Научная статья

УДК 519.24
DOI: 10.18101/2304-5728-2022-1-35-44

МОДИФИКАЦИЯ МЕТОДА ФЕХНЕРА ДЛЯ ПОВЫШЕНИЯ УСТОЙЧИВОСТИ АНАЛИЗА ДАННЫХ

© Демаков Владимир Иванович
кандидат технических наук, доцент,
Иркутский государственный медицинский университет
Россия, 664003, г. Иркутск, ул. Красного Восстания, 1
demakovvi@yandex.ru

© Демаков Алексей Владимирович
студент,
Иркутский государственный университет
Россия, 664003, г. Иркутск, ул. Карла Маркса, 1
swim_alex@mail.ru

Аннотация. В статье предложена модификация критерия согласованности отклонений значений признаков Фехнера посредством замены среднего арифметического медианой. Приведены ситуации, в которых применение медианного критерия Фехнера целесообразно. На тестовом примере показан механизм различия использования медианы вместо среднего значения при оценке корреляционной связи данным критерием. На основе данных о многолетних наблюдениях за результатами разновозрастных спортсменов Иркутской области, полученных авторами в предыдущих работах, проведены эксперименты с массивами разного объема. Показана разница в устойчивости медианного метода Фехнера и классического как в условиях, близких к нормальному распределению, так и при наличии незначительного количества выбросов в выборочных данных. В том числе проанализировано поведение традиционного и модифицированного критериев при малых выборках.

Ключевые слова: согласованность выборочный данных, коэффициент Фехнера, коэффициент корреляции Пирсона, t-критерий Стьюдента, средние показатели.

Для цитирования
Демаков В. И., Демаков А. В. Модификация метода Фехнера для повышения устойчивости анализа данных // Вестник Бурятского государственного университета. Математика, информатика. 2022. № 1. С. 35–44.

Введение

Одним из наиболее востребованных на практике направлений статистических исследований является изучение зависимостей между различными варьирующимися признаками. Решение данной задачи может предшествовать построению математической модели рассматриваемого

явления или служить обоснованием для реализации каких-либо методик обработки данных, интерпретации получаемых результатов, а также для поиска причинно-следственных связей между процессами. Не случайно спектр способов практического использования корреляционного анализа достаточно широк.

Для оценки линейной зависимости в условиях классических параметрических распределений количественных показателей, близких к нормальному, используют традиционный коэффициент корреляции Пирсона. Для случаев нелинейного влияния факторов применяют методы регрессионного анализа [1]. Если количество изучаемых факторов велико, то обращаются к средствам факторного анализа [2]. В непараметрических ситуациях, когда нарушены условия нормального распределения характеристик или при сложности обоснования данного факта [3; 4], чаще исследуют зависимость между признаками критерием ранговой корреляции Спирмена или при помощи коэффициента корреляции Кендалла [5; 6]. Изучение дихотомических совокупностей производят при помощи коэффициентов ассоциации Φ, тетрахорических коэффициентов, углового преобразования Фишера и т. д. [7].

Среди перечисленных способов исследования корреляционной связи относительной простотой и универсальностью выделяется метод Фехнера, позволяющий установить наличие и направление зависимости между наборами данных или ее отсутствие [8]. Этот метод является непараметрическим, может применяться в условиях малой выборки, что делает его привлекательным. В данной работе предлагается способ повышения устойчивости данной методики по отношению к наличию некорректных наблюдений, снижающих качество выборки.

