ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. ЛОМОНОСОВА»

МЕХАНИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ

КАФЕДРА ТЕОРИИ ВЕРОЯТНОСТЕЙ

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
специалиста

ПОДХОДЫ К ОЦЕНКЕ ОТКЛОНЕНИЙ РАСПРЕДЕЛЕНИЯ ДАННЫХ ОТ НОРМАЛЬНОГО ЗАКОНА

Выполнила студентка 608 группы
Дьячкова Екатерина Николаевна

_____

подпись студента

Научный руководитель:
д.ф.-м.н., профессор
Яровая Елена Борисовна

_____

подпись научного руководителя

Москва
2023 год

Содержание

Введение

В руководствах по прикладной статистике при использовании статистических критериев предлагается проводить предварительное тестирование данных на нормальность распределения [1], [2]. **Целью выпускной квалификационной работы** является исследование существующих подходов к оценке отклонения распределения данных от нормального закона с использованием инструментов, которые ранее были недоступны авторам пособий по прикладной статистике, – численных методов и практических результатов в области биостатистики.

В **главе 1** мы приведем теоретическое описание наиболее распространенных критериев нормальности и сравним их мощности.

В **главе 2** мы обратимся к прикладным данным. Мы обсудим результаты работы из области количественной генетики [10], в которой на данных Биобанка Великобритании показано, что рост взрослого человека, который исторически признавался нормальным [12], [13], [14], разумно считать логарифмически нормально распределенной величиной.

В **главе 3** мы исследуем влияние округления на результаты критериев нормальности, поскольку это один из факторов, который часто используют и который может нам помешать детектировать нормальность на практике.

Далее в **главе 4** мы продемонстрируем результаты систематического обзора статей из двух ведущих российских журналов по кардиологии: «Рациональная фармакотератия в кардиологии» и «Российский кардиологический журнал». Мы покажем, как часто используется двухэтапная процедура с предварительной проверкой на нормальность и критериями Стьюдента и Манна – Уитни и проведем эмпирическое исследование ошибки I рода такой процедуры.

В **главе 5** в качестве альтернативного подхода к оценке отклонения данных от нормальности, мы рассмотрим ряды Эджворда и изучим их асимптотическую сходимость. Поскольку на интересующих нас выборках до 100 элементов ошибка в оценке отклонения от нормальности оказалась слишком велика, мы перейдем к моделям, где данные извлекаются из логарифмически нормального распределения, а также смеси нормального и логарифмически нормального распределения, а к полученнным выборкам применяется критерий Стьюдента.

Глава 6 будет посвящена исследованию ошибки I рода и мощности критерия Стьюдента в рамках указанных выше моделей. Мы покажем, что при проверке одновыборочной гипотезы для контроля ошибки I рода достаточно использовать показатели асимметрии. Поэтому далее мы обратимся к модифицированному одновыборочному критерию Стьюдента с поправкой на асимметрию, приведем его теоретическое описание и сравним его показатели ошибки I рода и мощности с показателями для классического одновыборочного критерия Стьюдента.

1 Эмпирическое исследование мощности критериев нормальности распределения

1.1 Элементы проверки гипотез

Для лучшего восприятия материала приведём вступительную теорию из [3]. Пусть нам дана выборка $\mathcal{X} = (X_1, ..., X_n)$ из распределения P, которое нам неизвестно, но мы знаем, что оно принадлежит некоторому параметризованному семейству распределений: $P \in \mathcal{P} = \{P_\theta$, где $\theta \in \Theta\}$. Для проверки предположений о виде такого распределения используют статистические критерии.
Параметрические статистические гипотезы — это пара из предположения H_0 о неизвестном параметре и альтернативы H_1:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0 \end{cases} \tag{1}$$

Если множество Θ_0 (Θ_1) состоит из одной точки, то гипотезу H_0 (альтернативу H_1) называют *простой*; в противном случае гипотезу (или альтернативу) называют *сложной. Статистическим критерием* называется правило, по которому принимают или отклоняют гипотезу H_0. Это правило строится следующим образом: выбирается *критериальная статистика* $S = S(\mathcal{X})$ — функция, зависящая от выборки, но не от параметра, и *критическая область* G. Если критериальная статистика попала в критическое множество: $S \in G$, то мы отклоняем гипотезу H_0. При тестировании возможны два вида ошибок: *ошибка I рода* — принимаем H_1, когда на самом деле верна H_0, и, наоборот, *ошибка II рода* — принимаем H_0, когда на самом деле верна H_1. Введём обозначение: $P_{\theta'}(X) = P(X|\theta = \theta')$, где $\theta' \in \Theta$ — некоторое фиксированное значение. Тогда *мощностью* статистического критерия называют вероятность отвергнуть гипотезу H_0 при значении параметра θ:

$$w(\theta) = P_\theta(S \in G) \tag{2}$$

Задачей отбора подходящего теста является минимизация ошибок и, соответственно, максимизация мощности статистического критерия. Вероятности ошибок I, II рода — $\alpha(\theta)$, $\beta(\theta)$, уровень значимости — α и мощность связаны следующим образом:

$$\begin{cases} w(\theta) = P_\theta(S \in G) = P(H_1|H_0) = \alpha(\theta) \leq \alpha \ \forall \theta \in \Theta_0 \\ w(\theta) = P_\theta(S \in G) = P(H_1|H_1) = 1 - P(H_0|H_1) = 1 - \beta(\theta) \ \forall \theta \in \Theta_1 \end{cases} \tag{3}$$

Если зафиксированы уровень значимости α, критериальная статистика S и простые гипотезы:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases} \tag{4}$$

задача нахождения наиболее мощного критерия разрешима т.к., зависит только от вида критической области G. Перебирая различные множества, выбираем тот тест,

мощность которого будет наибольшей:

$$\begin{cases} P_{\theta_0}(S \in G) = P_0(G) \leq \alpha \\ P_{\theta_1}(S \in G) = P_1(G) = 1 - \beta \to max_G \end{cases} \tag{5}$$

В случае, когда альтернатива не является простой, а полученный тест оказывается наиболее мощным для каждого фиксированного $\theta_1 \in \Theta_1$, такой критерий называют *равномерно наиболее мощным.*

1.2 Теорема Неймана-Пирсона

На основании материала, изложенного в [4], рассмотрим базовый пример: тест Стьюдента. Пусть дана выборка из нормального распределения: $\mathcal{X} = (X_1, ..., X_n)$, где $X_i \sim N(\mu, \sigma^2)$ с неизвестным средним и известной дисперсией. Фиксируем уровень значимости α. Рассмотрим гипотезу с правосторонней критической областью, вида:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (\mu_0 < \mu_1) \end{cases} \tag{6}$$

В качестве критериальной статистики рассмотрим выборочное среднее $S = \bar{X}$: $\bar{X} \sim_{H_0} N(\mu_0, \frac{\sigma^2}{n})$, $\bar{X} \sim_{H_1} N(\mu_1, \frac{\sigma^2}{n})$. Найдём квантиль t_α для критической области.

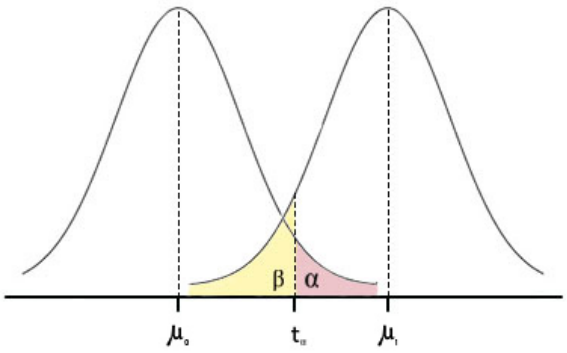

Рис. 1: Распределение статистики S при гипотезах H_0 и H_1.

Т.к., $\sqrt{n} \cdot \frac{\bar{X} - \mu_0}{\sigma} \sim_{H_0} N(0,1) \Rightarrow 1 - \alpha = P_0\left(\bar{X} \leq t_\alpha\right) = P_0\left(\sqrt{n} \cdot \frac{\bar{X} - \mu_0}{\sigma} \leq \sqrt{n} \cdot \frac{t_\alpha - \mu_0}{\sigma}\right) = = \Phi\left(\sqrt{n} \cdot \frac{t_\alpha - \mu_0}{\sigma}\right) = \Phi\left(q_{1-\alpha}\right)$, где Φ — функция стандартного нормального распределения, $q_{1-\alpha}$ — квантиль уровня $1 - \alpha$. Следовательно, искомый квантиль $t_\alpha = \mu_0 + \frac{(\sigma \cdot q_{1-\alpha})}{\sqrt{n}}$.
Получили критерий: выборочное среднее $\bar{X} \geqslant t_\alpha \Rightarrow$ отклоняем H_0.

Вернёмся к задаче вычисления мощности при фиксированной критериальной статистике. Аналогично примерам из [4] рассмотрим одновыборочный тест Стьюдента с правосторонней альтернативой. Пусть дана выборка размера $n = 100$ из нормального

распределения с неизвестным средним μ и дисперсией $\sigma^2 = 4$, $\alpha = 0.05$. Сформулируем гипотезы:

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu > 0 \end{cases} \tag{7}$$

В отличие от уже известных моделей, рассмотрим следующие нестандартные критические области: Зафиксируем альтернативу в гипотезах 7, получим тест Стьюдента

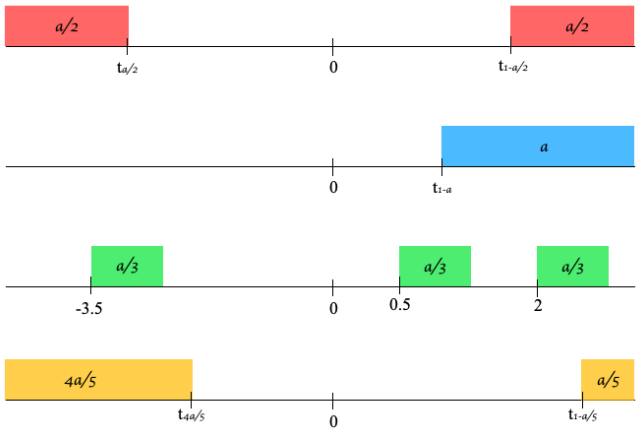

Рис. 2: Рассматриваемые критические области для теста Стьюдента.

с простыми гипотезами:

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu = \mu_1 \end{cases} \tag{8}$$

Рассмотрим график зависимости мощности тестов от односторонних альтернатив: Цвет каждой линии графика на рисунке 3 соответствует цвету критической области на рисунке 2. Как видно из рисунка 3, для произвольной фиксированной альтернативы тест с голубой критической областью будет иметь наибольшую мощность среди четырёх рассмотренных. Покажем, что произвольный статистический тест с такой критической областью и односторонними гипотезами вида 7 будет иметь наибольшую мощность среди тестов со всеми возможными критическими областями.
Напомним теорему Неймана-Пирсона, которую мы цитируем по учебнику [5].

Теорема 1 (Нейман-Пирсон) *Дана выборка рамера n. Вводим систему вложенных множеств $G_c := \{x \in \mathbb{R}^n : \frac{p_1(x)}{p_0(x)} \geq c\}$ и функцию $\varphi(c) := P_0(G_c)$, требуем выполнение двух условий:*

1. *плотности выборки $p_0(x)$ и $p_1(x)$, при H_0 и H_1 соответственно, положительны при всех $x \in \mathbb{R}^n$;*

2. *для заданного уровня $\alpha \in (0; 1)$ существует $c = c_\alpha : \varphi(c_\alpha) = \alpha$ (всегда выполнено при непрерывной φ).*

Тогда при заданных условиях (1) и (2) наиболее мощный критерий уровня $\alpha \in (0; 1)$ задаётся критическим множеством $G^* := G_{c_\alpha} = \{x \in \mathbb{R}^n : \frac{p_1(x)}{p_0(x)} \geq c_\alpha\}$.

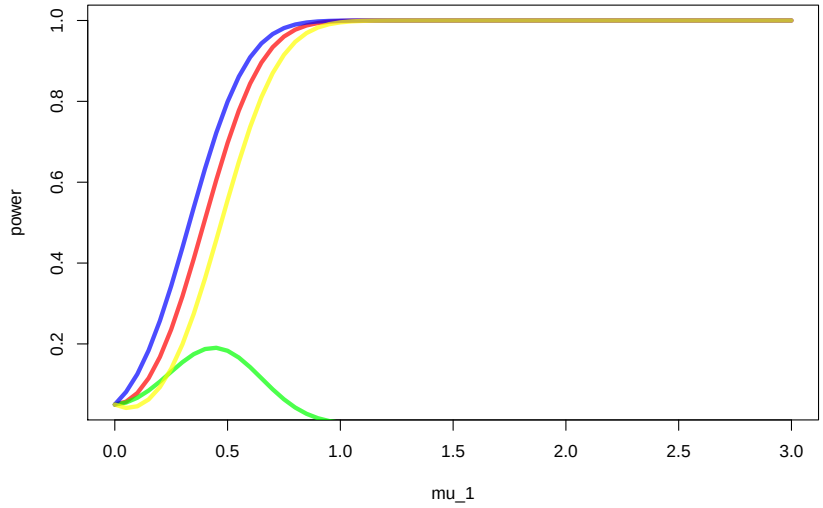

Рис. 3: График зависимости мощности теста Стьюдента от односторонней альтернативы.

В примере с односторонней альтернативой и тестом Стьюдента голубая критическая область задавалась следующим образом:

$$\{X \in \mathbb{R}^n : S = \sqrt{n} \cdot \frac{\bar{X} - \mu_0}{\sigma} \geq t_{1-\alpha}\} = \{X \in \mathbb{R}^n : \bar{X} \geq \frac{\sigma \cdot t_{1-\alpha}}{\sqrt{n}} + \mu_0\} = \tag{9}$$

$$= \{X \in \mathbb{R}^n : \bar{X} \geq c, \text{ где } c \in \mathbb{R}\} \tag{10}$$

По теореме Неймана-Пирсона критическая область наиболее мощного теста в условиях этой задачи задаётся множеством:

$$\{X \in \mathbb{R}^n : \frac{p_1(X)}{p_0(X)} = \exp^{-\frac{\sum_{i=0}^{n}(X_i-\mu_1)^2}{2\cdot\sigma^2} + \frac{\sum_{i=0}^{n}(X_i-\mu_0)^2}{2\cdot\sigma^2}} \geq c_\alpha\} = \tag{11}$$

$$= \{X \in \mathbb{R}^n : \bar{X} \geq \frac{2 \cdot \sigma^2 \cdot \ln c_\alpha + n \cdot (\mu_1^2 - \mu_0^2)}{2 \cdot n \cdot (\mu_1 - \mu_0)}\} = \tag{12}$$

$$= \{X \in \mathbb{R}^n : \bar{X} \geq c_\alpha^{'}, \text{ где } c_\alpha^{'} \in \mathbb{R}\} \tag{13}$$

Видим, что множества 10 и 13 являются односторонними для некоторых констант c и c'_α. Для фиксированного уровня значимости α эти константы совпадут. Условия теоремы Неймана-Пирсона выполнены. Критическая область наиболее мощного теста с гипотезами 8 имеет вид 10 и не зависит от альтернативы, т.е., величины $\mu_1 > 0$. Поэтому тест с гипотезами 7 и голубой критической областью является *равномерно наиболее мощным.*

Как мы показали в предыдущем абзаце, теорема Неймана-Пирсона явным образом доказывает, что равномерно наиболее мощный критерий в случае односторонней альтернативы существует. Рассмотрим пример, демонстрирующий, что в случае двусторонней альтернативы наиболее мощный тест для каждой фиксированной μ_1 существует, но равномерно наиболее мощного — нет.

Пусть дана выборка размера $n = 100$ из нормального распределения с неизвестным средним μ и дисперсией $\sigma^2 = 4$. Используем одновыборочный тест Стьюдента, $\alpha = 0.05$ и двусторонние гипотезы:

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases} \tag{14}$$

Аналогично примеру теста Стьюдента с односторонней альтернативой проводим тесты для каждой из четырёх критических областей на рисунке 2 с фиксированными альтернативами:

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu = \mu_1 \end{cases} \tag{15}$$

Строим графики зависимости мощности тестов от альтернативы. На рисунке 4 на

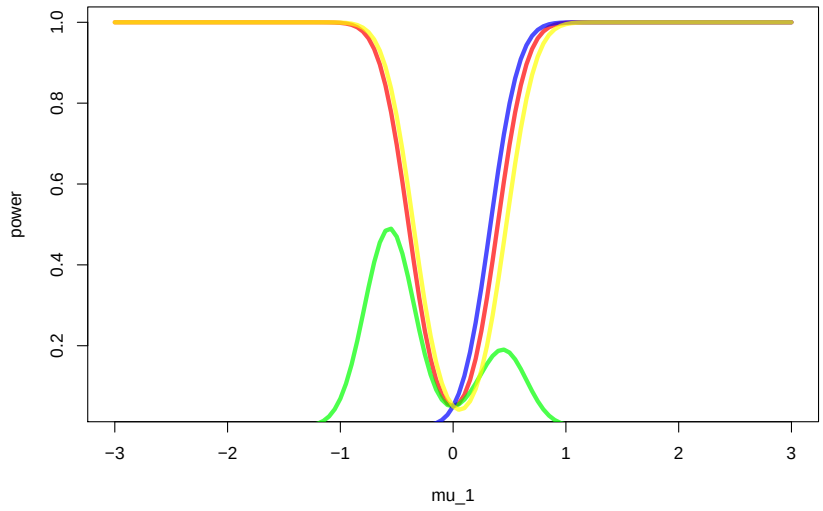

Рис. 4: Графики зависимости мощности тестов Стьюдента от двусторонней альтернативы.

правой полуоси $\mu_1 > 0$ наибольшая мощность у теста с правой критической областью, это вытекает из теоремы Неймана-Пирсона. Мощнее него при $\mu_1 > 0$ — нет. Тем не менее, на левой полуоси $\mu_1 < 0$ наибольшая мощность у теста с жёлтой критической областью. Получается, что на разных полуосях наибольшую мощность имеют разные тесты и в этом примере теста с наибольшей мощностью для любой μ_1 нет.
То есть, как и утверждалось, для теста с двусторонней альтернативой равномерно наиболее мощного теста не существует.

1.3 Метод Монте-Карло

В предыдущем подразделе примеры были связаны с тестом Стьюдента, а оценки мощности критериев вычислялись аналитически. На практике гипотеза может

иметь сложную критическую область из-за чего аналитическое вычисление может быть трудоёмким. Поэтому для вычисления оценки мощности статистических тестов можно использовать метод Монте-Карло, описанный в [6] и [7]. Этот метод при помощи закона больших чисел позволяет вычислять оценку мощности произвольного статистического критерия при фиксированной альтернативе и заданной критической области. Приведём алгоритм метода Монте-Карло:

1. Пусть заданы гипотеза $H_0 : \theta \in \Theta_0$ и альтернатива $H_1 : \theta \in \Theta_1$;
2. Фиксируем простую альтернативу $H_1 : \theta = \theta_1$;
3. Выбираем достаточно большое натуральное число K и генерируем K раз выборку $\mathcal{X}$ при условиях гипотезы H_1;
4. Для каждой выборки вычисляем статистику t_{H_1};
5. Выясняем, попала ли статистика в критическую область G. Если да, то $m_i = I(t_{H_1} \in G) = 1$;
6. Вычисляем количество отвержений гипотезы H_0: $M = \sum_{i=1}^{K} m_i$;
7. Вычисляем оценку мощности теста: $W = \frac{M}{K}$.

Краткая схема метода Монте-Карло:

$$\forall i = 1, \ldots, K : \mathcal{X}_{H_1} \longrightarrow t_{H_1} \longrightarrow m_i = I(t_{H_1} \in G^*) \longrightarrow M = \sum_{i=1}^{K} m_i \longrightarrow W = \frac{M}{K} \quad (16)$$

Покажем преимущества этого метода на примере теста Стьюдента с двусторонней альтернативой и зелёной критической областью. Для сравнения аналитического вычисления и метода Монте-Карло нами были написаны две программы в среде статистического анализа R 3.5.1. Коды программ представлены на рисунках 5-6: при аналитическом вычислении легко ошибиться, требуется подробное описание критической области при помощи формул, которые для каждой области свои; метод Монте-Карло — более универсальный. Также можно увидеть, как быстро этот метод

```
myp=seq(1,121,by=1)
pw=function(i){
  m=pt(q1, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)-
    pt(-3.5, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)+
    pt(q2, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)-
    pt(0.5, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)+
    pt(q3, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)-
    pt(2, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)
  return(m)
}

for (j in myp) {
  myp[j] = pw((j-1)/20 -3)
}
```

Рис. 5: Аналитическое вычисление

```
K=1000 #количество итераций
s=2 #sd выборки
mc_pwr=function(i)
{
  res_plus=0
  for( j in 1:K)
  {
    X=rnorm(n = N, mean = i, sd = s)
    test=t.test(X)
    win=(test$statistic>-3.5 & test$statistic<q1) |
      (test$statistic>0.5 & test$statistic<q2) |
      (test$statistic>2 & test$statistic<q3)
    res_plus=res_plus+win
  }
  return(res_plus/K)
}
myc=seq(1,121,by=1)
for (j in myc) {
  myc[j] = mc_pwr((j-1)/20 -3)
}
```

Рис. 6: Метод Монте-Карло

достигает необходимой точности. При $K = 1000$ оценка мощности методом Монте-Карло близка к оценке мощности, вычисленной аналитически: в 60% точек графика

на рисунке 7 абсолютная ошибка не превышает 5%. При $K = 10^4$ графики на рисунке 8 уже почти неотличимы: в 80% точек абсолютная ошибка не превышает 5%.

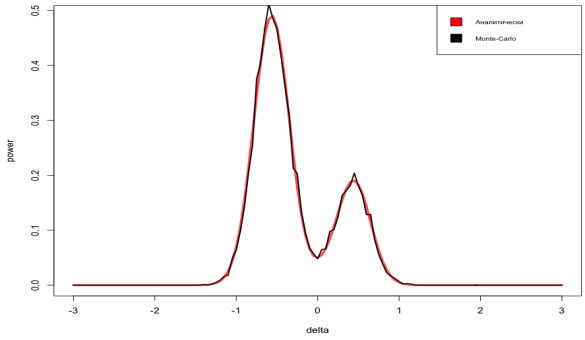

Рис. 7: $K = 1000$

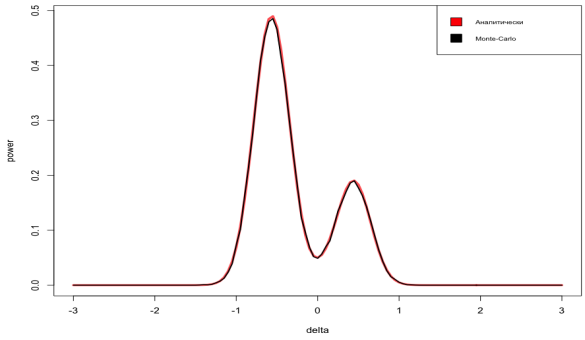

Рис. 8: $K = 10^4$

1.4 Критерий Лиллиефорса

Воспользуемся работой [8]. Критерий Лиллиефорса является модификацией теста Колмогорова. Пусть дана выборка $\mathcal{X} = (X_1, \ldots, X_n)$ с неизвестной функцией распределения $F(x)$. Фиксируем уровень значимости α, формулируем гипотезы:

$$\begin{cases} H_0 : F(x) \in \mathcal{F} = \{\Phi_{\mu,\sigma^2}(x) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\} \\ H_1 : F(x) \notin \mathcal{F} \end{cases} \tag{17}$$

Вычисляем оценки: $\hat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i = \bar{X}$, $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. Статистика теста вычисляется по формуле:

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - \Phi_{\hat{\mu},s^2}(x)| \tag{18}$$

где $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$ — эмпирическая функция распределения выборки $\mathcal{X}$, а $\Phi_{\mu,s^2}(x)$ — функция нормального распределения со средним $\hat{\mu}$ и дисперсией s^2. Из свойств нормального распределения:

$$\Phi_{\hat{\mu},s^2}(x) = \Phi_{0,1}\left(\frac{x - \hat{\mu}}{s}\right) \tag{19}$$

Для индикаторов в эмпирической функции распределения выполнено:

$$\forall i = 1, \ldots, n : I(X_i \leq x) = I\left(\frac{X_i - \hat{\mu}}{s} \leq \frac{x - \hat{\mu}}{s}\right) \tag{20}$$

Вводим обозначения: $Y_i = \frac{X_i - \hat{\mu}}{s}$, $i = 1, \ldots, n$ и $y = \frac{x - \hat{\mu}}{s}$ и получаем:

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - \Phi_{\hat{\mu},s^2}(x)| = \sup_{-\infty<y<+\infty} |\tilde{F}_n(y) - \Phi_{0,1}(y)| \tag{21}$$

где $\tilde{F}_n(y)$ — эмпирическая функция распределения выборки $\mathcal{Y} = (Y_1, \ldots, Y_n)$. Из свойств нормального распределения:

$$\hat{\mu} \sim \frac{1}{n} N(n\mu, n\sigma^2) = N\left(\mu, \frac{\sigma^2}{n}\right); \quad S \sim \sigma^2 \frac{1}{n-1} \chi^2_{n-1} \tag{22}$$

$$\Rightarrow Y_i = \frac{X_i - \frac{1}{n}\sum X_i}{S_X} \sim \frac{N_1(\mu, \sigma^2) - N_2\left(\mu, \frac{\sigma^2}{n}\right)}{\sqrt{\sigma^2 \frac{1}{n-1}\chi^2_{n-1}}} \sim \frac{N_1(0, \sigma^2) - N_2\left(0, \frac{\sigma^2}{n}\right)}{\sqrt{\sigma^2 \frac{1}{n-1}\chi^2_{n-1}}} \sim$$
$$\sim \frac{\sigma N_1(0,1) - \sigma N_2\left(0, \frac{1}{n}\right)}{\sigma\sqrt{\frac{1}{n-1}\chi^2_{n-1}}} \sim \frac{N_1(0,1) - N_2\left(0, \frac{1}{n}\right)}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}} \sim \frac{Z_i - \frac{1}{n}\sum Z_i}{S_Z} \tag{23}$$

где $Z_i \sim N(0,1)$. Если H_0 верна, критериальная статистика сходится по распределению к распределению Лиллиефорса. Критические значения находятся при помощи таблиц или метода Монте-Карло.

