Ошибки округления в статистических расчетах

© 2016 Левит Борис Юльевич
кандидат экономических наук, доцент
© 2016 Салин Виктор Николаевич
кандидат экономических наук, профессор
Финансовый университет при Правительстве Российской Федерации
125993, г. Москва, Ленинградский пр-т, д. 49
E-mail: tzeldner@gmail.com

По различным, в том числе методическим, соображениям в ряде вузов решение задач по общей теории статистики на аудиторных занятиях выполняется с помощью калькуляторов. Одновременно стимулируется решение подобных задач на компьютерах средствами Excel при выполнении домашних заданий и лабораторных работ. При этом, как показывается в статье, различная точность устройств и различия в технологии вычислений могут привести к существенно различающимся, а порой и бессмысленным результатам.

Ключевые слова: общая теория статистики, использование калькуляторов, погрешности вычислений, точность в Excel.

Постановка задачи, точность вычислений в Excel. Проблема, обозначенная в названии статьи, может показаться утратившей свою актуальность, поскольку выполнение статистических вычислений на компьютере практически сводит к нулю погрешность, обусловленную округлением промежуточных результатов и уменьшением их точности. Однако в процессе изучения курса "Общая теория статистики" (далее - ОТС) все статистические вычисления выполняются студентами с помощью простейших калькуляторов, как правило, с точностью девять десятичных разрядов и с использованием одного регистра памяти. Такой способ вычислений далее кратко будет называться *калькуляторным*. Несмотря на значительную трудоемкость таких вычислений, они представляются методически обоснованными, поскольку позволяют студентам лучше уяснить смысл и алгоритмы решения основных статистических задач. В процессе калькуляторных расчетов регистр памяти позволяет вычислить ряд показателей (например, средние значения, среднеквадратические отклонения и др.) с достаточно высокой точностью, однако при использовании этих показателей в последующих расчетах они обычно округляются до двух-трех десятичных знаков. Такое же округление часто производится и в примерах расчетов в учебниках и учебных пособиях по ОТС. Приводимые ниже примеры показывают, насколько подобные, казалось бы, малозначимые округления могут привести к отклонению получаемых при таком округлении результатов от их точных значений, вычисленных в Excel. Примеры, приведенные ниже, взяты из[1].

Чтобы продемонстрировать последствия округлений, напомним, что в Excel существуют следующие два режима вычислений:

- ***режим полной точности***, при котором все вычисления выполняются и все результаты сохраняются с 15 десятичными знаками. Значение числа в ячейке и в расчетах в этом режиме не изменяется при уменьшении количества десятичных знаков, с которыми оно представляется на мониторе ПК. В Excel данный режим обычно установлен по умолчанию;
- ***режим экранной точности***, в котором значение числа в ячейке и в расчетах всегда совпадает с его представлением на экране монитора. В режиме экранной точности истинные значения *необратимо и автоматически обрезаются до их экранной формы,* и такие урезанные значения используются в вычислениях. В Excel 2007 этот режим устанавливается пометкой флага, открывающегося последовательными щелчками на кнопках (командах): Кнопка Office/Параметры Excel/Дополнительно/Задать точность как на экране (группа флажков При пересчете этой книги).

Поясним влияние режима вычислений на результаты расчетов на примере таблицы, приведенной на рис. 1.

	B	C	D	E
10	Исходные данные		Полная точность	Точность как на экране
11	U	V	U+V	U+V
12	2,342	3,324	5,666	5,666
13	2,34	3,32	5,67	5,66
14	2,3	3,3	5,7	5,6
15	2	3	6	5

Рис. 1. **Зависимость суммы чисел от режима вычислений**

Во всех ячейках столбцов *B*:*C* на рисунке содержатся одни и те же значения *U*=2,342 и *V*=3,324, но представленные в различных строках с различной точностью. В столбце *D* приведена сумма *U*+*V*, вычисленная в режиме полной точностью и представленная с различным количеством десятичных знаков. Результаты в диапазоне *D*13:*D*15 кажутся странными (особенно 2+3=6), если не помнить, что в режиме полной точности во всех строках используются *одни и те же* значения *U*=2,342 и *V*=3,324, а полученная их сумма во всех строках равна 5,666, но округлена по правилам округления до соответствующего количества разрядов. В столбце же *E* показаны результаты, получаемые в режиме экранной точности. В этом случае в строках таблицы суммируются *различные* значения *U* и *V*, равные показанным в столбцах *B*:*C*, что и приводит к естественным результатам в столбце *E*.

Погрешность при вычислении коэффициентов корреляции. Рассмотрим влияние погрешностей округления на примере вычисления различных коэффициентов корреляции между переменными *y*, x_1, x_2, выборочные значения которых приведены на рис. 2 в диапазоне *C*4:*E*15. Исходные данные заданы с двумя десятичными знаками, а расчетные значения вычислены с четырьмя знаками после запятой (ради экономии места у расчетных значений в таблице, содержащих менее четырех знаков, отброшены последние незначащие нули).

