Проверка статистических гипотез с помощью метода перестановок

Гребенюк Алексей Сергеевич, 622-я группа

Санкт-Петербургский Государственный Университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель — д.ф.-м.н. **В. Б. Мелас**
Рецензент — д.т.н. **Ю. Д. Григорьев**

Санкт-Петербург
2022г.

Задачи работы:

1. Эмпирическое сравнение мощности энергетического и модифицированного критериев. Асимптотическая оценка мощности модифицированного критерия.
2. Исследование зависимости мощности модифицированного критерия в зависимости от значения вспомогательного параметра k.
3. Сравнение модифицированного и энергетического критериев с классическими тестами (t-Стьюдента, Колмогорова-Смирнова, Уилкоксона-Манна-Уитни).

Постановка задачи

Задача проверки гипотезы о равенстве двух распределений

$$H_0 : F_1 = F_2$$

против альтернативы

$$H_1 : F_1 \neq F_2,$$

где $X = (X_1, \ldots, X_n)$ и $Y = (Y_1, \ldots, Y_n)$ с функциями распределения F_1 и F_2 соответственно.

Постановка задачи

Предположим, что функции распределения F_1 и F_2 принадлежат к такому классу распределений, что случайные величины ξ

- имеют плотность, симметричную относительно некоторой точки;
- соответствуют свойству

$$\mathbb{E}[\ln^2(1+\xi^2)] < \infty.$$

Рассматриваемые распределения:

1. нормальное распределение;
2. распределение Лапласа;
3. распределение Коши.

Статистика энергетического теста (Aslan, Zech, 2005) имеет общий вид

$$\Phi_{n,m}(X,Y) = \Phi_{AB} + \Phi_A + \Phi_B,$$

$$\Phi_{AB} = -\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} g(X_i - Y_j),$$

$$\Phi_A = \frac{1}{n^2}\sum_{i<j}^{n} g(X_i - X_j),$$

$$\Phi_B = \frac{1}{m^2}\sum_{i<j}^{n} g(Y_i - Y_j).$$

Энергетическое расстояние

$$g(x) = \ln(|x|).$$

Энергетическое расстояние

$$g(x) = \ln(|x|).$$

Модифицированное расстояние (Мелас, 2021)

$$g(x) = \ln\left(1 + (kx)^2\right).$$

В работе (Мелас, 2022) найдены асимптотические формулы для широкого класса функций, в т.ч. для данной функции.

Обозначим интегралы (Мелас, 2022). Пусть $f(x)$ — это плотность F_1, тогда

$$J(h) = \int_{\mathbb{R}} -g\left(x - y - \frac{|h|}{\sqrt{n}}\right) f(x)f(y)dxdy,$$

$$J_1 = J(0) = \int_{\mathbb{R}} -g(x-y)f(x)f(y)dxdy,$$

$$J_2 = \int_{\mathbb{R}} g^2(x-y)f(x)f(y)dxdy,$$

$$J_3 = \int_{\mathbb{R}} g(x-y)g(x-z)f(x)f(y)f(z)dxdydz.$$

Обозначим через $J^*(h) = \lim_{n\to\infty} n(J(h) - J(0))$ существование предела, доказанно в работе (Мелас, 2021). Рассмотрим коэффициенты

$$\bar{b} = \sqrt{\frac{J^*(h)}{h^2}},$$

$$a^2 = \sqrt{J_2 + J_1^2 - 2J_3},$$

$$c = J_1 - a^2.$$

Теорема (Мелас, 2022)

Пусть $F_1(x) = F(x)$, $F_2 = F(x + \frac{h}{\sqrt{n}})$, *где* F *— функция распределения общего вида, симметричная относительно некоторой точки и обладающая свойством* $\mathbb{E}[\ln^2(1+\xi^2)] < \infty$, h *— произвольно заданное число. Тогда*

- *При верной* H_0 $nT_n \xrightarrow{n\to\infty} (aL)^2 + c$, *где* $L \sim N(0,1)$,
- *При верной* H_1 $nT_n \xrightarrow{n\to\infty} (aL+b)^2 + c$, *где* $L \sim N(0,1)$, $a^2 = \sqrt{J_2 + J_1^2 - 2J_3}$, $c = J_1 - a^2$.

Кроме того функция распределения nT_n *сходится при* H_1 *к функции распределения случайной величины*

$$(aL+b)^2 + c,$$

где $b = \bar{b}h$.

