МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. Ломоносова
МЕХАНИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ
КАФЕДРА ТЕОРИИ ВЕРОЯТНОСТЕЙ

КУРСОВАЯ РАБОТА

«Оценка отклонений распределения данных от нормального закона. Ряды Эджворта »

Студентки 5-го курса 508-ой группы
кафедры теории вероятностей
Дьячковой Екатерины

Научный руководитель:
доктор физико-математических наук, профессор
Яровая Елена Борисовна

МОСКВА 2021

Содержание

Введение

Предположение о нормальности распределения изучаемых данных лежит в основе многих статистических моделей, в рамках которых проверяются различные статистические гипотезы. Такое предположение может влиять на статистические выводы. В частности, долгое время предположение о нормальности распределения использовалось при анализе данных о росте взрослого человека. В разделе 1 мы проведем обзор последних результатов из области генетики количественных признаков. Будут рассматриваться различные модели наследования роста взрослого человека, и будет показано, что предположение о распределении роста по нормальному закону является достаточно грубым, и распределение роста точнее описывается логарифмически нормальным распределением.

При измерении непрерывных клинико-демографических характеристик на практике исследователи зачастую сталкиваются с ошибками измерений, округляют данные или вносят их в компьютер с определенной точностью. Это может приводить к ошибочному отклонению статистической гипотезы о нормальности распределения. Поэтому раздел 2 будет посвящен численному исследованию того, как применение критерия Шапиро — Уилка для проверки гипотезы о нормальном распределении к данным, имеющим исходно нормальное распределение, связано с точностью их округления и размером выборок.

Наконец, в разделе 3 на примере логарифмически нормального распределения при помощи численных методов мы исследуем скорость сходимости рядов Эджворта к реальной функции распределения.

1 Ограничения нормальной аппроксимации

1.1 Введение

В этом разделе мы проведем разбор статьи [1] С.А. Славского и дополнительных материалов к ней.

Человеческий рост является важным примером количественного биологического признака. Рост, в отличие, например, от давления, измерять легче, ошибки измерений менее грубые. А знание закономерностей наследования роста может помочь нам в построении моделей наследования других количественных биологических признаков.

С самого зарождения генетики человеческий рост описывался как признак, который является суммой индивидуального вклада различных факторов [2]. То есть предполагалась аддитивность роста. А аддитивность, в свою очередь, влечет нормальность. Поэтому рост взрослого человека в учебниках по статистике обычно служит эмпирическим примером нормально распределенного биологического признака [3] – –[5]. Наследование роста исторически описывалось так называемой аддитивной полигенной моделью, где суммируются многие генетические эффекты [6] [7]:

$$height = \mu + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon \tag{1}$$

где μ – математическое ожидание, b_i – эффект i-го фактора, X_i – значение фактора, а ε – вектор остатков, считается нормально распределенным.

Аддитивная модель активно используется до сих пор. В таблице 1 приведена часть списка статей, где использовалась аддитивная модель наследования роста

Таблица 1: Список журналов, в которых использовалась аддитивная модель наследования роста взрослого человека.

Title	Year	Journal
Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index	2015	Nature Genetics
Hundreds of variants clustered in genomic loci and biological pathways affect human height	201	Nature
Defining the role of common variation in the genomic and biological architecture of adult human height	2014	Nature Genetics
Genome Partitioning of Genetic Variation for Height from 11,214 Sibling Pairs	2007	AJHG
Population genetic differentiation of height and body mass index across Europe	2015	Nature Genetics
Genetic linkage of human height is confirmed to 9q22 and Xq24	2006	Human Genetics
Common variants in the GDF5-UQCC region are associated with variation in human height	2008	Nature Genetics
Genome-wide genetic homogeneity between sexes and populations for human height and body mass index	2015	Hum Mol Gen
A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits	2009	Nature Genetics
Genetic variation in GPR133 is associated with height: genome wide association study in the self-contained population of Sorbs	2009	Hum Mol Gen
Genome-wide association analysis identifies 20 loci that influence adult height	2008	Nature Genetics
A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling pathway in the etiology of human height variation	2009	Hum Mol Gen
Genetic influences on the difference in variability of height, weight and body mass index between Caucasian and East Asian adolescent twins	2008	International Journal of Obesity
...	...	...

взрослого человека и, соответственно, предполагалось нормальное распределение роста в пределе:

Рассмотрим коэффициент вариации CV_ξ случайной величины ξ:

$$CV_\xi = \frac{\sigma}{\mu} \tag{2}$$

где $\sigma = \sqrt{D\xi}$ – стандартное отклонение ξ, $\mu = E\xi$ – математическое ожидание ξ.

Если предположить, что аддитивная модель верна, то мы должны наблюдать как минимум два факта. Первый из них заключается в том, что в нашем предположении коэффициент вариации роста взрослого человека должен быть низким. Пусть $\xi_1, \ldots, \xi_n$ – н.о.р., случайные величины с $E\xi_1 = \mu$ и $\sqrt{D\xi_1} = \sigma$. Тогда:

$$CV_{\sum_{i=1}^n \xi_i} = \frac{\sqrt{D\left(\xi_1 + \ldots + \xi_n\right)}}{E\left(\xi_1 + \ldots + \xi_n\right)} = \frac{\sqrt{D\xi_1 + \ldots + D\xi_n}}{E\xi_1 + \ldots + E\xi_n} = \frac{\sqrt{n}\sigma}{n\mu} = \frac{1}{\sqrt{n}}\frac{\sigma}{\mu}$$

$$\Rightarrow \left(CV_{\sum_{i=1}^n \xi_i}\right)^2 \sim \frac{1}{n} \tag{3}$$

Выражение 3 верно и в более общем случае, для строго положительных, не идеально коррелированных случайных величин [8]. Второй факт заключается в том, что в пределе мы должны наблюдать нормальное распределение роста в популяции. Однако нормальное распределение – это лишь приближение к реальному распределению человеческого роста. Более того, когда мы говорим об аддитивной полигенной модели как о модели наследования роста, мы должны учитывать следующее. Если мы не

рассматриваем такие серьезные мутации, как, например, карликовость, то про оставшиеся мутации известно, что они примерно одинаково распределены. Помимо генетических факторов, на наследование роста также влияют социально-демографические характеристики, такие как, например, пол и достаток. Про них известно, что они по размеру своего эффекта совершенно не сопоставимы с генетикой. Поэтому центральная предельна теорема для такой модели не выполняется напрямую. В действительности мы скорее будем наблюдать смесь нормальных распределений роста и нормальность остатков от регрессии на социально-демографические параметры:

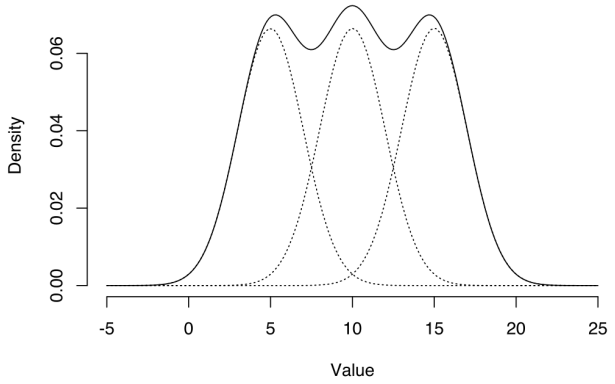

Рис. 1: График плотности смеси нормальных распределений.

Авторы статьи [1] проанализировали большое количество как классических, так и достаточно свежих исследований, связанных с моделями наследования роста. В аддитивной модели и аппроксимации нормальным распределением были выделены следующие проблемы.
Первая: при анализе больших выборок стандартное отклонение роста, как правило, было больше в более высоких популяциях, а коэффициент вариации между популяциями был низкий и довольно стабильный. На рисунке 2 из [9] показано, как дисперсия роста женщин увеличивается при увеличении среднего роста женщин. А на рисунке 3 из [10] построено распределение значений коэффициентов вариации роста людей и длины тела животных и показано, насколько коэффициент вариации роста человека низкий. Такие признаки не характерны для нормального распределения, его параметры – среднее и стандартное отклонение – независимы. Для логарифмически нормального распределения, наоборот, известно, что его коэффициент вариации стабильный. В некоторых исследованиях, например [11], напрямую постулировалось логнормальное распределение роста человека.
Второе замечание в процессе анализа исследований касается того, как в разных исследованиях учитывали эффекты пола. Поскольку рост женщин и мужчин отличается, мужчины в среднем выше чем женщины, этот факт необходимо отображать в модели наследования роста. Если мы рассматриваем аддитивную модель, то мы должны прибавлять к росту женщин некоторую абсолютную величину. Но как в классических, так и в современных исследованиях роста поступают по-разному. Зачастую выборки стратифицируют по полу. В части исследований в таблице 1 использовалась стратификация, в части – прибавление абсолютной величины. Но, например, Гальтон [2] в своей работе поступал иначе, он делал мультипликативную поправку на рост, умножая рост женщин на 1.08. Отметим, что если мы говорим об аддитив-

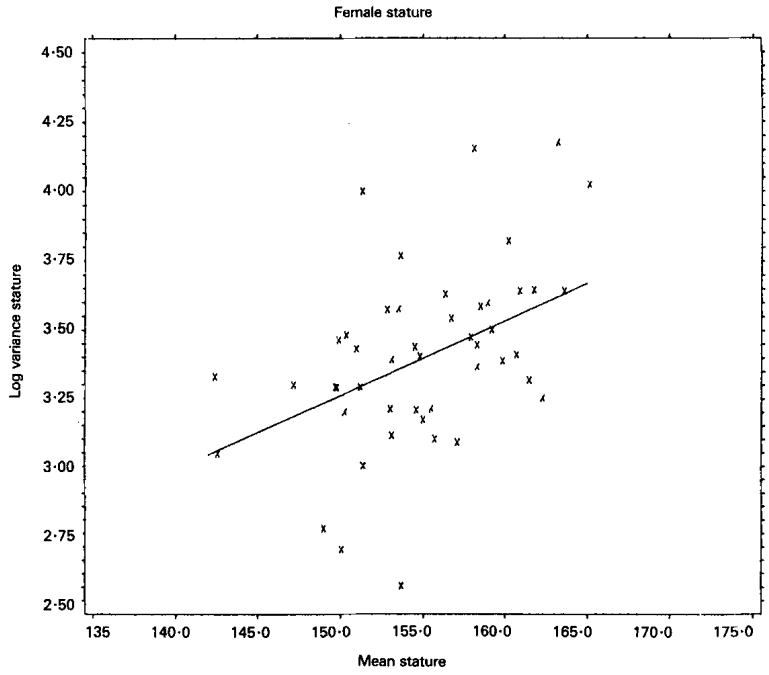

Рис. 2: График зависимости дисперсии роста женщин от среднего роста женщин [9].

