А. Ю. ДОЛГОВ

Молдова, г. Тирасполь, Приднестровский гос. ун-т им. Т. Г. Шевченко

Дата поступления в редакцию
18.05 1998 г.
Оппонент *к. т. н. Т. Д. БОРДЯ*

РОБАСТНАЯ МЕРА КОРРЕЛЯЦИОННОЙ СВЯЗИ

Найдена мера линейной корреляционной связи, устойчивая к наличию грубых промахов и к виду закона распределения случайных величин.

The measure of linear correlation relationship, stable to existence of gross blunders and to kind of normal variables has been found.

Одним из предварительных этапов математического моделирования любого технологического процесса массового и крупносерийного производства является корреляционный анализ собранной контрольно-измерительной информации. Такой анализ сокращает размерность факторного пространства путем удаления избыточных факторов без потери информации о процессе, например, в производстве кристаллов интегральных микросхем определяет степень однородности параметров по площади пластин, и т. п.

Для определения меры тесноты линейной корреляционной связи в математической статистике обычно используется коэффициент корреляции, который обладает рядом достоинств, позволяющих широко применять его на практике [1, с. 165]. Однако коэффициент корреляции требует, как правило, нормального закона распределения исходных случайных величин и очень чувствителен к наличию грубых промахов.

Давно доказано, что в цеховых условиях контрольные измерения обязательно содержат некоторое количество грубых промахов, а вид закона распределения может значительно отличаться от нормального. Поиск и удаление этих грубых промахов и преобразование вида закона распределения к нормальному представляют собой трудоемкую задачу, не всегда выполнимую в реальном процессе массового производства. Если учесть, что теоретических экспресс-методов определения грубых промахов в двухмерной совокупности не существует, то возникает потребность в такой мере тесноты связи, которая была бы свободна от закона распределения и устойчива к грубым промахам.

Одной из таких мер является индекс Фехнера [1], который, по свидетельству авторов, является показателем степени линейной взаимосвязи между двумя статистическими рядами. Суть его состоит в том, что числовые значения каждого из двух параметров заменяются знаком «—», если они оказались меньше своего среднего арифметического, и знаком «+» — если больше. Затем для каждой пары значений сравнивают присвоенные им знаки и подсчитывают количество одинаковых знаков у обоих параметров.

Обозначим это число через v, а количество пар с разными знаками — через w. Тогда индекс Фехнера находится как отношение

$$f = \frac{v - w}{v + w}. \qquad (1)$$

При $f>0$ имеем положительную корреляцию, при $f=0$ связь отсутствует.

Несомненное преимущество индекса Фехнера — простота вычисления. В силу того, что он учитывает только количество совпадений и несовпадений знаков, он свободен от закона распределения, но по этой же причине он грубее коэффициента корреляции и поэтому рекомендуется лишь для приблизительной оценки связи.

Предыдущими исследованиями [2] установлено, что при высокой степени тесноты связи (r=0,98) индекс Фехнера практически совпадает с коэффициентом корреляции и является более устойчивым к грубым промахам, чем коэффициент корреляции. Однако остался неясным вопрос о соответствии индекса Фехнера коэффициенту корреляции при меньшей степени тесноты связи. Выяснению соотношения этих показателей и посвящена настоящая работа.

Исследование проводилось в виде «машинного» (имитационного) эксперимента. Вначале были получены 130 парных выборок, каждая из которых состояла из 250 пар случайных чисел, сгенерированных с заданным коэффициентом корреляции. Затем они подвергались следующей обработке:

— каждый файл с выборкой, содержащей 250 пар чисел X и Y, проверялся на нормальность закона распределения как переменной X, так и переменной Y;

— каждая выборка подвергалась корреляционному анализу. Была получена таблица распределения частот для X и Y, а также вычислен реально получившийся коэффициент корреляции, корреляционное отношение, уравнение регрессии с коридором существования, а также построен график зависимости одного параметра от друго-

