Численное и аналитическое исследование мощности «энергетического» теста проверки гипотез.

Лобанова Полина Юрьевна

гр. 21.M03-мм
Санкт-Петербургский государственный университет
Математическое моделирование программирование и искусственный интеллект
Кафедра статистического моделирования
Научный руководитель: профессор Мелас Вячеслав Борисович
Рецензент: профессор Григорьев Юрий Дмитриевич

2023

Актуальность темы

Задача проверки гипотезы о равенстве двух распределений является классической задачей математической статистики и для её решения предложено значительное число различных методов.
Однако, t–критерий, требуют предположения о нормальности распределений, а универсальный критерии часто имеет низкую мощность.

«Энергетический» критерий

- «Энергетический» критерий для проверки статистических гипотез о равенстве двух распределений был предложен в работе (Аслан, Цех, 2005) и модифицирован в работе (Мелас, Сальников, 2021).
- Также в работе (Мелас,2023) введена формула для асимптотической мощности критерия в случае, когда распределение различаются только параметром сдвига.

Цели работы

- Разработать математическое и программное обеспечение для численного сравнения мощности «энергетического» метода с альтернативными классическими критериями.
- Провести численное сравнение мощности «энергетического» и альтернативных критериев.
- Разработать методы вычисления асимптотической мощности по формуле из статьи (Мелас, 2023). Исследовать применимость формулы.

Краткий обзор предшествующих результатов

- I.D. Reid, R.H.C. Lopes and P.R. Hobson. Comparison of Two-Dimensional Binned Data Distributions Using the Energy Test // CMS Note – 2008
- I.D. Reid, R.H.C. Lopes and P.R. Hobson. Non-parametric comparison of histogrammed two- dimensional data distributions using the Energy Test // Journal of Physics: Conference Series – 2012
- Cheng Huang and Xiaoming Huo. An Efficient and Distribution-Free Two-Sample Test Based on Energy Statistics and Random Projections // – 2017

Постановка задачи

Рассмотрим классическую задачу проверки гипотезы о равенстве двух распределений

$$H_0 : F_1 = F_2 \tag{1}$$

против альтернативы

$$H_1 : F_1 \neq F_2 \tag{2}$$

в случае двух независимых выборок $X = (X_1, \ldots, X_n)$ и $Y = (Y_1, \ldots, Y_m)$ с функциями распределения F_1 и F_2 соответственно, принадлежащим классу функций распределений случайных величин ξ и $g(\xi) = \ln(1 + |\xi|^2)$ таких, что

$$E[g(\xi)^2] < \infty. \tag{3}$$

Постановка задачи

Будем рассматривать случай, когда распределения различаются только одним параметром, а именно либо параметром сдвига: $F_2(x) = F_1(x - v_1)$, либо параметром масштаба: $F_2(x) = F_1(v_2 x)$. Положим $v_1 = h_1/\sqrt{n}$, $v_2 = 1 + h_2/\sqrt{n}$.
Тогда при $n \to \infty$ мощность критерия стремится к некоторому пределу, который назовем асимптотической мощностью.

Теорема [Мелас 2023]

Рассмотрим задачу проверки гипотезы (1)-(2), где обе функции обладают свойством (3) и имеют плотности распределения симметричные относительно некоторой точки. Тогда
(i) при условии $n \to \infty$ функция распределения nT_n сходится при H_0 к функции распределения случайной величины

$$(aL)^2 + c, \tag{4}$$

где L - случайная величина, которая имеет стандартное нормальное распределение,

$$c = J_1 - a^2, a^2 = \sqrt{J_2 + J_1^2 - 2J_3}. \tag{5}$$

Теорема [Мелас 2023]