Значения среднеквадратических отклонений в диапазоне *I*19:*K*19 таблицы вычислены по следующей известной формуле, в которой *n*=12 - это объем выборки:

$$\sigma_u = \sqrt{\sum u^2 / n - \bar{u}^2}. \qquad (1)$$

В данную формулу вместо *u* подставлялись, соответственно, значения *y*, x_1, x_2.

Приведенная на рис. 2 таблица воспроизводит калькуляторную технологию вычисления коэффициентов. Однако при использовании функций Excel СРЗНАЧ(...) и СТАНДОТКЛОНП(...) средние значения и среднеквадратические отклонения, приведенные в стр. 19, могут быть сразу вычислены по исходным данным с точностью 15 десятичных знаков независимо от установленного режима вычислений (полного или экранного). При использовании регистра памяти, имеющегося практически во всех калькуляторах, калькуляторные расчеты также обеспечивают приведенную в таблице точность средних и среднеквадратических значений. Значительная погрешность возникает, когда найденные с высокой точностью в стр. 19 средние и среднеквадратические значения округляются до 2-3 знаков при вычислении парных (r_{x_1y}, r_{x_2y}, $r_{x_1x_2}$), частных ($r_{x_1y(x_2)}$; $r_{x_2y(x_1)}$) и множественного $R_{y(x_1x_2)}$ коэффициентов корреляции по формулам:

$$r_{x_1y} = \frac{\overline{x_1 \cdot y} - \bar{x}_1 \cdot \bar{y}}{\sigma_{x_1}\sigma_y},\ r_{x_2y} = \frac{\overline{x_2 \cdot y} - \bar{x}_2 \cdot \bar{y}}{\sigma_{x_2}\sigma_y},\ r_{x_1x_2} = \frac{\overline{x_1 \cdot x_2} - \bar{x}_1 \cdot \bar{x}_2}{\sigma_{x_1}\sigma_{x_2}}; \quad (2)$$

$$r_{x_1y(x_2)} = \frac{r_{x_1y} - r_{x_2y} \cdot r_{x_1x_2}}{\sqrt{(1-r_{x_2y}^2)(1-r_{x_1x_2}^2)}},\ r_{x_2y(x_1)} = \frac{r_{x_2y} - r_{x_1y} \cdot r_{x_1x_2}}{\sqrt{(1-r_{x_1y}^2)(1-r_{x_1x_2}^2)}}; \quad (3)$$

$$R_{y(x_1x_2)} = \sqrt{\frac{r_{x_1y}^2 + r_{x_2y}^2 - 2r_{x_1y} \cdot r_{x_2y} \cdot r_{x_1x_2}}{1-r_{x_1x_2}^2}}. \qquad (4)$$

Средние и среднеквадратичные значения, используемые в формулах (2)-(4) при вычислении коэффициентов корреляции, будем далее для краткости называть *операндами*.

	B	C	D	E	F	G	H	I	J	K
2	Исходные данные				Расчетные значения					
3	№ п.п.	y	x_1	x_2	x_1y	x_2y	x_1x_2	y^2	x_1^2	x_2^2
4	1	38,70	0,99	85,70	38,313	3316,590	84,8430	1497,69	0,9801	7344,4900
5	2	39,40	1,01	77,00	39,794	3033,800	77,7700	1552,36	1,0201	5929,0000
6	3	41,40	1,03	97,30	42,642	4028,220	100,2190	1713,96	1,0609	9467,2900
7	4	44,00	1,00	68,80	44,000	3027,200	68,8000	1936,00	1,0000	4733,4400
8	5	46,20	0,98	53,58	45,276	2475,396	52,5084	2134,44	0,9604	2870,8164
9	6	51,70	0,94	108,30	48,598	5599,110	101,8020	2672,89	0,8836	11728,8900
10	7	58,90	0,94	93,70	55,366	5518,930	88,0780	3469,21	0,8836	8779,6900
11	8	59,70	0,91	130,40	54,327	7784,880	118,6640	3564,09	0,8281	17004,1600
12	9	64,60	0,96	153,58	62,016	9921,268	147,4368	4173,16	0,9216	23586,8164
13	10	68,30	0,97	113,70	66,251	7765,710	110,2890	4664,89	0,9409	12927,6900
14	11	75,40	0,89	140,18	67,106	10569,572	124,7602	5685,16	0,7921	19650,4324
15	12	76,00	0,93	157,32	70,680	11956,320	146,3076	5776,00	0,8649	24749,5824
17	Всего	664,30	11,55	1279,56	634,369	74996,996	1221,4780	38839,85	11,1363	148772,2976
18	*n*	$\bar{y}$	$\bar{x}_1$	$\bar{x}_2$	$\overline{x_1y}$	$\overline{x_2y}$	$\overline{x_1x_2}$	σ_y	σ_{x1}	σ_{x2}
19	12	55,3583	0,9625	106,630	52,8641	6249,7497	101,7898	13,1190	0,0402	32,0583

