Министерство сельского хозяйства РФ
Департамент научно-технологической политики и образования
ФГОУ ВПО
Волгоградская государственная сельскохозяйственная академия
Кафедра высшей математики

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА. РЕШЕНИЕ ЗАДАЧ ПО ТЕМЕ: «ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ»

Методическая разработка

Волгоград
ИПК «Нива»
2010

УДК 519.2(075.3)
ББК 22.17я72
К 67

Рекомендовано к изданию кафедрой высшей математики ВГСХА
и учебно-методическим советом факультета электрификации с.х.

Корниенко, В.С.
К 67 **Математическая статистика. Решение задач по теме «Однофакторный дисперсионный анализ». Методическая разработка** [Текст] /В.С. Корниенко; Волгогр. гос. с.-х. акад. Волгоград, 2010. 20 с.

Приведены необходимые теоретические сведения. Решено (в «ручном» режиме и с помощью Mathcad) достаточное число задач.

Для студентов специальности 110302 - «Электрификация и автоматизация сельского хозяйства».

УДК 519.2(0.75.3)
ББК 22.17я72

Дисперсионный анализ дает общую схему проверки статистических гипотез, основанную на тщательном изучении различных источников вариации [изменчивости, неоднородности] в сложной ситуации. Он позволяет оценить влияние одного или нескольких факторов на результирующий признак.

Предположения, лежащие в основе дисперсионного анализа, довольно жесткие и подчеркивают тот факт, что данный метод следует использовать только для таких зависимых переменных, которые были тщательно изучены и точно измерены. До тех пор, пока объемы выборок приблизительно равны, дисперсионный анализ может мириться с некоторым нарушением допущений модели. Но в ситуации выборок, сильно отличающихся по объему, следует воспользоваться другими методами (например, хи-квадрат).

На практике часто встречается ситуация, когда можно указать один фактор, влияющий на конечный результат, и этот фактор принимает конечное число значений. Такая ситуация может быть проанализирована при помощи **однофакторного дисперсионного анализа.**

I. Теоретические сведения

1. Условия применимости

Дисперсионный анализ был предложен Р. Фишером* для решения некоторых задач в области биологических исследований, в частности в сельскохозяйственной статистике. В настоящее время *дисперсионный анализ* определяется как *статистический метод*, предназначенный для *оценки влияния различных факторов на результат эксперимента*, в связи с чем, область применения этого метода становится значительно шире. Результатом эксперимента является некоторая случайная величина *X*, называемая также *результативным признаком*. На значения случайной величины *X* влияет фактор *A*, состоящий из нескольких уровней [групп] A_i, $i=\overline{1,r}$.

Рассмотрим простой пример. Директора фирмы интересует зависимость выполненных работ за смену от работающей на стройке бригады. Предположим, что на стройке работают r бригад. Объем выполненных работ является результативным признаком *X*, работающую бригаду назовем фактором A, а через A_i обозначим $i-$й уровень [группу] фактора A ($i-$ю бригаду, $i=\overline{1,r}$).

* Фишер Роналд Эймлер (1890-1962) – английский математик, генетик и статистик. Исследования относятся к математической статистике.

В дисперсионном анализе наблюдаемые величины разбиваются на r групп, причем i – я группа содержит выборку из n_i, $i = \overline{1,r}$, величин $X_i \in N(a + m_i, \sigma_0)$, где σ_0 является постоянной, хотя и неизвестной величиной, не зависящей от i.

Обозначим через $x_{i,j}$ значение j – й величины в i – й группе. Модель однофакторного дисперсионного анализа можно записать в виде

$$x_{i,j} = a + m_i + \varepsilon_{i,j}, \tag{1}$$

где a - генеральное среднее всех мыслимых результатов наблюдений, т.е. $M(X)$, m_i - эффект влияния на X, вызванный i-м уровнем фактора A, или, иначе, отклонение математического ожидания a_i результативного признака при i-м уровне фактора от общего математического ожидания a, т.е. $m_i = a_i - a$; $\varepsilon_{i,j}$ - случайный остаток, отражающий влияние на величину $x_{i,j}$ всех других неконтролируемых факторов.

Основными предпосылками дисперсионного анализа являются:

1) Остатки $\varepsilon_{i,j}$ взаимно независимы для любых i и j.

2) $\varepsilon_{i,j} \in N(0, \sigma_0)$ и σ_0 не зависит от i и j.

Средние значения m_i в (1) могут меняться под влиянием некоторых факторов, например, под влиянием различных способов обработки, различных видов животных или растений, неоднородности почвы и т.д. Целью эксперимента является исследование этой изменчивости средних значений (например, гипотеза H_0 о их равенстве).

2. Разложение суммы квадратов отклонений

Несмещенной оценкой для неизвестной дисперсии σ^2 является, как известно, сумма квадратов

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(x_{i,j} - \bar{x}\right)^2, \tag{2}$$

деленная на $n - 1$, где

$$n = \sum_{i=1}^{r} n_i \tag{3}$$

количество всех наблюдений.

Основная идея дисперсионного анализа заключается в разбиении этой суммы квадратов отклонений на несколько компонент, каждая из которых соответствует предполагаемой причине изменения средних значений m_i.

