

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ТОЧНОЙ И ПРИБЛИЖЕННОЙ ФОРМУЛ ВЫЧИСЛЕНИЯ КОЭФФИЦИЕНТА РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЕНА

© 2016 Левит Борис Юльевич

кандидат экономических наук, профессор кафедры “Статистика”

© 2016 Салин Виктор Николаевич

кандидат экономических наук, профессор

Финансовый университет при Правительстве Российской Федерации

E-mail: tzeldner@gmail.com

Наряду с формулой вычисления коэффициента ранговой корреляции Ч. Спирмена (1904) в современной статистической литературе часто и без обоснований предлагается использовать более простую приближенную формулу. В статье изучается обоснованность такой замены и возникающая при этом погрешность.

Ключевые слова: ранговый коэффициент корреляции Спирмена, точное и приближенное значения, оценка погрешности, аналитические оценки и оценки на имитационной модели.

1. Аналитические оценки

Как известно, анализ корреляции с помощью рангового коэффициента корреляции Спирмена (далее для краткости - РККС или $\rho(X, Y)$) является одним из непараметрических методов изучения корреляционной связи между показателями X и Y . Математически X, Y - это векторы размерности n , компоненты x, y которых получены в результате выборочного наблюдения с объемом выборки, равным n . Случайность выборки обуславливает и случайность компонент x, y . РККС применяется в следующих двух случаях:

а) компоненты X, Y - количественные величины и закон распределения значений компонент хотя бы для одного из этих векторов не является нормальным. Применение РККС в таком случае часто обусловлено небольшим объемом выборки и, соответственно, малым значением n , при котором нормальность распределения значений x или y не устанавливается;

б) компоненты x, y хотя бы одного из векторов X, Y являются качественными и ранжируемыми (порядковыми, ординалистскими) величинами.

В любом из данных случаев применение РККС основано на замене компонент x, y их рангами Rx, Ry , которые далее и используются для вычисления $\rho(X, Y)$. Тот факт, что в формулах должны применяться *средние ранги*, часто в литературе или не оговаривается, или упоминается мельком.

В литературе приводятся две формулы вычисления РККС, которые мы назовем, соответ-

ственно, приближенной и точной. Приближенная форма РККС имеет вид

$$\rho_{\Pi} = 1 - \frac{6\sum(Rx - Ry)^2}{n^3 - n}. \quad (1)$$

Применение данной формулы по умолчанию предполагает, что среди компонент x и y нет повторяющихся значений, хотя в ряде литературных источниках она рекомендуется для вычисления РККС независимо от выполнения этого условия.

Точная формула вычисления РККС, предложенная Ч. Спирменом в 1904 г., явно учитывает наличие повторяющихся значений среди компонент хотя бы одного из векторов X, Y . В настоящее время в литературе точная формула представлена различными ее модификациями. Во всех таких модификациях группа одинаковых значений x или y называется *связкой*. Количество компонент, содержащихся в i -й связке, называется ее *мощностью* и обозначается k_{xi}, k_{yi} , при этом подчеркивается, что рассматриваются только $k_{xi} > 1, k_{yi} > 1$, т.е. значения x или y , встречающиеся в выборке однократно, ни в одну связку не входят. Так, если один из векторов имеет, например, вид $X = \{2, 5, 1, 2, 4, 6, 4, 2\}$, то в составе его компонент имеются две связки. Одна связка образована числом 2 и имеет мощность $k_1 = 3$, а другая образована числом 4 и имеет мощность $k_2 = 2$.

В соответствии с данными замечаниями ниже приводятся различные модификации точной формулы вычисления РККС, в которых точное значение РККС далее обозначается ρ . Так в¹ формула для вычисления ρ дается в виде

$$\rho = \frac{n^3 - n - 6\Sigma(Rx - Ry)^2 - 0.5(A + B)}{\sqrt{(n^3 - n - A)(n^3 - n - B)}}, \quad (2)$$

где $A = \Sigma(k_{xi}^3 - k_{xi})$, $B = \Sigma(k_{yi}^3 - k_{yi})$. (3)

Данная формула получена преобразованием аналогичной формулы из² и упрощением содержащихся там обозначений.

