделей, отличных от модели Кокса. Некоторые обобщения получены также в [14]. Следует отметить, что предельная матрица ковариации в рассмотренных тестах обычно оценивается эмпирически исходя из нулевой гипотезы, поэтому рассмотренные критерии, построенные на базе семипараметрических моделей, по сути являются непараметрическими. Также следует отметить, что в работе [26] рассмотрены асимптотические свойства оценок максимального правдоподобия в модели Кокса для интервально-цензурированных данных с $k = 2$ границами.

Непараметрические методы в модели с интервальным цензурированием идеологически гораздо более сложны, чем в случае правого цензурирования, а полученные оценки часто

[1] Interval censoring case k (*англ.*).
[2] Simplified (*англ.*).

являются неоднозначными. В задаче непараметрического оценивания распределения T используют принцип максимизации непараметрической функции правдоподобия, однако, в отличие от случая правого цензурирования, неоднозначность оценки возникает не только на правом хвосте распределения. Метод построения непараметрической оценки максимального правдоподобия, основанный на принципе состоятельной устойчивости (self-consistency) и использующий EM-алгоритм, разработан в [38]. Groneboom & Wellner [19] предлагают использовать ICM[1]-алгоритм для построения непараметрической оценки максимального правдоподобия. Различные методы построения непараметрической оценки максимального правдоподобия обсуждаются в работе [18]. В работе [25] обсуждаются обобщения логранк-тестов на случай интервально цензурированных данных. Обобщения G^{ρ} -критериев на случай интервально-цензурированных данных получены в работе [33].

5. АНАЛИЗ КОРОТКИХ ВРЕМЕННЫХ РЯДОВ

Классические методы анализа временных рядов [1] в основном ориентированы на достаточно большое число точек, в которых измеряется исследуемая характеристика, и нормальное распределение наблюдаемой переменной. В биомедицинских исследованиях обычно используются модели коротких временных рядов, техника анализа данных в которых отличается от классического анализа временных рядов [22, 11]. Особенностью модели коротких временных рядов является наличие нескольких измерений наблюдаемой переменной, соответствующих каждому индивиду, в различные моменты времени, что подразумевает зависимость этих наблюдений. Наблюдаемая переменная может быть как непрерывного, так и ординального или категориального типа. Наблюдения, соответствующие различным индивидам, остаются независимыми. Обозначим Y_{ij} – наблюдение i-го индивида в j-й момент времени t_{ij}, $j = 1, ..., n_i$, $i = 1, ..., n$. Наиболее благоприятный случай, когда наблюдения проводятся в одни и те же моменты времени, однако на практике поставить такой эксперимент достаточно сложно и времена измерений обычно оказываются различными. Более того, различным индивидам может соответствовать различное число измерений.

5.1. ОБОБЩЕННЫЕ МОДЕЛИ АНАЛИЗА КОРОТКИХ ВРЕМЕННЫХ РЯДОВ

Теория обобщенных линейных моделей [31] разработана в основном для экспоненциальных семейств определенного вида. Она легко обобщается на многомерный случай [12], однако практическая ценность таких результатов невелика, поскольку экспоненциальные семейства многомерных распределений с зависимыми компонентами подобрать достаточно сложно и обосновать применимость таких моделей на практике весьма затруднительно. Исключение составляет многомерное нормальное распределение, но и в этом случае из нормальности распределения компонент не следует нормальность распределения всего вектора.

Отметим, что целью статистического анализа являются одномерные распределения компонент, структура зависимости между компонентами представляет вспомогательный интерес. В связи с этим, широкое распространение получили се́мипараметрические обобщенные модели, называемые GEE (Generalized Estimation Equations [29]). Для описания распределений компонент многомерного распределения используются параметрические модели, а зависимость между компонентами (копула) предполагается полностью неизвестной и считается мешающим параметром. Вместо оценок максимального правдоподобия, используются М-оценки, которые строятся с использованием выбранных параметрических моделей для компонент совместного распределения и так называемой «рабочей» матрицы ковариаций. Полученные М-оценки состоятельны и асимптотически нормальны с допускающей

[1] Iterative convex minorant (*англ.*).

