Предлагаемый метод будет предпочтителен также там, где имеется очень много входных данных, в которых скрыты закономерности. В этом случае можно учесть различные взаимодействия между показателями-признаками, характеризующими такие данные. Это особенно важно в системах обработки информации (распределенных базах данных, телекоммуникационных и экспертных системах), в частности, для ее предварительного анализа или отбора, выявления «выпадающих» фактов или грубых ошибок человека, принимающего решения. Целесообразно использовать предлагаемый метод в задачах с неполной информацией, а также в тех случаях, когда решение можно подтвердить интуитивными соображениями.

ЛИТЕРАТУРА

1. Тюрин Ю. Н., Макаров А. А. Статистический анализ данных на компьютере. М.: ИНФРА-М., 1998. 528 с.

2. Справочник по прикладной статистике / Под ред. Э. Ллойда, У. Ледермана, Ю. Н. Тюрина. М.: Финансы и статистика, 1990. 830 с.

3. Александров В. В., Горский Н. Д. Алгоритмы структурного метода обработки данных. Л.: Наука, 1983. 208 с.

4. Александров В. В., Алексеев А. И., Горский Н. Д. Анализ данных на ЭВМ. (На примере системы СИТО). М.: Финансы и статистика, 1990. 192 с.

Галина Викторовна РУБЛЕВА — старший преподаватель кафедры математического анализа и теории функций математического факультета

УДК 519.2 (075.8)

ОСОБЕННОСТИ ИСПОЛЬЗОВАНИЯ ПАКЕТА «АНАЛИЗ ДАННЫХ» (MICROSOFT EXCEL)

АННОТАЦИЯ. В статье предложен вариант дополнения к «Справке» пакета «Анализ данных» (Microsoft Excel) по дисперсионному анализу.

In the article the supplement to «Information» of the package «The Analysis of the data» (Microsoft Excel) on dispersing analysis is offered.

В настоящее время существуют пакеты прикладных программ (Microsoft Excel — пакет «Анализ данных», статистические функции, финансовые функции, Statistica, SPSS), позволяющие выполнять расчеты для изучения взаимосвязей между экономическими явлениями и процессами, построения прогнозов в экономике и социологии.

Соответствующая литература представляет собой справочную информацию, предназначенную для специалистов, и не может быть использована в качестве учебных пособий по таким дисциплинам, как «математическая статистика», «эконометрика», «прикладная статистика». Поэтому представляется целесообразным издание специальной методической литературы, содержащей необходимые теоретические сведения, четкие указания по вводу нужной информации и пояснения к выходным данным.

В данной статье предложен вариант методических указаний по теме «Дисперсионный анализ. Использование пакета «Анализ данных» (Microsoft Excel)».

Рассмотрим программы из пакета «Анализ данных» (Microsoft Excel), позволяющие выполнять дисперсионный анализ.

Дисперсионный анализ — это статистический метод анализа результатов, зависящих от действия качественных признаков. Суть метода состоит в том, что общая вариация результативного признака разбивается на части, соответствующие раздельному и совместному влиянию различных качественных факторов, и остаточную вариацию, отражающую влияние всех неучтенных факторов. Статистическое изучение этих частей позволяет делать выводы о том, действительно ли оказывает влияние на результативный признак тот или иной качественный фактор.

С помощью пакета анализа можно выполнять дисперсионный анализ трех видов:

Однофакторный дисперсионный анализ.
Двухфакторный анализ с повторениями.
Двухфакторный анализ без повторений.

1. Однофакторный дисперсионный анализ

В этом случае исследуется наличие или отсутствие влияния на результативный признак одного качественного фактора. Наблюдаемые значения результативного признака группируются по значениям факторного признака. Строится модель:

$y_{ji} = a_i + \varepsilon_{ji}$; $j = 1,2,...,n$; $i = 1,2,...,M$; $\sum_{i=1}^{M} n_i = n$ — число наблюдений;

y_{ji} — наблюдаемые значения результативного признака;
a_i — среднее результативного признака;
ε_{ij} — случайные отклонения.

Ставится задача: на уровне значимости α проверить гипотезу H_0 о равенстве групповых средних при допущении, что групповые теоретические дисперсии хотя и неизвестны, но одинаковы. Другими словами, H_0 : качественный признак не влияет на результативный.

