Исследование непараметрических методов проверки гипотез с помощью статистического моделирования

Мелас В. Б., доктор физ.-мат. наук, профессор кафедры статистического моделирования СПбГУ, vbmelas@yandex.ru[1],

Сальников Д.И., студент-магистр кафедры статистического моделирования СПбГУ, st013309@student.spbu.ru

Аннотация

Перестановочные тесты — важный класс методов проверки статистических гипотез, ставший доступным с развитием вычислительной техники. Их применение почти не требует выполнения каких–либо предположений о данных, тестовую статистику достаточно просто подстроить под конкретные задачи, а благодаря построению возможно точно достигнуть заданного уровня значимости.

В данной работе вводятся новые перестановочные тесты, а так же с помощью статистического моделирования исследуются их мощности в рамках задачи сравнения распределений двух выборок.

Введение

Задача проверки гипотезы равенства распределений двух выборок является классической задачей математической статистики и часто встречается на практике в различных областях. Классическим тестом проверки данной гипотезы является тест Колмогорова-Смирнова.

Хорошо известно (см. Леман, 1979 [1]), что для выборок из нормальных распределений в случае равных дисперсий оптимальным является тест Стьюдента. Для проверки гипотезы равенства центров распределений часто применяется ранговый тест Манна-Уитни, который обладает высокой мощностью для распределений с «тяжелыми хвостами».

Наряду с классическими параметрическими и непараметрическими тестами важное место занимают перестановочные тесты, обладающие высокой мощностью, гибкостью и универсальностью. Так, например, в работе Ludbrook, Dudley (1998, [7]) рекомендуется применение перестановочных

[1]Работа выполнена при частичной поддержке РФФИ (грант № 17-01-00161).

тестов в медико-биологических исследованиях, где объемы выборок зачастую слишком малы для применения классических тестов. В работе Huo и др. (2014, [8]) приведен большой обзор применения перестановочных тестов в педагогических и поведенческих науках.

В общем случае теоретическое сравнение мощностей статистических тестов является сложной задачей, часто неразрешимой. В данной работе мы рассмотрим широкий класс распределений и проведем сравнительный анализ мощностей перестановочных тестов для выборок из этих распределений.

Постановка задачи

Рассмотрим классическую задачу проверки гипотезы однородности

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

против альтернативы

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

в случае двух независимых выборок $X = (X_1 \dots X_{n_1})$ и $Y = (Y_1 \dots Y_{n_2})$ из генеральных совокупностей с функциями распределения F_1 и F_2 соответственно. Для упрощения обозначений и без потери общности положим $n_1 = n_2 = n$.

Определим вектор

$$Z = (X_1 \dots X_n, Y_1 \dots Y_n).$$

Определим следующие статистики:

$$K_1(Z) = (\overline{X} - \overline{Y})^2, \tag{3}$$

$$K_6(Z) = \sum_{i,j=1}^{n} |X_i - Y_j|, \tag{4}$$

$$L_\gamma(Z) = \sum_{i,j=1}^{n} \ln(1 + |X_i - Y_j|^\gamma), \quad \gamma = \{1, 2, 0.5\}, \tag{5}$$

$$L_\infty(Z) = \sum_{i,j=1}^{n} \ln(|X_i - Y_j|), \tag{6}$$

где под $\overline{X}$, $\overline{Y}$ подразумеваются выборочные средние значения.

Обозначим множество всевозможных перестановок элементов вектора Z через

$$\left\{Z(\pi_k) = (\underbrace{X_1(\pi_k)\dots X_n(\pi_k)}_{X(\pi_k)}, \underbrace{Y_1(\pi_k)\dots Y_n(\pi_k)}_{Y(\pi_k)})\right\}_{k=1}^{(2n)!}. \tag{7}$$

Приведем алгоритм проверки гипотезы 1, 2 на примере статистики $K_1(\cdot)$. Вычисляя значения $K_1(Z(\pi_k))$, $k = 1\dots(2n)!$, мы получаем перестановочное распределение величины $K_1(Z)$, которое позволяет нам принять решение относительно поставленной гипотезы, а именно:

- пусть d — общее число перестановок, r — число перестановок π_k, для которых $K_1(Z(\pi_k)) \geq K_1(Z)$;
- если отношение $\frac{r}{d} < \alpha$, то гипотеза H_0 (1) отвергается в пользу альтернативной гипотезы H_1 (2) с уровнем значимости α.

