Министерство образования Республики Беларусь
Учреждение образования «Витебский государственный
университет имени П.М. Машерова»
Кафедра математики

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ В MS EXCEL

Методические рекомендации

Витебск
ВГУ имени П.М. Машерова
2023

УДК 004.51(075.8)
ББК 32.972.131я73
С78

Печатается по решению научно-методического совета учреждения образования «Витебский государственный университет имени П.М. Машерова». Протокол № 7 от 26.04.2023.

Составитель: заведующий кафедрой математики ВГУ имени П.М. Машерова, кандидат физико-математических наук **Т.Б. Караулова**

Р е ц е н з е н т ы :
старший преподаватель кафедры математики
и информационных технологий УО «ВГТУ» *А.В. Коваленко*;
доцент кафедры прикладного и системного программирования
ВГУ имени П.М. Машерова,
кандидат физико-математических наук *Е.А. Витько*

С78 **Статистическая обработка данных в MS EXCEL :** методические рекомендации / сост. Т.Б. Караулова. – Витебск : ВГУ имени П.М. Машерова, 2023. – 21 с.

Данное издание подготовлено в соответствии с учебными программами дисциплин «Статистические методы обработки данных», «Статистические методы анализа данных» и «Методы статистического анализа данных» I ступени высшего образования. Кратко излагается теоретический материал, разобраны примеры решения ключевых задач в пакете MS EXCEL, предложены задания для самостоятельного решения.

УДК 004.51(075.8)
ББК 32.972.131я73

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Современная производственная и научная деятельность людей связана с необходимостью анализа непрерывно нарастающего потока данных. Это требуется для создания моделей изучаемых явлений и правильного принятия решения. Появление электронных таблиц (табличных процессоров) привело к тому, что статистические методы, ранее доступные лишь узкому кругу математиков, стали использоваться широким кругом специалистов разных областей. Дальнейшее развитие программного обеспечения привело к созданию большого количества прикладных пакетов по статистике. Удобной универсальной вычислительной средой для решения задач является табличный процессор *MS Excel*.

Данные методические рекомендации содержат основные теоретические сведения о ключевых понятиях статистического анализа, таких как дисперсионный, корреляционный и регрессионный. Изложение материала осуществляется следующим образом. Вначале приводятся основные определения и формулы, затем рассматриваются решения типовых примеров с помощью пакета Microsoft Excel.

Адресовано студентам Витебского государственного университета имени П.М. Машерова, обучающимся по специальностям:

1-31 03 07-01 Прикладная информатика;

1-40 05 01-07 Информационные системы и технологии (в здравоохранении);

1-40 01 01 Программное обеспечение информационных технологий. Базы данных и программное обеспечение информационных систем.

Данное учебное издание может использоваться для подготовки к занятиям по дисциплинам: «Статистические методы обработки данных», «Статистические методы анализа данных» и «Методы статистического анализа данных».

1. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

Математическая статистика – раздел математики, посвященный методам сбора, анализа и обработки статистических данных для научных и практических целей. Выделяют две основные области: описательную и аналитическую статистику.

Описательная статистика охватывает методы описания статистических данных, представления их в форме таблиц, распределений и т. п.

Аналитическая статистика, или теория статистических выводов, ориентирована на обработку данных, полученных в ходе эксперимента, с целью формулировки выводов, имеющих прикладное значение для самых различных областей человеческой деятельности.

Пакет *MS Excel* оснащен средствами статистической обработки данных. И хотя *Excel* существенно уступает специализированным статистическим пакетам обработки данных, тем не менее этот раздел математики представлен в *Excel* наиболее полно. В него включены основные, часто используемые статистические процедуры: средства описательной статистики, критерии различия, корреляционные и другие методы, позволяющие проводить необходимый статистический анализ экономических, медико-биологических и иных типов данных.

Статистическая совокупность – множество единиц, обладающих массовостью, типичностью, качественной однородностью и наличием вариации.

Изучение статистических свойств совокупности можно организовать двумя способами: с помощью сплошного и несплошного наблюдения. Сплошное наблюдение предусматривает обследование всех единиц изучаемой совокупности (*генеральной совокупности*), а несплошное (выборочное) наблюдение – только его части.

Выборочный метод – статистический метод исследования общих свойств совокупности каких-либо объектов на основе изучения свойств лишь части этих объектов, взятых на выборку.

Выборка – это группа элементов, выбранная для исследования из всей совокупности элементов. Задача выборочного метода состоит в том, чтобы сделать правильные выводы относительно всей совокупности объектов.

Основная цель изучения выборочной совокупности – это получение информации о генеральной совокупности. Поэтому необходимо стремиться делать выборку такой, чтобы она наилучшим образом отражала всю генеральную совокупность. В случае, если генеральная совокупность недостаточно известна, обычно в качестве представительной выборки, предлагается случайный выбор, то есть из генеральной совокупности случайным образом извлекается по одному объекту.