1.5 Критерий χ^2 Пирсона

Опираемся на теорию из [5]. Пусть дана выборка: $\mathcal{X} = (X_1, \ldots, X_n)$ с неизвестной функцией распределения $F(x)$. Фиксируем уровень значимости α, формулируем гипотезы:

$$\begin{cases} H_0 : F(x) \in \mathcal{F} = \{\Phi_{\mu,\sigma^2}(x) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\} \\ H_1 : F(x) \notin \mathcal{F} \end{cases} \tag{24}$$

Разбиваем область значений X_1 на N промежутков: $\Delta_j = (a_i; b_j]$, $j = 1, ..., N$ и перегруппировываем выборку т.е., переходим к случайным величинам $\nu_j = \sum\limits_{i=1}^{n} I(X_i \in \Delta_j)$ — количество X_i, попавших в Δ_j, $j = 1 \ldots, N$. Вектор $\nu = (\nu_1, \ldots, \nu_N)$ имеет мультиномиальное распределение. Вводим обозначение $p_j(\mu, \sigma^2) = P(X_1 \in \Delta_j)$. Вычисляем оценку максимального правдоподобия по сгруппированным данным:

$$(\hat{\mu}, \hat{\sigma^2}) = argmax_{(\mu,\sigma^2)} P\left(\nu_1 = l_1, ..., \nu_N = l_N\right) =$$
$$= argmax_{(\mu,\sigma^2)} \frac{n!}{l_1! \cdot \ldots \cdot l_N!} \cdot [p_1(\mu, \sigma^2)]^{l_1} \cdot \ldots \cdot [p_N(\mu, \sigma^2)]^{l_N} =$$
$$= argmax_{(\mu,\sigma^2)} \sum_{j=1}^{N} l_j \cdot \ln p_j(\mu, \sigma^2) \tag{25}$$

Критериальная статистика выглядит следующим образом:

$$\chi^2_n = \sum_{j=1}^{N} \frac{(Observed_j - Expected_j)^2}{Expected_j} = \sum_{j=1}^{N} \frac{\left(\nu_j - n \cdot p_j(\hat{\mu}, \hat{\sigma^2})\right)^2}{n \cdot p_j(\hat{\mu}, \hat{\sigma^2})} \tag{26}$$

По теореме Фишера, доказанной в 1924 году [3], если H_0 верна, $\chi^2_n \xrightarrow[n\to\infty]{d} \chi^2(N-3)$.

1.6 Критерий Харке – Бера

Этот тест основан на поиске отклонений выборочного распределения от нормального при помощи коэффициентов асимметрии и эксцесса. У нормального распределения они принимают нулевые значения. Пусть дана выборка: $\mathcal{X} = (X_1, \ldots, X_n)$. Выборочные коэффициенты асимметрии S и эксцесса K вычисляются по следующим формулам:

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^{3/2}}, \; K = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^2} - 3 \tag{27}$$

Интуитивно ΔK и ΔS можно изобразить как на рисунке 9, где синим цветом изображена плотность нормального распределения, а оранжевым цветом – плотность произвольного выборочного распределения. Фиксируем уровень значимости α, фор-

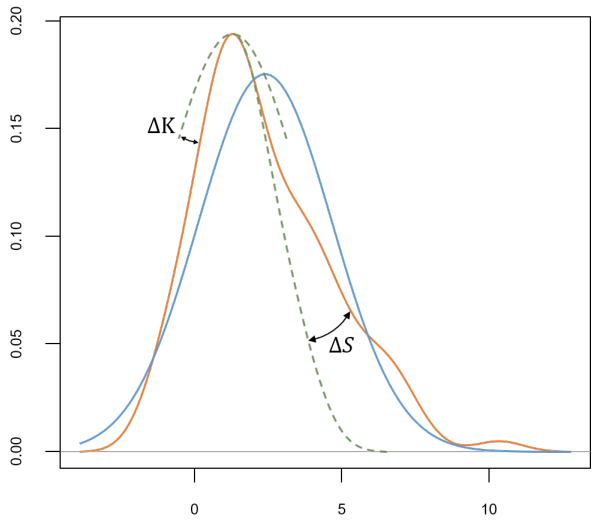

Рис. 9: Идея критерия Харке – Бера.

мулируем гипотезы:

$$\begin{cases} H_0 : S = 0, K = 0 \\ H_1 : S \neq 0 \text{ и(или) } K \neq 0 \end{cases} \tag{28}$$

Критериальная статистика вычисляется следующим образом:

$$JB = n\left(\frac{S^2}{6} + \frac{K^2}{24}\right) \tag{29}$$

Если H_0 выполнена, коэффициенты S и K асимптотически нормальны и $JB \xrightarrow[n\to\infty]{d} \chi^2(2)$.

1.7 Критерий Шапиро – Уилка

Воспользуемся работой [9]. Пусть дана выборка: $\mathcal{X} = (X_1, \ldots, X_n)$ с неизвестной функцией распределения $F(x)$. Фиксируем уровень значимости α, формулируем гипотезы:

$$\begin{cases} H_0 : F(x) \in \mathcal{F} = \{\Phi_{\mu,\sigma^2}(x) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\} \\ H_1 : F(x) \notin \mathcal{F} \end{cases} \tag{30}$$

Пусть $\mathcal{Y} = (Y_1, \ldots, Y_n)$ – выборка из стандартного нормального распределения и, соответственно, $Y_{(1)} < \ldots < Y_{(n)}$ – вариационный ряд порядковых статистик. Пусть $m = (m_1, ..., m_n)$ — вектор математических ожиданий порядковых статистик из стандартного нормального распределения и $V = ||v_{i,j}||_{i,j=1}^n$ — ковариационная матрица порядковых статистик из стандартного нормального распределения. То есть:

$$EY_{(i)} = m_i,\ i = 1 \ldots, n. \tag{31}$$

$$cov(Y_{(i)}, Y_{(j)}) = v_{i,j},\ i, j = 1, \ldots, n. \tag{32}$$

Если выборка $\mathcal{X}$ была извлечена из нормального распределения со средним μ и дисперсией σ^2, тогда:

$$X_{(i)} = \mu + \sigma \cdot Y_{(i)},\ i = 1, \ldots n. \tag{33}$$

Согласно обобщенной теореме наименьших квадратов [5], а также того, что нормальное распределение симметрично, наилучшие линейные несмещенные оценки для μ и σ вычисляются следующим образом:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \hat{\sigma} = \frac{m^T V^{-1} \mathcal{X}}{m^T V^{-1} m} \tag{34}$$

Пусть $s^2 = \sum\limits_{i=1}^{n}(X_i - \bar{X})^2$ – несмещённая оценка для $(n-1) \cdot \sigma^2$.
Тогда статистика критерия Шапиро – Уилка вычисляется следующим образом:

$$W = \frac{\left(\sum\limits_{i=1}^{n} a_i \cdot X_{(i)}\right)^2}{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2} = \left(\frac{R^2\hat{\sigma}}{C}\right)^2 \frac{1}{s^2} = \frac{b^2}{s^2} \tag{35}$$

где коэффициенты a_i, R^2 и C вычисляются так:

$$(a_1, \ldots, a_n) = \frac{m^T \cdot V^{-1}}{\sqrt{m^T \cdot V^{-1} \cdot V^{-1} \cdot m}} \tag{36}$$

$$R^2 = m^T V^{-1} m \tag{37}$$

$$C = \sqrt{m^T V^{-1} V^{-1} m} \tag{38}$$

Рассмотрим график нормальной вероятности на рисунке 10, который является частным случаем qq-plot для нормального распределения. По оси абсцисс откладываем $m_1, ..., m_n$, по оси ординат – $X_{(1)}, \ldots, X_{(n)}$. Используя обобщенный метод наимень-

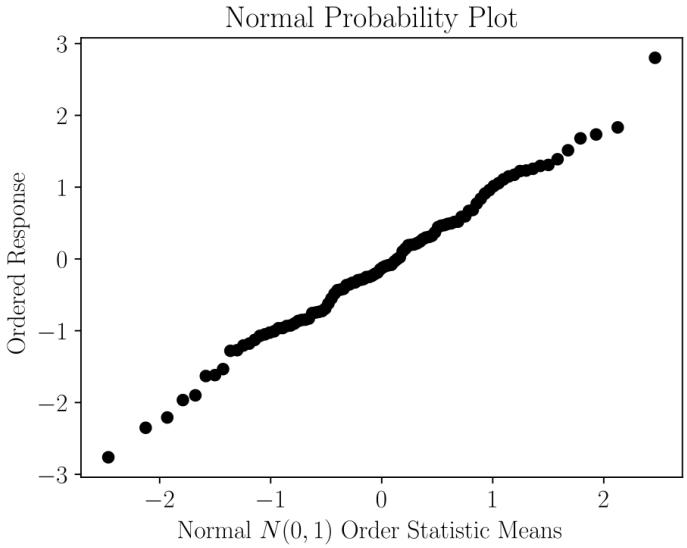

Рис. 10: Идея критерия Шапиро – Уилка.

ших квадратов, проведем линию регрессии. Обычный метод наименьших квадратов в данной ситуации неприменим в силу коррелированности порядковых статистик. Из определения коэффициента b получаем, что b, с точностью до константы C, является наилучшей линейной несмещенной оценкой коэффициента наклона полученной линии регрессии. Константа C является в данном случае нормирующим множителем. Именно эта идея лежит в основе критерия Шапиро – Уилка. Заметим, что, если нулевая гипотеза H_0 на самом деле верна, то и b^2, и s^2, с точностью до константы, являются оценками одной и той же величины – дисперсии σ^2 популяции, из которой была извлечена выборка $\mathcal{X}$. Если же выборка была извлечена из не нормально распределенной генеральной совокупности, то, в общем случае, b^2 и s^2 оценивают разные величины. Критические значения для заданного уровня значимости α, с которыми необходимо сравнить полученное значение критериальной статистики W, находятся из таблиц.

1.8 Эмпирическое исследование мощности

Для исследование мощности тестов на нормальность, рассмотренных в разделах 1.4 – 1.7, нами был использован метод Монте-Карло. Уровень значимости фиксировали $\alpha = 0.05$, далее генерировали $K = 10^4$ раз выборки размера $n = 10, \ldots, 2000$ из четырёх распределений:

1. Бета с параметрами $\alpha = 2$ и $\beta = 2$, его плотность:

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{39}$$

2. Гамма с параметрами $k = 4$ и $\theta = 5$, его плотность:

$$p(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} \exp -\frac{x}{\theta} \tag{40}$$

3. Гамма с параметрами $k = 1$, $\theta = 5$;

4. Распределение Лапласа с параметрами $\alpha = 1$ и $\beta = 0$, его плотность:

$$p(x) = \frac{\alpha}{2} \exp -\alpha|x - \beta| \tag{41}$$

Далее мы применяли метод Монте-Карло и вычисляли оценки мощности тестов. Результаты представлены на рисунках 11-14.

Таким образом, из рассмотренных в разделах 1.4 – 1.7 критериев наибольшей мощ-

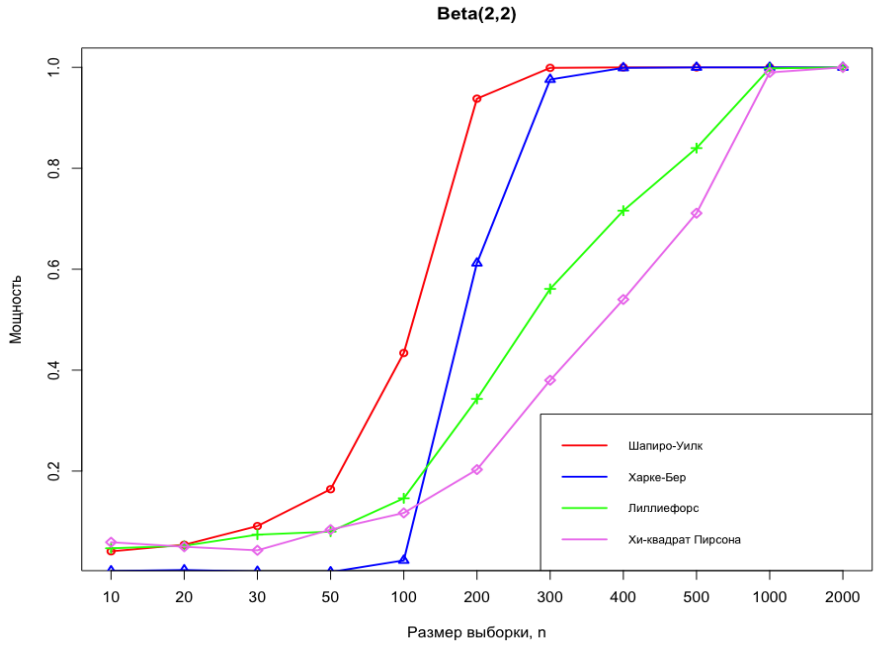

Рис. 11: Сравнение мощности тестов на нормальность на распределении Бета(2,2).

ностью на исследуемых распределениях обладает тест Шапиро — Уилка.

2 Ограничения нормальной аппроксимации

2.1 Введение

В этом разделе мы проведем разбор статьи [10] 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 [11]. То есть

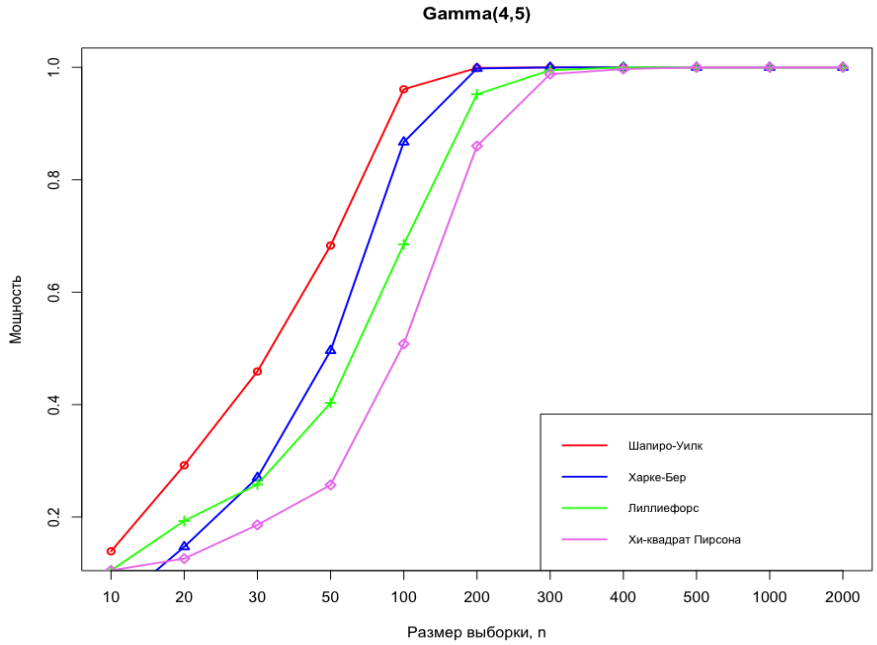

Рис. 12: Сравнение мощности тестов на нормальность на распределении Гамма(4,5).

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 [12], [13], [14]. Наследование роста исторически описывалось так называемой аддитивной полигенной моделью, где суммируются многие генетические эффекты [15], [16]:

$$height = \mu + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon \tag{42}$$

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

Рассмотрим коэффициент вариации CV_ξ случайной величины ξ:

$$CV_\xi = \frac{\sigma}{\mu} \tag{43}$$

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

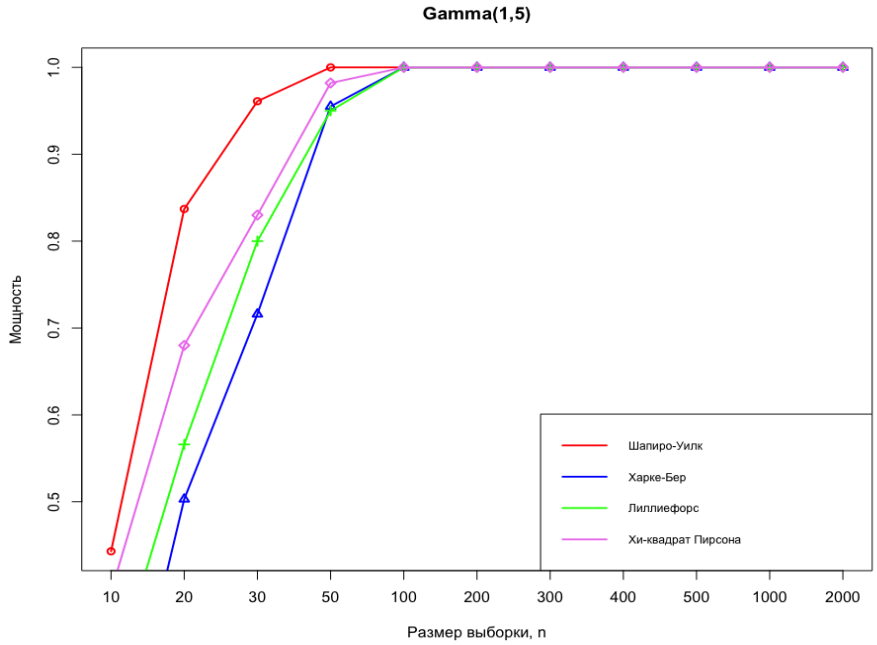

Рис. 13: Сравнение мощности тестов на нормальность на распределении Гамма(1,5).

$\xi_1, \ldots, \xi_n$ – н.о.р., случайные величины с $E\xi_1 = \mu$ и $\sqrt{D\xi_1} = \sigma$. Тогда:

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

$$\Rightarrow (CV_{\sum_{i=1}^n \xi_i})^2 \sim \frac{1}{n} \quad (44)$$

Выражение 44 верно и в более общем случае, для строго положительных, не идеально коррелированных случайных величин [17]. 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 ситуацию как на рисунке 15 смесь нормальных распределений роста и нормальность остатков от регрессии на социально-демографические параметры.

Авторы статьи [10] 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,

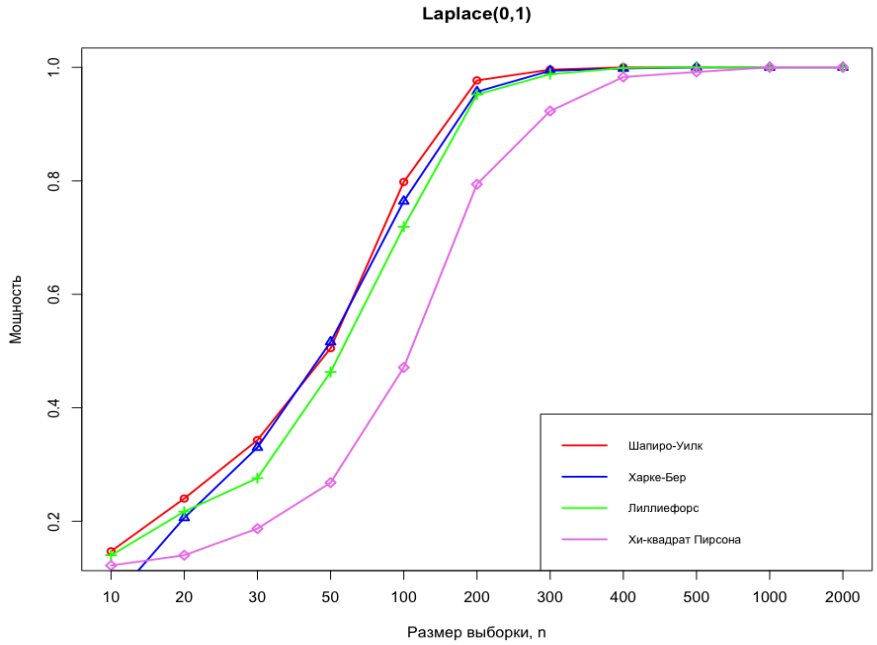

Рис. 14: Сравнение мощности тестов на нормальность на распределении Лапласа(0,1).

было больше в более высоких популяциях, а коэффициент вариации между популяциями был низкий и довольно стабильный. На рисунке 16 из [18] показано, как дисперсия роста женщин увеличивается при увеличении среднего роста женщин. А на рисунке 17 из [19] 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 [20], 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 [11] в своей работе поступал иначе, он делал мультипликативную поправку на рост, умножая рост женщин на 1.08. Отметим, что если мы говорим об аддитивной модели и нормально аппроксимации модели наследования роста человека, то такие поправки делать запрещено. В работе [21] 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

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
...	...	...

рост женщин должен умножаться на некоторую константу, а дисперсия роста женщин, соответственно, на квадрат этой константы. Поэтому в работе [21] 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 [10] 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 [22] 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 объяснением может быть то, что воздействие окружающей среды на рост распределяется по-разному для мужчин и женщин, что приводит к различию в величине различий,

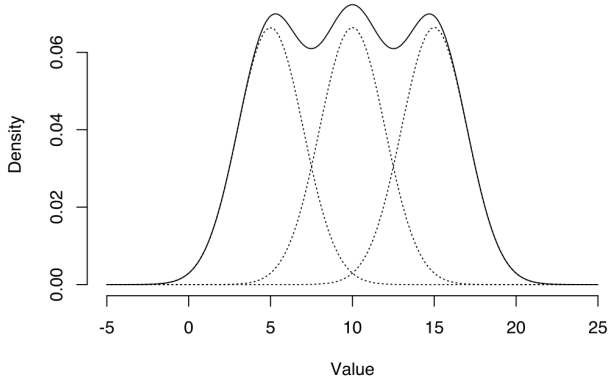

Рис. 15: График плотности смеси нормальных распределений.

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 [23], [24].
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 [12], [13], [14]?
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 [25] 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 [26].
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 [27].

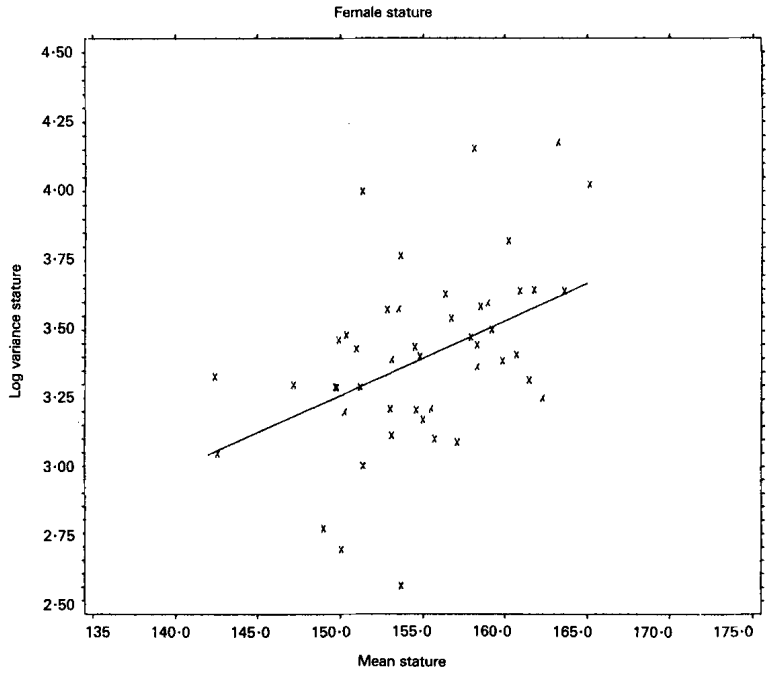

Рис. 16: График зависимости дисперсии роста женщин от среднего роста женщин [18].

2.2 Исследование моделей наследования роста

У нас есть две основные идеи:

1. Аддитивная модель наследования роста:

$$height = \mu + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n + \varepsilon \tag{45}$$

2. Мультипликативная модель наследования роста:

$$height = \mu \cdot b_1^{X_1} \cdot b_2^{X_2} \cdot \ldots \cdot b_n^{X_n} \cdot \varepsilon \tag{46}$$

   или что то же самое:

$$log_{10}(height) = \mu' + b'_1 X_1 + b'_2 X_2 + \ldots + b'_n X_n + \varepsilon' \tag{47}$$

   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 [10], был проведен регрессионный анализ между стандартным отклонением (SD), коэффициентом вариации (CV) и средним ростом женщин из 54 развивающихся стран по данным из [28]. 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).

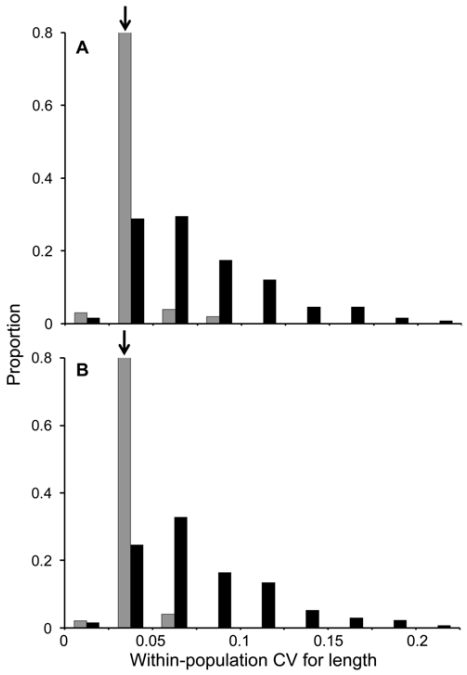

Рис. 17: Распределения коэффициентов вариации (CV) длины тела животных и роста человека внутри популяции. Показаны видовые значения для животных (черный) и популяционные значения для людей (серый) для самцов (A) и самок (B). Стрелки указывают расположение CV для среднего роста человека [19].

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 [29]. 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 графике (A) на рисунке 18 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.

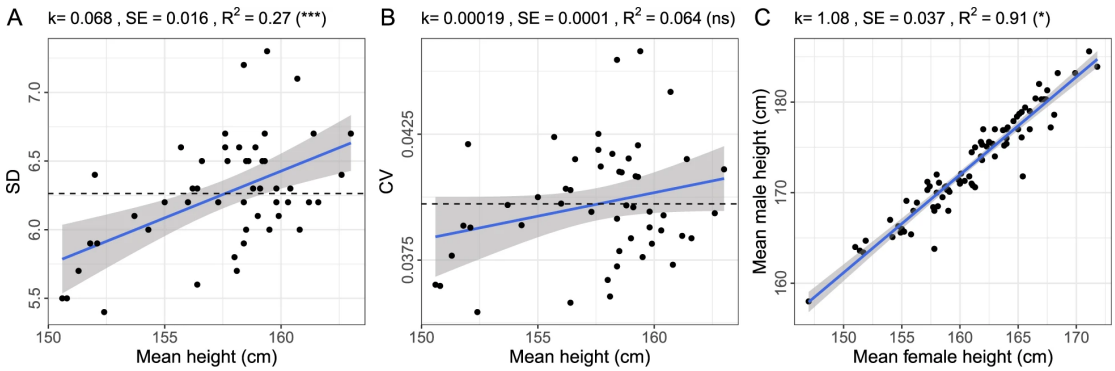

Рис. 18: **Связь между параметрами распределения роста взрослого человека по популяциям.** Линейная регрессия SD (A) и CV (B) роста на средний рост женщин из [28]. Пунктирная линия показывает общее среднее значение. (C) Линейная регрессия среднего роста мужчин на средний рост женщин в популяциях из [29]. 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).

2.3 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 [10]). 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 [10].