В³ точная формула для ρ приведена в следующей форме:

$$\rho = 1 - \frac{\Sigma(Rx - Ry)^2}{\frac{1}{6}(n^3 - n) - (Tx + Ty)}. \quad (4)$$

В этой формуле $Tx = A/12$, $Ty = B/12$, (5)

а значения A и B вычисляются по (3).

Все упомянутые точные формулы эквивалентны и приводят к одинаковым результатам.

Если среди x и y нет повторяющихся значений, то, соответственно, $A=B=Tx=Ty=0$ и формулы преобразуются в (1), т.е. в этом случае $\rho = \rho_n$. Как указывалось выше, проблема состоит в том, что в ряде, даже очень серьезных, источников, например, таких как⁴ и др., точное значение ρ даже не упоминается и РККС предлагается считать по формуле (1) во всех случаях. Такая рекомендация кажется естественной особенно для количественных векторов X , Y , поскольку повторяющиеся компоненты можно заменить за очень малую величину так, чтобы они стали бы различными. При этом, казалось бы, коэффициент РККС должен также измениться очень незначительно. Однако это не очевидно, поскольку РККС не является непрерывной функцией компонент векторов X , Y . Поэтому целью данной работы выступает анализ допустимости замены ρ на ρ_n и оценка возникающей при этом погрешности.

В качестве отправной формулы для анализа используем (4), которую очевидным преобразованием и с учетом (5) представим в виде

$$\begin{aligned} \rho &= 1 - \frac{\Sigma(Rx - Ry)^2}{\frac{1}{6}(n^3 - n) - (Tx + Ty)} = \\ &= 1 - \frac{6\Sigma(Rx - Ry)^2}{n^3 - n - 6(Tx + Ty)} = \\ &= 1 - \frac{6\Sigma(Rx - Ry)^2}{n^3 - n - 0.5(Ax + By)}. \end{aligned} \quad (6)$$

Далее отметим, что если в число связей включить и координаты, значение которых встре-

чается однократно и при этом естественным образом считать мощность таких связей $k=1$, то значения A и B не изменятся, поскольку $1^3-1=0$.

Однако в этом случае значения k_i будут представлять собой частоты f_{xi} , f_{yi} , с которыми каждая компонента содержится в векторах X и Y . Как это часто делается в статистической литературе, далее индекс i для краткости опустим и указанные частоты будем записывать как f_x , f_y . Обозначение же просто f без индекса будет означать, что не важно, о какой из частот f_x , f_y идет речь.

С учетом сказанного, а также того, что $\Sigma f_x = \Sigma f_y = n$, выражения для A и B могут быть записаны следующим образом:

$$\begin{aligned} A &= \Sigma(f_x^3 - f_x) = \Sigma f_x^3 - n, \\ B &= \Sigma(f_y^3 - f_y) = \Sigma f_y^3 - n. \end{aligned} \quad (7)$$

В данном случае (6) можно представить в виде

$$\rho = 1 - \frac{6\Sigma(Rx - Ry)^2}{n^3 - 0.5(\Sigma f_x^3 + \Sigma f_y^3)}. \quad (8)$$

Отметим, что при условиях $\Sigma f = n$ и $0 < f \leq 1$ всегда $\max \Sigma f^3 \geq \Sigma f = n$ и при этом $\Sigma f^3 = \Sigma f$, только если все $f=1$. Поэтому, если в (8) суммы Σf^3 заменить на n , то знаменатель дроби в правой части (8) не уменьшится, а может и увеличиться, сама дробь при этом уменьшится, а вся правая часть (8) увеличится. В результате получаем:

$$\begin{aligned} \rho &= 1 - \frac{6\Sigma(Rx - Ry)^2}{n^3 - 0.5(\Sigma f_x^3 + \Sigma f_y^3)} \leq 1 - \\ &= \frac{6\Sigma(Rx - Ry)^2}{n^3 - 0.5(n + n)} = 1 - \frac{6\Sigma(Rx - Ry)^2}{n^3 - n} = \rho_n. \end{aligned} \quad (9)$$

Опустив среднюю часть выкладок в (9), окончательно имеем:

$$\rho \leq \rho_n. \quad (10)$$

Как указывалось, точное равенство $\rho = \rho_n$ возможно, только если все $f = 1$. Если же отбросить такую ситуацию, то $\rho < \rho_n$ и в этом случае доказано следующее:

а) при $\rho > 0$ переход от ρ к ρ_n завышает силу связи;

б) при $\rho < 0$ переход от ρ к ρ_n занижает силу связи.