Из полученной выборки можно построить приблизительные значения для функции распределения и других характеристик случайной величины.

Выборочной или *эмпирической* функцией распределения случайной величины ξ, построенной по выборке x_1, x_2, ..., x_n, называется функция $F_n(x)$, равная доле таких значений x_i, что $x_i < x$, $i = 1, \ldots, n$. Таким образом, $F_n(x)$ есть частота события $x_i < x$ в ряду x_1, x_2, ..., x_n.

Для построения *выборочной функции* распределения весь диапазон изменения случайной величины X разбивают на ряд интервалов одинаковой ширины. Число интервалов обычно выбирают не менее 5 и не более 15. Затем определяют число значений случайной величины X, попавших в каждый интервал. Поделив эти числа на общее количество наблюдений n, находят относительную частоту попадания случайной величины X в заданные интервалы. По найденным относительным частотам строят гистограммы выборочных функций распределения. Если соответствующие точки относительных частот соединить ломаной линией, то полученная диаграмма будет называться *полигоном* частот.

В *Excel* для построения выборочных функций распределения используются специальная функция **ЧАСТОТА()** и процедура Пакета анализа **Гистограмма**. Функция **ЧАСТОТА** вычисляет частоты появления случайной величины в интервалах значений и выводит их как массив чисел.

Выборочные характеристики. Замена теоретической функции распределения $F(x)$ на ее выборочный аналог $F_n(x)$ в определении математического ожидания, дисперсии, стандартного отклонения и т. п. приводят к выборочному среднему, выборочной дисперсии, выборочному стандартному отклонению и т. д. Выборочные характеристики являются оценками соответствующих характеристик генеральной совокупности и должны удовлетворять следующим требованиям:

- *Состоятельность.* При неограниченном увеличении объема выборки оценка стремится к истинному значению характеристики генеральной совокупности.
- *Несмещенность.* Оценка не содержит систематической ошибки, то есть среднее значение оценки, определенное по многократно повторенной выборке объема n из одной и той же генеральной совокупности, стремится к истинному значению соответствующего генерального параметра.
- *Эффективность.* Несмещенная оценка является эффективной, если она имеет наименьшую дисперсию по сравнению с другими несмещенными оценками того же параметра генеральной совокупности.

Мода – это элемент выборки с наиболее часто встречающимся значением.

Средним значением выборки, или выборочным аналогом математического ожидания, называется величина

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

При увеличении числа наблюдений среднее приближается к математическому ожиданию.

Выборочная медиана – это число, которое является серединой выборки, то есть половина чисел имеет значения большие, чем медиана, а половина чисел имеет значения меньшие, чем медиана. Для нахождения медианы обычно выборку ранжируют – располагают элементы в порядке возрастания. Если количество членов ранжированного ряда нечетное, медианой является значение ряда, которое расположено посередине, то есть элемент с номером $(n + 1)/2$. Если число членов ряда четное, то медиана равна среднему значению членов ряда с номерами $n/2$ и $n/2 + 1$.

Основными показателями рассеяния вариант являются интервал, дисперсия выборки, стандартное отклонение и стандартная ошибка.

Интервал – это разница между максимальным и минимальным значениями элементов выборки. Интервал является простейшей и наименее надежной мерой вариации или рассеяния элементов в выборке. Более точно отражают рассеяние показатели, учитывающие не только крайние, но и все значения элементов выборки.

Дисперсией выборки, или выборочным аналогом дисперсии, называется величина

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$$

Данный параметр характеризует степень разброса элементов выборки относительно среднего значения.

Выборочным стандартным отклонением называется величина

$$\sigma = \sqrt{s^2}$$

Параметр аналогичен дисперсии и используется в тех случаях, когда необходимо, чтобы показатель разброса случайной величины выражался в тех же единицах, что и среднее значение этой случайной величины.

Стандартная ошибка или *ошибка среднего* вычисляется по формуле

$$m = \frac{s}{\sqrt{n}}$$

Данный параметр характеризует степень возможного отклонения среднего значения, полученного на исследуемой ограниченной выборке, от истинного среднего значения, полученного на всей совокупности элементов.

Выборочной квантилью называется решение уравнения

$$F_n(x)=p.$$

Далее рассмотрим показатели, которые характеризуют форму распределения.

Эксцесс – это степень частоты появления удаленных от среднего значений.

Асимметрия – величина, характеризующая несимметричность распределения элементов выборки относительно среднего значения. Принимает

значения от -1 до 1. В случае симметричного распределения асимметрия равна 0.

В электронной таблице *Excel* имеется ряд специальных функций, предназначенных для вычисления выборочных характеристик.