Теорема (Мелас, 2022)

В этом случае асимптотическая мощность критерия T_n с уровнем значимости α асимптотически равна

$$\mathbb{P}\left\{L \geq z_{1-\alpha/2} - \frac{\bar{b}h}{a}\right\} + \mathbb{P}\left\{L \leq -z_{1-\alpha/2} - \frac{\bar{b}h}{a}\right\} =$$
$$= 1 - \Phi\left(z_{1-\alpha/2} - \frac{\bar{b}h}{a}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\bar{b}h}{a}\right),$$

где $z_{1-\alpha/2}$ — это $(1-\alpha/2)$-квантиль нормального распределения:

$$\mathbb{P}\left\{L \geq z_{1-\alpha/2}\right\} = \alpha/2.$$

Для расчета эмпирических мощностей критериев написана программа на Python. Алгоритм перестановочного метода:

- $Z(\pi_0)$ — изначальная выборка;
- $Z(\pi_r)$ — выборка после перестановки;
- $r_2 = 800$ — общее число перестановок;
- r_1 — число перестановок π_r, для которых $K_i(Z(\pi_r)) > K_i(Z(\pi_0))$.

Если $\frac{r_1}{r_2} > \alpha$, то нулевая гипотеза H_0 не отвергается.
Если $\frac{r_1}{r_2} < \alpha$, то нулевая гипотеза H_0 отвергается.

Исследуются распределения со стандартным значением параметра масштаба и различными параметрами сдвига альтернативного распределения для размера выборок одинакового размера $n = 100$. Рассмотренные распределения:

- нормальное распределение;
- распределени Лапласа;
- распределени Коши.

Полученные результаты:

- Эмпирические мощности энергетического критерия и модифицированного критерия при $k = 1$

$$g(x) = \ln\left(1 + (kx)^2\right);$$

- Асимптотическая мощность для модифицированного критерия.

Мощности энергетического (E) и модифицированного (M) критериев для разных параметров сдвига альтернативного распределения $\dfrac{h}{\sqrt{n}}$, $n = 100$.

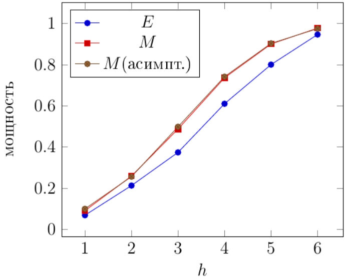

Рис. 1: Эмпирические и асимптотические мощности для **норм. распределения**

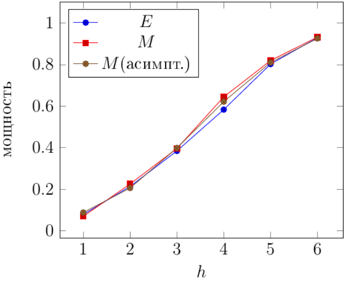

Рис. 2: Эмпирические и асимптотические мощности для **распределения Лапласа**

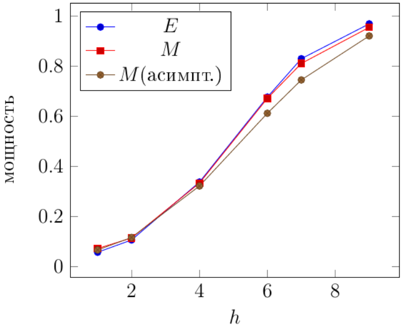

Рис. 3: Эмпирические и асимптотические мощности для **распределения Коши**

Продемонстрировано преимущество модифицированного критерия над энергетическим. Асимптотические формулы для модифицированного критерия хорошо аппроксимируют эмпирические мощности для нормального распределения и распределения Лапласа.

2. Исследование зависимости мощности модифицированного критерия от значения параметра k

Проведено исследование зависимости мощности модифицированного критерия от значения параметра k в случае неизвестного распределения. Модифицированное расстояние

$$g(x) = \ln\left(1 + (kx)^2\right).$$

Эффективность критерия $eff(k) = \dfrac{\hat{\hat{b}}}{a}$. Согласно теореме от $eff(k)$ монотонно зависит асимптотическая мощность модифицированного критерия.

2. Исследование зависимости мощности модифицированного критерия от значения параметра k

Эффективность модифицированного критерия $eff(k) = \frac{\hat{\bar{b}}}{a}$ для разных параметров k.

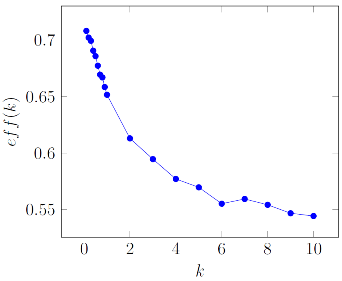

Рис. 4: Эффективность модифицированного критерия для **норм. распределения**

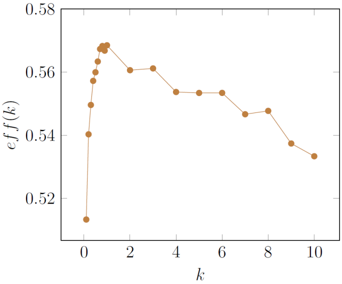

Рис. 5: Эффективность модифицированного критерия для **распределения Лапласа**

2. Исследование зависимости мощности модифицированного критерия от значения параметра k