Модификация метода Фехнера

Рассмотрим подробнее метод Фехнера. Пусть даны два массива данных $X = \{x_1, x_2, \ldots, x_n\}$, $x_i \in R$ и $Y = \{y_1, y_2, \ldots, y_n\}$, $y_i \in R$ объёмом n. Для каждой пары наблюдений $\{x_i, y_i\}$ определяется, совпадает ли знак разности между текущим элементом и средним арифметическим данного массива. Подсчитывается количество таких совпадений v и несовпадений w:

$$v = \sum_{i=1}^{n} i, \quad \text{при } sign\left(x_i - \overline{x}\right) = sign\left(y_i - \overline{y}\right) \tag{1}$$

и

$$w = \sum_{i=1}^{n} i, \quad \text{при } sign\left(x_i - \overline{x}\right) \neq sign\left(y_i - \overline{y}\right), \tag{2}$$

где $\overline{x}$ и $\overline{y}$ — средние значения соответствующих массивов X и Y. Причем нулевые отклонения наблюдений от среднего считаются положительными.

Далее вычисляется коэффициент Фехнера (K_ϕ):

$$K_\phi = \frac{v - w}{v + w} = \frac{v - w}{n}. \tag{3}$$

Величина коэффициента Фехнера может принимать значения от -1 до 1 и интерпретация данного показателя аналогична классическому коэффициенту корреляции. Близость к нулю говорит об отсутствии линейной зависимости массивов *X* и *Y*, что происходит при примерно одинаковом количестве *v* и *w*. Наличие же корреляционной связи определяется в виде прямой зависимости при приближении $K_ф$ к 1 в случае значительного преобладания *v* или обратной зависимости при стремлении $K_ф$ к -1 в случае роста *w*.

Преимуществами данного метода является простота вычислений и программной реализации, а также независимость от закона распределения изучаемых массивов данных.

Ферстер и Ренц в работе [7] показывают, что величина коэффициента Фехнера вполне сопоставима с коэффициентом корреляции Пирсона (*r*). Однако данное утверждение подтверждается другими исследователями лишь при сильной тесноте связи ($|K_ф|$ и $|r| \in [0{,}8;\ 1]$) [9]. Авторами [10] было построено уравнение регрессии между $K_ф$ и *r* при доверительной вероятности 0,95 и интервалом существования $\Delta K_ф = \pm 0{,}0702$:

$$K_ф = 1{,}237r - 0{,}438 .$$

В работе [10] предлагается модифицированный индекс Фехнера, рассчитываемый по формуле:

$$K_ф^* = \pm 0{,}949\sqrt{\frac{v-w}{v+w}} \pm 0{,}051 ,$$

где знаки слагаемых положительны при $v > w$ и отрицательны при $v < w$.

Без сомнения, приведенные выше вариации метода Фехнера носят частный характер и были получены эмпирическим путем на основе пусть и многочисленных, но все же конкретных наборов данных.

Вместе с тем можно предложить скорректировать методику определения знаков отклонений в выражениях (1) и (2), заменив среднее арифметическое на медиану.

Соотношения (1) и (2) в этом случае примут вид:

$$v = \sum_{i=1}^{n} i, \quad \text{при } sign\left(x_i - x_{\text{мед}}\right) = sign\left(y_i - y_{\text{мед}}\right) \tag{4}$$

и

$$w = \sum_{i=1}^{n} i, \quad \text{при } sign\left(x_i - x_{\text{мед}}\right) \neq sign\left(y_i - y_{\text{мед}}\right), \tag{5}$$

где $x_{\text{мед}}$ и $y_{\text{мед}}$ — медианы массивов *X* и *Y*.

Действительно, в случае если единицы исследуемой совокупности расположены близко к нормальному распределению, то медиана незначительно отличается от среднего арифметического. Если же в наблюдаемых значениях прослеживается асимметричность или в выборке присутствуют

«выбросы» [13], то медиана нивелирует влияние таких данных. В отличие от среднего арифметического медиана является несмещенной оценкой [11; 12].

Приведем пример, иллюстрирующий бóльшую устойчивость (робастность) метода Фехнера при использовании медианы (табл. 1 и 2).