Показатель	Функция Excel
Средним значением выборки	**СРЗНАЧ**(число1; число2; ...) вычисляет среднее арифметическое из нескольких чисел.
Мода	**МОДА**(число1;число2; ...) вычисляет наиболее часто встречающееся значение.
Медиана	**МЕДНА**(число1;число2;...) позволяет получить медиану заданной выборки.
Дисперсия выборки	**ДИСП**(число1;число2; ...) позволяет оценить дисперсию по выборочным данным.
Среднее квадратичное отклонение (выборочное стандартное отклонение)	**СТАНДОТКЛОН**(число1; число2; ...) вычисляет стандартное отклонение.
Эксцесс	**ЭКСЦЕСС**(число1;число2; ...) вычисляет оценку эксцесса по выборочным данным.
Асимметрия	**СКОС**(число1;число2; ...) позволяет оценить асимметрию выборочного распределения.

Для проведения статистических расчетов в электронной таблице *Microsoft Excel* имеется пакет анализа. *Анализ данных* является надстройкой, которую перед первым использованием необходимо установить. Выполним для этого последовательность действий: *Файл → Параметры → Надстройки → Имя → Пакет анализа → Перейти → откроется меню Надстройки → Доступные настройки → выберите Пакет анализа → ОК.*

В результате на вкладке *Данные* в группе *Анализ* появится пиктограмма, которая обеспечит доступ к *Инструментам анализа*. Данная надстройка предназначена для выполнения базовых операций статистического анализа данных. Используется она и при проведении инженерных расчетов. При запуске надстройки открывается диалоговое окно, в котором можно выбрать необходимый инструмент анализа.

Пример 1.1. В нашем распоряжении имеются выборочные данные о температуре воздуха, собранные на некоторой опытной станции за три года на небольшом острове. Нас интересует климат этого острова и насколько он комфортен для проживания. Исходные данные приведены в таблице 1.1.

Таблица 1.1 – Выборка.

10,1	-4,3	-2,2	10,6	8,9	9,4	16,6	7,3	4,7	-2,7
15,2	17,1	-4,5	8,2	15,2	6,2	19,7	11,5	1,3	15,5
4,5	8,8	9,8	13,8	4	4,5	4,5	8,1	-3,1	14,8
13,5	20	12	7,4	8,2	6,7	-0,6	18,5	7	16,5
6,5	9,4	2,6	3,2	6,8	4,7	7,9	-1	6,6	17,2
-1,1	16,5	-9	2,7	10,6	9,5	11,5	-2	2,3	7,8
-3	12,3	5,7	0,5	3,8	5	2,3	7,9	20	-7
7,1	4,9	9	8	7,6	18,1	1,7	20	11,3	15
8	14,3	15,4	3,2	16,5	2,6	13,4	9,6	15,5	6
-2	-1,5	9,5	19,2	16,2	3,3	8,2	9,4	1	13

Решение.

Для проведения расчетов данные необходимо поместить в один столбец. Объем выборки равен $n = 100$. Чтобы получить некоторые предварительные данные об изучаемой величине, воспользуемся описательной статистикой Пакета анализа (рисунок 1.1)

Столбец1	
Среднее	7,864
Стандартная ошибка	0,667866375
Медиана	7,95
Мода	4,5
Стандартное отклонение	6,678663751
Дисперсия выборки	44,60454949
Эксцесс	-0,500580597
Асимметричность	-0,150349262
Интервал	29
Минимум	-9
Максимум	20
Сумма	786,4
Счет	100
	0

Рисунок 1.1 – Описательная статистика

2. ДИСПЕРСИОННЫЙ АНАЛИЗ

Задачей дисперсионного анализа является изучение влияния одного или нескольких качественных факторов на рассматриваемый признак (наблюдаемую случайную величину). Рассмотрим только однофакторный дисперсионный анализ.

При этом суть заключается в том, чтобы сравнить дисперсию, обусловленную случайными причинами, с дисперсией, вызываемой наличием исследуемого фактора. Если они существенно отличаются друг от друга, то считают, что фактор оказывает статистически значимое влияние на исследуемую переменную. Значимость различий проверяется по критерию Фишера.

Влияние случайной составляющей характеризует *внутригрупповая дисперсия*, а влияние изучаемого фактора – *межгрупповая*. Внутригрупповая дисперсия рассчитывается по формуле:

$$s_2^2 = \frac{1}{m(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - M_i)^2 ,$$

межгрупповая:

$$s_1^2 = \frac{1}{m-1} \sum_{i=1}^{m} (M_i - M)^2 ,$$

$$M_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij} ,$$

где M – общее среднее, m – количество групп, n – количество элементов в группе.

В *MS Excel* для проведения однофакторного дисперсионного анализа используется процедура *Однофакторный дисперсионный анализ.* Рассмотрим работу данной процедуры на примере.

Пример 2.1. Результаты наблюдений за расходом сырья при производстве одинаковой продукции по одной и той же технологии на пяти различных заводах равных мощностей представлены в таблице 2.1. Известно, что расход сырья является нормально распределенной случайной величиной и дисперсии наблюдений по каждому заводу равны.

При уровне значимости 0,05 требуется выяснить, зависит ли расход сырья от того, на каком заводе произведена продукция.

Таблица 2.1 – Данные наблюдений за расходом сырья на пяти заводах, усл. ед.