(ii) Пусть $F_1(x) = F(x), F_2(x) = F(x(1 + h_2/\sqrt{n}) + h_1/\sqrt{n})$, где F — произвольная функция распределения,с плотностью $f(x)$ и обладающая свойством (3), h_1, h_2 - произвольные заданные числа .
Тогда функция распределения nT_n сходится при выполнении гипотезы H_1 к распределению случайной величины

$$(aL + b)^2 + \rho(h_1, h_2)L + c. \qquad (6)$$

Мощность критерия nT_n с уровнем значимости α приближённо равна

$$Pr\{L \geq z_{1-\alpha/2} - b/a\} + Pr\{L \leq -z_{1-\alpha/2} - b/a\}, \qquad (7)$$

где $b = \sqrt{b_1^2 + b_2^2}$, $z_{1-\alpha/2}$ является таким, что

$$Pr\{L \geq z_{1-\alpha/2}\} = \alpha/2.$$

Метод отношения правдоподобия

Рассмотрим также критерий $\widetilde{T}_n$ предложенный научным руководителем:

$$\widetilde{T}_n = \widetilde{T}_n(Z) = \ln(\widetilde{Q}(Z)/\widetilde{P}(Z)), \ln \widetilde{P}(Z) = -\frac{1}{2n} \min_{t \in R}(\sum_{i=1}^{2n}(g(Z_i - t))) \tag{8}$$

$$\ln \widetilde{Q}(Z) = -\frac{1}{n} \min_{t \in R}(\sum_{1 \le i \le n}(g(X_i - t))) - \min_{t \in R}(\sum_{1 \le i \le n}(g(Y_i - t))), \tag{9}$$

где $g = \ln(1 + x^2)$. Заметим, что критерий $\widetilde{T}_n$ по построению эквивалентен критерию отношения правдоподобия с параметром сдвига, оцененным по методу максимального правдоподобия.

Описание программы

Программа разделена на 2 части:

- Проведение численного эксперимента — моделирование эмпирических мощностей с помощью перестановочного метода.
- Вычисление интегралов и подсчёт асимптотической мощности.

Приемлемое время работы осуществляется за счет:

- Использования векторов из numpy и векторных операций.
- Использование параллельных вычислений.

Эмпирические мощности для нормального распределения

h_1	T_n	$\widetilde{T}_n$	KS	WMW	t	Anderson–Darling
1	0.12	0.1	0.08	0.18	0.15	0.11
2	0.27	0.23	0.18	0.42	0.39	0.28
3	0.52	0.48	0.36	0.68	0.67	0.53
4	0.77	0.68	0.64	0.87	0.87	0.75
5	0.93	0.86	0.83	0.97	0.98	0.92
6	0.98	0.97	0.94	0.99	0.99	0.98
7	0.99	0.99	0.98	0.99	1.0	0.99
8	1.0	0.99	0.99	1.0	1.0	0.99
9	1.0	1.0	0.99	1.0	1.0	1.0
10	1.0	1.0	1.0	1.0	1.0	1.0

Таблица: Для моделирования соответствующей H_1 ситуации использовались распределения $N(0,1)$ и $N(h_1/(\sqrt{n}),1)$. Число итераций $N = 1000$, число перестановок $K = 700$, $n = 100$, $\alpha = 0.05$.

Эмпирические мощности для распределения Коши

h_1	T_n	$\widetilde{T}_n$	KS	WMW	t	Anderson–Darling
1	0.06	0.08	0.05	0.11	0.05	0.06
2	0.11	0.17	0.11	0.2	0.05	0.11
3	0.18	0.32	0.22	0.32	0.06	0.22
4	0.32	0.5	0.38	0.47	0.06	0.36
5	0.48	0.72	0.55	0.61	0.08	0.53
6	0.65	0.85	0.7	0.72	0.1	0.68
7	0.8	0.94	0.82	0.83	0.12	0.79
8	0.9	0.97	0.92	0.91	0.13	0.88
9	0.95	0.99	0.96	0.96	0.15	0.94
10	0.98	0.99	0.98	0.97	0.18	0.97

Таблица: Для моделирования соответствующей H_1 использовались распределения *Cauchy*(0,1) и *Cauchy*$(h_1/(\sqrt{n}),1)$. Число итераций $N = 1000$, число перестановок $K = 700$, $n = 100$, $\alpha = 0.05$.

