Исследование мощности перестановочных тестов для проверки гипотез о равенстве двух распределений

Анна Андреевна Белкова, группа 17.Б04-мм

Санкт-Петербургский государственный университет
Кафедра статистического моделирования

Научный руководитель: д.ф.-м.н., проф. Мелас В. Б.

Рецензент: д.т.н., проф. Григорьев Ю. Д.

Санкт-Петербург
2021г.

Краткая постановка задачи

Недавно [Melas, Salnikov 2020] был предложен новый тест для проверки гипотезы о равенстве двух распределений.

С помощью стохастического моделирования было показано, что тест дает более хорошие результаты, чем многие другие.

Для случая распределений, отличающихся только параметром сдвига, была найдена асимптотическая формула для мощности теста.

Задача

Проведение численных исследований нового теста.

Статистика нового критерия [Melas, Salnikov 2020]

Пусть $X_1, ..., X_n$ и $Y_1, ..., Y_m$ — две независимые выборки с функциями распределения соответственно F_1 и F_2, причем F_1, F_2 принадлежат одному классу распределений, но могут отличаться сдвигом или масштабом.
$H_0 : F_1(x) = F_2(x)$ для любого $x \in \mathbb{R}$,
$H_1 : F_1(x) \neq F_2(x)$ для хотя бы одного $x \in \mathbb{R}$.

Статистика теста

$$T_{nm} = \Phi_A + \Phi_B + \Phi_{AB},$$

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

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

где $g(|u|) = \ln(1 + |u|^2)$.

Перестановочные методы

Для численного исследования теста и проверки выведенных формул авторы [M., S. 2020] использовали перестановочный (permutation) метод нахождения распределения статистики критерия при верной H_0.

Пусть мы имеем:

- Две независимые выборки: $X = (X_1, ..., X_n)$ и $Y = (Y_1, ..., Y_n)$;
- Гипотезу о равенстве двух распределений, например: $H_0 : F(x) = G(x)$ для любого $x \in \mathbb{R}$, $H_1 : F(x) \neq G(x)$ для хотя бы одного $x \in \mathbb{R}$;
- Статистику критерия $\varphi = \varphi(X, Y)$.

Хотим найти критическую область значений статистики критерия.

Перестановочные методы

Идея

Если H_0 верна, то элементы выборок X и Y можно воспринимать как элементы из одной выборки. В этом случае, если поменять в выборках X и Y некоторые элементы X_i и Y_j местами, получатся новые выборки X' и Y', статистика критерия от которых не будет отличаться от $\varphi(X, Y)$.

Алгоритм [Efron, Tibshirani 1993]:

- Объединим X и Y в единую выборку Z;
- Найдем K случайных разбиений Z на $(X^{(i)}, Y^{(i)})$, таких что: $\#X^{(i)} = \#Y^{(i)}$, $X^{(i)} \cup Y^{(i)} = Z$, $X^{(i)} \cap Y^{(i)} = \varnothing$;
- Найдем $ASL_{perm} = \frac{\#\{\varphi(X,Y) \leq \varphi(X^{(i)},Y^{(i)}), i=1,..,K\}}{K}$ – достигнутый уровень значимости теста;
- Сравним выбранный нами уровень значимости α с ASL_{perm}.

Теорема о мощности нового теста [M., S. 2020]

Введем обозначения:
Пусть $f_1(x)$ и $f_2(x)$ плотности, соответствующие ф-ям распр. F_1 и F_2, где $F_2 = F_1(x-\theta)$, $\theta = h/\sqrt{n}$. Обозначим

$$J_h = \int_R g(x-y) f_1(x) f_2(y) dx dy,$$

$$J^*(h) = \lim_{n\to\infty} n(J_h - J_0),$$

$$\bar{b} = \sqrt{J^*(h)/h^2}.$$

Теорема. Часть 1

Рассмотрим поставленную выше задачу тестирования гипотезы о равенстве двух распределений.
(1) при верной H_0 и $n \to \infty$ распределение nT_n стремится к распределению с.в. $(aZ)^2$, где $Z \sim N(0,1)$; $a \in \mathbb{R}, a > 0$.