Рис. 2. **Исходные данные и расчетные значения, необходимые для вычисления коэффициентов корреляции**

Отметим, что *точные значения* парных коэффициентов, вычисленные в Excel по исходным данным с помощью функции КОРРЕЛ(...) как в режиме полной, так и экранной точности, соответственно, равны:

r_{x_1y}=-0,7925; r_{x_2y}=0,8248; $r_{x_1x_2}$=-0,6524. (5)

Расчеты в режиме полной точности. Результаты, имитирующие калькуляторные вычисления коэффициентов корреляции по формулам (2)-(4), приведены на рис. 3. В эти формулы средние и среднеквадратические значения (операнды), вычисленные в стр. 19 на рис. 2, подставлялись с различной точностью (с различным количеством десятичных знаков после запятой), а вычисления с округленными значениями операндов выполнялись в режиме *полной точности*.

Числа в диапазоне *C*5:*E*13 представлены на рисунке с четырьмя десятичными знаками, с тем чтобы явно показать изменение значений операндов при их округлении.

В столбцах *F*:*K* приведены фактические *t* и *F*-значения, а также соответствующие им *P*-значения, используемые для определения значимости найденных значений коэффициентов корреляции при уровне значимости α=0,05. Формулы и функции Excel, используемые для их вычисления, приведены в табл. 1, где *n*=12 - объем выборки, а *k*=2 - количество факторных переменных.

	B	C	D	E	F	G	H	I	J	K
2	Режим вычислений "с полной точностью"									
3	Показатели	Точность показателей (число десятичных знаков)			Комментарий					
4										
5		4	3	2						
6	yср	55,3583	55,3580	55,3600	Среднее значение y					
7	x1ср	0,9625	0,9630	0,9600	Среднее значение x1					
8	x2ср	106,6300	106,6300	106,6300	Среднее значение x2					
9	x1yср	52,8641	52,8640	52,8600	Среднее значение x1*y					
10	x2yср	6249,7497	6249,7500	6249,7500	Среднее значение x2*y					
11	x1x2ср	101,7898	101,7900	101,7900	Среднее значение x1*x2					
12	σy	13,1190	13,1190	13,1200	Среднеквадратическое отклонение y					
13	σx1	0,0402	0,0400	0,0400	Среднеквадратическое отклонение x1					
14	σx2	32,0583	32,0580	32,0600	Среднеквадратическое отклонение x2					
15	Парные коэффициенты корреляции				t-фактическое			P-значения для α=0,05		
16	rx1y	-0,7931	-0,8494	-0,5442	-4,117	-5,091	-2,051	0,002	0,000	0,067
17	rx2y	0,8248	0,8249	0,8243	4,613	4,615	4,604	0,001	0,001	0,001
18	rx1x2	-0,6530	-0,6977	-0,4482	-2,727	-3,080	-1,586	0,021	0,012	0,144
19	Частные коэффициенты корреляции				t-фактическое			P-значения для α=0,05		
20	rx1;y(x2)	-0,5943	-0,6764	-0,3453	-2,217	-2,755	-1,104	0,054	0,022	0,298
21	rx2;y(x1)	0,6653	0,6143	0,7738	2,674	2,336	3,665	0,025	0,044	0,005
22	Множественный коэффициент корреляции				F-фактическое			P-значения для α=0,05		
23	Ry(x1,x2)	0,8906	0,9092	0,8471	17,262	21,458	11,437	0,001	0,000	0,003
24	R^2	0,7932	0,8266	0,7176						

Рис. 3. **Коэффициенты корреляции, вычисленные в режиме полной точности при различных способах округления средних и среднеквадратических значений переменных**

Таблица 1. **Формулы расчета *t*, *F* и *P*-значений**

Для парных коэффициентов корреляции
$t = r\sqrt{\frac{n-k}{1-r^2}}$, P ⇒ =СТЬЮДРАСП(ABS(r);n-k;2) (6)
Для частных коэффициентов корреляции
$t = r\sqrt{\frac{n-k-1}{1-r^2}}$, P ⇒ =СТЬЮДРАСП(ABS(r);n-k-1;2) (7)
Для коэффициента R^2
$F = \frac{(n-k-1)R^2}{k(1-R^2)}$, P ⇒ =FРАСП(F;k;n-k-1) (8)

Анализ таблицы на рис. 3 показывает, что при точности операндов в четыре десятичных знака калькуляторные вычисления дают точные значения парных коэффициентов корреляции в (5), которые за единственным исключением оказываются значимыми, поскольку все их *P*-значения меньше α. Единственным незначащим коэффициентом при этом оказывается $r_{x_1y(x_2)}$, для которого P=0,054>α. Погрешности (в процентах к точному значению) и изменение значимости менение значимости некоторых коэффициентов. Уменьшение же точности средних и среднеквадратических значений до двух знаков вообще приводит к неприемлемым результатам.