Эмпирические мощности для нормального распределения

h_2	T_n	Anderson–Darling	WMW	KS	$\widetilde{T}_n$
1	0.060	0.070	0.053	0.046	0.043
2	0.143	0.102	0.050	0.062	0.045
3	0.298	0.252	0.049	0.124	0.050
4	0.511	0.401	0.056	0.196	0.054
5	0.732	0.608	0.049	0.284	0.055
6	0.897	0.805	0.054	0.396	0.065
7	0.948	0.869	0.055	0.477	0.070
8	0.978	0.956	0.052	0.637	0.076
9	0.998	0.986	0.064	0.766	0.082
10	1.000	0.996	0.060	0.789	0.089

Таблица: Для моделирования соответствующей $H1$ ситуации использовались распределения $N(0,1)$ и $N(0,1+h_2/\sqrt{n})$. Число итераций $N = 1000$, число перестановок $K = 700$, $n = 100$, $\alpha = 0.05$.

Эмпирические мощности для распределения Коши

h_2	T_n	Anderson–Darling	WMW	KS	$\widetilde{T}_n$
1	0.071	0.050	0.048	0.048	0.047
2	0.106	0.059	0.055	0.057	0.052
3	0.171	0.092	0.055	0.088	0.054
4	0.266	0.124	0.049	0.109	0.055
5	0.382	0.152	0.057	0.130	0.058
6	0.483	0.235	0.054	0.178	0.063
7	0.601	0.326	0.054	0.247	0.065
8	0.680	0.360	0.057	0.271	0.070
9	0.797	0.474	0.053	0.353	0.075
10	0.847	0.568	0.056	0.407	0.078

Таблица: Для моделирования соответствующей H_1 использовались распределения *Cauchy*(0,1) и *Cauchy*$(0, 1 + h_2/\sqrt{n})$.Число итераций $N = 1000$, число перестановок $K = 700$, $n = 100$, $\alpha = 0.05$.

Сравнение эмпирических и теоретических мощностей

h_2	T_n teor	T_n emp 100	T_n emp 25
1	0.082	0.060	0.067
2	0.183	0.143	0.110
3	0.351	0.298	0.221
4	0.557	0.511	0.356
5	0.748	0.732	0.534
6	0.884	0.897	0.675
7	0.957	0.948	0.786
8	0.988	0.978	0.863
9	0.997	0.998	0.923
10	1.000	1.000	0.960

Таблица: Сравнение аналитических и численных результатов. Для моделирования соответствующей H_1 ситуации использовались распределения $N(0,1)$ и $N(0,1 + h_2/\sqrt{n})$

Сравнение эмпирических и теоретических мощностей

h_2	T_n teor	T_n emp 100	T_n emp 25
1	0.066	0.071	0.068
2	0.115	0.106	0.155
3	0.200	0.171	0.112
4	0.319	0.266	0.218
5	0.460	0.382	0.296
6	0.607	0.483	0.368
7	0.748	0.601	0.438
8	0.845	0.680	0.509
9	0.918	0.797	0.576
10	0.961	0.847	0.630

Таблица: Сравнение аналитических и численных результатов. Для моделирования соответствующей $H1$ ситуации использовались распределения *Cauchy*(0,1) и *Cauchy*$(0,1+h_2/\sqrt{n})$

Заключение

- Разработано математическое и программное обеспечение для сравнения «энергетического» метода с альтернативными классическими методами.
- С помощью статистического моделирования было показано, что этот метод превосходит по мощности альтернативные методы в случае, когда распределения различаются параметром масштаба.
- Также установлено, что в этом случае, аналитическая асимптотическая формула позволяет достаточно точно предсказывать мощность.