Следующим шагом является оценка относительной погрешности δ , возникающей при переходе от ρ к ρ_n . При этом погрешность δ вычисляется следующим образом:

$$\delta = \max |(\rho_n - \rho) / \rho| = \max |\rho_n / \rho - 1|. \quad (11)$$

Однако в аналитическом виде сделать этого не удастся, и потому оценка δ выполняется ме-

тодом Монте-Карло посредством имитационной модели, описанной ниже.

2. Имитационная модель оценки δ

2.1. Базовая модель

Для обеспечения наглядности модели она строится в Excel с приведением всех формул, используемых в расчетах. Основой модели служит таблица на рис. 1, где значения ρ , ρ_n и δ вычисля-

чения далее будут называться *метками*. Метки n , ρ , ρ_n , содержащиеся, соответственно, в ячейках **A10**, **C16**, **C17**, обозначают ячейки, расположенные справа от них. Отметим также, что все метки представлены в обычном математическом виде с использованием верхних и нижних индексов.

2. Для обеспечения наглядности расчетных формул Excel ссылки в них созданы не в форме адресов ячеек, а посредством содержательных имен, назначенных диапазонам или отдельным

	A	B	C	D	E	F	G	H	I	J
1										
2	Исходные данные			Средние ранги		$(R_x - R_y)^2$	Мощности связок		Учет связок	
3	n_{ij}	X	Y	R_x	R_y	d^2	k_x	k_y	k_x^2	k_y^2
4	1	18,0	1,42	6,0	5,0	1,00	1	1	1	1
5	2	10,2	1,28	9,0	9,5	0,25	1	2	1	4
6	3	10,5	1,28	7,5	9,5	4,00	2	2	4	4
7	4	10,5	1,30	7,5	7,5	0,00	2	2	4	4
8	5	9,7	1,30	10,0	7,5	6,25	1	2	1	4
9	6	19,0	1,37	5,0	6,0	1,00	1	1	1	1
10	7	20,0	1,54	4,0	2,0	4,00	1	1	1	1
11	8	21,0	1,50	2,5	3,5	1,00	2	2	4	4
12	9	21,0	1,50	2,5	3,5	1,00	2	2	4	4
13	10	22,0	1,56	1,0	1,0	0,00	1	1	1	1
14	n	10				Σd^2			Σk_x^2	Σk_y^2
15						18,50			22	28
16		ρ	0,8862	=1-6* $\Sigma d^2 / (n^3 - 0,5 * (\Sigma k_x^2 + \Sigma k_y^2))$						
17		ρ_n	0,8879	=1-6* $\Sigma d^2 / (n^3 - n)$						
18		$\delta = \rho_n - \rho \cdot 100\%$	0,195%	=ABS($\rho_n / \rho - 1$)						

Рис. 1. Вычисление точного и приближенного значений коэффициента Спирмена

ются для некоторых векторов X, Y в примере небольшой размерности с $n=10$. Относительно этой таблицы и всех последующих необходимо сделать следующие замечания.

1. В стр. 3 содержатся обозначения, введенные в первой части изложения и поясняющие содержание конкретных столбцов. Аналогичные обозначения содержатся в стр. 14. Такие обозна-

чения далее будут называться *метками*. Имена ячеек/диапазонов создаются с помощью обозначающих их меток. Однако имена в Excel не могут содержать верхние и нижние индексы. Поэтому при использовании меток в качестве имен происходит некоторое преобразование меток, однако не снижающее смысловую наглядность имен. Так, метки следующим образом преобразованы в имена:

Метки	X	Y	d^2	k_x	k_y	k_x^2	k_y^2	n	Σd^2	Σk_x^2	Σk_y^2	ρ	ρ_n
Имя	X	Y	d2	kx	ky	kx2	ky2	n	Σd^2	Σkx^2	Σky^2	ρ	ρ_n
Что обозначает	B4:B13	C4:C13	F4:F13	E4:E13	G4:G13	H4:H13	I4:I13	B14	F15	I15	J15	D16	D17

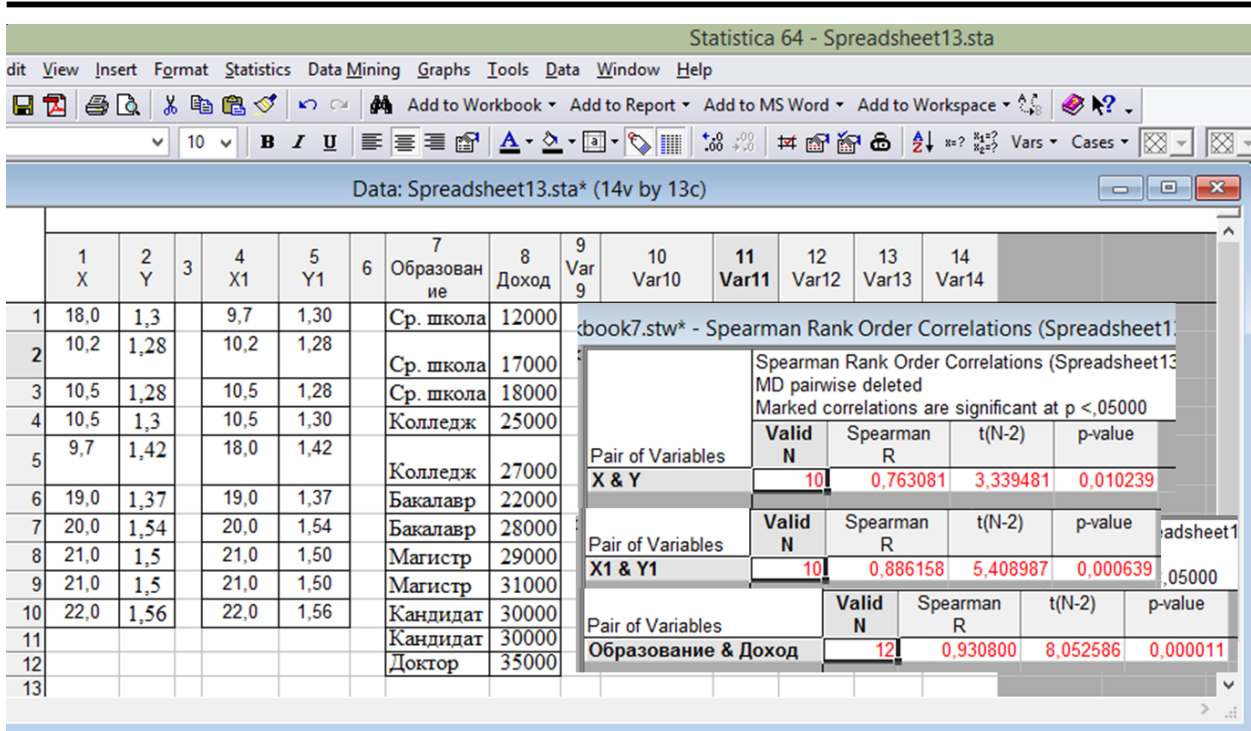


Рис. 2. Расчеты РККС по программе Statsoft Statistica

3. Расчетные формулы, используемые для построения таблицы, приведены или в выносках, концы которых показывают на ячейку с формулой (см. стр. 2-4), или рядом с ячейкой, содержащей формулу (см. ячейки E16, E17, в которых показано, как вычислены ρ и ρ_{Π}). Двойная линия, окаймляющая диапазон D4:J13, означает, что ячейки первой строки диапазона скопированы вниз.

