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

КУРСОВАЯ РАБОТА

«Вычисление мощности некоторых статистических критериев методом Монте-Карло. Примеры»

Студентки 4-го курса 408-ой группы
кафедры теории вероятностей
Дьячковой Екатерины

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

МОСКВА 2021

Содержание

Введение

Предположение о нормальности распределения изучаемых данных лежит в основе большого количества статистических тестов. При использовании таких методов бывает важно проверить, подчиняется ли выборка нормальному закону. В разделе 1 мы приведем теоретическое описание наиболее распространённых критериев нормальности. Раздел 2 будет посвящен сравнению их мощности при помощи метода Монте-Карло. В разделе 3 мы рассмотрим пример двухэтапной процедуры, цель которой – поиск отличий между двумя независимыми выборками и на первом этапе которой проводится предварительное тестирование входных данных на нормальность. Используя результаты раздела 2, мы выберем соответствующий критерий нормальности для нашего примера и затем исследуем ошибку I рода всей двухэтапной процедуры. На практике широко используется тест Стьюдента. Его использование обусловлено доказанной асимптотической сходимостью статистики к нормальному распределению. В этой работе мы сделаем первый шаг к исследованию применимости теста Стьюдента, основываясь на результатах о скорости сходимости в центральной предельной теореме. Существуют разные варианты формулировки ЦПТ, все они, в том или ином виде, утверждают, что при определенных условиях сумма независимых или слабо зависимых случайных величин аппроксимируется нормальным распределением. На практике зачастую возникают вопросы о точности такой аппроксимации: какой размер выборки можно считать достаточно большим для того, чтобы было законно использование утверждения ЦПТ. Ряды Эджворта являются одним из методов повышения точности аппроксимации при помощи моментов высокого порядка. В разделе 4 мы приведем теоретическое описание рядов Эджворта. Для лучшего восприятия материала приведём вступительную теорию из [1].

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:

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

1 Теоретическое описание критериев нормальности распределения

1.1 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:

$$\begin{cases} H_0 : F(x) \in \mathcal{F} = \{\Phi_{\mu,\sigma^2}(x) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\} \\ H_1 : F(x) \notin \mathcal{F} \end{cases} \tag{3}$$

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

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - \Phi_{\hat{\mu},s^2}(x)| \tag{4}$$

где $F_n(x) = \frac{1}{n} \sum\limits_{i=1}^{n} I(X_i \leq x)$ — 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:

$$\Phi_{\hat{\mu},s^2}(x) = \Phi_{0,1}\left(\frac{x - \hat{\mu}}{s}\right) \tag{5}$$

Для индикаторов в эмпирической функции распределения выполнено:

$$\forall i = 1, \ldots, n : I(X_i \leq x) = I\left(\frac{X_i - \hat{\mu}}{s} \leq \frac{x - \hat{\mu}}{s}\right) \tag{6}$$

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

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - \Phi_{\hat{\mu},s^2}(x)| = \sup_{-\infty<y<+\infty} |\tilde{F}_n(y) - \Phi_{0,1}(y)| \tag{7}$$

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

$$\hat{\mu} \sim \frac{1}{n} N(n\mu, n\sigma^2) = N\left(\mu, \frac{\sigma^2}{n}\right); \quad S \sim \sigma^2 \frac{1}{n-1} \chi^2_{n-1} \tag{8}$$

$$\begin{aligned} \Rightarrow Y_i = \frac{X_i - \frac{1}{n}\sum X_i}{S_X} &\sim \frac{N_1(\mu, \sigma^2) - N_2\left(\mu, \frac{\sigma^2}{n}\right)}{\sqrt{\sigma^2 \frac{1}{n-1}\chi^2_{n-1}}} \sim \frac{N_1(0, \sigma^2) - N_2\left(0, \frac{\sigma^2}{n}\right)}{\sqrt{\sigma^2 \frac{1}{n-1}\chi^2_{n-1}}} \sim \\ &\sim \frac{\sigma N_1(0,1) - \sigma N_2\left(0, \frac{1}{n}\right)}{\sigma\sqrt{\frac{1}{n-1}\chi^2_{n-1}}} \sim \frac{N_1(0,1) - N_2\left(0, \frac{1}{n}\right)}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}} \sim \frac{Z_i - \frac{1}{n}\sum Z_i}{S_Z} \end{aligned} \tag{9}$$

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.

1.2 Критерий χ^2 Пирсона

Опираемся на теорию из [3]. Пусть дана выборка: $\mathcal{X} = (X_1, \ldots, X_n)$ с неизвестной функцией распределения $F(x)$. Фиксируем уровень значимости α, формулируем гипотезы:

$$\begin{cases} H_0 : F(x) \in \mathcal{F} = \{\Phi_{\mu,\sigma^2}(x) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\} \\ H_1 : F(x) \notin \mathcal{F} \end{cases} \tag{10}$$

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:

$$\begin{aligned} (\hat{\mu}, \hat{\sigma^2}) &= argmax_{(\mu,\sigma^2)} P\left(\nu_1 = l_1, ..., \nu_N = l_N\right) = \\ &= argmax_{(\mu,\sigma^2)} \frac{n!}{l_1! \cdot \ldots \cdot l_N!} \cdot [p_1(\mu, \sigma^2)]^{l_1} \cdot \ldots \cdot [p_N(\mu, \sigma^2)]^{l_N} = \\ &= argmax_{(\mu,\sigma^2)} \sum_{j=1}^{N} l_j \cdot \ln p_j(\mu, \sigma^2) \end{aligned} \tag{11}$$