Приведенный выше алгоритм будем называть перестановочным тестом K_1. Тесты K_6 и L_γ, $\gamma = \{1, 2, 0.5, \infty\}$ вводятся по аналогии. В случае L_∞ необходимо, чтобы все элементы вектора Z были различны. Заметим, что при проверке гипотезы мы можем использовать лишь случайное подмножество перестановок $Z(\pi_k)$ размера d (рекомендации для выбора величины d, основанные на эмпирических исследованиях, можно найти в работах Keller-McNulty, Higgins, 1987 [5] и Marozzi, 2004 [6]).

Мощность K_1 была изучена численными методами в работе Sturino и др. (2010, [2]), K_6 был рассмотрен в работе Sirsky (2012, [3]), также мощности этих двух тестов исследовались в работе Мелас и др. (2016, [4]). Согласно выводам этих работ, K_6 обладает высокой мощностью для широкого класса распределений и особенно эффективен в случае, когда центры сравниваемых распределений совпадают, а мощность K_1 наиболее близка к мощности классического теста Стьюдента. Тесты L_i, $i = \{1, 2, 0.5, \infty\}$, насколько известно авторам данной статьи, введены впервые.

Заметим, что статистики $L_\gamma(\cdot)$, $\gamma = \{1, 2, 0.5\}$ (5), в отличие от $K_1(\cdot)$, $K_6(\cdot)$ и $L_\infty(\cdot)$, зависят от нормировки аргумента. Рассмотрим следующие пределы:

$$\frac{\sum_{i,j=1}^n \ln(1 + |X_i - Y_j|^\gamma)}{\sum_{i,j=1}^n |X_i - Y_j|^\gamma} \xrightarrow[\max\limits_{1\leq i,j\leq n} |X_i - Y_j| \to 0]{} 1,$$

$$\frac{\sum_{i,j=1}^n \ln(1 + |X_i - Y_j|^\gamma)}{\gamma \sum_{i,j=1}^n \ln(|X_i - Y_j|)} \xrightarrow[\min\limits_{1\leq i,j\leq n} |X_i - Y_j| \to \infty]{} 1.$$

Таким образом, при уменьшении расстояния между элементами выборок тест L_1 сходится к тесту K_6, а L_2 — к K_1. При увеличении расстояния между элементами выборок тесты $L_\gamma(\cdot)$, $\gamma = \{1, 2, 0.5\}$ сходятся к тесту L_∞ (согласно алгоритму проверки гипотезы умножение статистики теста L_∞ на константу γ не влияет на результат тестирования).

Рассмотрим выражение

$$\sum_{i,j=1}^{n} |X_i - Y_j| = \underbrace{\sum_{1\le i<j\le 2n} |Z_i - Z_j|}_{C'} - \sum_{1\le i<j\le n} |X_i - X_j| - \sum_{1\le i<j\le n} |Y_i - Y_j|.$$

Сумма C' состоит из $n(2n-1)$ слагаемых и является инвариантной относительно любой перестановки π_k. В качестве нормировочной константы было решено взять величину $C = C'/n(2n-1)$. Таким образом, к рассмотрению добавилось еще 3 теста со следующими статистиками:

$$L_\gamma^C(Z) = \sum_{i,j=1}^{n} \ln\left(1 + \left(\frac{|X_i - Y_j|}{C}\right)^\gamma\right), \quad i = \{1, 2, 0.5\}. \tag{8}$$

Задача заключается в исследовании мощности данных тестов для широкого класса типичных распределений с помощью статистического моделирования, а также в сравнении их мощности с классическими тестами Стьюдента (t.test), Колмогорова-Смирнова (ks.test) и Манна-Уитни (w.test).

Описание экспериментов

В исследовании были рассмотрены следующие распределения:

- нормальное распределение $N(\mu, \sigma)$;
- распределение Коши $C(x_0, \gamma)$;
- логнормальное распределение $LN(\mu, \sigma)$;
- распределение Парето $P(x_m, k)$;
- распределение Фишера $F(d_1, d_2)$;
- распределение Вейбулла $W(k, \lambda)$;
- бета-распределение $B(\alpha, \beta)$;

- гамма-распределение $G(k, \theta)$.

В каждом эксперименте были рассмотрены выборки размера $n = 5$ и $n = 20$, уровень значимости выбран равным $\alpha = 0.05$.

Для оценки мощности тестов в каждом случае было проведено по $m = 2500$ независимых испытаний. Стандартная ошибка среднего в m испытаниях Бернулли оценивается как $\hat{\sigma}_m = \sqrt{\frac{\hat{p}(1-\hat{p})}{m}}$. Согласно центральной предельной теореме при $m \to \infty$ истинная величина мощности p с вероятностью более 0.95 находится в интервале $(\hat{p} - 2\hat{\sigma}_m,\ \hat{p} + 2\hat{\sigma}_m) \overset{m=2500}{\subseteq} (\hat{p} - 0.02,\ \hat{p} + 0.02)$, где $\hat{p}$ — полученная оценка мощности.