Описание действия функций, используемых в таблице и показанных, например, в выносках, выходит за рамки статьи. Представляется, что их смысл достаточно ясен. Покажем только, что $\Sigma k_x^2 = \Sigma f_x^3$, $\Sigma k_y^2 = \Sigma f_y^3$, но при этом в сумме, например, Σk_x^2 слагаемые k_x^2 заданы для каждой компоненты x и их число равно n . Количество же слагаемых в правой части равенств Σf_x^3 , Σf_y^3 равно количеству различных компонент в X , Y , т.е. количеству вариантов для координат этих векторов. Сказанное следует из того, что k_{xi}^2 - это квадрат частоты f_i компоненты x_i . В сумме же Σk_x^2 по всем компонентам x каждое слагаемое k_x^2 встречается k_x раз. Поэтому, сгруппировав слагаемые для одинаковых значений x (для вариантов X), можно записать $\Sigma k_x^2 = \Sigma(k_x * k_x^2) = \Sigma k_x^3 = \Sigma f_x^3$, где все суммы в приведенных равенствах, кроме первой, берутся для вариантов X . Аналогичным образом объясняется и равенство $\Sigma k_y^2 = \Sigma f_y^3$.

Далее таблицу на рис. 1 назовем базовой моделью и будем обозначать БМ. Отметим, что в БМ

не требуется какой-либо упорядоченности компонент X , Y , что существенно для дальнейшего.

Для проверки правильности работы БМ для той же пары (X, Y) значение ρ было вычислено программой Statsoft Statistica (v. 12), результаты которой приведены на рис. 2. Расчеты по ней показали, что для правильного вычисления коэффициента ρ требуется, чтобы пары (X, Y) были упорядочены по возрастанию или убыванию значений компонент X . Так, на рис. 2 для неупорядоченных пар (X, Y) неправильно вычислен коэффициент $\rho=0,7631$. Но при их упорядочении по X (см. пары X1, Y1) рассчитанное значение $\rho=0,8861$ совпадает с полученным в БМ.

Попутно отметим, что программа Statistica правильно вычисляет ρ в случае, когда X является качественным вектором, а его компоненты x упорядочены в порядке их ранжирования. Так, на рис. 2 на условном примере определяется РККС между уровнем образования респондента и его доходом (переменные 7, 8). При этом получено $\rho=0,9308$.

2.2. Имитационная модель

“Голая” идея имитационной модели (далее - ИМ) состоит в том, чтобы увеличить размерность БМ, например до $n=50$, в ней случайным образом многократно генерировать пары (X, Y) , наблюдая при этом за изменением ρ, ρ_{Π}, δ и набирая необходимую статистику. Однако для этого реализация ИМ требует выполнения следующих дополнительных условий:

1) случайные векторы (X, Y) необходимо генерировать таким образом, чтобы среди координат x и y имелись группы совпадающих значений (связки);

2) в процессе генерации необходимо получать малые, средние и большие значения ρ , с тем, чтобы определить влияние величины ρ на погрешность δ .

1. Способ генерации случайного вектора X . Назовем пары (X, Y) и (U, V) эквивалентными, если:

а) размерности векторов в этих парах совпадают и равны n ;

б) $\rho(X, Y) = \rho(U, V)$.

Утверждение. Для любой пары (X, Y) можно построить эквивалентную пару (U, V) , в которой компоненты векторов U, V состоят из первых чисел натурального ряда, не превосходящие n .

Действительно, последовательно пронумеруем компоненты вектора X в порядке их возрастания, присваивая одинаковым компонентам (компонентам связки) один и тот же номер, и эти номера возьмем в качестве компонент вектора U . Максимальное значение компонент U будет не более n , а при наличии связок строго меньше n . Поскольку средние ранги Ru будут такими же, как и Rx , то значение $\rho(U, Y)$ не изменится. Продолжив аналогичную операцию для вектора Y , получим пару (U, V) , эквивалентную (X, Y) .