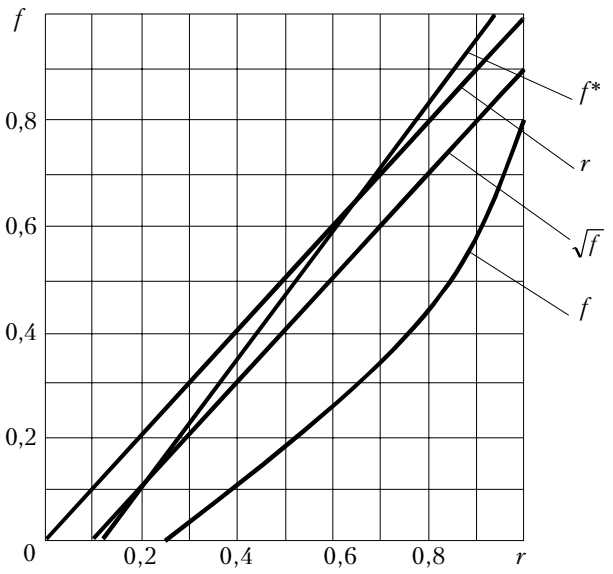

Рис. 1

го. В полученных 130 выборках использовались различные по величине коэффициенты корреляции, начиная с 0,33 (т. к. меньшая корреляция не имеет практического значения) и до 0,99;

— кроме коэффициента корреляции, для каждой выборки был найден и индекс Фехнера.

Таким образом была получена таблица для сравнения двух исследуемых величин. График, построенный по этой таблице, приведен на **рис. 1**.

Анализ таблицы и графика позволяет отметить, что индекс Фехнера в отдельных случаях отличается от коэффициента корреляции в 1,5—2 раза. Кроме того, индекс Фехнера становится сравним с коэффициентом корреляции только при значениях последнего около 0,95—0,99.

Проверка полученных параметров r и f на вид закона распределения и на тесноту взаимосвязи между ними дала следующие результаты:

— поскольку коэффициент корреляции в 130 выборках задавался таким образом, чтобы в итоге получился нормальный закон распределения, то это и подтвердилось (а индекс Фехнера имеет закон, отличный от нормального);

— используя метод Чебышева [3], получаем меру тесноты связи $\hat{r}$=0,975 и уравнение регрессии между коэффициентом корреляции r и индексом Фехнера f

$$f=1{,}237r-0{,}438 \qquad (2)$$

с коридором существования Δf =±0,070 (здесь и далее расчеты будут проводиться при доверительной вероятности P=0,95).

Приведение индекса Фехнера к нормальному закону распределения можно осуществить путем извлечения из него корня квадратного. Тогда мера тесноты связи между коэффициентом корреляции и корнем из индекса Фехнера станет равной r=0,992, а уравнение регрессии примет новый вид:

$$\sqrt{f} = 0{,}906r + 0{,}011 \qquad (3)$$

с коридором существования $\Delta\sqrt{f}$ =±0,029.

Поскольку в производственных условиях наличие грубых промахов неизбежно, а их выявление (и удаление) в парной выборке при различных величинах коэффициентов корреляции является трудоемкой процедурой, а также в силу того, что индекс Фехнера не оправдал ожиданий в полной мере, нами предлагается другая мера тесноты линейной корреляционной связи, робастная к наличию грубых промахов при парных выборках и достигающая 7% от общего объема парных выборок, — модифицированный индекс Фехнера (МИФ):

$$f^*=\sqrt{f}+0{,}051=0{,}906r+0{,}051 \qquad (4)$$

или

$$f^* = \sqrt{\frac{|v-w|}{v+w}} + 0{,}051. \qquad (5)$$

Предложенная новая мера тесноты линейной корреляционной связи (5) требует анализа ее поведения по сравнению с коэффициентом корреляции при нарушении классических предпосылок корреляционного анализа.

Для начала рассмотрим сравнительную величину МИФ и коэффициента корреляции в условиях *изменения объема парной выборки, распределенной по нормальному закону*. Для этого при обработке массивов исходных данных для двухмерных случайных величин X и Y были найдены коэффициенты корреляции r и модифицированные индексы Фехнера f. В результате выделились 12 двухмерных выборок f и r, обработка же этих выборок привела к целому ряду уравнений связи между ними. Из них были выбраны следующие наиболее подходящие уравнения (числовые индексы при МИФ соответствуют объему парной выборки):

1) $f^*_{250}=0{,}82r+0{,}15$ с коридором существования $\Delta f^*=\pm 0{,}046$;

2) $f^*_{200}=0{,}86r+0{,}10$ с коридором существования $\Delta f^*=\pm 0{,}045$;

3) $f^*_{100}=1{,}25r-0{,}23$ с коридором существования $\Delta f^*=\pm 0{,}147$;

4) $f^*_{60}=1{,}41r-0{,}38$ с коридором существования $\Delta f^*=\pm 0{,}144$;