ной модели и нормально аппроксимации модели наследования роста человека, то такие поправки делать запрещено. В работе [12] сравнивали аддитивную и мультипликативную поправки на рост. При рассмотрении аддитивной модели наследования роста мы должны предполагать, что у роста мужчин и женщин равные дисперсии. В то время как при рассмотрении мультипликативной модели наследования средний рост женщин должен умножаться на некоторую константу, а дисперсия роста женщин, соответственно, на квадрат этой константы. Поэтому в работе [12] тестировала гипотеза о равенстве дисперсий роста мужчин и женщин. Эта гипотеза отверглась и в работе было сказано, что мультипликативная поправка на рост статистически лучше, чем аддитивная. Однако важно отметить, что тестирование гипотезы проводилось при помощи F-теста, который не является устойчивым к незначительным отклонениям распределения от нормального закона. Также в рассматриваемой нами статье [1] говорится о том, что на достаточно больших выборках можно заметить, что стандартное отклонение роста мужчин больше, чем у женщин, и отношение между стандартным отклонением роста мужчин и женщин близко к отношению между средними роста мужчин и женщин.

Таким образом, эти два факта плохо согласуются с аддитивной моделью и аппроксимацией нормальным распределением модели наследования роста взрослого человека.

Прежде чем отклонять аддитивную модель, предоставим возможные объяснения двум фактам, описанным выше.

Первое возможное объяснение касается эволюционной биологии. Например, в [13] постулируется более узкая норма реакции у самцов и, следовательно, для признака при стабилизирующем отборе в беспородной популяции большая общая дисперсия. Этим можно было бы объяснить большее стандартное отклонение роста мужчин, по сравнению со стандартным отклонением роста женщин. Другим широким объясне-

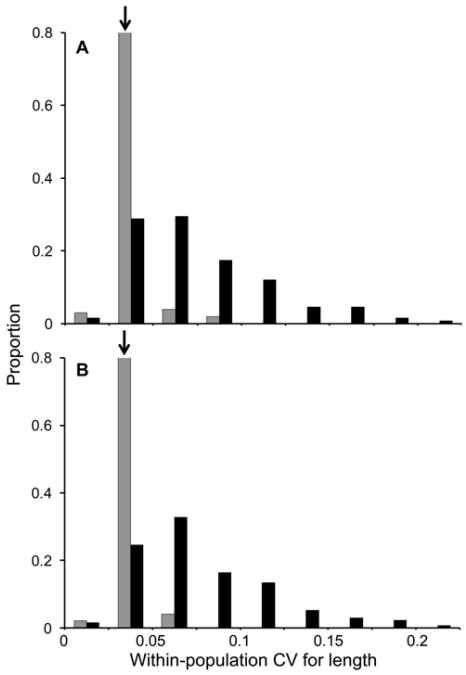

Рис. 3: Распределения коэффициентов вариации (CV) длины тела животных и роста человека внутри популяции. Показаны видовые значения для животных (черный) и популяционные значения для людей (серый) для самцов (A) и самок (B). Стрелки указывают расположение CV для среднего роста человека [10].

нием может быть то, что воздействие окружающей среды на рост распределяется по-разному для мужчин и женщин, что приводит к различию в величине различий, обусловленных окружающей средой. В этом случае близость соотношения между стандартным отклонением роста мужчин и женщин к соотношению между средним ростом мужчин и женщин может быть совпадением.

Оба эти объяснения, однако, также предсказывали бы различия в наследуемости роста между мужчинами и женщинами, а эта разница, если таковая имеется, очень мала [14], [15].

Другое объяснение состоит в том, что может быть верна мультипликативная модель наследования роста, которая привела бы к логарифмически нормальному приближению распределения роста. Но, если это так, то почему логарифмически нормальное распределение роста не было обнаружено ранее? И рост так хорошо описывался нормальным распределением [3] – [5]?

Помимо масштабирования стандартного отклонения вместе со средним значением у логарифмически нормального распределения, когда логарифмически нормально распределенный признак рассматривается в аддитивном приближении, могут включаться мультипликативные взаимодействия генов с окружающей средой и, так называемые, эпистатические взаимодействия генов [16] для улучшения соответствия модели данным. То есть, если рост на самом деле распределен логнормально, а мы будем использовать аддитивную модель, то нам придется включать помимо простой суммы эффектов, произведения генетических факторов с социо-демографическими, а также неаддитивные генетические взаимодействия в нашу модель. Ранее значимость таких эффектов в аддитивной модели не детектировалась [17].

Наконец, третье объяснение. Можно было бы предложить гибридную гипотезу, предполагающую, что некоторые эффекты в модели наследования роста умножаются, а другие эффекты суммируются.

Стоит также отметить, что доказательства, говорящие в пользу мультипликативной модели и логнормальной аппроксимации роста, в основном получены в результате сравнения различий в росте между различными популяциями. А аддитивная модель, как правило, используется в генетических исследованиях межиндивидуальных различий в однородных популяциях.

До совсем недавнего момента не было доступных данных, которые обеспечивали бы как уровень детализации, необходимый для генетического исследования, так и разнообразие выборки. Этот пробел был недавно ликвидирован проектом биобанка Великобритании [18].

1.2 Исследование моделей наследования роста

У нас есть две основные идеи:
1. Аддитивная модель наследования роста:

$$height = \mu + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n + \varepsilon \tag{4}$$

2. Мультипликативная модель наследования роста:

$$height = \mu \cdot b_1^{X_1} \cdot b_2^{X_2} \cdot \ldots \cdot b_n^{X_n} \cdot \varepsilon \tag{5}$$

или что то же самое:

$$log_{10}(height) = \mu' + b'_1 X_1 + b'_2 X_2 + \ldots + b'_n X_n + \varepsilon' \tag{6}$$

где μ и μ' – математическое ожидание, b_i и b'_i – эффект i-го фактора, X_i и X'_i – значение фактора, а ε и ε' – вектор остатков, считается нормально распределенным.

На первом этапе исследования, проделанного в рассматриваемой нами статье [1], был проведен регрессионный анализ между стандартным отклонением (SD), коэффициентом вариации (CV) и средним ростом женщин из 54 развивающихся стран по данным из [19]. Из рассмотрения были исключены страны, для которых размер выборки составлял менее 1000 человек (Коморские Острова) и для которых наблюдения отклонялись более чем на 3 стандартных отклонения от общего среднего значения (Демократическая Республика Конго, Республика Конго, Гватемала). Полученные 50 популяций были взяты для дальнейшего анализа. Чтобы избежать доминирования нескольких очень больших выборок, был использован равный вес для наблюдений, поступающих из разных популяций.

Далее была проанализирована зависимость между средним ростом мужчин и женщин в мировых популяциях при использовании данных интернет-ресурса [20]. Первый этап фильтрации включал удаление строк с отсутствующими значениями для мужчин или женщин, а затем удаление строк со значениями, отклоняющимися от среднего значения более чем на 3 стандартных отклонения для соответствующего пола. На втором этапе фильтрации были исключены повторяющиеся данные по одним и тем же странам и сохранен один результат опроса для каждой страны и/или национальной группы. Критерии фильтрации были следующими: если были доступны данные о городском/сельском населении и населении в целом, сохранялись общие

данные; если были доступны разные возрастные интервалы, сохранялся более широкий; если были доступны данные для нескольких возрастов, сохранялся тот, который ближе к 21 году. Этнические группы в одной стране рассматривались как отдельные группы населения. В конце концов, 80 групп населения прошли все фильтры. На

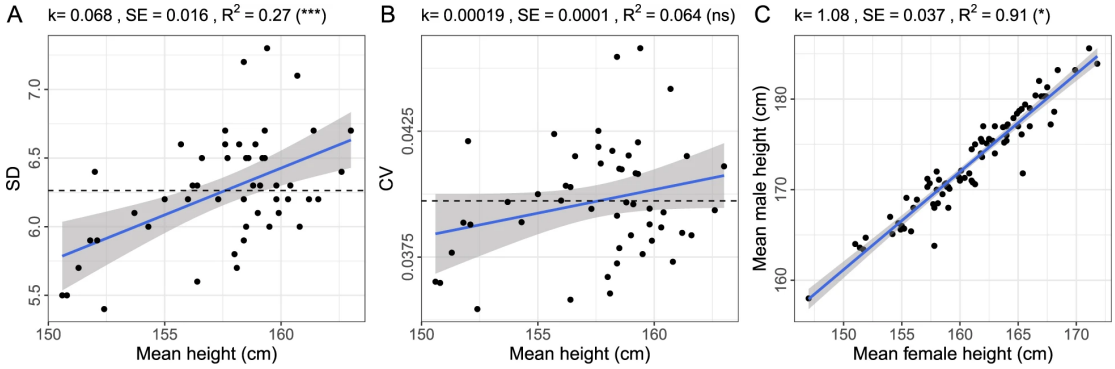

Рис. 4: **Связь между параметрами распределения роста взрослого человека по популяциям.** Линейная регрессия SD (A) и CV (B) роста на средний рост женщин из [19]. Пунктирная линия показывает общее среднее значение. (C) Линейная регрессия среднего роста мужчин на средний рост женщин в популяциях из [20]. Невзвешенная линейная регрессия использовалась для оценки тренда (k), его стандартной ошибки (SE), скорректированного R^2 и, в скобках, значимости отклонения коэффициента регрессии от нуля для (A), (B) и от единицы для (C) ($p < 0.001$ – ***; $p < 0.01$ – *; $p > 0.05$ – ns).

графике (A) на рисунке 4 показано, что стандартное отклонение (SD) роста женщин увеличивается с увеличением среднего роста женщин. На графике (B) коэффициент вариации (CV) низкий и не значимо увеличивается, то есть он стабильный. На графике (C) показано, что рост мужчин и женщин связан мультипликативно, поскольку наклон кривой в этой модели был $k = 1.08$, что значимо отличается от 1 ($p = 0.003$).