Месяцы	Расход сырья				
	1 завод	2 завод	3 завод	4 завод	5 завод
1	114	112	132	124	124
2	124	119	124	114	116
3	110	124	129	119	119
4	116	116	129	124	119
5	119	116	129	116	132
6	119	124	124	116	129
7	129	112	114	129	116
8	124	119	119	124	119
9	110	119	124	114	-
10	124	112	-	116	-
11	119	-	-	129	-
12	124	-	-	-	-

Решение.

Введем исходные данные на лист *MS Excel*, как показано в таблице 2.1. Во вкладке *Данные* выберем *Анализ данных* → *Однофакторный анализ.* Заполним диалоговое окно, как показано ниже (рисунок 2.1).

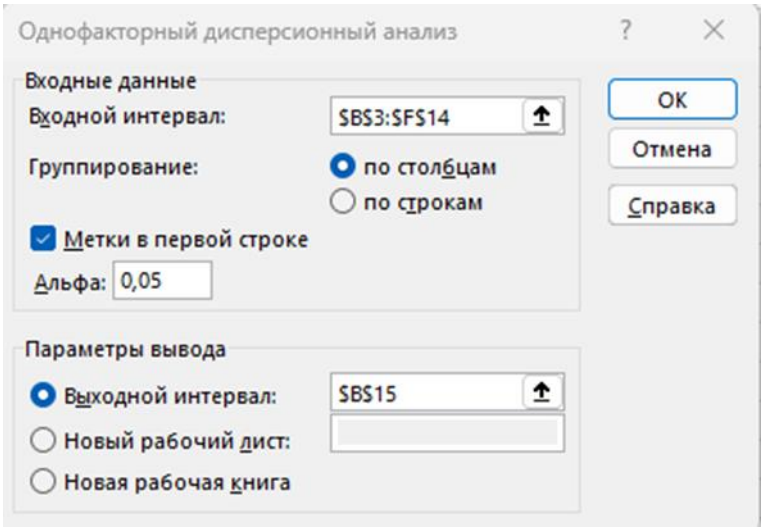

Рисунок 2.1 – Диалоговое окно «Однофакторный дисперсионный анализ»

Флажок *Метки в первой строке* поставлен потому, что входной интервал включает заголовки столбцов, и они будут использованы для формирования результата. Уровень значимости задан равным 0,05 (поле *Альфа*). Результат работы этого инструмента анализа представлен на рисунке 2.2.

Однофакторный дисперсионный анализ						
ИТОГИ						
Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
1 завод	12	1432	119,3333333	35,87878788		
2 завод	10	1173	117,3	20,67777778		
3 завод	9	1124	124,8888889	32,11111111		
4 завод	11	1325	120,4545455	32,87272727		
5 завод	8	974	121,75	35,92857143		
Дисперсионный анализ						
Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	304,437172	4	76,10929293	2,415515664	0,062563703	2,578739184
Внутри групп	1417,88283	45	31,5085073			
Итого	1722,32	49				

Рисунок 2.2 – Результат работы инструмента *Однофакторный дисперсионный анализ*

Интерпретация результатов. В таблице *Дисперсионный анализ* на пересечении строки *Между группами* и столбца *P-Значение* находится величина 0,062563703. Величина *P-Значение* > 0,05, следовательно, расход сырья существенно не зависит от того, на каком заводе произведена продукция.

3. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Важным разделом статистического анализа является корреляционный анализ, который служит для выявления взаимосвязей между выборками.

Корреляционный анализ – определение степени и направления связи между двумя явлениями. Корреляция рассматривается как признак, указывающий на взаимосвязь ряда числовых последовательностей. Иначе говоря, корреляция характеризует силу взаимосвязи в данных. Если это касается взаимосвязи двух числовых массивов x_i и y_i, то такую корреляцию называют *парной*.

При поиске корреляционной зависимости обычно выявляется вероятная связь одной измеренной величины x (для какого-то ограниченного диапазона ее изменения, например от x_1 до x_n) с другой измеренной величиной y (также изменяющейся в каком-то интервале от y_1 до y_n). В таком случае мы будем иметь дело с двумя числовыми последовательностями, между которыми и надлежит установить наличие статистической (корреляционной) связи. На этом этапе пока не ставится задача определить, является ли одна из этих случайных величин функцией, а другая – аргументом. Отыскание количественной зависимости между ними в форме конкретного аналитического выражения $y=f(x)$ – это задача уже другого анализа, *регрессионного*.

Для количественной оценки существования связи между изучаемыми совокупностями случайных величин используется специальный статистический показатель – *коэффициент корреляции r* (параметр, характеризующий степень линейной взаимосвязи между двумя выборками). Коэффициент *r* – это безразмерная величина, она может меняться от 0 до ±1. Чем ближе значение коэффициента к единице (неважно, с каким знаком), тем с большей уверенностью можно утверждать, что значение какой-то одной из этих случайных величин (*y*) существенным образом зависит от того, какое значение принимает другая (*x*). Если окажется, что *r* = 1(или –1), то имеет место классический случай чисто функциональной зависимости.

