УДК 006.91

Г.Г. САФАРЯН, М.П.СЕРГИЕНКО

ИССЛЕДОВАНИЕ ПАРАМЕТРИЧЕСКИХ И НЕПАРАМЕТРИЧЕСКИХ МЕТОДОВ ОПРЕДЕЛЕНИЯ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ ДАННЫХ С АНОРМАЛЬНЫМИ ЗАКОНАМИ РАСПРЕДЕЛЕНИЯ

Исследованы особенности определения коэффициента корреляции двух функционально независимых коррелированных величин параметрическими и непараметрическими методами для случаев большого и малого числа наблюдений. Выявлены причины расхождения результатов, получаемых разными методами, и предложены возможности их устранения. Исследовано влияние закона распределения исходных величин на точность расчета коэффициента корреляции между ними. Даны рекомендации по оптимальному использованию разных методов для исследования корреляционных связей различных величин.

Актуальность исследования. Во многих областях жизнедеятельности человека существует необходимость математической обработки взаимосвязанных данных (это большинство технических отраслей, медицина, химия, биология, социология, психология и т.д.). При этом выделяют два вида взаимосвязи – функциональную и корреляционную. Под функциональной связью понимают связь, где существует полное соответствие между факторными и результативными признаками, то есть определенному значению факторного признака соответствует одно и только одно значение результативного признака. Корреляционная связь – это связь, где воздействие отдельных факторов проявляется только как тенденция (в среднем) при массовом наблюдении фактических данных.

В настоящее время для определения функциональной взаимосвязи с заданной необходимой точностью разработано большое количество аналитических и численных методов (методы наименьших квадратов, максимального правдоподобия и др. [1]). Результаты исследования корреляционных зависимостей получены для случая нормально распределенных данных [2, 3], нет единого подхода к выбору метода определения коэффициента корреляции (что вызвано разными результатами при использовании разных методов для одних и тех же исходных данных), предполагается использование опыта и навыков у исследователя, что предает результатам субъективный и рекомендательный характер. Такое состояние проблемы нельзя считать удовлетворительным, поскольку корреляция может иметь сильное влияние на точность результата измерения. При повышении точности и разрешающей способности средств измерительной техники и постоянном совершенствовании вычислительной техники и программного обеспечения к ней появляются возможности для более детального исследования физических величин и процессов, и одной из актуальных и важных задач является исследование корреляционных связей между исходными данными и способов их определения.

Постановка задачи. Для нахождения корреляции между двумя выборками наиболее часто используются коэффициент корреляции и выборочный коэффициент корреляции (параметрические), коэффициенты ранговой корреляции Спирмена и Кендалла (непараметрические) [2, 4]. В технической литературе [2,3] отмечается, что выборочный коэффициент корреляции может быть применен только при нормальном распределении обеих выборок. Таким образом, *задачей* является исследование точности параметрических и непараметрических методов определения коэффициентов корреляции; исследрование влияния закона распределения на значение параметрических коэффициентов корреляции; исследование корреляции при малом объеме выборок.

Цель исследования – определение границ применимости и усовершенствование разных способов вычисления коэффициента корреляции.

Способы расчета коэффициента корреляции. Как было отмечено выше, существуют параметрические и непараметрические способы определения коэффициента корреляции. В общем случае при параметрическом подходе коэффициент корреляции R случайных величин x и y рассчитывают по формуле [2]

$$R = \frac{M[(x - Mx)(y - My)]}{\sigma_x \sigma_y}, \qquad (1)$$

где M[…] - математическое ожидание (МО) величины […];

Mx , My - МО исследуемых величин x и y соответственно;

σ_x , σ_y - средние квадратические отклонения (СКО) величин x и y соответственно.

При этом величины x и y должны быть «...взаимно независимыми, одинаково распределенными двумерными случайными величинами, подчиняющимися нормальному распределению» [2]. Если же параметры функции нормального совместного распределения неизвестны, то в качестве оценки коэффициента корреляции используют выборочный коэффициент корреляции r

$$r = \frac{\sum_{i=1}^{n}[(x_i - \overline{x})(y_i - \overline{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}, \qquad (2)$$

где $\overline{x} = \sum_{i=1}^{n} x_i \Big/ n$, $\overline{y} = \sum_{i=1}^{n} y_i \Big/ n$.

