Санкт-Петербургский Государственный Университет
Факультет Прикладной математики - процессов управления

Кектеева Ангира Игоревна

СТАТИСТИЧЕСКИЙ АНАЛИЗ КАТЕГОРИАЛЬНЫХ ДАННЫХ ГЕНЕТИЧЕСКИХ ИССЛЕДОВАНИЙ

Выпускная работа бакалавра

Научный руководитель:
д.ф.-м.н, профессор
Андрианов С.Н.

Рецензент:
ст. преп. Севрюков С.Ю.

Санкт-Петербург — 2017

Saint-Petersburg State University
Faculty of Applied Mathematics and Control Processes

Angira Kekteeva

STATISTICAL ANALYSIS OF CATEGORIAL DATA OF GENETIC RESEARCH

Bachelor's Thesis

Научный руководитель:
D.Sc., Professor
Andrianov Sergey N.

Reviewer:
Senior Lecturer
Sevryukov Sergei Yu.

Saint-Petersburg — 2017

Оглавление

Стр.

Введение

Одним из крупнейших научных достижений на рубеже XX и XXI веков является разработка метода секвенирования ДНК, предложенного Ф.Сенгером в 1977 году [1]. Позднее с помощью этого метода был прочитан первый геном, основанный на ДНК – геном бактеориофага $\varphi X174$ длиной в 5 386 нуклеотидов. В 1986 году лаборатория Лероя Худа улучшила метод Сенгера, что дало возможность разработать первый полу-автоматический секвенатор [2]. Этот способ будет использован для многих проектов секвенирования полных геномов. Первым клеточным организмом, геном которого был полностью прочитан, стала бактерия Haemophilus influenzae или гемофильная палочка [3], вызывающая некоторые формы пневмонии и менингита. Длина генома этой бактерии составляет 1 830 137 нуклеотидов. Затем, в 1998-ом году секвенируется первый геном многоклеточного животного, круглого червяка Caenorhabditis elegans [4] длиной в 98 000 000 нуклеотидов. Геном резуховидки Таля становится первым секвенированным геномом растения в 2000 году [5], его длина составляет уже 157 000 000 нуклеотидов. Скорость и масштабы секвенирования росли все быстрее и быстрее, а базы данных геномов пополнялись все с большей скоростью.

В 1990 году под руководством Джеймса Уотсона был запущен проект «Геном человека» (The Human Genome Project, HGP) - международный научно-исследовательский проект, цель которого заключалась в определении последовательности нуклеотидов, которые составляют ДНК человека и в идентификации 20-25 тысяч генов в геноме. [6]. В результате исследователи смогли секвенировать 99% человеческой ДНК длиной $3{,}1467 \times 10^6$ нуклеотидов. В дальнейшем темпы секвенирования только возрастали.

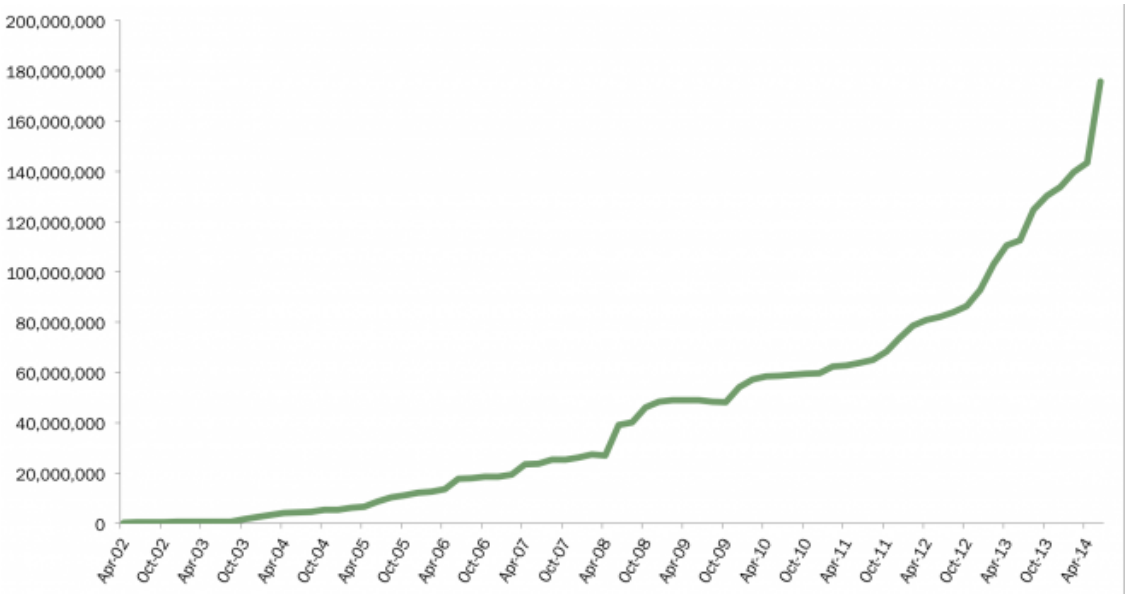

Рис. 1 — Количество полностью секвенированных геномов в базе данных GenBank.

Несомненные успехи в секвенировании геномов поспособствовали бурному развитию таких областей науки, как молекулярная биология, эволюционная биология, медицина и т.д. Одним из методов исследования данных в перечисленных разделах биологии является полногеномный скрининг ассоциаций (Genome-Wide Association Studies, GWAS), который заключается в исследовании зависимости между геномными вариантами и фенотипическими признаками [7].

Геномы всех людей различны. Это могут быть как однонуклеотидные полиморфизмы [1], так и более крупные изменения. Любое из этих различий может отвечать за популяционные, этнические или индвивидуальные особенности индивида. Полногеномный скрининг ассоциаций основан на проведении множества статистических тестов, с помощью которых проводится анализ частоты аллелей различных генов среди индивидуумов. Если при сравнении те или иные аллели генов встречаются у людей с определенным фенотипом значимо чаще, чем у других, то есть основания предполагать, что именно эти аллели ответственны за проявление этого фенотипа.

В связи с бурным ростом объема статистических данных, полученных в результате генетических исследований, целесообразно создание программного продукта для их анализа с использованием статистических критериев.

[1] 8

Постановка задачи

Цель данной работы заключается в реализации наиболее востребованных статистических методов множественного тестирования для изучения зависимости фенотипа с каждым из имеющихся генетических маркеров [2] на языке, допускающем эффективные вычисления. Для достижения цели были сформулированы следующие задачи:

1) Разработка программного продукта, проводящего статистический анализ генетических данных с помощью критерия Хи-квадрат, точного теста Фишера и коэфициента корреляции. Вычисления должны выполняться в модуле, написанном на языке программирования C++, и вызываемом из программы, реализованной на языке R.
2) Реализация программы на языке программирования R , выполняющей множество выше описанных статистических тестов, с использованием стандартных методов используемого языка.
3) Сравнение полученных результатов вычислений и скорости работы собственного R-пакета и стандартных функций языка R.

