Корреляционный анализ

Содержание

Парная корреляция. Основные понятия

Корреляционная связь – это согласованное изменение двух признаков, отражающее тот факт, что изменчивость одного признака находится в соответствии с изменчивостью другого.

Корреляционные связи различаются ***по форме, направлению и степени*** (силе).

По форме корреляционная связь может быть прямолинейной или криволинейной. Прямолинейной может быть, например, связь между количеством тренировок на тренажере и количеством правильно решаемых задач в контрольной сессии. Криволинейной может быть, например, связь между уровнем мотивации и эффективностью выполнения задачи. При повышении мотивации эффективность выполнения задачи сначала возрастает, затем достигается оптимальный уровень мотивации, которому соответствует максимальная эффективность выполнения задачи; дальнейшему повышению мотивации сопутствует уже снижение эффективности.

По направлению корреляционная связь может быть положительной (прямой) и отрицательной (обратной). При положительной прямолинейной корреляции более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака – низкие значения другого. При отрицательной корреляции соотношения обратные. При положительной корреляции коэффициент корреляции имеет положительный знак, например r_{xy} = 0,207, при отрицательной корреляции – отрицательный знак, например, r_{xy} = -0,207.

Степень, сила или теснота корреляционной связи определяется по величине коэффициента корреляции. Сила связи не зависит от её направленности и определяется по абсолютному значению коэффициента корреляции. Максимальное возможное абсолютное значение коэффициента корреляции r_{xy} = 1; минимальное r_{xy} = 0.

Рассмотрим общую классификация корреляционных связей:

сильная, или тесная при коэффициенте корреляции $|r_{xy}| > 0,70$;

средняя	при $0,50 < \|r_{xy}\| < 0,69$;
умеренная	при $0,30 < \|r_{xy}\| < 0,49$;
слабая	при $0,20 < \|r_{xy}\| < 0,29$;
очень слабая	при $\|r_{xy}\| < 0,19$.

Коэффициент корреляции Пирсона

Термин «корреляция» был введен в науку выдающимся английским естествоиспытателем Ф. Гальтоном в 1886 г. Однако точную формулу для подсчёта коэффициента корреляции разработал его ученик К. Пирсон.

Коэффициент характеризует наличие только линейной связи между признаками, обозначаемыми, как правило, символами *X* и *Y*. Формула расчёта коэффициента корреляции построена таким образом, что, если связь между признаками имеет линейный характер, коэффициент Пирсона точно устанавливает тесноту этой связи. Поэтому он называется также коэффициентом линейной корреляции Пирсона. Если же связь между переменными *X* и *Y* не линейна, то Пирсон предложил для оценки тесноты этой связи так называемое корреляционное отношение.

Величина коэффициента линейной корреляции Пирсона не может превышать +1 и быть меньше, чем –1. Эти два числа +1 и –1 являются границами для коэффициента корреляции. Если при расчёте получается величина, большая +1 или меньшая –1, то произошла ошибка в вычислениях.

Знак коэффициента корреляции очень важен для интерпретации полученной связи. Подчеркнём ещё раз, что если знак коэффициента линейной корреляции положительный, то связь между коррелирующими признаками такова, что большей величине одного признака (переменной) соответствует б*о*льшая величина другого признака (другой переменной). Иными словами, если один показатель (переменная) увеличивается, то соответственно увеличивается и другой показатель (переменная). Такая зависимость носит название прямой зависимости.

Если же коэффициент корреляции отрицателен, то большей величине одного признака соответствует меньшая величина другого. Иначе говоря, при наличии знака минус увеличению одной переменной (признака, значения) соответствует уменьшение другой переменной. Такая зависимость носит название обратной зависимости.

Коэффициент корреляции Пирсона вычисляется по формуле:

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sigma(x) \cdot \sigma(y)},$$

где x_i – значения, принимаемые в выборке *X*, y_i – значения, принимаемые в выборке *Y*; $\overline{x}$ – среднее значение по *X*, $\overline{y}$ - среднее значение по *Y*.

Замечание. Описание применения коэффициента корреляции Пирсона в MS Excel приведено в лабораторной работе по корреляционному анализу.

Расчёт коэффициента корреляции Пирсона предполагает, что переменные X и Y измеряются в ***шкале отношений***, распределены **нормально** и число значений переменной X равно числу значений переменной Y.