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 19. На графике A рисунка 19 для среднего роста стандартное 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

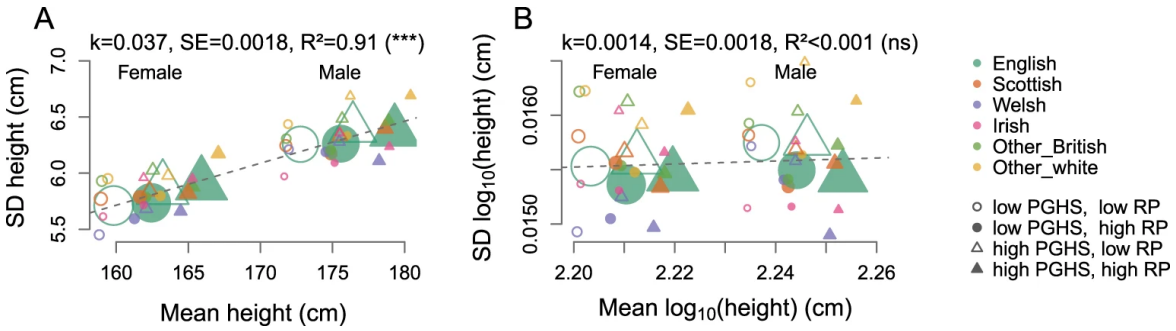

Рис. 19: Графики зависимости SD от среднего роста и от среднего $\log_{10}$ роста в британском Биобанке. Отношение SD к среднему росту (A) и логарифмическому росту (B) для шести групп британских лиц белого происхождения из британского Биобанка, определенных на основе места рождения и разделенных по полу, 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).

монопопуляции, внутри одной, английской, национальности.

Для создания рисунка 20 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 20 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 19 и 20 подтверждают предположения о некорректности аддитивной модели для наследования роста человека. При использовании аддитивной модели мы не должны наблюдать увеличения стандартного отклонения и размеров эффектов при увеличении зависимой переменной. И, наоборот, для среднего логарифмического роста таких противоречий с аддитивной моделью не наблюдалось.

2.4 Мультипликативные и эпистатические взаимодействия

Как было ранее отмечено в разделе 2.1, 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 [10] были рассмотрены модели линейной регрессии для наследования роста 48

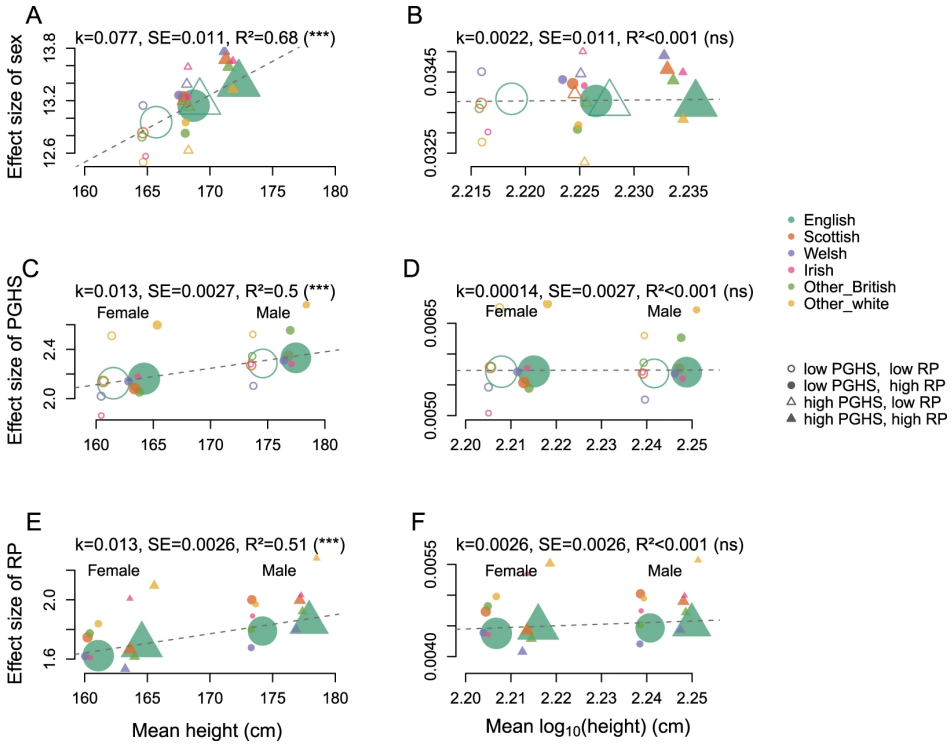

Рис. 20: Графики зависимости влияния различных факторов на средний рост и на средний $\log_{10}$ роста в британском Биобанке. Связь между оценкой величины эффекта пола (A, B), генотипа (C, D; генотип определялся как полигенный показатель роста, PGHS), других факторов (E, F; линейный остаточный предиктор, 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).

и логарифма роста 49, где в качестве предикторов были не только пол (sex), полигенный показательно роста (PGHS) и остаточный предиктор (RP), но и произведения этих трех факторов – sex·PGHS, sex·RP, RP·PGHS. Уровень значимости фиксировали $\alpha = 0.05$. Результаты представлены в дополнительных таблицах [10]. Приводим часть этих результатов (для полной выборки) в таблице 2 для модели 48 и таблице 3 для модели 49 соответственно.

$$height = \mu + \beta_1 \cdot sex + \beta_2 \cdot PGHS + \beta_3 \cdot RP + \\ + \beta_4 \cdot \mathbf{sex} \cdot \mathbf{PGHS} + \beta_5 \cdot \mathbf{sex} \cdot \mathbf{RP} + \beta_6 \cdot \mathbf{PGHS} \cdot \mathbf{RP} + \varepsilon \quad (48)$$

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	

$$log_{10}(height) = \mu + \beta_1 \cdot sex + \beta_2 \cdot PGHS + \beta_3 \cdot RP + \\ + \beta_4 \cdot \mathbf{sex} \cdot \mathbf{PGHS} + \beta_5 \cdot \mathbf{sex} \cdot \mathbf{RP} + \beta_6 \cdot \mathbf{PGHS} \cdot \mathbf{RP} + \varepsilon \quad (49)$$

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 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 [10], понимается любое отклонение от аддитивности генетических эффектов. Поэтому для анализа значимости эпистатических взаимодействий в [10] была рассмотрена модель 50 для роста и модель 51 для логарифма роста, где в качестве эпистатического взаимодействия рассматривается квадрат полигенного показателя PGHS·PGHS. p-значения для фактора PGHS·PGHS представлены в формулах 50 и 51 соответственно:

$$height = \mu + \beta_1 \cdot PGHS + \beta_2 \cdot PGHS^2 + \varepsilon \ (p_{PGHS^2} = 4 \cdot 10^{-7}) \quad (50)$$

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

$$log_{10}(height) = \mu + \beta_1 \cdot PGHS + \beta_2 \cdot PGHS^2 + \varepsilon \ (p_{PGHS^2} = 0.48) \tag{51}$$

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.

2.5 Итоговое сравнение моделей

На последнем этапе исследования в статье [10] 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 [10], 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%.

3 Влияние округления на результаты критериев нормальности

3.1 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.

3.2 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:

$$f_0(x) = x \tag{52}$$

$$f_1(x) = round\left(\frac{x}{10}\right) * 10 \tag{53}$$

$$f_2(x) = round\left(\frac{x}{5}\right) * 5 \tag{54}$$

$$f_3(x) = round\left(\frac{x}{2}\right) * 2 \tag{55}$$

$$f_4(x) = round(x) \tag{56}$$

$$f_5(x) = round(x, 1) \tag{57}$$

$$f_6(x) = round(x, 2) \tag{58}$$

$$f_7(x) = round(x, 3) \tag{59}$$

Далее для каждого фиксированного $\sigma_i \in \sigma_{mesh}$, для каждого фиксированного $N_j \in N_{mesh}$ и для каждой фиксированной функции $f_l(x), l = 0 \dots 7$ из (52)-(59) 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. В нашем случае $K = 10^4$. и Генерируем K раз выборку размера N_j (при условиях гипотезы H_0). Обозначим полученные выборки: $\mathcal{X}_k, k = 1 \dots K$;

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

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

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 (52)-(59) 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 (52)-(59). Пример таблицы для $N_1 = 5$ представлен в таблице 5. В наших обозначениях каждое значение в таблице

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
...	...	...	...	...	...	...	...	...

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:

$$\Delta_l = \begin{pmatrix} \delta_{l,1,1} & \delta_{l,2,1} & \ldots & \delta_{l,L,1} \\ \delta_{l,1,2} & \delta_{l,2,2} & \ldots & \delta_{l,L,2} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{l,1,L} & \delta_{l,2,L} & \ldots & \delta_{l,L,L} \end{pmatrix}, l = 0, \ldots, 7 \tag{60}$$

Строки матрицы 60 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 21 и 22.

3.3 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. Сравнивая графики для разных видов округления, видно как получаются разные исходы. Заметим, что с увеличением объема выборки, размер красной области на графиках рисунков 21 и 22 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 округ-

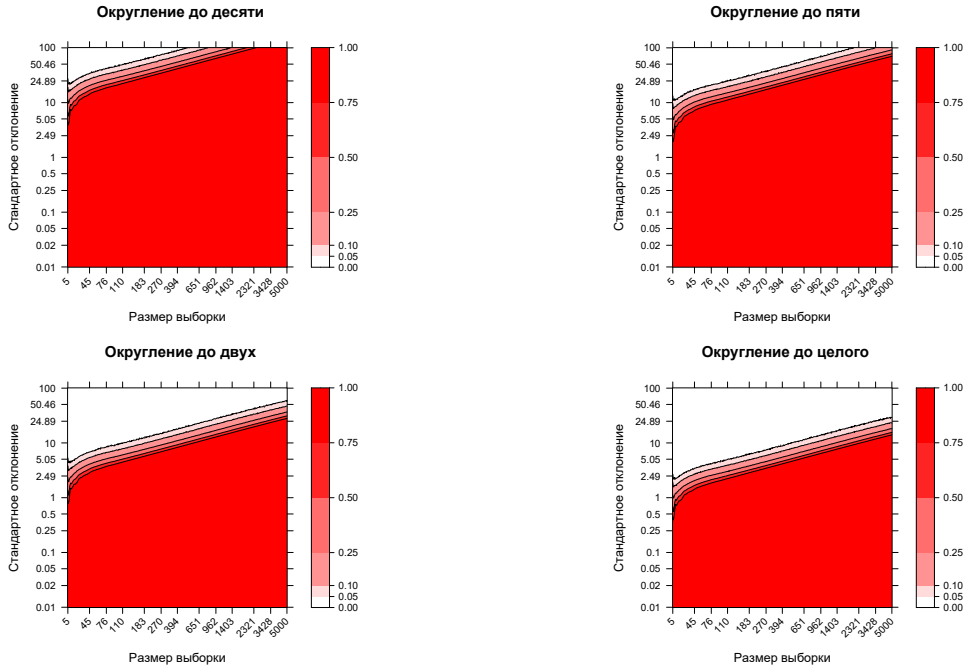

Рис. 21: Тепловые карты результатов моделирования для $f_1(x)$ – $f_4(x)$.

ляются каким-либо образом, критерий нормальности отвергнет нулевую гипотезу о нормальности.

3.4 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 23 на графиках линиями изображены клинико-демографические характеристики из исследования [30]. Видим, что на левом графике рисунка 23 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 23 изображены линии, соответствующие измерениям роста, веса и талии. Видим, что, например, рост, попадает в красную зону уже на размере выборки $N = 400$. 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.

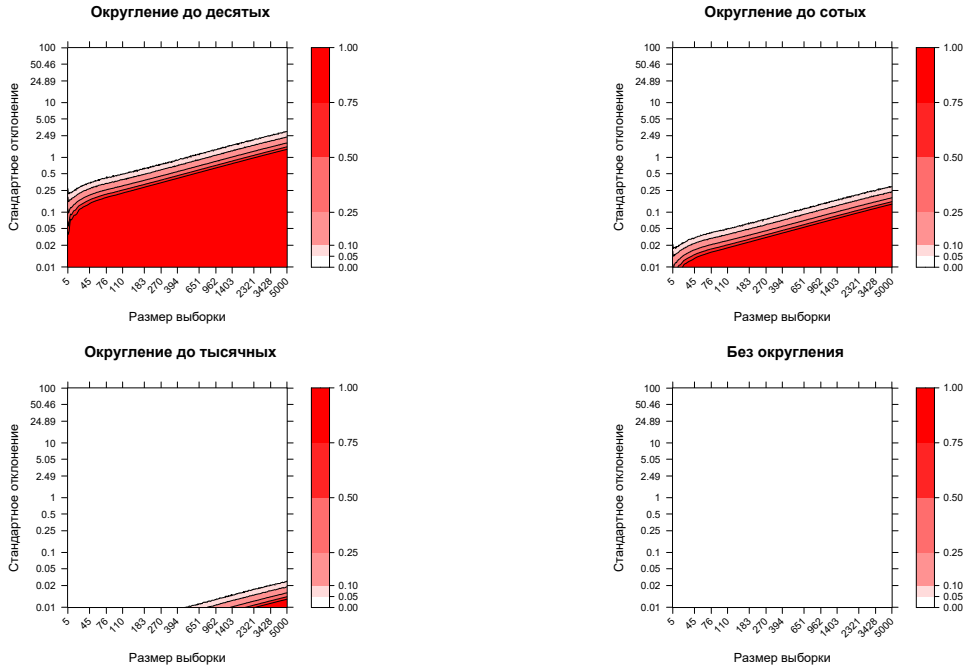

Рис. 22: Тепловые карты результатов моделирования для $f_0(x)$ и $f_5(x)$ – $f_7(x)$.

4 Исследование ошибки I рода двухэтапного тестирования

4.1 Обзор статей в медицинских журналах

В предыдущей главе 3 на примере влияния округления мы показали, что критерии нормальности необходимо использовать с осторожностью. Далее мы решили выяснить, как часто их применяют сейчас на практике. Поэтому мы провели обзор статей за 2021 год из двух ведущих российских журналов по кардиологии, входящих в списки «Web of Science» и «Scopus»:

1. «Рациональная фармакотерапия в кардиологии»;
2. «Российский кардиологический журнал».

Полученная статистика:

- 356 статей было проанализировано;
- 192 статьи содержали статистический анализ;
- 107 статей содержали предварительную проверку на нормальность;
- В 56 статьях предварительная проверка на нормальность использовалась для последующего выбора между критериями. Из них в 49 статьях выбор производился между критерием Стьюдента и U-критерием Манна – Уитни;

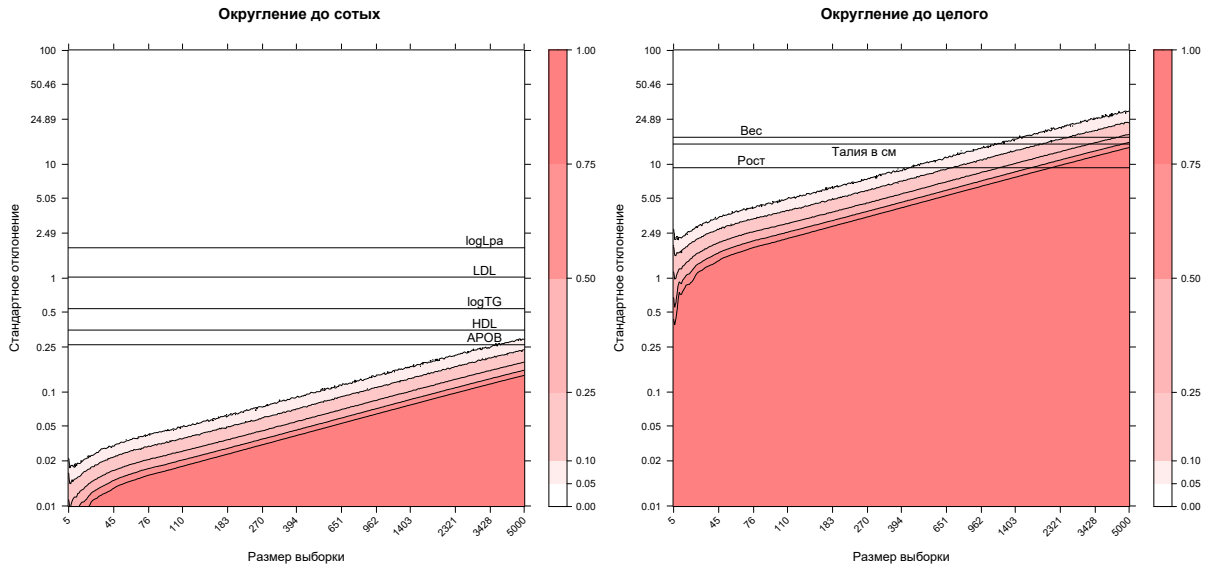

Рис. 23: Тепловые карты результатов моделирования для реальных данных.

- В 66 статьях предварительная проверка на нормальность использовалась или по неизвестным причинам или для выбора формата представления данных: среднее $\pm$ стандартное отклонение или медиана и интерквантильный размах.

Таким образом, зачастую критерии нормальности использовались для последующего выбора критерия. И чаще всего этот выбор был между критерием Стьюдента и критерием Манна – Уитни. Поэтому далее мы исследуем такую двухэтапную процедуру более подробно.

4.2 Схема двухэтапного тестирования

Рассмотрим ситуацию, когда необходимо проверить гипотезу о различии между двумя независимыми выборками. Например, когда необходимо сравнить средние μ_1 и μ_2 двух генеральных совокупностей, из которых извлекаются выборки:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \tag{61}$$

При использовании двухвыборочного критерия Стьюдента для проверки таких гипотез необходимо учитывать ограничение на начальные данные: выборки должны извлекаться из нормально распределенных генеральных совокупностей. Поэтому, прежде чем применять критерий Стьюдента, необходимо выяснить, соблюдается ли это условие, или, как минимум, провести предварительное тестирование на нормальность. Если условие нормальности не выполняется, необходимо проверять гипотезу о различии между выборками при помощи другого критерия, например, при помощи непараметрического U-критерия Манна – Уитни.

Пусть даны две независимые выборки $\mathcal{X} = (X_1, ..., X_n)$ и $\mathcal{Y} = (Y_1, ..., Y_n)$, извлеченные из двух генеральных совокупностей с равными дисперсиями σ^2 и со средними μ_1 и μ_2 соответственно. Исследуем поведение ошибки I рода следующей двухэтапной процедуры:

1. Проводим предварительное тестирование на нормальность при помощи критерия Шапиро – Уилка. Критерий нормальности Шапиро – Уилка был выбран, поскольку в разделе 1.8 мы показали, что этот тест является наиболее мощным среди рассмотренных;

2. Если обе выборки прошли предварительное тестирование на нормальность, используем двухвыборочный критерий Стьюдента для проверки гипотез 61;

3. Если хотя бы одна выборка не прошла предварительное тестирование на нормальность, используем непараметрический U-критерий Манна – Уитни.

Вероятность ошибки первого рода всей двухэтапной процедуры представляется следующим образом:

$$P(err_I) = P(err_I | SW \text{ не отверг гипотезу о нормальности}) \cdot P(\text{неотвержения } SW) + \\ + P(err_I | SW \text{ отверг гипотезу о нормальности}) \cdot P(\text{отвержения } SW) \quad (62)$$

где $P_{I,SW} = P(err_I | SW$ не отверг гипотезу о нормальности) – вероятность ошибки I рода критерия Стьюдента при условии неотвержения гипотезы о нормальности критерием Шапиро – Уилка; $P_{SW} = P(\text{неотвержения } SW)$ – вероятность неотвержения гипотезы о нормальности критерием Шапиро – Уилка; $P_{I,U} = P(err_I | SW$ отверг гипотезу о нормальности) – вероятность ошибки I рода U-критерия Манна – Уитни при условии отвержения гипотезы о нормальности критерием Шапиро – Уилка и $P_U = P(\text{отвержения } SW)$ – вероятность отвержения гипотезы о нормальности критерием Шапиро – Уилка.

4.3 Исследование ошибки I рода параметрического тестирования методом Монте-Карло

Уровень значимости для критерия Шапиро – Уилка фиксировали $\alpha_{pre} = 0.05$. Уровень значимости для критерия Стьюдента также фиксировали $\alpha = 0.05$. При помощи метода Монте-Карло мы построили график зависимости оценок для $P_{I,SW}$, P_{SW} и их произведения – первого члена разложения 62, от размера выборок. В качестве модельного распределения выбрали экспоненциальное распределение с параметром $\lambda = 1$. Размеры выборок брали от 10 до 30 с шагом 2. Алгоритм вычисления оценки для $P_{I,SW}$:

1. Генерируем пары выборок из распределения Exp(1) пока K = 1000 пар не пройдут предварительное тестирование на нормальность;

2. Применяем к каждой паре полученных выборок критерий Стьюдента. Если на i-ой итерации гипотеза о равенстве средних отвержена, то $m_i = 1$, иначе: $m_i = 0$;

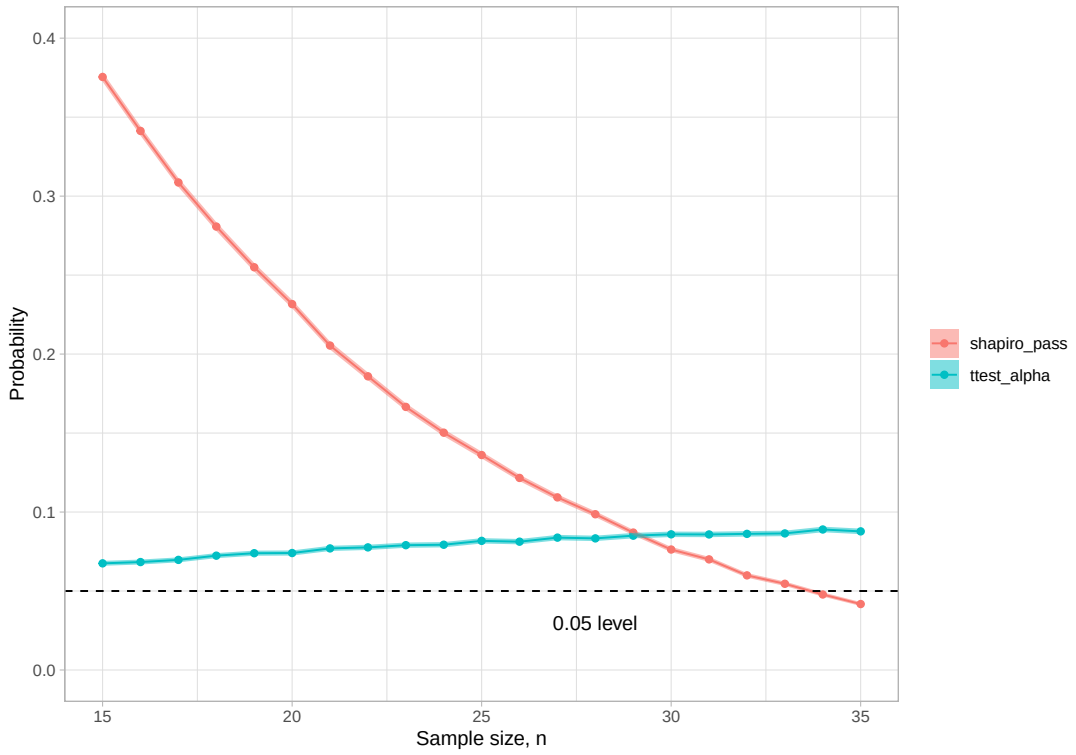

Рис. 24: Уловная ошибка первого рода для критерия Стьюдента (синий) и вероятность прохождения предварительного тестирования (красный) с соответствующими доверительными интервалами.

3. Получаем оценку $\widehat{P_{I,SW}} = \frac{\sum_{i=1}^{K} m_i}{K}$.

На рисунке 24 оценка условной вероятности $\widehat{P_{I,SW}}$ превышает заранее фиксированный уровень значимости 5 % на всей сетке размеров выборок. Эта оценка возрастает к 10%, и даже ее доверительный интервал не задевает уровня 0.05.

4.4 Исследование ошибки I рода непараметрического тестирования методом Монте-Карло

Уровни значимости для критерия Шапиро – Уилка и U-критерия Манна– Уитни аналогично разделу 4.3 фиксировали $\alpha_{pre} = 0.05$ и $\alpha = 0.05$. При помощи метода Монте-Карло на рисунке 25 мы построили график зависимости оценок для $P_{I,U}$, P_U и их произведения – второго члена разложения 62, от размера выборок. Модельное распределение и размеры выборок брали такие же, как и в разделе 4.3. Алгоритм вычисления оценки для $P_{I,U}$:

1. Генерируем пары выборок из распределения Exp(1) пока не получим K = 1000 пар, которые не прошли предварительное тестирование на нормальность;

2. Применяем к каждой паре полученных выборок U-критерий Манна – Уитни. Если на i-ой итерации гипотеза об отсутствии различий между выборками отвержена, то $m_i = 1$, иначе: $m_i = 0$;

3. Получаем оценку $\widehat{P_{I,U}} = \frac{\sum_{i=1}^{K} m_i}{K}$.

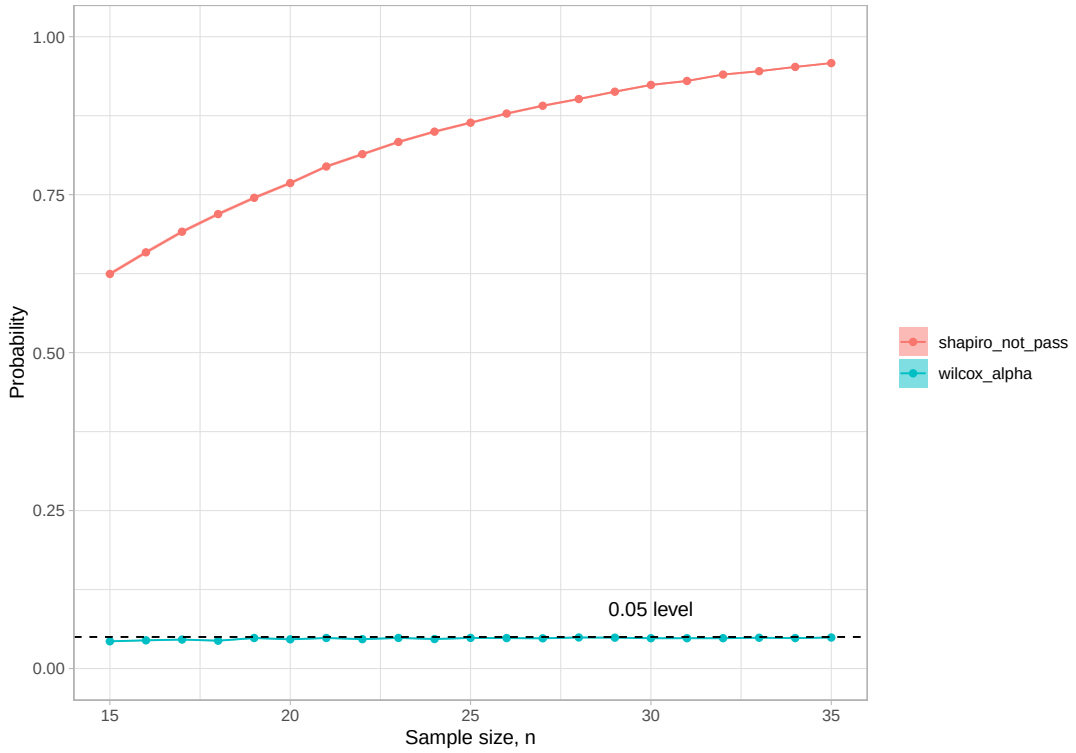

Рис. 25: Уловная ошибка первого рода для U-критерия Манна – Уитни (синий) и вероятность отвержения предварительной гипотезы о нормальности (красный) с соответствующими доверительными интервалами.