Исходные данные

Данные генетических исследований состоят из генетической и клинической части. Генетические данные, с которыми проводится работа, представляют собой набор значений генетических маркеров, измеренных у каждого индивида, каждое из которых представляет собой категориальную величину с определенным числом уровней. Простейшие генетические маркеры представляют собой категориальные величины с тремя уровнями, значения которых определяются парами бинарных величин, характеризующих наличие мутации в соответствующей позиции на одной или обеих хромосомах без учета порядка. Для генетических данных с тремя уровнями проводится четыре теста, соответствующие кодоминантной, доминантной и рецессивной альтернативам, а также, с использованием аллельного подхода.

[2] 9

Глава 1. Обзор предметной области

В данной главе описывается методы статистического анализа, реализованные в программе, а также вводятся основные понятия и общие принципы теории проверки гипотез.

§1.1 Основные понятия

Определение 1. *Распределение* χ^2 *(хи-квадрат) с* k *степенями свободы* [8] – это распределение суммы квадратов n независимых стандартных нормальных величин. То есть, пусть $X_1, X_2, \ldots, X_k$ – независимые случайные величины, $X_i \sim N(0,1)$, тогда случайная величина $X = X_1^2 + X_2^2 + \cdots + X_k^2$ имеет распределение χ^2 с k степенями свободы. Является частным случаем гамма-распределения.

Функция плотности распределения имеет вид:

$$f_{\chi^2(k)}(x) \equiv \Gamma(2, \frac{k}{2}) = \frac{(1/2)^{\frac{k}{2}}}{\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}},$$

функция распределения:

$$F_{\chi^2(k)}(x) = \frac{\gamma(\frac{k}{2}, \frac{x}{2})}{\Gamma(\frac{k}{2})} \tag{1}$$

где Γ и γ — полная и нижняя неполная гамма-функции соответственно, имеющие вид:

1) полная:

$$\Gamma(\lambda) = \int_0^\infty t^{\lambda-1} e^{-t} dt, \qquad \lambda > 0,$$

2) нижняя неполная:

$$\gamma(\lambda, x) = \int_0^x t^{\lambda-1} e^{-t} dt, \qquad \lambda > 0.$$

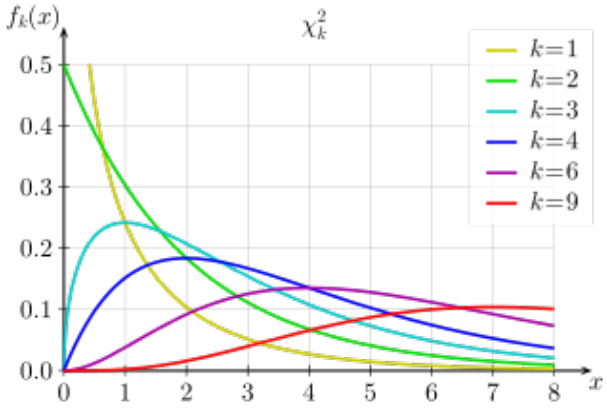

Рис. 2 — Плотность распределения χ^2

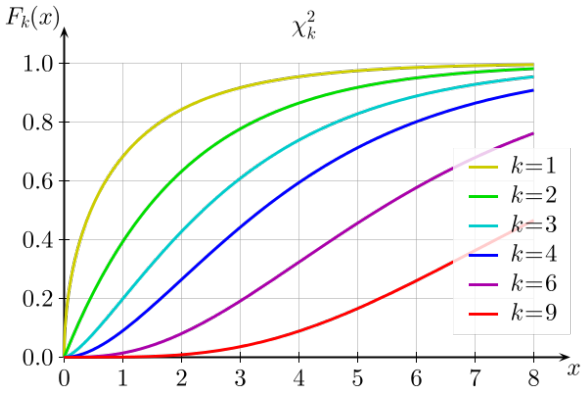

Рис. 3 — Функция распределения χ^2

О п р е д е л е н и е 2. *Мультиномиальное (полиномиальное) распределение* $M(n, \mathbf{p})$ [9] – совместное распределение вероятностей независимых случайных величин $\xi_1, \ldots, \xi_k$, принимающих целые неотрицательные значения $n_1, \ldots, n_k$, удовлетворяющие условиям $n_1 + \cdots + n_k = n$, с вероятностями

$$\mathbf{P}(\xi_1 = n_1, \ldots, \xi_k = n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k}, \tag{2}$$

где $p_i \geqslant 0, \sum\limits_{i=1}^{n} p_i = 1$; является многомерным дискретным распределением случайного вектора $(\xi_1, \ldots, \xi_k)$ такого, что $\xi_1 + \cdots + \xi_n = n$; естественным образом обобщает биномиальное распределение и совпадает с ним при $n = 2$.

Мультиномиальное распределение появляется в так называемой *мультиномиальной схеме* случайных экспериментов: результатом каждого эксперимента является одно из взаимоисключающих событий $A_1, \cdots, A_k$. Проводится n экспериментов и считается число событий каждого типа. Случайная величина ξ_i - это число наступлений события A_i в n экспериментах. Иными словами, результат каждого эксперимента можно представить в виде случайной $\eta \in 1, \cdots, k$. Пусть $\eta_1, \cdots .\eta_n$ – выборка из распределения η, тогда

$$\xi_i = \sum_{j=1}^{n} \mathbb{1}\{\eta_j = i\}$$

$$\mathbb{1}_A = \begin{cases} 1, & A - \text{выполнена} \\ 0, & A - \text{не выполнена} \end{cases}$$

Если в каждом эксперименте вероятность наступления события A_j равна p_j, то мультиномиальная вероятность (2) равна вероятности того, что в n независимых экспериментах события $A_1, \ldots, A_k$ наступят $n_1, \ldots, n_k$ раз соответственно. Каждая из случайных величин ξ_i имеет биномиальное распределение с математическим ожиданием np_i и дисперсией $np_i(1-p_i)$.

Определение 3. *Статистической гипотезой*[10] H называют утверждение о виде или свойствах распределения наблюдаемых случайных величин:

$$H : \mathcal{F} = \mathcal{F}_\xi \quad \text{или} \quad H : \mathcal{F} \in \mathbb{F}$$

где $\mathbb{F}$ – некоторое подмножество в множестве распределений. Гипотеза H называется *простой*[10], если она фиксирует распределение: $\mathcal{F} = \mathcal{F}_\xi$, т.е. $\mathbb{F} = \{\mathcal{F}_\xi\}$. Иначе H называется *сложной*: $\mathcal{F} \in \mathbb{F}$.

Если для исследуемого явления сформулирована та или иная гипотеза (называемая *нулевой или основной гипотезой* и обозначаемая H_0), то задача заключается в том, чтобы сформулировать правило, которое позволило бы по имеющимися статистическим данным принять или отклонить нулевую гипотезу. Правило, согласно которому отклоняется или принимается гипотеза H_0, называется *статистическим критерием.*

Одним из способов проверки гипотезы является сравнение P - значения, соответствующего статистике критерия, с ранее заданным уровнем значимости.

Определение 4. *Уровень значимости* статистического критерия - допустимая для данной задачи вероятность отвергнуть гипотезу H_0, когда она верна. Наиболее распространенные значения $\alpha = 0.005,\ 0.01,\ 0.05,\ 0.1$.

Определение 5. P-значение статистического критерия – наименьшее значение уровня значимости при котором основная гипотеза отвергается.