Таким образом, в этой части исследования подтвердились результаты предыдущих работ про распределение и модели наследования роста взрослого человека.

1.3 Анализ данных биобанка Великобритании

На втором этапе исследования были проанализированы 369153 участников британского биобанка европейского происхождения, принадлежащих к шести группам, определенным по этническому происхождению и месту рождения (см. Дополнительную таблицу 2 в [1]). Были рассмотрены влияния пола, генотипа и остаточных эффектов. Генотип был включен в анализ в виде полигенного показателя роста (PGHS – polygenic height score), определяемой как взвешенная распространенность аллелей, увеличивающих рост, в генотипе. Факторы, связанные с социально-экономическим статусом и другими ковариатами исследования, были использованы для построения единого линейного предиктора, далее называемого остаточным предиктором (RP – residual predictor). Все три предиктора были тесно связаны с ростом взрослого человека. Результаты для каждой группы представлены в дополнительных таблицах в [1].

Для анализа масштабирования стандартного отклонения (SD) роста в зависимости от среднего роста каждая из шести групп анализа была разделена на восемь подгрупп, определенных по полу, высокому и низкому PGHS, высокому и низкому RP. В каждой из полученных 48 подгрупп были оценены влияние среднего роста на стандартное отклонение с помощью модели линейной регрессии с весами, определяемыми как размер группы. В каждой из шести групп анализа был рассчитан медианный полигенный показатель и медианный остаточный предиктор. Результаты представлены на рисунке 5: На графике A рисунка 5 для среднего роста стандартное

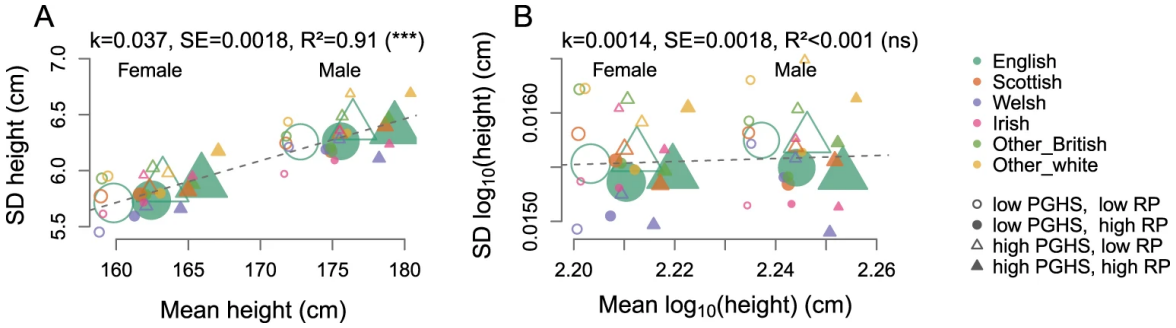

Рис. 5: Графики зависимости SD от среднего роста и от среднего $\log_{10}$ роста в британском Биобанке. Отношение SD к среднему росту (A) и логарифмическому росту (B) для шести групп британских лиц белого происхождения из британского Биобанка, определенных на основе места рождения и разделенных по полу, медианному полигенному показателю и медианному остаточному предиктору (всего 48 групп). Размер символа пропорционален весу регрессии, определяемому как удвоенный размер группы. Взвешенная линейная регрессия использовалась для оценки тренда (k), его стандартной ошибки (SE), скорректированного значения R^2 и, в скобках, значимости отклонения коэффициента регрессии от нуля ($p < 0.001$ – *** ; $p > 0.05$ – ns).

отклонение значимо увеличивается при увеличении значения среднего роста, стандартное отклонение для мужчин больше, чем для женщин. Для логарифма роста этого не происходит. Также важно отметить, что если ранее подобные результаты были получены для различных популяций, то в этом случае исследование проводилось на монопопуляции, внутри одной, английской, национальности.

Для создания рисунка 6 были оценены влияния определенных факторов (пол, полигенный показатель и остаточный предиктор) на средний рост и на средний логарифмический рост в шести этнических группах, дополнительно разделенных на два других фактора: медианный полигенный показатель и медианный остаточный предиктор. В общей сложности были рассмотрены 24 подгруппы по каждому фактору. В каждой подгруппе влияние фактора на средний рост (средний логарифмический рост) оценивалось с использованием взвешенной одномерной линейной регрессии. Веса были определены как размер группы. На графиках A, C, E рисунка 6 для среднего роста размеры эффектов всех трех факторов (пол, PGHS, RP) значимо увеличиваются при увеличении среднего роста. Для среднего логарифмического роста этого снова не происходит.

Оба рисунка 5 и 6 подтверждают предположения о некорректности аддитивной модели для наследования роста человека. При использовании аддитивной модели мы не должны наблюдать увеличения стандартного отклонения и размеров эффектов

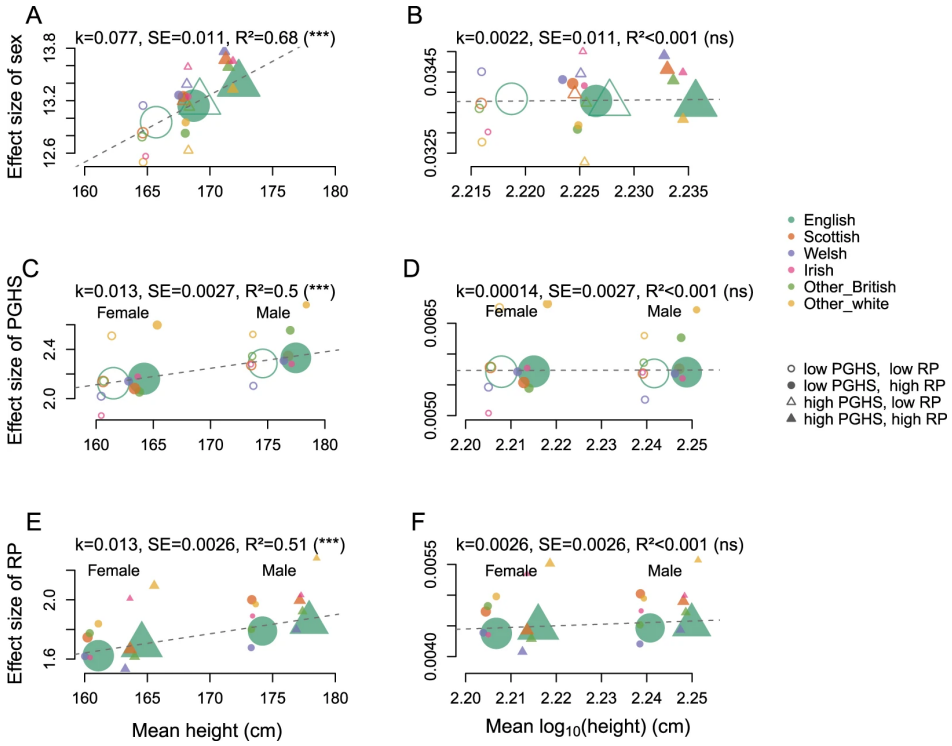

Рис. 6: Графики зависимости влияния различных факторов на средний рост и на средний $\log_{10}$ роста в британском Биобанке. Связь между оценкой величины эффекта пола (A, B), генотипа (C, D; генотип определялся как полигенный показатель роста, PGHS), других факторов (E, F; линейный остаточный предиктор, RP, объединяющий социально-демографические и исследовательские ковариаты) и среднего роста (A, C, E) и логарифмического роста (B, D, F) для 6 групп британских лиц белого происхождения из британского Биобанка, определенных на основе места рождения. Шесть групп дополнительно разделены по полу (C–F), медианному полигенному показателю (A, B, E, F) и медианному остаточному предиктору (A–D). Размер символа пропорционален размеру группы (используется в качестве веса регрессии). Взвешенная линейная регрессия использовалась для оценки тренда (k), его стандартной ошибки (SE), скорректированного R^2 и, в скобках, значимости отклонения коэффициента регрессии от нуля ($p < 0.001$ – ***; $p > 0.05$ – ns).

при увеличении зависимой переменной. И, наоборот, для среднего логарифмического роста таких противоречий с аддитивной моделью не наблюдалось.

1.4 Мультипликативные и эпистатические взаимодействия

Как было ранее отмечено в разделе 1.1, при рассмотрении логарифмически нормально распределенного признака в нормальном приближении, мы должны наблюдать в модели попарные (мультипликативные) и эпистатические взаимодействия генетических факторов.

Для анализа значимости попарных взаимодействий генетических факторов в статье [1] были рассмотрены модели линейной регрессии для наследования роста 7 и логарифма роста 8, где в качестве предикторов были не только пол (sex), полигенный показательно роста (PGHS) и остаточный предиктор (RP), но и произведения

Таблица 2: Результаты линейной регрессии для модели наследования роста, включающей мультипликативные взаимодействия (для полной выборки).

Предиктор	Оценка	Ст.ошибка	t-статистика	p-значение	R^2	N
Своб.коэф.	162.7	0.013	12480.859	$< 10^{-100}$		369153
Sex	13.157	0.019	684.473	$< 10^{-100}$		
PGHS	2.132	0.013	163.634	$< 10^{-100}$		
RP	1.668	0.013	128.012	$< 10^{-100}$		
Sex·PGHS	0.16	0.019	8.305	$< 10^{-16}$		
Sex·RP	0.142	0.019	7.398	$1.38 \cdot 10^{-13}$		
PGHS·RP	0.042	0.01	4.396	$1.1 \cdot 10^{-5}$	0.603	

Таблица 3: Результаты линейной регрессии для модели наследования логарифма роста, включающей мультипликативные взаимодействия (для полной выборки).