оценивание матрицей ковариации, что позволяет использовать их в задачах доверительного оценивания и проверки гипотез. Правильный выбор «рабочей» матрицы ковариации повышает эффективность статистического анализа. На практике обычно выбирают определенную корреляционную структуру, то есть «рабочая» матрица ковариации параметризована и параметр также подлежит оцениванию. Для определения формы зависимости одномерных распределений от ковариаты обычно используют регрессию. Покоординатные регрессионные соотношения для поставленной задачи проверки влияния генотипа на течение измеряемого процесса могут быть выбраны следующим образом:

$$g(\mathbb{E}_\theta(Y|z_{ij})) = g(\mathbb{E}_\theta Y_{ij}) = \mu + \beta_i + \gamma t_{ij} + \gamma_i t_{ij}, \quad (8)$$

где θ – параметр модели, включающий в себя μ – базовый уровень (взвешенное среднее), β_i – главный эффект генотипа, γ – параметр линейной регрессии по времени, γ_i – взаимодействие уровня фактора генотипа и времени; z_{ij} – ковариата, включающая в себя момент времени обследования t_{ij} и элементарные переменные простой группировки, характеризующие наличие того или иного генотипа и не зависящие от j; g – функция связи. Выбор параметрического семейства распределений Y и функции связи производится с учетом природы имеющихся наблюдений. В частности, если Y_{ij} – бинарные величины, то выбор семейства распределений ограничен биномиальным, а в качестве функции связи удобно использовать логистическую регрессию $g(u) = \log(u/(1-u))$.

Если временные интервалы между соседними обследованиями одинаковы у всех индивидов, то можно использовать неструктурированную (произвольную) форму «рабочей» матрицы ковариации, в противном случае обычно используют параметрическую зависимость корреляции наблюдений от времен, в которые проводились измерения.

5.2. СМЕШАННЫЕ МОДЕЛИ АНАЛИЗА КОРОТКИХ ВРЕМЕННЫХ РЯДОВ

При анализе коротких временных рядов часто используют смешанные модели с так называемым простым эффектом индивида

$$g(\mathbb{E}_\theta(Y|z_{ij}, v_i)) = g(\mathbb{E}_\theta Y_{ij}|v_i) = \mu + \beta_i + \gamma t_{ij} + \gamma_i t_{ij} + \sigma_v v_i, \quad (9)$$

где первые слагаемые в левой части совпадают с левой частью (8); $v_1, ..., v_n$ – независимые и одинаково распределенные случайные величины с плотностью распределения h, $i = 1, ..., n$. Считаем также, что при фиксированных значениях $\boldsymbol{v} = (v_1, ..., v_n)$ величины $Y_{i1}, ..., Y_{in_i}$ являются независимыми.

В предположении, что совместное распределение наблюдаемых величин является нормальным, выбирают $g(u) = u$, и можно использовать следующую модель:

$$Y_{ij} = \mu + \beta_i + \gamma t_{ij} + \gamma_i t_{ij} + \sigma_v v_i + \sigma_e \varepsilon_{ij},$$

где v_i – независимые и одинаково распределенные случайные величины, имеющие стандартное нормальное $N(0, 1)$ распределение, а $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{in_i})$ $i = 1, ..., n$ – независимые нормально-распределенные случайные векторы, имеющие определенную корреляционную структуру.

Литература

1. *Андерсон Т.* Статистический анализ временных рядов. М.: Мир, 1976.

2. *Инге-Вечтомов С.Г.* Генетика с основами селекции. 2-е изд, перераб. и доп. СПб.: Изд-во Н-Л, 2010.