АЛГОРИТМ

1. Сформулировать для выбранного уровня значимости гипотезы.

H_0: коэффициент корреляции статистически не значим (случайно отличается от нуля).

H_1: коэффициент корреляции статистически значим (не случайно отличается от нуля).

2. Оценить достоверность коэффициента корреляции, используя эмпирическое значение коэффициента Стьюдента:

$$t_r = \frac{|r_{xy}|}{S_r} = |r_{xy}| \cdot \sqrt{\frac{n-2}{1-r_{xy}^2}} \,. \qquad (57)$$

3. Найти критическое значение критерия Стьюдента для числа степеней свободы $k = n - 2$ по одноимённой статистической таблице из Приложения: $t_{крит} = t(\alpha; k)$.

Если $t_r < t_{крит}$, то принимают нулевую гипотезу, иначе – отвергают нулевую гипотезу.

Пример. Изучалась зависимость между массой шимпанзе-матерей x_i, измеряемой в начале беременности (кг), и массой новорождённых детёнышей y_i (кг) (табл. 1).

Таблица 1

i	1	2	3	4	5	6	7	8	9	10
x_i	10	10	10,1	10,2	10,8	11	11,1	11,3	11,3	11,4
y_i	0,7	0,7	0,65	0,61	0,73	0,65	0,65	0,75	0,7	0,7

i	11	12	13	14	15	16	17	18	19	20
x_i	11,8	12	12	12,1	12,3	13	13,4	13,5	14,5	15,6
y_i	0,69	0,72	0,6	0,75	0,63	0,8	0,78	0,7	0,7	0,85

Решение. Для расчёта необходимых сумм и произведений составим вспомогательную таблицу 2.

Таблица 2

Масса матерей (x_i)	Масса детёнышей (y_i)	$x_i \cdot y_i$	x_i^2	y_i^2
10	0,7	7	100	0,49
10	0,7	7	100	0,49
10,1	0,65	6,565	102,01	0,4225
10,2	0,61	6,222	104,04	0,3721
10,8	0,73	7,884	116,64	0,5329
11	0,65	7,15	121	0,4225
11,1	0,65	7,215	123,21	0,4225
11,3	0,75	8,475	127,69	0,5625
11,3	0,7	7,91	127,69	0,49
11,4	0,7	7,98	129,96	0,49
11,8	0,69	8,142	139,24	0,4761
12	0,72	8,64	144	0,5184
12	0,6	7,2	144	0,36
12,1	0,75	9,075	146,41	0,5625
12,3	0,63	7,749	151,29	0,3969
13	0,8	10,4	169	0,64
13,4	0,78	10,452	179,56	0,6084
13,5	0,7	9,45	182,25	0,49
14,5	0,7	10,15	210,25	0,49
15,6	0,85	13,26	243,36	0,7225
Сумма 237,4	14,06	167,919	2861,6	9,9598

Сформулируем нулевую и альтернативную гипотезы. H_0: коэффициент корреляции статистически не значим (случайно отличается от нуля). H_1: коэффициент корреляции статистически значим (не случайно отличается от нуля).

Находим значение коэффициента корреляции по формуле (56):

$$r_{xy} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \left(\bar{x}\right)^2} \cdot \sqrt{\overline{y^2} - \left(\bar{y}\right)^2}} = \frac{\frac{167{,}919}{20} - \frac{237{,}4}{20} \cdot \frac{14{,}06}{20}}{\sqrt{\frac{2861{,}6}{20} - \left(\frac{237{,}4}{20}\right)^2} \cdot \sqrt{\frac{9{,}9598}{20} - \left(\frac{14{,}06}{20}\right)^2}} =$$

$$= \frac{0{,}05134}{\sqrt{2{,}1831} \cdot \sqrt{0{,}003781}} = 0{,}565 .$$

Следовательно, между массой шимпанзе матерей и массой их новорождённых детёнышей существует линейная положительная средняя связь.

Оценим достоверность коэффициента корреляции:

$$t_r = \left| r_{xy} \right| \cdot \sqrt{\frac{n-2}{1-r_{XY}^2}} = 0{,}565 \cdot \sqrt{\frac{20-2}{1-(0{,}565)^2}} = 2{,}9.$$

Для уровня значимости α=0,05 найдём критическое значение критерия Стьюдента из Приложения: $t_{крит} = t(\alpha; k) = t(0{,}05; 18) = 2{,}1$. Таким образом, $t_r \geq t_{крит}$, так как $2{,}9 > 2{,}1$, и на уровне значимости 0,05 делаем вывод о статистической значимости коэффициента корреляции.