Предиктор	Оценка	Ст.ошибка	t-статистика	p-значение	R^2	N
Своб.коэф.	2.211	$3.35 \cdot 10^{-5}$	65911.831	$< 10^{-100}$		369153
Sex	0.034	$4.95 \cdot 10^{-5}$	682.693	$< 10^{-100}$		
PGHS	0.006	$3.35 \cdot 10^{-5}$	169.696	$< 10^{-100}$		
RP	0.004	$3.35 \cdot 10^{-5}$	132.887	$< 10^{-100}$		
Sex·PGHS	$-3.04 \cdot 10^{-5}$	$4.95 \cdot 10^{-5}$	-0.614	0.539		
Sex·RP	$2.04 \cdot 10^{-5}$	$4.95 \cdot 10^{-5}$	0.411	0.68		
PGHS·RP	$4.44 \cdot 10^{-5}$	$2.46 \cdot 10^{-5}$	1.803	0.071	0.602	

этих трех факторов – sex·PGHS, sex·RP, RP·PGHS. Уровень значимости фиксировали $\alpha = 0.05$. Результаты представлены в дополнительных таблицах [1]. Приводим часть этих результатов (для полной выборки) в таблице 2 для модели 7 и таблице 3 для модели 8 соответственно.

$$height = \mu + \beta_1 \cdot sex + \beta_2 \cdot PGHS + \beta_3 \cdot RP + \\ + \beta_4 \cdot \mathbf{sex} \cdot \mathbf{PGHS} + \beta_5 \cdot \mathbf{sex} \cdot \mathbf{RP} + \beta_6 \cdot \mathbf{PGHS} \cdot \mathbf{RP} + \varepsilon \quad (7)$$

$$log_{10}(height) = \mu + \beta_1 \cdot sex + \beta_2 \cdot PGHS + \beta_3 \cdot RP + \\ + \beta_4 \cdot \mathbf{sex} \cdot \mathbf{PGHS} + \beta_5 \cdot \mathbf{sex} \cdot \mathbf{RP} + \beta_6 \cdot \mathbf{PGHS} \cdot \mathbf{RP} + \varepsilon \quad (8)$$

Анализируя p-значения в таблицах, приходим к выводу, что все три мультипликативных фактора оказались значимы в модели наследования роста и не значимы в модели наследования логарифма роста. Таким образом, если в качестве модели

Таблица 4: Значения R^2 для аддитивной и мультипликативной моделей (в процентах).

Группа	R^2_{or}	$R^2_{log+exp}$	R^2_{log}	R^2_{or+log}	$R^2_{log+exp} - R^2_{or}$	$R^2_{log} - R^2_{or+log}$	**N**
Англичане	60.174	60.188	60.101	60.087	0.014	0.014	282509
Шотландцы	60.701	60.716	60.519	60.505	0.015	0.014	29133
Др.Британцы	59.910	59.946	59.779	59.744	0.036	0.036	18073
Валлийцы	61.761	61.764	61.669	61.663	0.003	0.006	16397
Др.европейцы	59.101	59.100	58.794	58.802	-0.001	-0.009	14593
Ирландцы	61.263	61.277	61.126	61.111	0.013	0.015	8448

наследования роста выбирается аддитивная модель, то необходимо учитывать мультипликативные взаимодействия. В случае выбора мультипликативной модели (т.е., аддитивной модели для логарифма роста) модель получается более простой, мультипликативные эффекты включать не требуется.

Под эпистазом, как было указано в статье [1], понимается любое отклонение от аддитивности генетических эффектов. Поэтому для анализа значимости эпистатических взаимодействий в [1] была рассмотрена модель 9 для роста и модель 10 для логарифма роста, где в качестве эпистатического взаимодействия рассматривается квадрат полигенного показателя PGHS·PGHS. p-значения для фактора PGHS·PGHS представлены в формулах 9 и 10 соответственно:

$$height = \mu + \beta_1 \cdot PGHS + \beta_2 \cdot PGHS^2 + \varepsilon \ (p_{PGHS^2} = 4 \cdot 10^{-7}) \tag{9}$$

$$log_{10}(height) = \mu + \beta_1 \cdot PGHS + \beta_2 \cdot PGHS^2 + \varepsilon \ (p_{PGHS^2} = 0.48) \tag{10}$$

Таким образом, для аддитивной модели такое эпистатическое взаимодействие, как PGHS·PGHS, оказалось значимо, а значит, его необходимо учитывать при использовании аддитивной модели и приближения роста нормальным распределением. Модель вновь оказывается более сложной. В случае мультипликативной модели наследования роста и приближении роста логарифмически нормальным распределением фактор PGHS·PGHS значимым не оказался.

1.5 Итоговое сравнение моделей

На последнем этапе исследования в статье [1] для аддитивной и мультипликативной моделей наследования роста была составлена таблица 4 со значениями R^2 – долей объясненной дисперсии (в процентах). Все результаты были порядка 60%, для мультипликативной модели R^2 был больше, чем для аддитивной модели менее, чем на 0.1%. Таким образом, итоговая разница аддитивной и мультипликативной моделей наследования роста небольшая, но мультипликативная модель объясняет дисперсию роста лучше, чем аддитивная. Обозначения в таблице 4: $R^2_{or}(\%)$ – доля дисперсии роста в исходном масштабе, объясняемая предсказанием линейной модели для роста; $R^2_{log+exp}(\%)$ – доля дисперсии роста в исходном масштабе, объясненная экспоненциальным предсказанием линейной модели для логарифма роста; $R^2_{log}(\%)$ – доля дисперсии логарифма роста, объясненная предсказанием линейной модели для логарифма роста; $R^2_{or+log}(\%)$ – доля дисперсии логарифма роста, объясненная логарифмическим предсказанием линейной модели для роста.

Как было указано в рассмотренной нами статье [1], проведенное исследование в основном имеет концептуальную ценность и может помочь лучше интерпретировать результаты анализа больших данных. При исследований больших выборок в рамках аддитивной модели появляются неоднородность дисперсии и разные взаимодействия (попарные и эпистатические), которые необходимо учитывать. Модель получается более сложной. Часто это объясняют с точки зрения эволюционной биологии. Но в статье показано, что если мы воспользуемся логарифмически нормальным приближением, то мы по прежнему сможем использовать простую аддитивную модель, в которой эффекты суммируются (в логарифмическом масштабе). Обе модели достаточно хорошо предсказывают рост: доля объясненной дисперсии R^2 порядка 60%, разница меньше 0.1%.

2 Влияние округления на результаты критериев нормальности

2.1 Введение

В этом разделе исследуется вопрос влияния округления данных на результаты критериев нормальности. Для моделирования использовался компьютерный кластер с 48 процессорными ядрами и статистическая среда R 3.6.3. В качестве модельного критерия нормальности был выбран критерий Шапиро — Уилка.

Приводим далее кратко алгоритм моделирования. К выборкам из нормального распределения разного размера и с разной дисперсией, но с одинаковым фиксированным нулевым средним, применялись разные функции округления. Далее при помощи метода Монте-Карло вычислялась оценка вероятности отвержения нулевой гипотезы о нормальности критерия Шапиро — Уилка. Заметим, что для неокругленной выборки такая оценка вероятности совпадает с оценкой вероятности ошибки первого рода критерия Шапиро — Уилка. Далее мы сравнивали оценки вероятностей для каждой округленной выборки с оценкой вероятности для неокругленной выборки и представили результаты графически.

2.2 Алгоритм моделирования

Пусть round(x, n) — функция округления вещественного числа x до n знаков после запятой по стандарту IEC 60559. Рассмотрим две сетки значений переменных. Первая сетка N_{mesh} генерируется следующим образом:

1. Берем арифметическую последовательность от $\log_{10}(5)$ до $\log_{10}(5000)$ длины 550. Обозначим эту последовательность $A = \{a_n\}_{n=1}^{550}$;

2. Из последовательности A получим последовательность $B = \{b_n = 10^{a_n}\}_{n=1}^{550}$;

3. Применяем к B функцию округления round(x, 0), получаем последовательность $C = \{c_n = round(b_n)\}_{n=1}^{550}$;

4. Выбираем из последовательности C только уникальные значения, получаем последовательность N_{mesh} длины $L = 405$.

Вторая сетка σ_{mesh} генерируется следующим образом:

1. Берем арифметическую последовательность от $\log_{10}(0.01)$ до $\log_{10}(100)$ той же длины, что и N_{mesh}, $L = 405$. Обозначим эту последовательность $A = \{a_n\}_{n=1}^{405}$;

2. Из последовательности A получим последовательность $\sigma_{mesh} = \{\sigma_n = 10^{a_n}\}_{n=1}^{405}$.

Для удобства обозначений будем использовать восемь функций $f_i(x), i = 0, \ldots, 7$, из которых $f_0(x)$ тождественная, а остальные семь функций соответствуют различным видам округления:

$$f_0(x) = x \tag{11}$$

$$f_1(x) = round\left(\frac{x}{10}\right) * 10 \tag{12}$$

$$f_2(x) = round\left(\frac{x}{5}\right) * 5 \tag{13}$$

$$f_3(x) = round\left(\frac{x}{2}\right) * 2 \tag{14}$$

$$f_4(x) = round(x) \tag{15}$$

$$f_5(x) = round(x, 1) \tag{16}$$

$$f_6(x) = round(x, 2) \tag{17}$$

$$f_7(x) = round(x, 3) \tag{18}$$

Далее для каждого фиксированного $\sigma_i \in \sigma_{mesh}$, для каждого фиксированного $N_j \in N_{mesh}$ и для каждой фиксированной функции $f_l(x), l = 0 \ldots 7$ из (11)-(18) мы действовали по следующему алгоритму, который является модификацией алгоритма метода Монте-Карло:

1. Пусть заданы гипотеза $H_0 : F(x) \in \mathcal{F} = \{\Phi_{\mu,\sigma^2}(x) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\}$ и альтернатива $H_1 : F(x) \notin \mathcal{F}$;

2. Фиксируем функцию распределения, которая удовлетворяет нулевой гипотезе H_0: $F(x) = \Phi_{0,\sigma_i^2}(x)$;

3. Выбираем достаточно большое натуральное число K и генерируем K раз выборку размера N_j (при условиях гипотезы H_0). Обозначим полученные выборки: $\mathcal{X}_k, k = 1 \ldots K$;

4. К каждой из $\mathcal{X}_k, k = 1 \ldots K$ применяем функцию $f_l(x)$, получаем выборки $f_l(\mathcal{X}_k), k = 1 \ldots K$;

5. Для каждой $f_l(\mathcal{X}_k)k = 1 \ldots K$ выясняем, отверглась ли гипотеза H_0 о нормальности критерием Шапиро — Уилка. Если да, то $m_k = 1$, иначе $m_k = 0$;

Таблица 5: Результаты моделирования.