Таблица 1

Пример реализации метода Фехнера без «выброса»

№	X	Y	Знак отклонения				Совпадения знаков	
			X		Y			
			$x_i - \overline{x}$	$x_i - x_{\text{мед}}$	$y_i - \overline{y}$	$y_i - y_{\text{мед}}$	среднее	медиана
1	951	83	–	+	+	+	не совп.	совп.
2	874	76	–	–	–	–	совп.	совп.
3	957	84	+	+	+	+	совп.	совп.
4	1084	89	+	+	+	+	совп.	совп.
5	903	79	–	–	–	–	совп.	совп.

Для данных массивов X и Y найдем средние величины:

$$\overline{x} = 953{,}8\ ;\ \overline{y} = 82{,}2\ ;\ x_{\text{мед}} = 951 \text{ и } y_{\text{мед}} = 83\ .$$

В таблице 1 приведены знаки разности наблюдений и средних, а также совпадения и несовпадения этих знаков.

В данном случае отличие между средними и медианами незначительно. Определим коэффициенты Фехнера для обоих вариантов (3):

$$K_{\text{ф(среднее)}} = 0{,}6;\ K_{\text{ф(медиана)}} = 1.$$

Коэффициент корреляции Пирсона здесь составил $r = 0{,}966$.

Для полноты примера оценим статистическую значимость найденной корреляционной связи

$$T_{\text{набл}} = K_{\phi\,(\text{среднее})} \frac{\sqrt{n-2}}{\sqrt{1 - K^2_{\phi\,(\text{среднее})}}} = 0{,}6 \frac{\sqrt{8}}{\sqrt{1 - 0{,}6^2}} \approx 2{,}12\ .$$

Табличное значение критерия Стьюдента для вероятности 0,1 составляет $t_{\text{табл}} = 1{,}86$.

Так как $T_{\text{набл}} > t_{\text{табл}}$, то гипотеза о нулевой корреляции отвергается. Полученный $K_{\text{ф (среднее)}}$ статистически значим.

Оценим значимость коэффициента Фехнера, рассчитанного по медиане.

$$T_{\text{набл}} = K_{\phi\,(\text{медиана})} \frac{\sqrt{n-2}}{\sqrt{1 - K^2_{\phi\,(\text{медиана})}}} = 1 \frac{\sqrt{8}}{\sqrt{1 - 1^2}} \to \infty\ .$$

$T_{\text{набл}} > t_{\text{табл}}$, следовательно $K_{\text{ф (медиана)}}$ также статистически значим.

Теперь придадим исходным данным асимметричность (табл. 2), увеличив одно из значений множества X (в 4-й строке).

Таблица 2

Пример реализации метода Фехнера с «выбросом»

№	X	Y	Знак отклонения				Совпадения знаков	
			X		Y			
			$x_i - \overline{x}$	$x_i - x_{\text{мед}}$	$y_i - \overline{y}$	$y_i - y_{\text{мед}}$	среднее	медиана
1	951	83	–	+	+	+	не совп.	совп.
2	874	76	–	–	–	–	совп.	совп.
3	957	84	–	+	+	+	не совп.	совп.
4	***1123***	89	+	+	+	+	совп.	совп.
5	903	79	–	–	–	–	совп.	совп.

Увеличение разброса наблюдений привело к изменению среднего арифметического, в то время как значение медианы осталось прежним:

$$\overline{x} = 961{,}6\ ;\ \overline{y} = 82{,}2\ ;\ x_{\text{мед}} = 951 \text{ и } y_{\text{мед}} = 83\ .$$

Коэффициент Фехнера, основанный на среднем арифметическом, также изменился:

$$K_{ф(\text{среднее})} = 0{,}2;\ K_{ф(\text{медиана})} = 1.$$

Коэффициент корреляции Пирсона в этом случае составил $r = 0{,}944$.

Безусловно, приведенный пример не является репрезентативным, он лишь демонстрирует особенности использования медианы вместо среднего арифметического в случае асимметричности данных.

Для оценки значимости полученного результата мы использовали *t*-критерий Стьюдента, в основе которого также лежит сопоставление различий наблюдений и их среднего арифметического. В [15] показано, что параметрические критерии достаточно устойчивы к отклонениям выборочных законов распределения от нормального при проверке предположений о равенстве математических ожиданий. Анализ мощности и устойчивости таких критериев для различных ситуаций приведен в работах [16–18].