Были проведены исследования, направленные на изучение влияния законов распределения исходных величин на точность определения коэффициента корреляции в соответствии с выражениями (1), (2). При этом в формуле (1) в качестве МО и СКО для случайной величины ξ , распределенной по нормальному закону, использовались выражения

$$M\xi = \overline{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i \; ; \; \sigma_\xi = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2} \; , \qquad (3)$$

и тогда (1) и (2) совпадают;
для величины, распределенной по равномерному закону и закону арксинуса, МО и СКО имеют вид [5]

$$M\xi = \frac{\xi_{max} + \xi_{min}}{2} \; ; \; \sigma_\xi = \frac{\xi_{max} - \xi_{min}}{2t_p} \; , \qquad (4)$$

где ξ_{min} , ξ_{max} - наименьшее и наибольшее значения случайной величины ξ ;

t_p - доверительный коэффициент ($t_p = \sqrt{3}$ для равномерного закона распределения, $t_p = \sqrt{2}$ для арксинусного закона распределения).

Также широкое распространение при исследованиях корреляционных связей различного происхождения получили непараметрические методы, оперирующие не параметрами законов распределения исходных величин (МО, СКО), а рангами этих величин по каждому признаку. Выборочными мерами связи в этом случае служат ранговые коэффициенты корреляции. Наиболее используемыми являются коэффициент ранговой корреляции ρ Спирмена

$$\rho = 1 - \frac{6S_\rho}{n(n^2-1)}, \qquad (5)$$

где $S_\rho = \sum_{i=1}^{n}(d_i - i)^2$: d_i - порядковый номер (по второму признаку) той пары признаков, которая по первому признаку имеет номер i; n - число наблюдений (пар рангов);
и коэффициент ранговой корреляции τ Кендалла

$$\tau = \frac{4N}{n(n-1)} - 1, \qquad (6)$$

где N - количество тех пар рангов, для которых одновременно $j > i$ и $d_j > d_i$.

Основными преимуществами непараметрических методов расчета коэффициента корреляции является независимость от закона распределения исследуемых величин и нечувствительность к возможным выбросам в исходных данных, чего нельзя сказать о параметрических методах. При этом коэффициент ранговой корреляции ρ Спирмена является аналогом выборочного коэффициента корреляции r и эти критерии сравнимы по мощности, в то же время коэффициент ранговой корреляции τ Кендалла отличается повышенной сложностью, поскольку количество используемых для расчета данных нелинейно возрастает с увеличением исходных данных, что существенно увеличивает время расчета и необходимые программные ресурсы.

Исследования проводились методом Монте-Карло генерирования совместного (двумерного) закона распределения коррелированных входных величин с произвольными законами распределения в соответствии с рекомендациями работы [6], включающими следующие операции:

1) генерирование двух последовательностей нормально распределенных некоррелированных случайных чисел ξ_1 и ξ_2 с нулевым математическим ожиданием и единичным стандартным отклонением.

2) формирование из этих последовательностей третьей последовательности $\xi_3 = r^* \xi_1 + \sqrt{1-(r^*)^2}\xi_2$. В этом случае ξ_1 и ξ_3 представляют собой нормально распределенные коррелированные случайные величины с заданным коэффициентом корреляции r^*;

3) преобразование от ξ_1 и ξ_3 в виде интегральной функции нормированного нормального распределения $\upsilon = F_н(\xi)$ с получением последовательностей равномерно распределенных в диапазоне от 0 до 1 коррелированных случайных чисел υ_1 и υ_2 с коэффициентом корреляции $r_{1,2}$, близким по значению к исходному коэффициенту r^*;

4) получение нормированных равномерно распределенных случайных чисел $\upsilon_{1н}$ и $\upsilon_{2н}$ с нулевым МО и единичным СКО в соответствии с выражением $\upsilon_{1,2н} = (2\upsilon_{1,2} - 1)\sqrt{3}$;

5) получение двух последовательностей коррелированных случайных чисел x и y с заданными законами распределения методом обратных функций $x = F^{-1}(\upsilon_{1н})$, $y = F^{-1}(\upsilon_{2н})$, где F - интегральная функция заданного закона распределения.

Таким образом, были получены совместные (двумерные) функции распределения величин с нормальными, равномерными и арксинусными законами распределения и их комбинациями.

Для моделирования исходных данных согласно описанному алгоритму и расчета коэффициентов корреляции в соответствии с выражениями (1) – (6) были использованы пакеты Mathcad 13 и Statistica 6.0.