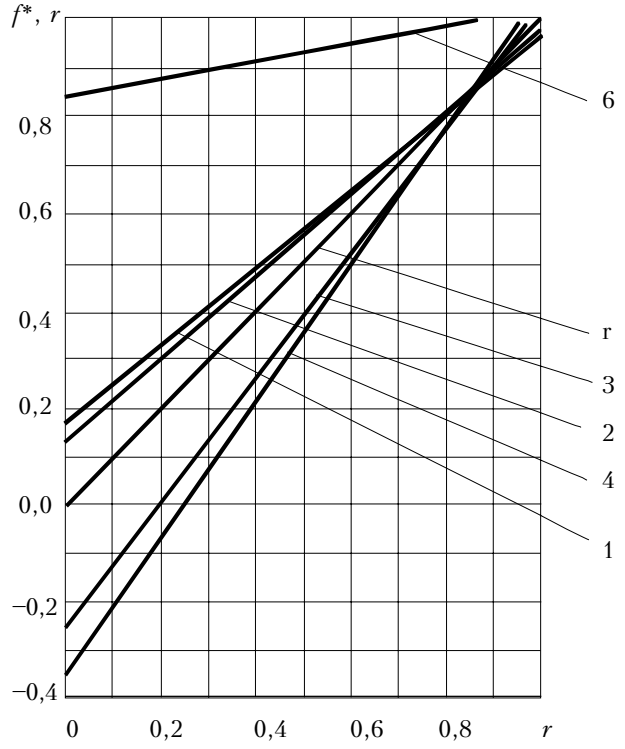

Рис. 2

5) f^*_{40}=0,94r+0,11 с коридором существования Δf^*=±0,130;

6) f^*_{20}=0,18r+0,86 с коридором существования Δf^*=±0,090.

Из **рис. 2** видно, что эти уравнения можно условно разделить на две группы. К одной из них относятся графики регрессий (*1*—*5*) с весовыми коэффициентами от 40 до 250, а к другой — график регрессии *6* с весовым коэффициентом 20. Кроме того, анализ графиков показывает достаточное совпадение коэффициента корреляции с модифицированным индексом Фехнера в пределах коридора существования, а также что между выборочными коэффициентом корреляции и МИФ существует связь (коэффициент корреляции между ними равен 0,955—0,994). Это означает практически полное совпадение двух мер тесноты корреляционной связи случайных величин для выборок большого объема (от 40 до 250) и резкое их отличие при объеме выборки, равном 20, следовательно, граница значимости находится в районе 30.

Отсюда можно сделать вывод о том, что при больших объемах выборок можно говорить о практической идентичности коэффициента корреляции и модифицированного индекса Фехнера, а при уменьшении объема выборки до тридцати границы значимости начинают сужаться и значение МИФ становится неприемлемым.

Следующее исследование относилось к поведению МИФ в условиях *изменения закона распределения.* Оно осуществлялось на материалах «машинного» (имитационного) эксперимента. В наличии имелись выборки пар случайных чисел X и Y, распределенных по нормальному закону, коэффициент корреляции между которыми задавался в диапазоне от 0,30 до 0,99. Объем выборок составлял соответственно 200 и 100 пар случайных чисел.

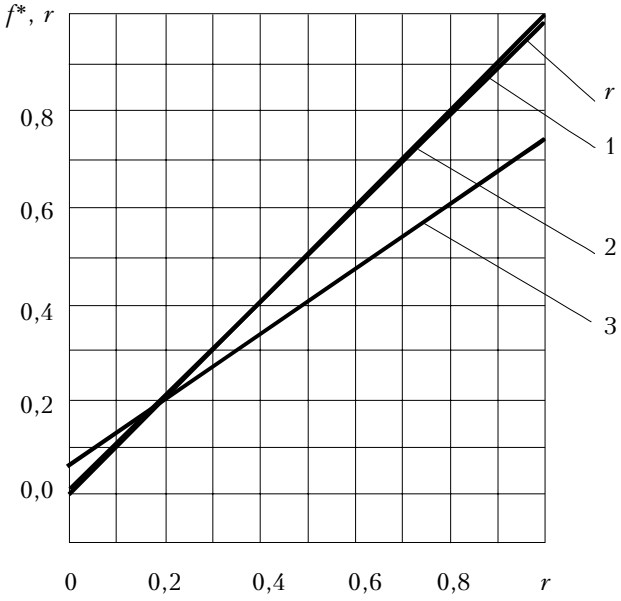

Рис. 3

В каждой из исследуемых выборок параметр Y, распределенный по нормальному закону, приводился к виду, отличному от нормального. Вершина искусственным путем сдвигалась вправо и влево. Для сдвига вправо значения Y логарифмировались, для сдвига влево — возводились в степень.