Обозначим через

$$\bar{x}_i = \frac{1}{n_i} \sum_{i=1}^{n_i} x_{i,j} \tag{4}$$

- среднее арифметическое величин i-й группы, через

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{i,j} \qquad (5)$$

- среднее арифметическое всех величин. Тогда справедливо тождество

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x}\right)^2 = \sum_{i=1}^{r} n_i\left(\bar{x}_i - \bar{x}\right)^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x}_i\right)^2, \qquad (6)$$

или

$$Q = Q_1 + Q_2. \qquad (7)$$

Таким образом, полная сумма квадратов отклонений от общего среднего Q разбивается на две компоненты: Q_1 - сумма квадратов между группами, Q_2 - сумма квадратов внутри групп. Если поделить обе части равенства (6) на число наблюдений n, то получим известное правило сложения дисперсий:

$$\bar{D}_{ОБЩ} = \bar{D}_{ВНГР} + \bar{D}_{МЕЖГР},$$

где

$$\bar{D}_{ОБЩ} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot n_i; \quad \bar{D}_{ВНГР} = \frac{1}{n}\sum N_j \cdot \bar{D}_{ГР_j};$$

$$\bar{D}_{ГР} = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}_j\right)^2; \quad \bar{D}_{МЕЖГР} = \frac{1}{n}\sum N_j\left(\bar{x}_j - \bar{x}\right)^2.$$

П р и м е р. Дана совокупность, состоящая из следующих двух групп:

x	3	1	4	**n1**		2	6	**n2**	**n**
Частота	2	5	3	**10**		4	5	**9**	**19**

Необходимо доказать, что $\bar{D}_{ОБЩ} = \bar{D}_{ВНГР} + \bar{D}_{МЕЖГР}$.

Р е ш е н и е. Дано: $n_1 = 10$, $n_2 = 9$.

Найдем групповые средние:

$$\bar{x}_1 = \frac{1}{n_1}(3\cdot 2 + 1\cdot 5 + 4\cdot 3) = 2{,}3; \quad \bar{x}_2 = \frac{1}{n_2}(2\cdot 4 + 6\cdot 5) = 4{,}222.$$

Найдем групповые дисперсии:

$$\bar{D}_{ГР1} = \frac{1}{10}\left((3-\bar{x}_1)^2\cdot 2 + (1-\bar{x}_1)^2\cdot 5 + (4-\bar{x}_1)^2\cdot 3\right) = 1{,}81;$$

$$\bar{D}_{ГР2} = \frac{1}{9}\left((2-\bar{x}_2)^2\cdot 4 + (6-\bar{x}_2)^2\cdot 5\right) = 3{,}951.$$

Найдем внутригрупповую дисперсию:

$$\bar{D}_{ВНГР} = \frac{1}{n_1+n_2}(n_1\cdot\bar{D}_{ГР1} + n_2\cdot\bar{D}_{ГР2}) = 2{,}824.$$

Найдем общую среднюю:

$$\bar{x} = \frac{1}{n_1+n_2}(3\cdot 2 + 1\cdot 5 + 4\cdot 3 + 2\cdot 4 + 6\cdot 5) = 3{,}2.$$

Найдем общую дисперсию:

$$\overline{D}_{ОБЩ} = \frac{1}{19}\left((3-\overline{x})^2 \cdot 2 + (1-\overline{x})^2 \cdot 5 + (4-\overline{x})^2 \cdot 3 + (2-\overline{x})^2 \cdot 4 + (6-\overline{x})^2 \cdot 5\right) = 3{,}745.$$

Найдем межгрупповую дисперсию:

$$\overline{D}_{МЕЖГР} = \frac{1}{n_1 + n_2}\left((\overline{x}_1 - \overline{x})^2 \cdot 10 + (\overline{x}_2 - \overline{x})^2 \cdot 9\right) = 0{,}921.$$

Убедимся, что общая дисперсия равна сумме внутригрупповой и межгрупповой дисперсий:

$$\overline{D}_{ОБЩ} = \overline{D}_{ВНГР} + \overline{D}_{МЕЖГР} = 2{,}824 + 0{,}921 = 3{,}745.$$

Решим этот пример с помощью Mathcad.

Введем данные:

$$ORIGIN := 1$$

$$G1 := \begin{pmatrix} 3 & 2 \\ 1 & 5 \\ 4 & 3 \end{pmatrix} \qquad G2 := \begin{pmatrix} 2 & 4 \\ 6 & 5 \end{pmatrix} \qquad G := \mathrm{stack}(G1, G2)$$

$$n_1 := \sum G1^{\langle 2 \rangle} \qquad n_2 := \sum G2^{\langle 2 \rangle}$$

Найдем средние:

$$xs_1 := \frac{G1^{\langle 1 \rangle} \cdot G1^{\langle 2 \rangle}}{n_1} \qquad xs_2 := \frac{G2^{\langle 1 \rangle} \cdot G2^{\langle 2 \rangle}}{n_2} \qquad xso := \frac{G^{\langle 1 \rangle} \cdot G^{\langle 2 \rangle}}{\sum n}$$

Найдем дисперсии:

$$D_1 := \frac{\overrightarrow{\left(G1^{\langle 1 \rangle} - xs_1\right)^2 \cdot G1^{\langle 2 \rangle}}}{n_1} \qquad D_2 := \frac{\overrightarrow{\left(G2^{\langle 1 \rangle} - xs_2\right)^2 \cdot G2^{\langle 2 \rangle}}}{n_2}$$

$$Do := \frac{\overrightarrow{\left(G^{\langle 1 \rangle} - xso\right)^2 \cdot G^{\langle 2 \rangle}}}{\sum n} \qquad DvG := \frac{n \cdot D}{\sum n} \qquad DmG := \frac{\overrightarrow{(xs - xso)^2 \cdot n}}{\sum n}$$

