Однофакторный дисперсионный анализ для несвязных выборок

Пусть на количественный нормально распределенный признак X действует некоторый фактор F, имеющий p постоянных уровней $F_1, F_2, \ldots, F_p$. На каждом уровне осуществлено по q испытаний над разными выборками. Результаты наблюдений - числа x_{ij}, где i - номер испытания, $i = 1, 2, \ldots, q$; j - номер уровня фактора, $j = 1, 2, \ldots, p$, - записывают в виде таблицы

Номер испытания i	Уровень фактора			
	F_1	F_2	$\cdots$	F_p
1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
q	x_{q1}	x_{q2}	$\cdots$	x_{qp}
Групповое среднее $\overline{x}_{\text{гр}}$	$\overline{x}_{\text{гр. }1}$	$\overline{x}_{\text{гр. }2}$	$\cdots$	$\overline{x}_{\text{гр. }p}$

Задача заключается в следующем: при уровне значимости α про- верить основную гипотезу о равенстве групповых средних при условии, что групповые генеральные дисперсии хотя и неизвестны, но одинаковы. Чтобы решить эту задачу, находят:

а) *общую сумму* квадратов отклонений наблюдаемых значений признака от общего выборочного среднего:

$$S_{\text{общ}} = \sum_{j=1}^{p}\sum_{i=1}^{q}(x_{ij} - \overline{x}_{\text{в}})^2$$

б) *факторную сумму* квадратов отклонений групповых средних от общего выборочного среднего (характеризует рассеяние «между группами»):

$$S_{\text{факт}} = q\sum_{j=1}^{p}(\overline{x}_{\text{гр. }j} - \overline{x}_{\text{в}})^2$$

в) *остаточную сумму* квадратов отклонений наблюдаемых значений группы от своего группового среднего (характеризует рассеяние «внутри группы»):

$$S_{\text{ост}} = \sum_{j=1}^{p}\sum_{i=1}^{q}(x_{ij} - \overline{x}_{\text{гр. }j})^2.$$

Практически остаточную сумму находят по формуле

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}}.$$

Общую и факторную суммы удобнее вычислять по таким формулам:

$$S_{\text{общ}} = \sum_{j=1}^{p} P_j - \frac{\left(\sum_{j=1}^{p} R_j\right)^2}{pq};$$

$$S_{\text{факт}} = \frac{\sum_{j=1}^{p} R_j^2}{q} - \frac{\left(\sum_{j=1}^{p} R_j\right)^2}{pq},$$

где $P_j = \sum_{i=1}^{q} x_{ij}^2$ — сумма квадратов наблюдаемых значений признака на уровне F_j; $R_j = \sum_{i=1}^{q} x_{ij}$ — сумма наблюдаемых значений признака на уровне F_j.

Если наблюдаемые значения признака сравнительно большие, то для упрощения вычислений вычитают от каждого наблюдаемого значения одно и то же число *C*, которое приблизительно равняется общему среднему. Если уменьшенные значения обозначить $y_{ij} = x_{ij} - C,$

$$S_{\text{общ}} = \sum_{j=1}^{p} Q_j - \frac{\left(\sum_{j=1}^{p} T_j\right)^2}{pq}$$

$$S_{\text{факт}} = \frac{\sum_{j=1}^{p} T_j^2}{q} - \frac{\left(\sum_{j=1}^{p} T_j\right)^2}{pq}$$

где $Q_j = \sum_{i=1}^{q} y_{ij}^2$ — сумма квадратов уменьшенных значений признака на уровне F_j; $T_j = \sum_{i=1}^{q} y_{ij}$ — сумма уменьшенных значений признака на уровне F_j.