Расчеты в режиме экранной точности. Обращаем внимание: результаты, приведенные в таблице на рис. 3, получены в режиме вычислений с *полной точностью*, что даже при понижении точности средних и среднеквадратических значений не полностью имитируют калькуляторные

Таблица 2. **Зависимость погрешности и значимости коэффициентов корреляции от точности средних и среднеквадратических значений**

Коэффициенты	Погрешность, % к точному значению		Изменилась ли значимость	
	Точность показателей (число десятичных знаков)		Точность показателей (число десятичных знаков)	
	3	2	3	2
rx1y	7,11	31,38	нет	да
rx2y	0,01	0,07	нет	нет
rx1x2	6,84	31,36	нет	да
rx1;y(x2)	13,81	41,90	да	нет
rx2;y(x1)	7,66	16,31	нет	нет
Ry(x1,x2)	2,09	4,88	нет	нет

	B	C	D	E	F	G	H	I	J	K
2	Режим вычислений "с экранной точностью"									
3	Показатели	Точность показателей (число десятичных знаков)			Комментарий					
4										
5		4	3	2						
6	yср	55,3583	55,3580	55,3600	Среднее значение y					
7	x1ср	0,9625	0,9630	0,9600	Среднее значение x1					
8	x2ср	106,6300	106,6300	106,6300	Среднее значение x2					
9	x1yср	52,8641	52,8640	52,8600	Среднее значение x1*y					
10	x2yср	6249,7497	6249,7500	6249,7500	Среднее значение x2*y					
11	x1x2ср	101,7898	101,7900	101,7900	Среднее значение x1*x2					
12	σy	13,1190	13,1190	13,1200	Среднеквадратическое отклонение y					
13	σx1	0,0402	**0,0260**	**0,0800**	Среднеквадратическое отклонение x1					
14	σx2	32,0583	32,0580	32,0600	Среднеквадратическое отклонение x2					
15	Парные коэффициенты корреляции				t-фактическое			P-значения для α=0,05		
16	rx1y	-0,7931	**-1,3068**	-0,2721	-4,118	#ЧИСЛО!	-0,894	0,002	#ЧИСЛО!	0,392
17	rx2y	0,8248	0,8249	0,8243	4,613	4,615	4,604	0,001	0,001	0,001
18	rx1x2	-0,6530	**-1,0734**	-0,2241	-2,727	#ЧИСЛО!	-0,727	0,021	#ЧИСЛО!	0,484
19	Частные коэффициенты корреляции				t-фактическое			P-значения для α=0,05		
20	rx1;y(x2)	-0,5943	#ЧИСЛО!	-0,1584	-2,217	#ЧИСЛО!	-0,481	0,054	#ЧИСЛО!	0,642
21	rx2;y(x1)	0,6653	-1,7606	0,8140	2,673	#ЧИСЛО!	4,204	0,025	#ЧИСЛО!	0,002
22	Множественный коэффициент корреляции				F-фактическое			P-значения для α=0,05		
23	Ry(x1,x2)	0,8906	#ЧИСЛО!	0,8292	17,257	#ЧИСЛО!	9,903	0,001	#ЧИСЛО!	0,005
24	R^2	0,7932	#ЧИСЛО!	0,6876						

Рис. 4. **Коэффициенты корреляции, вычисленные в режиме экранной точности при различных способах округления средних и среднеквадратических значений переменных**

коэффициентов корреляции, возникающие при уменьшении точности используемых операндов, приведены в табл. 2.

Анализ таблицы показывает, что понижение точности операндов до трех знаков после запятой уже вызывает заметную погрешность и из-

вычисления. Более адекватно такие вычисления имитирует режим экранной точности. Возникающие при этом разительные погрешности показаны в таблице на рис. 4.

Бросаются в глаза существенные различия значений σ_{x1} в ячейках *C*13:*E*13, вычисленных

по формуле (1) при различной точности среднего значения $\overline{x_1}$. Такие же значения σ_{x1} были получены авторами и при их вычислении на калькуляторе посредством следующих операций:

|CM|(11,1363)|/|(12)|M+|($\overline{x_1}$)|*|=|M-|RM|$\sqrt{}$|. (9)

Для пояснения данной формулы напомним, что большинство простейших калькуляторов оперирует со следующими тремя видами чисел:

- текущее число, высвечиваемое на экране калькулятора, которое получается в результате выполнения очередной операции или вводится с клавиатуры;
- число, хранящееся в арифметическом регистре (обозначается буквой *R*), куда помещается результат большинства операций;
- число, хранящееся в специальной ячейке памяти (обозначается буквой *M*).