Теорема о мощности нового теста [M., S. 2020]

Теорема. Часть 2

(2) Положим $F_1 = F(x)$, $F_2 = F(x+\theta)$, где F функция распределения, которая соответствует некоторому симметричному распределению, обладающему свойством $E(\ln^2(1+\xi^2)) < \infty$, $\theta = h/\sqrt{n}$, h — произвольное число. Тогда распределение nT_n при $n \to \infty$ стремится к распределению случайной величины

$$(aZ+b)^2,$$

где $Z \sim N(0,1)$; $a, b \in \mathbb{R}$, $b = 0$, если верна H_0 и $b = \bar{b}h$ при верной H_1. И в этом случае мощность критерия T_n при уровне значимости α асимптотически эквивалентна величине, получаемой по формуле

$$Pr\{Z \geq z_{1-\alpha/2} - \bar{b}h/a\} + Pr\{Z \leq -z_{1-\alpha/2} - \bar{b}h/a\},$$

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

Оценки параметров b, a, $\bar{b}/a$

В ходе ВКР получены оценки параметра $\bar{b}/a$ для случаев нормального распределения, распределения Коши и Лапласа. Это оценки по МНК, то есть

$$\arg\min_{\bar{b}/a} \sum_{i=1}^{s} (Q_{h_i}(\bar{b}/a) - \widetilde{Q}_{h_i}(\bar{b}/a))^2,$$

где Q — мощность по формуле из Теоремы 1, $\widetilde{Q}$ — эмпирическая мощность.

Результат

- Для нормального распределения $\widehat{\bar{b}/a} = 0.651$;
- Для распределения Коши $\widehat{\bar{b}/a} = 0.339$;
- Для распределения Лапласа $\widehat{\bar{b}/a} = 0.558$;

Работа программы

- В ходе работы программы N раз генерируются 2 выборки заданных распределений (например, $Cauchy(0, 1)$ и $Cauchy(0, 1 + h/\sqrt{n})$);
- Для каждой пары считается значение нового теста (с помощью перестановок) и тестов Колмогорова – Смирнова, Андерсона – Дарлинга и Уилкоксона – Манна – Уитни, и сохраняется информация о том, отвергается ли гипотеза о равенстве распределений;
- После N итераций подсчитывается мощность каждого из тестов против заданной альтернативы.

Результаты работы программы. Таблицы

Таблица: Объем выборок $n = 100$, уровень значимости $\alpha = 0.05$. Для моделирования соответствующей H_1 ситуации использовались распределения $Cauchy(0, 1)$ и $Cauchy(0, 1 + h/\sqrt{n})$

h	Power (T_n)	Power (K-S test)	Power (A-D test)	Power (W-M-W test)
0	0.052	0.045	0.048	0.052
2	0.104	0.059	0.066	0.054
4	0.208	0.098	0.124	0.060
5	0.403	0.115	0.146	0.061
6	0.394	0.159	0.232	0.065
8	0.692	0.261	0.367	0.071
10	0.899	0.365	0.528	0.064

Результаты работы программы. Таблицы

Таблица: Объем выборок $n = 100$, уровень значимости $\alpha = 0.05$. Для моделирования соответствующей H_1 ситуации использовались распределения $N(0, 1)$ и $N(0, 1 + h/\sqrt{n})$

h	Power (T_n)	Power (K-S test)	Power (A-D test)	Power (W-M-W test)
0	0.051	0.039	0.048	0.050
4	0.461	0.174	0.399	0.060
6	0.803	0.354	0.770	0.062
8	0.983	0.609	0.940	0.051
10	0.999	0.800	0.994	0.066

Результаты работы программы. Таблицы

Таблица: Объем выборок $n = 100$, уровень значимости $\alpha = 0.05$. Для моделирования соответствующей H_1 ситуации использовались распределения $Cauchy(0, 1)$ и $Cauchy(h/\sqrt{n}, 1)$

h	Power (T_n)	Power by formula	Power (K-S test)	Power (A-D test)	Power (W-M-W test)
0	0.045	0.050	0.050	0.049	0.052
2	0.084	0.104	0.101	0.118	0.108
4	0.274	0.274	0.335	0.384	0.342
6	0.544	0.530	0.694	0.623	0.639
8	0.810	0.774	0.928	0.888	0.851
10	0.941	0.924	0.980	0.972	0.953