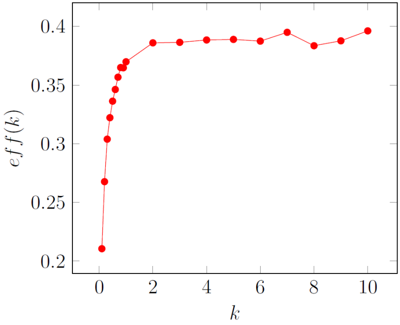

Рис. 6: Эффективность модифицированного критерия для **распределения Коши**

2. Исследование зависимости мощности модифицированного критерия от значения параметра k

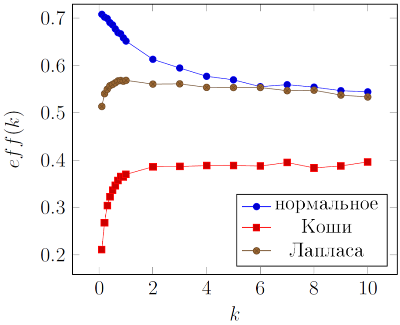

Рис. 7: Эффективность модифицированного критерия для **нормального распределения, распределения Коши, распределения Лапласа**

Если распределение неизвестно, то стандартное значение параметра $k = 1$ является хорошим компромиссом.

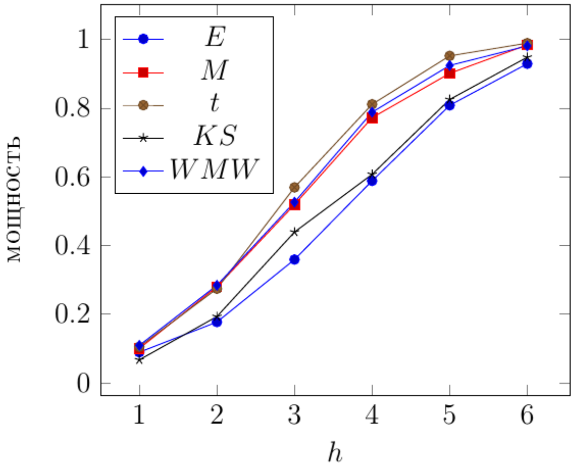

Рис. 8: Эмпирические мощности критериев E, M, t, KS, WMW для **нормального распределения**

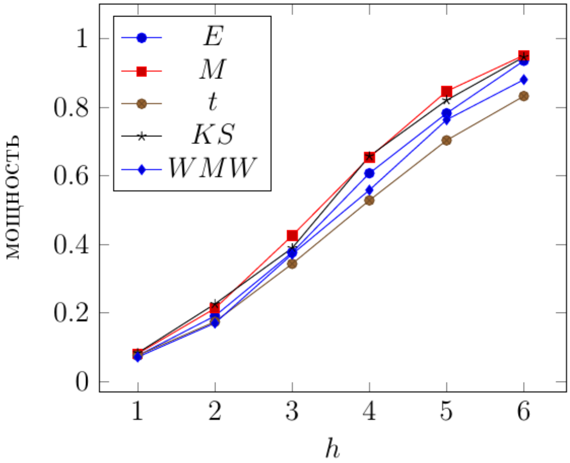

Рис. 9: Эмпирические мощности критериев E, M, t, KS, WMW для **распределения Лапласа**

3. Сравнение с классическими критериями

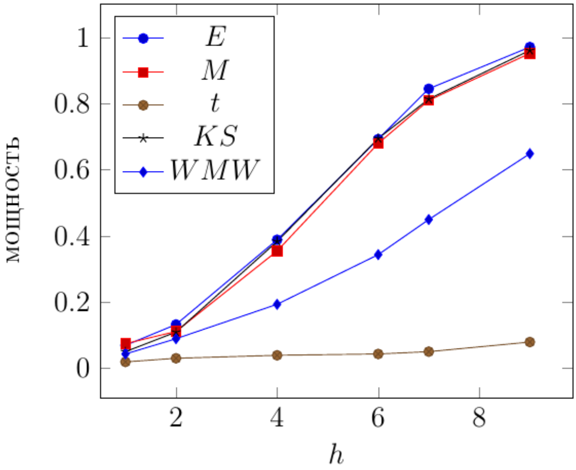

Рис. 10: Эмпирические мощности критериев E, M, t, KS, WMW для **распределения Коши**

Заключение

1. Для **нормального распределения** и **распределения Лапласа** модифицированный тест превосходит энергетический во всех рассмотренных случаях, а для **распределения Коши** мощности близки. Показано, что асимптотические формулы хорошо аппроксимируют эмпирические мощности для модифицированного критерия;
2. Для модифицированного критерия в случае неизвестного распределения хорошим компромиссом значения параметра является стандартное значение $k = 1$;
3. Модифицированный критерий не уступает лучшему из классических критериев во всех рассмотренных случаях.