Основываясь на утверждении Б. Ю. Лемешко [15–21] о том, что критерий Стьюдента с ростом объемов выборок становится устойчивым к наличию асимметрии в законе распределения и не приводит к серьезным ошибкам, сопоставим степень восприимчивости медианного метода Фехнера по сравнению с классическим.

Для анализа робастности были выбраны данные, полученные на основе многолетних наблюдений за результатами разновозрастных спортсменов Иркутской области в некоторых циклических видах спорта. В работе [13] была предпринята попытка построить зависимость уровня работоспособности или показываемых результатов от возраста. Используя ту же математическую модель и решая обратную задачу оценки возраста по

имеющемуся спортивному результату, придадим исходным данным некоторый ограниченный псевдослучайный разброс. Средний предполагаемый результат составил 12 минут, что соответствует времени преодоления дистанции 3 км человеком, не занимающимся легкой атлетикой профессионально, в среднем возрасте — 30 лет. Диапазон разброса выбран ±1,5 минуты. Генерация разброса осуществлялась по методике Макото Мацумото алгоритмом «Вихрь Мерсенна» [22], применяемом в реализации функции выбора случайного числа MS Excel. Рассматривались выборки объемом 10, 50 и 100 наблюдений. Было проведено 1000 испытаний. При этом анализировались две выборки: полученная на основе описанной выше модели и с «выбросом» – некоторым искусственно завышенным показателем времени. Завышение в выборке объема 10 составило 1 наблюдение (10%), в выборках объема 50 и 100 — по 2 намеренно высоких данных (4 и 2% соответственно). Величина завышения — случайное число в пределах 90-100%. Полученные усредненные результаты приведены в таблице 3.

Таблица 3

Усредненные результаты реализации
1000 повторений метода Фехнера (классического и с «выбросом»)

Объем наблюдений	10	50	100
Коэф. Фехнера классический	0,314	0,392	0,314
Коэф. Фехнера медианный	0,294	0,391	0,314
Коэф. Фехнера классический с выбросом	-0,045	0,357	0,296
Коэф. Фехнера медианный с выбросом	0,284	0,394	0,317
Коэф. корреляции	0,281	0,297	0,276
Коэф. корреляции с выбросом	0,288	0,279	0,256

Из таблицы 3 видно, как меняется значение классического коэффициента Фехнера при наличии асимметрии данных. В малой выборке его величина изменилась с 0,314 до -0,045, в то время как медианный коэффициент Фехнера остался достаточно стабильным. Не так наглядна эта тенденция при росте объема выборки, тем не менее она видна. В совокупности в 100 единиц наличие 2% выбросов привело к изменению традиционного показателя Фехнера с 0,314 до 0,296, а медианный в тех же условиях изменился с 0,314 до 0,317. Для сопоставления в таблице также приведены усредненные показатели корреляции Пирсона для всех выборок.

На каждом шаге исследования проводилась оценка значимости получаемых результатов критерием Стьюдента. Критерий Стьюдента реализовывался с вероятностью получения ошибки первого рода $p = 0,1$. Усредненный анализ значимости результатов из табл. 3 приведен в следующей таблице:

Таблица 4

Усредненные результаты анализа значимости
1000 повторений метода Фехнера (классического и с «выбросом»)

Объем наблюдений	10	50	100
Средняя значимость классического коэф. Фехнера	0,580	0,388	0,309
% значимых наблюдений	17,6	74,3	92,3
Средняя значимость медианного коэф. Фехнера	0,569	0,349	0,284
% значимых наблюдений	19,3	74,2	91,2
Средняя значимость классического коэф. Фехнера с «выбросом»	0,632	0,312	0,261
% значимых наблюдений с «выбросом»	9	25	35,6
Средняя значимость классического коэф. Фехнера с «выбросом»	0,578	0,267	0,233
% значимых наблюдений с «выбросом»	15,6	67,3	88,7