Результаты работы программы. Таблицы

Таблица: Объем выборок $n = 100$. Для моделирования соответствующей H_1 ситуации использовались распределения $Cauchy(0, 1)$ и $Logistic(0, 1)$

α	Power (T_n)	Power (K-S test)	Power (A-D test)
0.01	0.137	0.014	0.015
0.02	0.189	0.030	0.039
0.05	0.363	0.058	0.115
0.1	0.497	0.114	0.230

Результаты работы программы. График

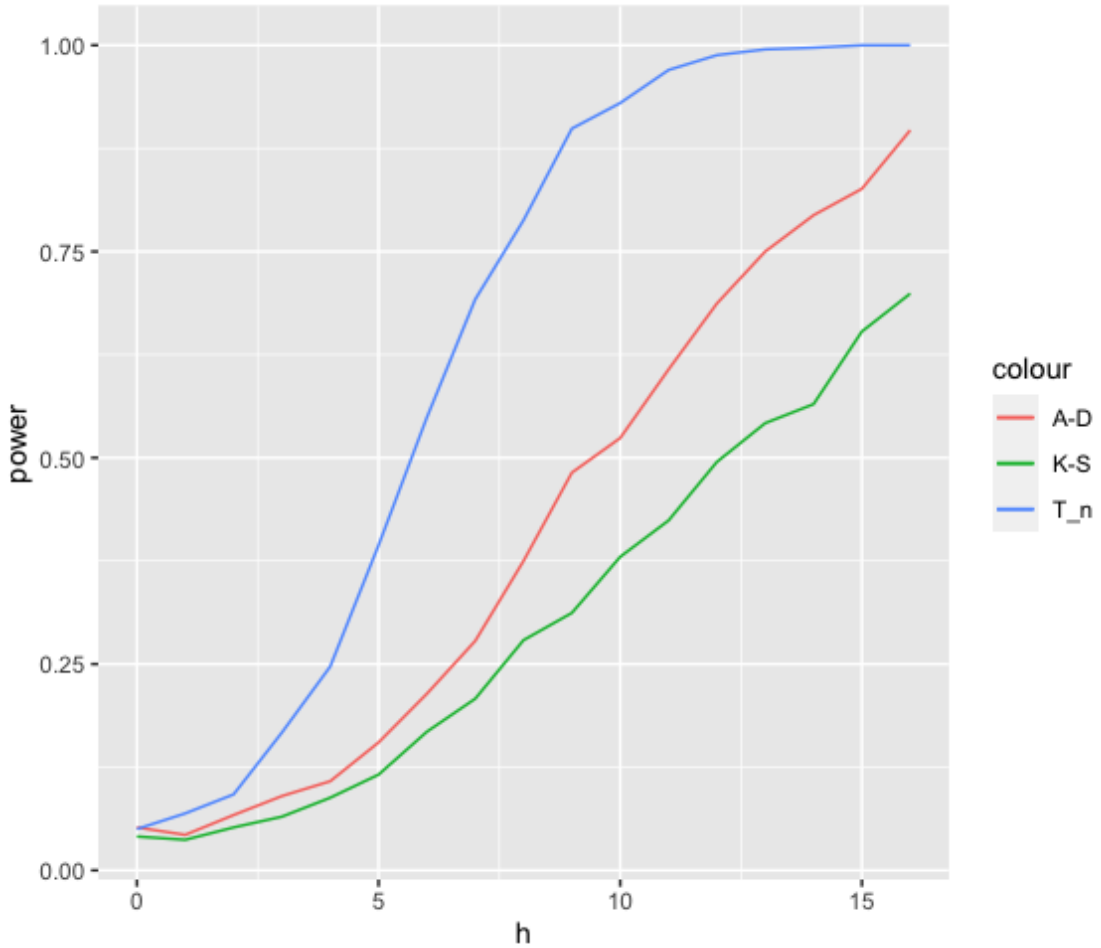

Рис.: Сравнение мощностей в зависимости от параметра h. Объем выборок $n = 100$, $\alpha = 0.05$. H_1: $X \sim Cauchy(0, 1)$ и $Y \sim Cauchy(0, 1 + h/\sqrt{n})$