Пусть $T(X)$ - статистика критерия, используемая для проверки нулевой гипотезы H_0. Статистический критерий построен следующим образом: гипотеза H_0 принимается при уровне значимости α, если $T \leqslant x_\alpha$ и отвергается, если $T > x_\alpha$. Предполагается, что при справедливости гипотезы H_0

распределение статистики $T(X)$ известно. Обозначим функцию распределения статистики T при справедливости H_0 $F(t) = P(T \leqslant t)$. При проверке альтернативы P - значение определяется как:

$$P(T) = 1 - F(T). \tag{3}$$

Подставим (1) в (3)

$$P(x) = P(X > x) = 1 - \frac{\gamma(\frac{k}{2}, \frac{x}{2})}{\Gamma(\frac{k}{2})}, \tag{4}$$

где $\Gamma(\frac{k}{2})$ и $\gamma(\frac{k}{2}, \frac{x}{2})$ — полная и нижняя неполная гамма-функции соответственно.

Таким образом, *P-значение* [11] есть случайная величина, зависящая от статистики критерия.

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из распределения $\mathcal{L}(\xi)$ с неизвестной функцией распределения $F_\xi(x)$, о которой выдвинута простая гипотеза $H_0 : F_\xi(x) = F(x)$.

§1.2 Критерий согласия χ^2 К. Пирсона

Одним из методов проверки различных статистических гипотез является *критерий* χ^2 [12]. Он может использоваться для любых распределений, в том числе многомерных. Перед тем, как воспользоваться этим критерием, выборочные данные предварительно группируют, т.е. область значений предполагаемого распределения $F(x)$ делят на множество непересекающихся интервалов $\{\Delta_i\}_{i=1}^k$. Тогда ν_i - число элементов выборки, попавших в интервал Δ_i : $\nu_i = \#\{i : X_i \in \Delta_i\}$ – число X_i, попавших в Δ_i, $(\nu_1 + \cdots + \nu_k = n)$. Обозначим p_i как вероятность попадания в интервал Δ_i : $p_i = \mathbf{P}(\xi \in \Delta_i)$. Пусть $\nu = (\nu_1, \ldots, \nu_k)$ – вектор частот попадания выборочных точек в соответствующие интервалы группировки и $p^0 = (p_1^0, \ldots, p_k^0)$, где $p_i^0 = P(\xi \in \Delta_i | H_0), i = 1, \ldots, k$. Тогда $\nu = (\nu_1, \ldots, \nu_k)$ имеет мультиномиальное распределение, и при справедливости гипотезы H_0 вероятности мультиномиального распределения построенного вектора частот ν имеют заданные значения $p_i^0, i = 1, \ldots, k$. В качестве статистики, характеризующей отклонение выборочных данных (т.е. частот ν_i) от

соответствующих гипотетических значений принимают величину:

$$X_n^2 = X_n^2(\nu) = \sum_{i=1}^{k} \frac{(\nu_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^{k} \frac{\nu_i^2}{np_i^0} - n. \tag{5}$$

На этой статистике и основывается критерий χ^2 К.Пирсона. В основе этого лежат следующие соображения. Если гипотеза H_0 справедлива, то, поскольку частота $\frac{\nu_i}{n}$ события $\{\xi = i\}$ является состоятельной оценкой его вероятности $p_i^0 (i = 1, \ldots, k)$, то разности $|\nu_i/n - p_i^0|$ должны быть малы, следовательно, и значение статистики χ^2 не должно быть слишком большим. Поэтому естественно задать критическую область для гипотезы H_0 в виде $\{\chi_n^2 > t_\alpha\}$, где критическая граница t_α при заданном уровне значимости выбрана из условия:

$$P\{X_n^2 > t_\alpha | H_0\} = \alpha$$

Такой критической областью и определяется критерий χ^2. Главная проблема здесь — вычисление границы t_α, для чего надо знать распределение статистики X_n^2 при нулевой гипотезе H_0. Точное распределение $\mathcal{L}(X_n^2|H_0)$, зависящее от набора $p_1, \ldots, p_k$, неудобно для расчета критерия, но для достаточно больших объемов выборок n при справедливости гипотезы H_0 статистика X_n^2 имеет простое предельное распределение при нулевой гипотезе H_0. А именно, справедливо следующее утверждение:

Т е о р е м а 1. Если $0 < p_i^0 < 1, i = 1, \ldots, k$, то при $n \to \infty$

$$\mathcal{L}(X_n^2|H_0) \to \chi_{k-1}^2$$

Таким образом, критерий согласия χ^2 можно сформулировать следующим образом:

Пусть заданы уровень значимости α и объем выборки n и наблюдавшиеся значения $\mathbf{h} = (h_1, \ldots, h_k)$ вектора частот $\nu = (\nu_1, \ldots, \nu_k)$; тогда если наблюдавшееся значение $T = X_n^2(\mathbf{h})$ статистики (5) удовлетворяет неравенству $T > t_\alpha$ такого, что $F_{\chi_{k-1}^2}(t_\alpha) = 1 - \alpha$, то гипотезу H_0 отвергают; в противном случае гипотеза H_0 не противоречит результатам испытаний.

§1.3 Критерий согласия χ^2 для сложной гипотезы

Критерий согласия χ^2 применим и для проверки гипотезы о принадлежности неизвестной функции распределения наблюдаемой случайной величины ξ заданному семейству функций распределения [10]. Таким образом, задача формулируется так: пусть $\mathscr{F} = \{F(x,\theta), \theta \in \Theta\}$ - заданное параметрическое семейство функций распределения (параметр θ может быть как скалярным, так и векторным) и $\mathbf{X} = (X_1, \ldots, X_n)$ – выборка из распределения $\mathcal{L}(\xi)$ с неизвестной функцией распределения. Проверяется сложная гипотеза

$$H_0 : \mathcal{F} \in \{\mathcal{F}_\theta; \theta \in \Theta \subseteq \mathbb{R}^d\},$$

где θ – неизвестный параметр, d – его размерность.

Пусть исходные данные сгруппированы и $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_k)$ – соответствующий вектор частот попадания наблюдений в интервалы группировки. Составим статистику, аналогичную (5). В данном случае вероятности попадания в интервалы группировки при гипотезе H_0 уже не будут заданы однозначно, а представляют собой некоторые функции от параметра θ:

$$p_i(\theta) = P_\theta(X_i \in \Delta_i), \quad \theta \in \Theta$$

Поэтому значение X^2 имеет вид:

$$X_n^2 = X_n^2(\theta) = \sum_{i=1}^{k} \frac{(\boldsymbol{\nu}_j - np_j(\theta))^2}{np_j(\theta)} \tag{6}$$

Оно зависит от неизвестного параметра; следовательно, непосредственно использовать его в качестве статистики для построения критерия нельзя – требуется предварительно исключить в (7) неопределенность, связанную с неизвестным параметром θ. Для этого поступают следующим образом: заменяют θ некоторой оценкой $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X})$ и получают, таким образом, статистику:

$$\hat{X}_n^2 = X_n^2(\hat{\theta}) = \sum_{i=1}^{k} \frac{(\boldsymbol{\nu}_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})}. \tag{7}$$

Значение этой статистики можно однозначно вычислить для каждой заданной реализации выборки $\mathbf{X}$.