Выведем результаты:

$$n = \begin{pmatrix} 10 \\ 9 \end{pmatrix} \qquad xs = \begin{pmatrix} 2.3 \\ 4.222 \end{pmatrix} \qquad xso = 3.211 \qquad D = \begin{pmatrix} 1.81 \\ 3.951 \end{pmatrix}$$

$$Do = 3.745 \qquad DvG = 2.824 \qquad DmG = 0.921$$

Проверим выполнимость правила:

Do = DvG + DmG = 1 ∎

Здесь G1 – первая группа; G2 – вторая группа; G – общая группа; xs_1 – средняя группы G1; xs_2 – средняя группы G2; xso – средняя группы G; D_1 – дисперсия группы G1; D_2 – дисперсия группы G2; Do – дисперсия группы G; DvG – внутригрупповая дисперсия; DmG – межгрупповая дисперсия.

3. Критерий Бартлетта

Одним из условий применения дисперсионного анализа является равенство генеральных групповых дисперсий $\sigma_i^2 = \sigma_0^2$, $i = \overline{1,r}$.

Проверим гипотезу H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_r^2$ с помощью критерия Бартлетта.

Для этого выполним следующую последовательность расчетов.

1. Найдем несмещенные оценки $\overline{S_i^2}$ групповых дисперсий по формуле

$$\overline{S_i^2} = \frac{n_i S_i^2}{n_i - 1}, \quad i = \overline{1,r}. \tag{8}$$

2. Найдем общую несмещенную оценку дисперсии:

$$\overline{S^2} = \frac{(n_1 - 1)\overline{S_1^2} + (n_2 - 1)\overline{S_2^2} + \ldots + (n_r - 1)\overline{S_r^2}}{(n_1 - 1) + (n_2 - 1) + \ldots + (n_r - 1)}. \tag{9}$$

3. Вычислим

$$q = \left[1 + \frac{1}{3(r-1)}\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} + \ldots + \frac{1}{n_r - 1} - \frac{1}{(n_1 - 1)\cdot \ldots \cdot (n_r - 1)}\right)\right]^{-1}. \tag{10}$$

4. Вычислим статистику Бартлетта:

$$\psi = q\left[(n_1 - 1)\ln\left(\frac{\overline{S^2}}{\overline{S_1^2}} + \ldots + (n_r - 1)\ln\left(\frac{\overline{S^2}}{\overline{S_r^2}}\right)\right)\right]. \tag{11}$$

Статистика ψ при $n_i > 3$, $i = \overline{1,r}$, и справедливости гипотезы H_0 имеет распределение, близкое к χ_{r-1}^2, что дает возможность проверить гипотезу H_0 описанными ранее способами.

4. Проверка гипотезы о равенстве групповых средних

Пусть H_0: $m_i = m$, $i = \overline{1,r}$. Заметим, что величина $\overline{S^2} = \frac{Q}{n-1}$, являющаяся несмещенной оценкой для σ^2, всегда будет иметь распределение χ^2 с $n-1$ степенями свободы и по ней можно построить доверительный интервал для σ^2.

Если гипотеза H_0 верна, то величины

$$\overline{S_{\phi}^2} = \frac{1}{r-1} Q_1 \text{ и } \overline{S_0^2} = \frac{1}{n-r} Q_2 \qquad (12)$$

будут иметь распределение Фишера с $r-1$ и $n-r$ степенями свободы, соответственно, при этом $\overline{S_{\phi}^2}$ и $\overline{S_0^2}$ являются несмещенными оценками для межгрупповой дисперсии σ_0^2.

Отношение

$$\psi = \frac{\frac{1}{r-1} Q_1}{\frac{1}{n-r} Q_2} \qquad (13)$$

называется **дисперсионным отношением** и, если гипотеза H_0 верна, то статистика ψ имеет распределение Фишера с $r-1$, $n-r$ степенями свободы. В этом случае эффекты влияния уровней фактора A будут нулевыми, т.е. $m_1 = m_2 = ... = m_r = 0$, а оценка параметра a равна общему среднему $\overline{x}$, вычисленному по формуле (5). Проверка гипотезы H_0 о равенстве групповых средних проводится по схеме, изложенной ранее. Если же гипотеза H_0 отвергается, то параметр a по-прежнему вычисляется по формуле (5), а оценка эффекта m_i влияния *i*-го уровня фактора равна

$$\overline{m_i} = \overline{x_i} - \overline{x}, \qquad (14)$$

где $\overline{x_i}$ определяется по формуле (4), а $\overline{x}$ - по формуле (5). Проверка гипотезы H_0 о равенстве групповых средних проводится по схеме, изложенной ранее.

5. Коэффициент детерминации

Предположим, что фактор A влияет на результативный признак X.

Для измерения степени этого влияния используют **выборочный коэффициент детерминации**, равный

$$\overline{d} = \frac{Q_1}{Q}, \qquad (15)$$

который показывает, какую долю выборочной дисперсии составляет дисперсия групповых средних, иначе говоря, какая доля общей дисперсии объясняется зависимостью результативного признака X от фактора A.

6. Сводка формул

Изложенные выше формулы для решения задач однофакторного анализа приведем в табл. 1.