Заметим, что все введенные перестановочные тесты инвариантны относительно перестановок элементов внутри векторов $X(\pi_k)$ и $Y(\pi_k)$, а также, в силу того, что размеры выборок одинаковы, относительно перемены местами $X(\pi_k)$ и $Y(\pi_k)$. Таким образом мы можем сократить размер множества перестановок 7 до $\frac{(2n)!}{2(n!)^2}$. При размерах исходных выборок $n = 5$ мы получаем всего 126 перестановок и можем применять точные тесты, однако при $n = 20$ их количество резко возрастает.

Для избежания вычислительных трудностей при $n = 20$ будем рассматривать только $d = 1600$ случайных, возможно повторяющихся перестановок. Такое количество было выбрано на основе выводов из работ Keller-McNulty, Higgins, 1987 [5] и Marozzi, 2004 [6]. Для практического применения рекомендуется брать $d = 5000$.

Исследование мощности логарифмических тестов

Проведем сравнительный анализ введенных логарифмических тестов L_γ, L^C_γ, где $\gamma = \{0.5, 1, 2\}$, и L_∞, используя пакет для статистической обработки данных R. Приведем полученные значения мощности тестов для случая размера выборки $n = 20$, так как в случае $n = 5$ результаты похожи. Жирным шрифтом выделены тесты, оценки 95% доверительных интервалов мощности которых пересекаются с оценкой 95% доверительного интервала мощности лидирующего теста. В каждом случае было проведено по 5 различных экспериментов с контролем уровня значимости, в таблицы вошли наиболее наглядные результаты.

Таблица 1: Мощность логарифмических тестов при размерах выборок $n = 20$

F_1	F_2	L_1	L^C_1	L_2	L^C_2	$L_{0.5}$	$L^C_{0.5}$	L_∞
$N(0,1)$	$N(1,1)$	0.802	0.807	**0.83**	**0.841**	0.76	0.762	0.685

$N(0,1)$	$N(0,3)$	**0.86**	**0.856**	**0.871**	0.826	**0.846**	**0.847**	0.812
$N(0,1)$	$N(1,2)$	**0.734**	**0.736**	**0.748**	**0.738**	0.706	0.71	0.662
$C(0,1)$	$C(2,1)$	**0.906**	0.854	**0.902**	0.663	**0.911**	**0.902**	**0.903**
$C(0,1)$	$C(0,6)$	**0.912**	0.822	**0.91**	0.525	**0.916**	**0.904**	**0.914**
$LN(0,1)$	$LN(1,1)$	**0.776**	**0.781**	**0.774**	**0.762**	**0.764**	**0.77**	0.725
$LN(0,1)$	$LN(0,4)$	0.925	0.888	0.882	0.739	**0.952**	**0.953**	**0.956**
$P(1,1)$	$P(3,1)$	0.991	0.911	0.989	0.684	**0.994**	0.986	**0.997**
$P(1,2)$	$P(1,6)$	**0.86**	0.851	**0.878**	**0.873**	0.814	0.811	0.727
$F(40,2)$	$F(40,20)$	**0.894**	**0.901**	**0.881**	0.838	0.868	**0.879**	0.801
$F(2,40)$	$F(20,40)$	**0.796**	**0.802**	0.609	0.687	**0.819**	**0.819**	**0.809**
$W(2,2)$	$W(2,4)$	0.949	**0.956**	**0.962**	**0.97**	0.937	0.938	0.902
$W(2,2)$	$W(8,2)$	**0.959**	**0.963**	0.882	0.941	**0.961**	**0.96**	0.944
$G(3,1)$	$G(3,2)$	0.908	**0.923**	0.915	**0.941**	0.891	0.9	0.844
$G(1,1)$	$G(2,1)$	**0.741**	**0.743**	**0.745**	**0.748**	**0.723**	**0.726**	0.68
$B(2,2)$	$B(5,2)$	**0.904**	0.889	**0.924**	**0.909**	0.867	0.856	0.804
$B(1,1)$	$B(8,8)$	0.81	**0.864**	0.103	0.786	**0.869**	**0.874**	**0.857**

Усредненные по всем рассмотренным случаям ошибки первого рода каждого теста находятся в интервале $(0.047, 0.05)$, что хорошо согласуется с заданным уровнем значимости $\alpha = 0.05$.