Замечание. Если сравниваемые переменные, например X и Y, измеряются в ***дихотомической шкале*** (частный случай шкалы наименований), то для определения связи целесообразно использовать коэффициент Пирсона для дихотомических шкал.

В тех случаях, когда нет необходимости подсчитывать частоту появления различных значений переменных X и Y, удобно проводить вычисления коэффициента корреляции с помощью таблицы сопряжённости, показывающей количество совместных появлений пар значений по двум переменным (признакам). A – количество случаев, когда переменная X имеет значение, равное нулю, и, одновременно, переменная Y имеет значение, равное единице; B – количество случаев, когда переменные X и Y имеют одновременно значения, равные единице; C – количество случаев, когда переменные X и Y имеют одновременно значения, равные нулю; D – количество случаев, когда переменная X имеет значение, равное единице, а переменная Y имеет значение, равное нулю (табл. 3).

Таблица 3

Общая таблица сопряжённости

		Признак X		Всего
		0	1	
Признак Y	1	A	B	$A + B$
	0	C	D	$C + D$
Итого		$A + C$	$B + D$	N

В общем виде формула ***коэффициента Пирсона для дихотомических данных*** имеет вид:

$$\varphi = \frac{(BC - AD)}{\sqrt{(A+C)\times(B+D)\times(A+B)\times(C+D)}}.$$

Пример. Пусть две сравниваемые переменные X (семейное положение) и Y (исключение из университета) измеряются в дихотомической шкале (частный случай шкалы наименований). Выяснить, существует ли взаимосвязь между семейным положением и исключением из университета.

Шифр испытуемого	Переменная X	Переменная Y
1	0	0
2	1	1
3	0	1
4	0	0
5	1	1
6	1	0
7	0	0
8	1	1
9	0	0
10	0	1

Решение. Составим по таблице исходных данных таблицу сопряжённости:

		Признак X		Всего
		0	1	
Признак Y	1	2	3	6
	0	4	1	5
Итого		6	4	10

Подставим в формулу данные из таблицы сопряжённости, соответствующей рассматриваемому примеру:

$$\varphi = \frac{3\cdot 4 - 2\cdot 1}{\sqrt{(2+4)\cdot(3+1)\cdot(2+3)\cdot(4+1)}} = 0{,}32.$$

Таким образом, коэффициент корреляции Пирсона для выбранного примера равен 0,32, то есть зависимость между семейным положением студентов и фактами исключения из университета незначительная.

Коэффициент корреляции Спирмена

В том случае, когда сравниваемые переменные X и Y являются ранговыми, то в качестве меры связи целесообразно использовать ***коэффициент ранговой корреляции Спирмена*** ($\mathbf{r}_s$), число значений переменной X равно числу значений переменной Y (n).

АЛГОРИТМ

1. Определить, какие два признака или две иерархии признаков будут участвовать в сопоставлении как переменные X и Y.

2. Выдвигаем нулевую гипотезу H_0: коэффициент корреляции статистически не значим (случайно отличается от нуля); альтернативную гипотезу H_1: коэффициент корреляции статистически значим (не случайно отличается от нуля).

3. Проранжироватъ значения переменной X, начисляя первый ранг наименьшему значению, в соответствии с правилами ранжирования. Проранжироватъ значения переменной Y в соответствии с теми же правилами.

4. Подсчитать разности между рангами X и Y по каждой строке таблицы и обозначить их через d. Возвести каждую разность в квадрат (d^2) и подсчитать их сумму.

5. Рассчитать коэффициент ранговой корреляции r_s по формуле:

а) при отсутствии одинаковых рангов вычисляется по формуле:

$$r_s = 1 - 6 \cdot \frac{\sum_{i=1}^{n} d_i^2}{n^3 - n};$$

где $d = rankX - rankY$.