В формуле (9) в круглых скобках указаны *текущие значения*, вводимые с клавиатуры калькулятора, а в вертикальных черточках |...| - вычислительные операторы. В соответствии с этим фрагменты операций в (9) имеют следующий смысл:

- |CM|(11,1363)|/|(12)|M+| - очищается ячейка памяти (|CM|), выполняется деление 11,1363/12 (оба числа вводятся вручную) и результат заносится в ячейку памяти (|M+|);
- ($\overline{x_1}$)|*|=|M-| - с необходимой точностью вручную вводится значение $\overline{x_1}$, которое возводится в квадрат (|*|=|), и результат вычитается из ячейки памяти (|M-|);
- |RM|$\sqrt{}$| - число из ячейки памяти переносится в арифметический регистр (|RM|), становится текущим, и из него извлекается квадратный корень (|$\sqrt{}$|).

Таким способом по (9) последовательно и были вычислены значения σ_{x1} для $\overline{x_1}$, соответственно, равные 0,9625; 0,9630; 0,9600. При этом были получены те же значения, что и вычисленные в Excel в режиме экранной точности и показанные на рис. 4.

Значения коэффициентов парной корреляции в таблице на рис. 4, вычисленные калькуляторным способом или в режиме экранной точности Excel, не просто содержат большую погрешность, но даже *становятся по абсолютной величине больше единицы*! Связанные с такими коэффициентами подкоренные выражения в формулах (3), (4), (6)-(8) становятся отрицательными, что вызывает появление сообщения об ошибке #ЧИСЛО! в соответствующих ячейках на рис. 4. При этом, естественно, становится невозможным вычислить ни парные коэффициенты корреляции, ни коэффициент множественной корреляции, ни оценить их значимость.

Анализ источника погрешности. Провести полный аналитический анализ факторов, влияющих на возникновение погрешности, представляется достаточно сложным. При визуальном же и качественном анализе может показаться, что источником приведенных существенных погрешностей в вычислении коэффициентов корреляции служит вариация малого значения σ_{x_1}=0,0402, стоящего в знаменателе формул (2). Для проверки этого утверждения, а также анализа последствий округления различных операндов формул (2), были проведены следующие вычисления. Сначала все операнды в (2) фиксировались на уровне их точных значений, приведенных на рис. 3 в диапазоне *C*6:*C*14, после чего вносилась погрешность *только в один* какой-либо из операндов и отслеживалось изменение коэффициентов корреляции. В частности, округление только значения σ_{x_1}=0,0402 до 0,0400 (изменение на 0,0002 или на 0,5 %) привело к изменению r_{x_1y} с -0,7931 до -0,7971, т.е. тоже приблизительно на 0,5 %. Этот результат нетрудно обосновать аналитически. Действительно, если в формуле (2) округляется (варьируется) только значение σ_{x_1}, то остальные компоненты формулы можно считать константами и выражение для r_{x_1y} записать в виде

$$r_{x_1y}=\frac{a}{\sigma_{x_1}},\ \text{где}\ a=\frac{\overline{x_1\cdot y}-\overline{x_1}\cdot\overline{y}}{\sigma_y}=-0{,}032. \quad (10)$$

Поскольку производная rr_{x_1y} по σ_{x_1} равна

$$\frac{\partial r_{x_1y}}{\partial\sigma_{x_1}}=-\frac{a}{\sigma^2_{x_1}}, \quad (11)$$

то приращение Δr_{x_1y} и относительная погрешность $\delta r_{x_1y}=\Delta r_{x_1y}/r_{x_1y}$ приблизительно равны:

$$\Delta r_{x_1y}=\frac{\partial r_{x_1y}}{\partial\sigma_{x_1}}\Delta\sigma_{x_1}=-\frac{a}{\sigma^2_{x_1}}\Delta\sigma_{x_1}; \quad (12)$$

$$\delta r_{x_1y}=\frac{\Delta r_{x_1y}}{r_{x_1y}}=-\frac{a}{\sigma^2_{x_1}}\Delta\sigma_{x_1}:\frac{a}{\sigma_{x_1}}=-\frac{\Delta\sigma_{x_1}}{\sigma_{x_1}}. \quad (13)$$

При $\Delta\sigma_{x_1}$=0,0402 и $\Delta\sigma_{x_1}$=0,0002 (0,5 %) получаем, что δr_{x_1y}=0,0002/0,0402=0,5%, т.е. получена та же относительная погрешность, что и приведенная выше. Заметим, что в соответствии с (13) при большей погрешности σ_{x_1}, например $\Delta\sigma_{x_1}$=0,001 (1,99 %), в рассматриваемом примере получим δr_{x_1y}=2,49 %.