При проверке гипотезы используется F — критерий. Эмпирическое значение критерия рассчитывается по формуле:

$$F_{эмп} = \frac{S^2_{м/гр}}{M-1} : \frac{S^2_{вн/гр}}{n-M},$$

где M — количество групп (число различных значений качественного признака);

$S^2_{м/гр} = \sum_{i=1}^{M} n_i(\bar{y}_i - \bar{y})^2$ — сумма квадратов отклонений между группами;

характеризует вариацию, обусловленную качественным фактором;

$S^2_{вн/гр} = \sum_{i=1}^{M}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ — сумма квадратов отклонений внутри групп;

характеризует остаточную вариацию, обусловленную случайными отклонениями от групповых средних.

При этом:

$S^2_{м/гр} + S^2_{вн/гр} = S^2$ — сумма квадратов отклонений результативного признака от общего среднего;

Если $F_{эмп} > F_{кр}(\alpha; M-1; n-M)$, то гипотеза H_0 отвергается.

Эту программу можно использовать и как статистический тест для определения того, взяты несколько выборок из одной совокупности или нет.

Пример 1:

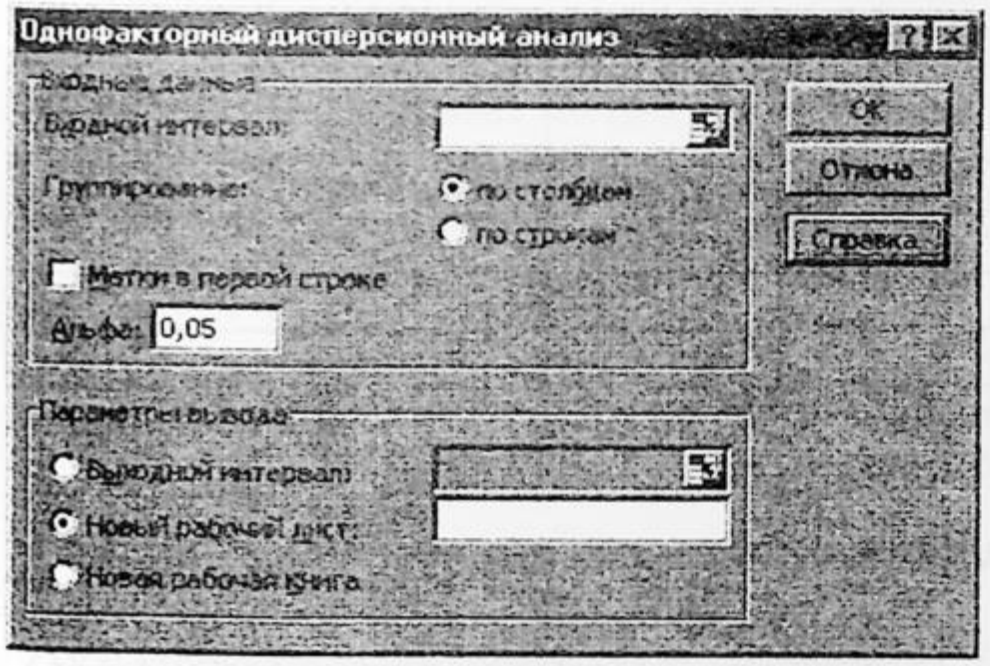
Предположим, что в результате эксперимента получены следующие значения результативного признака, сгруппированные по качественному фактору (табл. 1.1):

Таблица 1.1

Номер наблюдения	Уровни фактора			
	F1	F2	F3	F4
1	0,9	11,9	12	18
2	4,2	5,5	14,6	21,6
3	3,2	4,8	10,9	17,3
4	4,6	5,5	7,6	19,2
5		4,8	11,2	18,5
6		9,7	8,6	
7		7,2	8,4	
8		9,7		
9		4,9		
10		9,6		

При уровне значимости 0,05 проверить гипотезу H_0 о равенстве групповых средних.

Диалоговое окно «Однофакторный дисперсионный анализ» имеет вид (рис 1).



Входной диапазон

Ссылка на диапазон, содержащий анализируемые данные. Ссылка должна состоять не менее чем из двух смежных диапазонов, данные в которых расположены по строкам или столбцам.

Группирование

Установите переключатель в положение «По столбцам» или «По строкам» в зависимости от расположения данных во входном диапазоне.