При наличии большого числа наблюдений (моделирование осуществлялось при $n = 65000$ с усреднением 50 раз) для двух нормально распределенных величин были получены зависимости коэффициентов корреляции r, ρ и τ, рассчитанных по формулам (2), (5) и (6) по одним и тем же данным, от заданного коэффициента корреляции r^* (по модулю). Исследования показали, что для выборочного коэффициента корреляции (2) разность между ним и заданным значением коэффициента корреляции не превышает $2 \cdot 10^{-4}$. Формулы для расчета ранговых коэффициентов корреляции имеют систематическую составляющую погрешности и нуждаются во введении поправки при нахождении коэффициента корреляции, которая показана на рис.1. СКО коэффициентов корреляции не превысило $7 \cdot 10^{-4}$.

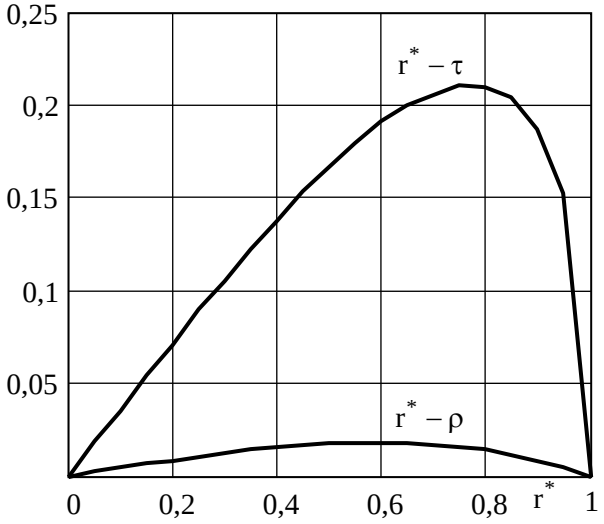


Рис.1. Поправки для расчета коэффициентов ранговой корреляции ρ Спирмена и τ Кендалла

Показанные зависимости являются очень важными, поскольку отражают несовершенство непараметрических методов и отвечают на вопросы, почему разные методы определения коэффициента корреляции не дают одинаковый результат, каким методом пользоваться, и можно ли использовать сразу несколько методов. Очевидно, что прибавление поправки в соответствии с рис.1 позволит разными методами получить одинаковый результат и оптимально использовать непараметрические методы, когда это необходимо.

Теперь обратимся к вопросу влияния закона распределения исходных величин на точность расчета коэффициента корреляции R (1) с учетом выражений (3), (4) и выборочного коэффициента корреляции (2). Полученные значения коэффициентов корреляции при использовании выражения (1) приведены в табл.1, при использовании выражения (2) – в табл.2. СКО коэффициентов корреляции не превысило $8 \cdot 10^{-4}$.

Таблица 1

R	Заданный коэффициент корреляции r^*										
Закон 1, закон 2	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Норм., норм.	0	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	1,00
Равн., равн.	0	0,09	0,19	0,29	0,39	0,48	0,58	0,68	0,78	0,89	1,00
Арксин.,арксин.	0	0,09	0,18	0,27	0,36	0,46	0,55	0,65	0,76	0,87	0,92
Норм., равн.	0	0,10	0,19	0,29	0,39	0,49	0,59	0,68	0,78	0,88	0,98
Норм., арксин.	0	0,09	0,19	0,28	0,38	0,47	0,57	0,66	0,76	0,85	0,95
Равн., арксин.	0	0,09	0,18	0,28	0,37	0,47	0,57	0,67	0,77	0,88	0,99

Таблица 2

r	Заданный коэффициент корреляции r^*										
Закон 1, закон 2	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Норм., норм.	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Равн., равн.	0	0,09	0,19	0,29	0,39	0,48	0,58	0,68	0,79	0,89	1
Арксин.,арксин.	0	0,09	0,18	0,27	0,37	0,46	0,56	0,66	0,77	0,88	1
Норм., равн.	0	0,10	0,19	0,29	0,39	0,49	0,59	0,68	0,78	0,88	0,98
Норм., арксин.	0	0,09	0,19	0,28	0,38	0,48	0,57	0,66	0,76	0,85	0,95
Равн., арксин.	0	0,09	0,19	0,28	0,37	0,47	0,57	0,67	0,77	0,88	0,99

Таблицы 1 и 2 показывают, что существует влияние закона распределения исходных величин на точность определения коэффициента корреляции, разность между заданным и полученным коэффициентами корреляции может достигать 0,08 при использовании выражения (1) и 0,05 при использовании выражения (2). Однако если в высокой точности определения коэффициента корреляции необходимости нет, и исходные величины не обременены случайными погрешностями и промахами, возможно использование этих выражений.