Эксперименты с варьированием (различным округлением) *только одного* какого-либо операнда в (2) показали, что существенное влияние на значение вычисляемых коэффициентов корреляции

	B	C	D	E	F	G
2	**Варьирование только X1ср**					
3	**Точные значения**	$\overline{x_1}$	0,9625	**Экспериментальные расчеты**		
4		rx1y	-0,7931	$\overline{x_1}$ округл	0,9630	0,9600
5	**Теоретические расчеты**			$\Delta\overline{x_1}$	0,0005	-0,0025
6	**c**	**d**	**c+d*Xср**	$\delta\ \overline{x_1}$ в %	0,052%	-0,260%
7	100,2384	-104,9678	-0,7931	r экспер	-0,8456	-0,5307
8	Δr теор	-0,0525	0,2624	Δr экспер	0,0525	-0,2624
9	δr теор	6,62%	-33,09%	δr экспер	-6,62%	33,09%

Рис. 5. **Теоретический и экспериментальный анализ влияния погрешности округления** $\overline{x_1}$

оказала *только* точность представления среднего значения $\overline{x_1}$. При его изменении с 0,9625 на 0,9630 и на 0,9600 (соответственно, на 0,052 % и 0,26 %) с сохранением значения σ_{x_1}=0,0402, значение r_{x_1y} изменяется с -0,7931, соответственно, на -0,8456 (на 6,62 %) и на -0,5307 (на 33,09 %), как это показано на рис. 5 в ячейках *E*7:*G*9 (расчеты выполнены с экранной точностью).

Теоретический анализ влияния точности $\overline{x_1}$ на r_{x_1y} выполняется посредством следующих выкладок, полностью аналогичных приведенным выше формулам (10)-(13). Если $\overline{x_1}$ единственный варьируемый вариант в (2), то формулу можно записать в виде

$$r_{x_1y} = c + d\cdot\overline{x_1}, \qquad (14)$$

где $c=\dfrac{\overline{x_1\cdot y}}{\sigma_{x_1}\sigma_y}=100{,}2384; d=-\dfrac{\overline{y}}{\sigma_{x_1}\sigma_y}=-104{,}9678.$ (15)

Производная r_{x_1y} по $\overline{x_1}$ и, соответственно, приблизительные абсолютное приращение Δr_{x_1y} и относительная погрешность δr_{x_1y} равны:

$$\frac{\partial r_{x_1y}}{\partial\overline{x_1}}=d;\ \Delta r_{x_1y}=d\cdot\Delta\overline{x_1};\ \delta r_{x_1y}=\frac{d\cdot\Delta\overline{x_1}}{c+d\cdot\overline{x_1}}=\frac{d\cdot\Delta\overline{x_1}}{r_{x_1}}. \qquad (16)$$

В формуле (16) знаменатель r_{x_1y} у последней дроби *всегда меньше единицы*, что усиливает влияние погрешности $\Delta\overline{x_1}$ на точность результата. Поэтому при большом абсолютном значении *d*, что в соответствии с (15) может быть обусловлено малыми значениями среднеквадратичных отклонений σ_x, σ_y, точность среднего значения будет сильно влиять на значение коэффициента корреляции, что и имеет место в данном примере. Все вычисления по формулам (15)-(16) приведены на рис. 5 в диапазоне *B*7:*D*9.

Влияние погрешности округления на значения коэффициентов уравнения регрессии. В таблице на рис. 6 приведены фрагменты выдачи инструмента Регрессия из надстройки Пакет анализа, содержащие коэффициенты уравнения регрессии

$$y_T(x_1,x_2)=a_0+a_1x_1+a_2x_2, \qquad (17)$$

а также показатели, позволяющие проверить значимость, как найденных значений коэффициентов, так и всего уравнения в целом.

Из данных таблицы на рисунке следует, что

$$a_0=170{,}99;\ a_1=-144{,}43;\ a_2=0{,}22. \qquad (18)$$

И хотя при уровне значимости α=0,05 коэффициент a_1=-144,43 незначим (для него

	B	C	D	E	F
51	**Регрессионная статистика**				
53	R-квадрат	0,7930			
58	**Дисперсионный анализ**				
60	Регрессия	**SS**	**MS**	**F**	**Значимость F**
61	Остаток	1637,728	818,86	17,24	0,000836
62	Итого	427,581	47,51		
64	Коэффициенты		Стандартная ошибка e_i	t-статистика	P-значение
65	**a0**	170,99	68,86	2,4832	0,0348
66	**a1**	-144,43	65,26	-2,2131	**0,0542**
67	**a2**	0,22	0,08	2,6773	0,0253

Рис. 6. **Фрагмент выдачи инструмента Пакет анализа/Регрессия**

P-значение=0,0542>α), однако показатель значимости *F* в ячейке *D*60, равный 0,00084, свидетельствует, что уравнение регрессии в целом значимо и при этом достаточно качественно, поскольку *R*-квадрат=0,7930.