Из таблицы 1 видно, что мощности всех тестов довольно близки, однако наименее мощным является тест L_∞. Для нормальных распределений, различающихся параметром сдвига $N(\mu, 1)$, тест L_2 является наиболее мощным, для распределений Парето и логнормальных распределений, различающихся параметром масштаба (случаи $P(x_m, 1)$ и $LN(0, \sigma)$ соответственно) — тест $L_{0.5}$. Тест L_1 является своеобразным компромиссом между этими двумя тестами, зачастую оценка его мощности находится между оценками мощности тестов L_2 и $L_{0.5}$.

Опираясь на численные результаты можно заметить, что введенная нормировка для теста L_1 уменьшает его мощность для распределений с тяжелыми хвостами, а именно распределений Коши и распределений Парето с параметром формы, равном единице $P(x_m, 1)$ (у этих распределений не существует математического ожидания), а также для логнормальных распределений, различающихся параметром масштаба $LN(0, \sigma)$, однако дает незначительный выигрыш в остальных рассмотренных случаях, а также делает тест независимым к разбросу данных.

Сравнение лучшего логарифмического теста с известными ранее тестами

На основе проведенного в предыдущей главе анализа проведем сравнительное исследование мощности теста L_1^C, который оказался наилучшим среди введенных логарифмических тестов, с классическими тестами Стьюдента (t.test), Колмогорова-Смирнова (ks.test), Манна-Уитни (w.test), а также с перестановочными тестами K_1 и K_6. В таблицах 2 и 3 представлены результаты экспериментов в случае выборок размера $n = 5$ и $n = 20$ соответственно, при этом из таблицы 2 исключено распределение Фишера, так как при таком малом обьеме выборок все тесты имеют низкую мощность при любом выборе параметров.

Таблица 2: Мощность тестов при размерах выборок $n = 5$

F_1	F_2	K_1	K_6	L_1^C	$t.test$	$w.test$	$ks.test$
$N(0,1)$	$N(2,1)$	**0.799**	**0.782**	0.738	0.765	0.672	0.375
$N(0,1)$	$N(0,9)$	0.111	0.249	**0.726**	0.052	0.055	0.038
$N(0,1)$	$N(4,4)$	0.56	0.655	**0.702**	0.41	0.402	0.266
$C(0,1)$	$C(5,1)$	0.595	0.725	**0.791**	0.452	0.51	0.406
$C(0,1)$	$C(0,20)$	0.106	0.222	**0.658**	0.026	0.061	0.045
$LN(0,1)$	$LN(2,1)$	**0.768**	**0.756**	0.734	0.286	0.667	0.37
$LN(0,1)$	$LN(0,40)$	0.659	0.686	**0.739**	0.012	0.058	0.053
$P(1,1)$	$P(5,1)$	0.638	**0.662**	**0.687**	0.222	0.587	0.416
$P(1,2)$	$P(1,20)$	**0.762**	**0.759**	**0.748**	0.176	0.605	0.363
$W(2,2)$	$W(2,6)$	**0.758**	**0.755**	**0.726**	0.581	0.596	0.336
$W(2,2)$	$W(20,2)$	0.142	0.293	**0.712**	0.086	0.089	0.054
$G(3,1)$	$G(3,5)$	**0.943**	**0.938**	**0.928**	0.724	0.85	0.635
$G(1,1)$	$G(5,1)$	**0.954**	**0.947**	**0.938**	0.895	0.894	0.648
$B(2,2)$	$B(9,2)$	**0.715**	**0.712**	**0.682**	0.618	0.564	0.287
$B(1,1)$	$B(40,40)$	0.105	0.213	**0.584**	0.073	0.062	0.035

Таблица 3: Мощность тестов при размерах выборок $n = 20$

F_1	F_2	K_1	K_6	L_1^C	$t.test$	$w.test$	$ks.test$
$N(0,1)$	$N(1,1)$	**0.868**	**0.849**	0.807	**0.868**	**0.854**	0.704
$N(0,1)$	$N(0,3)$	0.058	0.742	**0.856**	0.054	0.065	0.294