При вычислении сумм квадратов Q, Q_1, Q_2 часто удобно при $n_i = n_0$ использовать следующие формулы:

$$Q_1 = \frac{\sum_{i=1}^{r}\left(\sum_{j=1}^{n_0} x_{i,j}\right)^2}{n_0} - \frac{\left(\sum_{i=1}^{r}\sum_{j=1}^{n_0} x_{i,j}\right)^2}{n_0 r}, \tag{16}$$

$$Q_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_0} x_{i,j}^2 - \frac{\left(\sum_{i=1}^{r}\sum_{j=1}^{n_0} x_{i,j}\right)^2}{n_0}, \tag{17}$$

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{n_0} x_{i,j}^2 - \frac{\left(\sum_{i=1}^{r}\sum_{j=1}^{n_0} x_{i,j}\right)^2}{n_0 r}. \tag{18}$$

Таблица 1

Компоненты дисперсии	**Сумма квадратов**	**Число степеней свободы**	**Оценки дисперсии**
Межгрупповая	$Q_1 = \sum_{i=1}^{r} n_i \left(\overline{x}_i - \overline{x}\right)^2$	$r-1$	$\overline{S_{\phi}^2} = \frac{Q_1}{r-1}$
Внутригрупповая	$Q_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i} \left(x_{i,j} - \overline{x}_i\right)^2$	$n-r$	$\overline{S_0^2} = \frac{Q_2}{n-r}$
Общая	$Q = \sum_{i=1}^{r}\sum_{j=1}^{n_i} \left(x_{i,j} - \overline{x}\right)^2$	$n-1$	$\overline{S^2} = \frac{Q}{n-1}$

II. Задачи

Задача 1. В табл. 2 приведены данные по объемам работ, выполненных на стройке за смену для четырех бригад.

Таблица 2

Номер бригады	Объем выполненной работы	Групповое среднее	Выборочная смещенная дисперсия
1	140, 144, 142, 145	142,75	3,688
2	150, 149, 152, 152	150,75	1,688
3	148, 149, 146, 147	147,50	1,25
4	150, 155, 154, 152	152,75	3,688

Выполняются ли для этих данных условия проведения дисперсионного анализа?

Р е ш е н и е. Будем считать, что результаты выработок не зависят друг от друга и имеют нормальное распределение. Проверим по критерию Бартлетта гипотезу о равенстве групповых дисперсий. В нашем случае

$$r = 4,\quad n_i = 4,\quad \left(i = \overline{1,4}\right),\quad n = 16.$$

1. Найдем несмещенные оценки дисперсий по формулам (8):

$$\overline{S_1^2} = \frac{S_1^2 n_1}{n_1 - 1} = \frac{3{,}688 \cdot 4}{3} = 4{,}917\,; \quad \overline{S_2^2} = \frac{S_2^2 \cdot n_2}{n_2 - 1} = \frac{1{,}688 \cdot 4}{3} = 2{,}251\,;$$

$$\overline{S_3^2} = \frac{S_3^2 \cdot n_3}{n_3 - 1} = \frac{1{,}25 \cdot 4}{3} = 1{,}667\,; \quad \overline{S_4^2} = \frac{S_4^2 \cdot n_4}{n_4 - 1} = \frac{3{,}688 \cdot 4}{3} = 4{,}917\,.$$

2. По формуле (9) найдем общую несмещенную дисперсию:

$$\overline{S^2} = \frac{(n_1 - 1)S_1^2 + \ldots + (n_4 - 1)S_4^2}{(n_1 - 1) + \ldots + (n_4 - 1)} = \frac{3(4{,}917 + 2{,}251 + 1{,}667 + 4{,}917)}{3 \cdot 4} = 3{,}438\,.$$

3. По формуле (10) найдем параметр

$$q = \left[1 + \frac{1}{3(4-1)}\left(\frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} - \frac{1}{3+3+3+3}\right)\right]^{-1} = 0{,}878\,.$$

4. Вычислим статистику для критерия Бартлетта:

$$\psi = 0{,}878\left(3\ln\left(\frac{3{,}438}{4{,}917}\right) + 3\ln\left(\frac{3{,}438}{2{,}251}\right) + 3\ln\left(\frac{3{,}438}{1{,}667}\right) + 3\ln\left(\frac{3{,}438}{4{,}917}\right)\right) = 3{,}986\,.$$

Проверим гипотезу H_0: $\sigma_1^2 = \ldots = \sigma_1^4$ на уровне значимости $\alpha = 0{,}05$. По таблицам распределения χ_3^2 при $\alpha = 0{,}05$ находим $\chi_{3,кр}^2 = 7{,}82$. Так как $\psi = 3{,}986 < \chi_{3,кр}^2$, то гипотезу H_0 принимаем.

Mathcad-контроль:

Сперва проверим результаты 3-го и 4-го столбцов табл. 2:

$$G1 := \begin{pmatrix} 140 \\ 144 \\ 142 \\ 145 \end{pmatrix} \quad G2 := \begin{pmatrix} 150 \\ 149 \\ 152 \\ 152 \end{pmatrix} \quad G3 := \begin{pmatrix} 148 \\ 149 \\ 146 \\ 147 \end{pmatrix} \quad G4 := \begin{pmatrix} 150 \\ 155 \\ 154 \\ 152 \end{pmatrix}$$

$\mathrm{mean}(G1) = 142.75$ $\quad\mathrm{mean}(G2) = 150.75$ $\quad\mathrm{mean}(G3) = 147.5$

$\mathrm{mean}(G4) = 152.75$ $\quad\mathrm{var}(G1) = 3.688$ $\quad\mathrm{var}(G2) = 1.688$