σ_{mesh}	$f_0(x)$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_7(x)$
0.01	0.0471	1	1.0	1.0	1.0	1.0	0.3509	0.0503
0.0102	0.0460	1	1.0	1.0	1.0	1.0	0.3462	0.0486
0.0105	0.0495	1	1.0	1.0	1.0	1.0	0.3353	0.0512
...	...	...	...	...	...	...	...	...

6. Вычисляем количество отвержений гипотезы H_0 о нормальности критерия Шапиро — Уилка: $M = \sum_{k=1}^{K} m_k$;

7. Вычисляем оценку вероятности отвержения гипотезы H_0 о нормальности критерием Шапиро — Уилка: $W_{l,i,j} = \frac{M}{K}$.

Таким образом, для всех возможных размеров выборок из N_{mesh}, дисперсий из σ_{mesh} и функций (11)-(18) мы получили оценку вероятности отвержения гипотезы H_0 о нормальности критерием Шапиро — Уилка. Далее мы занесли результаты в таблицы. Каждая таблица соответствовала одному размеру выборки $N_j \in N_{mesh}$, строки соответствовали $\sigma_i \in \sigma_{mesh}$, столбцы – функциям из (11)-(18). Пример таблицы для $N_1 = 5$ представлен в таблице 5. В наших обозначениях каждое значение в таблице 5 соответствует одному из $W_{l,i,1}, l = 0, \ldots, 7, i = 1, \ldots, 405$.
Далее для каждой пары $\sigma_i \in \sigma_{mesh}$ и $N_j \in N_{mesh}$ мы вычислили модуль разности $\delta_{l,i,j} = |W_{l,i,j} - W_{0,i,j}|, l = 0, \ldots, 7$ и занесли их в соответствующие матрицы $\Delta_l, l = 0, \ldots, 7$ размера $L \times L = 405 \times 405$ следующим образом:

$$\Delta_l = \begin{pmatrix} \delta_{l,1,1} & \delta_{l,2,1} & \ldots & \delta_{l,L,1} \\ \delta_{l,1,2} & \delta_{l,2,2} & \ldots & \delta_{l,L,2} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{l,1,L} & \delta_{l,2,L} & \ldots & \delta_{l,L,L} \end{pmatrix}, l = 0, \ldots, 7 \tag{19}$$

Строки матрицы 19 соответствуют $N_j \in N_{mesh}$, столбцы соответствуют $\sigma_i \in \sigma_{mesh}$. Такой порядок необходим для удобства программирования.

Теперь мы можем представить полученные результаты в виде графиков хитмэп. Тепловая карта (heatmap, хитмэп) — это графическое представление данных, в котором отдельные значения, содержащиеся в матрице, представлены в виде цветов. Для каждой из восьми функций $f_l(x), l = 0 \ldots 7$ мы построили хитмэп, каждая точка которого соответствовала одному из значений $\delta_{l,i,j}$, а само значение отображалось цветом. Результаты представлены на рисунках 7 и 8.

2.3 Результаты моделирования

Белые зоны на графиках означают разницу между оценками вероятностей отклонения гипотезы о нормальности для округленной и неокругленной выборки меньше 5 %-ных пунктов. Ее можно считать несущественной. Красные зоны, наоборот, показывают, что эта разница может достигать значений близких к 1. То есть в красной зоне критерий Шапиро — Уилка всегда отвергает нулевую гипотезу о нормальности.

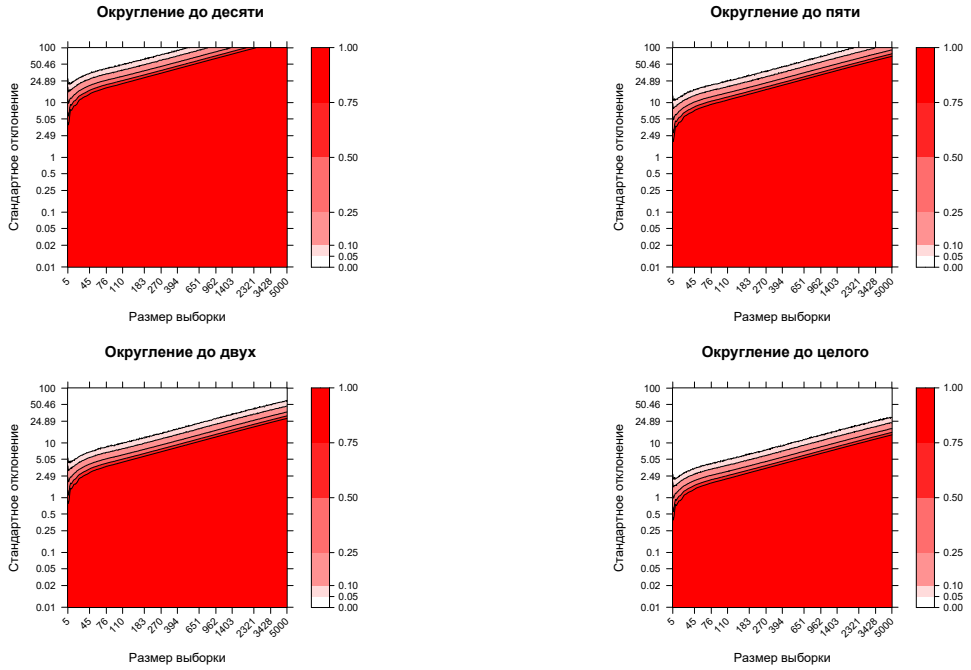

Рис. 7: Тепловые карты результатов моделирования для $f_1(x)$ – $f_4(x)$.

Сравнивая графики для разных видов округления, видно как получаются разные исходы. Заметим, что с увеличением объема выборки, размер красной области на графиках рисунков 7 и 8 увеличивается, по сравнению с белой областью. Это связано с тем, что при увеличении объема выборки, увеличивается мощность критерия Шапиро — Уилка и гипотеза о нормальности отвергается чаще на округленных выборках большего размера.

Можно сделать следующий вывод. Отвержение гипотезы о нормальности критерием Шапиро — Уилка может быть связано с ошибками округления. То есть выборка, на которой мы проводим тестирование, могла быть изначально извлечена из нормально распределенной генеральной совокупности. Но из-за того, что при внесении данных в компьютер числа вводятся с определенной точностью или вовсе округляются каким-либо образом, критерий нормальности отвергнет нулевую гипотезу о нормальности.

2.4 Приложения

Наиболее часто встречающиеся виды округления в медицине – это округление до целого, как округляются рост, вес и, например, объем талии; а также округление до сотых, как округляются всевозможные биомаркеры.

На рисунке 9 на графиках линиями изображены клинико-демографические характеристики из исследования [21]. Видим, что на левом графике рисунка 9 в границах применимости критерия Шапиро — Уилка округление до сотых практически не влияет на отвержение от нормальности для рассматриваемых биомаркеров, так как почти все эти прямые лежат в белой зоне. С другой стороны, на правом графике рисунка 9 изображены линии, соответствующие измерениям роста, веса и талии. Видим, что, например, рост, попадает в красную зону уже на размере выборки $N = 400$.

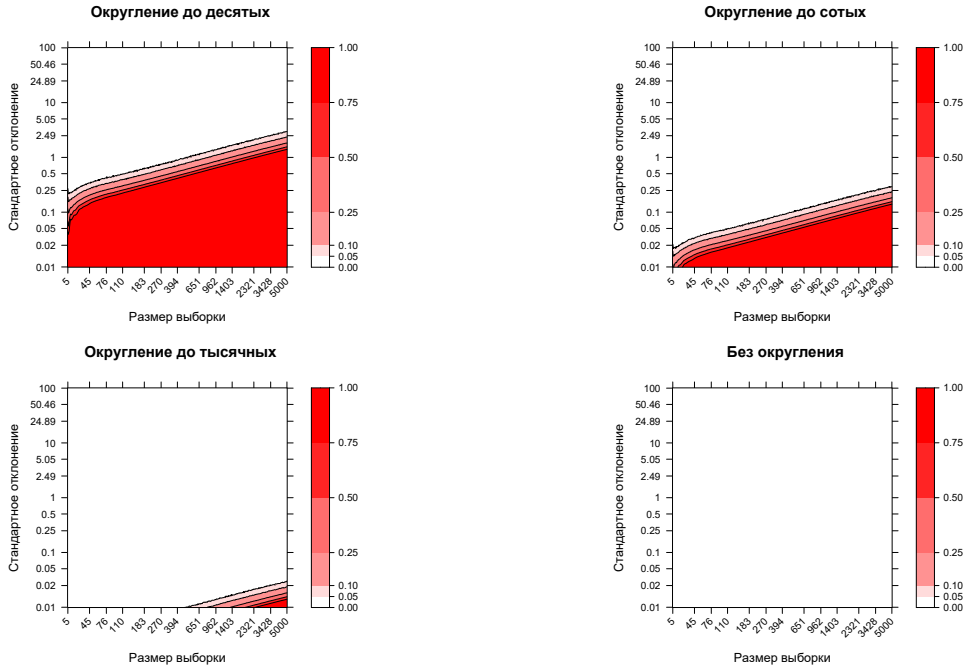

Рис. 8: Тепловые карты результатов моделирования для $f_0(x)$ и $f_5(x)$ – $f_7(x)$.

Однако, какой следует из этого вывод: имели ли исходные данные нормальное распределение, достоверно не известно.

Таким образом, при работе с биомаркерами можно использовать округление до сотых, оно не повлияет значимо на результаты критериев нормальности. Но при округлении роста, веса и талии необходимо применять округление с осторожностью. Наши расчеты могут быть полезны для самопроверки. Если мы знаем размер выборки и стандартное отклонение, мы можем открыть таблицу и посмотреть, в какой зоне мы находимся: в белой или красной, и из этого сделать соответствующие выводы.

3 Ряды Эджворта

В курсовой работе 4 курса я привела подробное теоретическое описание рядов Эджворта. Приведем основные выводы, которые понадобятся нам вдальнейшем.