Поделив факторную и остаточную суммы на соответствующее число степеней свободы

$$k_{\text{факт}} = p - 1, \qquad k_{\text{ост}} = p(q-1),$$

находят факторную и остаточную дисперсии:

$$s^2_{\text{факт}} = \frac{S_{\text{факт}}}{p-1}; \qquad s^2_{\text{ост}} = \frac{S_{\text{ост}}}{p(q-1)}.$$

Факторную и остаточную дисперсии сравнивают по *критерию Фишера-Снедекора*:

$$F_{\text{набл}} = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}}.$$

Далее по таблице критических точек распределения Фишера-Снедекора при заданном уровне значимости α и числе степеней свободы находят критическую точку Fкр(α; k1, k2).

$$k_1 = k_{\text{факт}}, \qquad k_2 = k_{\text{ост}}$$

Если $F_{\text{набл}} < F_{\text{кр}}$, различие групповых средних считают незначащим. Если $F_{\text{набл}} > F_{\text{кр}}$, то различие считают значащим.

Пример 1. Проведено по семь испытаний на каждом из четырёх уровней фактора над разными выборками. Методом дисперсионного анализа при уровне значимости 0,05 проверить основную гипотезу о равенстве групповых средних. Предполагается, что выборки получены из нормальных совокупностей с одинаковыми дисперсиями.
Результаты испытания приведены в таблице.

Номер испытания i	Уровень фактора			
	F_1	F_2	F_3	F_4
1	24	26	21	25
2	25	27	22	26
3	26	28	24	28
4	29	29	25	28
5	31	30	27	29
6	32	32	29	30
7	33	35	30	33
Групповое среднее $\overline{x}_{\text{гр}}$	28,57143	29,57143	25,42857	28,42857

Решение. Найдём общее среднее:
$x_в$ = (24 + 26 + 21 + 25 + 25 + 27 + 22 + 26 + 26 + 28 + 24
+ 28 + + 29 + 29 + 25 + 28 + 31 + 30 + 27 + 29 + 32 + 32
+ 29 + 30 +
+ 33 + 35 + 30 + 33)/(7 · 4) = 28.

Решение. Найдём общее среднее:
$x_в$ = (24 + 26 + 21 + 25 + 25 + 27 + 22 + 26 + 26 + 28 + 24 + 28 + + 29 + 29 + 25 + 28 +
31 + 30 + 27 + 29 + 32 + 32 + 29 + 30 +
+ 33 + 35 + 30 + 33)/(7 · 4) = 28.

Для упрощения расчётов вычтем от каждого наблюдаемого значения x_{ij} общее среднее $x_в$ = 28, т. е. перейдём к уменьшенным величинам $y_{ij} = x_{ij} - 28.$

Составим расчётную таблицу

Номер испытания i	Уровень фактора								Итоговый столбик
	F_1		F_2		F_3		F_4		
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	y_{i4}	y_{i4}^2	
1	−4	16	−2	4	−7	49	−3	9	
2	−3	9	−1	1	−6	36	−2	4	
3	−2	4	0	0	−4	16	0	0	
4	1	1	1	1	−3	9	0	0	
5	3	9	2	4	−1	1	1	1	
6	4	16	4	16	1	1	2	4	
7	5	25	7	49	2	4	5	25	
$Q_j = \sum_{i=1}^{7} y_{ij}^2$		80		75		116		43	$\sum_{j=1}^{4} Q_j = 314$
$T_j = \sum_{i=1}^{7} y_{ij}$	4		11		−18		3		$\sum_{j=1}^{4} T_j = 0$
T_j^2	16		121		324		9		$\sum_{j=1}^{4} T_j^2 = 470$

Используя итоговый столбик таблицы, находим общую и факторную суммы квадратов отклонений, учитывая, что количество уровней фактора p = 4, количество испытаний на каждом уровне q = 7:

$$S_{общ} = \sum_{j=1}^{p} Q_j - \frac{\left(\sum_{j=1}^{p} T_j\right)^2}{pq} = 314 - 0 = 314;$$

$$S_{факт} = \frac{\sum_{j=1}^{p} T_j^2}{q} - \frac{\left(\sum_{j=1}^{p} T_j\right)^2}{pq} = \frac{470}{7} - 0 \approx 67{,}14.$$