$\mathrm{var}(G3) = 1.25$ $\quad\mathrm{var}(G4) = 3.688$

Теперь найдем несмещенные оценки дисперсий:

$\mathrm{Var}(G1) = 4.917$ $\quad\mathrm{Var}(G2) = 2.25$ $\quad\mathrm{Var}(G3) = 1.667$

$\mathrm{Var}(G4) = 4.917$

Вычисляем общую несмещенную дисперсию:

$$\frac{\mathrm{Var}(G1) + \mathrm{Var}(G2) + \mathrm{Var}(G3) + \mathrm{Var}(G4)}{4} = 3.438$$

Вычислим параметр q:

$$q := \left[1 + \frac{1}{3\cdot(4-1)}\cdot\left(\frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} - \frac{1}{3+3+3+3}\right)\right]^{-1} \qquad q = 0.878$$

Вычислим статистику для критерия Бартлетта:

$$\psi := 3\cdot q\cdot \ln\left(\frac{3.438^4}{\mathrm{Var(G1)\cdot Var(G2)\cdot Var(G3)\cdot Var(G4)}}\right) \qquad \psi = 1.139$$

Найдем χ_{kr}:

$$\alpha := 0.05 \qquad \chi kr := \mathrm{qchisq}(1-\alpha, 3) \qquad \chi kr = 7.815$$

Задача 2. Для данных задачи 1 проверить гипотезу дисперсионного анализа о равенстве средних H_0: $m_1 = m_2 = m_3 = m_4$.

Р е ш е н и е. Для проверки гипотезы H_0 вычислим суммы Q_1 и Q_2.

Общее выборочное среднее равно

$$\bar{x} = \frac{140+144+142+\ldots+154+152}{16} = 148{,}438.$$

Тогда

$$Q_1 = \sum_{i=1}^{4} 4\left(\bar{x}_i - \bar{x}\right)^2 = 4\left[(142{,}75-148{,}31)^2 + \ldots + (152{,}75-148{,}31)^2\right] = 228{,}688;$$

$$Q_2 = \sum_{i=1}^{4}\sum_{j=1}^{4}\left(x_{i,j} - \bar{x}_i\right)^2 = (140-142{,}75)^2 + \ldots + (152-152{,}75)^2 = 41{,}25.$$

Вычислим статистику Фишера:

$$\psi = \frac{(n-r)Q_1}{(r-1)Q_2} = \frac{(16-4)\cdot 228{,}688}{(4-1)\cdot 41{,}25} = 22{,}176.$$

По таблицам распределения Фишера для $\alpha = 0{,}05$ и степеней свободы $k_1 = 3$, $k_2 = 12$ найдем критическое значение $F_{кр} = 3{,}49$. Так как $\psi > F_{кр}$, то гипотезу H_0 отклоняем, т.е. считаем, что объем ежедневной выработки зависит от работающей бригады. Оценим степень этой зависимости с помощью коэффициента детерминации. Для этого вычислим Q:

$$Q = \sum_{i=1}^{4}\sum_{j=1}^{4}\left(x_{i,j} - \bar{x}\right)^2 = (140-148{,}31)^2 + \ldots + (152-148{,}31)^2 = 269{,}938.$$

Контроль: $Q = Q1 + Q2$.

По формуле (15) получим

$$\bar{d} = \frac{Q_1}{Q} = \frac{220{,}19}{259{,}46} = 0{,}847.$$

Это означает, что 84,7 % общей вариации ежедневного объема выработки связано с работающей сменой.

Mathcad-контроль:

$$G := \mathrm{augment}(G1, G2, G3, G4) \qquad a := \mathrm{mean}(G) \qquad a = 148.438$$

$$Q1 := (142.75 - a)^2 + (150.75 - a)^2 + (147.5 - a)^2 + (152.75 - a)^2$$

$$Q1 := 4 \cdot Q1 \qquad Q1 = 228.688$$

$$\begin{aligned} Q2 := &\left[(140 - 142.75)^2 + (144 - 142.75)^2 + (142 - 142.75)^2\right] \ldots \\ &+ (145 - 142.75)^2 + (150 - 150.75)^2 + (149 - 150.75)^2 \ldots \\ &+ \left[(152 - 150.75)^2 + (152 - 150.75)^2 + (148 - 147.5)^2\right] \ldots \\ &+ \left[(149 - 147.5)^2 + (146 - 147.5)^2 + (147 - 147.5)^2\right] \ldots \\ &+ \left[(150 - 152.75)^2 + (155 - 152.75)^2 + (154 - 152.75)^2\right] .. \\ &+ (152 - 152.75)^2 \end{aligned}$$

$$Q2 = 41.25$$

$$\begin{aligned} Q := &\left[(140 - a)^2 + (144 - a)^2 + (142 - a)^2 + (145 - a)^2 + (150 - a)^2\right] \ldots \\ &+ \left[(149 - a)^2 + (152 - a)^2 + (152 - a)^2 + (148 - a)^2 + (149 - a)^2\right] .. \\ &+ \left[(146 - a)^2 + (147 - a)^2 + (150 - a)^2 + (155 - a)^2 + (154 - a)^2\right] .. \\ &+ (152 - a)^2 \end{aligned}$$

$$Q = 269.938 \qquad Q = Q1 + Q2 = 1$$

$$\alpha := 0.05 \qquad Fkr := qF(1 - \alpha, 3, 12) \qquad Fkr = 3.49$$

$$d := \frac{Q1}{Q} \qquad d = 0.847$$

Задача 3. По данным задач 1 и 2 найти оценки параметров модели (1) дисперсионного анализа.