На графике на рисунке 25 и на увеличенном графике на рисунке 26 видно, что оценка $\widehat{P_{I,U}}$, в отличие от оценки $\widehat{P_{I,SW}}$, с ростом n возрастает, но контролируется на уровне 0.05.

4.5 Исследование ошибки I рода двухэтапной процедуры методом Монте-Карло

Используя результаты, полученные в разделах 4.3 и 4.4, мы построили итоговый график оценки ошибки I рода всей двухэтапной процедуры от размера выборок. Результаты представлены на рисунке 27. Видим, что итоговая оценка ошибки I рода слабо не контролируется, и даже ее доверительный интервал не задевает заранее фикрсированного уровня значимости 0.05. То есть, с теоретической точки зрения, такая процедура кажется не очень корректной. Более того, в работе [38] показано, что критерий Манна – Уитни не является критерием для проверки гипотез о центральной тенденции. То есть, несмотря на то, какие гипотезы мы фиксировали в самом начале двухэтапной процедуры, мы не знаем, что именно тестирует критерий Манна – Уитни. А как мы показали в разделе 4.1, такую процедуру до сих пор используют и достаточно активно.

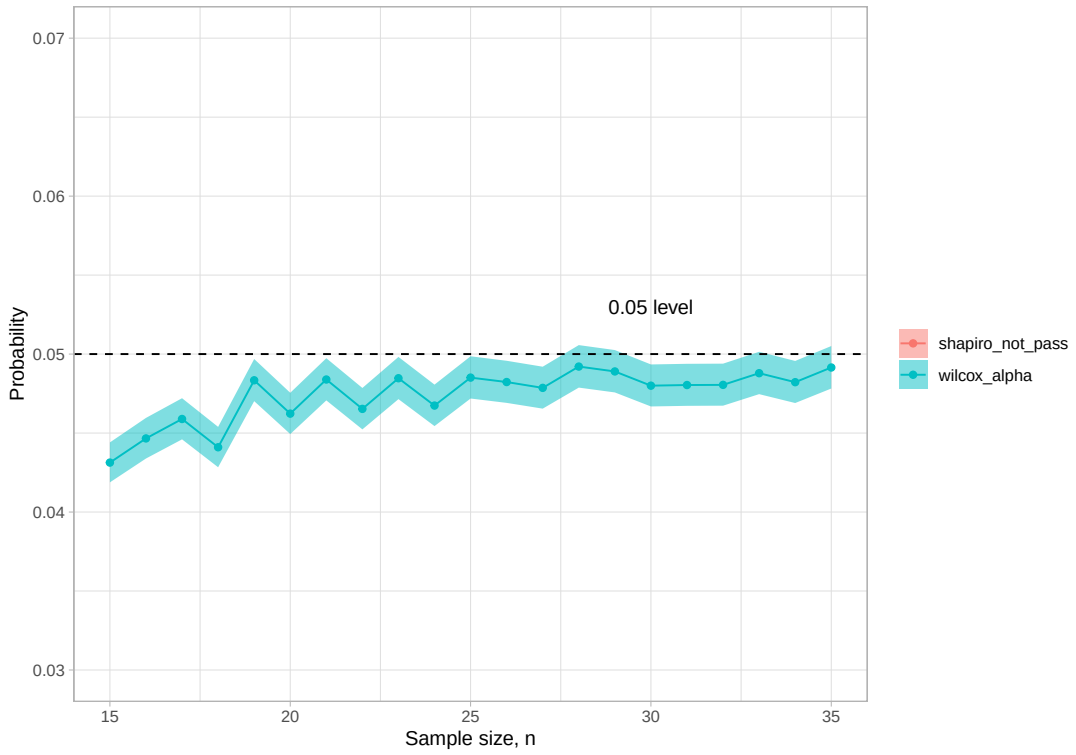

Рис. 26: Уловная ошибка первого рода для U-критерия Манна – Уитни (синий) и вероятность отвержения предварительной гипотезы о нормальности (красный) с соответствующими доверительными интервалами в увеличенном масштабе.

5 Ряды Эджворта

Опираясь на материал, изложенный в [5], [31], [32] и [33], приведем теоретическое описание рядов Эджворта.

Теорема 2 (Центральная предельная) *Пусть $X_1, \ldots, X_n$ – последовательность независимых одинаково распределенных случайных величин с математическим ожиданием $EX_1 = a$ и дисперсией $DX_n = \sigma^2$. Если $0 < \sigma^2 < \infty$, то*

$$\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - a}{\sigma} \xrightarrow[n\to\infty]{d} N(0,1)$$

Определение 1 (Семиинвариант) *Если $E|X|^n < \infty$, то в некоторой окрестности точки $t = 0$ логарифм характеристичекой функции случайной величины X $ln\phi_X(t)$ (ветвь логарифма, для которого $ln\phi_X(0) = 0$) непрерывно дифференцируема до порядка n включительно. Величина*

$$\kappa_n = (-i)^n \frac{d^n}{dt^n} ln\phi_X(t)|_{t=0}$$

называется семиинвариантом порядка k.

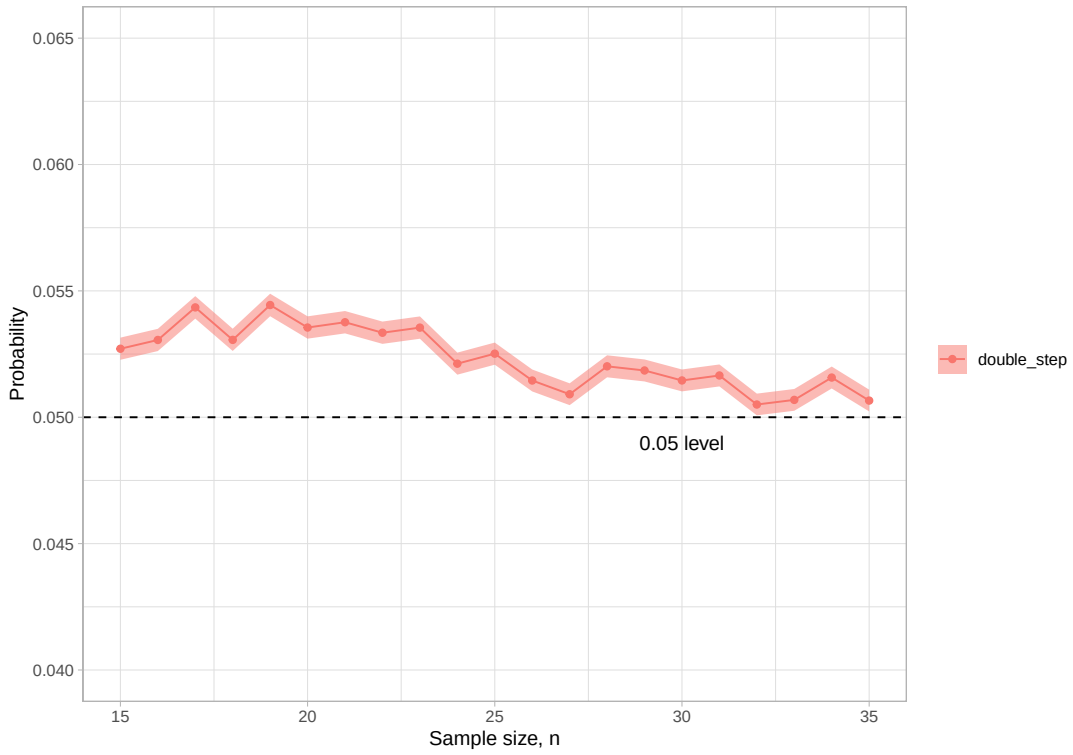

Рис. 27: Оценка вероятности ошибки I рода двухэтапной процедуры и её доверительный интервал в зависимости от размера выборки.

Определение 2 (Полиномы Эрмита)

$$H_n(z) = \frac{(-1)^n}{\varphi(z)} \frac{d^n}{dz^n} \varphi(z)$$

где $\varphi(z)$ – плотность стандартного нормального распределения.

Пусть $X_1, \ldots, X_n$ – независимые одинаково распределенные случайные величины, имеющие конечное математическое ожидание μ 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.
Пусть $\phi_Z(t) = E \exp(itZ)$ – характеристическая функция Z, $\phi(t) = E \exp(itX_1)$ – характеристическая функция X_1. Используем свойства характеристических функций:

$$\phi_Z(t) = \left(\phi \left(\frac{t}{\sqrt{n}\sigma} \right) \right)^n \exp \left(-\frac{\sqrt{n} i t \mu}{\sigma} \right) \tag{63}$$

Пусть семиинварианты любого порядка существуют. Тогда справедливо разложение:

$$\phi_Z(t) = \exp \left(\sum_{k=1}^{\infty} \frac{\kappa_k (it)^n}{k!} \right) \tag{64}$$

Используя 63, 64, а также разложение в ряд Тейлора для $\ln\phi\left(\frac{t}{\sqrt{n}\sigma}\right)$ в окрестности точки $t=0$, получаем:

$$\ln\phi_Z(t) = n\ln\phi\left(\frac{t}{\sqrt{n}\sigma}\right) - \frac{\sqrt{n}it\mu}{\sigma} = n\sum_{i=2}^{\infty}\left(\frac{it}{\sqrt{n}\sigma}\right)^i\frac{\kappa_i(X_1)}{i!} = \\ = \frac{t^2}{2} + \frac{(it)^3\kappa_3(X_1)}{6\sqrt{n}\sigma^3} + \frac{(it)^4\kappa_4(X_1)}{24n\sigma^4} + O\left(n^{-3/2}\right) \quad (65)$$

Далее используем формулу обращения для преобразования Фурье и 65, получаем формулу для плотности Z:

$$f_Z(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-itz)\phi_Z(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-itz)\exp(\ln\phi_Z(t))dt = \\ = \frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-itz)\exp\left(\frac{t^2}{2} + \frac{(it)^3\kappa_3(X_1)}{6\sqrt{n}\sigma^3} + \frac{(it)^4\kappa_4(X_1)}{24n\sigma^4} + O\left(n^{-3/2}\right)\right)dt = \\ = \frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-itz)\exp\left(\frac{t^2}{2}\right)\left(1 + \frac{(it)^3\rho_3(X_1)}{6\sqrt{n}} + \frac{(it)^4\rho_4(X_1)}{24n} + \frac{(it)^6\rho_3^2(X_1)}{72n} + O\left(n^{-3/2}\right)\right)dt \quad (66)$$

где $\rho_i(X_1) = \frac{\kappa_i(X_1)}{\sigma^i}$ – нормированный семиинвариант. Используем свойство производных:

$$\frac{d^r(exp(-itz))}{dz^r} = (-1)^r\exp(-itz)(it)^r \quad (67)$$

а также формулу обращения для плотности стандартного нормального распределения $\varphi(z)$:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-itz)\exp\left(-\frac{t^2}{2}\right) = \varphi(z) \quad (68)$$

Меняем порядок дифференцирования и интегрирования [31], получаем:

$$f_Z(z) = \varphi(z)\left(1 + \frac{\rho_3(X_1)}{6\sqrt{n}}H_3(z) + \frac{\rho_4(X_1)}{24n}H_4(z) + \frac{\rho_3^2(X_1)}{72n}H_6(z) + O(n^{-3/2})\right) \quad (69)$$

Раскрываем полиномы Эрмита:

$$f_Z(z) = \varphi(z)\left(1 + \frac{\rho_3(X_1)}{6\sqrt{n}}(z^3 - 3z) + \frac{\rho_4(X_1)}{24n}(z^4 - 6z^2 + 3)\right) + \\ + \varphi(z)\left(\frac{\rho_3^2(X_1)}{72n}(z^6 - 15z^4 + 45z^2 - 15) + O(n^{-3/2})\right) \quad (70)$$

Интегрируем 70, получаем выражение для функции распределения Z:

$$F_Z(x) = P(Z \le x) = \Phi(x) + \varphi(x)\left(\frac{\rho_3(X_1)}{6\sqrt{n}}(1 - x^2) + \frac{\rho_4(X_1)}{24n}(3x - x^3)\right) + \\ + \varphi(x)\left(+\frac{\rho_3^2(X_1)}{72n}(-x^5 + 10x^3 - 15x) + O(n^{-3/2})\right) \quad (71)$$

Где $\Phi(x)$ – функция стандартного нормального распределения. По определению семиинвариантов:

$$\rho_3(X_1) = \frac{E(X_1 - EX_1)^3}{\sigma^3} = \gamma_1, \tag{72}$$

$$\rho_4(X_1) = \frac{E(X_1 - EX_1)^4 - 3E(X_1 - EX_1)^2}{\sigma^4} = E(X_1 - EX_1)^4\sigma^4 - 3 = \gamma_2 \tag{73}$$

где γ_1 – коэффициент асимметрии X_1, γ_2 – коэффициент эксцесса X_1. Тогда формула для плотности распределения $f_Z(z)$ случайной величины Z:

$$f_Z(z) = \varphi(z)\left(1 + \frac{\gamma_1}{6\sqrt{n}}(z^3 - 3z) + \frac{\gamma_2}{24n}(z^4 - 6z^2 + 3)\right) + \\ + \varphi(z)\left(\frac{\gamma_1^2}{72n}(z^6 - 15z^4 + 45z^2 - 15) + O(n^{-3/2})\right) \tag{74}$$

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

$$F_Z(x) = \Phi(x) + \\ \varphi(x)\left(\frac{\gamma_1}{6\sqrt{n}}(1 - x^2) + \frac{\gamma_2}{24n}(3x - x^3) + \frac{\gamma_1^2}{72n}(-x^5 + 10x^3 - 15x) + O(n^{-3/2})\right) \tag{75}$$

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: $\{10, 12, \ldots, 100\} \bigcup \{100, 120, \ldots, 1000\} \bigcup \{1500, 2000, \ldots, 5000\} \bigcup \{6000, 7000, \ldots, 10^4\}$;
- асимметрии: $\{0.1, 0.5, 1, 2, 3.5, 5, 7.5, 10\}$.

Для каждой выборки вероятность $1 - F_Z(q)$ была оценена на 10^4 итераций метода Монте-Карло. Определенные выше значения асимметрии были выбраны не случайно и брались, опираясь на значения, оцененные по выборке исследования [30]:

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

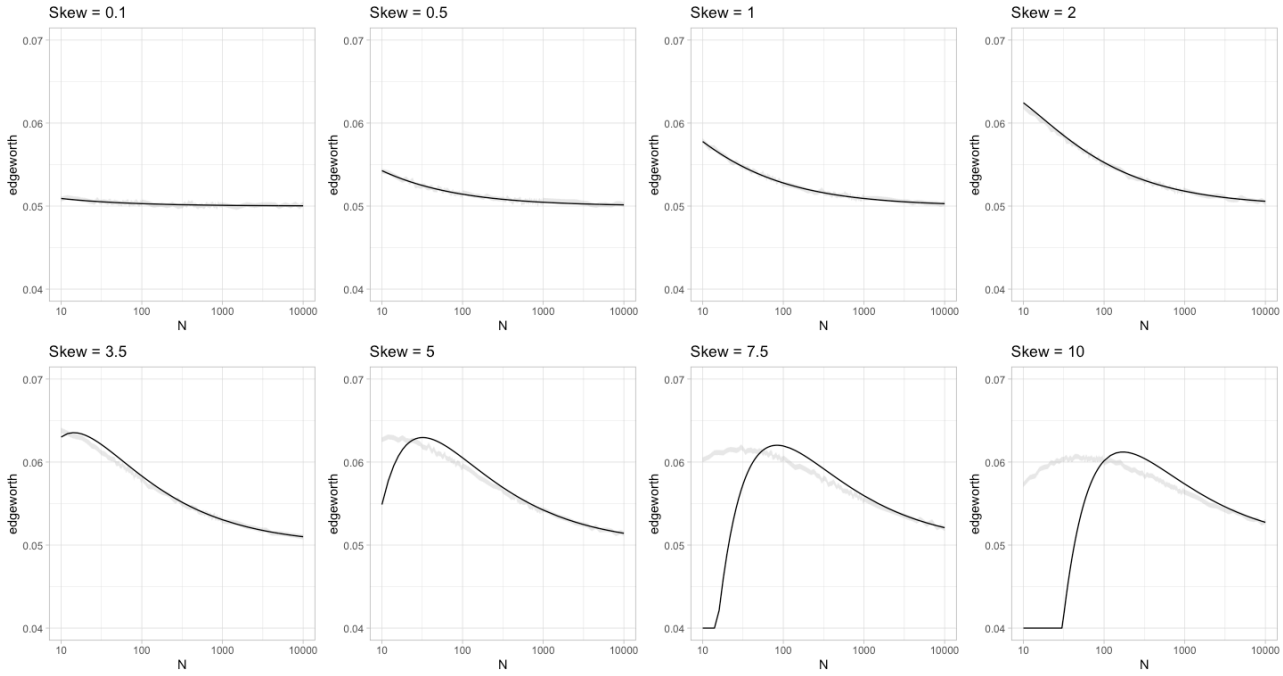

Рис. 28: Сравнение вероятностей, оцененных по выборке для различных значений асимметрии: для эмпирической функции распределения приведен 95 %-ный доверительный интервал оценки вероятности (серым), для разложения Эджворта приведена оценка вероятности (черным).

Результаты моделирования представлены на рисунке 28.
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 29 и 30.

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.

6 Критерий Стьюдента

6.1 Одновыборочный критерий Стьюдента

Опираясь на материал, изложенный в [4], приведем теоретическое описание одновыборочного критерия Стьюдента. Пусть дана выборка $\mathcal{X} = (X_1, ..., X_n)$, для которой известно, что она была извлечена из нормально распределенной генеральной

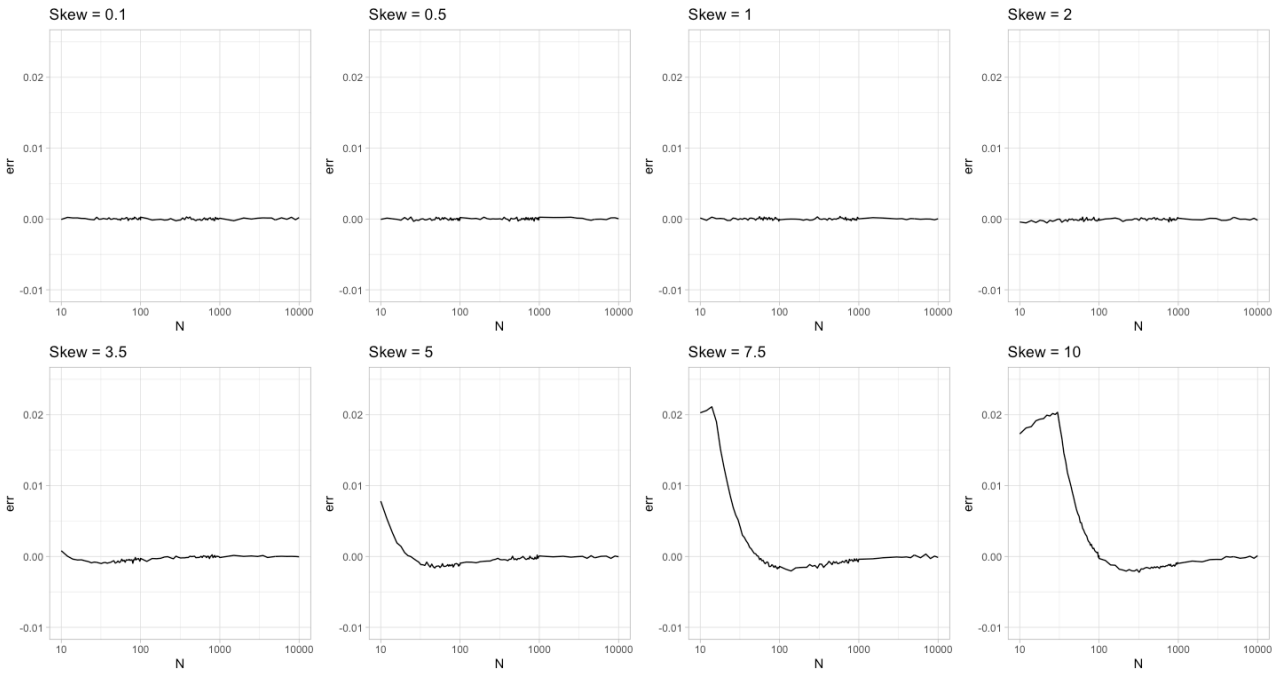

Рис. 29: График абсолютной ошибки.

совокупности со средним μ и дисперсией σ^2: $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$. Альтернативой этому условию может быть условие на выборочное среднее, оно должно быть извлечено из нормально распределенной генеральной совокупности со средним μ и дисперсией $\frac{\sigma^2}{n}$: $\bar{\mathcal{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$. Фиксируем уровень значимости α. Рассмотрим гипотезы о средних с двусторонней критической областью вида:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \tag{76}$$

Критериальная статистика имеет вид:

$$t = \sqrt{n} \cdot \frac{\bar{\mathcal{X}} - \mu}{s} \tag{77}$$

где $\bar{\mathcal{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$ и $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{\mathcal{X}})^2$ – выборочные среднее и дисперсия. Если H_0 верна, критериальная статистика имеет распределение Стьюдента с $(n-1)$ степенями свободы: $t_{H_0} = \sqrt{n} \cdot \frac{\bar{\mathcal{X}} - \mu_0}{s} \sim_{H_0} t(n-1)$.

Исследуем ситуацию, когда условие нормальности распределения входных данных нарушается. Рассмотрим логнормальное распределение $L(\mu, \sigma^2)$ с параметрами $\mu = 0$ и $\sigma = 0.1$. Математическое ожидание такого распределения выражается следующей формулой: $mean = \exp\left(\mu + \frac{\sigma^2}{2}\right) = \exp(0.005)$. Заметим, что коэффициент асимметрии такого распределения γ_1 сравнительно небольшой:

$$\gamma_1 = (\exp(0.01) + 2)\sqrt{\exp(0.01) - 1} \approx 0.3$$

Фиксируем уровень значимости $\alpha = 0.05$. $K = 10^4$ раз извлекаем выборку $\mathcal{X} =$

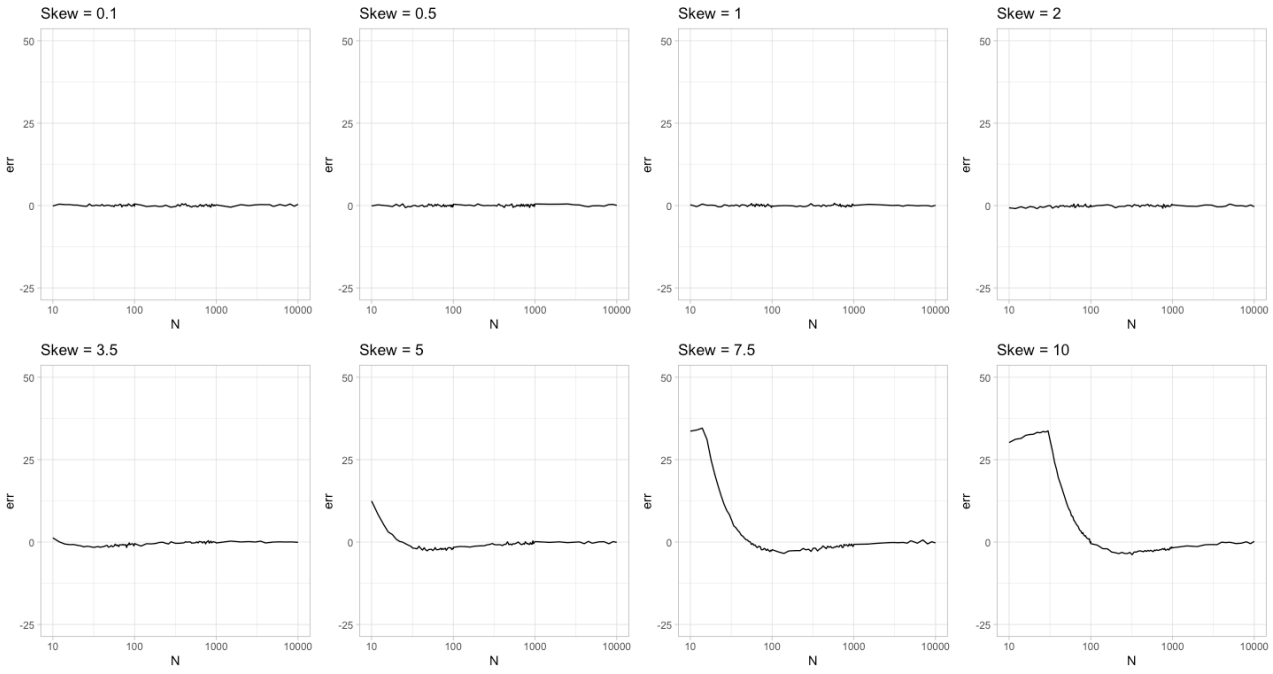

Рис. 30: График относительной ошибки (в процентах).

$(X_1, ..., X_{100})$ из рассмотренного нами логнормального распределения $L(0, 0.01)$ и применяем критерий Стьюдента с гипотезами:

$$\begin{cases} H_0 : mean = \exp(0.005) \\ H_1 : mean \neq \exp(0.005) \end{cases} \tag{78}$$

Из значений статистики критерия Стьюдента получаем выборку $\mathcal{T} = (T_1, ..., T_{10^4})$. На рисунке 31 представлен график ядерной оценки плотности распределения $\mathcal{T}$. Видим, что распределение достаточно симметрично.

Теперь рассмотрим логнормальное распределение $L(\mu, \sigma^2)$ с параметрами $\mu = 0$ и $\sigma = 1$. Математическое ожидание распределения: $mean = \exp\left(\mu + \frac{\sigma^2}{2}\right) = \exp(0.5)$. Коэффициент асимметрии такого распределения γ_1 принимает уже гораздо большие значения:

$$\gamma_1 = (\exp(1) + 2)\sqrt{\exp(1) - 1} \approx 6.2$$

Фиксируем уровень значимости $\alpha = 0.05$. $K = 10^4$ раз извлекаем выборку $\mathcal{X} = (X_1, ..., X_{100})$ из нашего нового логнормального распределения $L(0, 1)$ и применяем критерий Стьюдента с гипотезами:

$$\begin{cases} H_0 : mean = \exp(0.5) \\ H_1 : mean \neq \exp(0.5) \end{cases} \tag{79}$$

Из значений статистики получаем выборку $\mathcal{T} = (T_1, ..., T_{10^4})$. На рисунке 32 аналогично представлен график ядерной оценки плотности распределения $\mathcal{T}$. На этот раз распределение имеет тяжелый левый хвост и совсем не похоже на распредление Стьюдента.