б) при наличии одинаковых рангов:

$$r_s = 1 - 6 \cdot \frac{\sum_{i=1}^{n} d_i^2 + T_a + T_b}{n^3 - n},$$

где Σd^2 – сумма квадратов разностей между рангами; n – количество испытуемых или признаков, участвовавших в ранжировании; T_a и T_b – поправки на одинаковые ранги:

$$T_a = \frac{\sum\left(a^3 - a\right)}{12}, \quad T_b = \frac{\sum\left(b^3 - b\right)}{12},$$

где a – объём каждой группы одинаковых рангов в ранговом *ряду X;* b – объём каждой группы одинаковых рангов в ранговом ряду Y.

6. Оценим достоверность коэффициента корреляции по формуле:

$$t_r = \frac{\left|r_{xy}\right|}{S_r} = \left|r_{xy}\right| \cdot \sqrt{\frac{n-2}{1-r_{xy}^2}}.$$

7. Найдем критическое значение критерия для уровня значимости α и для числа степеней свободы $k = n - 2$ с помощью статистической функции **СТЬЮДЕНТ.ОБР.2Х(*α*; *k*)**.

8. Вывод:

1) если $t_r \geq t_{крит}$, то отвергают нулевую гипотезу H_0, то есть при выбранном уровне значимости делают вывод о статистической значимости коэффициента ранговой корреляции;

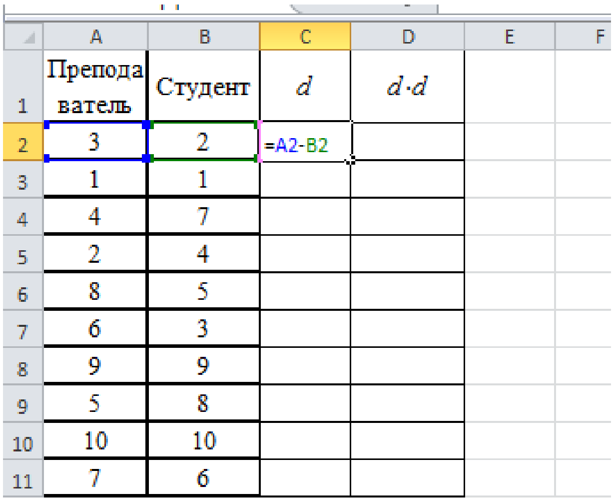
2) если $t_r < t_{крит}$, принимают нулевую гипотезу H_0, то есть при выбранном уровне значимости делают вывод о статистической значимости коэффициента ранговой корреляции.

Пример. Преподавателю и студенту было предложено расположить 10 профессий в порядке их общественной значимости. Выяснить, существует ли связь между мнениями преподавателя и студента (табл. 4).

Таблица 4

Преподаватель	Профессия	Студент
3	Профессор	2
1	Врач	1
4	Учитель школы	7
2	Директор магазина	4
8	Бухгалтер	5
6	Банкир	3
9	Водитель	9
5	Журналист	8
10	Ди-джей	10
7	Программист	6

Решение. Выполним расчеты в MS Excel.



	A	B	C	D	E	F
1	Преподаватель	Студент	d	$d \cdot d$		
2	3	2	=A2-B2			
3	1	1				
4	4	7				
5	2	4				
6	8	5				
7	6	3				
8	9	9				
9	5	8				
10	10	10				
11	7	6				

D14 f_x =1-6*D12/(D13^3-D13)

	A	B	C	D
1	Преподаватель	Студент	d	$d \cdot d$
2	3	2	1	1
3	1	1	0	0
4	4	7	-3	9
5	2	4	-2	4
6	8	5	3	9
7	6	3	3	9
8	9	9	0	0
9	5	8	-3	9
10	10	10	0	0
11	7	6	1	1
12			**сумма**	42
13			**счет**	10
14			**коэффициент**	**0,74545**

Таким образом, ранговый коэффициент корреляции равен 0,745, что говорит о тесной и прямой связи между мнениями преподавателя и студента. Найдем критическое значение критерия с помощью статистической функции **СТЬЮДЕНТ.ОБР.2Х(0,05; 8)**: $t_{крит} = 2{,}306$.

В силу того, что $t_r \geq t_{крит}$, так как $3{,}163 > 2{,}306$, при уровне значимости 0,05 отклоняется H_0 и коэффициент ранговой корреляции является статистически значимым.

Множественная корреляция

Показатель множественной корреляции характеризует тесноту связи рассматриваемого набора факторов с исследуемым признаком, или, иначе, оценивает тесноту совместного влияния факторов на результат.