$N(0,1)$	$N(1,2)$	0.5	**0.711**	**0.736**	0.486	0.46	0.524
$C(0,1)$	$C(2,1)$	0.316	0.742	**0.854**	0.21	0.81	**0.867**
$C(0,1)$	$C(0,6)$	0.05	0.663	**0.822**	0.02	0.07	0.353
$LN(0,1)$	$LN(1,1)$	0.746	0.78	0.781	0.667	**0.842**	0.708
$LN(0,1)$	$LN(0,4)$	0.71	0.836	**0.888**	0.065	0.064	0.452
$P(1,1)$	$P(3,1)$	0.525	0.768	0.911	0.284	0.98	**0.996**
$P(1,2)$	$P(1,6)$	**0.895**	**0.889**	0.851	0.673	0.804	0.678
$F(40,2)$	$F(40,20)$	0.724	**0.883**	**0.901**	0.276	0.241	0.449
$F(2,40)$	$F(20,40)$	0.073	0.626	**0.802**	0.072	0.232	0.461
$W(2,2)$	$W(2,4)$	**0.979**	**0.973**	0.956	**0.978**	0.946	0.894
$W(2,2)$	$W(8,2)$	0.098	0.898	**0.963**	0.086	0.158	0.489
$G(3,1)$	$G(3,2)$	**0.959**	**0.948**	0.923	**0.955**	0.931	0.839
$G(1,1)$	$G(2,1)$	0.742	0.757	0.743	0.736	**0.806**	0.656
$B(2,2)$	$B(5,2)$	**0.925**	**0.91**	0.889	**0.926**	0.893	0.787
$B(1,1)$	$B(8,8)$	0.054	0.708	**0.864**	0.052	0.067	0.324

Численные эксперименты показывают, что в обоих случаях тест L_1^C значительно превосходит остальные тесты для распределений, различающихся только параметрами масштаба ($N(0,\sigma)$, $LN(0,\sigma)$, $B(\alpha,\alpha)$), распределений Коши $C(x_0,\gamma)$, распределений Вейбулла с разным параметром формы $W(k,2)$, а также, при $n=20$, для распределений Фишера $F(d_1,d_2)$, особенно в случае, когда различается первый параметр распределения $F(d_1,40)$. В случае нормальных распределений, различающихся параметром сдвига $N(\mu,1)$, тесты K_1 и t.test являются наиболее мощными, однако, если у нормальных распределений одновременно изменяются оба параметра $N(\mu,\sigma)$, тест L_1^C является наиболее мощным.

Стоит также отметить, что для выборок малого объема ($n=5$, табл. 2) перестановочные тесты оказываются значительно более мощными, чем классические неперестановочные, что хорошо согласуется с выводами работы Ludbrook, Dudley (1998, [7]).

Заключение

В работе предложен ряд новых перестановочных тестов, основанных на сумме логарифмов разностей элементов двух выборок. В рамках задачи проверки гипотезы о равенстве двух распределений с помощью статистического

моделирования из предложенных тестов выбран наиболее мощный и универсальный, им оказался тест L_1^C (5).

Проведено исследование мощности теста L_1^C в сравнении с классическими тестами Стьюдента, Колмогорова-Смирнова и Манна-Уитни, а также с ранее исследованными в литературе перестановочными тестами. Из полученных результатов следует, что перестановочный тест L_1^C обладает высокой мощностью, в большом числе рассмотренных случаев оказываясь наиболее мощным. Преимущество этого теста над другими перестановочными тестами особенно велико в случае распределений с совпадающими центрами, а так же для распределений с тяжелыми хвостами.

Литература

[1] Леман Э. Проверка статистических гипотез. — М. : Наука, 1979.

[2] Statistical methods for comparative phenomics using high-throughput phenotype microarrays / J. Sturino, I. Zorych, B. Mallick et al. // The International Journal of Biostatistics. — 2010. — Vol. 6.

[3] Sirsky M. On the Statistical Analysis of Functional Data Arasing from Designed Experiments : Ph. D. thesis / M. Sirsky ; University of Manitoba. — 2012.

[4] Мелас В.Б., Сальников Д.И., Гудулина А.О. Численное сравнение перестановочных и классических методов проверки статистических гипотез // Вестник СПбГУ, сер.1, вып.3. — 2016.

[5] Keller-McNulty S., Higgins J. Effect of tail weight and outliers on power and type-I error of robust permutation tests for location // Communications in Statistics — Simulation and Computation. — 1987. — Vol. 16.

[6] Marozzi M. Some remarks about the number of permutations one should consider to perform a permutation test // Statistica — 2004.

[7] Ludbrook J., Dudley H. Why Permutation Tests Are Superior to t and F Tests in Biomedical Research // American Statistician — 1998 — Vol. 52.

[8] Permutation Tests in the Educational and Behavioral Sciences: A Systematic Review / M. Huo, M. Heyvaert, W. Van den Noortgate, P. Onghena. // Methodology European Journal of Research Methods for the Behavioral and Social Sciences. — 2014. — Vol. 10.