Если бы распределение статистики $\hat{X}_n^2$ при гипотезе H_0 можно было бы найти и при этом распределение не зависело бы от конкретных функций $F(x;\theta)$, составляющих гипотезу H_0, то, основываясь на $\hat{X}_n^2$, можно было бы построить критерий согласия для гипотезы H_0.

В данном случае величины $p_i(\hat{\theta})$ уже не постоянные, а представляют собой функции от выборки (случайные величины). Поэтому теорема 1 к статистике X_n^2 будет неприменима. Более того, следует ожидать, что распределение этой статистики (даже если оно существует) будет зависеть от способа построения оценки $\hat{\theta}_n$.

Проблема нахождения предельного распределения для $\hat{X}_n^2$ при этих усложненных условия впервые была рассмотрена в 1924 г. Р.Фишером, который показал, что существуют методы оценивания параметра θ, при которых предельное распределение имеет простой вид, а точнее, является распределением χ^2 с числом степеней свободы $k-1-r$, где r – размерность оцениваемого параметра θ. Одним из таких методов оценивания является метод максимального правдоподобия, основанный на частотах $\nu_1,\dots,\nu_k$, т.е. когда в качестве $\hat{\theta}_n$ в формуле (7) используют мультиномиальную оценку максимального правдоподобия.

Т е о р е м а 2. Пусть функции $p_i(\theta), i=1,\dots,k$ удовлетворяют условиям:

1) $\sum\limits_{i=1}^{k} p_i(\theta) = 1, \forall\theta\in\Theta$;

2) $p_i(\theta) \geqslant c > 0, \forall i$ и существуют непрерывные производные

$$\frac{\delta p_i(\theta)}{\delta\theta_j}, \frac{\delta^2 p_i(\theta)}{\delta\theta_j\delta\theta_l}, j,l = 1,\dots,r;$$

3) $(k\times r)$ - матрица $\left\|\frac{\delta p_i(\theta)}{\delta\theta_j}\right\|$ имеет ранг r для всех $\theta\in\Theta$.

Тогда

$$\mathcal{L}(\hat{X}_n^2|H_0) \to \chi^2_{k-r-1}$$

§1.4 Критерий χ^2 для проверки гипотезы о независимости двух признаков

Пусть данные представляют собой выборку (X_i, Y_i), $i = 1, \ldots, n$, из распределения некоторой двумерной случайной величины $\xi = (\xi_1,\xi_2)$ с неизвестной функцией распределения $F_\xi(x,y)$, для которой требуется проверить основную гипотезу о независимости компонент $H_0 : F_\xi(x,y) = F_{\xi_1}(x) \cdot F_{\xi_2}(y)$, где $F_{\xi_i}(x), i = 1, 2$ – некоторые одномерные функции распределения.

Разобьем [13] множество значений ξ_1 на s непересекающихся интервалов $\{\Delta_i^x\}_{i=1}^s$ и множество значений ξ_2 – на m сегментов$\{\Delta_j^y\}_{j=1}^m$, таким образом, множество значений двумерной величины $\xi = (\xi_1,\xi_2)$ разбито на $N = mk$ прямоугольников $\{\Delta_i^x \times \Delta_j^y\}_{i,j=1}^{m,k}$. Обозначим через ν_{ij} число элементов выборки, принадлежащих прямоугольнику $\Delta_i^x \times \Delta_j^y$, так что

$$\sum_{i=1}^{s}\sum_{j=1}^{m} \nu_{ij} = n.$$

Результаты наблюдений удобно представлять в виде *таблицы сопряженности* двух признаков (ξ_1 и ξ_2 обычно означают два признака, по которым производится группировка результатов наблюдений, и тогда H_0 есть гипотеза о независимости этих признаков).

Таблица 1 — Таблица сопряженности

ξ_1 \ ξ_2	Δ_1^y	Δ_2^y	$\cdots$	Δ_m^y	$\sum \nu_{i\bullet}$
Δ_1^x	ν_{11}	ν_{12}	$\cdots$	ν_{1m}	$\nu_{1\bullet}$
Δ_2^x	ν_{21}	ν_{22}	$\cdots$	ν_{2m}	$\nu_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Δ_s^x	ν_{s1}	ν_{s2}	$\cdots$	ν_{sm}	$\nu_{s\bullet}$
$\sum \nu_{\bullet j}$	$\nu_{\bullet 1}$	$\nu_{\bullet 2}$	$\cdots$	$\nu_{\bullet m}$	

Пусть $p_{ij} = P(\xi_1 \in \Delta_i^x, \xi_2 \in \Delta_j^y), i = \overline{1,s}, j = \overline{1,k}$ – теоретические вероятности попадания пары (ξ_1, ξ_2) в любую из областей $\Delta_i^x \times \Delta_j^y$. Имеет смысл заменить исходную гипотезу H_0 на категориальную гипотезу независимости $H_0^\Delta : p_{ij} = p_{i\bullet} \cdot p_{\bullet j} \quad i = 1, \cdots, s, j = 1, \cdots, m$, где $p_{i\bullet}, p_{\bullet j}$ – некоторые значения, удовлетворяющие условию:

$$\sum_{i=1}^{s} p_{i\bullet} = \sum_{j=1}^{k} p_{\bullet j} = 1, \quad \begin{cases} p_{i\bullet} \geqslant 0 \\ p_{\bullet j} \geqslant 0 \end{cases}$$

В общем случае,

$$p_{i\bullet} = \sum_{j} p_{ij}, \quad p_{\bullet j} = \sum_{i} p_{ij}$$

При гипотезе H_0 частоты $(\nu_{ij}, i = 1, \ldots, s, j = 1, \ldots, m)$ распределены по мультиномиальному закону с вероятностями исходов p_{ij} определяющимися значениями $r = m + s - 2$ неизвестных параметров $p_{i\bullet}, i = 1, \ldots, s-1, \quad p_{\bullet j}, j = 1, \ldots, m-1$.

Для проверки этой гипотезы, следовательно, можно применить критерий согласия χ^2 для проверки сложных параметрических гипотез с $m+s-2$ параметрами $p_{i\bullet}, p_{\bullet j}, i = 1, \ldots, s-1, j = 1, \ldots, m-1$. Найдем мультиномиальные оценки максимального правдоподобия для определяющих рассматриваемую модель параметров при справедливости гипотезы H_0, то есть максимизируя в данном случае функцию правдоподобия:

$$\frac{n!}{\nu_{11}! \ldots \nu_{sm}!} \prod_{i,j} (p_{i\bullet} p_{\bullet j})^{\nu_{ij}} = \frac{n!}{\nu_{11}! \ldots \nu_{sm}!} \prod_{i} p_{i\bullet}^{\nu_{i\bullet}} \prod_{j} p_{\bullet j}^{\nu_{\bullet j}}$$

по параметрам $p_{i\bullet}, p_{\bullet j}$, получим оценки для них:

$$\hat{p}_{i\bullet} = \frac{\nu_{i\bullet}}{n}, \qquad \hat{p}_{\bullet j} = \frac{\nu_{\bullet j}}{n} \tag{8}$$

т.е. это выборочные оценки параметров $p_{i\bullet} = P(\xi_1 \in \Delta_i^x), p_{\bullet j} = P(\xi_2 \in \Delta_j^y)$, при этом

$$\mathcal{L}(\nu_{1\bullet}, \ldots, \nu_{s\bullet}) = M(n; p_{1\bullet}, \ldots, p_{s\bullet}),$$
$$\mathcal{L}(\nu_{\bullet 1}, \ldots, \nu_{\bullet k}) = M(n; p_{\bullet 1}, \ldots, p_{\bullet k}).$$