Критериальная статистика выглядит следующим образом:

$$\chi_n^2 = \sum_{j=1}^{N} \frac{(Observed_j - Expected_j)^2}{Expected_j} = \sum_{j=1}^{N} \frac{\left(\nu_j - n \cdot p_j(\hat{\mu}, \hat{\sigma^2})\right)^2}{n \cdot p_j(\hat{\mu}, \hat{\sigma^2})} \tag{12}$$

По теореме Фишера, доказанной в 1924 году [1], если H_0 верна, $\chi_n^2 \xrightarrow[n\to\infty]{d} \chi^2(N-3)$.

1.3 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:

$$S = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{\left(\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^{3/2}}, \; K = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{\left(\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^2} - 3 \tag{13}$$

Интуитивно ΔK и ΔS можно изобразить как на рис.1, где синим цветом изображена плотность нормального распределения, а оранжевым цветом – плотность произвольного выборочного распределения:

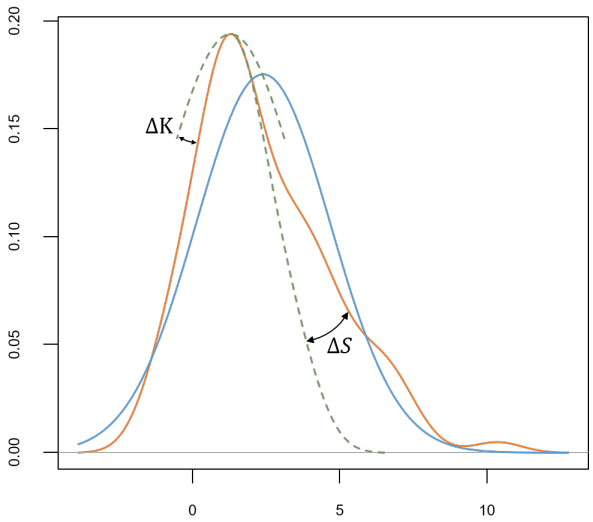

Рис.1: Идея критерия Харке – Бера.

Фиксируем уровень значимости α, формулируем гипотезы:

$$\begin{cases} H_0 : S = 0, K = 0 \\ H_1 : S \neq 0 \text{ и(или) } K \neq 0 \end{cases} \tag{14}$$

Критериальная статистика вычисляется следующим образом:

$$JB = n\left(\frac{S^2}{6} + \frac{K^2}{24}\right) \tag{15}$$

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

1.4 Критерий Шапиро – Уилка

Воспользуемся работой [12]. Пусть дана выборка: $\mathcal{X} = (X_1, \ldots, X_n)$ с неизвестной функцией распределения $F(x)$. Фиксируем уровень значимости α, формулируем гипотезы:

$$\begin{cases} H_0 : F(x) \in \mathcal{F} = \{\Phi_{\mu,\sigma^2}(x) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\} \\ H_1 : F(x) \notin \mathcal{F} \end{cases} \tag{16}$$

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:

$$EY_{(i)} = m_i,\ i = 1 \ldots, n. \tag{17}$$

$$cov(Y_{(i)}, Y_{(j)}) = v_{i,j},\ i, j = 1, \ldots, n. \tag{18}$$

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

$$X_{(i)} = \mu + \sigma \cdot Y_{(i)},\ i = 1, \dots n. \tag{19}$$

Согласно обобщенной теореме наименьших квадратов [3], а также того, что нормальное распределение симметрично, наилучшие линейные несмещенные оценки для μ и σ вычисляются следующим образом:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \hat{\sigma} = \frac{m^T V^{-1} \mathcal{X}}{m^T V^{-1} m} \tag{20}$$

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:

$$W = \frac{\left(\sum\limits_{i=1}^{n} a_i \cdot X_{(i)}\right)^2}{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2} = \left(\frac{R^2\hat{\sigma}}{C}\right)^2 \frac{1}{s^2} = \frac{b^2}{s^2} \tag{21}$$

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

$$(a_1, \dots, a_n) = \frac{m^T \cdot V^{-1}}{\sqrt{m^T \cdot V^{-1} \cdot V^{-1} \cdot m}} \tag{22}$$

$$R^2 = m^T V^{-1} m \tag{23}$$

$$C = \sqrt{m^T V^{-1} V^{-1} m} \tag{24}$$

Рассмотрим график нормальной вероятности на рис.2, который является частным случаем qq-plot для нормального распределения. По оси абсцисс откладываем $m_1, ..., m_n$, по оси ординат – $X_{(1)}, \dots, X_{(n)}$:

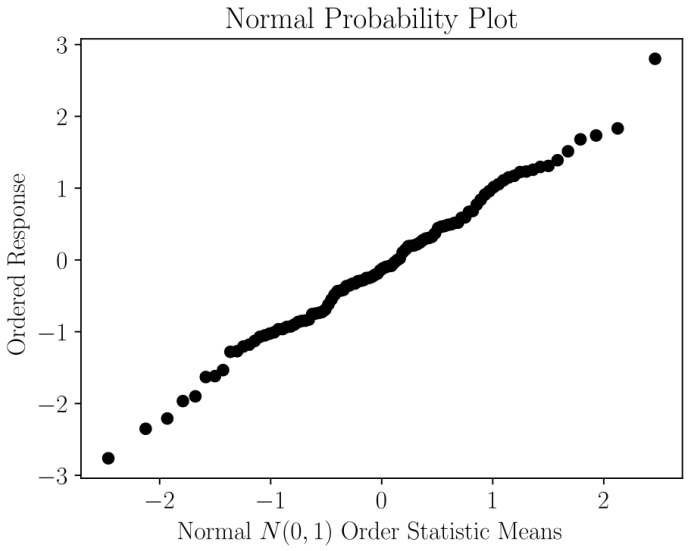

Рис.2: Идея критерия Шапиро – Уилка.

Используя обобщенный метод наименьших 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.

2 Исследование мощности наиболее распространенных критериев

2.1 Метод Монте-Карло

Статистическая гипотеза, в частности гипотеза проверки на нормальность, может иметь сложную структуру критической области, из-за чего аналитическое вычисление оценки мощности критерия может быть трудоемким. Поэтому для вычисления оценки мощности статистических тестов можно использовать метод Монте-Карло, описанный в [4] 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:

$$\forall i = 1, \ldots, K : \mathcal{X}_{H_1} \Rightarrow t_{H_1} \Rightarrow m_i = I(t_{H_1} \in G^*) \Rightarrow M = \sum_{i=1}^{K} m_i \Rightarrow W = \frac{M}{K} \qquad (25)$$

2.2 Эмпирическое исследование мощности

Для исследование мощности тестов на нормальность, рассмотренных в разделе 1, нами был использован метод Монте-Карло. Уровень значимости фиксировали $\alpha = 0.05$, далее генерировали $K = 10000$ раз выборки размера $n = 10, \ldots, 2000$ из четырёх рас-

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

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{26}$$

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

$$p(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} \exp -\frac{x}{\theta} \tag{27}$$

3.) Гамма с параметрами $k = 1$, $\theta = 5$;
4.) Распределение Лапласа с параметрами $\alpha = 1$ и $\beta = 0$, его плотность:

$$p(x) = \frac{\alpha}{2} \exp -\alpha|x-\beta| \tag{28}$$

Далее мы применяли метод Монте-Карло и вычисляли оценки мощности тестов. Результаты представлены на рис. 3-6:

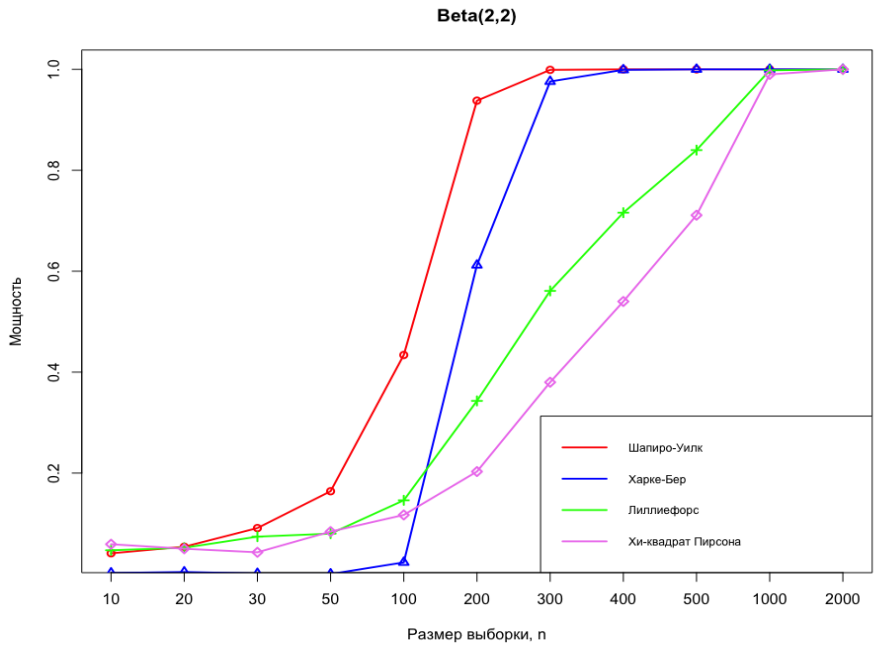

Рис.3 Сравнение мощности тестов на нормальность на распределении Бета(2,2).