Р е ш е н и е. Так как объем ежедневной выработки зависит от работающей смены, то оценка параметра a равна $\overline{x} = 148{,}31$. Оценки параметров m_i равны

$$\overline{m_i} = \overline{x_i} - \overline{x},$$

т.е.

$$\overline{m_1} = 142{,}75 - 148{,}31 = -5{,}56; \quad \overline{m_2} = 150{,}15 - 148{,}31 = 1{,}94;$$

$$\overline{m_3} = 147{,}5 - 148{,}31 = -0{,}81; \quad \overline{m_4} = 152{,}75 - 148{,}31 = 4{,}44.$$

Оценка параметра σ_0^2 равна

$$\overline{\sigma_0^2} = \frac{Q}{r-1} = \frac{220{,}19}{3} = 73{,}39.$$

Сформулируем базовую задачу однофакторного дисперсионного анализа, и укажем методику ее решения.

Базовая задача. Пусть имеется m партий A_1, $A_2, \ldots,$ A_m изделий*. Из каждой партии отобрано соответственно n_1, $n_2, \ldots,$ n_m изделий. Значения показателя качества этих изделий представлены в виде табл. 3 наблюдений. Необходимо на уровне значимости $\alpha = 0{,}05$ проверить существенность влияния партий (фактор A) на их качество.

Таблица 3

Партия **(фактор *A*)**	Номер наблюдения			
	1	2	…	
A_1	$a_{1,1}$	$a_{1,2}$	$\cdots$	a_{1,n_1}
A_2	$a_{2,1}$	$a_{2,2}$	$\cdots$	a_{2,n_2}
…	…	…	…	…
A_m	$a_{m,1}$	$a_{m,2}$	$\cdots$	a_{m,n_m}

Возможный путь решения в Mathcad.

1) Введем данные:

$ORIGIN := 1$

$A1 :=$ матрица размера $1 \times n_1$

$A2 :=$ матрица размера $1 \times n_2$

………………………………

$Am :=$ матрица размера $1 \times n_m$

$A := augment(A1, A2, ..., Am)$ матрица-строка данных табл. 3

$m :=$ число уровней фактора A

$$n := \begin{pmatrix} cols(A1) \\ cols(A2) \\ ... \\ cols(A3) \end{pmatrix} \qquad N := \sum n \quad \text{число столбцов матрицы } A$$

2) Проверим гипотезу о равенстве групповых дисперсий для всех уровней фактора. Для этого вычислим критерий Бартлетта b.

$$V := \begin{pmatrix} Var(A1) \\ Var(A2) \\ ... \\ Var(Am) \end{pmatrix} \text{ - матрица несмещенных дисперсий}$$

* Партию естественно назвать *фактором*, а каждую конкретную партию A_i $(i = \overline{1,m})$ *уровнем фактора*; m - полное число взятых партий.

$$S := \frac{(n-1)\cdot V}{\sum(n-1)} \qquad drob := \frac{1}{\prod_{i=1}^{m}(n_i - 1)} \qquad q := \left(1 - \frac{1}{3(m-1)} \cdot \left(\sum\left(\overrightarrow{\left(\frac{1}{n-1}\right)} + drob\right)\right)\right)^{-1}$$

$b := q\cdot(n-1)\cdot\ln\left(\frac{S}{V}\right)$ - критерий Бартлетта

$\alpha := 0.05 \qquad \chi_{\alpha,кр} := qchisq(1-\alpha, m-1)$

$b < \chi_{\alpha,кр} =$ сравниваем критерий Бартлетта с критическим значением $\chi_{\alpha,кр}$

Делаем **вывод:** если $b < \chi_{\alpha,кр}$ (т.е. если после знака равенства компьютер выдаст 1), то групповые дисперсии всех уровней фактора *совпадают*.

3) Продолжаем проводить дисперсионный анализ.

а) Вычисляем **межгрупповую дисперсию** $s12$:

$$sr := \begin{pmatrix} mean(A1) \\ mean(A2) \\ \dots \\ mean(Am) \end{pmatrix} \qquad Q1 := \begin{pmatrix} \sum\overrightarrow{(A1 - sr_1)^2} \\ \sum\overrightarrow{(A2 - sr_2)^2} \\ \dots \\ \sum\overrightarrow{(Am - sr_m)^2} \end{pmatrix} \qquad s12 := \frac{\sum Q1}{N-m}$$

б) Вычисляем **внутригрупповую дисперсию** $s22$:

$$sr1 := mean(A) \qquad Q2 := (sr - sr1)^2\cdot n \qquad s22 := \frac{Q2}{m-1}$$

в) Находим FN, $F_{\alpha,кр}$ и сравниваем их

$$FN := \frac{s22}{s12} \qquad F_{\alpha,кр} = qF(1-\alpha, m-1, N-m) \qquad FN < F_{\alpha,кр} =$$

Делаем **вывод:** если $FN < F_{\alpha,кр}$ (т.е. если после знака равенства компьютер поставил 1), то выбор партии (фактор A) на качество изделий *незначимо*, в противном случае *значимо*.

Задача 4. Выяснить на уровне значимости $\alpha = 0{,}05$, зависит ли урожайность сельскохозяйственной культуры от технологии обработки почвы, по результатам табл. 4.