Тогда статистика критерия χ^2 с числом степеней свободы $(s-1)(k-1)$ принимает вид:

$$\chi_n^2 = \sum_{i,j} \frac{(\nu_{ij} - n\hat{p}_{i\bullet}\hat{p}_{\bullet j})^2}{n\hat{p}_{i\bullet}\hat{p}_{\bullet j}} = n \sum_{i=1}^{s} \sum_{j=1}^{k} \frac{(\nu_{ij} - \nu_{i\bullet} \cdot \nu_{\bullet j}/n)^2}{\nu_{i\bullet} \cdot \nu_{\bullet j}}. \tag{9}$$

О п р е д е л е н и е 6. Значение

$$E_{ij} = \frac{\nu_{i\bullet} \cdot \nu_{\bullet j}}{n} \tag{10}$$

называется *ожидаемым значением*, тогда как ν_{ij} — *наблюдаемым*

Таким образом, гипотезу H_0 отвергают тогда и только тогда, когда вычисленное по наблюдаемым данным значение t статистики (9) удовлетворяет неравенству: $t \geqslant \chi^2_{1-\alpha,(s-1)(k-1)}$, где α - заданный уровень значимости. [10]

Так как для применения критерия производится замена распределения статистики X_n^2 при нулевой гипотезе H_0 на более более простое предельное распределение χ^2_{k-1} (см. теорему 1), т.е. критерий χ^2 является асимптотическим, то для корректного результата статистического теста необходимо, чтобы в анализируемой таблице сопряженности ожидаемые значения ячеек были больше 5. Если число наблюдений мало, целесообразно применение точного критерия Фишера.

§1.5 Точный критерий Фишера

Критерий Фишера в основном применяется для исследования значимости взаимосвязи между двумя категориальными переменными в таблице сопряженности размерности 2×2 [11]. Тест основан на переборе всех возможных вариантов заполнения таблицы и потому является точным (что и отражено в его названии), что позволяет использовать его в исследованиях с небольшим количеством наблюдений.

Вид таблицы сопряженности:

Таблица 2 — Таблица сопряженности T_0

ξ_1 \ ξ_2	Δ_1^y	Δ_2^y	$\sum$
Δ_1^x	ν_{11}	ν_{12}	$\nu_{1\bullet}$
Δ_2^x	ν_{21}	ν_{22}	$\nu_{2\bullet}$
$\sum$	$\nu_{\bullet 1}$	$\nu_{\bullet 2}$	n

Требуется проверить категориальную нулевую гипотезу о независимости переменных ξ_1 и ξ_2.

Как было сказано ранее, точный критерий Фишера основан на переборе всех возможных вариантов таблицы сопряженности T_i при фиксированных значениях сумм $\nu_{1\bullet} = n^i_{11} + n^i_{12}$, $\nu_{2\bullet} = n^i_{21} + n^i_{22}$, $\nu_{\bullet 1} = n^i_{11} + n_{21}$, $\nu_{\bullet 2} = n^i_{12} + n^i_{22}$.

Таблица 3 — Всевозможные варианты T_i таблицы сопряженности

ξ_1 \ ξ_2	Δ^y_1	Δ^y_2	$\sum$
Δ^x_1	n^i_{11}	n^i_{12}	$\nu_{1\bullet}$
Δ^x_2	n^i_{21}	n^i_{22}	$\nu_{2\bullet}$
$\sum$	$\nu_{\bullet 1}$	$\nu_{\bullet 2}$	n

Вероятность получения n^i_{11} наблюдений в первой ячейке таблицы T_i подчиняется гипергеометрическому распределению:

$$P_{n^i_{11}} = \frac{\binom{\nu_{1\bullet}}{n^i_{11}}\binom{\nu_{2\bullet}}{n^i_{21}}}{\binom{n}{\nu_{\bullet 1}}} = \frac{\nu_{1\bullet}!\,\nu_{2\bullet}!\,\nu_{\bullet 1}!\,\nu_{\bullet 2}!}{n!\,n^i_{11}!\,n^i_{12}!\,n^i_{21}!\,n^i_{22}!} \tag{11}$$

Критерий Фишера основан на расчете вероятности:

$$P = \sum_{T_i : P_{n^i_{11}} \leqslant P_{\nu_{11}}, i \neq 0} P_i + P_0, \tag{12}$$

где P_0 - вероятность получения исходной таблицы T_0, а P_i – таблицы T_i при фиксированных $\nu_{1\bullet}, \nu_{2\bullet}, \nu_{\bullet 1}, \nu_{\bullet 2}$.

Вероятность P и есть P - значимость различий сравниваемых групп. Полученное значение необходимо сопоставить с критическим уровнем значимости α. Если P не превосходит α, то нулевая гипотеза о независимости отклоняется. В противном случае отклонять нулевую гипотезу нет оснований.

§1.6 Коэффициент корреляции

В реализованном пакете R производится также вычисление коэффициента корреляции.

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ и $\mathbf{Y} = (Y_1, \ldots, Y_n)$ - – выборки из распределений некоторых случайных величин ξ_1 и ξ_2 соответственно.

О п р е д е л е н и е 7. *Коэффициентом корреляции* [14] $\rho(\xi_1,\xi_2)$ для любых случайных величин ξ_1 и ξ_2, дисперсии которых существуют и отличны от нуля, называется функция

$$\rho(X,Y) = \frac{\sum\limits_{ij}(X_i - \overline{X})(Y_j - \overline{Y})}{\sqrt{\sum\limits_{i}(X_i - \overline{X})^2}\sqrt{\sum\limits_{j}(Y_j - \overline{Y})^2}},$$

где $\overline{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$, $\overline{Y} = \frac{1}{n}\sum\limits_{i=1}^{n} Y_i$ – выборочные средние.

Коэффициент корреляции позволяет установить характер (направление) зависимости двух величин. Например, влечет ли увеличение величины ξ_1 увеличение ξ_2.

При анализе таблиц сопряженности вида:

Таблица 4 — Таблица сопряженности

ξ_1 \ ξ_2	b_1	b_2	$\cdots$	b_m	$\sum \nu_{i\bullet}$
a_1	ν_{11}	ν_{12}	$\cdots$	ν_{1m}	$\nu_{1\bullet}$
a_2	ν_{21}	ν_{22}	$\cdots$	ν_{2m}	$\nu_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_s	ν_{s1}	ν_{s2}	$\cdots$	ν_{sm}	$\nu_{s\bullet}$
$\sum \nu_{\bullet j}$	$\nu_{\bullet 1}$	$\nu_{\bullet 2}$	$\cdots$	$\nu_{\bullet m}$	

имеет смысл считать $\rho(\xi_1, \xi_2)$ по группированным данным по формуле:

$$\rho(X,Y) = \frac{\frac{1}{n}\sum\limits_{i,j}(\nu_{ij}\cdot a_i \cdot b_j) - \overline{X}^* \cdot \overline{Y}^*}{\sqrt{\frac{1}{n}\sum\limits_{i}\nu_{i\bullet}(a_i - \overline{X}^*)^2 \frac{1}{n}\sum\limits_{j}\nu_{\bullet j}(j - \overline{Y}^*)^2}} \tag{13}$$

где

$$\overline{X}^* = \frac{1}{n}\sum_{i}(\nu_{i\bullet}\cdot a_i),$$

$$\overline{Y}^* = \frac{1}{n}\sum_{j}(\nu_{\bullet j}\cdot b_j).$$

Глава 2. Детали реализации программы

§2.1 Описание используемого языка и технологий программирования

Выполнение вычислений

Вычисления внутри реализованного в ходе работы пакета R, решающие поставленную задачу, были написаны на языке программирования C++ как на языке, допускающем эффективные вычисления. Для реализации запуска C++ кода из программы, написанной на языке R, были использованы R-пакет Rcpp и одноименная библиотека C++ [15].