Следует отметить некоторые ситуации, описанные в научно-технической литературе. Так, в [2] отмечено, что «если корреляция случайных величин x и y отлична от нормальной (т.е. если функция распределения этих величин не принадлежит совместному нормальному распределению), то коэффициент корреляции может принимать значения, близкие или даже равные нулю в тех случаях, когда x и y зависимы», в [1] показаны рисунки, когда при явно зависимых переменных коэффициент корреляции близок к нулю, в [3] описан случай, когда зависимость величин $\xi_1 = \xi_2^2$, где ξ_2 - нормально распределенная случайная величина с нулевым МО, не вызывает сомнений, однако их корреляционный момент равен нулю. По нашему мнению, первые две ситуации возникают вследствие наличия между исходными величинами помимо корреляционной связи еще и функциональной связи или их связи с третьей неизвестной величиной. Поэтому в данных случаях необходимо сначала найти эту зависимость (например, с помощью метода наименьших квадратов), а потом между функционально независимыми величинами искать корреляционную связь. Ситуация, описанная в третьем примере, возникла вследствие равенства МО величины ξ_2 нулю, и естественно при возведении ξ_2 в квадрат корреляционный момент стал равен нулю. При искусственном изменении МО величины ξ_2, что не влияет на точность расчета коэффициента корреляции, такой результат возникнуть не может, и коэффициент корреляции между ξ_1 и ξ_2 будет равен 1.

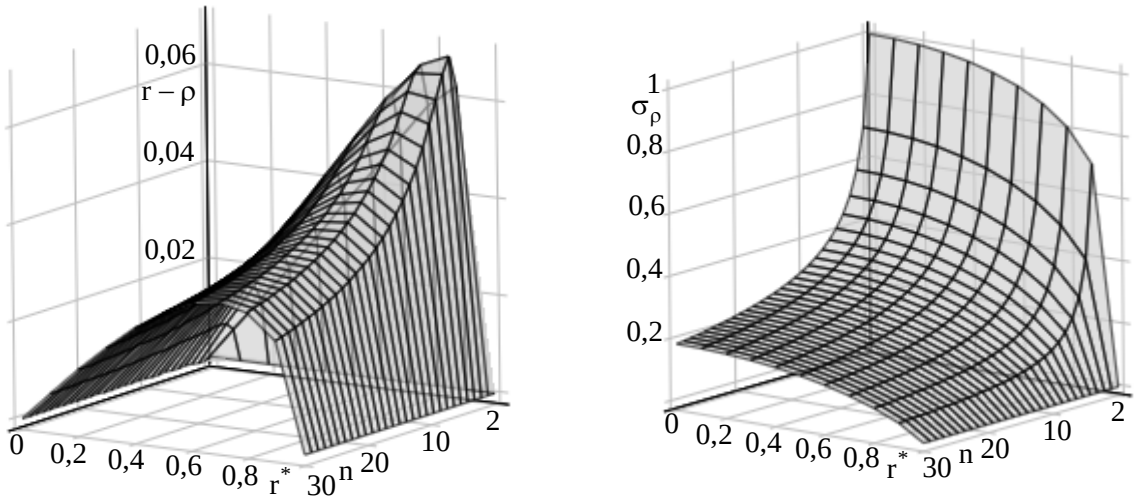
При малом числе наблюдений ($n = 2…30$ с усреднением $1 \cdot 10^6$ раз) было проведено исследование точности определения коэффициента корреляции с использованием выражений (1), (2) и (5). Коэффициент ранговой корреляции τ Кендалла для малых выборок рассмотрен не был, поскольку исследования показали его большую систематическую погрешность по сравнению с коэффициентом ρ Спирмена.

Рассмотрим случай, когда исходные величины распределены по нормальному закону, т.е. когда выражения (1) и (2) идентичны. Исследования показали, что при расчете коэффициента корреляции при малом числе наблюдений по формуле (1) для коррекции результата целесообразнее использование выражения

$$r^* = r\left[1 + \frac{1 - r^2}{2(n-1)}\right] \tag{7}$$

вместо выражения $r^* = r\left[1 + \frac{1 - r^2}{2(n-3)}\right]$, приведенного в [4].