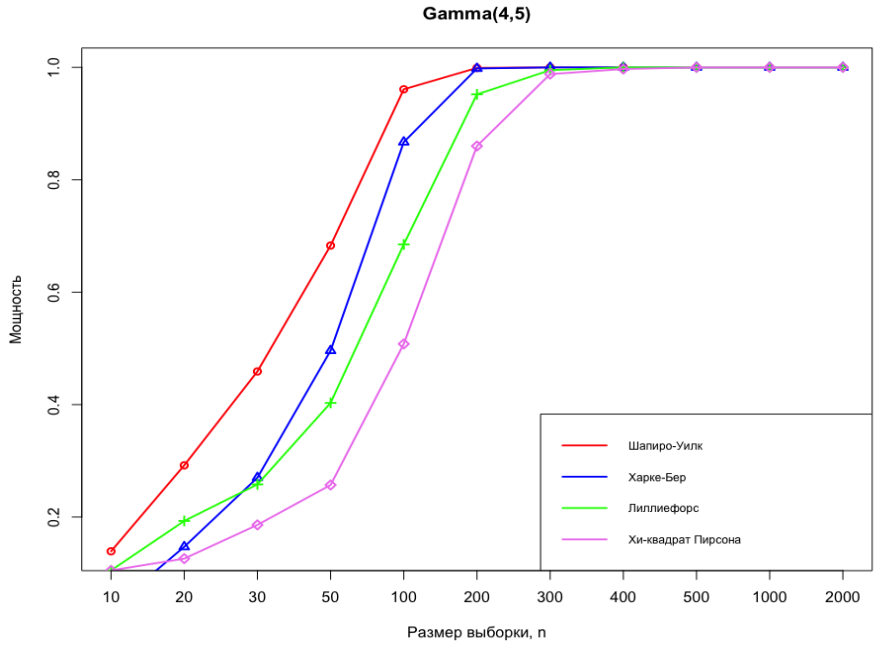

Рис.4 Сравнение мощности тестов на нормальность на распределении Гамма(4,5).

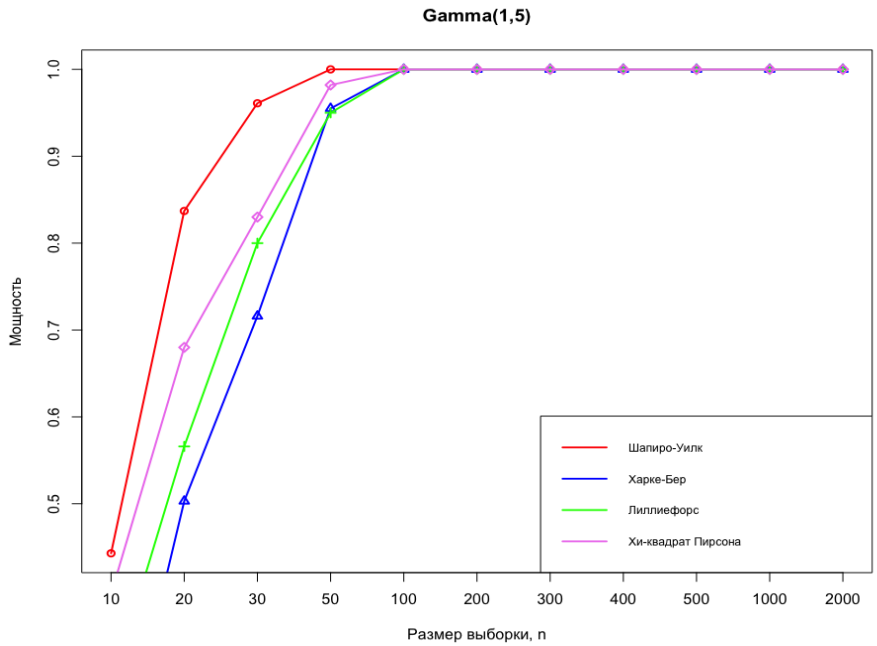

Рис.5 Сравнение мощности тестов на нормальность на распределении Гамма(1,5).

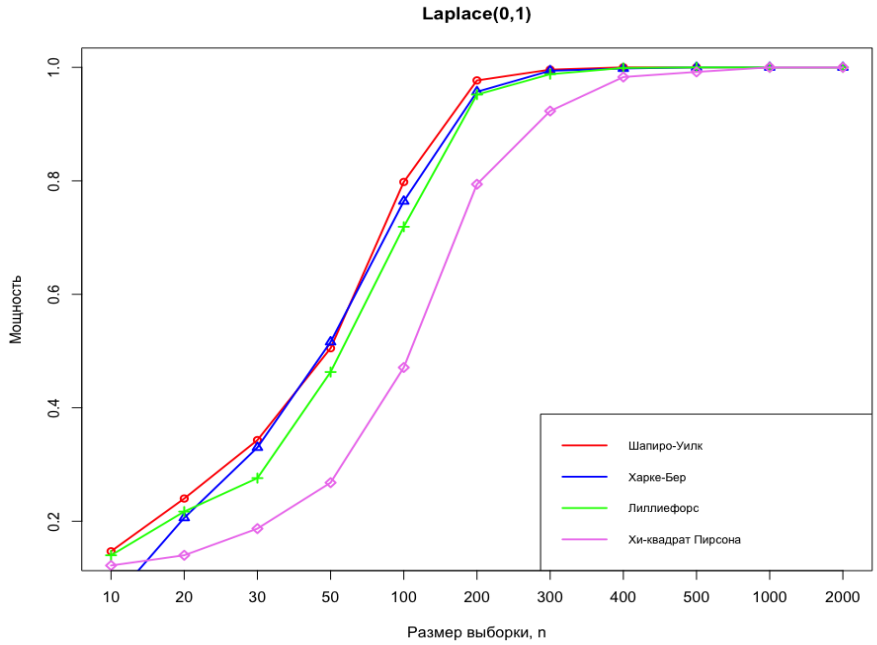

Рис.6 Сравнение мощности тестов на нормальность на распределении Лапласа(0,1).