Поскольку выполнение статистических тестов предполагает многократное выполнение однотипных операций, для повышения эффективности вычислений была использована технология OpenMP («Open Multi-Processing») - стандарт для создания многопоточных программ, выполняемых на многопроцессорных системах с общей памятью [16].

Для проверки корректности вычислений, а также для сравнения скорости работы была написана программа на языке программирования R, выполняющая вычисления статистических тестов с помощью стандартных функций языка.[17]

Для распараллеливания вычислений, выполняемых программой, написанной на R, были использованы пакеты программной среды R, размещенные в репозитории CRAN («Comprehensive R Archive Network»): foreach - пакет, позволяющий выполнять действия каждой итерации цикла foreach параллельно, doParallel - вспомогательный пакет, позволяющий пакету «foreach» запускать итерации одновременно [18].

Разработка, сборка и компиляция программы велись на операционной системе Debian GNU/Linux, поскольку в качестве среды реализации предполагается использовать ОС Linux.

Обработка входных данных и запись результатов вычислений

Данные представляют собой клинические данные k индивидов. Каждому индивиду сопоставляется целочисленное значение в диапазоне $0, \ldots, k-1$. Таким образом, одно наблюдение содержит:

1) идентификационный номер индивида
2) фенотип — категориальную величину с уровнями $0, \ldots, k-1$
3) генотип — категориальную величину с уровнями $0, \ldots, 3$, где 0, 1, 2 соответствуют различным комбинациям аллелей, а значение 3 означает отсутствие данных о генотипе.

Входные данные предоставляются в виде файлов в формате .gds («Genomic Data Structure») - компактном формате, предназначенном для хранения генотипных данных. В R работа с .gds - файлами реализована в пакете «gdsfmt» [19], с помощью которого в написанном в ходе работы приложении проводится чтение файла и обработка извлеченных данных для дальнейшей их передачи в модуль, реализованный на языке C++.

Результат вычислений записывается файл в формате .csv («Comma-Separated Values») - текстовый формат, предназначенный для хранения и представления табличных данных [20].

§2.2 Алгоритм работы программы

Общая схема работы R-пакета (рис. 4):

Рис. 4 — Общая схема программы

Как было сказано выше, извлечение данных из .gds файла происходит в R-программе с помощью пакета gdsfmt, в дальнейшем они обрабатываются и передаются в модуль для проведения вычислений.

На вход модуля данные подаются в виде (рис. 5):

1) вектора длины n, элементы которого являются фенотипами,
2) матрицы размером $n \times m$, где каждая i-ая строка соответствует i-му фенотипу из вектора.

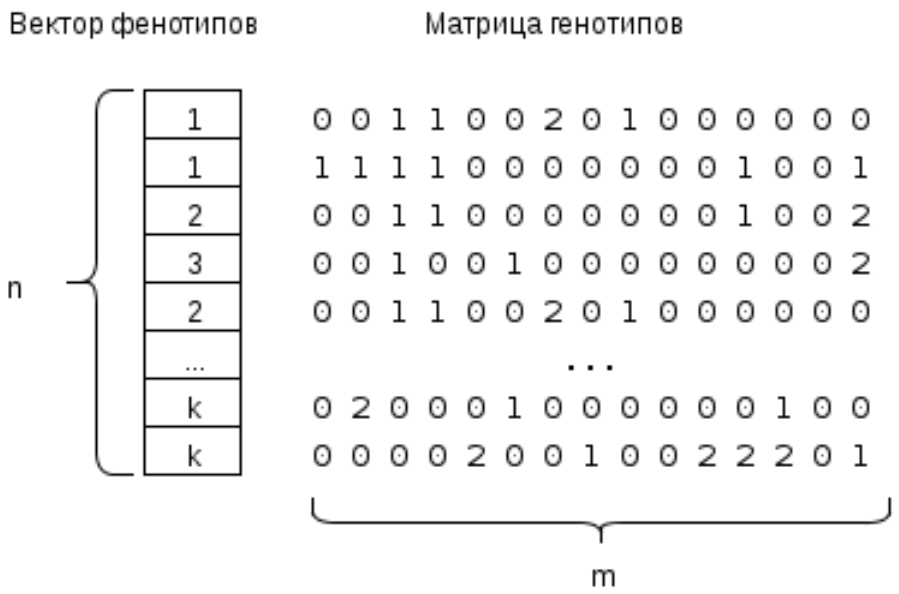

Рис. 5 — Представление данных о генотипах и фенотипах

Статистические тесты проводятся для каждого столбца матрицы генотипов отдельно.

Затем, для каждого столбца производится следующий порядок действий:

1) Для вектора фенотипов и одного столбца матрицы генотипов составляются 4 таблицы сопряженности, соответствующие кодоминантной, доминантной, рецессивной альтернативам, а так же аллельному подходу.

Таблица 5 — Таблица сопряженности для кодоминантной альтернативы

Фенотип	Генотип		
	0	1	2
0	n_{00}	n_{01}	n_{02}
1	n_{10}	n_{11}	n_{12}
...	...	...	...
k	n_{k0}	n_{k1}	n_{k2}

Таблица 6 — Остальные таблицы сопряженности: доминантная - (a), рецессивная - (b), аллельная - (c)

Фенотип	Генотип	
	0	1
0	$n_{00} + n_{01}$	n_{02}
1	$n_{10} + n_{11}$	n_{12}
...	...	...
k	$n_{k0} + n_{k1}$	n_{k2}

(a)

Фенотип	Генотип	
	0	1
0	n_{00}	$n_{01} + n_{02}$
1	n_{10}	$n_{11} + n_{12}$
...	...	...
k	n_{k0}	$n_{k1} + n_{k2}$

(b)

Фенотип	Генотип	
	0	1
0	$2n_{00} + n_{01}$	$n_{01} + 2n_{02}$
1	$2n_{10} + n_{11}$	$n_{11} + 2n_{12}$
...	...	...
k	$2n_{k0} + n_{k1}$	$n_{k1} + 2n_{k2}$

(c)

2) Для каждой таблицы сопряженности определяется, какой тест запускать: Хи-квадрат или точный тест Фишера. Решение принимается исходя из наличия в таблице ожидаемых значений меньше 5 и из размеров таблицы. Для таблиц размером 2×2, которые содержат хоть одно ожидаемое значение (10), не превосходящее 5, запускается точный тест Фишера, в противном случае высчитывается χ^2 - статистика.
    а) В первом случае, то есть если запускается точный тест Фишера, то в результате вычислений сразу получаем p - значение.
    б) Если же в таблице сопряженности не имеются ожидаемые значения меньше 5, то в результате запуска теста χ^2 вычисляются статистика критерия χ^2 и число степеней свободы, затем по полученным значениям высчитывается p - значение.
3) Наконец, считаем для данных векторов фенотипов и генотипов коэффициент корреляции.

