Т е м а 7 . ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Объяснение. Наряду с относительно простыми способами сравнения двух выборок (например, с помощью *t*-критерия) встречаются и более сложные задачи, когда необходимо сравнить несколько выборок, объединенных в единый статистический комплекс. Попарное сравнение в таких случаях неудобно, так как требует длительных и сложных вычислений. Учитывая это, английским ученым Р. А. Фишером был предложен метод комплексной оценки сравниваемых средних, получивший название дисперсионного анализа.

Исторически возникновение дисперсионного анализа связано с экспериментами в сельском хозяйстве по использованию различных доз удобрений. Требовалось доказать существенность «отклика» урожая на точно определенную дозу удобрений в сравнении с контролем. Однофакторный дисперсионный анализ широко используется в различных областях научных исследований и часто является их завершающим этапом. В то же время, он может быть составной частью других важных методов анализа данных. Поэтому целесообразно его подробное рассмотрение.

Например, измеряется влажность почвы в каждом генетическом горизонте. При этом проба с определенной влажностью называется элементом, почвенный горизонт – группой или классом. Если определяется численность какого-либо вида в каждом типе местообитания, то численность это элемент, а тип местообитания – заданный класс. Целью анализа является проверка гипотезы о принадлежности выборок в каждом классе по их математическим ожиданиям к одной общей генеральной совокупности. Анализ строится на основе сопоставления дисперсий выборок, с учетом принадлежности их к классам, с общей дисперсией всей совокупности измерений. Использование при оценивании только одного параметра (дисперсии) требует, чтобы распределения не сильно отличались от нормальных. Следовательно, перед использованием дисперсионного анализа необходима обязательная нормализация выборки, либо, если данные не подчиняются закону нормального распределения, необходимо использовать непараметрические аналоги дисперсионного анализа (например, в Statistica).

Пример 1. С помощью дисперсионного анализа необходимо установить достоверность различий массы проростков томата в

зависимости от температуры проращивания и выяснить степень влияния изучаемого фактора на общую изменчивость результирующего признака (массы проростка). Данные, полученные в ходе эксперимента представлены в табл. 6.

Таблица 6 – Масса проростков томата, мг

Вариант	Повторения, *x*				Число наблюдений *n*	Сумма по вариантам *V*	Среднее по вариантам *X*
	I	II	III	IV			
16°C	18,6	16,5	16,4	19,0	4	70,5	17,625
19°C	20,7	20,0	17,7	19,5	4	77,9	19,475
22°C	26,5	27,7	28,2	27,6	4	110,0	27,5
25°C	26,3	27,6	28,4	28,7	4	111,0	27,75
Суммы по повторениям *P*	92,1	91,8	90,7	94,8	$\sum x = N = 16$	$\sum X = 369,4$	$23,1 = x_0$

Здесь же начинаем вычислять необходимые показатели:

Суммы по вариантам *V* находят сложением данных о массе проростков в четырех повторениях каждого варианта (каждую строку):

18,6+16,5+16,4+19,0 = 70,5 и т.д.

Суммы по повторениям *P* находят сложением данных о массе проростков каждой повторности по всем вариантам (каждую колонку):

18,6+20,7+26,5+26,3 = 92,1 и т.д.

Среднее арифметическое по каждому варианту находят делением сумм по вариантам на число повторений: $\bar{x} = V \div n$; $70,5 \div 4 = 17,625$ и т.д. Правильность расчетов проверяют по равенству: $\sum X = \sum P = \sum V = 369,4$

Для нахождения средней массы проростка по всему опыту (x_0) сумму всех измерений $\sum X$ делят на общее число делянок в опыте *N*:

$x_0 = \sum X \div N = 369,4 \div 16 = 23,1$

Для облегчения дальнейших расчетов исходные данные целесообразно преобразовать по соотношению $X_1 = X - A$, приняв за *A* число 23, близкое к x_0. Преобразованные данные записывают в табл. 7. и рассчитывают суммы по вариантам V_1, суммы по повторениям P_1 и $\sum X_1$.

Правильность расчетов проверяют по равенству: $\sum X_1 = \sum V_1 = \sum P_1 = 1,4$

Таблица 7 – Таблица преобразованных данных

Вариант	$X_1 = X - A = X - 23$				Сумма по вариантам V_1
	I	II	III	IV	
16°C	-4,4	-6,5	-6,6	-4	-21,5
19°C	-2,3	-3	-5,3	-3,5	-14,1
22°C	3,5	4,7	5,2	4,6	18
25°C	3,3	4,6	5,4	5,7	19
Суммы по повторениям P_1	0,1	-0,2	-1,3	2,8	$\sum X_1 = 1,4$

Все данные табл. 7 возводят в квадрат и заполняют табл. 8.