Таким образом, из рассмотренных в разделе 1 критериев наибольшей мощностью на исследуемых распределениях обладает тест Шапиро-Уилка.

3 Исследование ошибки I рода двухэтапного тестирования

3.1 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:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \tag{29}$$

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 предыдущем разделе 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 (29).
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:

$$P\left(err_I\right) = P\left(err_I | SW \text{ отверг гипотезу о нормальности}\right) \cdot P\left(\text{отвержения } SW\right) + \\ + P\left(err_I | SW \text{ не отверг гипотезу о нормальности}\right) \cdot P\left(\text{неотвержения } SW\right) \quad (30)$$

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

3.2 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 (30), 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: $m_i = 0$;
3.) Получаем оценку $\widehat{P_{I,SW}} = \frac{\sum_{i=1}^{K} m_i}{K}$

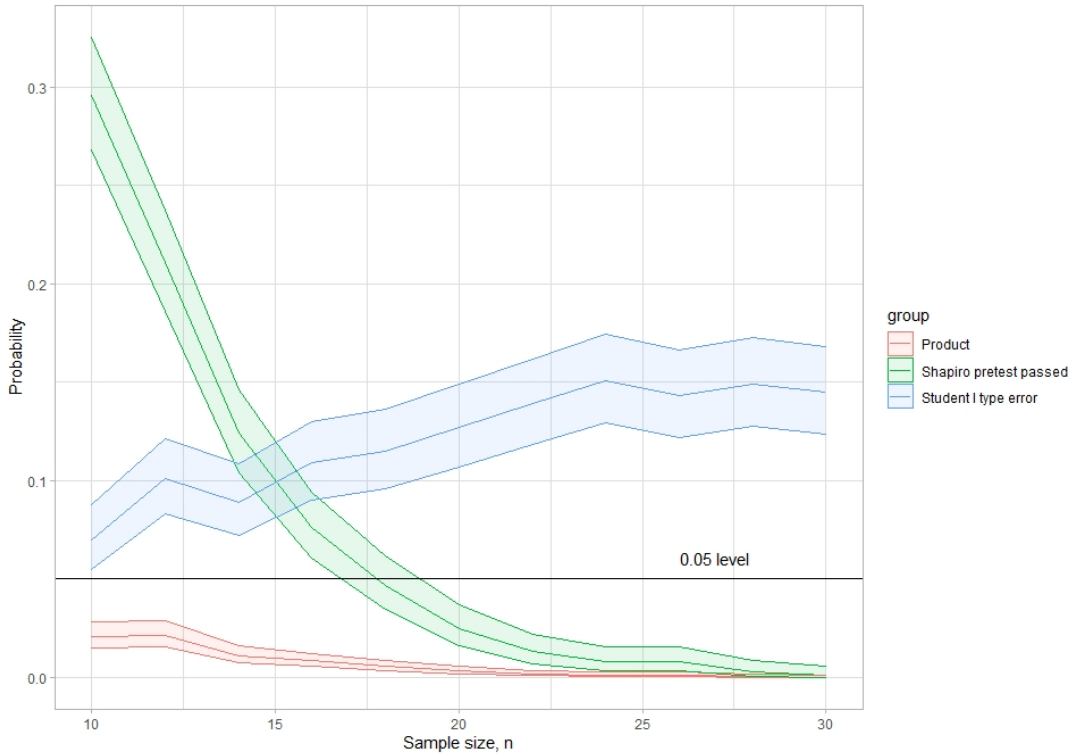

Рис.7 Уловная ошибка первого рода для критерия Стьюдента (синий), вероятность прохождения предварительного тестирования (зеленый) и их произведение (красный) с соответствующими доверительными интервалами.

Оценка условной вероятности $\widehat{P_{I,SW}}$ возрастает. При условии прохождения предварительного тестирования на нормальность на выборках размера $n = 25$ ошибка I рода критерия Стьюдента имеет порядок 15%, что в свою очередь втрое превышает заранее фиксированный уровень значимости. Однако оценка $\widehat{P_{SW}}$ с ростом n убывает так, что их итоговый вклад $\widehat{P_{I,SW}} \cdot \widehat{P_{SW}}$, оценивающий первый член разложения (30), наоборот, стремится к нулю с ростом n.

3.3 Исследование ошибки I рода непараметрического тестирования методом Монте-Карло

Уровни значимости для критерия Шапиро – Уилка и U-критерия Манна– Уитни аналогично разделу 3.2 фиксировали $\alpha_{pre} = 0.05$ и $\alpha = 0.05$. При помощи метода Монте-Карло мы построили график зависимости оценок для $P_{I,U}$, P_U и их произведения – второго члена разложения (30), от размера выборок. Модельное распределение и размеры выборок брали такие же, как и в разделе 3.2. 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 $\widehat{P_{I,U}} = \frac{\sum_{i=1}^{K} m_i}{K}$