Существуют различные аналитические приемы определения коэффициента *r*. Наиболее часто рекомендуется использовать выражение:

$$r = \frac{\sum (x - M_x)(y - M_y)}{\sqrt{\sum (x - M_x)^2 (y - M_y)^2}}.$$

Зная коэффициент корреляции, можно дать качественно-количественную оценку тесноты связи. Используются, например, специальные табличные соотношения (так называемая шкала Чеддока). Ее представление может иметь следующий вид (таблица 3.1):

Таблица 3.1– Качественная оценка тесноты связи.

Величина коэффициента парной корреляции	**Характеристика силы связи**
До 0,3	Практически отсутствует
0,3-0,5	Слабая
0,5-0,7	Заметная
0,7-0,9	Сильная
0,9-0,99	Очень сильная

Такие оценки носят общий характер и не претендуют на статистическую строгость, поскольку не дают гарантий на вероятностную достоверность. Поэтому в статистике принято использовать более надежные критерии для оценки степени тесноты связи, основываясь на рассчитанных значениях коэффициента парной корреляции.

Процедуру установления корреляционной зависимости принято называть проверкой гипотезы. Ее принято проводить в следующей последовательности:

- вычисление линейного коэффициента парной корреляции между совокупностями случайных величин x_i и y_i;
- его статистическая оценка (проверка значимости).

Статистическую оценку коэффициента парной корреляции проводят путем сравнения его абсолютной величины с табличным (или критическим) показателем $r_{крит}$, значения которого отыскиваются из специальной таблицы. Если окажется, что $|r_{расч} \geq r_{крит}|$, то с заданной степенью вероятности (обычно 95%) можно утверждать, что между рассматриваемыми числовыми совокупностями существует значимая линейная связь. В случае же обратного соотношения, т.е. при $|r_{расч} < r_{крит}|$, делается заключение об отсутствии значимой связи.

В *MS Excel* для вычисления парных коэффициентов линейной корреляции используется функция **КОРРЕЛ(*массив1*; *массив2*)**, где *массив1* и *массив2* – это диапазоны ячеек, содержащих выборочные значения первой и второй случайной величины. При исследовании связи между несколькими случайными величинами находят выборочные коэффициенты корреляции между парами всех исследуемых величин и строят *корреляционную матрицу*.

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

В *MS Excel* для расчета корреляционной матрицы используется инструмент *Корреляция* из *Пакета анализа*.

Пример 3.1. Имеются ежемесячные данные наблюдений за состоянием погоды и посещаемостью парков (таблица 3.2). Необходимо определить, существует ли линейная взаимосвязь между этими случайными величинами и оценить степень этой взаимосвязи.

Таблица 3.2 – Данные наблюдений за состоянием погоды и посещаемостью парков.

Число ясных дней	8	4	20	25	20	15
Количество посетителей парка	35	48	645	860	42	541

Решение.

Для выполнения корреляционного анализа введем данные наблюдений (рисунок 3.1).

	A	B	C	D	E	F	G	H
1	**Число ясных дней**	8	14	20	25	20	15	
2	**Количество посетителей парка**	135	348	645	860	742	541	
3								

Рисунок 3.1 – Исходные данные

Выполним команду *Данные* → *Анализ данных* и выберем инструмент *Корреляция*. Заполним диалоговое окно *«Корреляция»* как показано на рисунке 3.2.

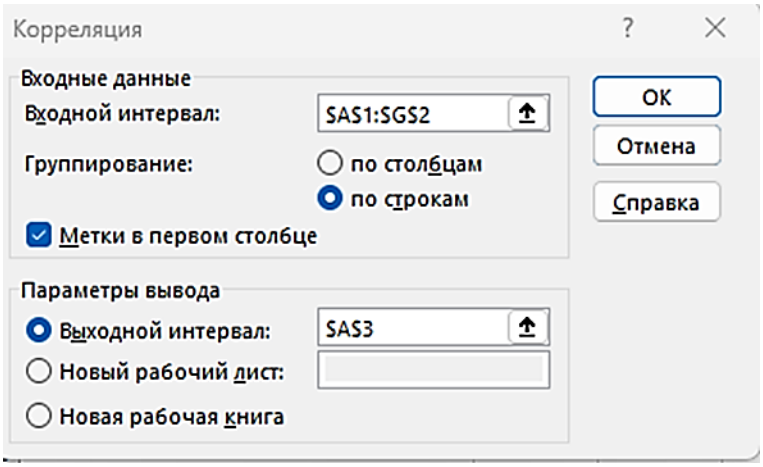

Рисунок 3.2 – Пример заполнения диалогового окна «Корреляция»

В выходном диапазоне получаем корреляционную матрицу (рисунок 3.3).