Методом Монте-Карло получим оценки вероятностей ошибки I рода $\hat{\alpha}$:

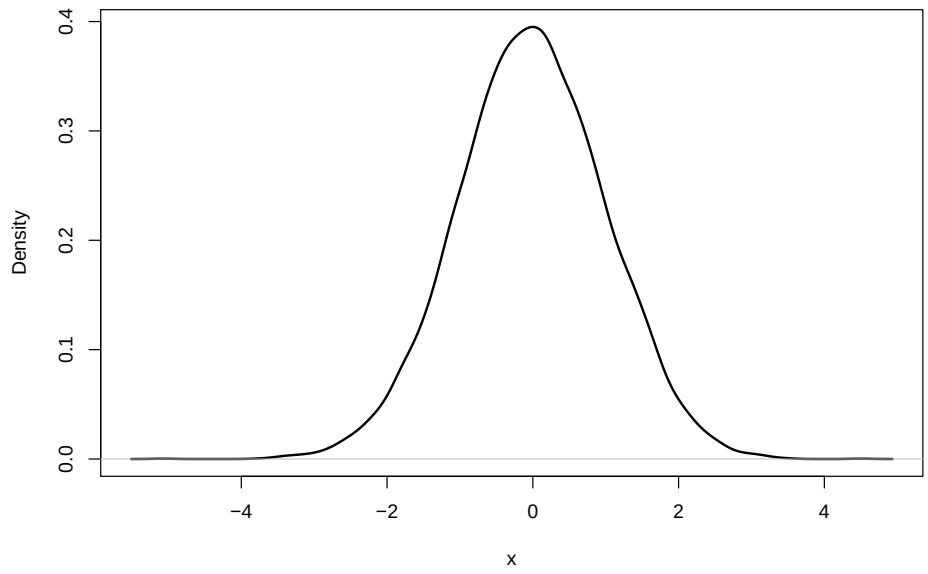

Рис. 31: График ядерной оценки плотности статистики критерия Стьюдента для случая выборки из распределения $LN(0, 0.01)$.

- для $L(0, 0.1)$ $\hat{\alpha}_{0.1} \approx 0.0495$;
- а для $L(0, 1)$ $\hat{\alpha}_1 \approx 0.0858$, что превышает фиксированный ранее уровень значимости $\alpha = 0.05$!

Таким образом, мы показали, как распределение с большой асимметрией ломает критерий Стьюдента.

Исследуем подробнее влияние асимметрии на статистику одновыборочного критерия Стьюдента. Рассмотим сетку для параметра логнормального распределения $\sigma \in \{0.1, 1, 3, 5, 7, 10\}$, влияющего на коэффициент асимметрии. Для каждого значения параметра будем генерировать $K = 1000$ раз выборку $\mathcal{X} = (X_1, \ldots, X_n)$ размера $n = 100$ из распредления $LN(0, \sigma^2)$ и вычислять:

1. выборочное среднее: $\bar{\mathcal{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$;
2. отношение выборочного среднего с выборочным стандартным отклонением: $\frac{\bar{\mathcal{X}}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{\mathcal{X}})^2}}$;
3. отношение среднего генеральной совокупности с выборочным стандартным отклонением, взятого с обратным знаком: $\frac{-mean}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{\mathcal{X}})^2}}$.

Обозначим выборки, полученные в пунктах 1.– 3., как $\mathcal{M}^{1,\sigma} = \left(M_1^{1,\sigma}, \ldots, M_K^{1,\sigma}\right)$, $\mathcal{M}^{2,\sigma} = \left(M_1^{2,\sigma}, \ldots, M_K^{2,\sigma}\right)$ и $\mathcal{M}^{3,\sigma} = \left(M_1^{3,\sigma}, \ldots, M_K^{3,\sigma}\right)$. На рисунке 33 построены гистограммы распределений выборочных средних $\mathcal{M}^{1,\sigma}$. Пунктирная линия соответствует значению среднего генеральной совокупности $mean = \exp\left(\frac{\sigma^2}{2}\right)$, а значение

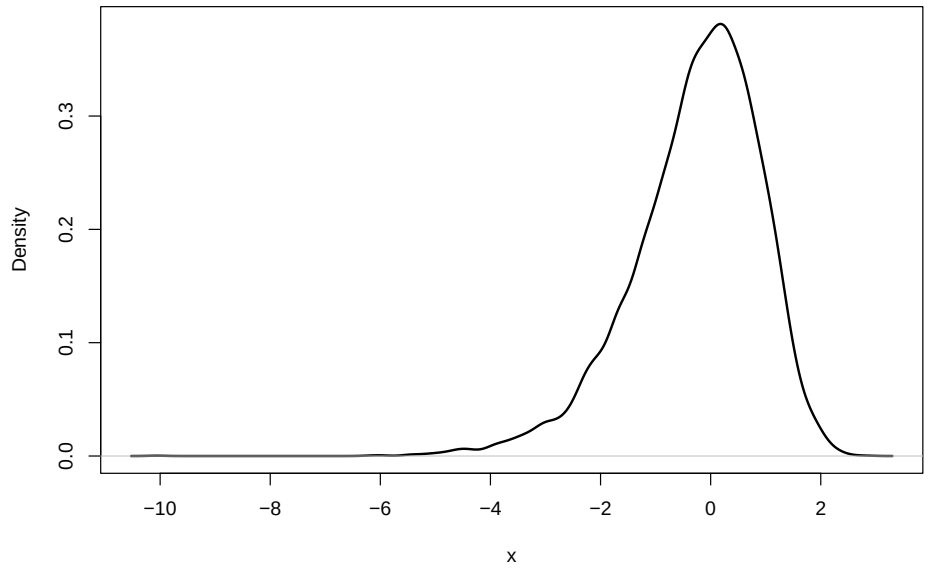

Рис. 32: График ядерной оценки плотности статистики критерия Стьюдента для случая выборки из распределения $LN(0,1)$.

$per = \frac{|\{M_k > mean, k=1,\ldots,K\}|}{K} \cdot 100\%$ – процент элементов выборки выборочных средних, превышающих значение среднего генеральной совокупности. Видим, что с увеличением значения параметра σ и, следовательно, коэффицикнта асимметрии логнормального распределения, распределение выборочного среднего также становится все более асимметричным. Распределение начинает съезжает влево от реального среднего, и при большой асимметрии выборочное среднее никогда не превысит среднего генеральной совокупности. Таким образом мы получим отрицательные значения статистики.

Для отношения выборочного среднего с выборочным стандартным отклонением $\mathcal{M}^{2,\sigma}$ мы построим графики ядерных оценок плотности. На рисунке 34 мы видим, что с увеличением параметра σ и, следовательно, коэффициента асимметрии логнормального распределения, распределение нашей статистики также становится асимметричным, появляется тяжелый правый хвост.

Аналогично мы поступим с нашей третьей статистикой, отношением среднего генеральной совокупности с выборочным стандартным отклонением, взятого с обратным знаком $\mathcal{M}^{3,\sigma}$ – построим графики ядерных оценок плотностей. На рисунке 35 мы вновь видим у распределения тяжелый хвост, но на этот раз слева.

Сложив наши статистики из пунктов 2. и 3., рассмотрим графики ядерных оценок плотности для итоговой статистики критерия Стьюдента. На рисунке 36, как и в нашем изначальном примере на рисунке 32, мы видим тяжелый левый хвост распределения. Возможным объяснением может быть то, что отношением выборочного среднего с выборочным стандартным отклонением (статистика из пункта 2.) становится асимметричным позже и не настолько сильно, как отношением среднего генеральной совокупности с выборочным стандартным отклонением (статистика из пункта 3.), из-за чего отрицательные значения в выборке "побеждают".

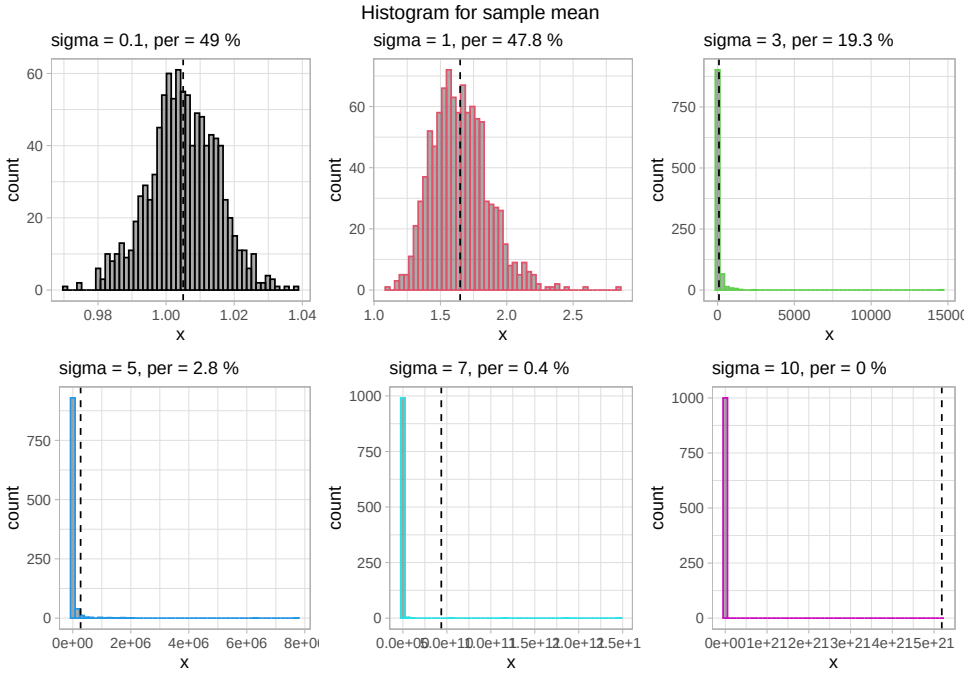

Рис. 33: Гистограммы распределения выборочных средних $\mathcal{M}^{1,\sigma}$ для разных значений параметра σ логнорального распределения $LN(0, \sigma^2)$.

Таким образом, на работоспособность одновыборочного критерия Стьюдента сильно влияет асимметрия. Этот факт согласуется с тем, что коэффициент асимметрии является первым членом разложения в ряду Эджворта для случайной величины, схожей со статистикой критерия Стьюдента. Для малых значений асимметрии все будет хорошо, критерий Стьюдента все еще можно применять. Но для больших — нет. И как найти эту грань — очень хороший вопрос, на который мы, надеюсь, попробуем найти ответ.

6.2 Двухвыборочный критерий Стьюдента для независимых выборок

Опираясь на материал, изложенный в [4], приведем теоретическое описание двухвыборочного криетрия Стьюдента для независимых выборок. Пусть даны две независимые выборки: $\mathcal{X} = (X_1, ..., X_{n_1})$, извлеченная из нормального распределения со средним μ_1 и дисперсией σ^2: $X_i \sim N(\mu_1, \sigma^2)$, $i = 1, \ldots, n_1$; и $\mathcal{Y} = (Y_1, ..., Y_{n_2})$, извлеченная из нормального распределения со средним μ_2 и дисперсией σ^2: $Y_i \sim N(\mu_2, \sigma^2)$, $i = 1, \ldots, n_2$. Заметим, что у распределений совпадают дисперсии. Фиксируем уровень значимости α. Рассмотрим гипотезы с двусторонней критической областью:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \text{ (или } \mu_1 - \mu_2 = 0) \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \tag{80}$$

Статистика критерия Стьюдента в таком случае вычисляется по следующей форму-

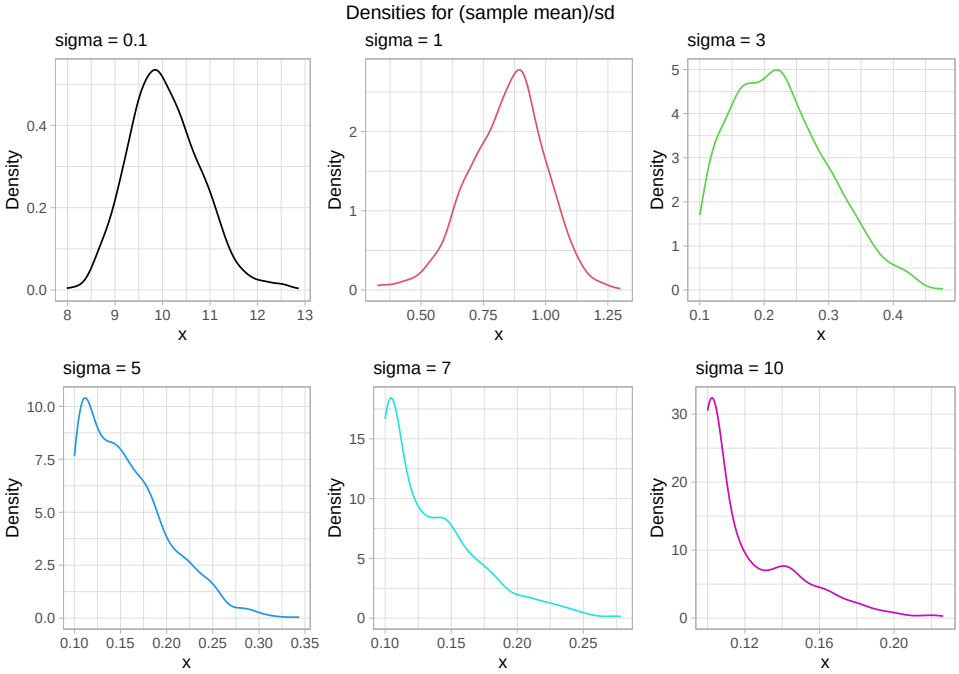

Рис. 34: Ядерные оценки плотности распределения $\mathcal{M}^{2,\sigma}$ – отношения выборочного среднего с выборочным стандартным отклонением, для разных значений параметра σ логнорального распределения $LN(0, \sigma^2)$.

ле:

$$t = \frac{\bar{\mathcal{X}} - \bar{\mathcal{Y}}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{81}$$

где $\bar{\mathcal{X}} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ и $\bar{\mathcal{Y}} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ – выборочные средние, $s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \bar{\mathcal{X}})^2$ и $s_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \bar{\mathcal{Y}})^2$ – выборочные дисперсии, а $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ – выборочная дисперсия для разности $\mathcal{X} - \mathcal{Y}$. Если H_0 верна, критериальная статистика имеет распределение Стьюдента с (n_1+n_2-2) степенями свободы: $t \sim_{H_0} t(n_1+n_2-2)$.

Опираясь на разложение Эджворта, рассмотрим коэффициенты асимметрии γ_1 и эксцесса γ_2 для разности независимых одинаково распределенных случайных величин ξ и η:

$$\gamma_1(\xi - \eta) = \frac{E(\xi - \eta - E(\xi - \eta))^3}{\left(\sqrt{D(\xi - \eta)}\right)^3} = \frac{E\xi^3 - 3E\xi^2 E\eta + 3E\xi E\eta^2 - E\eta^3}{\left(\sqrt{D\xi + D\eta}\right)^3} = \tag{82}$$

$$= \frac{E\xi^3 - 3E\xi^2 E\xi + 3E\xi E\xi^2 - E\xi^3}{\left(\sqrt{2D\xi}\right)^3} = 0 \tag{83}$$

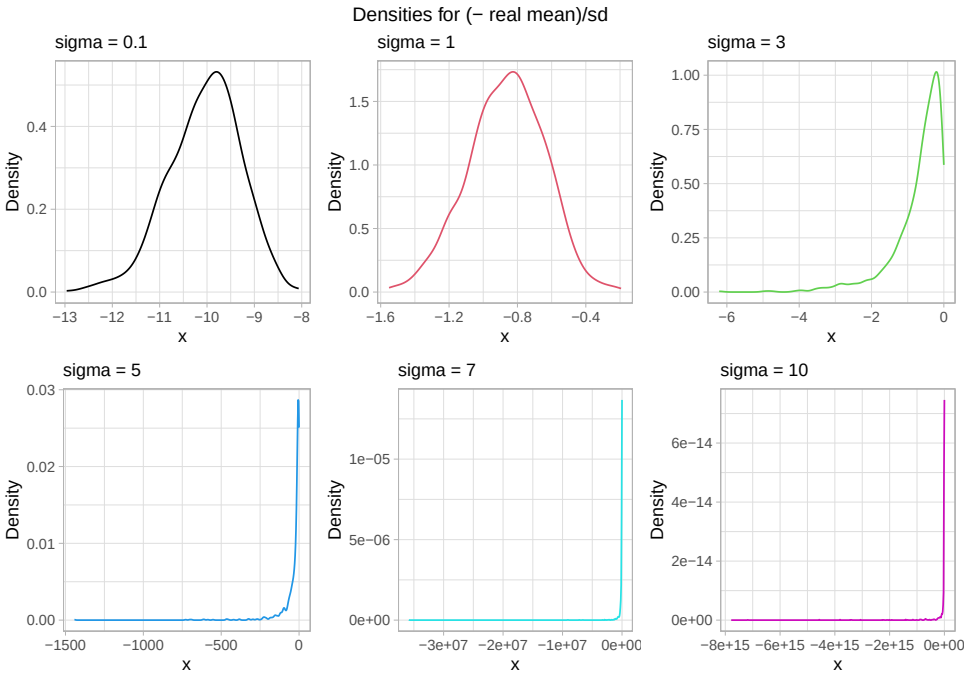

Рис. 35: Ядерные оценки плотности распределения $\mathcal{M}^{3,\sigma}$ – отношения среднего генеральной совокупности с выборочным стандартным отклонением, взятого с обратным знаком, для разных значений параметра σ логнорального распределения $LN(0, \sigma^2)$.

$$\gamma_2(\xi - \eta) = \frac{E(\xi - \eta - E(\xi - \eta))^4}{\left(\sqrt{D(\xi - \eta)}\right)^4} - 3 = \tag{84}$$

$$= \frac{E\xi^4 - 4E\xi^3 E\eta + 6E\xi^2 E\eta^2 - 4E\xi E\eta^3 + E\eta^4}{(D\xi + D\eta)^2} - 3 = \tag{85}$$

$$= \frac{2E\xi^4 - 8E\xi^3 E\xi + 6\left(E\xi^2\right)^2}{4\left(D\xi\right)^2} - 3 \neq 0 \tag{86}$$

Аналогично предыдущему случаю для одновыборочного критерия Стьюдента, рассмотрим ядерную оценку плотности статистики двувыборочного критерия Стьюдента для случая логнормальных распределений $LN(\mu_i, \sigma_i^2)$, $i = \overline{1,4}$, где $\mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$ и $\sigma_1 = 1$, $\sigma_2 = 3$, $\sigma_3 = 10$, $\sigma_4 = 100$.
Для каждого логнормального распределения $LN(\mu_i, \sigma_i^2)$, $i = \overline{1,4}$ повторяем следующую процедуру. Фиксируем уровень значимости $\alpha = 0.05$. $K = 10^4$ раз извлекаем выборки $\mathcal{X} = (X_1, ..., X_{100})$ и $\mathcal{Y} = (Y_1, ..., Y_{100})$ из соответствующего логнормального распределения и применяем критерий Стьюдента с гипотезами:

$$\begin{cases} H_0 : mean(\mathcal{X}) = mean(\mathcal{Y}) \\ H_1 : mean(\mathcal{X}) \neq mean(\mathcal{Y}) \end{cases} \tag{87}$$

Из значений статистики получаем выборку $\mathcal{T} = (T_1, ..., T_{10^4})$. На рисунке 37 представлен график ядерных оценок плотности распределения $\mathcal{T}$. Видим, что с увеличением

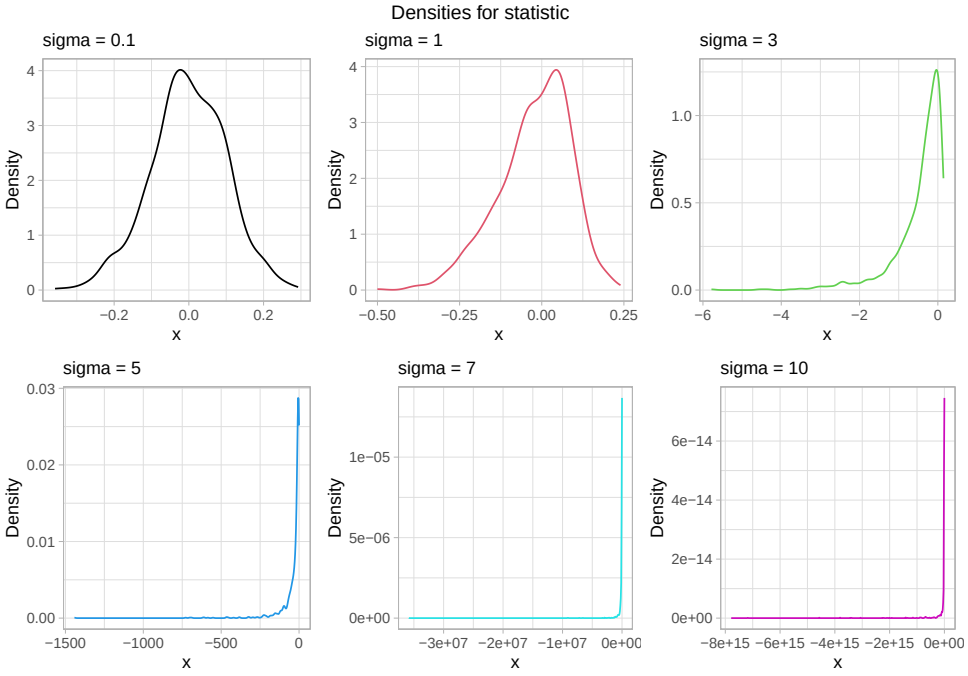

Рис. 36: Ядерные оценки плотности распределения статистики одновыборочного критерия Стьюдента для разных значений параметра σ логнорального распределения $LN(0, \sigma^2)$.

параметра σ получаются разные результаты: при малых значениях параметра распределение статистики еще похоже на распределение Стьюдента, но с увеличением значения параметра оно смещается к значениям ± 1.

В случае одновыборочного критерия Стьюдента проблема была в асимметрии распределения. В текущем же случае асимметрия разности двух распределений сокращается в ноль, чего нельзя сказать об эксцессе. Коэффициент эксцесса логнормального распределения выражается следующей формулой:

$$\gamma_2 = \exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 6 \tag{88}$$

и, соответственно, при увеличении параметра σ, коэффициента эксцесса также возрастает. Согласно [37], большие значения эксцесса говорят о тяжелых хвостах, то есть в выборке будут редко попадаться очень большие значения, которые будут находиться в числителе и знаменателе статистики критерия Стьюдента и давать итоговое ее значение $+1$ или -1.

С точки зрения ошибки I рода все будет в порядке: она не превысит заранее фиксированного уровня значимости α, но будет уменьшаться с увеличением σ и, следовательно, эксцесса. Оценки вероятностей ошибок I рода для наших распределений $LN(0, \sigma_i^2)$, $i = \overline{1,4}$, полученные методом Монте-Карло: $\hat{\alpha_1} = 0.047$, $\hat{\alpha_2} = 0.016$, $\hat{\alpha_3} = 0.0008$, $\hat{\alpha_4} = 0$.

Поскольку ранее у нас был фиксирован параметр $\mu = 0$, убедимся, что его изменение не повлияет на результаты. Аналогично, методом Монте-Карло, будем вычислять оценки ошибки I рода для двухвыборочного критерия Стьюдента для выборок из $LN(\mu_i, \sigma_j^2)$, $i = \overline{1,5}$, $j = \overline{1,6}$, где $\mu \in \{-100, -50, 0, 50, 100\}$, а $\sigma \in \{0.1, 1.08, 2.06, 3.04, 4.02, 5\}$. Результаты представлены на рисунке 38.

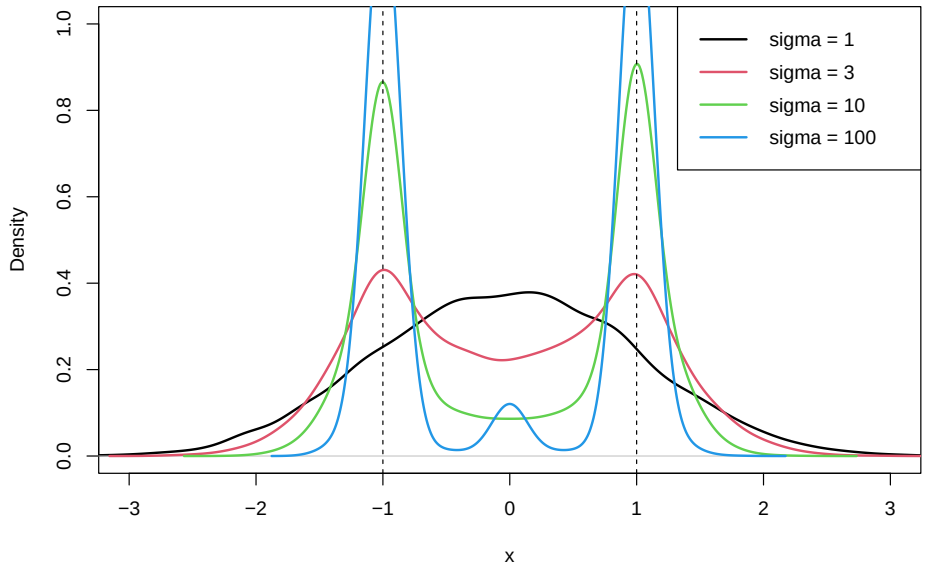

Рис. 37: График ядерных оценок плотности статистики двухвыборочного критерия Стьюдента для случая выборок из логнормальных распределений $LN(0, \sigma^2)$.

С мощностью критерия Стьюдента в такой постановке уже будут проблемы!

6.3 Эмпирическое исследование мощности двухвыборочного критерий Стьюдента для независимых выборок

Исследуем поведение мощности двухвыборочного критерия Стьюдента, когда нарушается предположение о нормальности распределения входных данных. Рассмотрим сетку $\sigma \in \{1, 2, 3, 4, 5, 6\}$ для параметра логнормального распределения, а также сетку вероятностей $p \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Будем рассматривать пары выборок $(\mathcal{X}, \mathcal{Y})$, где первая выборка извлечена из смеси нормального распределения $N(0.5, 1)$ со средним $\delta = 0.5$ и дисперсией 1 и логнормального распределения $LN(0, \sigma^2)$, центрированного относительно $\delta = 0.5$. Описанное распределение можно представить следующем образом:

$$\xi \sim (1 - p) \cdot N(0.5, 1) + p \cdot \left(LN(0, \sigma^2) - \exp\left(\frac{\sigma^2}{2}\right) + 0.5 \right) \tag{89}$$

Вторая выборка будет извлечена из смеси стандартного нормального распределения $N(0, 1)$ и логнормального распределения $LN(0, \sigma^2)$, центрированного относительно нуля:

$$\eta \sim (1 - p) \cdot N(0, 1) + p \cdot \left(LN(0, \sigma^2) - \exp\left(0 + \frac{\sigma^2}{2}\right) \right) \tag{90}$$

Для всех значений сетки по p и для всех значений сетки по σ методом Монте-Карло при $K = 1000$ будем вычислять оценки мощности критерия Стьюдента и U-критерия Манна – Уитни:

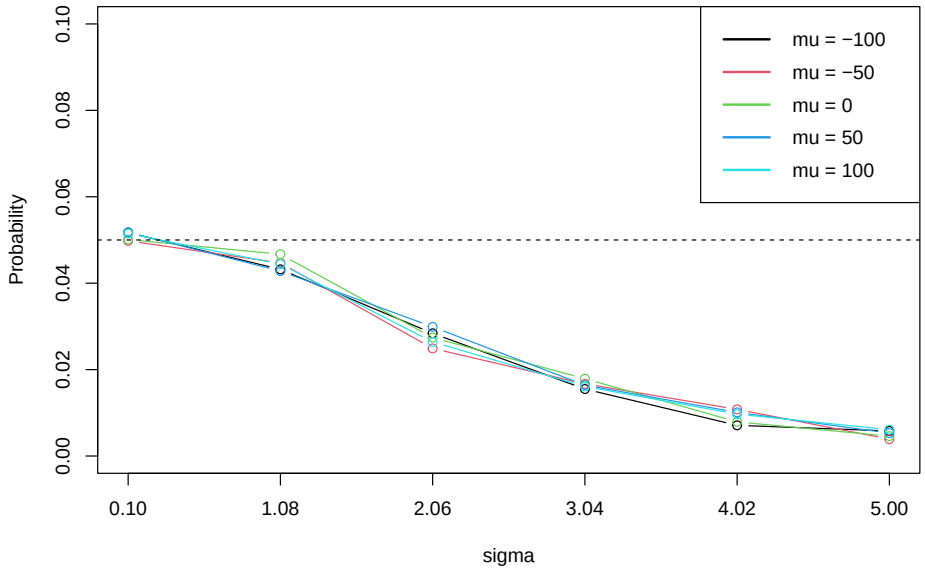

Рис. 38: Оценки ошибок I рода двухвыборочного критерия Стьюдента для логнормальных распределений $LN(\mu, \sigma^2)$.