Пусть $X_1, \ldots, X_n$ – независимые одинаково распределенные случайные величины, имеющие конечное математическое ожидание $\mu < \infty$ и дисперсию $0 < \sigma^2 < \infty$. Согласно центральной предельной теореме случайная величина $Z = \sqrt{n} \cdot \frac{\bar{X}-\mu}{\sigma} \xrightarrow[n\to\infty]{d} N(0,1)$, где $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Ряды Эджворта позволяют получить соответствующие разложения для плотности и функции распределения Z и, таким образом, определить, насколько быстро происходит эта сходимость.
Формула для плотности распределения $f_Z(x)$ случайной величины Z:

$$f_Z(z) = \varphi(z)\left(1 + \frac{\gamma_1}{6\sqrt{n}}(z^3 - 3z) + \frac{\gamma_2}{24n}(z^4 - 6z^2 + 3)\right) + \\ + \varphi(z)\left(\frac{\gamma_1^2}{72n}(z^6 - 15z^4 + 45z^2 - 15) + O(n^{-3/2})\right) \quad (20)$$

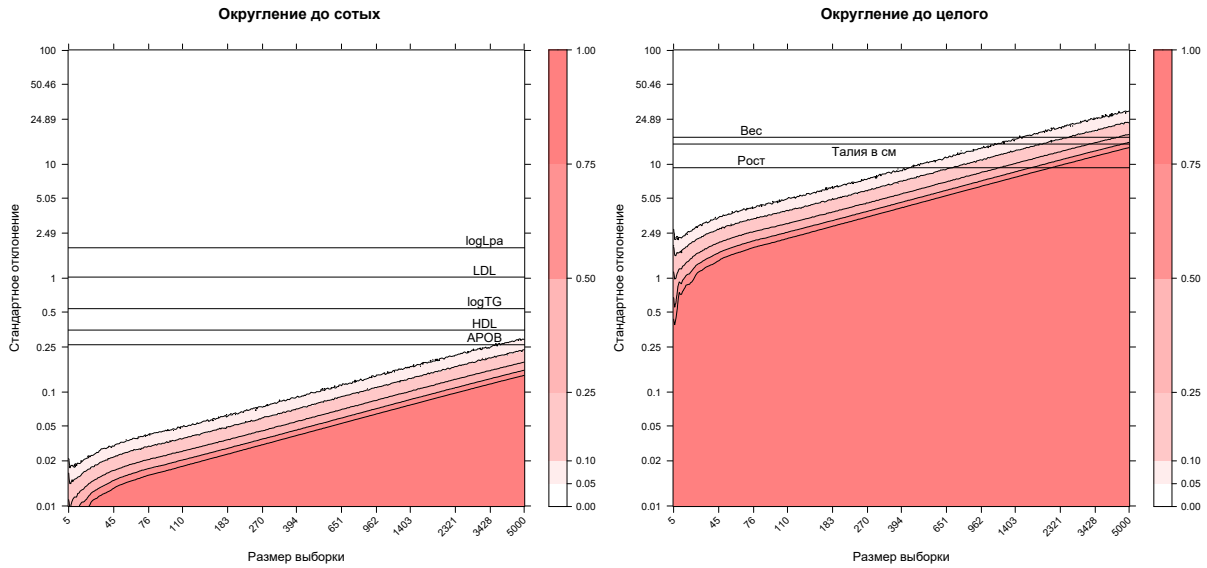

Рис. 9: Тепловые карты результатов моделирования для реальных данных.

Формула для функции распределения $F_Z(x)$ случайной величины Z:

$$F_Z(x) = \Phi(x) + \varphi(x)\left(\frac{\gamma_1}{6\sqrt{n}}(1-x^2) + \frac{\gamma_2}{24n}(3x-x^3) + \frac{\gamma_1^2}{72n}(-x^5+10x^3-15x) + O(n^{-3/2})\right) \quad (21)$$

где $\Phi(x)$ и $\phi(x)$ – соответственно, функция и плотность стандартного нормального распределения, $\gamma_1 = \frac{E(X_1-EX_1)^3}{\sigma^3}$ и $\gamma_2 = \frac{E(X_1-EX_1)^4}{\sigma^4} - 3$ – соответственно, коэффициенты асимметрии и эксцесса X_1.

Пусть теперь X_1 имеет логарифмически нормально распределение $LN(0, \sigma^2)$ с параметрами $\mu = 0$, $0 < \sigma^2 < \infty$, γ_1 и γ_2 – соответственно, коэффициенты асимметрии и эксцесса X_1. Пусть случайная величина $Z = \sqrt{n} \cdot \frac{\bar{X}-\mu}{\sigma}$, где $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ – аналогично определена как и раньше. Пусть $\hat{F}_Z(x) = \frac{1}{n}\sum_{i=0}^{n} I(X_i \leq x)$ – эмпирическая функция распределения случайной величины Z, $q : \Phi(q) = 0.95$ – квантиль стандартного нормального распределения уровня 0.95.
Будем сравнивать значение $1-\hat{F}_Z(q)$ со значением $1-F_Z(q)$, полученным при помощи разложения Эджворта. Для моделирования использовался компьютерный кластер с 48 процессорными ядрами и среда R 3.6.3.
Параметры выборок:

- размеры выборок: {10, 12, ..., 100}$\bigcup${100, 120, ..., 1000}$\bigcup${1500, 2000, ..., 5000}$\bigcup${6000, 7000, ..., 10000};
- асимметрии: {0.1, 0.5, 1, 2, 3.5, 5, 7.5, 10}.

Для каждой выборки вероятность $1 - F_Z(q)$ была оценена на 10000 итераций метода Монте-Карло. Определенные выше значения асимметрии были выбраны не случайно и брались, опираясь на значения, оцененные по выборке исследования [21]:

- триглицериды: 4.62;
- С-реактивный белок: 10.54;
- липопротеин (а): 2.50;
- липопротеины низкой плотности: 0.45.

Результаты моделирования представлены на рисунке 10.

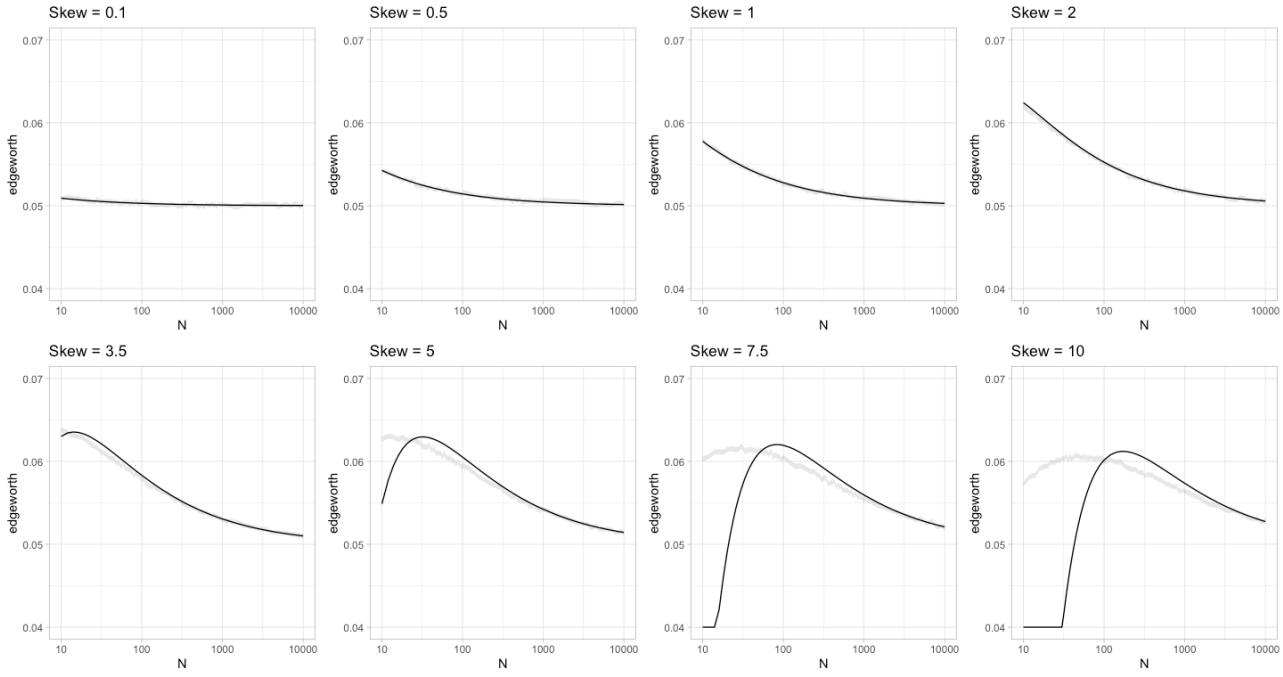

Рис. 10: Сравнение вероятностей, оцененных по выборке для различных значений асимметрии: для эмпирической функции распределения приведен 95 %-ный доверительный интервал оценки вероятности (серым), для разложения Эджворта приведена оценка вероятности (черным).

Обозначим $ECDF$ – значения вероятности, полученные при помощи эмпирической функции распределения, EDG – значения вероятности, полученные в разложении Эджворта. Тогда абсолютная погрешность $\Delta = ECDF - EDG$, относительная погрешность (в процентах) $\delta = \frac{ECDF-EDG}{ECDF} \cdot 100\%$. Графики соответствующих погрешностей представлены на рисунках 11 и 12.

Таким образом, с увеличением асимметрии отклонение оценок вероятностей для эмпирической функции распределения и для разложения Эджворта увеличивается. Заметим, что для значения асимметрии 10 наблюдается значительное отклонение даже для выборок размера 1000.