Таблица 4 показывает оценку значимости модели Фехнера для рассматриваемой задачи моделирования работоспособности спортсменов. Результаты исследования значимости классического метода Фехнера не сильно отличаются от медианного при всех трех объемах выборки. Однако наличие «выбросов» существенно снизило процент наблюдений, которые критерием Стьюдента признавались значимыми. Медианный критерий Фехнера в аналогичных условиях демонстрирует гораздо бóльшую робастность относительно выбросов по сравнению с традиционным критерием.

Заключение

Проведенное исследование показало возможность повысить устойчивость критерия Фехнера к наличию в выборочных данных незначительного количества некорректных наблюдений за счет использования медианы вместо среднего арифметического. Предложенная модификация классического метода Фехнера апробирована на модели зависимости возраста человека от показываемых им спортивных результатов. Реализация данной модели в количестве 1000 повторений с псевдослучайным разбросом исходных данных, а также внесение в эти данные небольшого числа искусственно завышенных значений позволяет сделать следующие выводы:

1. Модификация классического метода Фехнера оценивания зависимости между наборами выборочных наблюдений путем замены среднего арифметического медианой не приводит к усложнению процедуры его реализации.

2. Рост объема наблюдений способствует выявлению корреляционной зависимости или независимости как при использовании классического, так и медианного критерия.

3. Наличие в выборке даже незначительного количества «выбросов» способно существенно снизить устойчивость классического метода Фехнера. Медианный метод при этом остается достаточно невосприимчивым к асимметрии в наблюдениях.

Литература

1. Кендалл М., Стьюарт А. Статистические выводы и связи. Москва: Наука, 1973. 900 с. Текст: непосредственный.

2. Демаков В. И., Баранов С. А. Проблемы проведения криминологического анализа // Вестник Восточно-Сибирского института Министерства внутренних дел России. 2015. Т. 75, № 4. С. 28–35. Текст: непосредственный.

3. Петров А. А. Проверка гипотезы о нормальности распределений по малым выборкам // Доклады Академии наук. 1951. Т. 76, № 3. С. 355–358. Текст: непосредственный.

4. Леман Э. Проверка статистических гипотез. Москва: Наука, 1964. 498 с. Текст: непосредственный.

5. Большев Л. Н. К вопросу о различении по малым выборкам нормального и равномерного типов распределений // Теория вероятностей и ее применения. 1965. Т. 10, № 4. С. 764–765. Текст: непосредственный.

6. Митропольский А. К. Техника статистических вычислений. Москва: Наука, 1971. 576 с. Текст: непосредственный.

7. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов. Москва: Финансы и статистика, 1983. 302 с. Текст: непосредственный.

8. Гаскаров Д. В., Шаповалов В. И. Малая выборка. Москва: Статистика, 1978. 248 с. Текст: непосредственный.

9. Долгов А. Ю. Повышение эффективности статистических методов контроля и управления технологическими процессами изготовления микросхем: диссертация на соискание ученой степени кандидата технических наук. Москва: МГАПИ, 2000. 217 с. Текст: непосредственный.

10. Долгов Ю. А., Долгов А. Ю., Столяренко Ю. А. Метод повышения точности вычисления параметров выборки малого объема // Вестник Приднестровского государственного университета им. Т. Г. Шевченко. 2010. Юб. вып. С. 232–242. Текст: непосредственный.

11. Вентцель Е. С. Теория вероятностей. Москва: Наука, 1973. 576 с. Текст: непосредственный.

12. Крамер Г. Математические методы статистики. Москва: Мир, 1975. 648 с. Текст: непосредственный.

13. Моделирование уровня работоспособности и сопоставимости спортивных результатов в зависимости от возраста спортсменов / Я. А. Портная, В. И. Демаков, В. И. Рерке [и др.] // Успехи геронтологии. 2021. Т. 34, № 3. С. 419–424. Текст: непосредственный.