Метки в первой строке/Метки в первом столбце

Если первая строка исходного диапазона содержит названия столбцов, установите переключатель в положение Метки в первой строке. Если названия строк находятся в первом столбце входного диапазона, установите переключатель в положение Метки в первом столбце. Если входной диапазон не содержит меток, то необходимые заголовки в выходном диапазоне будут созданы автоматически.

Альфа

Введите уровень значимости, необходимый для оценки критических параметров F-статистики. Уровень альфа — это вероятность ошибки I рода (отклонить верную основную гипотезу).

Выходной диапазон

Введите ссылку на ячейку, расположенную в левом верхнем углу выходного диапазона. Размеры выходной области будут рассчитаны автоматически, и соответствующее сообщение появится на экране в том случае, если выходной диапазон занимает место существующих данных или его размеры превышают размеры листа.

Новый лист

Установите переключатель, чтобы открыть новый лист в книге и вставить результаты анализа, начиная с ячейки A1. Если в этом есть необходимость, введите имя нового листа в поле, расположенное напротив соответствующего положения переключателя.

Новая книга

Установите переключатель, чтобы открыть новую книгу и вставить результаты анализа в ячейку A1 на первом листе в этой книге.

Результат анализа — в табл. 1.2, 1.3.

Таблица 1.2

Итоги

Группы	Счет	Сумма	Среднее	Дисперсия
Столбец 1	4	12,9	3,225	2,749166667
Столбец 2	10	73,6	7,36	6,964888889
Столбец 3	7	73,3	10,47143	6,022380952
Столбец 4	5	94,6	18,92	2,727

Таблица 1.3

Дисперсионный анализ

Источник вариации	SS	df	MS	F	P-значение	F критическое
Между группами	651,4801	3	217,16	40,49645787	3,94691E-09	3,04912362
Внутри групп	117,9738	22	5,362445			
Итого	769,4538	25				

SS — сумма квадратов отклонений;

df — число степеней свободы;

MS — среднее суммы квадратов отклонений ($MS = \frac{SS}{df}$).

$S^2_{м/гр} = 651{,}4801$; $S^2_{вн/гр} = 117{,}9738$; $S^2 = 769{,}4538$; $F_{эмп} = 40{,}49645787$.

Так как $F_{эмп} > F_{кр}$, то гипотеза H_0 о равенстве групповых средних отклоняется.

Двухфакторный дисперсионный анализ с повторениями (с несколькими выборками для каждой группы данных)

Двухфакторный дисперсионный анализ представляет собой более сложный вариант анализа, включающий более чем одну выборку для каждой группы данных. В этом случае модель имеет вид:

$y_{jik} = a_{ik} + \varepsilon_{jik}$, где $j = 1, 2, \ldots, n_{ik}$; $i = 1, 2, \ldots, M$; $k = 1, 2, \ldots, K$;

y_{jik} — наблюдаемые значения результативного признака;

a_{ik} — средние значения результативного признака при i-ом значении первого качественного фактора и k-ом значении второго качественного фактора;

ε_{jik} — случайные отклонения; $n_{ik} = N$.

Среднее можно представить в виде:

$a_{ik} = a + \alpha_i + \beta_k + \gamma_{ik}$,

где a — общее среднее результативного признака;

α_i — главные эффекты первого качественного фактора;

β_k — главные эффекты второго качественного фактора;

γ_{ik} — эффекты взаимодействия.

В этом случае общая вариация результативного признака расщепляется на составляющие:

$S^2 = S_A^2 + S_B^2 + S_{AB}^2 + S_\varepsilon^2$, где

S_A^2 — вариация, обусловленная влиянием первого фактора;

S_B^2 — вариация, обусловленная влиянием второго фактора;

S_{AB}^2 — вариация, обусловленная взаимодействием первого и второго факторов;

S_ε^2 — остаточная вариация.

При этом с помощью F-критерия проверяются следующие гипотезы:

H_A: первый фактор не влияет на результат;

H_B: второй фактор не влияет на результат;

H_{AB}: взаимодействие изучаемых качественных факторов не оказывает существенного влияния на результат.

Если $F_{эмп}(A) = \dfrac{S_A^2}{M-1} : \dfrac{S_\varepsilon^2}{M \cdot K \cdot (N-1)} \geq F_{кр}(\alpha; M-1; M \cdot K \cdot (N-1))$,

то H_A отвергается.