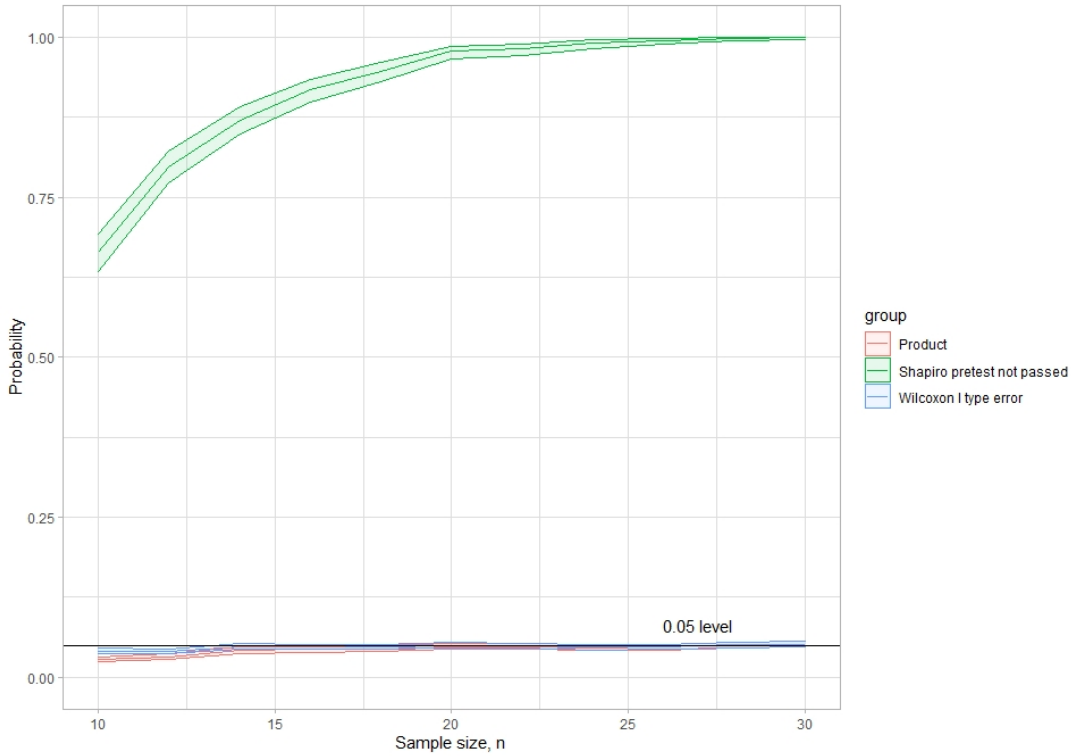

Рис.8 (а) Уловная ошибка первого рода для U-критерия Манна – Уитни (синий), вероятность отвержения предварительной гипотезы о нормальности (зеленый) и их произведение (красный) с соответствующими доверительными интервалами.

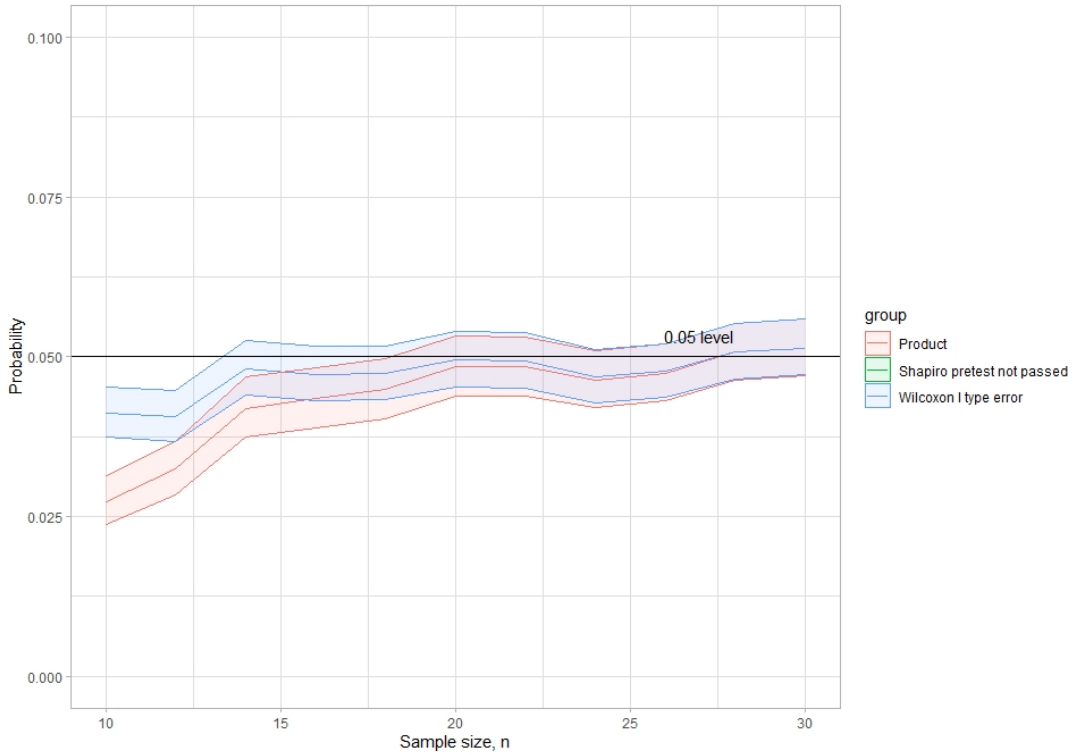

Рис.8 (б) Уловная ошибка первого рода для U-критерия Манна – Уитни (синий) и её произведение с вероятностью отвержения предварительной гипотезы о нормальности (красный) с соответствующими доверительными интервалами в увеличенном масштабе.

