Лабораторная работа

Использование электронных таблиц MS Excel при решении задач корреляционного анализа

Цель работы – научиться выполнять корреляционный анализ с применением MS Excel.

1. Парная корреляция

1.1. Коэффициент корреляции Пирсона

Расчёт коэффициента корреляции Пирсона предполагает, что переменные *X* и *Y* являются количественными переменными, распределены ***нормально***, число значений переменной *X* равно числу значений переменной *Y* (*n*). Найдем коэффициент корреляции Пирсона:

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

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

В MS Excel для вычисления парных коэффициентов линейной корреляции используется специальная функция **КОРРЕЛ (массив1; массив2),** где

массив1 – ссылка на диапазон ячеек первой выборки (*X*);

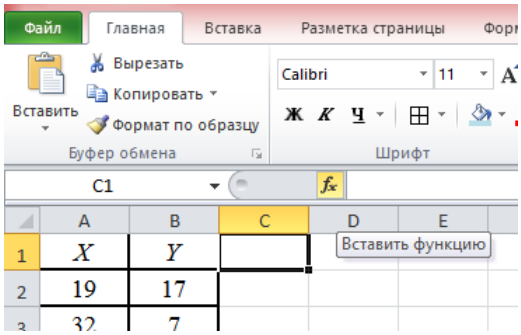
массив2 – ссылка на диапазон ячеек второй выборки (*Y*).

Пример 1. Десяти школьникам были даны тесты на наглядно-образное и вербальное мышление. Измерялось среднее время решения заданий теста в секундах. Исследователя интересует вопрос: существует ли взаимосвязь между временем решения этих задач? Переменная *X* обозначает среднее время решения наглядно-образных, а переменная *Y* – среднее время решения вербальных заданий тестов (табл. 1).

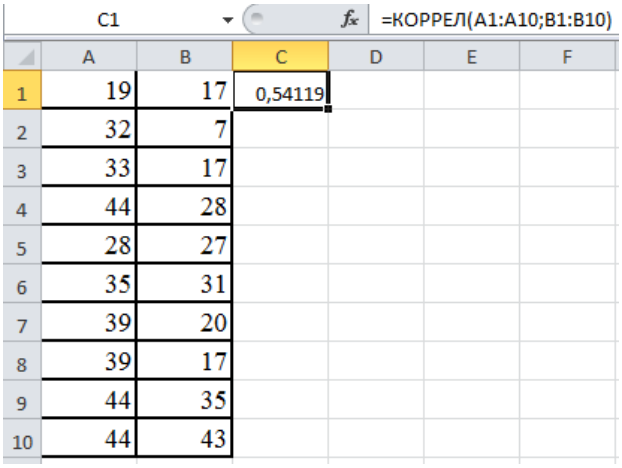
Таблица 1

№	1	2	3	4	5	6	7	8	9	10
X	19	32	33	44	28	35	39	39	44	44
Y	17	7	17	28	27	31	20	17	35	43

Решение. Введем данные в таблицу MS Excel. Затем вычислим значение коэффициента корреляции. Для этого курсор установим в ячейку С1 и активизируем кнопку f_x, находящуюся слева от строки формул.



В появившемся диалоговом окне выберем функцию **КОРРЕЛ** категории **Статистические**. Указателем мыши введем диапазон данных выборки *X* в поле *массив1* (A1:A10). В поле *массив2* введем диапазон данных выборки *Y* (B1:B10), нажмем кнопку **ОК**. В ячейке C1 появится значение коэффициента корреляции 0,54119.



	A	B	C	D	E	F
1	19	17	0,54119			
2	32	7				
3	33	17				
4	44	28				
5	28	27				
6	35	31				
7	39	20				
8	39	17				
9	44	35				
10	44	43				

Рис. 1. Результаты вычисления коэффициента корреляции

Таким образом, $r = 0{,}54$ связь между временем решения наглядно-образных и вербальных заданий теста прямая средняя.

2. Множественная корреляция

При большом числе наблюдений, когда коэффициенты корреляции необходимо последовательно вычислять для нескольких выборок, для удобства получаемые коэффициенты сводят в таблицы, называемые **корреляционными матрицами.**

Корреляционная матрица – это квадратная таблица, в которой на пересечении соответствующих строки и столбца находится коэффициент корреляции между соответствующими параметрами.

MS Excel для вычисления корреляционных матриц используется процедура **Корреляция** из пакета **Данные / анализ данных.** Процедура позволяет получить корреляционную матрицу, содержащую коэффициенты корреляции между различными параметрами.

Для реализации процедуры необходимо выполнить: **Анализ данных / корреляция.** В появившемся диалоговом окне указать **Входной интервал**, то есть ввести ссылку на ячейки, содержащие анализируемые данные. Входной интервал должен содержать не менее двух столбцов. В разделе **Группировка** переключатель установить в соответствии с введенными данными (по столбцам или по строкам). Нажать кнопку **ОК.**

В выходной диапазон будет выведена корреляционная матрица, в которой на

пересечении каждых строки и столбца находится коэффициент корреляции между соответствующими параметрами. Ячейки выходного диапазона, имеющие совпадающие координаты строк и столбцов, содержат значение 1, так как каждый столбец во входном диапазоне полностью коррелирует сам с собой.