Таблица 8 – Таблица квадратов

Вариант	X_1^2				V_1^2
	I	II	III	IV	
16°C	19,36	42,25	43,56	16,00	462,25
19°C	5,29	9,00	28,09	12,25	198,81
22°C	12,25	22,09	27,04	21,16	324,00
25°C	10,89	21,16	29,16	32,49	361,00
P_1^2	0,01	0,04	1,69	7,84	$1,96 = (\sum X_1)^2$

Для дальнейших расчетов путем сложения находят соответствующие суммы:

$$\sum X_1^2 = (19,36 + 42,25 + \ldots + 32,49) = 352,04$$

$$\sum V_1^2 = (462,25 + 198,81 + 324,00 + 361,00) = 1346,06$$

$$\sum P_1^2 = (0,01 + 0,04 + 1,69 + 7,84) = 9,58$$

Общее число наблюдений в опыте определяется умножением числа вариантов ($l = 4$) на число повторений ($n = 4$): $N = l \times n = 4 \quad 4 = 16$

Затем определяют корректирующий фактор (C):

$$C = (\sum X_1)^2 / N = 1,96 / 16 = 0,12$$

Сумму квадратов отклонений рассчитываем по следующим формулам:

Для общего варьирования

$$C_y = \sum X_1^2 - C = 352,04 - 0,12 = 351,92$$

Для варьирования вариантов

$$C_v = \sum V_1^2 / n - C = 1346,06 / 4 - 0,12 = 336,4$$

Для варьирования повторений

$$C_p = \sum P_1^2 / l - C = 9,58 / 4 - 0,12 = 2,28$$

Для варьирования ошибки (остатка)

$$C_z = C_y - C_v - C_p = 351,92 - 336,40 - 2,28 = 13,24$$

Для выяснения достоверности различий между вариантами данные о суммах квадратов отклонений сводим в табл. 9. Для заполнения этой таблицы необходимо определить число степеней

свободы для общего варьирования, для повторений, для вариантов, для остаточного варьирования (ошибки). Число степеней свободы для общего варьирования определяется общим числом наблюдений минус единицу:

$$N - 1 = 16 - 1 = 15$$

Так же определяют степени свободы для повторений и вариантов:

$$n - 1 = 4 - 1 = 3 \text{ и } l - 1 = 4 - 1 = 3$$

Число степеней свободы ошибки определяют умножением степеней свободы вариантов и повторений: $(n - 1) \times (l - 1) = 3 \times 3 = 9$

Дисперсия для различных видов варьирования находится делением суммы квадратов вида варьирования на число степеней свободы:

Для вариантов $S_v^2 = C_v / (n - 1) = 336{,}4 / 3 = 112{,}13$

Для ошибки $S_z^2 = C_z / ((n - 1) \times (l - 1)) = 13{,}24 / 9 = 1{,}47$

Критерий Фишера находят делением дисперсии вариантов на дисперсию ошибок: $F_{факт.} = S_v^2 / S_z^2 = 112{,}13 / 1{,}47 = 76{,}28$

Критерий F_{05} находят в таблицах приложения для доверительной вероятности 95% (при 5%-ом уровне значимости). По вертикали откладывают число степеней свободы для меньшей дисперсии (ошибки), а по горизонтали – число степеней свободы для большей дисперсии (вариантов). На пересечении находят число, показывающее табличное отношение дисперсий. В нашем случае $F_{05} = 3{,}86$.

Таблица 9 – Результаты дисперсионного анализа

Источник вариации	Сумма квадратов	Степени свободы	Дисперсия S^2	Критерий Фишера	
				$F_{факт.}$	F_{05}
Общая C_y	351,92	15			
Повторений C_p	2,28	3			
Вариантов C_v	336,40	3	112,13	76,28	3,86
Ошибки C_z	13,24	9	1,47		

Сравнивая теоретическое и фактическое значения критерия F, подтверждают или отвергают нулевую гипотезу, которая состоит в предположении об отсутствии различий между вариантами опыта. Если $F_{факт.} \geq F_{теор.}$, то нулевая гипотеза отвергается. Следовательно, между изучаемыми вариантами есть существенные различия. Если $F_{факт.} < F_{теор.}$, нулевая гипотеза подтверждается и между вариантами существенных различий нет. В последнем случае дисперсионный анализ заканчивается и вычисляется только ошибка опыта.