3. *Кокс, Оукс.* Анализ данных типа времени жизни. М.: Финансы и статистика, 1988.

4. *Леман Э.* Проверка статистических гипотез. 2-е изд. М.: Наука, 1979.

5. *Малов С.В.* Регрессионный анализ: теоретические основы и практические рекомендации. СПб.: Изд-во СПбГУ, 2013.

6. *Agresti A.* Categorical data analysis. 2-nd edition. Hoboken, New Jersey: Wiley & Sons, Inc., 2002.

7. *Bagdonavičius V. & Nikulin M.S.* Transfer Functionals and Semiparametric Regression Models // Biometrika, 1997. Vol. 84, № 2. P. 365–378.

8. *Bagdonavičius, V., Levuliene, R, Nikulin, M.S. & Tran, Q.X.* On Chi-square Type Tests and Their Applications in Survival Analysis and Reliability // Zapiski nauchnih seminarov POMI, 2012. Vol. 408. P. 43–61.

9. *Bagdonavičius, V. & Nikulin, M.S.* Chi-squared Goodness-of-fit Test for Right Censored Data // International Journal of Applied Mathematics and Statistics, 2011. Vol 24. P. 30–50.

10. *Cox D.R.* Regression models and life tables (with discussion) // Journal of the Royal Statistical Society, 1972. Ser. B, Vol 34. P. 187–220.

11. *Diggle P.J., Heagerty P., Liang K.-Y., Zeger S.L.* Longitudinal data analysis. 2-nd edition. New York: Oxford University Press Inc., 2002.

12. *Fahrmeir L., Kaufman H.* Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear model // The Annals of Statistics, 1985. Vol. 13, № 1. P. 342–368.

13. *Fay M.P.* Rank invariant tests for interval censored data under the grouped continuous model. Biometrics, 1996. Vol 52. P. 811–822.

14. *Fay M.P.* Comparing several score tests for interval-censored data // Statistics in Medicine, 1999. Vol. 18. P. 273–285.

15. *Fay M.P. & Shaw P.A.* Exact and asymptotic weighted logrank tests for interval censored data: the interval R Package // Journal of Statistical Software, 2010. Vol. 36, № 2. P. 1–34.

16. *Fleming T.R. & Harrington D.P.* Counting Processes and Survival Analysis. 2-nd edition. New Jersey: Wiley & Sons, Inc., 2005.

17. *Finkelstein D.M.* A proportional hazards model for interval censored failure time data // Biometrics, 1986. Vol. 42. P. 845–854.

18. *Gentleman R. & Vandal A.C.* Computational algorithms for censored data problems using intersection graphs. Journal of Computationaal nd Graphical Statistics, 2001. Vol 10. P. 403–421.

19. *Groeneboom P. & Wellner J.A.* Information Bounds and Nonparametric Maximum Likelihood Estimation. DMV Seminar Band 19. Basel: Birkhäuser, 1992.

20. *Habib M.G. & Thomas D.R.* Chi-Square Goodness-of-Fit Tests for Randomly Censored Data // The Annals of Statistics, 1986. Vol. 14, № 2. P. 759–765.

21. *Harrington D.P. & Fleming T.R.* A class of rank test procedures for censored survival data. Biometrika, 1982. Vol. 69. P. 553–566.

22. *Hedeker R.D. & Gibbons R.D.* Longitudinal data analysis. New Jersey: Wiley & Sons, Inc., 2006.

23. *Hershey A. & Chase M.* Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. // The Journal of General Physiology, 1952. Vol. 36, № 1. P. 39–56.

24. *Hollander & Pena.* A Chi-Squared Goodness-of-Fit Test for Randomly Censored Data // Journal of the American Statistical Association, 1992. Vol. 87, № 418. P. 458–463.

25. *Huang J., Lee C. & Yu Q.* A generalized log-rank test for interval-censored failure time data via multiple imputation // Statistics in Medicine, 2008. Vol. 27. P. 3217–3226.