1. Для $k = 1, \ldots, K$ генерируем пары выборок $(\mathcal{X}_k, \mathcal{Y}_k)$, применяем к каждой паре двухвыборочный критерий Стьюдента для независимых выборок. Если на k-ой итерации гипотеза о равенстве средних отвержена, то $st_k = 1$, иначе: $st_k = 0$;

2. Вновь для $k = 1, \ldots, K$ генерируем пары выборок $(\mathcal{X}_k, \mathcal{Y}_k)$, применяя к каждой паре U-критерий Манна – Уитни. Аналогично, если гипотеза о равенстве средних на k-ой итерации отвержена, то $w_k = 1$, иначе: $w_k = 0$;

3. Получаем оценки мощности критерия Стьюдента $P^p_{st}(\sigma) = \frac{1}{K}\sum_{k=1}^{K} st_k$ и U-критерия Манна – Уитни $P^p_w(\sigma) = \frac{1}{K}\sum_{k=1}^{K} w_k$.

Результаты моделирования представлены на рисунке 39. В случае, когда выборки были извлечены из нормально распределенной генеральной совокупности, мощности критериев равны 1. Когда в смесь к нормальному распределению добавляется логнормальное, критерий Стьюдента перестает справляться, и его мощность с увеличением σ, а, следовательно, и эксцесса падает до нуля. Мощность U-критерия Манна – Уитни также падает, но не до нулевых значений, он все еще неплохо справляется.

6.4 Бутстрэп

Одной из фундаментальных проблем в статистике является оценка изменчивости статистических оценок, полученных на основе выборочных данных. Не всегда у исследователей есть возможность повторять свои эксперименты снова и снова для определения погрешности измерений – зачастую это непрактично и дорого.

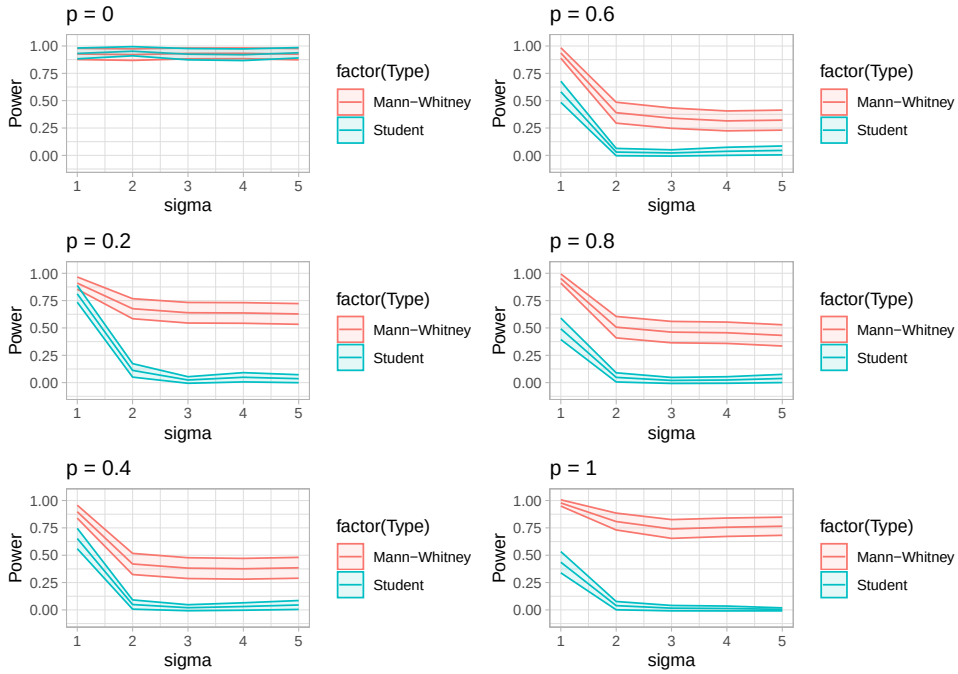

Рис. 39: Оценки мощности двухвыборочного критерия Стьюдента (синим) и U-критерия Манна – Уитни (красным) для выборок из смеси нормального и логнормального распределений.

До широкого распространения больших вычислительных мощностей эту проблему решали при помощи математических расчетов распределения выборки. Это легко сделать для простых показателей, таких как доля выборки, но не так просто для более сложных статистик.

Большие вычислительные мощности открыли новые возможности для решения проблемы определения выборочного распределения. В 1977 году Брэдли Эфрон ввел понятие бутстрэпа – практической компьютерной методики для исследования распределения статистик вероятностных распределений. Бутстрэп основан на многократной генерации выборок на базе имеющейся выборки, и позволяет просто и быстро оценивать самые разные статистики для сложных моделей: доверительные интервалы, дисперсию, корреляцию и так далее. В [36] показана состоятельность бутстрэп-оценки выборочного распределения для многих распространенных видов статистик. Также в [36] приведена теорема, согласно которой оценка методом бутстрэпа является более точной, чем аппроксимация нормальным распределением для стандартизированной случайной величины (Сила метода бутстрэпа заключается в том, что его можно применить почти к любой статистике, независимо от того, насколько она сложна.)

Чтобы проиллюстрировать идею и силу бутстрэпа, мы приведем два примера из [35], где оцениваются выборочное среднее и медиана для набора данных о годовом доходе $n = 25$ взрослых мужчин (в тысячах долларов), собранных в вымышленном округе в Северной Каролине: $\mathcal{X} = (1, 4, 6, 12, 13, 14, 18, 19, 20, 22, 23, 24, 26, 31, 34, 37, 46, 47, 56, 61, 63, 65, 70, 97, 385)$. Выборка $\mathcal{X}$ генерировалась следующим образом: $X_i = 30 \exp(Z_i)$ ($\times\$1000$), где Z_i – независимые одинаково распределенные случайные величины из стандартного нормального распределения $N(0, 1)$, $i = 1, \ldots, 25$.

6.4.1 Бутстрэп-оценка выборочного среднего

Выборочное среднее рассматриваемой выборки: $\bar{\mathcal{X}} = \frac{1}{25}\sum_{i=1}^{25} X_i = 47.76$. Поскольку мы знаем истинное распределение популяции, из которой была извлечена выборка $\mathcal{X}$, мы можем сгенерировать еще 1000 таких выборок и, на их основе, – оценок выборочного среднего, оценив тем самым распределение $\bar{\mathcal{X}}$. На рисунке 40 слева представлена полученная гистограмма распределения выборочного среднего $\bar{\mathcal{X}}$.
В случае, когда распределение популяции не известно, для оценки распределения $\bar{\mathcal{X}}$ мы можем воспользоваться методом бутстрэпа. Выборку $\mathcal{X}$ мы представляем как псевдопопуляцию и на ее основе генерируем повторные выборки: извлекаем случайным образом из $\mathcal{X}$ 25 элементов с возвращением. На рисунке 40 справа представлена гистограмма распределения выборочного среднего $\bar{\mathcal{X}}$, полученная на основе 1000 таких повторных выборок. Эта гистограмма является бутстрэп оценкой распределения, представленного на рисунке 40 слева. Оба графика действительно похожи, хоть и имеют отличия из-за того, что бутстрэп-оценка имеет информацию лишь об одной выборке и не знает всей популяции.

6.4.2 Бутстрэп-оценка выборочной медианы

Если построить гистограмму распределения выборки $\mathcal{X}$, то можно увидеть, что оно имеет положительную асимметрию. Это также подтверждается фактом, что среднее значение выборки, равное $47,8$, намного больше, чем медиана выборки, равная 26. В таких ситуациях для измерения центральной тенденции чаще отдают предпочтение именно медиане.
Как и в случае со средним значением из предыдущего подраздела, на рисунке 41 слева представлена гистограмма распределения выборочной медианы, полученная на основе 1000 выборок из известного нам распределения генеральной совокупности. В правой части рисунка 41 представлена гистограмма распределения выборочной медианы, полученная методом бутстрэпа из 1000 повторных выборок из $\mathcal{X}$. Видим, что масштаб по оси Y у левого и правого графиков отличается. Оценить распределение медианы по выборке оказывается сложнее, чем оценить распределение среднего значения, но бутстрэп по-прежнему оценивает это распределение достаточно хорошо.

6.5 Одновыборочный критерий Стьюдента с поправкой на асимметрию

Опираясь на материал, изложенный в [34], приведем теоретическое описание одновыборочного критерия Стьюдента с поправкой на асимметрию. Пусть дана выборка $\mathcal{X} = (X_1, \ldots, X_n)$ со средним μ. Фиксируем уровень значимости α, формулируем гипотезы:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \tag{91}$$

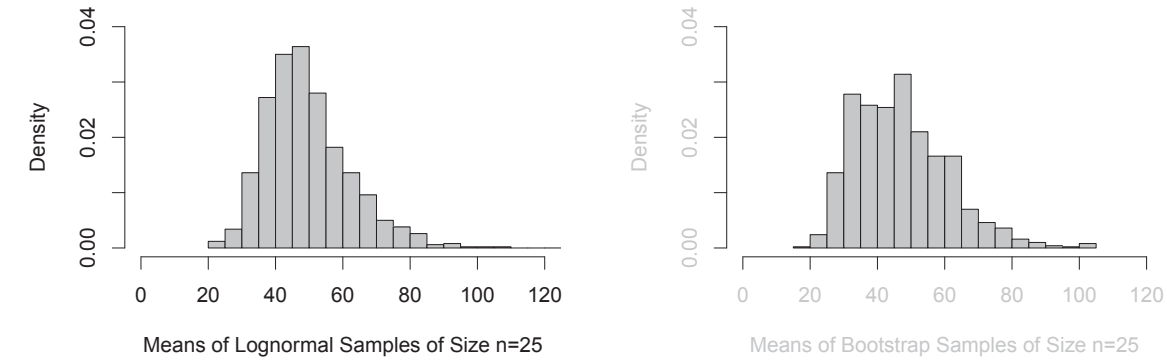

Рис. 40: (Слева) Гистограмма из 1000 выборочных средних, полученных на основе повторных выборок из теоретической логнормальной совокупности. (Справа) Гистограмма 1000 средних значений бутстрэп выборок на основе X.

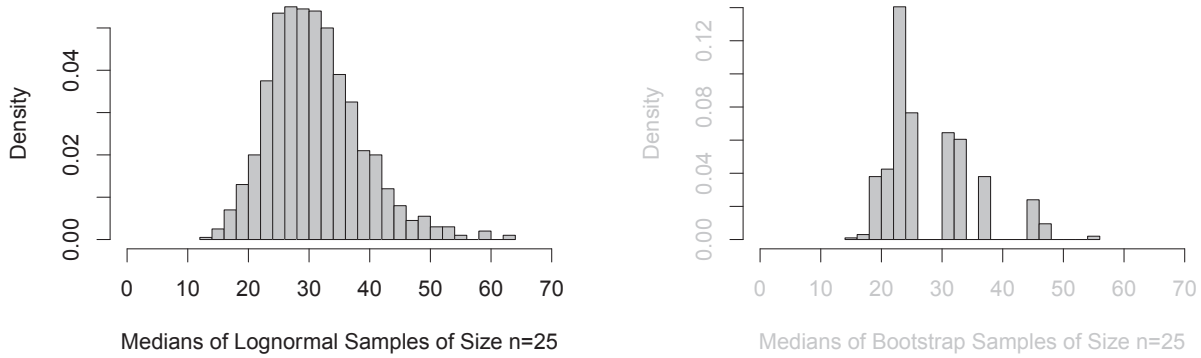

Рис. 41: (Слева) Гистограмма из 1000 выборочных медиан, полученных на основе повторных выборок из теоретической логнормальной совокупности. (Справа) Гистограмма 1000 медиан бутстрэп выборок на основе X.

Статистика модифицированного критерия Стьюдента:

$$t_{sa} = t + \frac{\gamma_1}{6n\sqrt{n}}\left(2t^2 + 1\right) \tag{92}$$

где, как и ранее, $t = \sqrt{n}\frac{\bar{\mathcal{X}}-\mu}{s}$ – статистика классического одновыборочного критерия Стьюдента, $\bar{\mathcal{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{\mathcal{X}})^2$ и $\gamma_1 = \frac{1}{s^3}\sum_{i=1}^{n}(X_i - \bar{\mathcal{X}})^3$ – выборочные среднее, дисперсия и коэффициент асимметрия для $\mathcal{X}$.
Методом бутстрэпа генерируем B псевдовыборок $\mathcal{X}_b = \left(X_1^b, \ldots, X_{n_B}^b\right), b = 1, \ldots, B$ размера n_B из исходной выборки $\mathcal{X}$ с возвращением. В [34] было использовано значение $n_b = \frac{n}{4}$, и такой выбор был основан на эмпирическом анализе. Для каждой выборки $\mathcal{X}_b$ вычисляем значение статистики:

$$t_{sa}^b = t_b + \frac{\gamma_1^b}{6n_b\sqrt{n_b}}\left(2t_b^2 + 1\right) \tag{93}$$

где $t_b = \sqrt{n_B}\frac{\bar{\mathcal{X}}_b - \bar{\mathcal{X}}}{s_b}$, – бутстрэп аналог статистики t, $\bar{\mathcal{X}}_b = \frac{1}{n_B}\sum_{k=1}^{n_B} X_k^b$, $s_b^2 = \frac{1}{(n_B-1)}\sum_{k=1}^{n_B}(X_k^b - \bar{\mathcal{X}}_b)^2$, $\gamma_1^b = \frac{1}{s_b^3}\sum_{k=1}^{n_B}(X_k^b - \bar{\mathcal{X}}_b)^3$ – соответственно, выборочные среднее,

дисперсия и асимметрия для $\mathcal{X}_b$.
Из B полученных значений t^b_{sa} вычисляем критические значения x^*_l и x^*_u для исходной статистики t_{sa}, решая следующие уравнения:

$$P(t^b_{sa} \leq x^*_l) = P(t^b_{sa} \geq x^*_u) = \frac{\alpha}{2},\ b = 1, \ldots, B \tag{94}$$

Если $t_{sa} < x^*_l$ или $t_{sa} > x^*_u$, то мы отвергаем гипотезу H_0 о равенстве средних.

6.6 Эмпирическое исследование ошибки I рода одновыборочного критерия Стьюдента

Исследуем подробнее влияние коэффициента асимметрии на ошибку I рода одновыборочного критерия Стьюдента и сравним полученные результаты со случаем одновыборочного критерия Стьюдента с поправкой на асимметрию.
Рассмотрим сетку размеров выборок $n \in \{10, 40, 70, 100, 130, 160, 190, 220, 250, 280, 310\}$ и сетку коэффициентов асимметрии $\gamma_1 \in \{0.1, 1, 4, 7, 10, 13, 16, 19\}$ для логнормальных распределений $LN(0, \sigma^2_i)$, где σ^2_i, $i = \overline{1,8}$ соответсвуют разным значениям параметра γ_1.
Фиксируем уровень значимости $\alpha = 0.05$. Для каждого размера выборки n_i и для каждого значения коэффициента асимметрии $\gamma_{1,j}$, $i, j = \overline{1,8}$ будем повторять следующую процедуру:

1. Генерируем $K = 10^400$ выборок $\mathcal{X}^{i,j}_k$, $k = \overline{1,K}$ размера n_i из распределения $LN(0, \sigma^2_j)$;

2. При помощи одновыборочного критерия Стьюдента проверяем гипотезы о средних:
$$\begin{cases} H_0 : mean(\mathcal{X}^{i,j}_k) = \exp(\frac{\sigma^2_j}{2}) \\ H_1 : mean(\mathcal{X}^{i,j}_k) \neq \exp(\frac{\sigma^2_j}{2}) \end{cases} \tag{95}$$
Заметим, что формально гипотеза H_0 верна;

3. Если на k шаге гипотеза H_0 отвергается, $a_k = 1$;

4. Получаем оценку вероятности ошибки I рода одновыборочного критерия Стьюдента методом Монте-Карло: $\hat{\alpha_{st}} = \frac{1}{K}\sum_{k=1}^{K} a_k$.

На рисунке 42 цветом отображено значение оценки вероятности ошибки I рода одновыборочного критерия Стьюдента. Видим, что при значении коэффициента асимметрии логнормального распределения $\gamma_1 = 1$ вероятность ошибки I рода уже превышает уровень значимости 0.05. При еще больших значениях коэффициента асимметрии оценки вероятности ошибки I рода возрастают и могут достигать значений от 0.1 до 1.
Повторим процедуру 1. – 4. для одновыборочного критерия Стьюдента с поправкой на асимметрию, изменив количество выборок $K = 10^4$. Количество выборок для бустинга в одновыборочном критерии Стьюдента с поправкой на асимметрию $N = 10^4$. Результаты представлены на рисунке 43. Заметим, что красная область на рисунке отсутствует, оценки вероятности ошибки I рода для всей сетки параметров не превышают 0.075.

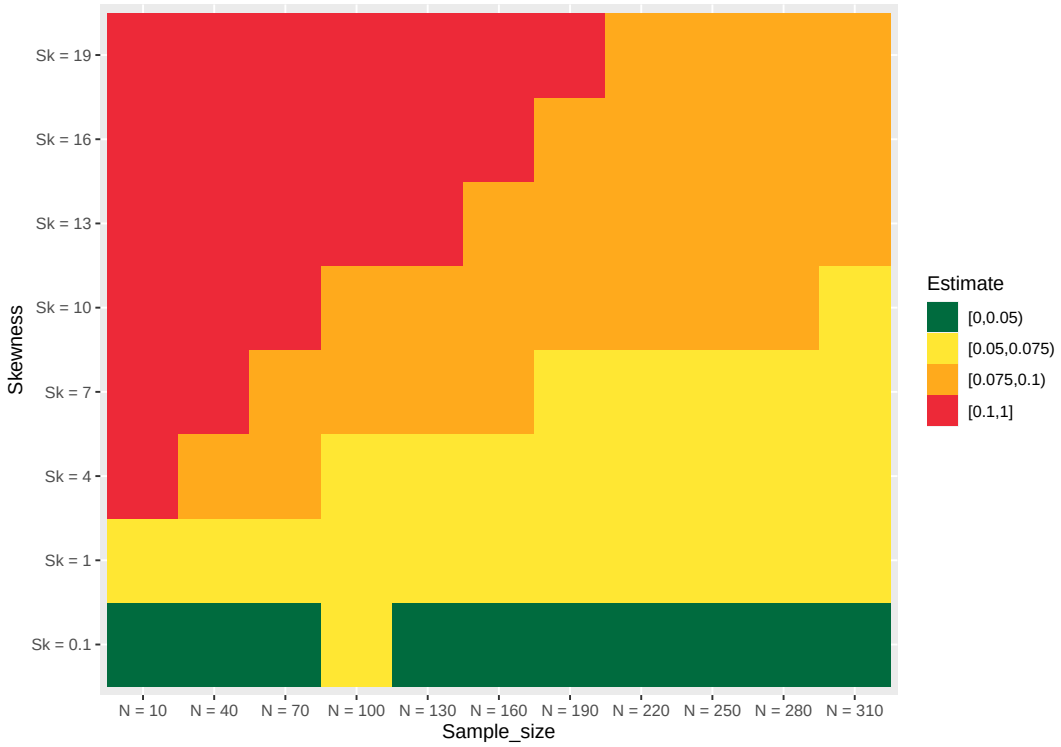

Рис. 42: Оценка вероятности ошибки I рода одновыборочного критерия Стьюдента в зависимости от размера выборки и коэффициента асимметрии для выборок из логнормального распределения.

6.7 Эмпирическое исследование мощности одновыборочного критерия Стьюдента

Исследуем теперь подробнее влияние коэффициента асимметрии на мощность одновыборочного критерия Стьюдента и сравним полученные результаты со случаем одновыборочного критерия Стьюдента с поправкой на асимметрию.
Рассмотрим такую же как и в предыдущем разделе сетку размеров выборок $n \in \{10, 40, 70, 100, 130, 160, 190, 220, 250, 280, 310\}$ и сетку коэффициентов асимметрии $\gamma_1 \in \{0.1, 1, 4, 7, 10, 13, 16, 19\}$ для логнормальных распределений $LN(0, \sigma_i^2)$, где σ_i^2, $i = \overline{1,8}$ соответствуют разным значениям параметра γ_1.
Фиксируем уровень значимости $\alpha = 0.05$. Для каждого размера выборки n_i и для каждого значения коэффициента асимметрии $\gamma_{1,j}$, $i, j = \overline{1,8}$ будем повторять следующую процедуру:

1. Генерируем $K = 10^6$ выборок $\mathcal{X}_k^{i,j}$, $k = \overline{1,K}$ размера n_i из распределения $LN(0, \sigma_j^2)$;

2. При помощи одновыборочного критерия Стьюдента проверяем гипотезы о средних:
$$\begin{cases} H_0 : mean(\mathcal{X}_k^{i,j}) = (\exp(\frac{\sigma_j^2}{2}) - \frac{1}{2}) \\ H_1 : mean(\mathcal{X}_k^{i,j}) \neq (\exp(\frac{\sigma_j^2}{2}) - \frac{1}{2}) \end{cases} \tag{96}$$
Заметим, что формально гипотеза H_0 не верна;

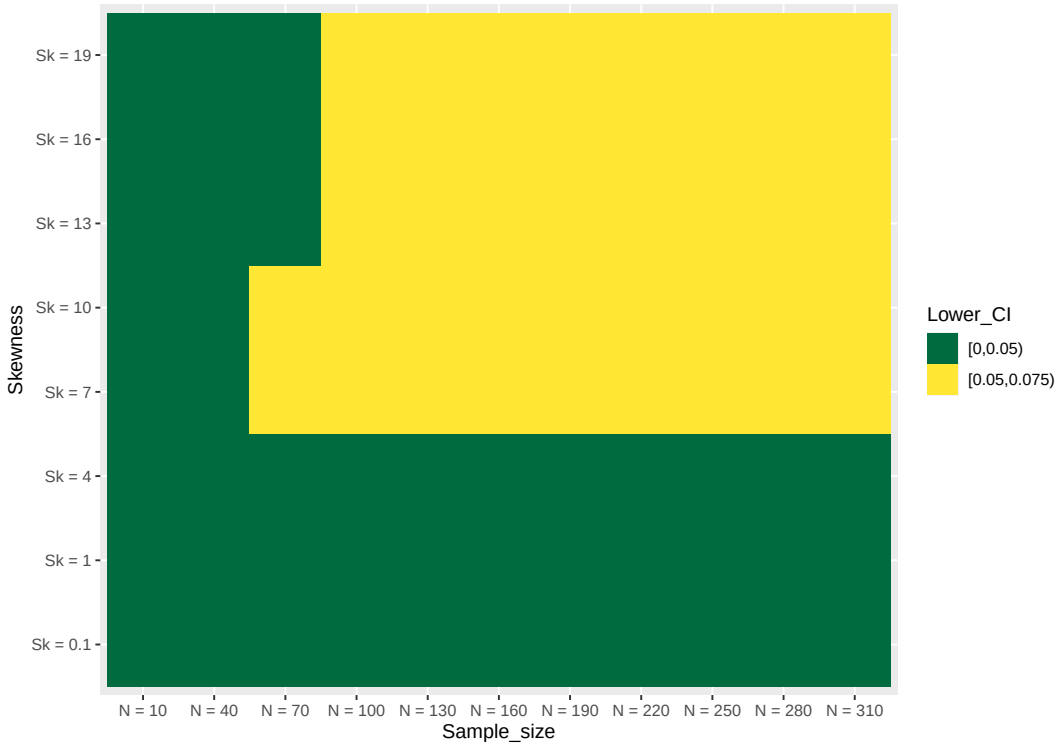

Рис. 43: Оценка вероятности ошибки I рода одновыборочного критерия Стьюдента с поправкой на асимметрию в зависимости от размера выборки и коэффициента асимметрии для выборок из логнормального распределения.

3. Если на k шаге гипотеза H_0 отвергается, $p_k = 1$;

4. Получаем оценку мощности одновыборочного критерия Стьюдента методом Монте-Карло: $\hat{P_{st}} = \frac{1}{K}\sum_{k=1}^{K} p_k$.

На рисунке 44 цветом отображено значение оценки мощности одновыборочного критерия Стьюдента: в белой зоне оценка близка к 0, а в зеленой, наоборот, – к 1.

Повторим процедуру 1. – 4. для одновыборочного критерия Стьюдента с поправкой на асимметрию. Количество выборок для бустинга в одновыборочном критерии Стьюдента с поправкой на асимметрию $N = 10^4$. Результаты представлены на рисунке 45: видим, что площадь зеленой зоны, где оценка мощности $\in (0.9, 1]$, увеличилась, по сравнению с предыдущим случаем для классического критерия Стьюдента. То есть показатели мощности у модифицированного критерия в нашем исследовании оказались лучше.