Для каждой выборки с внесенными в нее изменениями были найдены модифицированные индексы Фехнера для всего заданного диапазона коэффициентов корреляции. В процессе исследования сравнительной величины МИФ и коэффициента корреляции в условиях изменения закона распределения выделились 4 двухмерных выборки f^* и r, обработка которых привела к следующим уравнениям связи между ними:

при сдвиге вправо —

1) f^*_{200}=0,90r+0,08 с коридором существования Δf^*=±0,036;

2) f^*_{100}=1,10r–0,18 с коридором существования Δf^*=±0,047;

при сдвиге влево —

1) f^*_{200}=0,41r+0,33 с коридором существования Δf^*=±0,033;

2) f^*_{100}=0,83r–0,13 с коридором существования Δf^*=±0,112.

Чтобы лучше представить все вышесказанное, на **рис. 3** и **4** в одной системе координат приведены графики зависимостей каждого из параметров от закона распределения.

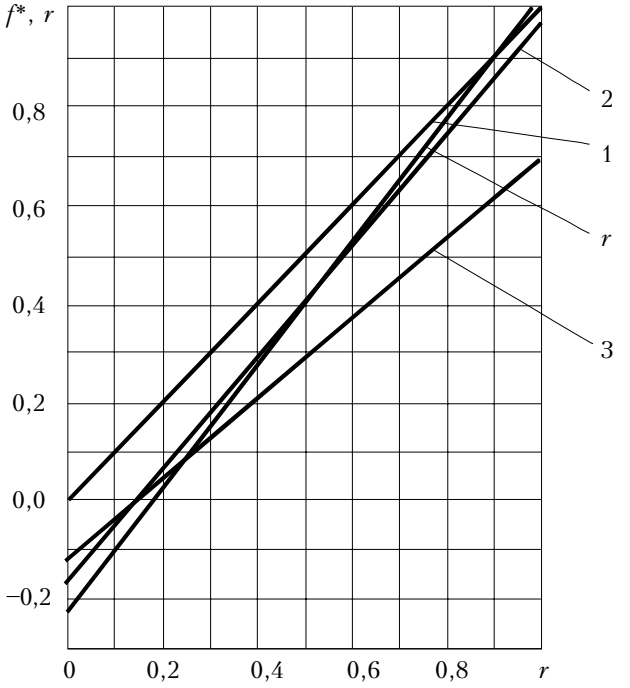

Рис. 4

Исследование этих рисунков дало следующие результаты. На рис. 3 график *2* регрессии, составленный по результатам обработки выборки объемом 200 при отличном от нормального законе распределения (сдвиг вправо — логарифмирование), практически полностью совпадает с графиком *1* модифицированного индекса Фехнера при нормальном законе распределения. Изменение закона распределения сдвигом влево (возведение в степень) дает некоторое отклонение графика регрессии *3* от графика МИФ *1* при нормальном законе распределения. То же можно сказать и о графиках регрессий на рис. 4, составленных по результатам обработки выборки объемом 100.

Итак, при сдвиге вправо коэффициент корреляции и модифицированный индекс Фехнера практически совпадают в пределах коридора существования, а при сдвиге влево наблюдается смещение и некоторое отклонение от биссектрисы прямого угла.

В результате можно сделать вывод о том, что при законе распределения, отличном от нормального, с изменением объема выборки модифицированный индекс Фехнера остается достаточно стабильным, кроме того, его численное значение ненамного отличается от его же собственных значений при нормальном законе распределения. Кроме того, МИФ по своему численному значению близок и к коэффициенту корреляции во всем диапазоне от 0,33 до 0,99.

В заключение можно сказать, что нами найдена мера тесноты линейной корреляционной связи, робастная к наличию грубых промахов (до 7% от объема парной выборки) и к виду закона распределения случайных величин (для унимодальных законов). При этом объем парных выборок в диапазоне свыше 100 незначительно влияет на ошибку определения модифицированного индекса Фехнера, а в диапазоне 30—100 составляет до 10% от истинного.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа.— М. : Финансы и статистика, 1983.

2. Шестакова Т. В., Долгов А. Ю. Повышение устойчивости метода корреляционных плеяд // Тез. докл. XXXIV СНТК вузов республик Прибалтики, Белоруссии и Молдавии, 30 янв.— 3 февр. 1990 г.— Каунас : КПИ, 1990.— Т. 2.— С. 108.

3. Митропольский А. К. Техника статистических вычислений.— М. : Наука, 1971