Для расчета коэффициента ранговой корреляции ρ Спирмена необходимо вводить поправку, показанную на рис. 2 а), и далее пользоваться выражением (7). СКО коэффициентов корреляции σ_ρ, близкие по значениям σ_r, показаны на рис. 2 б).



а) б)

Рис.2. Поправки к коэффициенту ранговой корреляции ρ Спирмена (а) и его СКО (б)

Преимуществом данного подхода, когда определяется коэффициент ρ Спирмена с введением соответствующей поправки, является возможность его применения при полном отсутствии информации о законе распределения исходных величин и наличии грубых погрешностей и промахов. Однако следует отметить, что СКО коэффициентов корреляции при малом числе наблюдений может быть достаточно большим (рис.2 б)), поэтому у исследователя должна быть возможность проведения повторных наблюдений для последующего усреднения.

Рассмотрим влияние законов распределения исходных величин на точность определения коэффициента корреляции по формуле (1) с использованием (3), (4) и формуле (2). В табл.3 приведены максимальные отклонения полученного с использованием (1) коэффициента корреляции от коэффициента корреляции, полученного для нормально распреде-

ленных исходных величин (r из выражения (7)). В табл.4 приведены те же данные для случая использования выражения (2) для расчета коэффициента корреляции.

Таблица 3

	Заданный коэффициент корреляции r^*										
Закон 1, закон 2	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Равн., равн.	0	0,12	0,26	0,39	0,53	0,66	0,82	0,99	1,18	1,43	2,00
Арксин.,арксин.	0	0,06	0,13	0,19	0,26	0,33	0,41	0,49	0,59	0,72	1,00
Норм., равн.	0	0,02	0,03	0,05	0,07	0,09	0,11	0,12	0,15	0,17	0,22
Норм., арксин.	0	-0,01	-0,01	-0,02	-0,03	-0,04	-0,04	-0,05	-0,06	-0,07	-0,08
Равн., арксин.	0	0,09	0,19	0,28	0,38	0,48	0,59	0,72	0,85	1,03	1,45

Таблица 4

	Заданный коэффициент корреляции r^*										
Закон 1, закон 2	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Равн., равн.	0	0	-0,01	-0,01	-0,01	-0,02	-0,02	-0,02	-0,01	-0,01	0
Арксин.,арксин.	0	-0,01	-0,02	-0,03	-0,03	-0,04	-0,04	-0,04	-0,03	-0,02	0
Норм., равн.	0	0	0	-0,01	-0,01	-0,01	-0,01	-0,01	-0,02	-0,02	-0,02
Норм., арксин.	0	-0,01	-0,01	-0,01	-0,02	-0,02	-0,03	-0,03	-0,04	-0,04	-0,05
Равн., арксин.	0	-0,01	-0,01	-0,02	-0,02	-0,03	-0,03	-0,03	-0,03	-0,02	-0,01

Сравнение табл. 3 и 4 показывает, что использование выражения (1) с учетом эффективных оценок МО и СКО (3), (4) для разных законов распределения не является адекватным. При этом, как показали проведенные исследования, СКО коэффициента корреляции может возрасти до 3 раз по сравнению с приведенными на рис.2 б). Выражение (2) с достаточной точностью может быть использовано при разных законах распределения исходных величин. СКО коэффициента корреляции при этом не изменяется.

Выводы. В работе решена важнейшая задача выбора способа расчета коэффициента корреляции между функционально независимыми величинами для случаев большого и малого числа наблюдений.

Для случая большого числа наблюдений возможны два пути получения коэффициента корреляции с необходимой точностью:

1) использование непараметрических коэффициентов ранговой корреляции ρ Спирмена или τ Кендалла с обязательным введением поправок. В этом случае нет необходимости в идентификации закона распределения исходных величин, а также нахождении грубых погрешностей и промахов;

2) использование параметрических (в частности, выборочного) коэффициентов корреляции. Однако в этом случае рекомендуется определить законы распределения исходных величин для того, чтобы определить, насколько возможно смещение результата вычисления коэффициента корреляции, хотя, как показали проведенные исследования, это смещение не является существенным для нормальных, равномерных и арксинусных законов распределения исходных величин, а также их комбинаций. Особое внимание необходимо уделять исключению грубых погрешностей и промахов, поскольку от этого сильно зависит точность определения МО и СКО исходных величин, посредством которых рассчитывают коэффициент корреляции.