Независимо от формы связи показатель множественной корреляции может быть найден как индекс множественной корреляции:

$$R_{yx_1x_2\dots x_m} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}},$$

где σ^2_y – общая дисперсия результативного признака; $\sigma^2_{\text{ост}}$ – остаточная дисперсия.

Границы изменения индекса множественной корреляции от 0 до 1. Чем ближе его значение к 1, тем теснее связь результативного признака со всем набором исследуемых факторов. Величина индекса множественной корреляции должна быть больше или равна максимальному парному индексу корреляции:

$$R_{yx_1x_2\dots x_m} \geq r_{yx_i\,(\max)} \quad \left(i = \overline{1,m}\right).$$

При правильном включении факторов в регрессионную модель величина индекса множественной корреляции будет существенно отличаться от индекса корреляции парной зависимости. Если же дополнительно включенные в уравнение множественной регрессии факторы третьестепенны, то индекс множественной корреляции может практически совпадать с индексом парной корреляции (различия в третьем-четвертом знаках). Отсюда ясно, что, сравнивая индексы множественной и парной корреляции, можно сделать вывод о целесообразности включения в уравнение регрессии того или иного фактора.

Возможно также при линейной зависимости определение совокупного коэффициента корреляции через матрицу парных коэффициентов корреляции:

$$R_{yx_1,\dots,x_m} = \sqrt{1-\frac{\Delta r}{\Delta r_{11}}},$$

где

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & \dots & r_{yx_m} \\ r_{yx_1} & 1 & \dots & r_{x_m x_1} \\ \dots & \dots & \dots & \dots \\ r_{yx_m} & r_{x_m x_1} & \dots & 1 \end{vmatrix}$$

– определитель матрицы парных коэффициентов корреляции;

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x_1 x_2} & \dots & r_{x_1 x_m} \\ r_{x_2 x_1} & 1 & \dots & r_{x_2 x_m} \\ \dots & \dots & \dots & \dots \\ r_{x_m x_1} & r_{x_m x_2} & \dots & 1 \end{vmatrix}$$

– определитель матрицы межфакторной корреляции.

Величина множественного коэффициента корреляции зависит не только от корреляции результата с каждым из факторов, но и от межфакторной корреляции. Рассмотренная формула позволяет определять совокупный коэффициент корреляции, не обращаясь при этом к уравнению множественной регрессии, а используя лишь парные коэффициенты корреляции.

При полной зависимости результативного признака от исследуемых факторов коэффициент совокупного их влияния равен единице. Из единицы вычитается доля остаточной вариации результативного признака $\left(1-r^2\right)$, обусловленная последовательно включенными в анализ факторами. В результате подкоренное выражение характеризует совокупное действие всех исследуемых факторов.

Критические значения критерия Стьюдента

k	$\alpha\,(p)$			k	$\alpha\,(p)$		
	0,1 (0,9)	0,05 (0,95)	0,01 (0,99)		0,1 (0,9)	0,05 (0,95)	0,01 (0,99)
1	6,31375	12,70615	63,65590	18	1,73406	2,10092	2,87844
2	2,91999	4,30266	9,92499	19	1,72913	2,09302	2,86094
3	2,35336	3,18245	5,84085	20	1,72472	2,08596	2,84534
4	2,13185	2,77645	4,60408	21	1,72074	2,07961	2,83137
5	2,01505	2,57058	4,03212	22	1,71714	2,07388	2,81876
6	1,94318	2,44691	3,70743	23	1,71387	2,06865	2,80734
7	1,89458	2,36462	3,49948	24	1,71088	2,0639	2,79695
8	1,85955	2,30601	3,35538	25	1,70814	2,05954	2,78744
9	1,83311	2,26216	3,24984	26	1,70562	2,05553	2,77872
10	1,81246	2,22814	3,16926	27	1,70329	2,05183	2,77068
11	1,79588	2,20099	3,10582	28	1,70113	2,04841	2,76326
12	1,78229	2,17881	3,05454	29	1,69913	2,04523	2,75639
13	1,77093	2,16037	3,01228	30	1,69726	2,04227	2,74998
14	1,76131	2,14479	2,97685	40	1,68385	2,02107	2,70446
15	1,75305	2,13145	2,94673	60	1,67065	2,00030	2,66027
16	1,74588	2,11990	2,92079	120	1,65765	1,97993	2,61742
17	1,73961	2,10982	2,89823	∞	1,64484	1,95996	2,57583

Замечания. В Microsoft Excel для расчёта критических значений критерия используется функция СТЬЮДЕНТ.ОБР.2Х(α; k).