Вывод: Так как в нашем примере $F_{факт.} > F_{теор.}$, то нулевая гипотеза опровергается. Между вариантами опыта есть достоверные различия.

Чтобы оценить существенность частных различий вычисляют:

а) Ошибку опыта:

$$S_{\bar{x}} = \sqrt{(S_z^2/n)} = \sqrt{(1{,}47/4)} = 0{,}61 \text{ мг}$$

б) Ошибку разности средних:

$$S_d = \sqrt{(2 \times S_z^2/n)} = \sqrt{(2 \times 1{,}47/4)} = 0{,}86 \text{ мг}$$

в) Наименьшую существенную разность ($НСР_{05}$) в абсолютных и относительных показателях:

$$НСР_{05} = t_{05} \times S_d = 2{,}26 \times 0{,}86 = 1{,}94 \text{ мг}$$

$$НСР_{05} = (t_{05} \times S_d / x_0) \times 100 = 2{,}26 \times 0{,}86 / 23{,}1 \times 100 = 8{,}4\%$$

Значение критерия t_{05} берут из таблицы (приложение 1) для 9 степеней свободы ошибки. В нашем случае t_{05} = 2,26. Итоги результатов опыта и статистической обработки данных записывают в табл. 10.

Таблица 10 – Результаты анализа

Вариант (температура)	Масса, мг	Отклонение от контроля		Группа
		мг	%	
25°С (контроль)	27,75	-	-	II
16°С	17,625	-10,125	36,5	III
19°С	19,475	-8,275	29,8	III
22°С	27,5	-0,25	0,9	II
$НСР_{05}$		1,94	8,4	

В колонку «масса» переносят средние значения вариантов из табл. 6. Отклонения от контроля находят, отнимая от средней массы по вариантам среднюю массу проростка в контроле. Затем сравнивают полученные значения с НСР. В том случае, если величина отклонения от контроля имеет положительный знак и по модулю больше величины НСР, делают вывод о том, что данный вариант существенно превышает контроль, и его относят к I группе. Если отклонение от контроля имеет отрицательный знак, но по модулю больше НСР, значит данный вариант существенно хуже контроля, и его относят к третьей группе.

Варианты, у которых отклонение от контроля по модулю меньше НСР, несущественно отличаются от контроля, и они относятся ко второй группе.

По результатам анализа делают вывод: Масса проростков томата, полученных в контроле при 25°С и в варианте опыта при 22°С, существенно не отличается между собой – варианты

относятся во II группу. Проращивание же семян при более низкой температуре (16°С и 19°С) позволяет получить проростки с достоверно меньшей массой – варианты относятся в III группу.

Дисперсионный анализ дает возможность получить представление о степени влияния того или иного фактора в общей дисперсии (изменчивости) признака. Это можно определить по следующим формулам:

а) для влияния вариантов

$\eta_v^2 = C_v / C_y$ =336,4 / 351,92 = 0,956 или 95,6%

б) для влияния повторений

$\eta_p^2 = C_p / C_y$ =2,28 / 351,92 = 0,006 или 0,6%

в) для влияния случайных факторов

$\eta_z^2 = C_z / C_y$ =13,24 / 351,92 = 0,038 или 3,8%

г) для влияния всех факторов

$\eta_y^2 = \eta_v^2 + \eta_p^2 + \eta_z^2$ = 1,0 (100%)

далее следует выполнить индивидуальное задание (приложение 8).

Пример 2. Для выяснения роли обработки инсектицидом были выделены два участка с яблонями, причём только первый был обработан, а второй оставлен контрольным. Количество яблок, собранных с каждой яблони приведено в таблице 11. Можно ли считать различия достоверными, т.е. можно ли считать, что применение инсектицида увеличивает урожай? Статистическую обработку данных проведите при помощи программы Microsoft Excel.