14. Справочник по прикладной статистике. Том 2 / под редакцией Э. Ллойда, У. Ледермана. Москва: Финансы и статистика, 1990. 526 с. Текст: непосредственный.

15. Лемешко Б. Ю., Лемешко С. Б. Об устойчивости и мощности критериев проверки однородности средних // Измерительная техника. 2008. № 9. С. 23–28. Текст: непосредственный.

16. Лемешко Б. Ю., Лемешко С. Б., Постовалов С. Н. Мощность критериев согласия при близких альтернативах // Измерительная техника. 2007. № 2. С. 22–27. Текст: непосредственный.

17. Лемешко Б. Ю., Лемешко С. Б., Постовалов С. Н. Сравнительный анализ мощности критериев согласия при близких конкурирующих гипотезах. I. Проверка простых гипотез // Сибирский журнал индустриальной математики. 2008. Т. 11, № 2(34). С. 96–111. Текст: непосредственный.

18. Лемешко Б. Ю., Лемешко С. Б., Постовалов С. Н. Сравнительный анализ мощности критериев согласия при близких альтернативах. II. Проверка сложных гипотез // Сибирский журнал индустриальной математики. 2008. Т. 11, № 4(36). С. 78–93. Текст: непосредственный.

19. Лемешко Б. Ю., Помадин С. С. Корреляционный анализ наблюдений многомерных случайных величин при нарушении предположений о нормальности // Сибирский журнал индустриальной математики. 2002. Т. 5, № 3. С. 115–130. Текст: непосредственный.

20. Лемешко Б. Ю. Критерии проверки отклонения распределения от нормального закона. Руководство по применению. Москва: ИНФРА-М, 2015. 160 с. Текст: непосредственный.

21. Лемешко Б. Ю. Непараметрические критерии согласия. Руководство по применению. Москва: ИНФРА-М, 2014. 163 с. DOI: 10.12737/11873. Текст: непосредственный.

22. Matsumoto M., Nishimura T. Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudo-random Number Generator // ACM Transactions on Modeling and Computer Simulations. 2017. Vol. 8, No. 1. Pp. 3–30. DOI:10.1145/272991.272995.

Статья поступила в редакцию 25.01.2022; одобрена после рецензирования 04.03.2022; принята к публикации 15.03.2022.

MODIFICATION OF THE FECHNER METHOD TO INCREASE THE ROBUSTNESS OF DATA ANALYSIS

Vladimir I. Demakov
candidate of sciences (technical), associate professor,
Irkutsk State Medical University
1 Krasnogo Vosstaniya str., Irkutsk 664003, Russia
demakovvi@yandex.ru

Alexey V. Demakov
student,
Irkutsk State University
1 Karl Marks str., Irkutsk, 664003, Russia
swim_alex@mail.ru

Abstract. The article proposes a modification of the criterion for the consistency of the deviations of the values of the Fechner's features by replacing the arithmetic mean with the median. Situations are given in which the use of the Fechner's median criterion is expedient. The test example shows the difference mechanism for using the median instead of the mean value when evaluating the correlation with this criterion. Based on the data on long-term observations of the results of athletes of different ages in the Irkutsk region, obtained by the authors in previous works, experiments were carried out with arrays of different sizes. The difference in the stability of the median Fechner method and the classical one is shown, both under conditions close to the normal distribution and in the presence of an insignificant amount of outliers in the sample data. Including analyzed the behavior of the traditional and modified criteria for small samples.

Keywords: consistency of sample data, Fechner's law, Pearson correlation coefficient, Student's t-test, average indicators.

For citation

Demakov V. I., Demakov A. V. Modification of the Fechner Method to Increase the Robustness of Data Analysis // Bulletin of Buryat State University. Mathematics, Informatics. 2022. N. 1. Pp. 35–44.

The article was submitted 25.01.2022; approved after reviewing 04.03.2022; accepted for publication 15.03.2022.