2.2.1 Вычисление статистики χ^2

Перед вычисление статистики проводится проверка применимости критерия χ^2 к рассматриваемой таблице сопряженности.

Находим строки и столбцы таблицы, полностью заполненные нулями. Если таких строк (или столбцов) такое количество, что в таблице остается только одна ненулевая строка (один ненулевой столбец), то вычисление корректного значения статистики χ^2 невозможно, и функция возвращает значение inf для данного столбца матрицы генотипов. Если же ненулевых столбцов и строк больше превышает 1, то выполняется вычисление статистики χ^2 по формуле (9). Вдобавок, вычисляется число степеней свободы = (количество_строк_таблицы-1)·(количество_столбцов_таблицы-1).

2.2.2 Точный тест Фишера

Вычисление p - значения с помощью точного теста Фишера производится в соответствии с формулами, описанными в параграфе 1.5 главы 1. Однако алгоритм был реализован с некоторыми особенностями:

1) в формуле (11) требуется считать факториалы чисел, однако, в программе вычисляются не именно факториалы, а их логарифмы. Как известно, диапазон чисел типа long double зависит от платформы и компилятора, поэтому максимальное число этого типа может быть равно как $1.18973 \cdot 10^{4932}$, так и $1.79769 \cdot 10^{308}$ (что соответствует максимально возможному числу типа double). То есть потенциально максимальное число, факториал которого мы можем вычислить, равно 170, что довольно мало. Таким образом, модифицированная формула (11) имеет вид:

$$\log(P_i) = \log(m_1)! + \log(m_2)! + \log(k_1)! + \log(k_2)! - \\ - \log(n)! - \log(n_{11}^i)! - \log(n_{12}^i)! - \log(n_{21}^i)! - \log(n_{22}^i)!$$

Тогда итоговая вероятность (12) вычисляется по формуле:

$$\log(P) = \log\left[\sum_{T_i:P_i\leqslant P_0, i\neq 0} P_i\right] = \log\left[\frac{P_0}{P_0}\sum_{T_i:P_i\leqslant P_0, i\neq 0} P_i\right] =$$
$$= \log(P_0) + \log\left[\sum_{T_i:P_i\leqslant P_0, i\neq 0} \frac{P_i}{P_0}\right] =$$
$$= \log(P_0) + \log\left[\sum_{T_i:P_i\leqslant P_0, i\neq 0} \exp\left(\log(P_i) - logP_0\right)\right]$$

2) большинство вычислений составляют повторяющиеся вычисления факториалов, чтобы каждый раз не считать факториал одного и того же числа, создадим «таблицу логарифмов факториалов», то есть одномерный массив размером n, где элемент массива является логарифмом факториала соответствующего индекса. Так как обращение к элементу массива занимает меньше времени, тем самым мы повышаем эффективность программы.

2.2.3 p - значение

Вычисление p - значения производилось так, как показано в формуле (4). В программе использовалась реализация Γ и γ функций из набора библиотек Boost. По полученным в результате проведения теста χ^2 значению χ^2 и числу степеней свободы вычисляется p значение с помощью формулы (4). Вычисление полной и нижней неполной гамма-функций производится с помощью функций из библиотеки Math набора библиотек Boost.

2.2.4 Коэффициент корреляции

Вычисление коэффициента корреляции производится для группированных данных, таблица сопряженности которых имеет вид:

Таблица 7 — Таблица сопряженности

ξ_1 \ ξ_2	0	1	2	$\sum \nu_{i\bullet}$
0	n_{00}	n_{01}	n_{02}	n_1
1	n_{10}	n_{11}	n_{12}	n_2
...	...	...	...	...
k	n_{k0}	n_{k1}	n_{k2}	n_k
$\sum$	m_1	m_2	m_3	n

происходит по модифицированной формуле (13):

$$\rho(X,Y) = \frac{\frac{1}{n}\sum\limits_{i,j}(n_{ij}\cdot i\cdot j) - \overline{X}^*\cdot\overline{Y}^*}{\sqrt{\frac{1}{n}\sum\limits_i n_i(i-\overline{X}^*)^2\,\frac{1}{n}\sum\limits_j m_j(j-\overline{Y}^*)^2}}$$

где

$$\overline{X}^* = \frac{1}{n}\sum_i (n_i\cdot i)\,,$$
$$\overline{Y}^* = \frac{1}{n}\sum_j (m_j\cdot j)\,.$$

§2.3 Проведения вычислительного эксперимента и его итоги. Результаты работы программы.

Были проведены подсчеты времени работы программ для разных размеров данных и, так как для вычисления использовались библиотека OpenMP и пакет doParallel, то и для разного количества ядер.

Наборы тестовых данных делятся на 2 категории:

1) набор данных с двумя уровнями фенотипов - для проверки эффективности точного теста Фишера, так как только в этом случае применяется данный тест.
2) набор с тремя уровня фенотипов - для проверки реализации алгоритма критерия χ^2.

1) **Таблица размера** 190 × 6810380 Набор данных содержит два уровня фенотипов. Вычисления проводились на сервере: Supermicro - Intel(R) Xeon(R) CPU E5-2690 0 2.90GHz, 32 cores, 377.82 GB, Linux 3.13.0-36-generic on x86_64, дистрибутив Ubuntu. Время работы:

Таблица 8 — Скорость вычислений для таблицы 190 x 6 810 380

Язык реализации программы	Время выполнения, с	Количество потоков
R	28477.447	10
C++	51.41762	10

2) **Таблица размера** 190 × 50000 Набор данных с двумя уровнями фенотипов.

Таблица 9 — Скорость вычислений для таблицы 190 x 50 000

Количество	Время выполнения, с	
потоков	R	C++
7	1.465	0.372
5	1.698	0.432
3	2.114	0.487
1	4.633	1.069

	stat_cd	deg_freedom_cd	p_value_cd	corr_cd	typetest_cd	elements_less_than5_cd
1	4.917714453	2	0.085532640	-0.140447484	Chi2	TRUE
2	4.917714453	2	0.085532640	-0.140447484	Chi2	TRUE
3	1.364342105	2	0.505518293	0.059873031	Chi2	TRUE
4	0.359447316	2	0.835501064	0.042657636	Chi2	TRUE
5	1.395155109	2	0.497789713	-0.061866374	Chi2	TRUE
6	2.348043523	1	0.125440354	0.111167099	Chi2	TRUE
7	1.815221182	2	0.403487169	-0.094067164	Chi2	TRUE
8	5.439467939	2	0.065892281	0.113597126	Chi2	FALSE
9	1.093668237	1	0.295659873	0.075869286	Chi2	TRUE
10	0.769820953	2	0.680511555	0.058971161	Chi2	TRUE
11	0.924055368	1	0.336412413	-0.069738435	Chi2	TRUE
12	3.314337055	2	0.190678116	-0.131457179	Chi2	FALSE
13	0.646215685	1	0.421468929	-0.058319252	Chi2	TRUE
14	1.636004623	2	0.441312379	-0.079675089	Chi2	TRUE
15	2.541093672	2	0.280678095	-0.110500140	Chi2	FALSE
16	3.393164986	1	0.065467170	-0.133636683	Chi2	TRUE