Для случая малого числа наблюдений оптимальным является использование выборочного коэффициента корреляции (2) с введением поправки (7). Если же в исходных дан-

ных присутствуют грубые погрешности и промахи, исключение которых сильно влияет на объем выборки, следует воспользоваться коэффициентом ранговой корреляции ρ Спирмена с введением соответствующей поправки.

Научная новизна проведенных исследований заключается в сравнении по точности различных методов (параметрических и непараметрических) определения коэффициента корреляции двух случайных величин и нахождении поправочных коэффициентов, что позволило выработать единый подход к исследованию коррелированных данных. Оптимизирована формула для расчета коэффициента выборочной корреляции при малом числе наблюдений, рассчитаны поправочные коэффициенты для расчета коэффициента ранговой корреляции Спирмена для этого случая. Показано, что влияние формы закона распределения исходных величин на точность определения выборочного коэффициента корреляции не настолько сильно, как принято считать.

Практическая значимость полученных результатов состоит в приобретении возможности с одинаковой точностью использовать параметрические и непараметрические методы обработки коррелированных величин в зависимости от условий проведения измерительного эксперимента, количества полученных данных, наличия в результатах эксперимента помех и др. Дальнейшим этапом в развитии этой темы может стать исследование корреляционной связи функционально зависимых величин.

Список литературы: 1. *Бронштейн И.Н., Семендяев К.А.* Справочник по математике для инженеров и учащихся втузов. М.: Наука, 1981. 720 с. **2.** *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. М.: Наука, 1983.416 с. **3.** *Фрумкин В.Д., Рубичев Н.А.* Теория вероятностей и статистика в метрологии и измерительной технике. М.: Машиностроение, 1987. 168 с. **4.** *Степнов М.Н.* Статистические методы обработки результатов механических испытаний: Справочник. М.: Машиностроение, 1985. 232 с. **5.** *Захаров И.П.* Теоретическая метрология. Харьков: ХТУРЭ, 2000.172 с. **6.** *Захаров И.П.* Моделирование коррелированных данных при обработке результатов измерений// Моделювання та інформаційні технології: Наук.–техн. зб. 2005. Вип. 33. С. 35 – 40.

Поступила в редколлегию 00.00.00

Сафарян Григорий Гагикович, инженер кафедры МИТ ХНУРЭ. Научные интересы: исследование погрешностей вычислительных операций при цифровой обработке сигналов, статистическая обработка результатов измерений. Адрес: Украина, 61166, г. Харьков, пр. Ленина-14, тел. 702-1331.

Сергиенко Марина Петровна, канд. техн. наук, с.н.с. кафедры МИТ ХНУРЭ. Научные интересы: метрологическая идентификация многопараметрических средств измерений. Адрес: Украина, 61166, г. Харьков, пр. Ленина-14, тел. 702-1331.

УДК 006.91

Дослідження параметричних та непараметричних методів визначення коефіцієнту кореляції даних з анормальними законами розподілу/ Г.Г. Сафарян, М.П. Сергієнко// АСУ та прилади автоматики. 2000. Вип. 00. С.000-000.

Проведено аналіз точності визначення коефіцієнту кореляції випадкових величин параметричними та непараметричними методами. Отримані корегуючі коефіцієнти для оптимізації цих методів. На прикладі величин з нормальними, рівномірними та арксинусними законами розподілу, а також їх комбінацій показано, що форма закону розподілу не впливає критично на результат визначення коефіцієнту кореляції. Дослідження проведені для випадків з великим та малим об'ємами вихідних даних.

Табл. 4. Іл. 2. Бібліогр.: 6 назв.

UDC 006.91

The research of parametric and nonparametric methods for determination data's with anomalistic distribution correlation coefficient/ G. Safaryan, M. Sergienko// Management Information System and Devices. All-Ukr. Sci. Interdep. Mag. 2000. N 00. P. 000-000.

There were analyzed an exactness of accidental data's correlation coefficient determination by parametric and nonparametric methods. The correction coefficients for these methods optimization were received. The distribution law does not have an influence on the determination correlation coefficient result. It was demonstrated by the example of data with normal, uniform, arc sine distributions and its combinations. The researchers are getting in cases of big and little reference quantity volumes.

Tab. 4. Fig. 2. Ref.: 6 items.