Из сделанного утверждения следует, что без потери общности результатов вектор X в ИМ можно генерировать так, чтобы его компонентами были случайные целые числа, не превосходящие заданного числа m_x . Такие числа генерируются в Excel с равными вероятностями посредством функции

$$=\text{СЛУЧМЕЖДУ}(1; m_x). \quad (12)$$

При таком способе генерации случайных значений X очевидно, что при $m_x \ll n$ среди n сгенерированных чисел будут повторяющиеся, т.е. будут образовываться связки. Количество связок случайно, но стохастически зависит от соотношения m_x и n . Если $m_x \approx 0,3n$, то мощности связок будут колебаться вокруг $k=3$. В этом случае при размерности вектора X (или, что то же самое, размерности БМ) $n=50$ можно принять $m_x=15$.

Важно отметить, что при любой операции на рабочем листе функция СЛУЧМЕЖДУ($a; b$), где $a < b$, каждый раз заново генерирует случайное целое число $\lambda \in [a; b]$.

2. Способ генерации случайного вектора Y . Случайный целочисленный вектор Y будет

генерироваться в виде $Y=X+\Delta_y$, где Δ_y - случайное целое число, удовлетворяющее условию $1 < \Delta_y < m_y$. Величина m_y - это положительный (пока) целочисленный параметр управления моделью. Случайное значение Δ_y для заданного m_y генерируется функцией =СЛУЧМЕЖДУ(1; m_y). Очевидно, что при $m_x > 0$ имеем $\rho > 0$ и чем меньше m_y , тем больше ρ . Отрицательные значения ρ будут получаться, если значения Y генерировать в виде $Y=-X+\Delta_y$. По абсолютной величине и в этом случае значение ρ уменьшалось с ростом m_y .

С целью упрощения управлением ИМ в ней сделано так, чтобы знак ρ соответствовал знаку m_y . Достигается это посредством следующей функции Excel:

$$=\text{ЕСЛИ}(m_y > 0; X + \text{СЛУЧМЕЖДУ}(1; m_y); \text{СЛУЧМЕЖДУ}(1; -m_y) - X). \quad (13)$$

Следует отметить, что генерируемые таким образом компоненты X и Y будут целочисленными, но по абсолютной величине не обязательно составят все значения начального отрезка натурального ряда, что в нашем случае не имеет значения.

В соответствии со сказанным базовая модель будет иметь вид, приведенный на рис. 1, со следующими отличиями:

- для компонент векторов X, Y отводится по 50 строк и потому в ячейке В14 будет $n=50$;

- во всех ячейках столбца X содержится формула (12), а в ячейках столбца Y формула (13).

- значения m_x и m_y задаются, соответственно, в ячейках В70 и В1072, как это показано на рис. 3.

3. Многократная работа БМ. Следующая задача состоит в том, чтобы для заданных значений m_x и m_y “заставить” БМ работать многократно, каждый раз генерируя различные случайные пары векторов (X, Y) , для которых вычислять ρ, ρ_p, δ и собирать для них статистику. Такой процесс может быть осуществлен посредством инструмента Excel (называемого Таблица данных), который организуется посредством команды Данные/Анализ “что если”/Таблица Данных (далее просто ТД). Перед выполнением этой команды на рабочем листе должна быть организована “заготовка”, управляющая работой команды ТД. Эта “заготовка” для рассматриваемой задачи приведена на рис. 3 в диапазоне В70-G1070, обведенном на рисунке двойной линией (строки 74-1067 скрыты, чтобы сократить размер рисунка). Все, что находится вне указанного диапазона, в ТД не входит и служит только для пояснения ее работы. Так, метки в диапазоне

	A	B	C	D	E	F	G	H
68	Таблица данных							
69	№	m_x	ρ	$\rho_{п}$	δ	Σk_x^2	Σk_y^2	
	п.п.							
70		15	0,774	0,776	0,256%	1724	566	
71	1	15	0,7981	0,7996	0,19%	1226	740	
72	2	15	0,814	0,8152	0,15%	1136	608	
73	3	15	0,7802	0,7815	0,17%	914	698	
1068	998	15	0,8278	0,8291	0,15%	1322	608	
1069	999	15	0,8524	0,8534	0,11%	1004	686	
1070	1000	15	0,7596	0,7611	0,19%	860	704	
1071	Статистика, собираемая по Таблице данных							
1072	m_y	10	ρ	$\rho_{п}$	δ	Σk_x^2	Σk_y^2	
1073	Статистики	Мин	0,6286	0,6313	0,058%	704	404	
1074		Макс	0,8946	0,8951	0,517%	2366	1766	
1075		Средн	0,7985	0,7999	0,180%	1084	747	
1076		σ	0,0505	0,0501	0,075%	236	195	
1077		Δ	0,0044	0,0044	0,007%	21	17	
1078		P	0,00%	=СЧЁТЕСЛИ(E71:E1070;">3%")/1000				