Оценка $\widehat{P_U}$ быстро возрастает к единице, поэтому итоговый вклад $\widehat{P_U} \cdot \widehat{P_{I,U}}$, оценивающий второй член разложения (30), начиная с некоторого n, практически не отличим от $\widehat{P_{I,U}}$. Как $\widehat{P_{I,U}}$, так и его произведение с $\widehat{P_U}$ с ростом n возрастают, но контролируются на уровне 0.05.

3.4 Исследование ошибки I рода двухэтапной процедуры методом Монте-Карло

Используя результаты, полученные в разделах 3.2 и 3.3, мы построили итоговый график зависимости оценок двух членов разложения (30) и общей оценки ошибки I рода всей двухэтапной процедуры от размера выборок. Результаты представлены на рис.9:

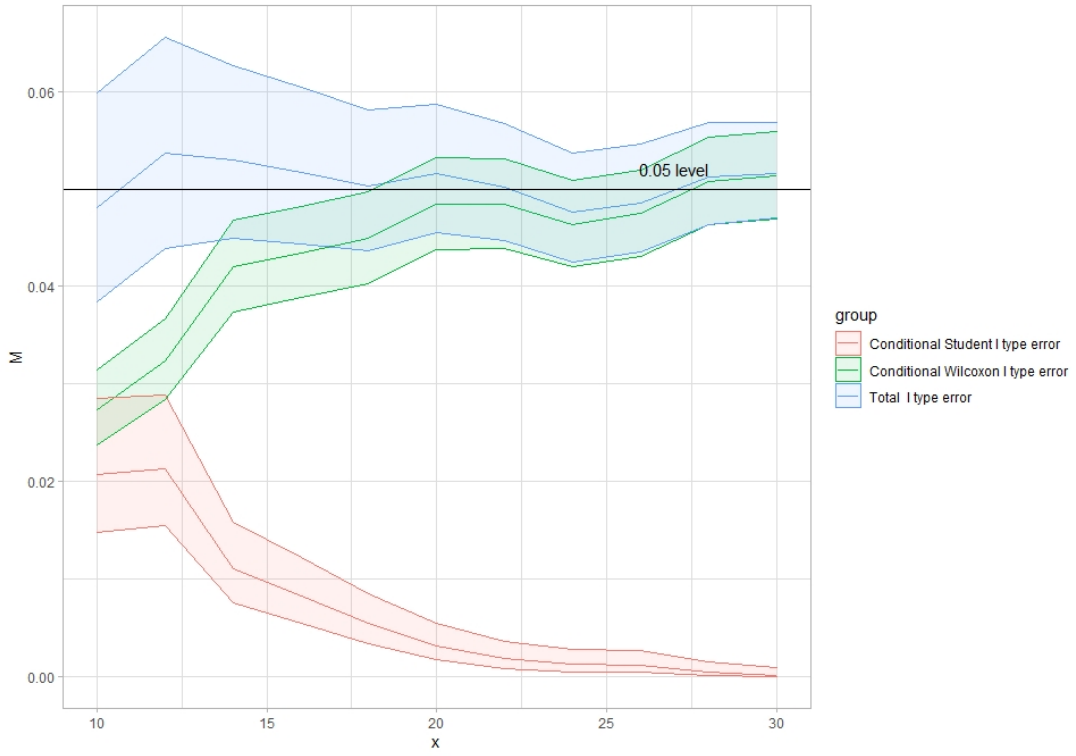

Рис.9 Оценки членов разложения ошибки I рода с соответствующими доверительными интервалами.

Несмотря на поведение оценок $\widehat{P_{I,SW}} \cdot \widehat{P_{SW}}$ и $\widehat{P_{I,U}} \cdot \widehat{P_U}$ двух членов разложения (30), видим, что итоговая оценка ошибки I рода контролируется.

4 Ряды Эджворта

Полученный в разделе 3 эффект требует объяснения. Для начала мы обратимся к рядам Эджворта. Опираясь на материал, изложенный в [3] и [9] – [11], приведем их теоретическое описание.

Теорема 1 (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.

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:

$$\phi_Z(t) = \left(\phi\left(\frac{t}{\sqrt{n}\sigma}\right)\right)^n \exp\left(-\frac{\sqrt{n}it\mu}{\sigma}\right) \tag{31}$$

Пусть семиинварианты любого порядка существуют. Тогда справедливо разложение:

$$\phi_Z(t) = \exp\left(\sum_{k=1}^{\infty} \frac{\kappa_k (it)^n}{k!}\right) \tag{32}$$

Используя (31), (32), а также разложение в ряд Тейлора для $\ln\phi\left(\frac{t}{\sqrt{n}\sigma}\right)$ в окрестности точки $t = 0$, получаем:

$$\ln\phi_Z(t) = n\ln\phi\left(\frac{t}{\sqrt{n}\sigma}\right) - \frac{\sqrt{n}it\mu}{\sigma} = n\sum_{i=2}^{\infty}\left(\frac{it}{\sqrt{n}\sigma}\right)^i \frac{\kappa_i(X_1)}{i!} =$$
$$= \frac{t^2}{2} + \frac{(it)^3\kappa_3(X_1)}{6\sqrt{n}\sigma^3} + \frac{(it)^4\kappa_4(X_1)}{24n\sigma^4} + O\left(n^{-3/2}\right) \tag{33}$$