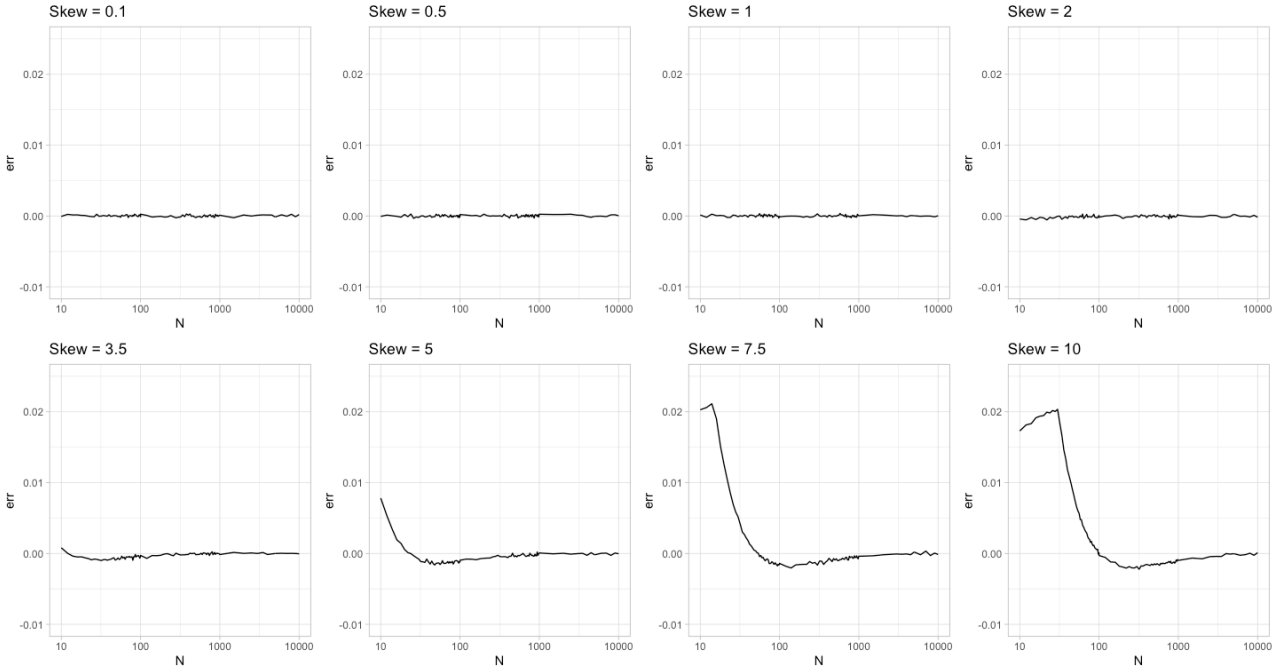

Рис. 11: График абсолютной ошибки.

Заключение

В разделе 1 мы показали, что предположение о распределении роста взрослого человека по Гауссовскому закону является достаточно грубым, и распределение роста точнее описывается логарифмически нормальным распределением. В разделе 2 мы численно исследовали связь между применением критерия Шапиро — Уилка для проверки гипотезы о нормальном распределении к данным, имеющим исходно нормальное распределение, точностью округления и размерами выборок. И, наконец, в разделе 3 мы численно исследовали скорость сходимости рядов Эджворта к реальной функции распределения на примере логарифмически нормального распределения.

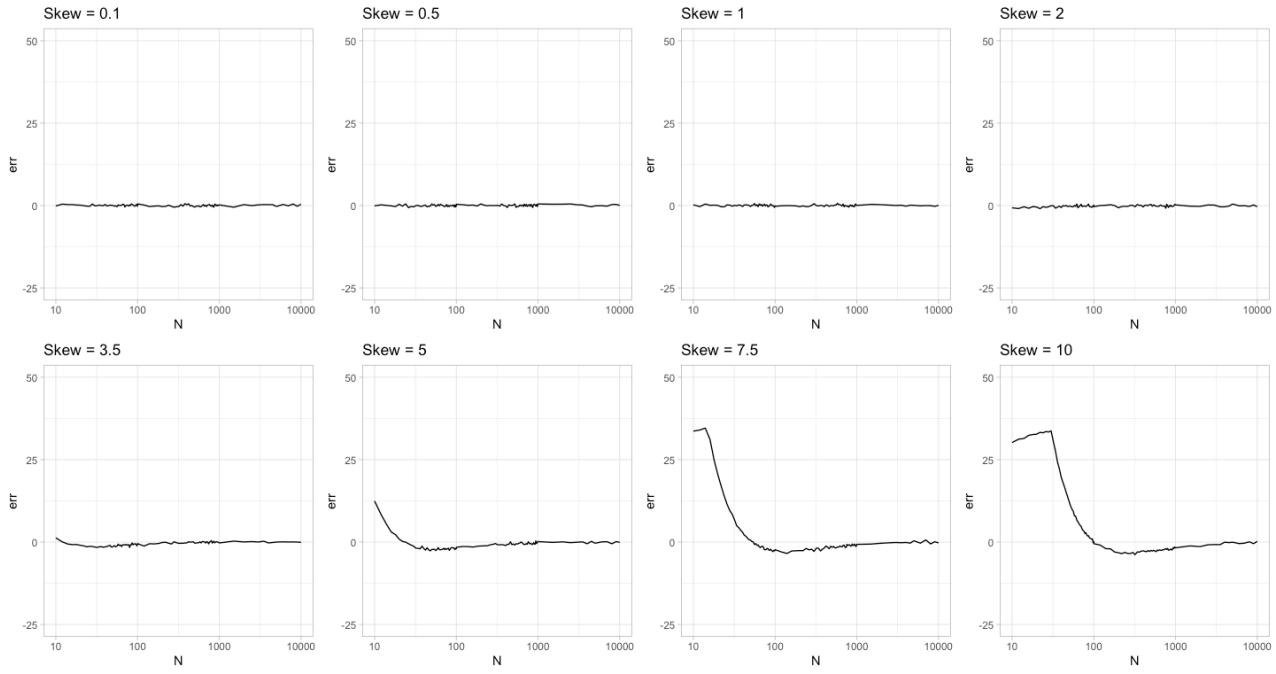

Рис. 12: График относительной ошибки (в процентах).

Список литературы

[1] Slavskii, S. A., Kuznetsov, I. A., Shashkova, T. I., Bazykin, G. A., Axenovich, T. I., Kondrashov, F. A., Aulchenko, Y. S. (2021). The limits of normal approximation for adult height. European Journal of Human Genetics, 29(7), 1082-1091.

[2] Galton F. Regression Towards Mediocrity in Hereditary Stature. The Journal of the Anthropological Institute of Great Britain and Ireland. 1886;15:246–63.

[3] Snedecor GW. Statistical Methods: By George W. Snedecor and William G. Cochran. Iwoa State University Press; 1989. 503 p.

[4] Devore JL, Berk KN. Modern Mathematical Statistics with Applications. Springer, New York, NY; 2012.

[5] Wright S. Evolution and the genetics of populations. Vol. 1. Genetic and biométrie foundations. London and Chicago: University of Chicago Press.; 1968.

[6] Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth Environ Sci Trans R Soc Edinb. 1918;52(2):399–433.

[7] Visscher PM. Sizing up human height variation. Nat Genet. 2008 May;40(5):489–90.

[8] Landau LD, Livshits EM. Statistical physics (in Russian). Gosudarstv. Izdat. Tehn.-Teor. Lit., Moscow; 1938.

[9] Schmitt LH, Harrison GA. Patterns in the within-population variability of stature and weight. Ann Hum Biol. 1988 Sep;15(5):353–64.

[10] McKellar AE, Hendry AP. How Humans Differ from Other Animals in Their Levels of Morphological Variation [Internet]. Vol. 4, PLoS ONE. 2009. p. e6876. Available from: http://dx.doi.org/10.1371/journal.pone.0006876

[11] Soltow L. Inequalities in the Standard of Living in the United States,1798-1875. In: American Economic Growth and Standards of Living before the Civil War. University of Chicago Press; 1992. p. 121–72.

[12] Solomon PJ, Thompson EA, Rissanen A. The inheritance of height in a Finnish population. Ann Hum Biol. 1983 May;10(3):247–56.

[13] Geodakyan VA. Differential mortality and the norm of reaction of males and females (in Russian). Zhurnal Obshey Biologii. 1974;35(3):376–85.

[14] Rawlik K, Canela-Xandri O, Tenesa A. Evidence for sex-specific genetic architectures across a spectrum of human complex traits. Genome Biol. 2016 Jul 29;17(1):166.

[15] Zillikens MC, Yazdanpanah M, Pardo LM, Rivadeneira F, Aulchenko YS, Oostra BA, et al. Sex-specific genetic effects influence variation in body composition. Diabetologia. 2008 Dec;51(12):2233–41.

[16] Falconer DS, Mackay TFC. Introduction to quantitative genetics. 1996.

[17] Aulchenko YS, Struchalin MV, Belonogova NM, Axenovich TI, Weedon MN, Hofman A, et al. Predicting human height by Victorian and genomic methods. Eur J Hum Genet. 2009 Aug;17(8):1070–5.

[18] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.

[19] Subramanian SV, Özaltin E, Finlay JE. Height of nations: a socioeconomic analysis of cohort differences and patterns among women in 54 low- to middle-income countries. PLoS One. 2011;6:e18962.

[20] Langtree I. Height Chart of Men and Women in Different Countries - DisabledWorld. Disabled World. Disabled World; 2017. https://www.disabled-world.com/calculators-charts/height-chart.php.

[21] Бойцов С. А. и др. Исследование ЭССЕ-РФ (Эпидемиология сердечно-сосудистых заболеваний и их факторов риска в регионах Российской Федерации). Десять лет спустя //Кардиоваскулярная терапия и профилактика. – 2021. – Т. 20. – №. 5. – С. 143-152.