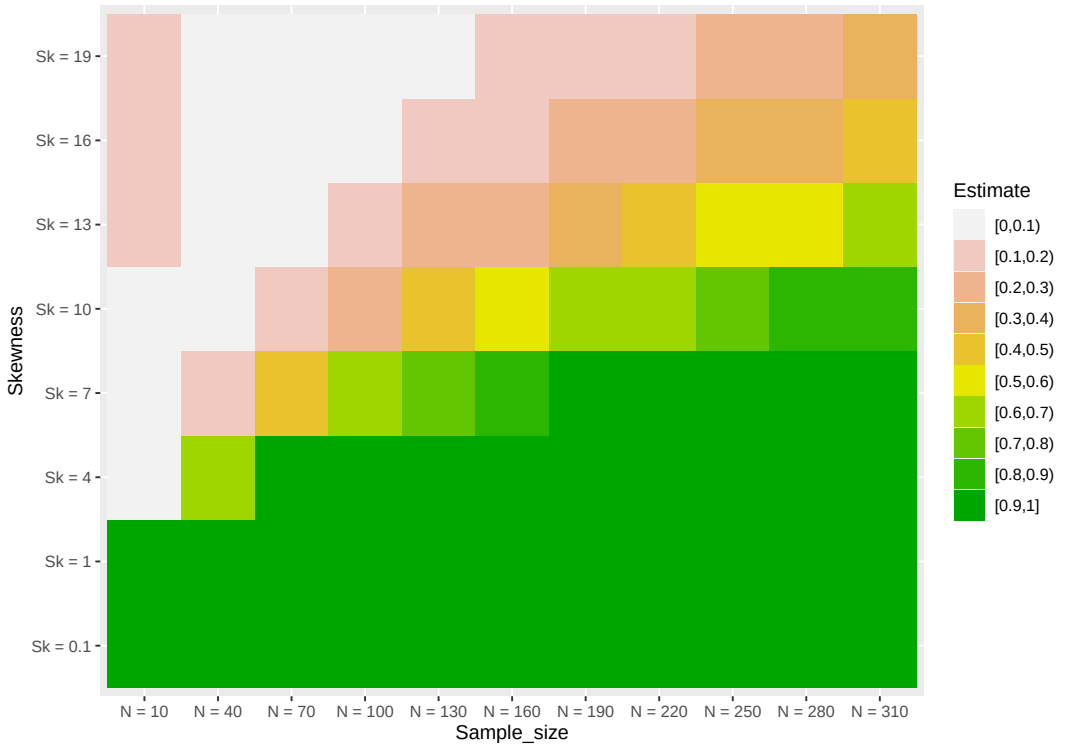

Рис. 44: Оценка мощности одновыборочного критерия Стьюдента в зависимости от размера выборки и коэффициента асимметрии для выборок из логнормального распределения.

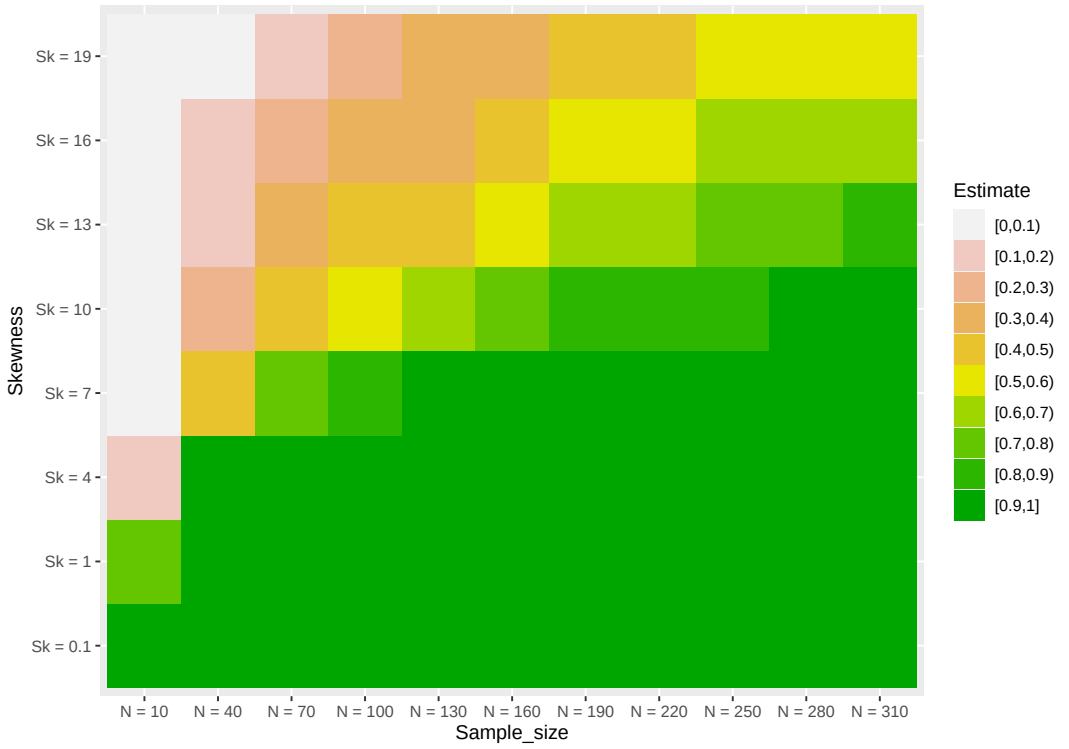

Рис. 45: Оценка мощности одновыборочного критерия Стьюдента с поправкой на асимметрию в зависимости от размера выборки и коэффициента асимметрии для выборок из логнормального распределения.

Заключение

Как было отмечено в главе 1, из рассмотренных в работе методов наибольшей мощностью на исследуемых распределениях обладал критерий Шапиро — Уилка. Эти результаты также подтверждаются в работах [39] и [40].
В главе 2 мы показали, что предположение о распределении роста взрослого человека по гауссовскому закону является достаточно грубым, и распределение роста точнее описывается логарифмически нормальным распределением.
Далее в главе 3 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.
Систематический обзор статей, проведенный в разделе 4, показал, что в половине статей, содержащих статистический анализ, используется предварительная проверка данных на нормальность. Зачастую эта проверка проводилась в рамках двухэтапной процедуры тестирования с критериями Стьюдента и Манна – Уитни. Наше исследование ошибки I рода такой двухэтапной процедуры показало, что с формальной точки зрения предварительная проверка на нормальность неверна. Таким образом, мы показали, что применение критерия Стьюдента на практике оправдано.
Затем мы обратились к рядам Эджворта и изучили их асимптотическую сходимость в главе 5. Оказалось, что на интересующих нас выборках до 100 элементов ошибка в оценке отклонения от нормальности слишком велика.
С другой стороны, ряды Эжворта обратили наше внимание на свои первые коэффициенты в разложении – асимметрию и эксцесс распределения. В главе 6 мы подробно исследовали их влияние на статистики одновыборочного и двухвыборочного критериев Стьюдента в модели, где данные извлекались из логарифмически нормального распределения, и выяснили, что при проверке одновыборочной гипотезы для контроля ошибки I рода достаточно использовать показатель асимметрии. Модифицированный критерий Стьюдента с поправкой на асимметрию, использующий метод бутстрэпа, имел в нашем исследовании хорошие показатели ошибки I рода и мощности. Из этого мы делаем вывод, что как сам по себе модифицированный критерий Стьюдента, так и подход с использованием метода бутстрэпа, – это достойные кандидаты для замены популярной двухэтапной процедуры тестирования гипотезы о центральной тенденции с предварительной проверкой на нормальность.

Список литературы

[1] *Берестнева О. Г., Марухина О. В., Шевелев Г. Е.* Прикладная математическая статистика //Томск: Изд-во Томского политех. ун-та. – 2012.

[2] *Гланц С.* Медико-биологическая статистика. Электронная книга. – 1999.

[3] *Ивченко Г. И., Медведев Ю.И.*, Математическая статистика: Учеб. пособие для втузов.— М.: Высш.шк., 1984. — 248 с.

[4] *Zar, Jerrold H.*, Biostatistical analysis: Pearson new international edition. Pearson Higher Ed, 2013.

[5] *М.Б. Лагутин*, Наглядная математическая статистика: учебное пособие - 2-е изд., испр.— М.: БИНОМ. Лаборатория знаний, 2009. — 472 с.

[6] *Rizzo M. L.*, Statistical computing with R. – CRC Press, 2019.

[7] *Постовалов С. Н.*, Применение компьютерного моделирования для расширения прикладных возможностей классических методов проверки статистических гипотез //Дисс. на соискание уч. степенид. т. н., НГТУ, 2013г.–298с. – 2013.

[8] *Lilliefors H. W.*, On the Kolmogorov-Smirnov test for normality with mean and variance unknown //Journal of the American statistical Association. – 1967. – Т. 62. – №. 318. – С. 399-402.

[9] *Shapiro S. S., Wilk M. B.*, An analysis of variance test for normality (complete samples) //Biometrika. – 1965. – Т. 52. – №. 3/4. – С. 591-611.

[10] *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.

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

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

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

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

[15] *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.

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

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

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

[19] *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

[20] *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.

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

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

[23] *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.

[24] *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.

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

[26] *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.

[27] *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.

[28] *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.

[29] *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.

[30] *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.

[31] *Sonderegger D.*, The Central Limit Theorem, Edgeworth Expansions : дис. – Montana State University, 2004.

[32] *Боровков А. А.*, Теория вероятностей. - 1999.

[33] *Прохоров Ю.В., Розанов Ю. А.*, Теория вероятностей. Основные понятия. Предельные теоремы. Случайные процессы. – Наука, 1987.

[34] *Lyon J. D., Barber B. M., Tsai C. L.*, Improved methods for tests of long-run abnormal stock returns //The Journal of Finance. – 1999. – Т. 54. – №. 1. – С. 165-201.

[35] *Boos D., Stefanski L.*, Efron's bootstrap //Significance. – 2010. – Т. 7. – №. 4. – С. 186-188.

[36] *Shao J., Tu D.*, The jackknife and bootstrap. – Springer Science Business Media, 1995.

[37] *Westfall P. H.*, Kurtosis as peakedness, 1905–2014. RIP //The American Statistician. – 2014. – Т. 68. – №. 3. – С. 191-195.

[38] *Корнеев А. А., Кричевец А. Н.*, Условия применимости критериев Стъюдента и Манна-Уитни //Психологический журнал. – 2011. – Т. 32. – №. 1. – С. 97-110.

[39] *Razali N. M. et al.*, Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests //Journal of statistical modeling and analytics. – 2011. – Т. 2. – №. 1. – С. 21-33.

[40] *Кобзарь А. И.*, Прикладная математическая статистика. Для инженеров и научных работников. – 2012.