Далее используем формулу обращения для преобразования Фурье и (33), получаем формулу для плотности Z:

$$f_Z(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itz)\phi_Z(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itz)\exp(\ln\phi_Z(t))dt =$$
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itz)\exp\left(\frac{t^2}{2} + \frac{(it)^3\kappa_3(X_1)}{6\sqrt{n}\sigma^3} + \frac{(it)^4\kappa_4(X_1)}{24n\sigma^4} + O\left(n^{-3/2}\right)\right)dt =$$
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itz)\exp\left(\frac{t^2}{2}\right)\left(1 + \frac{(it)^3\rho_3(X_1)}{6\sqrt{n}} + \frac{(it)^4\rho_4(X_1)}{24n} + \frac{(it)^6\rho_3^2(X_1)}{72n} + O\left(n^{-3/2}\right)\right)dt \tag{34}$$

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

$$\frac{d^r(exp(-itz))}{dz^r} = (-1)^r\exp(-itz)(it)^r \tag{35}$$

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

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itz)\exp\left(-\frac{t^2}{2}\right) = \varphi(z) \tag{36}$$

Меняем порядок дифференцирования и интегрирования [9], получаем:

$$f_Z(z) = \varphi(z)\left(1 + \frac{\rho_3(X_1)}{6\sqrt{n}}H_3(z) + \frac{\rho_4(X_1)}{24n}H_4(z) + \frac{\rho_3^2(X_1)}{72n}H_6(z) + O(n^{-3/2})\right) \tag{37}$$

Раскрываем полиномы Эрмита:

$$f_Z(z) = \varphi(z)\left(1 + \frac{\rho_3(X_1)}{6\sqrt{n}}(z^3 - 3z) + \frac{\rho_4(X_1)}{24n}(z^4 - 6z^2 + 3) + \right.$$
$$\left. + \frac{\rho_3^2(X_1)}{72n}(z^6 - 15z^4 + 45z^2 - 15) + O(n^{-3/2})\right. \tag{38}$$

Интегрируем (38), получаем выражение для функции распределения Z:

$$F_Z(x) = P(Z \le x) = \Phi(x) + \varphi(x)\left(\frac{\rho_3(X_1)}{6\sqrt{n}}(1 - x^2) + \frac{\rho_4(X_1)}{24n}(3x - x^3) + \right.$$
$$\left. + \frac{\rho_3^2(X_1)}{72n}(-x^5 + 10x^3 - 15x) + O(n^{-3/2})\right. \tag{39}$$

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

$$\rho_3(X_1) = \frac{E(X_1 - EX_1)^3}{\sigma^3} = \gamma_1, \tag{40}$$

$$\rho_4(X_1) = \frac{E(X_1 - EX_1)^4 - 3E(X_1 - EX_1)^2}{\sigma^4} = E(X_1 - EX_1)^4\sigma^4 - 3 = \gamma_2 \tag{41}$$

где γ_1 – коэффициент асимметрии X_1, γ_2 – коэффициент эксцесса X_1. Тогда функция распределения Z:

$$F_Z(x) = \Phi(x) + \frac{\varphi(x)}{\sqrt{n}}\left(\frac{\gamma_1}{6}(1-x^2)\right) + \\ + \frac{\varphi(x)}{n}\left(\frac{\gamma_2}{24}(3x-x^3) + \frac{\gamma_1^2}{72}(-x^5+10x^3-15x)\right) + O(n^{-3/2}) \quad (42)$$

Заключение

Как было отмечено в разделе 2, из рассмотренных в работе методов наибольшей мощностью на исследуемых распределениях обладал тест Шапиро-Уилка. Эти результаты также подтверждаются в работах [5] и [6]. Наше исследование двухэтапной процедуры в разделе 3 показало, что с формальной точки зрения предварительная проверка на нормальность неверна. Однако метод Монте-Карло для тех случаев, которые мы рассматривали, показал, что итоговая ошибка I рода все же контролируется. Таким образом, мы показали, что применение t-теста на практике оправдано. Мы не смогли найти значимого отклонения ошибки I рода от номинального уровня. В работе [2] получен схожий результат. Для дальнейшего исследования этого явления мы обратились к рядам Эджворта и привели необходимую теорию в разделе 4.

Список литературы

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

[2] Rochon J., Gondan M., Kieser M. To test or not to test: Preliminary assessment of normality when comparing two independent samples //BMC medical research methodology. – 2012. – Т. 12. – №. 1. – С. 1-11.

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

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

[5] Razali, Nornadiah Mohd, and Yap Bee Wah. "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests."Journal of statistical modeling and analytics 2.1 (2011): 21-33.

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

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

[10] Боровков А. А. Теория вероятностей. - 1999.

[11] Прохоров Ю.В., Розанов Ю. А. Теория вероятностей. Основные понятия. Предельные теоремы. Случайные продессы. – Наука, 1987.

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

Листинг программного кода

```
library(moments)
library(nortest)

K = 10000

gen_sample = function(name, N) #function for quick generation{
  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)


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")
```