Листинг программного кода

```
library(dplyr)
library(plotly)
library(export)
library(zoo)
require(lattice)
require(gridExtra)
require(rasterVis)
library(tictoc)
library(foreach)
library(doParallel)


cl <- makeCluster(47)
registerDoParallel(cl)

LEN = 550
N_REP = 10000
ALPHA_LEVEL = .05

N_mesh = round(10^seq(from = log10(5), to = log10(5000), length.out = LEN))  %>% uniq
a_mesh = 0
LEN = length(N_mesh)
s_mesh = (10^seq(from = log10(0.01), to = log10(100), length.out = LEN))


round_custom = function(z, dec_round = 10)
{
  round(z/dec_round)*dec_round
}

round_mesh = c((-1):4)

shapiro_p = function(z)
{
  if (length(unique(z)) == 1)
  {
    return(0)
  }

  shapiro.test(z)$p.value
}

process_point = function(s,N)
{
  #set.seed(1)
```

```
  sample = rnorm(n = N, mean = 0, sd = s)

  sample_list = list()
  sample_list[[1]] = sample
  sample_list[[2]] = round_custom(sample, dec_round = 10)
  sample_list[[3]] = round_custom(sample, dec_round = 5)
  sample_list[[4]] = round_custom(sample, dec_round = 2)

  sample_list[[5]] = round(sample)
  sample_list[[6]] = round(sample,1)
  sample_list[[7]] = round(sample,2)
  sample_list[[8]] = round(sample,3)

  p_vector = sapply(sample_list, shapiro_p)
  p_vector
}

tic()

pb = txtProgressBar(min = 1, max = LEN, style = 3)
result_over_N = list(LEN)
for (i in 1:LEN)
{
  setTxtProgressBar(pb, value = i)


  sd_row = foreach (j=1:LEN) %dopar%
    {
      s_tmp = s_mesh[j]
      N_tmp = N_mesh[i]
      rep_matrix = replicate(process_point(s_tmp, N_tmp), n = N_REP)
      reject_matrix = (rep_matrix < ALPHA_LEVEL)
      reject_prob = rowSums(reject_matrix)/N_REP

      grid_matrix = t(as.matrix(reject_prob))

      rownames(grid_matrix) = s_mesh[j]
      colnames(grid_matrix) = c("Pure", "dec_10", "dec_5", "dec_2", 0:3)
      grid_matrix
    }

  result = do.call(rbind, sd_row)

  result_over_N[[i]] = result
}

toc()
```

```
stopCluster(cl)
########################

res_dec10_pure <- as.data.frame(matrix(FALSE, nrow = LEN, ncol = LEN))

for (i in c(1:LEN))
  for (j in c(1:LEN))
    res_dec10_pure[i,j] <- result_over_N[[i]][j, 2] - result_over_N[[i]][j, 1]

res_dec5_pure <- as.data.frame(matrix(FALSE, nrow = LEN, ncol = LEN))

for (i in c(1:LEN))
  for (j in c(1:LEN))
    res_dec5_pure[i,j] <- result_over_N[[i]][j, 3] - result_over_N[[i]][j, 1]

res_dec2_pure <- as.data.frame(matrix(FALSE, nrow = LEN, ncol = LEN))

for (i in c(1:LEN))
  for (j in c(1:LEN))
    res_dec2_pure[i,j] <- result_over_N[[i]][j, 4] - result_over_N[[i]][j, 1]

res_round_pure <- as.data.frame(matrix(FALSE, nrow = LEN, ncol = LEN))

for (i in c(1:LEN))
  for (j in c(1:LEN))
    res_round_pure[i,j] <- result_over_N[[i]][j, 5] - result_over_N[[i]][j, 1]

res_round1_pure <- as.data.frame(matrix(FALSE, nrow = LEN, ncol = LEN))

for (i in c(1:LEN))
  for (j in c(1:LEN))
    res_round1_pure[i,j] <- result_over_N[[i]][j, 6] - result_over_N[[i]][j, 1]

res_round2_pure <- as.data.frame(matrix(FALSE, nrow = LEN, ncol = LEN))

for (i in c(1:LEN))
  for (j in c(1:LEN))
    res_round2_pure[i,j] <- result_over_N[[i]][j, 7] - result_over_N[[i]][j, 1]


res_round3_pure <- as.data.frame(matrix(FALSE, nrow = LEN, ncol = LEN))

for (i in c(1:LEN))
  for (j in c(1:LEN))
    res_round3_pure[i,j] <- result_over_N[[i]][j, 8] - result_over_N[[i]][j, 1]
```

```
res_dec10_pure_matrix <- res_dec10_pure %>% as.matrix()
res_dec5_pure_matrix <- res_dec5_pure %>% as.matrix()
res_dec2_pure_matrix <- res_dec2_pure %>% as.matrix()

res_round_pure_matrix <- res_round_pure %>% as.matrix()
res_round1_pure_matrix <- res_round1_pure %>% as.matrix()
res_round2_pure_matrix <- res_round2_pure %>% as.matrix()
res_round3_pure_matrix <- res_round3_pure %>% as.matrix()


numbers_s <- c(0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1, 2.5, 5, 10, 25, 50, 100)
index_s <- c()
for (i in 1:length(numbers_s)){
  index_s[i] <- which.min( abs(s_mesh - numbers_s[i]) )
}
index_s


window_mean = function(X, k)
{
  X1= rollmean(X,k)
  rollmean(t(X1),k) %>% t() %>% as.matrix()
}

fast_process = function(z)
{
  window_mean(abs(z), k = 1)+0.0001
}

cut_mesh = c(0,0.05,0.1,0.25,0.5,0.75,1)

levelplot_function = function(z,name)
{
  lvl = levelplot(fast_process(z), scales=list(x = list(at=index_s,labels=N_mesh[inde
                                                        y=list(at=index_s,labels=
            col.regions=hsv(1, c(seq(0,1,length.out = length(cut_mesh)+1)) , 1),
            colorkey = list(at=cut_mesh,
                            labels=list(at=cut_mesh)), contour = T,
            at = cut_mesh, xlab = "Размер выборки", ylab = "Стандартное отклонение",
            main = name)
  lvl
}

lvl_list = list()
lvl_list[[1]] = levelplot_function(res_dec10_pure_matrix, "Округление до десяти")
lvl_list[[2]] = levelplot_function(res_dec5_pure_matrix, "Округление до пяти")
```

```
lvl_list[[3]] = levelplot_function(res_dec2_pure_matrix, "Округление до двух")
lvl_list[[4]] = levelplot_function(res_round_pure_matrix, "Округление до целого")
lvl_list[[5]] = levelplot_function(res_round1_pure_matrix, "Округление до десятых")
lvl_list[[6]] = levelplot_function(res_round2_pure_matrix, "Округление до сотых")
lvl_list[[7]] = levelplot_function(res_round3_pure_matrix, "Округление до тысячных")
lvl_list[[8]] = levelplot_function(matrix(data = 0, nrow = NROW(res_round3_pure_matri
                                          ncol = NCOL(res_round3_pure_matrix)), "Без


grid.arrange(grobs = lvl_list[1:4] ,ncol = 2, nrow = 2)
graph2eps(width = 16, height = 9, file = "fig1")

grid.arrange(grobs = lvl_list[5:8] ,ncol = 2, nrow = 2)
graph2eps(width = 16, height = 9, file = "fig2")


process_coord = function(z)
{
  which.min(abs(s_mesh - z))
}

OFFSET = 0.2
F1 = levelplot(fast_process(res_round2_pure_matrix), scales=list(x = list(at=index_s,
                                        y=list(at=index_s,labels=round(s_mesh[index_s]
          col.regions=hsv(1, seq(0,0.5,length.out = length(cut_mesh)+1) ),
          colorkey = list(at=cut_mesh,
                          labels=list(at=cut_mesh)), contour = T,
          at = cut_mesh, xlab = "Размер выборки", ylab = "Стандартное отклонение",
          main = "Округление до сотых",
          panel = function(...){
            panel.levelplot(...)
            panel.abline(h = process_coord(1.02), col = 1, lwd = 1)
            panel.text(370,process_coord(1.02),"LDL",pos=3, offset = OFFSET)

            panel.abline(h = process_coord(0.35), col = 1, lwd = 1)
            panel.text(370,process_coord(0.35),"HDL",pos=3, offset = OFFSET)

            panel.abline(h = process_coord(0.54), col = 1, lwd = 1)
            panel.text(370,process_coord(0.54),"logTG",pos=3, offset = OFFSET)

            panel.abline(h = process_coord(0.26), col = 1, lwd = 1)
            panel.text(370,process_coord(0.26),"APOB",pos=3, offset = OFFSET)

            panel.abline(h = process_coord(1.83), col = 1, lwd = 1)
            panel.text(370,process_coord(1.83),"logLpa",pos=3, offset = OFFSET)

            })
```

```
F2 = levelplot(fast_process(res_round_pure_matrix), scales=list(x = list(at=index_s,l
                                                                y=list(at=index_s,lab
               col.regions=hsv(1, seq(0,0.5,length.out = length(cut_mesh)+1) ),
               colorkey = list(at=cut_mesh,
                               labels=list(at=cut_mesh)), contour = T,
               at = cut_mesh, xlab = "Размер выборки", ylab = "Стандартное отклонение"
               main = "Округление до целого",
               panel = function(...){
                 panel.levelplot(...)
                 panel.abline(h = process_coord(9.26), col = 1, lwd = 1)
                 panel.text(70,process_coord(9.26),"Рост",pos=3, offset = OFFSET)

                 panel.abline(h = process_coord(17.2), col = 1, lwd = 1)
                 panel.text(70,process_coord(17.2),"Вес",pos=3, offset = OFFSET)

                 panel.abline(h = process_coord(15), col = 1, lwd = 1)
                 panel.text(170,process_coord(15),"Талия в см",pos=1, offset = OFFSET)

               })


grid.arrange(grobs = list(F1,F2) ,ncol = 2, nrow = 1)
graph2eps(width = 16, height = 9, file = "fig3")

rotate <- function(x) t(apply(x, 2, rev))


tmp_mat = fast_process(res_round_pure_matrix)
dim(tmp_mat)
grid = expand.grid(y=s_mesh, x=N_mesh)
grid$z = as.numeric(t(tmp_mat))

c(0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1, 2.5, 5, 10, 25, 50, 100)

levelplot(z~x*y,grid,
          col.regions=hsv(1, seq(0,0.5,length.out = length(cut_mesh)+1) ),
          colorkey = list(at=cut_mesh,
                          labels=list(at=cut_mesh)), contour = T,
          at = cut_mesh, xlab = "Размер выборки", ylab = "Стандартное отклонение",
          main = "Округление до целого")

levelplot(z~x*y,grid,
          col.regions=hsv(1, seq(0,0.5,length.out = length(cut_mesh)+1) ),
          scales = list(x = list(log = 2.7),y = list(log = 2.7)),
          colorkey = list(at=cut_mesh,
```

```
                        labels=list(at=cut_mesh)), contour = T,
          at = cut_mesh, xlab = "Размер выборки", ylab = "Стандартное отклонение",
          main = "Округление до целого")

levelplot(fast_process(res_round_pure_matrix),
          col.regions=hsv(1, seq(0,0.5,length.out = length(cut_mesh)+1) ),
          colorkey = list(at=cut_mesh,
                        labels=list(at=cut_mesh)), contour = T,
          at = cut_mesh, xlab = "Размер выборки", ylab = "Стандартное отклонение",
          main = "Округление до целого")


A = matrix(data = 1:9, ncol = 3)
grid = expand.grid(y=1:3, x=1:3)
grid$z = as.numeric(t(A))
levelplot(A)
levelplot(z~x*y, grid)
```