Листинг программного кода

```
##### NEYMAN-PEARSON #####

N=100
alpha=0.05
nu=seq(0,3,by=0.05)

qw1=qt(alpha/2, df=N-1, ncp=0, lower.tail = T, log.p = FALSE)
qw2=qt(1-alpha/2, df=N-1, ncp=0, lower.tail = T, log.p = FALSE)
mypow_t=seq(1,61,by=1)
powf_t=function(i){
  return(1-pt(qw2, df=N-1, ncp=((i) * sqrt(N)/2), lower.tail = T, log.p = FALSE)+
           pt(qw1, df=N-1, ncp=((i)* sqrt(N)/2), lower.tail = T, log.p = FALSE))
}
for (j in mypow_t){
  mypow_t[j] = powf_t((j-1)/20)
}

q=qt(1-alpha, df=N-1, ncp=0, lower.tail = T, log.p = FALSE)
mypow_o=seq(1,61,by=1)
powf_o=function(i){
  return(1-pt(q, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE))
}
for (j in mypow_o){
  mypow_o[j] = powf_o((j-1)/20)
}

f1=pt(-3.5, df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
q1=qt(f1+(alpha/3), df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
f2=pt(0.5, df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
q2=qt(f2+(alpha/3), df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
f3=pt(2, df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
q3=qt(f3+(alpha/3), df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
myp=seq(1,61,by=1)
pw=function(i){
  m=pt(q1, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)-
    pt(-3.5, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)+
    pt(q2, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)-
    pt(0.5, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)+
    pt(q3, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)-
    pt(2, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)
  return(m)
}
for (j in myp){
  myp[j] = pw((j-1)/20)
}
```

```
qw041=qt((4*alpha/5), df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
qw042=qt((1-alpha/5), df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
myp04=seq(1,61,by=1)
pw04=function(i){
  return(return(1-pt(qw042, df=N-1, ncp=((i) * sqrt(N)/2), lower.tail = T,
  log.p = FALSE)+ pt(qw041, df=N-1, ncp=((i) * sqrt(N)/2), lower.tail = T,
  log.p = FALSE)))
}
for (j in myp04){
  myp04[j] = pw04((j-1)/20)
}

plot(nu,mypow_t,xlab="mu_1",ylab="power",col=rgb(1,0,0,0.7),type = "l",lwd=4)
lines(nu,myp,type="l",col=rgb(0,1,0,0.7),lwd=4)
lines(nu,mypow_o,col=rgb(0,0,1,0.7),type = "l",lwd=4)
lines(nu,myp04,type="l",col=rgb(1,1,0,0.7),lwd=4)

nu=seq(-3,3,by=0.05)

qw1=qt(alpha/2, df=N-1, ncp=0, lower.tail = T, log.p = FALSE)
qw2=qt(1-alpha/2, df=N-1, ncp=0, lower.tail = T, log.p = FALSE)
mypow_t=seq(1,121,by=1)
powf_t=function(i){
  return(1-pt(qw2, df=N-1, ncp=((i) * sqrt(N)/2), lower.tail = T, log.p = FALSE)+
           pt(qw1, df=N-1, ncp=((i) * sqrt(N)/2), lower.tail = T, log.p = FALSE))
}
for (j in mypow_t){
  mypow_t[j] = powf_t((j-1)/20 -3)
}

q=qt(1-alpha, df=N-1, ncp=0, lower.tail = T, log.p = FALSE)
mypow_o=seq(1,121,by=1)
powf_o=function(i){
  return(1-pt(q, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE))
}
for (j in mypow_o){
  mypow_o[j] = powf_o((j-1)/20 -3)
}

f1=pt(-3.5, df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
q1=qt(f1+(alpha/3), df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
f2=pt(0.5, df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
q2=qt(f2+(alpha/3), df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
f3=pt(2, df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
q3=qt(f3+(alpha/3), df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
myp=seq(1,121,by=1)
```

```
pw=function(i){
  m=pt(q1, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)-
    pt(-3.5, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)+
    pt(q2, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)-
    pt(0.5, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)+
    pt(q3, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)-
    pt(2, df=N-1, ncp=(i) * sqrt(N)/2, lower.tail = T, log.p = FALSE)
  return(m)
}

for (j in myp){
  myp[j] = pw((j-1)/20 -3)
}
qw041=qt((4*alpha/5), df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
qw042=qt((1-alpha/5), df=N-1, ncp=0, lower.tail = TRUE, log.p = FALSE)
myp04=seq(1,121,by=1)
pw04=function(i){
  return(return(1-pt(qw042, df=N-1, ncp=((i) * sqrt(N)/2), lower.tail = T,
  log.p = FALSE)+ pt(qw041, df=N-1, ncp=((i) * sqrt(N)/2), lower.tail = T,
  log.p = FALSE)))
}
for (j in myp04){
  myp04[j] = pw04((j-1)/20 -3)
}

plot(nu,mypow_t,xlab="mu_1",ylab="power",col=rgb(1,0,0,0.7),type = "l",lwd=4)
lines(nu,myp,type="l",col=rgb(0,1,0,0.7),lwd=4)
lines(nu,mypow_o,col=rgb(0,0,1,0.7),type = "l",lwd=4)
lines(nu,myp04,type="l",col=rgb(1,1,0,0.7),lwd=4)

##########

##### POWER ANALYSIS #####

library(moments)
library(nortest)
library(rmutil)

K = 10^4

gen_sample = function(name, N){
  if (name == "beta_22"){
    return(rbeta(shape1 = 2, shape2 = 2, n = N))
  }
  if (name == "gamma45"){
    return(rgamma(shape = 4, scale = 5, n = N))
  }
```

```
  if (name == "gamma15"){
    return(rgamma(shape = 1, scale = 5, n = N))
  }
  if (name == "laplace01"){
    return(rlaplace(m=0, s=1, n = N))
  }
}

quick_pearson = function(x, alpha = 0.05)
  return(pearson.test(x)$p.value < 0.05)

pearson_power = function (K = 1000, type = "beta_22", N = 100){
  sample_matrix = replicate(n = K, expr =  gen_sample(name = type, N=N))
  sample_matrix = as.data.frame(sample_matrix)
  p_vals = lapply(sample_matrix, quick_pearson)
  p_vals = as.numeric(p_vals)
  power = sum(p_vals)/length(p_vals)
  return(power)
}

quick_jarque = function(x, alpha = 0.05)
  return(jarque.test(x)$p.value < 0.05)

jarque_power = function (K = 1000, type = "beta_22", N = 100){
  sample_matrix = replicate(n = K, expr =  gen_sample(name = type, N=N))
  sample_matrix = as.data.frame(sample_matrix)
  p_vals = lapply(sample_matrix, quick_jarque)
  p_vals = as.numeric(p_vals)
  power = sum(p_vals)/length(p_vals)
  return(power)
}

quick_slillie = function(x, alpha = 0.05)
  return(lillie.test(x)$p.value < 0.05)

lillie_power = function (K = 1000, type = "beta_22", N = 100){
  sample_matrix = replicate(n = K, expr =  gen_sample(name = type, N=N))
  sample_matrix = as.data.frame(sample_matrix)
  p_vals = lapply(sample_matrix, quick_slillie)
  p_vals = as.numeric(p_vals)
  power = sum(p_vals)/length(p_vals)
  return(power)
}

quick_shapiro = function(x, alpha = 0.05)
  return(shapiro.test(x)$p.value < 0.05)
```

```
Shapiro_wilk_power = function (K = 1000, type = "beta_22", N = 100){
  sample_matrix = replicate(n = K, expr =  gen_sample(name = type, N=N))
  sample_matrix = as.data.frame(sample_matrix)
  p_vals = lapply(sample_matrix, quick_shapiro)
  p_vals = as.numeric(p_vals)
  power = sum(p_vals)/length(p_vals)
  return(power)
}

N_vector = list(10, 20, 30, 50, 100, 200, 300, 400, 500, 1000,2000)
X=seq(1,length(N_vector),by=1)
p_b22=sapply(N_vector, function(n) pearson_power(type = "beta_22", N = n))
j_b22=sapply(N_vector, function(n) jarque_power(type = "beta_22", N = n))
l_b22=sapply(N_vector, function(n) lillie_power(type = "beta_22", N = n))
s_b22=sapply(N_vector, function(n) Shapiro_wilk_power(type = "beta_22", N = n))
plot(x=X,y=s_b22,type = "o",main="Beta(2,2)",ylab = "Мощность",
     xlab="Размер выборки, n",col="red",pch=1,lwd=2,xaxt="n")
lines(x=X,y=j_b22,type = "o",col="blue",pch=2,lwd=2,xaxt="n")
lines(x=X,y=l_b22,type = "o",col="green",pch=3,lwd=2,xaxt="n")
lines(x=X,y=p_b22,type = "o",col="violet",pch=5,lwd=2,xaxt="n")
legend("bottomright",legend=c("Шапиро-Уилк","Харке-Бер","Лиллиефорс",
    "Хи-квадрат Пирсона"), col=c("red","blue","green","violet"), lwd=2,cex = 0.75)
axis(1, at=X,labels=N_vector, las=1)

p_l01=sapply(N_vector, function(n) pearson_power(type = "laplace01", N = n))
j_l01=sapply(N_vector, function(n) jarque_power(type = "laplace01", N = n))
l_l01=sapply(N_vector, function(n) lillie_power(type = "laplace01", N = n))
s_l01=sapply(N_vector, function(n) Shapiro_wilk_power(type = "laplace01", N = n))
plot(x=X,y=s_l01,type = "o",main="Laplace(0,1)",ylab = "Мощность",
     xlab="Размер выборки, n",col="red",pch=1,lwd=2,xaxt="n")
lines(x=X,y=j_l01,type = "o",col="blue",pch=2,lwd=2,xaxt="n")
lines(x=X,y=l_l01,type = "o",col="green",pch=3,lwd=2,xaxt="n")
lines(x=X,y=p_l01,type = "o",col="violet",pch=5,lwd=2,xaxt="n")
legend("bottomright",legend=c("Шапиро-Уилк","Харке-Бер","Лиллиефорс",
    "Хи-квадрат Пирсона"), col=c("red","blue","green","violet"), lwd=2,cex = 0.75)
axis(1, at=X,labels=N_vector, las=1)

p_g45=sapply(N_vector, function(n) pearson_power(type = "gamma45", N = n))
j_g45=sapply(N_vector, function(n) jarque_power(type = "gamma45", N = n))
l_g45=sapply(N_vector, function(n) lillie_power(type = "gamma45", N = n))
s_g45=sapply(N_vector, function(n) Shapiro_wilk_power(type = "gamma45", N = n))
plot(x=X,y=s_g45,type = "o",main="Gamma(4,5)",ylab = "Мощность",
     xlab="Размер выборки, n",col="red",pch=1,lwd=2,xaxt="n")
lines(x=X,y=j_g45,type = "o",col="blue",pch=2,lwd=2,xaxt="n")
lines(x=X,y=l_g45,type = "o",col="green",pch=3,lwd=2,xaxt="n")
lines(x=X,y=p_g45,type = "o",col="violet",pch=5,lwd=2,xaxt="n")
legend("bottomright",legend=c("Шапиро-Уилк","Харке-Бер","Лиллиефорс",
```

```
    "Хи-квадрат Пирсона"), col=c("red","blue","green","violet"), lwd=2,cex = 0.75)
axis(1, at=X,labels=N_vector, las=1)

p_g15=sapply(N_vector, function(n) pearson_power(type = "gamma15", N = n))
j_g15=sapply(N_vector, function(n) jarque_power(type = "gamma15", N = n))
l_g15=sapply(N_vector, function(n) lillie_power(type = "gamma15", N = n))
s_g15=sapply(N_vector, function(n) Shapiro_wilk_power(type = "gamma15", N = n))
plot(x=X,y=s_g15,type = "o",main="Gamma(1,5)",ylab = "Мощность",
     xlab="Размер выборки, n",col="red",pch=1,lwd=2,xaxt="n")
lines(x=X,y=j_g15,type = "o",col="blue",pch=2,lwd=2,xaxt="n")
lines(x=X,y=l_g15,type = "o",col="green",pch=3,lwd=2,xaxt="n")
lines(x=X,y=p_g15,type = "o",col="violet",pch=5,lwd=2,xaxt="n")
legend("bottomright",legend=c("Шапиро-Уилк","Харке-Бер","Лиллиефорс",
    "Хи-квадрат Пирсона"), col=c("red","blue","green","violet"), lwd=2,cex = 0.75)
axis(1, at=X,labels=N_vector, las=1)

##########

##### ROUNDING #####

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 = 10^4
ALPHA_LEVEL = .05

N_mesh = round(10^seq(from = log10(5), to = log10(5000), length.out = LEN))  %>%
    unique()
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[index_s], rot = 45), y=list(at=index_s,labels=
    round(s_mesh[index_s],2))), 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_matrix), ncol = NCOL(res_round3_pure_matrix)),
    "Без округления")


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,labels=N_mesh[index_s], rot = 45),
                           y=list(at=index_s,labels=round(s_mesh[index_s],2))),
               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)

##########

##### TWO-STEP TEST #####

library(dplyr)
library(tictoc)
library(ggplot2)

RATE <- 1
```

```
LEN <- seq(from = 10, to = 30, by = 2)
ITERS = 1e3

shapiro_pass <- function(N = 10, p_shapiro = 0.05){
  data_1 <- rexp(N, rate = RATE)
  data_2 <- rexp(N, rate = RATE)

  if(shapiro.test(data_1)$p.value > p_shapiro &&
  shapiro.test(data_2)$p.value > p_shapiro)
    res <- 1
  else
    res <- 0

  res
}

p_pass <- function(N = 10){
  x <- replicate(n = ITERS, shapiro_pass(N)) %>% unlist
  bt = binom.test(sum(x), ITERS)
  res <- c(bt$conf.int[1], bt$estimate, bt$conf.int[2])

  res
}

matrix_conf_int_pass <- lapply(LEN, p_pass) %>% as.data.frame() %>% t()
rownames(matrix_conf_int_pass) = LEN
colnames(matrix_conf_int_pass) = c("LI", "M","UI")

ggplot_df = matrix_conf_int_pass %>% as.data.frame()
ggplot_df$x = rownames(ggplot_df)
ggplot(ggplot_df, aes(x = x)) + geom_line(aes( y = M, group = 1),col = "red") +
geom_ribbon(aes(group = 1, ymin=LI, ymax=UI), fill = "red", alpha =0.1) +
theme_light()

p_student <- function(N = 10){
  res  = -1

  while (res == -1)
  {
    data_1 <- rexp(N, rate = RATE)
    data_2 <- rexp(N, rate = RATE)

    if(shapiro.test(data_1)$p.value > 0.05 && shapiro.test(data_2)$p.value > 0.05)
      res <- t.test(data_1, data_2)$p.value
  }
  res
}
```

```
alpha_student <- function(N = 10){
  M <- ITERS
  pb <- txtProgressBar(min = 0, max = M, style = 3)
  p_vec <- vector(len = M)

  for(i in 1:M){
    setTxtProgressBar(pb, i)
    p_vec[i] <- p_student(N)
  }

  close(pb)

  res = (p_vec < 0.05)
  bt = binom.test(x = sum(res), n = M)
  c(bt$conf.int[1], bt$estimate, bt$conf.int[2])
}

tic()
matrix_alpha_student <- lapply(LEN, alpha_student) %>% as.data.frame() %>% t()
rownames(matrix_alpha_student) = LEN
colnames(matrix_alpha_student) = c("LI_2", "M_2","UI_2")
toc()

matrix_product = matrix_conf_int_pass*matrix_alpha_student

ggplot_df_2 = rbind(matrix_conf_int_pass, matrix_alpha_student , matrix_product) %>%
as.data.frame()

ggplot_df_2$x = rep(LEN,3)
ggplot_df_2$group = rep(c("Shapiro pretest passed", "Student I type error",
"Product"),each = NROW(LEN)) %>% factor()

ggplot_df_2
ggplot(ggplot_df_2, aes(x = x, col = group)) + xlab("Sample size, n") +
ylab("Probability") + geom_line(aes( y = M, group = group)) +
  geom_ribbon(aes(ymin=LI, ymax=UI, group = group, fill = group), alpha =0.1) +
  theme_light() + ylim(c(0, NA)) + geom_hline(yintercept = 0.05) +
  annotate("text", max(ggplot_df_2$x)-3, 0.05, vjust = -1, label = "0.05 level")

shapiro_not_pass <- function(N = 10){
  data_1 <- rexp(N, rate = RATE)
  data_2 <- rexp(N, rate = RATE)

  if(shapiro.test(data_1)$p.value < 0.05 || shapiro.test(data_2)$p.value < 0.05)
    res <- 1
  else
```

```
    res <- 0

  res
}

p_not_pass <- function(N = 10){
  x <- replicate(n = ITERS, shapiro_not_pass(N)) %>% unlist
  bt = binom.test(sum(x), ITERS)
  res <- c(bt$conf.int[1], bt$estimate, bt$conf.int[2])

  res
}

matrix_conf_int_not_pass <- lapply(LEN, p_not_pass) %>% as.data.frame() %>% t()
rownames(matrix_conf_int_not_pass) = LEN
colnames(matrix_conf_int_not_pass) = c("LI", "M","UI")

p_wilcox <- function(N = 10){

  res = -1
  while(res == -1)
  {
    data_1 <- rexp(N, rate = RATE)
    data_2 <- rexp(N, rate = RATE)
    if(shapiro.test(data_1)$p.value < 0.05 || shapiro.test(data_2)$p.value < 0.05)
      res = wilcox.test(data_1,data_2)$p.value
  }

  res
}

alpha_wilcox <- function(N = 10){
  M <- 1e4
  pb <- txtProgressBar(min = 0, max = M, style = 3)
  p_vec <- vector(len = M)

  for(i in 1:M){
    setTxtProgressBar(pb, i)
    p_vec[i] <- p_wilcox(N)
  }

  close(pb)

  res = (p_vec < 0.05)
  bt = binom.test(x = sum(res), n = M)
  c(bt$conf.int[1], bt$estimate, bt$conf.int[2])
}
```

```
tic()
matrix_alpha_wilcox <- lapply(LEN, alpha_wilcox) %>% as.data.frame() %>% t()
rownames(matrix_alpha_wilcox) = LEN
colnames(matrix_alpha_wilcox) = c("LI_2", "M_2","UI_2")
toc()

matrix_product_2 = matrix_alpha_wilcox*matrix_conf_int_not_pass

ggplot_df_3 = rbind(matrix_conf_int_not_pass, matrix_alpha_wilcox ,
matrix_product_2) %>% as.data.frame()

ggplot_df_3$x = rep(LEN,3)
ggplot_df_3$group = rep(c("Shapiro pretest not passed", "Wilcoxon I type error",
"Product"),each = NROW(LEN)) %>% factor()

ggplot_df_3
ggplot(ggplot_df_3, aes(x = x, col = group)) + xlab("Sample size, n") +
ylab("Probability") + geom_line(aes( y = M, group = group)) +
geom_ribbon(aes(ymin=LI, ymax=UI, group = group, fill = group), alpha =0.1) +
theme_light() + ylim(c(0.0, 1.))+ geom_hline(yintercept = 0.05) +
annotate("text", max(ggplot_df_2$x)-3, 0.05, vjust = -1, label = "0.05 level")

total_error = matrix_product+matrix_product_2
ggplot_df_4 = rbind(matrix_product, matrix_product_2, total_error) %>%
as.data.frame()

ggplot_df_4$x = rep(LEN,3)
ggplot_df_4$group = rep(c("Conditional Student I type error",
"Conditional Wilcoxon I type error", "Total  I type error"), each = NROW(LEN)) %>%
factor()

ggplot_df_4
ggplot(ggplot_df_4, aes(x = x, col = group))  +
geom_line(aes( y = M, group = group)) +
geom_ribbon(aes(ymin=LI, ymax=UI, group = group, fill = group), alpha =0.1) +
theme_light() + ylim(c(0, NA))+ geom_hline(yintercept = 0.05) +
annotate("text", max(ggplot_df_2$x)-3, 0.05, vjust = -1, label = "0.05 level")

##########

##### EDGEWORTH #####

library(dplyr)
library(tictoc)
library(ggplot2)
library(foreach)
```

```
library(doParallel)


gglist = list()
flag = FALSE
N_for_skew = list()
for ( j in 1:length(skew_vect))
{
  tmp = res_list[[j]]
  m_result = apply(X = tmp,MARGIN = 1 , mean)
  se_result = apply(X = tmp,MARGIN = 1 , sd)/sqrt(LEN)


  edj_ln01 = sapply(N_vector, function(n) fun2(N = n,skew_ln = skew_vect[[j]]))
  edj_ln01 = pmax(edj_ln01,0.04)


  df = data.frame(edgeworth = edj_ln01, m = m_result, se = se_result,
                  N = N_vector)

  for (i in 2:nrow(df)){
    if ((df[i, 2]  - df[i, 1]) < 0 & (df[i-1, 2] - df[i-1, 1]) > 0 & flag == FALSE){
      N_for_skew[[j]] = df[i, 4]
      flag = TRUE
    }
  }
  if (flag == FALSE){
    N_for_skew[[j]] = -1
  }
  flag = FALSE

  gglist[[j]] = ggplot(df, aes(x = N, y = edgeworth)) + theme_light() +
  geom_line() + geom_ribbon(aes(ymin = m - 1.96*se, ymax =  m + 1.96*se),
  fill = "black", alpha = .1) + scale_x_continuous(trans = "log10",
  limits = c(10,NA)) + ggtitle(paste0("Skew = ", skew_vect[[j]])) +
  ylim(c(0.04,0.07)) + geom_vline(xintercept = N_for_skew[[j]],
  linetype = 'dashed', colour = 'gray')
}

multiplot(plotlist = gglist, cols = 4)


##### EDGEWORTH ABS.ERR #####

gglist = list()

delta  = 1
```

```
N_for_skew = list()
flag = FALSE

for ( j in 1:length(skew_vect))
{
  tmp = res_list[[j]]
  m_result = apply(X = tmp,MARGIN = 1 , mean)
  se_result = apply(X = tmp,MARGIN = 1 , sd)/sqrt(LEN)

  edj_ln01 = sapply(N_vector, function(n) fun2(N = n,skew_ln = skew_vect[[j]]))
  edj_ln01 = pmax(edj_ln01,0.04)

  error = (m_result - edj_ln01) * 100

  df = data.frame(err = error,
                  N = N_vector)

  for (i in 2:nrow(df)){
    if (abs(df[i, 1]) < delta & abs(df[i-1, 1]) > delta){
      N_for_skew[[j]] = df[i, 2]
      flag = TRUE
    }
  }
  if (flag == FALSE){
    N_for_skew[[j]] = 0
  }
  flag = FALSE

  gglist[[j]] = ggplot(df, aes(x = N, y = err))  + theme_light() + geom_line() +
  scale_x_continuous(trans = "log10", limits = c(10,NA)) +
  ggtitle(paste0("Skew = ", skew_vect[[j]])) + ylim(c(-0.5,2.5)) +
  geom_hline(yintercept = delta, colour = 'red', alpha = .1) +
  geom_hline(yintercept = -delta, colour = 'red', alpha = .1) +
  geom_vline(xintercept = N_for_skew[[j]], linetype = 'dashed', colour = 'gray')
}

multiplot(plotlist = gglist, cols = 4)

##########

##### t-TEST #####

library(dplyr)
library(ggplot2)
library(moments)
library(Rmisc)
library(tictoc)
```

```
my_sdlog = 1 # 0.01
extract_stat = function()
{
  A = rlnorm(100, sdlog = my_sdlog)
  sqrt(100)*((mean(A) - exp(0.5*my_sdlog**2))/ sd(A))
}

tic()
stat_distr = replicate(extract_stat(), n = 1e4)
toc()

hist(stat_distr)
plot(density(stat_distr), xlab = 'x', ylab = 'Density', main = '', lwd = 2)

passed = ( (stat_distr > qt(p = .975,df = 99)) | (stat_distr < qt(p = .025,df = 99)))
binom.test(x = sum(passed), n = length(passed))

##########

extract_stat = function(my_sdlog)
{
  A = rlnorm(100, sdlog = my_sdlog)
  B = rlnorm(100, sdlog = my_sdlog)
  (mean(A) - mean(B))/( sqrt(2/100) * sqrt((99*((sd(A))^2 + (sd(B))^2 ))/198) )
}

stat_sample = function(my_sdlog)
{
  replicate(extract_stat(my_sdlog), n = 1e4)
}

stat_distr_1 = stat_sample(1)
stat_distr_3 = stat_sample(3)
stat_distr_10 = stat_sample(10)
stat_distr_100 = stat_sample(100)

plot(density(stat_distr_1), xlab = 'x', ylab = 'Density', main = '', col = 1,
lwd = 2, ylim = c(0,1), xlim = c(-3, 3)) #, adj = .3
lines(density(stat_distr_3), col = 2, lwd = 2)
lines(density(stat_distr_10), col = 3, lwd = 2)
lines(density(stat_distr_100), col = 4, lwd = 2)
abline(v = -1, lwd = 1, lty = 2)
abline(v = 1, lwd = 1, lty = 2)
legend(x = 'topright', legend = c('sigma = 1', 'sigma = 3', 'sigma = 10',
'sigma = 100'), col = c(1,2,3,4), lwd = 2)
```

```
passed_1 = ( (stat_distr_1 > qt(p = .975,df = 198)) |
(stat_distr_1 < qt(p = .025,df = 198)))
binom.test(x = sum(passed_1), n = length(passed_1))

passed_3 = ( (stat_distr_3 > qt(p = .975,df = 198)) |
(stat_distr_3 < qt(p = .025,df = 198)))
binom.test(x = sum(passed_3), n = length(passed_3))

passed_10 = ( (stat_distr_10 > qt(p = .975,df = 198)) |
(stat_distr_10 < qt(p = .025,df = 198)))
binom.test(x = sum(passed_10), n = length(passed_10))

passed_100 = ( (stat_distr_100 > qt(p = .975,df = 198)) |
(stat_distr_100 < qt(p = .025,df = 198)))
binom.test(x = sum(passed_100), n = length(passed_100))

##########

extract_stat = function(my_meanlog, my_sdlog)
{
  A = rlnorm(100, sdlog = my_sdlog, meanlog = my_meanlog)
  B = rlnorm(100, sdlog = my_sdlog, meanlog = my_meanlog)

  tt = t.test(A, B, var.equal = T)
  tt$p.value
}


rej_est = function(my_meanlog, my_sdlog)
{
  p_vector = replicate(extract_stat(my_meanlog, my_sdlog), n = 1e4)

  mean(p_vector<0.05)
}

sd_mesh = seq(0.1, 5, length.out = 6)
mean_mesh = seq(-100, 100, length.out = 5)

res = outer(X = mean_mesh, Y = sd_mesh, FUN = Vectorize(rej_est))

colnames(res) = sd_mesh
res

for ( i in 1:NROW(res))
{
  if (i==1){
```

```
    plot(sd_mesh, res[i,], type = "b", ylab = 'Probability', xlab = 'sigma',
    ylim = c(0,0.1), col = i, axes = F)
    axis(1, sd_mesh)
    axis(2)
    box()
  }

  if (i>=2)
  {lines(sd_mesh, res[i,], type = "b", ylim = c(0,0.1), col = i)}
}
legend(x = 'topright', legend = c('mu = -100', 'mu = -50', 'mu = 0',
'mu = 50', 'mu = 100'), col = c(1,2,3,4,5), lwd = 2)
abline(h = 0.05, lwd = 1, lty = 2)

##########

library(gridExtra)

sdlog_mesh = c(0.1, 1, 3, 5, 7, 10)

samples = list()
means = list()
sds = list()

smp_mean_sd = list()
mean_sd = list()
stat = list()
means_for_hist = list()
means_del_percent = list()

for (i in 1:length(sdlog_mesh)){
  samples[[i]] = replicate(n = 1000, expr = rlnorm(sdlog = sdlog_mesh[i],
  n = 100), simplify = F)
  means[[i]] = sapply(X = samples[[i]], FUN = mean)
  sds[[i]] = sapply(X = samples[[i]], FUN = sd)
  smp_mean_sd[[i]] = ggplot(means[[i]]/sds[[i]] %>% as.data.frame(),
  aes(.)) + geom_density(color = i) + labs(x = 'x', y = 'Density',
  subtitle = paste0('sigma = ', sdlog_mesh[i])) + theme_light()
  mean_sd[[i]] = ggplot(-exp(0.5*sdlog_mesh[i]**2)/sds[[i]] %>%
  as.data.frame(), aes(.)) + geom_density(color = i) +
  labs(x = 'x', y = 'Density', subtitle = paste0('sigma = ',
  sdlog_mesh[i])) + theme_light()
  stat[[i]] = ggplot((means[[i]] - exp(0.5*sdlog_mesh[i]**2))/sds[[i]] %>%
  as.data.frame(), aes(.)) + geom_density(color = i) + labs(x = 'x',
  y = 'Density', subtitle = paste0('sigma = ', sdlog_mesh[i])) + theme_light()
  means_del_percent[i] = 100*sum(means[[i]] > exp(0.5*sdlog_mesh[i]**2)) /
  length(means[[i]])
```

```
  means_for_hist[[i]] = ggplot(means[[i]] %>% as.data.frame(), aes(.)) +
  geom_histogram(bins = 50, color = i, alpha = 0.5) + labs(x = 'x',
  subtitle = paste0('sigma = ', sdlog_mesh[i], ', per = ',
  means_del_percent[i], ' %')) + theme_light() +
  geom_vline(xintercept = exp(0.5*sdlog_mesh[i]**2), linetype = 'dashed')
}

multiplot(plotlist = smp_mean_sd, cols = 3)
multiplot(plotlist = mean_sd, cols = 3)
multiplot(plotlist = stat, cols = 3)

grid.arrange(smp_mean_sd[[1]], smp_mean_sd[[2]], smp_mean_sd[[3]],
smp_mean_sd[[4]], smp_mean_sd[[5]], smp_mean_sd[[6]], top = 'Densities
for (sample mean)/sd', ncol = 3)
grid.arrange(mean_sd[[1]], mean_sd[[2]], mean_sd[[3]], mean_sd[[4]],
mean_sd[[5]], mean_sd[[6]], top = 'Densities for (- real mean)/sd',
ncol = 3)
grid.arrange(stat[[1]], stat[[2]], stat[[3]], stat[[4]], stat[[5]],
stat[[6]], top = 'Densities for statistic', ncol = 3)

grid.arrange(means_for_hist[[1]], means_for_hist[[2]],
means_for_hist[[3]], means_for_hist[[4]], means_for_hist[[5]],
means_for_hist[[6]], top = 'Histogram for sample mean', ncol = 3)

##########

##### TYPE I ERR. t-TEST #####

library(ggplot2)
library(moments)
library(Rmisc)
library(tictoc)
library(dplyr)
library(RColorBrewer)
library(remotes)
library(future.apply)
library(future)

plan(multisession, workers = 12)

extract_p = function(K = 100, skew = 6.18, N_MC = 100)
{
  sk_x = function(x){ (x+2)*(x-1)^.5 }

  s2_root = uniroot( function(z) {sk_x(z) - skew} , interval = c(1,10e5))

  my_sdlog = s2_root$root %>% log() %>% sqrt()
```

```
  theor_mean = exp(0+0.5*my_sdlog**2)

  tmp_f = function()
  {
    A = rlnorm(K, sdlog = my_sdlog)
    tt = t.test(A,mu = theor_mean)
    tt$p.value
  }
  future_replicate(tmp_f(), n = N_MC) %>% as.numeric()
}

N_list = seq(from = 10, to = 310, by = 30)
sk_list = c(0.1,seq(from = 1, to = 21, by = 3))

i_N = length(N_list)
j_N = length(sk_list)
res_est = matrix(nrow = i_N, ncol = j_N)
res_low = matrix(nrow = i_N, ncol = j_N)
res_upp = matrix(nrow = i_N, ncol = j_N)

tic()
for (i in 1:i_N)
{
  for (j in 1:j_N)
  {
    tmp = extract_p(N_list[i],sk_list[j], N_MC = 1e6)
    test_res = binom.test(x= sum(tmp<0.05), n = length(tmp<0.05))
    res_low[i,j] = test_res$conf.int[1]
    res_upp[i,j] = test_res$conf.int[2]
    res_est[i,j] = test_res$estimate
  }
}
toc()

  colnames(res_low) = paste("Sk =",sk_list)
  rownames(res_low) = paste("N =",N_list)

  ggres = reshape2::melt(res_low)
  names(ggres) = c("Sample_size", "Skewness","error")

  ggres$Estimate = cut(ggres$error, breaks = c(0,0.05,0.075,0.1,1),
  include.lowest = T, right = F)

  ggplot(ggres, aes(Sample_size, Skewness, fill= Estimate)) +
    geom_tile() +
    scale_fill_manual(values=c( "#006B3E","#FFE733","#FFAA1C","#ED2938"))
```

```
##########

##### TYPE I ERR. skewt-TEST #####

library(ggplot2)
library(moments)
library(Rmisc)
library(tictoc)
library(dplyr)
library(RColorBrewer)
library(remotes)
library(future.apply)
library(future)

plan(multisession, workers = 12)

skewt.test <-
  function (x, mu = 0, conf.level = 0.95, b.frac = 1 / 4, N = 10^4, ...)
  {
    if (!missing(mu) && (length(mu) != 1 || is.na(mu)))
      stop("'mu' must be a single number")
    if (!missing(conf.level) &&
        (length(conf.level) != 1 || !is.finite(conf.level) ||
         conf.level < 0 || conf.level > 1))
      stop("'conf.level' must be a single number between 0 and 1")
    nx <- length(x)
    mx <- mean(x)
    vx <- var(x)
    if (nx < 2)
      stop("not enough 'x' observations")
    stderr <- sqrt(vx / nx)
    if (stderr < 10 * .Machine$double.eps * abs(mx))
      stop("data are essentially constant")
    df <- nx - 1
    tstat.calc <-
      function(n,S,gamma)
        sqrt(n) * (S + 1 / 3 * gamma * S ^ 2 + 1 / (6 * n) * gamma)
    S <- (mx - mu) / sqrt(vx)
    gamma <- sum((x - mx) ^ 3) / (nx * sqrt(vx) ^ 3)
    tstat <- tstat.calc(nx,S,gamma)
    nb <- max(nx * b.frac, 3) # Always sample at least 3
    # Calculate N boostrapped statistics, under the null hypothesis
    tsab <- replicate(N,{
      xb <- sample(x,nb,replace = T)
      m_xb <- mean(xb)
      sd_xb <- sd(xb)
```

```
      S_b <- (m_xb - mx) / sd_xb
      gamma_b <- sum((xb - m_xb) ^ 3) / (nb * sd_xb ^ 3)
      tstat.calc(length(xb),S_b,gamma_b)
    })
    # Remove NAs from the boostrapped vector
    tsab <- tsab[!is.na(tsab)] # (caused by constant subsamples from x)
    alpha <- 1 - conf.level
    tail <- if (tstat > 0)
      tsab > tstat
    else
      tsab <= tstat
    pval <- min(sum(tail) / (1 + N) * 2, 1)
    cint <- c(quantile(tsab,alpha / 2), quantile(tsab,1 - alpha / 2))
    sd <- sd(tsab)
    dname <- deparse(substitute(x))
    names(tstat) <- "t"
    names(sd) <- "sd"
    names(mu) <- "mean"
    attr(cint, "conf.level") <- conf.level
    method <- "Bootrsapped skewness-adjusted t-test (Lyon et al, 1999)"
    alternative <- "two.sided"
    rval <- list(
      statistic = tstat, parameter = sd, p.value = pval,
      conf.int = cint, estimate = mx, null.value = mu,
      alternative = alternative, method = method, data.name = dname
    )
    class(rval) <- "htest"
    return(rval)
  }

extract_p = function(K = 100, skew = 6.18, N_MC = 100)
{
  sk_x = function(x){ (x+2)*(x-1)^.5 }

  s2_root = uniroot( function(z) {sk_x(z) - skew} , interval = c(1,10e5))

  my_sdlog = s2_root$root %>% log() %>% sqrt()

  theor_mean = exp(0+0.5*my_sdlog**2)

  tmp_f = function()
  {
    A = rlnorm(K, sdlog = my_sdlog)
    tt = skewt.test(A,mu = theor_mean)
    tt$p.value
  }
```

```
  res = future_replicate(tmp_f(), n = N_MC) %>% as.numeric()
  res
}

N_list = seq(from = 10, to = 310, by = 30)
sk_list = c(0.1,seq(from = 1, to = 21, by = 3))

i_N = length(N_list)
j_N = length(sk_list)
res_est = matrix(nrow = i_N, ncol = j_N)
res_low = matrix(nrow = i_N, ncol = j_N)
res_upp = matrix(nrow = i_N, ncol = j_N)

tic()
for (i in 1:i_N)
{
  for (j in 1:j_N)
  {
    tmp = extract_p(N_list[i],sk_list[j], N_MC = 10^4)
    test_res = binom.test(x= sum(tmp<0.05), n = length(tmp<0.05))
    res_low[i,j] = test_res$conf.int[1]
    res_upp[i,j] = test_res$conf.int[2]
    res_est[i,j] = test_res$estimate
  }
}
toc()

  colnames(res_low) = paste("Sk =",sk_list)
  rownames(res_low) = paste("N =",N_list)

  ggres = reshape2::melt(res_low)
  names(ggres) = c("Sample_size", "Skewness","error")

  ggres$Lower_CI = cut(ggres$error, breaks = c(0,0.05,0.075,0.1,1),
  include.lowest = T, right = F)

  ggplot(ggres, aes(Sample_size, Skewness, fill= Lower_CI)) +
    geom_tile() +
    scale_fill_manual(values=c( "#006B3E","#FFE733","#FFAA1C","#ED2938"))

##########

##### POWER t-TEST #####

library(ggplot2)
library(moments)
library(Rmisc)
```

```
library(tictoc)
library(dplyr)
library(RColorBrewer)
library(remotes)
library(future.apply)
library(future)

plan(multisession, workers = 12)

extract_p = function(K = 100, skew = 6.18, N_MC = 100)
{
  sk_x = function(x){ (x+2)*(x-1)^.5 }

  s2_root = uniroot( function(z) {sk_x(z) - skew} , interval = c(1,10e5))

  my_sdlog = s2_root$root %>% log() %>% sqrt()

  theor_mean = exp(0+0.5*my_sdlog**2)

  tmp_f = function()
  {
    A = rlnorm(K, sdlog = my_sdlog)
    tt = t.test(A,mu = (theor_mean - 0.5))
    tt$p.value
  }
  future_replicate(tmp_f(), n = N_MC) %>% as.numeric()
}


N_list = seq(from = 10, to = 310, by = 30)
sk_list = c(0.1,seq(from = 1, to = 21, by = 3))

i_N = length(N_list)
j_N = length(sk_list)
res_est = matrix(nrow = i_N, ncol = j_N)
res_low = matrix(nrow = i_N, ncol = j_N)
res_upp = matrix(nrow = i_N, ncol = j_N)

tic()
for (i in 1:i_N)
{
  for (j in 1:j_N)
  {
    tmp = extract_p(N_list[i],sk_list[j], N_MC = 1e6)
    test_res = binom.test(x= sum(tmp<0.05), n = length(tmp<0.05))
    res_low[i,j] = test_res$conf.int[1]
    res_upp[i,j] = test_res$conf.int[2]
```

```
    res_est[i,j] = test_res$estimate
  }
}
toc()

  colnames(res_low) = paste("Sk =",sk_list)
  rownames(res_low) = paste("N =",N_list)

  ggres = reshape2::melt(res_low)
  names(ggres) = c("Sample_size", "Skewness","power")

  ggres$Estimate = cut(ggres$power, breaks =
  c(0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1),
  include.lowest = T, right = F)

  ggplot(ggres, aes(Sample_size, Skewness, fill= Estimate)) + geom_tile() +
  scale_fill_manual(values=c( "#F2F2F2", "#F0C9C0", "#EDB48E", "#EBB25E",
  "#E8C32E","#E6E600", "#A0D600", "#63C600", "#2DB600", "#00A600"))

##########

##### POWER skewt-TEST #####

library(ggplot2)
library(moments)
library(Rmisc)
library(tictoc)
library(dplyr)
library(RColorBrewer)
library(remotes)
library(future.apply)
library(future)

plan(multisession, workers = 12)

skewt.test <-
  function (x, mu = 0, conf.level = 0.95, b.frac = 1 / 4, N = 10^4, ...)
  {
    if (!missing(mu) && (length(mu) != 1 || is.na(mu)))
      stop("'mu' must be a single number")
    if (!missing(conf.level) &&
        (length(conf.level) != 1 || !is.finite(conf.level) ||
         conf.level < 0 || conf.level > 1))
      stop("'conf.level' must be a single number between 0 and 1")
    nx <- length(x)
    mx <- mean(x)
    vx <- var(x)
```

```
  if (nx < 2)
    stop("not enough 'x' observations")
  stderr <- sqrt(vx / nx)
  if (stderr < 10 * .Machine$double.eps * abs(mx))
    stop("data are essentially constant")
  df <- nx - 1
  tstat.calc <-
    function(n,S,gamma)
      sqrt(n) * (S + 1 / 3 * gamma * S ^ 2 + 1 / (6 * n) * gamma)
  S <- (mx - mu) / sqrt(vx)
  gamma <- sum((x - mx) ^ 3) / (nx * sqrt(vx) ^ 3)
  tstat <- tstat.calc(nx,S,gamma)
  nb <- max(nx * b.frac, 3) # Always sample at least 3
  # Calculate N boostrapped statistics, under the null hypothesis
  tsab <- replicate(N,{
    xb <- sample(x,nb,replace = T)
    m_xb <- mean(xb)
    sd_xb <- sd(xb)
    S_b <- (m_xb - mx) / sd_xb
    gamma_b <- sum((xb - m_xb) ^ 3) / (nb * sd_xb ^ 3)
    tstat.calc(length(xb),S_b,gamma_b)
  })
  # Remove NAs from the boostrapped vector
  tsab <- tsab[!is.na(tsab)] # (caused by constant subsamples from x)
  alpha <- 1 - conf.level
  tail <- if (tstat > 0)
    tsab > tstat
  else
    tsab <= tstat
  pval <- min(sum(tail) / (1 + N) * 2, 1)
  cint <- c(quantile(tsab,alpha / 2), quantile(tsab,1 - alpha / 2))
  sd <- sd(tsab)
  dname <- deparse(substitute(x))
  names(tstat) <- "t"
  names(sd) <- "sd"
  names(mu) <- "mean"
  attr(cint, "conf.level") <- conf.level
  method <- "Bootrsapped skewness-adjusted t-test (Lyon et al, 1999)"
  alternative <- "two.sided"
  rval <- list(
    statistic = tstat, parameter = sd, p.value = pval,
    conf.int = cint, estimate = mx, null.value = mu,
    alternative = alternative, method = method, data.name = dname
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
extract_p = function(K = 100, skew = 6.18, N_MC = 100)
{
  sk_x = function(x){ (x+2)*(x-1)^.5 }

  s2_root = uniroot( function(z) {sk_x(z) - skew} , interval = c(1,10e5))

  my_sdlog = s2_root$root %>% log() %>% sqrt()

  theor_mean = exp(0+0.5*my_sdlog**2)

  tmp_f = function()
  {
    A = rlnorm(K, sdlog = my_sdlog)
    tt = skewt.test(A,mu = (theor_mean - 0.5), N = 10^4)
    tt$p.value
  }

  res = future_replicate(tmp_f(), n = N_MC) %>% as.numeric()
  res
}

N_list = seq(from = 10, to = 310, by = 30)
sk_list = c(0.1,seq(from = 1, to = 21, by = 3))

i_N = length(N_list)
j_N = length(sk_list)
res_est = matrix(nrow = i_N, ncol = j_N)
res_low = matrix(nrow = i_N, ncol = j_N)
res_upp = matrix(nrow = i_N, ncol = j_N)

tic()
for (i in 1:i_N)
{
  for (j in 1:j_N)
  {
    tmp = extract_p(N_list[i],sk_list[j], N_MC = 10^4)
    test_res = binom.test(x= sum(tmp<0.05), n = length(tmp<0.05))
    res_low[i,j] = test_res$conf.int[1]
    res_upp[i,j] = test_res$conf.int[2]
    res_est[i,j] = test_res$estimate
  }
}
toc()

  colnames(res_low) = paste("Sk =",sk_list)
  rownames(res_low) = paste("N =",N_list)
```

```
ggres = reshape2::melt(res_low)
names(ggres) = c("Sample_size", "Skewness","power")

ggres$Estimate = cut(ggres$power, breaks =
c(0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1),
include.lowest = T, right = F)

ggplot(ggres, aes(Sample_size, Skewness, fill= Estimate)) + geom_tile() +
scale_fill_manual(values=c( "#F2F2F2", "#F0C9C0", "#EDB48E",
  "#EBB25E", "#E8C32E","#E6E600", "#A0D600", "#63C600", "#2DB600", "#00A600"))
```