Результаты работы программы. График

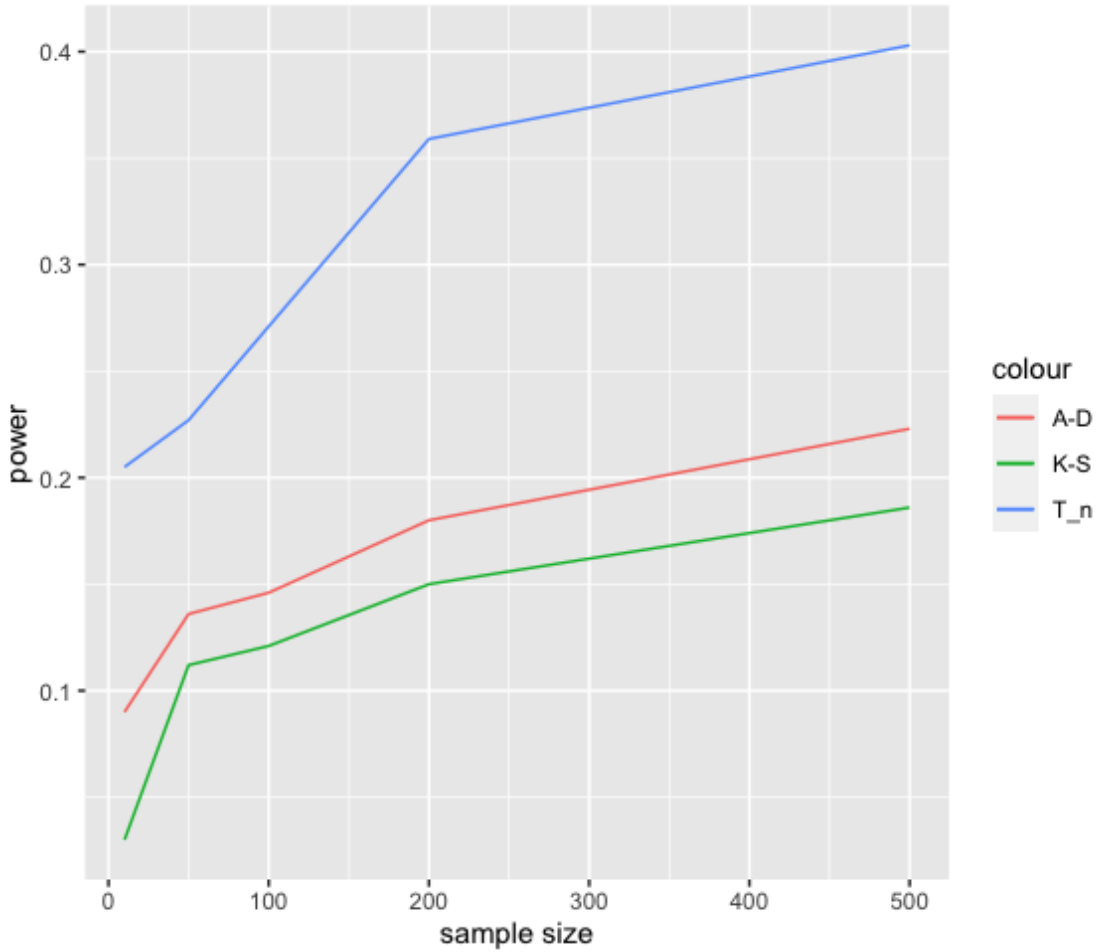

Рис.: Сравнение мощностей в зависимости от объема выборок. Параметр $h = 5$, $\alpha = 0.05$. H_1: $X \sim Cauchy(0, 1)$ и $Y \sim Cauchy(0, 1 + h/\sqrt{n})$

Итоги работы

- Изучены перестановочные методы нахождения критической и доверительной области значений статистики критерия;
- Изучена работа, посвященная новому тесту;
- Получены оценки параметра $\bar{b}/a$ для случаев нормального распределения, распределения Коши и Лапласа;
- Реализована программа на R, позволяющая эмпирически проверить аналитические результаты исследования мощности нового теста, собраны результаты численных исследований теста;
- Численными результатами подтверждена высокая эффективность метода, в особенности в случае распределений с тяжелыми хвостами, отличающихся в параметре масштаба.