Таблица 11 – Количество яблок, собранных с одной яблони, кг

Повторности	Обработанные деревья	Необработанные деревья
1	210	174
2	240	162
3	197	154
4	205	173
5	183	148
6	191	157
7	197	150
8	201	163
9	193	170
10	203	169

В состав Microsoft Excel входит набор средств анализа данных (так называемый пакет анализа), предназначенный для решения сложных статистических и инженерных задач. Для проведения

анализа данных с помощью этих инструментов следует указать входные данные и выбрать параметры; анализ будет проведен с помощью подходящей статистической или инженерной макрофункции, а результат будет помещен в выходной диапазон. Другие средства позволяют представить результаты анализа в графическом виде. Чтобы просмотреть список доступных инструментов анализа, выберите команду Анализ данных в меню Сервис. Если команда Анализ данных в меню Сервис отсутствует — необходима установка пакета анализа. Пакет анализа включает в себя три средства дисперсионного анализа. Выбор конкретного инструмента определяется числом факторов и числом выборок в исследуемой совокупности данных.

Решение задачи начинается с создания исходной таблицы на рабочем листе Excel, после чего в меню Сервис выбирается пункт Анализ данных, а в открывшемся списке функций выделяется курсором имя Однофакторный дисперсионный анализ. В открывшемся списке функций указывается диапазон ячеек, содержащий входные данные, для чего необходимо протащить указатель мыши при нажатой левой кнопке по всем ячейкам таблицы с данными на рабочем листе, а также задаётся выходной интервал. Указав группировку данных по столбцам и уровень значимости α = 0,05, щёлкаем по кнопке ОК, после чего на рабочем листе, начиная с указанной ячейки, выводятся результаты вычислений, приведённые ниже.

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
1	9	54	6	7,5
210	9	1810	201,1	257,6
174	9	1446	160,7	81

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	190877,6	2	95438,8	827,24	7,3E-23	3,40
Внутри групп	2768,9	24	115,4			
Итого	193646,5	26				

Рисунок 2 – Результаты дисперсионного анализа в Microsoft Excel

Как видно из полученных данных, фактическая величина *F*-критерия значительно больше табличного (критического) значения,

что свидетельствует о существенных различиях между групповыми средними, которые не могут быть результатами случайного варьирования.

Задания для самостоятельного выполнения:

1. В пресноводном озере на разных глубинах отобрали одинаковые пробы грунта для изучения количества пиявок. Определить, влияет ли глубина взятия пробы на количество пиявок.

Глубина, см	Повторности		
	1	2	3
0–10	20	19	21
10–20	15	14	13
20–30	10	11	12
30–40	5	6	5
40–50	1	2	1

2. Изучали вес телят (кг) одного возраста на разных фермах. Достоверны ли различия по этому показателю на разных фермах?

Номер фермы	Повторности		
	1	2	3
1	180	190	187
2	168	173	171
3	220	215	210
4	202	200	195

3. В пробах почвы подсчитали количество семян сорняков (шт.). Достоверно ли различие между разными горизонтами по числу семян?

Глубина, см	Повторности			
	1	2	3	4
0 – 20	839	695	744	812
20 – 40	238	292	254	303
40 – 60	89	104	95	91

4. Подсчитали число крольчат в помете у крольчих разных пород. Достоверно ли зависит число крольчат в помете от породы?

Порода	Повторности			
	1	2	3	4
1	12	11	9	10
2	7	6	7	8
3	5	7	5	6

5. Изучали число микроорганизмов в разных пробах воды (шт./м^3). Достоверны ли различия по изучаемому показателю для разных водоемов?

Водоем	Повторности

	1	2	3	4
Оршанское озеро	138	117	123	110
Нижнее озеро	98	87	92	85
Полящицы	34	42	54	40

6. Изучали зависимость урожайности томата от вида удобрений. Достоверны ли различия по урожайности томатов (ц/га) на разных участках?

Вариант	Повторности		
	1	2	3
Естественное плодородие	230	250	240
Органические удобрения	390	400	388
Минеральные удобрения	377	380	368
Органика + минеральные удобрения	580	610	625

7. В опыте по определению концентрации нитратов получены данные, приведенные в таблице. Достоверно ли отличается содержание нитратов (мг/кг продукции) в разных овощах?

Вид продукции	Повторности		
	1	2	3
Картофель	86	92	101
Морковь	93	95	86
Огурцы	52	31	59
Корневой сельдерей	176	140	153

8. Достоверно ли отличается содержание селена (мкг/100 г) в грибах в различных районах Могилевской области?

Район	Повторности			
	1	2	3	4
Шкловский	15,8	15,3	13,1	11,3
Горецкий	9,2	8,3	9,7	9,3
Мстиславский	4,1	5,7	3,9	4,2