Найдём остаточную сумму квадратов отклонений: $S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}} = 314 - 67{,}14 = 246{,}86.$

Определим количество степеней свободы:

$$k_{\text{факт}} = p - 1 = 4 - 1 = 3, \qquad k_{\text{ост}} = p(q-1) = 4 \cdot (7-1) = 24.$$

Вычислим факторную и остаточную дисперсии:

$$s^2_{\text{факт}} = \frac{S_{\text{факт}}}{p-1} = \frac{67{,}14}{3} \approx 22{,}38;$$

$$s^2_{\text{ост}} = \frac{S_{\text{ост}}}{p(q-1)} = \frac{246{,}86}{24} \approx 10{,}29.$$

Сравним факторную и остаточную дисперсии с помощью критерия Фишера-Снедекора. Для этого сначала определим наблюдаемое значение критерия:

$$F_{\text{набл}} = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}} = \frac{22{,}38}{10{,}29} \approx 2{,}175.$$

Учитывая, что число степеней свободы числителя $k_1 = 3$, а знаменателя - $k_2 = 24$, при уровне значимости $\alpha = 0{,}05$ находим критическую точку: $F_{\text{кр}}(0{,}05; 3; 24) = 3{,}01.$

Поскольку $F_{\text{набл}} < F_{\text{кр}}$, основную гипотезу о равенстве групповых средних не отвергаем. Другими словами, групповые средние различаются незначаще.

Однофакторный дисперсионный анализ для связных выборок

Пусть на количественный нормально распределенный признак X действует фактор F, который имеет p постоянных уровней $F_1, F_2, \ldots, F_p$. На каждом уровне проведены испытания над одной и той же выборкой, которая состоит из q элементов. Результаты наблюдений - числа x_{ij}, где i - номер испытания, $i = 1, 2, \ldots, q$, j - номер уровня фактора F, $j = 1, 2, \ldots, p$, - записывают в виде таблицы

Номер испытания i	Уровень фактора				Индивидуальные средние $\overline{x}_i$
	F_1	F_2	$\ldots$	F_p	
1	x_{11}	x_{12}	$\ldots$	x_{1p}	$\overline{x}_1$
2	x_{21}	x_{22}	$\ldots$	x_{2p}	$\overline{x}_2$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
q	x_{q1}	x_{q2}	$\ldots$	x_{qp}	$\overline{x}_q$
Групповое среднее $\overline{x}_{гр}$	$\overline{x}_{гр.\ 1}$	$\overline{x}_{гр.\ 2}$	$\ldots$	$\overline{x}_{гр.\ p}$	

В этом случае возникают две задачи:

1) при уровне значимости α проверить основную гипотезу о равенстве групповых средних при условии, что групповые генеральные дисперсии хотя и неизвестные, но одинаковые;
2) на уровне значимости α проверить основную гипотезу о равенстве индивидуальных средних при условии, что индивидуальные генеральные дисперсии хотя и неизвестные, но одинаковые.

Чтобы решить эти задачи, находят:

а) *общую сумму* квадратов отклонений наблюдаемых значений признака от общего выборочного среднего:

$$S_{общ} = \sum_{j=1}^{p}\sum_{i=1}^{q}(x_{ij} - \overline{x}_{в})^2$$

б) *факторную сумму* квадратов отклонений групповых средних от общего выборочного среднего (характеризует рассеяние «между группами»):

$$S_{\text{факт}} = q\sum_{j=1}^{p}(\overline{x}_{\text{гр. } j} - \overline{x}_{\text{в}})^2;$$

в) *индивидуальную сумму* квадратов отклонений индивидуальных средних от общего выборочного среднего (характеризует рассеяние «между индивидами»):

$$S_{\text{инд}} = p\sum_{i=1}^{q}(\overline{x}_i - \overline{x}_{\text{в}})^2;$$

г) *остаточную сумму* по формуле $S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}} - S_{\text{инд}}.$