Если $F_{эмп}(B) = \dfrac{S_B^2}{K-1} : \dfrac{S_\varepsilon^2}{M \cdot K \cdot (N-1)} \geq F_{кр}(\alpha; K-1; M \cdot K(N-1))$;

то H_B отвергается;

Если

$$F_{эмп}(AB) = \frac{S_{ab}^2}{(M-1)\cdot(K-1)} : \frac{S_\varepsilon^2}{M \cdot K \cdot (N-1)} \geq F_{кр}(\alpha; (K-1)(M-1); M \cdot K(N-1)),$$

то H_{AB} отвергается;

Пример 2:

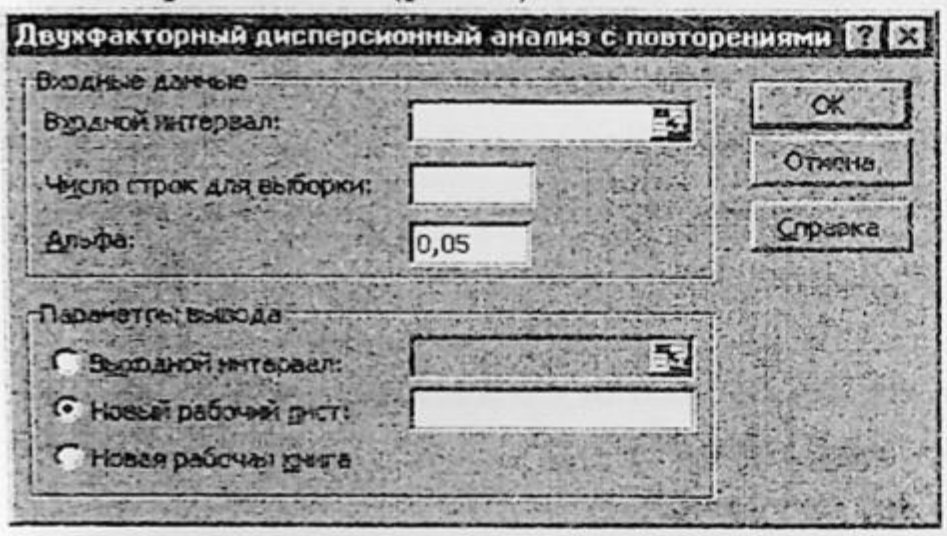
Магазин «Стройматериалы» получает керамическую плитку от 3-х поставщиков. В один контейнер входит 100 коробок. Наудачу из каждой партии и из каждого контейнера было взято по три коробки. Количество бракованных плиток в каждой из них приводится в таблице 2.1:

Таблица 2.1

	Поставщик 1	Поставщик 2	Поставщик 3
Контейнер 1	1	4	6
	2	3	3
	1	5	7
Контейнер 2	1	2	6
	1	0	3
	1	1	7
Контейнер 3	0	0	3
	1	0	3
	1	3	4
Контейнер 4	3	1	3
	1	3	0
	3	3	2

Проверьте гипотезу о том, что среднее количество бракованных плиток не изменяется от поставщика к поставщику и от контейнера к контейнеру. Проверьте гипотезу отсутствия взаимодействия. Уровень значимости α равен 0,05.

Параметры диалогового окна «Двухфакторный дисперсионный анализ с повторениями» (рис 2).



Входной диапазон

Введите ссылку на ячейки, содержащие анализируемые данные. Ссылка должна состоять как минимум из двух смежных диапазонов данных, организованных в виде столбцов или строк.

Число строк для выборки

Введите число строк, содержащихся в одной выборке. Поскольку каждая строка представляет повторение данных, то каждая выборка должна содержать одно и тоже количество строк.

Альфа

Введите уровень значимости, необходимый для оценки критических параметров F-статистики. Уровень альфа – это вероятность ошибки I рода (опровержение верной гипотезы).

Выходной диапазон

Введите ссылку на ячейку, расположенную в левом верхнем углу выходного диапазона. Размеры выходной области будут рассчитаны автоматически, и соответствующее сообщение появится на экране в том случае, если выходной диапазон занимает место существующих данных или его размеры превышают размеры листа.

Новый лист

Установите переключатель, чтобы открыть новый лист в книге и вставить результаты анализа, начиная с ячейки A1. Если в этом есть необходимость, введите имя нового листа в поле, расположенное напротив соответствующего положения переключателя.

Новая книга

Установите переключатель, чтобы открыть новую книгу и вставить результаты анализа в ячейку A1 на первом листе в этой книге.