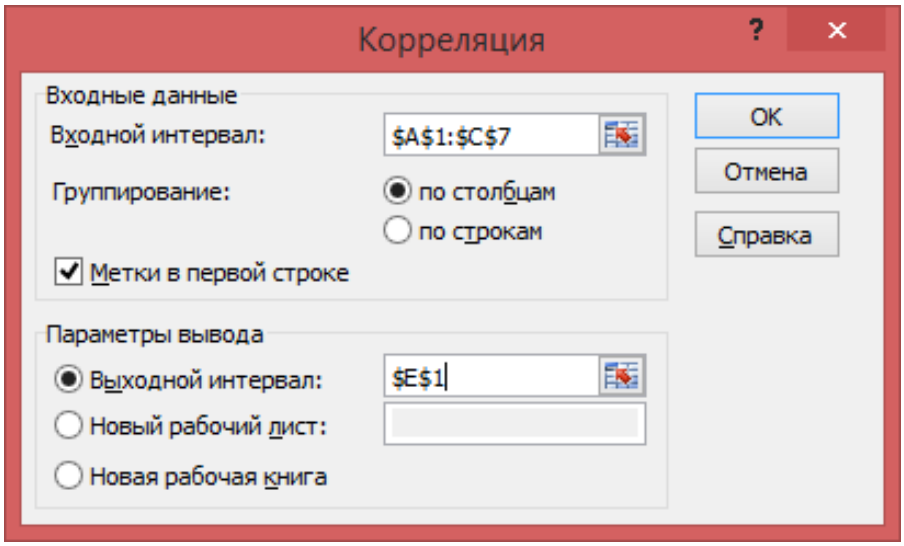
Рассматривается отдельно каждый коэффициент корреляции между соответствующими параметрами. Отметим, что в результате будет получена треугольная матрица, на самом же деле корреляционная матрица симметрична относительно главной диагонали.

Пример 2. Имеются ежемесячные данные наблюдений за состоянием погоды и посещаемостью музеев и парков (см. табл. 2). Необходимо определить, существует ли взаимосвязь между состоянием погоды и посещаемостью музеев и парков.

Таблица 2

Число ясных дней	Количество посетителей музея	Количество посетителей парка
8	495	132
14	503	348
20	380	643
25	305	865
20	348	743
15	465	541

Решение. Для выполнения корреляционного анализа введите в диапазон A1:G3 исходные данные (рис. 2). Затем в меню **Данные** выберите пункт **Анализ данных** и далее укажите строку **Корреляция**. В появившемся диалоговом окне укажите **Входной интервал** (A2:C7). Укажите, что данные рассматриваются по столбцам. Укажите выходной диапазон (E1) и нажмите кнопку **ОК**.



	A	B	C	D	E	F	G	H
1	Число ясных дней	Количество посетителей музея	Количество посетителей парка			*Число ясных дней*	*Количество посетителей музея*	*Количество посетителей парка*
2	8	495	132		Число ясных дней	1		
3	14	503	348		Количество посетителей музея	-0,92185	1	
4	20	380	643		Количество посетителей парка	0,974576	-0,91938	1
5	25	305	865					
6	20	348	743					
7	15	465	541					

Рис. 2. Результаты вычисления корреляционной матрицы

На втором рисунке видно, что корреляция между состоянием погоды и посещаемостью музея равна -0,92, между состоянием погоды и посещаемостью парка 0,97, а между посещаемостью парка и музея -0,92.

Таким образом, в результате анализа выявлены зависимости, а именно сильная степень обратной линейной взаимосвязи между посещаемостью музея и количеством солнечных дней и практически линейная (очень сильная прямая) связь между посещаемостью парка и состоянием погоды. Между посещаемостью музея и парка имеется сильная обратная взаимосвязь.

Задачи для самостоятельного решения

Задача 1. По некоторым территориям районов края известны значения средней суточного душевого дохода в у. е. (фактор X) и процент от общего дохода, расходуемого на покупку продовольственных товаров (фактор Y). Необходимо установить тесноту связи между переменными.

N	X	Y
1	68,8	45,1
2	61,2	59,0
3	59,9	57,2
4	56,7	61,8
5	55,0	58,8
6	54,3	47,2
7	49,3	55,2

Задача 2. Пусть имеются следующие данные о 10 студентах: количество решённых задач (из 10) на экзамене по математике в зимнюю сессию студентом-первокурсником y; количество решённых заданий (из 15) на вступительном экзамене по математике тем же студентом x_1; доля интерактивных занятий (в процентах) в первую сессию по математике x_2.

x_1	8	11	12	9	8	8	9	9	8	12
x_2	5	8	8	5	7	8	6	4	5	7
y	5	10	10	7	5	6	6	5	6	8

Необходимо установить тесноту связи между переменными.