	A	B	C	D	E	F	G	H
1	**Число ясных дней**	8,00	14,00	20,00	25,00	20,00	15,00	
2	**Количество посетителей парка**	135,00	348,00	645,00	860,00	742,00	541,00	
3		*Число ясных дней*	*Количество посетителей парка*					
4	Число ясных дней	1						
5	Количество посетителей парка	0,97419388	1					
6								

Рисунок 3.3 – Результаты вычисления корреляционной матрицы

Таким образом, в результате анализа выявлена сильная связь между посещаемостью парка и состоянием погоды. Подразумевается, что в пустой клетке в правой верхней половине таблицы находится тот же коэффициент корреляции, что и в нижней левой.

4. РЕГРЕССИОННЫЙ АНАЛИЗ

Корреляцию и регрессию принято рассматривать как совокупный процесс статистического исследования и поэтому их использование в статистике часто именуют корреляционно-регрессионным анализом.

Регрессия используется для анализа воздействия на отдельную зависимую переменную значений одной или более независимых переменных. Регрессионный анализ устанавливает формы зависимости между случайной величиной *Y* (зависимой) и точно заданными значениями одной или нескольких переменных величин (независимых). Данная зависимость

определяется математической моделью (уравнением регрессии), содержащей несколько неизвестных параметров.

Рассмотрим простейший случай линейной однофакторной регрессии. Уравнение линейной однофакторной регрессии имеет вид: $Y = aX+b$. С помощью регрессионного уравнения можно предсказать ожидаемое значение зависимой величины Y_0, которое соответствует заданному значению независимой переменной X_0.

Если рассматривается зависимость между одной зависимой переменной Y и несколькими независимыми переменными X_1, X_2, ..., X_n, говорят о *множественной линейной регрессии*.

Уравнение множественной линейной регрессии имеет вид

$$Y = a_0 + a_1X_1 + a_2X_2 + ... + a_nX_n,$$

где a_0, a_1, …, a_k – параметры регрессии, которые необходимо определить по выборочным данным.

Эффективность регрессионной модели определяется коэффициентом детерминации R^2 (R-квадрат). Коэффициент детерминации определяет, с какой степенью точности регрессионное уравнение описывает исходные данные. Значимость регрессионной модели исследуется также с помощью F-критерия (*критерия Фишера*). Если величина F-критерия значима ($p < 0,05$), то регрессионная модель является значимой. Достоверность отличия коэффициентов a_0, a_1, a_2, ..., a_n от нуля проверяется с помощью *критерия Стьюдента*. В случаях, когда $p > 0,05$, коэффициент может считаться нулевым, а это означает, что влияние соответствующей независимой переменной на зависимую переменную недостоверно, и эта независимая переменная может быть исключена из уравнения.

В пакете *MS Excel* для определения параметров линейной регрессии можно использовать функцию **ЛИНЕЙН()**, а также процедуру *Регрессия* из *Пакета анализа*.

Пример 4.1. Построить регрессионную модель для предсказания изменений уровня заболеваемости органов дыхания (Y) в зависимости от содержания в воздухе двуокиси углерода (X_1) и степени запыленности (X_2). В таблице 4.1 приведены данные наблюдений в течение 28 месяцев. Предсказать уровень заболеваемости при содержании двуокиси углерода, равной 0,7, и запыленности – 1,5.

Таблица 4.1 – Данные об уровне заболеваемости.

Месяц	1	2	3	4	5	6	7	8	9	10	11	12	13	14
X_1	1	1	1,1	1,1	1,1	1,1	1	1	1,2	1,2	0,6	0,6	0,7	0,7
X_2	1,3	1,3	1,4	1,4	1,5	1,5	1,4	1,5	1,6	1,7	1	1	1,1	1,15
Y	1160	1155	1158	1157	1160	1161	1157	1159	1256	1260	1040	1039	1039	1040

Месяц	15	16	17	18	19	20	21	22	23	24	25	26	27	28
X_1	0,75	0,7	0,7	0,7	0,8	0,8	0,78	0,8	0,78	0,78	0,8	0,8	0,75	0,78
X_2	1,2	1,2	1,3	1,3	1,4	1,4	1,5	1,5	1,5	1,6	1.7	1.8	1,8	1,9
Y	1040	1039	1040	1039	1140	1138	1240	1239	1241	1240	1239	1239	1240	1238

Решение.

Введем исходные данные. Выполним команду *Данные → Анализ данных* и выберем инструмент *Регрессия*. Заполним диалоговое окно *«Регрессия»* как показано на рисунке 4.1.