Полученные результаты представлены в таблицах 2.2 и 2.3:

Таблица 2.2

Двухфакторный дисперсионный анализ с повторениями

ИТОГИ	Поставщик 1	Поставщик 2	Поставщик 3	Итого
Коробка 1				
Счет	3	3	3	9
Сумма	4	12	16	32
Среднее	1,333333	4	5,333333	3,555556
Дисперсия	0,333333	1	4,333333	4,527778
Коробка 2				
Счет	3	3	3	9
Сумма	3	3	16	22
Среднее	1	1	5,333333	2,444444
Дисперсия	0	1	4,333333	6,027778
Коробка 3				
Счет	3	3	3	9
Сумма	2	3	10	15
Среднее	0,666667	1	3,333333	1,666667
Дисперсия	0,333333	3	0,333333	2,5
Коробка 4				
Счет	3	3	3	9
Сумма	7	7	5	19
Среднее	2,333333	2,333333	1,666667	2,111111
Дисперсия	1,333333	1,333333	2,333333	1,361111
Итого				
Счет	12	12	12	
Сумма	16	25	47	
Среднее	1,333333	2,083333	3,916667	
Дисперсия	0,787879	2,810606	4,628788	

Таблица 2.3

Дисперсионный анализ

Источник вариации	SS	Df	MS	F	P-значение	F критическое
Выборка	17,55556	3	5,851852	3,570621	0,028863	3,008786
Столбцы	42,38889	2	21,19444	12,9322	0,000155	3,402832
Взаимодействие	33,61111	6	5,601852	3,418079	0,013952	2,508187
Внутри	39,33333	24	1,638889			
ИТОГО	132,8889	35				

Так как:

$F_{эмп}(A) = 3,570621 > 3,008786 = F_{кр}$, то H_A отвергается,

$F_{эмп}(B) = 12,9322 > 3,402832 = F_{кр}$, то H_B отвергается,

$F_{эмп}(AB) = 3,418079 > 2,508187 = F_{кр}$, то гипотеза о том, что взаимодействие качественных факторов не оказывает существенного влияния на результативный признак, отклоняется.

Двухфакторный дисперсионный анализ без повторений

Статистическая модель:

$Y_{ij} = a + \alpha_i + \beta_j + \varepsilon_{ij}$, где

a — общее среднее результативного признака;

α_i — главный эффект первого качественного фактора, $i = 1, 2, \ldots, M$;

β_j — главный эффект второго качественного фактора, $j = 1, 2, \ldots, K$;

ε_{ij} — случайная компонента ($M\varepsilon = 0$; $D\varepsilon = \sigma^2$).

$S^2 = S_A^2 + S_B^2 + S_\varepsilon^2$.

Гипотезы: равенство эффектов строк $H_A: \alpha_1 = \alpha_2 = \ldots \alpha_M$;

равенство эффектов столбцов $H_B: \beta_1 = \beta_2 = \ldots \beta_M$;

Если $F_{эмп}(A) = \dfrac{S_A^2}{M-1} \div \dfrac{S_\varepsilon^2}{(M-1)(K-1)} > F_{кр}(\alpha; M-1; (M-1)(K-1))$,

то H_A отклоняем.

Если $F_{эмп}(B) = \dfrac{S_B^2}{K-1} \div \dfrac{S_\varepsilon^2}{(M-1)(K-1)} > F_{кр}(\alpha; K-1; (M-1)(K-1))$,

то H_B отклоняем.

Пример 3:

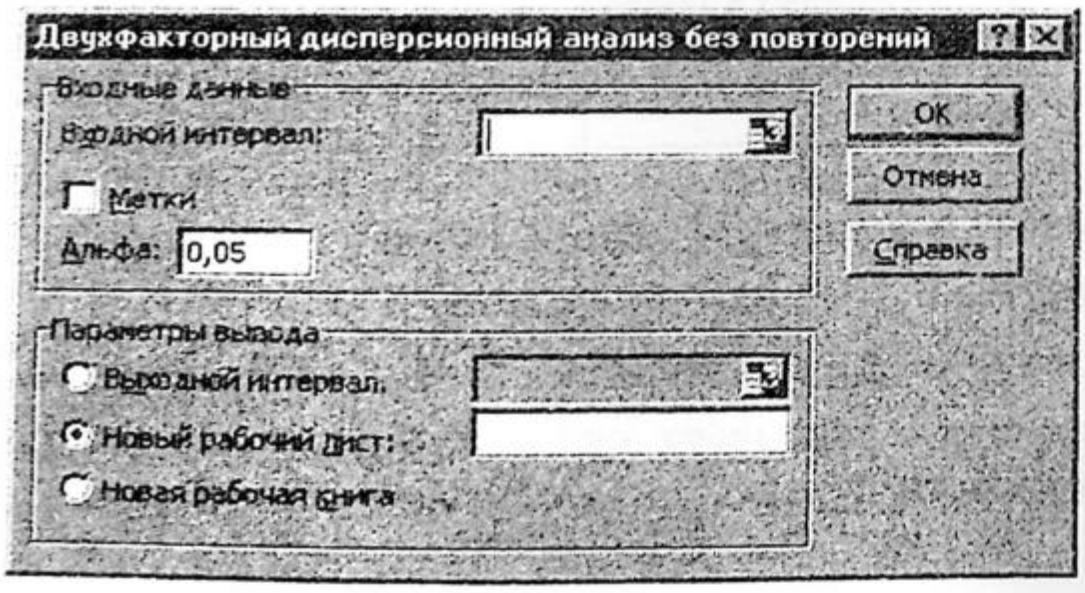
Некто Ельцов имеет магазины в пяти городах (однотипных) и периодически «запускает» рекламу различным образом. Пусть первый факторный признак — код города, второй факторный признак — способ рекламы. Требуется выяснить, отличаются ли три способа рекламирования товара по влиянию на объем его продаж (см табл. 3.1).

Таблица 3.1

Уровни второго фактора	Уровни первого фактора				
	1	2	3	4	5
1	86,7	102,7	204,6	184,6	578,6
2	88,4	108,1	213,2	183,4	593,1
3	81,2	99,8	201,1	179	561,1

Диалоговое окно «Двухфакторный дисперсионный анализ без повторения» имеет вид (рис. 3):

Параметры диалогового окна «Двухфакторный дисперсионный анализ без повторения»



Входной диапазон

Введите ссылку на ячейки, содержащие анализируемые данные. Ссылка должна состоять как минимум из двух смежных диапазонов данных, организованных в виде столбцов или строк.

Заголовки

Снимите флажок, если входной диапазон не содержит названий строк или столбцов, в этом случае подходящие заголовки в выходном диапазоне будут созданы автоматически.

Альфа

Введите уровень значимости, необходимый для оценки критических параметров F-статистики. Уровень альфа связан с вероятностью возникновения ошибки типа I (опровержение верной гипотезы).

Выходной диапазон

Введите уровень значимости, необходимый для оценки критических параметров F-статистики.

Новый лист

Установите переключатель, чтобы открыть новый лист в книге и вставить результаты анализа, начиная с ячейки A1. Если в этом есть необходимость, введите имя нового листа в поле, расположенное напротив соответствующего положения переключателя.

Новая книга

Установите переключатель, чтобы открыть новую книгу и вставить результаты анализа в ячейку A1 на первом листе в этой книге.

Полученные результаты представлены в таблицах 3.2 и 3.3.

Таблица 3.2

Двухфакторный дисперсионный анализ без повторений

ИТОГИ	Счет	Сумма	Среднее	Дисперсия
Строка 1	5	1157,2	231,44	40239,523
Строка 2	5	1186,2	237,24	42235,873
Строка 3	5	1122,2	224,44	38000,583
Столбец 1	3	256,3	85,43333333	14,16333333
Столбец 2	3	310,6	103,5333333	17,74333333
Столбец 3	3	618,9	206,3	38,77
Столбец 4	3	547	182,3333333	8,693333333
Столбец 5	3	1732,8	577,6	256,75

Таблица 3.3

Дисперсионный анализ

Источник вариации	SS	Df	MS	F	P-Значение	F критическое
Строки	410,8	2	205,4	6,285189718	0,022876561	4,458968306
Столбцы	481642,5	4	120410,619	3684,535465	4,3294E-13	3,837854479
Погрешность	261,44	8	32,68			
Итого	482314,7	14				

Так как $F_{эмп}(A) = 6,285189718 > 4,458968306 = F_{кр}$, то гипотезу H_A отклоняем.

Так как $F_{эмп}(B) = 3684,535465 > 3,837854479 = F_{кр}$, то гипотезу H_B отклоняем.

Итак, оба качественных фактора оказывают влияние на результативный признак.