Рис. 3. Исходная информация для команды Таблица данных и статистика, собранная в результате ее работы

C69-G69 поясняют, что лежащие под ними в следующей строке данные берутся из соответствующих ячеек базовой модели на **рис. 1** и такой же смысл имеют нижележащие по столбцу данные.

Работа ТД в нашем случае состоит в том, что последовательно следующим образом обрабатывается каждая строка диапазона **B71-G1070**:

- значение m_x из очередной обрабатываемой (далее текущей) строки заносится в ячейку **B70**, которое используется при генерации вектора X по формуле (12). В нашем случае в **B70** заносится одно и то же заданное значение m_x (на рисунке $m_x=15$), хотя в принципе они могут быть различными;

- поскольку при этом на листе происходят изменения, в БМ заново генерируются векторы (X, Y) ;

- для сгенерированных пар (X, Y) базовой моделью вычисляются показатели, указанные в первой строке ТД, и их значения заносятся в текущую строку ТД.

Таким образом, генерируется столько случайных пар (X, Y) и для них вычисляются соответствующие показатели, сколько строк содержится в ТД (в нашем случае 1000).

В стр. **1073-1078** по данным столбцов ТД вычисляются статистики соответствующих показателей, смысл которых представляется ясным. Отметим только, что Δ - это доверительный интервал по Стьюденту для соответствующего показателя, а P - доля случаев, в которых величина погрешности превышала 3 %.

Отметим, что при работе ТД значение m_x оставалось постоянным, равным $m_x=10$ на рис. 3 (см. ячейку **B1072**). На рис. 4 показана статистика результатов работы ТД при различных m_y .

Результаты на рисунке подтверждают приведенное в теоретической части доказательство, что всегда $\rho < \rho_{п}$. Кроме того, наблюдается рост средней ошибки δ при уменьшении ρ . Однако главный интерес представляет не само среднее значение ошибки (хотя и это интересно), а вероятность того, что ошибка δ превысит некоторое пороговое значение, например равное 3%. Таким образом, для различных m_y необходимо вычислить $P\{\delta > 3\%\}$. Для этого следует при соответствующем значении m_y многократно повторять работу ТД и набирать статистику значений P , вычисляемых на рис. 3 в ячейке **C1078**.

4. Вычисление вероятности $P\{\delta > 3\%\}$ для различных значений m_y (т.е. для различных ρ). “Заставить” ТД проработать нужное количество раз при фиксированном m_y возможно с помощью инструмента Excel “**Диспетчер сценариев**”, который вызывается командой **Данные/Анализ “что если”/Диспетчер сценариев**. Выполнению этой команды должна предшествовать подготовка сценариев расчетов. Не вдаваясь в подробности описания этой команды, скажем, что в каждом сценарии указывается, какая ячейка или ячейки должны быть изменены и на что, а также данные из каких ячеек следует

	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
60	<i>m_y</i>	10					20					40				
61		ρ	ρn	δ	Σk_x^2	Σk_y^2	ρ	ρn	δ	Σk_x^2	Σk_y^2	ρ	ρn	δ	Σk_x^2	Σk_y^2
62	Мин	0,700	0,703	0,057%	650	392	0,247	0,252	0,142%	680	230	-0,074	-0,068	0,210%	668	146
63	Макс	0,914	0,914	0,506%	2396	1490	0,796	0,798	2,178%	2672	1166	0,685	0,687	191,029%	2090	722
64	Средн	0,829	0,831	0,147%	1063,2	769,3	0,582	0,584	0,458%	1066,0	498,7	0,327	0,330	1,896%	1064,9	272,9
65	σ	0,037	0,037	0,056%	236	185,6	0,093	0,093	0,239%	241	125,8	0,123	0,122	8,866%	219,8	71,5
66	Δ	0,003	0,003	0,005%	21	16,3	0,008	0,008	0,021%	21	11,1	0,011	0,011	0,779%	19,3	6,3