Значения коэффициентов (18) далее будут называться *точными*. Отметим, что такие значения будут получены как в режиме полной, так и в режиме экранной точности работы Excel.

Проанализируем теперь точность, с какой могут быть получены коэффициенты уравнения регрессии при калькуляторных вычислениях. С этой целью напомним, что коэффициенты уравнения регрессии являются решением системы нормальных уравнений вида

$$Aa=b, \qquad (19)$$

где $A=\begin{pmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1x_2 \\ \sum x_2 & \sum x_1x_2 & \sum x_2^2 \end{pmatrix}$, (20)

$$b=(\Sigma y, \Sigma x_1 y, \Sigma x_2 y)^T,\ a=(a_0, a_1, a_2)^T. \qquad (21)$$

Символ *T* в последней формуле означает транспонирование. Значения всех компонент матрицы *A* и вектора *b* приведены в стр. 17 (Всего) на рис. 2 с четырьмя десятичными знаками (у чисел на рисунке с меньшим числом десятичных знаков далее следуют нули). Вектор *a* находится из выражения:

$$a=A^{-1}b, \qquad (22)$$

в котором A^{-1} - это матрица, обратная к *A*. В Excel вектор *a* находится с помощью формулы

$$a \Rightarrow =\text{МУМНОЖ(МОБР(A);b)}, \qquad (23)$$

результат вычисления по которой не зависит от режима работы Excel.

На рис. 7 формула (23) применяется к трем парам (*A*,*b*) матриц *A* - векторов *b*, компоненты которых (20) и (21) получены округлением данных стр. 17 на рис. 2 до соответствующего количества десятичных знаков, указанных в диапазоне *B*71:*B*79. Полученные для каждой такой пары (*A*, *b*) коэффициенты a_0, a_1, a_2 показаны в диапазоне *I*71:*I*79.

В таблице на рис. 8, являющейся продолжением вправо таблицы на рис. 7, приведены дан-

	B	C	D	E	F	G	H	I
70	Десятичных знаков	A			b	Определитель A	Коэффициенты уравнения регрессии	
71	4	12	11,5500	1279,5600	664,3000	1651	a0	170,89
72		11,5500	11,1363	1221,4780	634,3700		a1	-144,34
73		1279,5600	1221,4780	148772,2976	74996,9960		a2	0,22
74	3	12	11,5500	1279,5600	664,3000	1607	a0	175,08
75		11,550	11,1360	1221,4780	634,3700		a1	-148,32
76		1279,560	1221,4780	148772,2980	74996,9960		a2	0,22
77	2	12	11,5500	1279,5600	664,3000	2199	a0	133,17
78		11,5500	11,1400	1221,4800	634,3700		a1	-108,41
79		1279,5600	1221,4800	148772,3000	74997,0000		a2	0,25

Рис. 7. **Коэффициенты уравнения регрессии, рассчитанные при различной точности компонент матрицы *A* и вектора *b***

	H	I	J	K	L	M	N	O	P	Q	R
70	Коэффициенты уравнения регрессии		Номер коэффициента, i	SSR, SSE, F	Значимость F	Стандартная ошибка e_i	t-статистика	P-значение	Z		
71	a0	170,89	1	1637,44		68,856	2,48	0,035	1	0,99	85,70
72	a1	-144,34	2	427,58	0,000836	65,258	-2,21	**0,054**	1	1,01	77,00
73	a2	0,22	3	17,23		0,082	2,68	0,025	1	1,03	97,30
74	a0	175,08	1	1650,46		68,870	2,54	0,032	1	1,00	68,80
75	a1	-148,32	2	427,75	0,000814	65,271	-2,27	0,049	1	0,98	53,58
76	a2	0,22	3	17,36		0,082	2,64	0,027	1	0,94	108,30
77	a0	133,17	1	1536,29		70,012	1,90	**0,090**	1	0,94	93,70
78	a1	-108,41	2	442,05	0,001178	66,353	-1,63	**0,137**	1	0,91	130,40
79	a2	0,25	3	15,64		0,083	2,99	0,015	1	0,96	153,58

Рис. 8. **Оценка значимости уравнения регрессии и его коэффициентов**

	B	C	D	E	F	G	H
82	Количество десятичных знаков	Относительная погрешность коэффициентов			Изменилась ли значимость коэффициентов		
83		a_0	a_1	a_2	a_0	a_1	a_2
84	3	2,45%	2,76%	1,49%	Нет	Да	Нет
85	2	22,07%	24,89%	13,42%	Да	Нет	Нет

Рис. 9. **Влияние точности компонент матрицы *A* и вектора *b* на значение и значимость коэффициентов уравнения регрессии**

ные, необходимые для оценки значимости уравнения регрессии в целом и его коэффициентов.