Таблица 4

Уровни технологии	Годы					
	1	2	3	4	5	6
1	140	141	140	141	142	145
2	150	149	150	147		
3	147	147	145	150	150	
4	144	147	142	146		

Решение.

1) Введем данные

$ORIGIN := 1$

$A1 := (140 \quad 141 \quad 140 \quad 141 \quad 142 \quad 145)$

$A2 := (150 \quad 149 \quad 150 \quad 147)$

$A3 := (147 \quad 147 \quad 145 \quad 150 \quad 150)$

$A4 := (144 \quad 147 \quad 142 \quad 146) \qquad A := \mathrm{augment}(A1, A2, A3, A4)$

$$m := 4 \qquad n := \begin{pmatrix} \mathrm{cols}(A1) \\ \mathrm{cols}(A2) \\ \mathrm{cols}(A3) \\ \mathrm{cols}(A4) \end{pmatrix} \qquad N := \sum n$$

2) Проверим гипотезу о равенстве групповых дисперсий для всех уровней фактора

$$s := \begin{pmatrix} \mathrm{Var}(A1) \\ \mathrm{Var}(A2) \\ \mathrm{Var}(A3) \\ \mathrm{Var}(A4) \end{pmatrix} \qquad S := \frac{(n-1)\cdot s}{\sum (n-1)} \qquad drob := \frac{1}{\prod_{i=1}^{m} (n_i - 1)}$$

$$q := \left[1 - \frac{1}{3(m-1)} \cdot \left[\sum \left(\overrightarrow{\left(\frac{1}{n-1} \right)} + drob \right) \right] \right]^{-1} \qquad b := q \cdot (n-1) \cdot \ln\left(\frac{S}{s}\right)$$

$\alpha := 0.05 \qquad \chi kr\alpha := \mathrm{qchisq}(1 - \alpha, m - 1) \qquad b < \chi kr\alpha = 1$

Вывод: групповые дисперсии для всех уровней фактора *совпадают.*

3) Продолжаем проводить дисперсионный анализ

$$sr := \begin{pmatrix} \mathrm{mean}(A1) \\ \mathrm{mean}(A2) \\ \mathrm{mean}(A3) \\ \mathrm{mean}(A4) \end{pmatrix}$$

$$Q1 := \begin{pmatrix} \sum \overrightarrow{(A1 - sr_1)^2} \\ \sum \overrightarrow{(A2 - sr_2)^2} \\ \sum \overrightarrow{(A3 - sr_3)^2} \\ \sum \overrightarrow{(A4 - sr_4)^2} \end{pmatrix} \qquad s12 := \frac{\sum Q1}{N - m}$$

$$sr1 := \mathrm{mean}(A) \qquad Q2 := (sr - sr1)^2 \cdot n \qquad s22 := \frac{Q2}{m - 1}$$

$$FN := \frac{s22}{s12} \qquad Fkr\alpha := qF(1 - \alpha, m - 1, N - m) \qquad FN < Fkr\alpha = 0$$

Вывод: гипотеза о том, что все технологии одинаково эффективны, *отвергается*. Делаем вывод о том, что выбор технологии *влияет* на урожайность.

Оценим степень этого влияния. Для этого проведем вычисление *коэффициента детерминации* $r2$, который показывает, какую часть в общей дисперсии величин фактора A составляет часть, обусловленная зависимостью от фактора технологии.

$$r2 := \frac{s22 \cdot (m - 1)}{s12 \cdot (N - m) + s22 \cdot (m - 1)} \qquad r2 = 0.753$$

$$i := 1 .. m \qquad a_i := sr_i \qquad a = \begin{pmatrix} 141.5 \\ 149 \\ 147.8 \\ 144.75 \end{pmatrix} \qquad \sigma := \sqrt{s12} \qquad \sigma = 1.95$$

В нашем случае r2 = 0,753, т.е. 75,3 % общей вариации урожайности обусловлены технологией.

Вывод: из приведенных вычислений следует, например, что урожайность при второй технологии обработки представляет собой случайную величину, имеющую нормальное распределение $N(149;1,95)$.

Задача 5. В течение шести лет использовались пять различных технологий по выращиванию сельскохозяйственной культуры. Данные по урожайности (в ц/га) приведены в табл. 5. Необходимо на уровне значимости $\alpha = 0{,}05$ установить влияние различных технологий на урожайность культуры.

Таблица 5

Номер технологии	Годы					
	1	2	3	4	5	6
A_1	1,2	1,1	1,0	1,3	1,1	0,8
A_2	0,6	1,1	0,8	0,7	0,7	0,9
A_3	0,9	0,6	0,8	1,0	1,0	1,1
A_4	1,7	1,4	1,3	1,5	1,2	1,3
A_5	1,0	1,4	1,1	0,9	1,2	1,5

Р е ш е н и е. Понятно, какие изменения следует произвести в Mathcad-программе при решении задачи 4.

Приведем полное решение задачи 5.