Showing 1 to 17 of 50,000 entries

Рис. 6 — Часть результатов вычислений C++ программы с двумя уровнями фенотипов

	cd.X-squared	df_cd.df	p_values_cd	corr_cd
result.1	4.91771445313838	2	0.0855326395666859	-0.140447484140268
result.2	4.91771445313838	2	0.0855326395666859	-0.140447484140268
result.3	1.36434210508285	2	0.505518293354983	0.0598730309946607
result.4	0.359447315969055	2	0.835501063560563	0.0426576364889276
result.5	1.39515510944082	2	0.497789712734892	-0.0618663743231549
result.6	2.34804352274232	1	0.125440354142228	0.111167098567458
result.7	1.81522118220151	2	0.403487168973165	-0.0940671638355141
result.8	5.43946793946794	2	0.0658922814361018	0.113597125940554
result.9	1.09366823652538	1	0.295659873094735	0.0758692863633992
result.10	0.769820952658268	2	0.680511555370197	0.0589711614597229
result.11	0.924055368499813	1	0.336412413057923	-0.0697384349400942
result.12	3.31433705455232	2	0.190678116352012	-0.13145717897273
result.13	0.646215684975375	1	0.421468929100548	-0.0583192522592747
result.14	1.63600462348226	2	0.441312379064699	-0.0796750885973695
result.15	2.54109367150328	2	0.280678094988195	-0.110500140459684
result.16	3.39316498552167	1	0.0654671702188609	-0.133636683143862

Showing 1 to 17 of 50,000 entries

Рис. 7 — Часть результатов вычислений стандартных функций R с двумя уровнями фенотипов

3) **Таблица размера** 286×50000 Набор данных содержит 3 уровня фенотипов.

Таблица 10 — Скорость вычислений для таблицы 286 x 50 000

Количество	Время выполнения, с	
потоков	R	C++
7	1.804	0.316
5	1.961	0.360
3	2.411	0.445
1	5.175	0.696

Глава 3. Заключение

Выводы

В данной работе была рассмотрена задача проведения множества статистических тестов для изучения зависимости фенотипа с каждым из имеющихся генетических маркеров. Были рассмотрены наиболее распространенные методы статистического анализа, в частности, критерий χ^2 и точный тест Фишера для проверки гипотезы о независимости двух признаков, а также исследование характера зависимости категориальных данных с помощью вычисления коэффициента корреляции.

Был реализован программный продукт в виде R-пакета, вычисления в котором выполняются в отдельном модуле, написанном на языке программирования C++ как на языке, допускающем эффективные вычисления.

Так же была реализована программа на языке R, в которой статистический анализ данных выполнялся с помощью стандартных функций R.

Было проведено сравнение скорости вычисления и итоговых результатов двух программ. Численные результаты программ совпадают. Однако вычисления, реализованные на языке C++, оказались эффективнее реализации на R. Главная причина такого результата – особенности программной среды R, выполнение вычислений в которой особенно неэффективно при многократном выполнении однотипной операции (цикла).

Перспективы развития

Полученные результаты показывают, что итоговый программный продукт справляется с поставленной целью. Однако имеются некоторые ограничения в области применения написанных мною алгоритмов:

1) программная реализация точного критерия Фишера применяется только для таблиц сопряженности размера 2×2. Для таблиц,у которых количество столбцов и строк больше двух, существует алгоритм точного критерия Фишера с применением методов Монте-Карло.
2) возможность анализа зависимости фенотипа с простейшими генетическими маркерам, представляющими собой категориальные величины с тремя уровнями. В дальнейшем предполагается рассмотреть возможности использования генетических маркеров с большим числом уровней.

Список литературы

[1] S. Nicklen F. Sanger and A.R. Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 1977.

[2] Smith L.M. Sanders J.Z. Kaiser R.J. Hughes P. Dodd C. Connell C.R. Heiner C. Kent S.B. Hood L.E. Fluorescence detection in automated dna sequence analysis. *Nature*, 1986.

[3] Fleischmann R.D. Adams M.D. White O. Clayton R.A. Kirkness E.F. Kerlavage A.R. Bult C.J. Tomb J.F. Dougherty B.A. Merrick J.M. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 1995.

[4] Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science*, 1998.

[5] Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 2000.

[6] An overview of the human genome project. https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/, May 2016.

[7] Jason H. Moore William S. Bush. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 2012.

[8] Гмурман В.Е. *Теория вероятности и математическая статистика.* Издательство «Высшая школа», 1972.

[9] Прохоров Ю.В. *Вероятность и математическая статистика. Энциклопедия.* Научное издательство «Большая Российская энциклопедия», 1999.

[10] Медведев Ю.И. Ивченко Г.И. *Математическая статистика.* Издательство «Высшая школа», 1984.

[11] С. Гланц. *Медико-биологическая статистика.* Издательский дом «Практика», 1998.

[12] Медведев Ю.И. Ивченко Г.И. *Введение в математическую статистику.* Издательство ЛКИ, 2010.

[13] Чернова Н.И. *Математическая статистика: Учеб. пособие.* Новосиб. гос. ун-т, 2007.

[14] Чернова Н.И. *Теория вероятностей: Учеб. пособие.* Новосиб. гос. ун-т, 2007.

[15] Dirk Eddelbuettel. *Seamless R and C++ Integration with Rcpp (Use R!).* Springer, 2013.

[16] Guide into openmp: Easy multithreading programming for c++. http://bisqwit.iki.fi/story/howto/openmp/, June 2016.

[17] *Creating R Packages: A Tutorial*, 2009.

[18] Steve Weston and Rich Calaway. Getting started with doparallel and foreach. https://cran.r-project.org/web/packages/doParallel/vignettes/gettingstartedParallel.pdf, October 2015.

[19] Xiuwen Zheng, Stephanie Gogarten, and Jean loup Gailly. Package 'gdsfmt'. http://bioconductor.org/packages/release/bioc/manuals/gdsfmt/man/gdsfmt.pdf, May 2017.

[20] R data import/export. https://cran.r-project.org/doc/manuals/r-release/R-data.pdf, April 2017.

[21] Snp definition. https://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295.

[22] Лисовенко Л.А. Арефьев В.А. *Англо-русский толковый словарь генетических терминов.* Изд-во ВНИРО, 1995.

Приложение А.
Словарь терминов

Определение 8. *Однонуклеотидный полиморфизм (Single nucleotide polymorphism, SNP)*[21] – отличия последовательности ДНК, размер которых составляет один нуклеотид, в геноме представителей одной популяции.

Определение 9. *Генетический маркер*[22] – ген, детерминирующий отчетливо выраженный фенотипический признак, используемый для генетического картирования и индивидуальной идентификации организмов или клеток.

Приложение В.
Ссылки на исходный код

1) C++ реализация статистических тестов (Rcpp-package):https://github.com/masterSplinterIO/chi2fishercpp
2) R-реализация (R-package): https://github.com/masterSplinterIO/chi2fisher