26. *Huang J. & Wellner J.A.* Interval Censored Survival Data: A Review of Recent Progress // Proceedings of the First Seattle Symposium in Biostatistics. Lecture Notes in Statistics, 1997. Vol. 123. P. 123–169.

27. *Kalbfleisch J.D. & Prentice R.L.* The Statistical Analysis of Failure Time Data. 2-nd edition. Hoboken, New Jersey: Wiley & Sons, Inc., 2002.

28. *Lin D. Y. & Ying Z.* Semiparametrical analysis of the general additive-multiplicative hazard models for counting processes // The Annals of Statistics, 1996. Vol. 23. P. 1712–1734.

29. *Liang K.-Y. & Zeger S.L.* Longitudinal data analysis using generalized linear models // Biometrika, 1986. Vol. 73, № 1. P. 13–22.

30. *Malov S.V. & O'Brien S.J.* On Survival Categorical Methods with Applications in Epidemiology and AIDS Research. // Proceedings of the conference AMSA2013, 2013, in press.

31. *Nelder J.A. & Wedderburn R.W.M.* Generalized linear models // Journal of the Rojal Statistical Society, 1972. Vol. 135, № 3. P. 370–384.

32. *Oller R., Gómez G. & Calle M.L.* Interval censoring: model characterizations for the validity of the simplified likelihood // The Canadian Journal of Statistics, 2004. Vol. 32. P. 315–326.

33. *Oller R. & Gómez G.* A generalized Fleming and Harrington's class of tests for interval-censored data // The Canadian Journal of Statistics, 2012. Vol. 40, № 3. P. 501–516.

34. *Pollard K.S. & van der Laan M.J.* Resampling-based multiple testing: Asymptotic control of type I error and applications to gene expression data // J. Statist. Plann. Inference, 2002. Vol. 125. P. 85–100.

35. *Ridley M.* Genome: The Autobiography of a Species in 23 Chapters. New York, NY: Harper Perennial, 2006.

36. *Sun J.* A non-parametric test for interval-censored failure time data with applications to AIDS studies // Statistics in Medicine, 1996. Vol. 15. P. 1387–1395.

37. *Svitin et al.* Gene Discovery and Data Sharing in Disease Association Analyses Across the Genome. To appear.

38. *Turnbull B.W.* The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data // Journal of the Royal Statistical Society, 1976. Ser. B. Vol. 38. P. 290–295.

39. *Wang Z., Gardiner J.C. & Ramamoorthi R.V.* Identifiability in interval censorship model // Statistics & Probability Letters, 1994. Vol. 21. P. 215–221.

40. *Watson J.D. & Crick F.H.C.* A Structure for Deoxyribose Nucleic Acid // Nature, 1953. Vol. 171, № 4356. P. 737–738.

GENOME ASSOCIATIONS DISCOVERY. PART 1: STATISTICAL METHOD

Abstract

Methodology of genome association discovery is discussed comprehensively in this paper. In this part we consider main types of statistical data arises in genome association study experiments and wide range of statistical tests for these types of data analysis.

Keywords: Whole genome association discovery, genome wide association study (GWAS), categorical data, survival data, Cox model, longitudinal data, generalized linear models.

Малов Сергей Васильевич,
кандидат физико-математических наук, доцент, старший научный сотрудник лаборатории «Центр геномной биоинформатики им. Ф.Г. Добржанского»,
malovs@sm14820.spb.edu,

Шевченко Андрей Константинович,
лаборант-исследователь лаборатории «Центр геномной биоинформатики им. Ф.Г. Добржанского»,
andrey.k.shevchenko@gmail.com

О'Брайен Стефан Джеймс (Stephen J. O'Brien),
доктор философии в области биологии (PhD in biology), главный научный сотрудник – научный руководитель лаборатории «Центр геномной биоинформатики им. Ф.Г. Добржанского»,
lgdchief@gmail.com.