1)

ORIGIN := 1

$A1 := (1.2 \quad 1.1 \quad 1.0 \quad 1.3 \quad 1.1 \quad 0.8)$

$A2 := (0.6 \quad 1.1 \quad 0.8 \quad 0.7 \quad 0.7 \quad 0.9)$

$A3 := (0.9 \quad 0.6 \quad 0.8 \quad 1.0 \quad 1.0 \quad 1.1)$

$A4 := (1.7 \quad 1.4 \quad 1.3 \quad 1.5 \quad 1.2 \quad 1.3)$

$A5 := (1.0 \quad 1.4 \quad 1.1 \quad 0.9 \quad 1.2 \quad 1.5)$

$A := \mathrm{augment}(A1, A2, A3, A4, A5)$

$$m := 4 \qquad n := \begin{pmatrix} \mathrm{cols}(A1) \\ \mathrm{cols}(A2) \\ \mathrm{cols}(A3) \\ \mathrm{cols}(A4) \\ \mathrm{cols}(A5) \end{pmatrix} \qquad N := \sum n$$

2)

$$s := \begin{pmatrix} \mathrm{Var}(A1) \\ \mathrm{Var}(A2) \\ \mathrm{Var}(A3) \\ \mathrm{Var}(A4) \\ \mathrm{Var}(A5) \end{pmatrix} \qquad S := \frac{(n-1)\cdot s}{\sum (n-1)} \qquad drob := \frac{1}{\prod_{i=1}^{m} (n_i - 1)}$$

$$q := \left[1 - \frac{1}{3(m-1)} \cdot \left[\sum \left(\overrightarrow{\left(\frac{1}{n-1}\right)} + drob\right)\right]\right]^{-1} \qquad b := q\cdot(n-1)\cdot \ln\left(\frac{S}{s}\right)$$

$$\alpha := 0.05 \qquad \chi kr\alpha := qchisq(1-\alpha, m-1) \qquad b < \chi kr\alpha = 1$$

Вывод: групповые дисперсии для всех уровней *совпадают*.
3)

$$sr := \begin{pmatrix} \mathrm{mean}(A1) \\ \mathrm{mean}(A2) \\ \mathrm{mean}(A3) \\ \mathrm{mean}(A4) \\ \mathrm{mean}(A5) \end{pmatrix} \qquad Q1 := \begin{pmatrix} \sum \overrightarrow{(A1 - sr_1)^2} \\ \sum \overrightarrow{(A2 - sr_2)^2} \\ \sum \overrightarrow{(A3 - sr_3)^2} \\ \sum \overrightarrow{(A4 - sr_4)^2} \\ \sum \overrightarrow{(A5 - sr_5)^2} \end{pmatrix} \qquad s12 := \frac{\sum Q1}{N - m}$$

$$sr1 := \mathrm{mean}(A) \qquad Q2 := (sr - sr1)^2 \cdot n \qquad s22 := \frac{Q2}{m-1}$$

$$FN := \frac{s22}{s12} \qquad Fkr\alpha := qF(1-\alpha, m-1, N-m) \qquad FN < Fkr\alpha = 0$$

Вычислим коэффициент детерминации $r2$:

$$r2 := \frac{s22\cdot(m-1)}{s12\cdot(N-m) + s22\cdot(m-1)} \qquad r2 = 0.599$$

Вывод: влияние типа технологии (фактора A) на урожайность *значимо.*

Поскольку $r2 = 0{,}599$, то 59,9 % общей вариации урожайности обусловлено технологией.

Задача 6. (Решить самостоятельно). Предприятие решает вопрос о том, какую из трех систем контроля качества выбрать. Все три системы были тестированы. Результаты тестов были отображены в табл. 6.

Таблица 6

Номер системы	Число выявленных бракованных изделий партии продукции
1	1, 2, 3, 0, 2, 1
2	2, 3, 1, 0, 1
3	2, 2, 3, 2

Проверить гипотезу об отсутствии влияния различий между системами на результаты тестирования систем. Доверительная вероятность равна 95 %. Предполагается, что выборки получены из независимых нормальных генеральных совокупностей с одной и той же генеральной дисперсией.

Задача 7. (Решить самостоятельно). На заводе установлено четыре линии по выпуску облицовочной плитки. С каждой линии случайным образом в течение смены отобрано 10 плиток и сделаны замеры их толщины (мм). Отклонения от номинального размера приведены в табл. 7. Требуется на уровне значимости 5 % установить зависимость выпуска качественных плиток от линии выпуска (фактор A).

Таблица 7

Линия по выпуску плитки **(фактор A)**	Номер испытания									
	1	2	3	4	5	6	7	8	9	10
A_1	0,6	0,2	0,4	0,5	0,8	0,2	0,1	0,6	0,8	0,8
A_2	0,2	0,2	0,4	0,3	0,3	0,6	0,8	0,2	0,5	0,5
A_3	0,8	0,6	0,2	0,4	0,9	1,1	0,8	0,2	0,4	0,8
A_4	0,7	0,7	0,3	0,3	0,2	0,8	0,6	0,4	0,2	0,6

Ответ: влияние линии (фактор A) на качество облицовочных плиток *незначимо*. Только 9,7 % общей вариации качества плиток обусловлено линией по выпуску плиток.

Задача 8. (Решить самостоятельно). Имеются четыре партии сырья: A_1, A_2, A_3, A_4. Из каждой партии отобрано по пять образцов и проведены испытания на определение величины разрывной нагрузки. Результаты испытаний приведены в табл. 8. Необходимо на уровне значимости 5 % выяснить, существенно ли влияние различных партий сырья на величину разрывной нагрузки.

Т а б л и ц а 8

Партия **(фактор A)**	Разрывная нагрузка (кг/см^2)				
A_1	200	140	170	145	165
A_2	190	150	210	150	150
A_3	230	190	200	190	200
A_4	150	170	150	170	180

Ответ: различие между партиями сырья оказывает *существенное влияние* на величину разрыва нагрузки. Лишь 40,7 % изменчивости исследуемой случайной величины обусловлено изменением фактора A.