Критические значения коэффициента корреляции Пирсона

Число степеней свободы k	Уровень значимости α		
	0,05	0,01	0,001
2	0,9500	0,9900	0,9900
3	0,8783	0,9587	0,9911
4	0,8114	0,9172	0,9741
5	0,7545	0,8745	0,9509
6	0,7067	0,8343	0,9249
7	0,6664	0,7977	0,8983
8	0,6319	0,7646	0,8721
9	0,6021	0,7348	0,8471
10	0,5760	0,7079	0,8233
11	0,5529	0,6833	0,8010
12	0,5324	0,6614	0,7800
13	0,5139	0,6411	0,7604
14	0,4973	0,6226	0,7419
15	0,4821	0,6055	0,7247
16	0,4683	0,5897	0,7084
17	0,4555	0,5751	0,6932
18	0,4438	0,5614	0,6788
19	0,4329	0,5487	0,6625
20	0,4227	0,5368	0,6524
21	0,4132	0,5256	0,6402
22	0,4044	0,5151	0,6287
23	0,3961	0,5052	0,6177
24	0,3882	0,4958	0,6073
25	0,3809	0,4869	0,5974
26	0,3739	0,4785	0,5880
27	0,3673	0,4705	0,5790
28	0,3610	0,4629	0,5703
29	0,3550	0,4556	0,5620
30	0,3494	0,4487	0,5541
31	0,3440	0,4421	0,5465
32	0,3388	0,4357	0,5392
33	0,3338	0,4297	0,5322
34	0,3291	0,4238	0,5255

Число степеней свободы k	Уровень значимости α		
	0,05	0,01	0,001
35	0,3246	0,4182	0,5189
36	0,3202	0,4128	0,5126
37	0,3160	0,4076	0,5066
38	0,3120	0,4026	0,5007
39	0,3081	0,3978	0,4951
40	0,3044	0,3932	0,4896

Замечание. $k = n - 2$, где n – количество данных в коррелируемых рядах.

Критические значения коэффициента корреляции Спирмена

n	*α*		*n*	*α*		*n*	*A*	
	0,05	0,01		0,05	0,01		0,05	0,01
5	0,94	–	17	0,48	0,62	29	0,37	0,48
6	0,85	–	18	0,47	0,60	30	0,36	0,47
7	0,78	0,94	19	0,46	0,58	31	0,36	0,46
8	0,72	0,88	20	0,45	0,57	32	0,36	0,45
9	0,68	0,83	21	0,44	0,56	33	0,34	0,45
10	0,64	0,79	22	0,43	0,54	34	0,34	0,44
11	0,61	0,76	23	0,42	0,53	35	0,33	0,43
12	0,58	0,73	24	0,41	0,52	36	0,33	0,43
13	0,56	0,70	25	0,49	0,51	37	0,33	0,43
14	0.54	0,68	26	0,39	0,50	38	0,32	0,41
15	0,52	0,66	27	0,38	0,49	39	0,32	0,41
16	0,50	0,64	28	0,38	0,48	40	0,31	0,40

Литература

1. Гласс Дж. Статистические методы в педагогике и психологии / Дж. Гласс, Дж. Стенли. – М.: Прогресс, 1976. – 496 с.

2. Гланц С. Медико-биологическая статистика / С. Гланц. - М.: Практика, 1998. - 459 с.

3. Гмурман В. Е. Теория вероятностей и математическая статистика: учебное пособие для вузов / В. Е. Гмурман. – М.: Высш. шк., 2003. – 479 с.

4. Новиков Д. А. Статистические методы в педагогических исследованиях (типовые случаи) / Д. А. Новиков. – М.: МЗ-Пресс, 2004. – 67 с.

5. Новиков Д. А. Статистические методы в медико-биологическом эксперименте (типовые случаи) / Д. А. Новиков, В. В. Новочадов. – Волгоград: Изд-во ВГМУ, 2005. – 84 с.

6. Шилова З. В. Теория вероятностей и математическая статистика: учебное пособие / З. В. Шилова, О. И. Шилов. – Киров: Изд-во ВГГУ, 2015. – 158 с.