Рис. 4. Результаты работы ТД при различных значениях m_y

Фрагмент отчета Диспетчера сценариев с вероятностями $P\{\delta > 3\%\}$				Статистика вероятностей $P\{\delta > 3\%\}$			
<i>m_y</i>	10	20	40	<i>m_y</i>	10	20	40
ρ	0,829	0,582	0,327	ρ	0,829	0,582	0,327
$P\{\delta > 3\%\}$				$P\{\delta > 3\%\}$			
1	0,00%	0,00%	7,60%	Мин	0,00%	0,00%	6,50%
2	0,00%	0,00%	7,90%	Макс	0,00%	0,20%	9,60%
...	Средн	0,00%	0,03%	7,95%
49	0,00%	0,00%	8,00%	σ	0,00%	0,06%	0,70%
50	0,00%	0,10%	7,50%	Δ	#ЧИСЛО!	0,02%	0,20%

Рис. 5. Отчет Диспетчера сценариев и статистики для $P\{\delta > 3\%\}$

“наблюдать”. В нашем случае было создано 50 сценариев, в каждом из которых в ячейку m_y (в ячейку **V1072** на рис. 3) заносилось одно и то же значение m_y , а в качестве наблюдаемого указывалось значение P на рис. 3. В результате 50 раз прорабатывала ТД и Диспетчер сценариев выдавал отчет с наблюдаемыми значениями P . В процессе моделирования посредством Диспетчера сценариев было создано три группы сценариев для $m_y=10, 20, 40$, и для каждой такой группы было получено по 50 значений P . Фрагмент отчета Диспетчера сценариев со значениями P , полученными для различных m_y , приведен в левой таблице на рис. 5, а в правой - статистика этих значений.

Анализ правой таблицы показывает, что с уменьшением ρ погрешность δ заметно возрастает. Объем же выборки в 50 значений P позволяет получить достаточно небольшой доверительный интервал для среднего значения δ при уровне значимости 5 %. Значение #ЧИСЛО в таблице справа означает, что при большом значении ρ ошибка δ ни разу не превысила 3 %, при этом среднеквадратическое отклонение σ для P , естественно, оказалось равным нулю и доверительный интервал функцией ДОВЕРИТ.СТЮДЕНТ не может быть вычислен.

Заметим, что хотя выше приводятся результаты для положительных $\rho > 0$, однако сказанное остается справедливым и при отрицательной корреляции.

Выводы

1. Теоретически и экспериментально доказано, что использование приближенной формулы вычисления ρ завышает силу корреляционной связи при $\rho > 0$ и уменьшает при $\rho < 0$.

2. Относительная ошибка δ при использовании приближенной формулы для вычисления ρ пренебрежимо мала при достаточно больших значениях ρ и весьма заметна при его малых значениях. Однако, поскольку априорная величина ρ неизвестна, при получении малого приближенного значения ρ это значение рекомендуется пересчитать по точной формуле.

¹ Статистика : учеб. пособие / кол. авт. ; под ред. В.Н. Салина, Е.П. Шпаковской. 2-е изд., перераб. и доп. Москва, 2014.

² Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей : справ. изд. / под ред. С.А. Айвазяна. Москва, 1985.

³ Теория статистики : учебник / Р.А. Шмойлова [и др.] ; под ред. Р.А. Шмойловой. Москва, 2003.

⁴ См.: Справочник по прикладной статистике. В 2 т. Т. 2 / под ред. Э. Ллойда, У. Ледермана ; пер. с англ. под ред. С.А. Айвазяна, Ю.Н. Тюрина. Москва, 1990; Математическая статистика : учеб. для вузов / В.Б. Горяинов [и др.] ; под ред. В.С. Зарубина, А. П. Крищенко. 2-е изд., стереотип. Москва, 2002. Серия “Математика в техническом вузе”. Вып XVII.