Ниже приводятся известные математические формулы и реализующие их формулы Excel, посредством которых рассчитаны данные в таблице.

$$SSR=\Sigma(y_т-\bar{y})^2, \Rightarrow\{=КВАДРОТКЛ(\$I\$71+ +\$I\$72*x1\_+\$I\$73*x2\_)\}; \quad (24)$$

$$SSE=\Sigma(y_т-y)^2, \Rightarrow\{=СУММКВРАЗН(y;I71+ +I72*x1\_+I73*x2\_)\}; \quad (25)$$

$$F=(SSR/k)/(SSE/(n-k-1). \quad (26)$$

Значимость $F \Rightarrow$ =FРАСП(F;k;n-k-1). (27)

Напомним, что формулы Excel, заключенные в фигурные скобки {...}, должны вводиться как формулы массива.

Как известно, стандартная ошибка e_i (i=1,2,3) для i-го коэффициента уравнения регрессии вычисляется по формуле

$$e_i = \sqrt{\frac{Er_{i,i} \cdot SSE}{n-k-1}}; \; Er = (Z^T Z)^{-1}. \quad (28)$$

В формуле (28) e_{ii} - диагональные элементы матрицы *Er*, способ вычисления которой показан в (28), при этом *Z* - это расширенная факторная матрица, т.е. матрица значений факторов x_1 и x_2 (диапазон *D*4:*E*15 на рис. 2), дополненная слева вектором, состоящим из единичных значений. Фрагмент этой матрицы (отсутствуют последние три строки) приведен на рис. 8 в диапазоне *P*71:*R*79. Вычисление e_i реализовано посредством следующей формулы Excel:

{=(ИНДЕКС((МОБР(МУМНОЖ(ТРАНСП(Z);Z)));i;i)* *SSE/(n-k-1))^0,5}.

В данной формуле коэффициенты уравнения регрессии имеют номера *i*, указанные на рис. 8 в диапазоне *J*71:*J*79.

Значения в графе *t*-статистика получены делением соответствующих коэффициентов a_i в столбце *H* на их стандартную ошибку e_i. *P*-значение вычислялось по следующей формуле Excel:

=СТЬЮДРАСП(ABS(N71);n-k-1;2). (29)

Расчеты, посредством которых вычислены значения в таблицах на рис. 7, 8, достаточно хорошо имитируют расчет соответствующих значений калькуляторным способом. Полученные при этом результаты показывают, что при точности вычислений в четыре десятичных знака все значения, вычисленные на рис. 8 в стр. 71:73 калькуляторным способом, совпали с аналогичными *точными* значениями, полученными посредством Пакета анализа и приведенными на рис. 6. Совпали также оценки значимости как уравнения регрессии в целом, так и его коэффициентов. Относительная же погрешность значений этих коэффициентов, возникающая при округлении компонент *A* и *b* до трех и двух десятичных знаков, приведена на рис. 9.

Анализ данных таблицы рис. 9 показывает, что, несмотря на хорошую обусловленность всех матриц *A* (их определители существенно больше нуля), казалось бы, небольшие погрешности округления могут привести к существенной потере точности результатов или к неправильной оценке их значимости.

Выводы и рекомендации. В большинстве случаев технология калькулярных статистических вычислений предполагает первоначальное вычисление ряда значений (например, средних или вариации) с достаточно высокой точностью. Однако при использовании этих значений в последующих вычислениях они обычно округляются студентами до трех-двух десятичных знаков значений. На конкретных примерах в работе показано, что такие, казалось бы, малозначащие погрешности округления, могут привести к существенной потере точности или полному искажению результатов, а также к неправильной оценке их статистической значимости. Поэтому в общем случае можно считать необходимым, чтобы при задании исходных данных с двумя десятичными знаками все вычисления производились с точностью не меньшей, чем четыре десятичные цифры, что соответствует правилам вычислений с приближенными числами.

[1] См.: Статистика : учеб. пособие / кол. авт.; под ред. В.Н. Салина, Е.П. Шпаковской. 2-е изд., перераб. и доп. Москва, 2014; *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Исследование зависимостей : справ. изд. / под ред. С.А. Айвазяна. Москва, 1985; Теория статистики : учебник / Р.А. Шмойлова [и др.]; под ред. Р.А. Шмойловой. Москва, 2003.

Поступила в редакцию 05.02.2016 г.