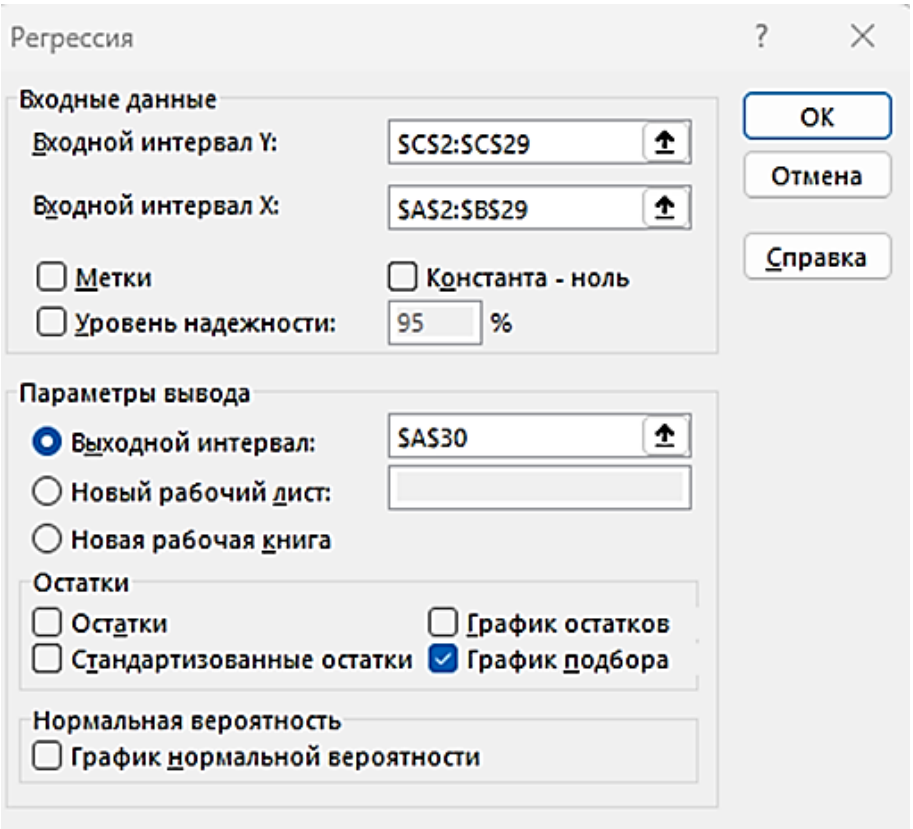

Рисунок 4.1 – Пример заполнения диалогового окна «Регрессия»

Результаты работы инструмента *Регрессия* из *Пакета анализа* показаны на рисунке 4.2.

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,869142							
R-квадрат	0,755407							
Нормированный R-квадр	0,73584							
Стандартная ошибка	44,0066							
Наблюдения	28							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	2	149524,7	74762,36	38,60534	2,2675E-08			
Остаток	25	48414,52	1936,581					
Итого	27	197939,3						
	Коэффициен	*артная о*	*татисти*	*-Значени*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	663,0847	57,35608	11,56084	1,59E-11	544,9576053	781,2117168	544,9576053	781,2117168
Переменная X 1	86,85819	50,82169	1,709077	0,099824	-17,81104317	191,5274246	-17,81104317	191,5274246
Переменная X 2	288,6511	39,22015	7,359764	1,04E-07	207,8756474	369,4264767	207,8756474	369,4264767

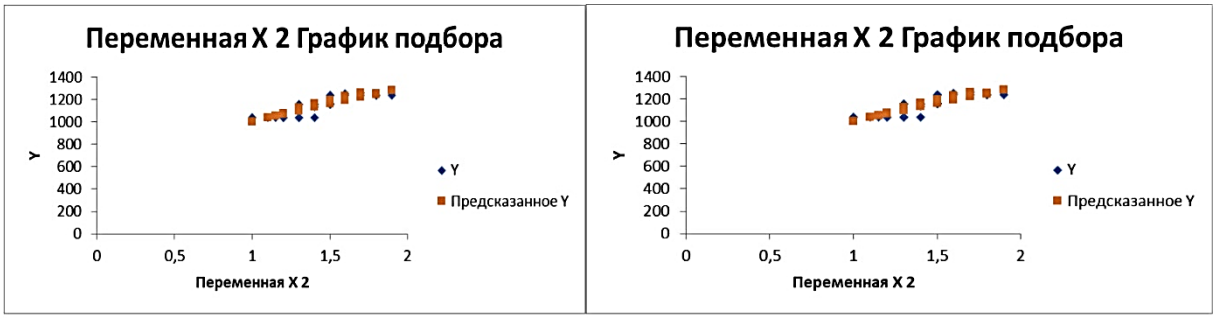

Рисунок 4.2 – Результаты регрессионного анализа

В таблице *Дисперсионный анализ* оценивается достоверность полученной модели по уровню значимости критерия Фишера (строка Регрессия, столбец Значимость F, в примере – 2,3E-08 (1,4*10-8), то есть p < 0,05 и модель значима) и степень описания моделью процесса – *R*-квадрат (вторая строка сверху в таблице *Регрессионная статистика*, в примере *R*-квадрат = 0,755). Можно говорить о довольно высокой точности аппроксимации. Далее необходимо определить значения коэффициентов модели. Они определяются из таблицы в столбце *Коэффициенты* – в строке *Y-пересечение* приводится свободный член a_0 = 663; в строках соответствующих переменных приводятся значения коэффициентов при этих переменных a_1 = 87 и a_2 = 288. В столбце *p-значение* приводится достоверность отличия соответствующих коэффициентов от нуля. Все коэффициенты значимы, то есть p < 0,05, и коэффициенты могут считаться не равными нулю. Поэтому выражение для определения уровня заболеваемости органов дыхания в зависимости от содержания углекислого газа и пыли в воздухе будет иметь вид:

$$Y = 663 + 87X_1 + 288*X_2.$$

5. ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

1. Найти выборочные среднее, медиану, моду, дисперсию и стандартное отклонение для следующей выборки: 25, 33, 28, 27, 36, 34, 31, 33, 31, 28.