Поделив факторную, индивидуальную и остаточную суммы на соответствующее число степеней свободы $k_{\text{факт}} = p - 1, \quad k_{\text{инд}} = q - 1, \quad k_{\text{ост}} = (p-1)(q-1),$

находят факторную, индивидуальную и остаточную дисперсии:

$$s^2_{\text{факт}} = \frac{S_{\text{факт}}}{p-1}; \qquad s^2_{\text{инд}} = \frac{S_{\text{инд}}}{q-1}; \qquad s^2_{\text{ост}} = \frac{S_{\text{ост}}}{(p-1)(q-1)}.$$

Сравнивают факторную и индивидуальную дисперсии с остаточной дисперсией с помощью *критерия Фишера-Снедекора*: $F^{\text{факт}}_{\text{набл}} = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}}, \qquad F^{\text{инд}}_{\text{набл}} = \frac{s^2_{\text{инд}}}{s^2_{\text{ост}}}.$

Далее по таблице критических точек распределения Фишера-Снедекора при заданном уровне значимости α и числе степеней свободы ($k_{факт}$ - числитель, $k_{инд}$ - числитель, $k_{ост}$ - знаменатель) находят критические точки

$$F_{кр}^{факт}(\alpha; k_{факт}, k_{ост}) \qquad F_{кр}^{инд}(\alpha; k_{инд}, k_{ост}).$$

Если $F_{набл}^{факт} < F_{кр}^{факт}$ различие групповых средних считают незначащим.

Если $F_{набл}^{факт} > F_{кр}^{факт}$ различие групповых средних считают значащим.

Если $F_{набл}^{инд} < F_{кр}^{инд}$ различие индивидуальных средних считают незначащим.

Если $F_{набл}^{инд} > F_{кр}^{инд}$ различие индивидуальных средних считают значащим.

Пример 2. Проведены испытания на каждом из четырёх уровней фактора над одной и той же выборкой из шести элементов. Методом дисперсионного анализа при уровне значимости 0,01 проверить основную гипотезу о равенстве групповых средних и основную гипотезу о равенстве индивидуальных средних. Предполагается, что результирующий признак нормально распределён в исследуемой выборке. Результаты испытания приведены в таблице

Номер испытания i	Уровень фактора				Индивидуальные средние $\overline{x}_i$
	F_1	F_2	F_3	F_4	
1	5	11	7	13	9
2	3	9	8	12	8
3	5	12	5	14	9
4	6	11	6	17	10
5	4	8	5	11	7
6	7	15	5	17	11
Групповое среднее $\overline{x}_{\text{гр}}$	5	11	6	14	

Решение. Найдём общее среднее:

$x_{в}$ = (5 + 11 + 7 + 13 + 3 + 9 + 8 + 12 + 5 + 12 + 5 + 14 + 6 + 11 + 6 + 17 + 4 + 8 + 5 + 11 + 7 + 15 + 5 + 17)/(6 · 4) = 9.

Для вычисления общей суммы квадратов отклонений наблюдаемых значений признаков от общего выборочного среднего составим расчётную таблице, в которую вместо значений x_{ij} запишем значения $y_{ij} = (x_{ij} - \overline{x}_в)^2$.

Номер испытания i	Уровень фактора				$(\overline{x}_i - \overline{x}_в)^2$
	F_1	F_2	F_3	F_4	
1	16	4	4	16	0
2	36	0	1	9	1
3	16	9	16	25	0
4	9	4	9	64	1
5	25	1	16	4	4
6	4	36	16	64	4
$(\overline{x}_{гр} - \overline{x}_в)^2$	16	4	9	25	$S_{общ} = 404$

По таблице определим:

$$S_{факт} = q \sum_{j=1}^{p} (\overline{x}_{гр.\ j} - \overline{x}_в)^2 = 6 \cdot (16 + 4 + 9 + 25) = 324;$$

$$S_{инд} = p \sum_{i=1}^{q} (\overline{x}_i - \overline{x}_в)^2 = 4 \cdot (0 + 1 + 0 + 1 + 4 + 4) = 40.$$