2. Найти наиболее популярный туристический маршрут из четырех реализуемых фирмой (моду), если за неделю последовательно были реализованы следующие маршруты (приводятся номера маршрутов): 1, 3, 2, 2, 1, 1, 1, 4, 3, 4, 1, 3, 4, 4, 1, 4, 3, 3, 1, 2, 3, 1, 1, 1, 3.

3. В рабочей зоне производились замеры концентрации вредного вещества. Получен ряд значений (в мг/м3): 11, 15, 14, 13, 9, 21, 15, 13, 17, 14, 15, 17, 22, 16. Необходимо определить основные выборочные характеристики.

4. Построить линейную модель для двух наблюдаемых величин (например, объем реализованных подержанных автомобилей фирмой за указанное число недель).

Неделя	1	2	3	4	5	6	7	8
Количество	9	15	24	29	38	52	53	58

5. Определите, влияет ли фактор образования на уровень заработной платы работников гостиницы. Используйте уровень значимости 10%. Определите, можно ли достоверно считать, что фактор образования имеет влияние при уровне значимости 5%.

Образование	Заработная плата сотрудников, тыс. р.					
	1 месяц	2 месяц	3 месяц	4 месяц	5 месяц	6 месяц
Высшее	3200	3000	2600	2000	1900	1900
Среднее специальное	2600	2000	2000	1900	1800	1700
Среднее	2000	2000	1900	1800	1700	1700

6. В исследовании изучалась эффективность трех рекламных роликов А, Б, В. Для оценки рекламы по девятибалльной шкале выбрали 10 потребителей. При уровне значимости 0,05 выясните, какой ролик можно считать более эффективным.

Рекламные ролики	Потребители									
	1	2	3	4	5	6	7	8	9	10
А	4	5	3	4	3	7	4	3	5	5
Б	7	5	6	5	4	6	5	5	4	4
В	8	7	7	6	8	6	5	8	5	6

7. Данные о прибыли, оборотных средствах, стоимости основных фондов шести предприятий приведены в таблице. Установите наличие линейной взаимосвязи между указанными показателями.

Предприятия	Прибыль	Величина оборотных средств	Стоимость основных фондов
1	188	130	791
2	78	64	910
3	93	69	800
4	152	86	695
5	55	53	936
6	162	141	750

8. Определите наличие линейной зависимости объема потребления по домохозяйству (Y) в зависимости от располагаемого дохода (X) его членов по данным 12 наблюдений, приведенных в таблице.

Наблюдения	1	2	3	4	5	6	7	8	9	10	11	12
X	107	108	110	120	115	122	123	136	140	145	145	156
Y	100	104	109	115	112	120	119	125	136	136	129	136

ЛИТЕРАТУРА

1. Еськова, Л.П. Методы статистической обработки информации в MS Excel: пособие для студентов специальности 1-26 03 01 «Управление информационными ресурсами» / авт.-сост.: О.И. Еськова, Л. П. Авдашкова. – Гомель: учреждение образования «Белорусский торгово-экономический университет потребительской кооперации», 2012. – 132 с.

2. Борздова, Т.В. Основы статистического анализа и обработка данных с применением Microsoft Excel: учеб. пособие / Т.В. Борздова. – Минск: ГИУСТ БГУ, 2011. – 75 с.

3. Бараз, В.Р. Использование MS Excel для анализа статистических данных: учеб. пособие / В.Р. Бараз, В.Ф. Пегашкин; М-во образования и науки РФ; ФГАОУ ВПО «УрФУ им. Первого Президента России Б.Н. Ельцина», Нижнетагил. техн. ин-т (филиал). − 2-е изд., перераб. и доп. − Нижний Тагил: НТИ (филиал) УрФУ, 2014. – 181с.

4. Агапова, Е.Г. Обработка экспериментальных данных в MS Excel: методические указания к выполнению лабораторных работ для студентов дневной формы обучения / сост. Е.Г. Агапова, Е.А. Битехтина. – Хабаровск: Изд-во Тихоокеан. гос. ун-та, 2012. – 32 с.

5. Бурнаева, Э.Г. Статистический пакет анализа данных в Excel 2013: учебное пособие / Э.Г. Бурнаева, С. Н. Леора. – СПб.: СПбГУ, 2020. – 40 с.