Остаточная сумма квадратов отклонений $S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}} - S_{\text{инд}} = 404 - 324 - 40 = 40.$

Найдём число степеней свободы:

$$k_{\text{факт}} = p - 1 = 4 - 1 = 3,$$
$$k_{\text{инд}} = q - 1 = 6 - 1 = 5,$$
$$k_{\text{ост}} = (p-1)(q-1) = 3 \cdot 5 = 15.$$

Найдём факторную, индивидуальную и остаточную дисперсии:

$$s^2_{\text{факт}} = \frac{S_{\text{факт}}}{p-1} = \frac{324}{3} = 108;$$
$$s^2_{\text{инд}} = \frac{S_{\text{инд}}}{q-1} = \frac{40}{5} = 8;$$
$$s^2_{\text{ост}} = \frac{S_{\text{ост}}}{(p-1)(q-1)} = \frac{40}{15} = \frac{8}{3}.$$

Сравним факторную и индивидуальную дисперсии с остаточной дисперсией по критерию Фишера-Снедекора:

$$F^{\text{факт}}_{\text{набл}} = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}} = \frac{108 \cdot 3}{8} = 40{,}5;$$
$$F^{\text{инд}}_{\text{набл}} = \frac{s^2_{\text{инд}}}{s^2_{\text{ост}}} = \frac{8 \cdot 3}{8} = 3.$$

Далее по таблице критических точек распределения Фишера-Снедекора при заданном уровне значимости $\alpha = 0{,}01$ и числе степеней свободы ($k_{\text{факт}}$ - числитель, $k_{\text{инд}}$ - числитель, $k_{\text{ост}}$ - знаменатель) находим критические точки:

$$F^{\text{факт}}_{\text{кр}}(0{,}01; 3; 15) = 5{,}42;$$
$$F^{\text{инд}}_{\text{кр}}(0{,}01; 5; 15) = 4{,}56.$$

Поскольку $F^{\text{факт}}_{\text{набл}} > F^{\text{факт}}_{\text{кр}}$, основную гипотезу о равенстве групповых средних отвергаем.

Поскольку $F^{\text{инд}}_{\text{набл}} < F^{\text{инд}}_{\text{кр}}$, основную гипотезу о равенстве индивидуальных средних принимаем

Задача 1. Проведено по семь испытаний на каждом из четырёх уровней фактора над разными выборками. Методом дисперсионного анализа при уровне значимости 0,01 проверить основную гипотезу о равенстве групповых средних. Предполагается, что выборки по- лучены из нормальных совокупностей с одинаковыми дисперсиями. Результаты испытания приведены в таблице.

Номер испытания i	Уровень фактора			
	F_1	F_2	F_3	F_4
1	12	11	10	13
2	15	12	13	15
3	16	13	14	17
4	19	15	16	19
5	20	16	18	22
6	21	18	20	23
7	23	20	21	24
Групповое среднее $\overline{x}_{\text{гр}}$	18	15	16	19

Задача 2. Проведены испытания на каждом из четырёх уровней фактора над одной и той же выборкой из шести элементов. Методом дисперсионного анализа при уровне значимости 0,05 проверить основную гипотезу о равенстве групповых средних и основную гипотезу о равенстве индивидуальных средних. Предполагается, что результирующий признак нормально распределён в исследуемой выборке. Результаты испытания приведены в таблице.

Номер испытания i	Уровень фактора				Индивидуальные средние $\overline{x}_i$
	F_1	F_2	F_3	F_4	
1	22	23	22	21	22
2	25	24	22	24	23,75
3	19	20	22	25	21,5
4	14	17	20	24	18,75
5	20	21	21	21	20,75
6	14	15	19	23	17,75
Групповое среднее $\overline{x}_{гр}$	19	20	21	23	

Домашнее задание:

В.Е. Гмурман. РУКОВОДСТВО К РЕШЕНИЮ ЗАДАЧ ПО ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Задачи: 669-673, 675-678