СТАТИСТИЧЕСКИЙ АНАЛИЗ ВЛИЯНИЯ ФАКТОРОВ НА ИССЛЕДУЕМУЮ ВЕЛИЧИНУ: ДИСПЕРСИОННЫЙ АНАЛИЗ

1 Однофакторный дисперсионный анализ

1.1 Постановка и алгоритм решения задачи однофакторного дисперсионного анализа

Пример 1 – Имеются данные о средней заработной плате на предприятиях трех отраслей промышленности (пять предприятий отрасли П1, шесть – отрасли П2, пять – отрасли П3):

- отрасль П1: 640, 637, 648, 624, 639;
- порода П2: 638, 625, 662, 645, 640, 637;
- порода П3: 625, 617, 631, 629, 619.

Требуется выяснить, имеются ли значимые различия в заработной плате между отраслями.

В данном случае анализируется один *фактор* – отрасль, т.е. это задача однофакторного дисперсионного анализа. Требуется выяснить, зависит ли зарплата от отрасли, т.е. определить, значимо ли различие между зарплатами в разных отраслях. В задаче исследуются три *группы* – отрасли П1, П2, П3. Обозначим это как m=3. Объемы выборок в группах и общий объем выборки следующее: n_1=5, n_2=6, n_3=5, n=16.

Конкретные значения исследуемой величины (зарплаты) будем обозначать как x_{ik}, i=1,…,n_k, k=1,…,m, где k – номер группы (отрасли), i – номер элемента данных в группе.

Для дальнейших расчетов потребуются средние значения по каждой группе, а также среднее значение по всем группам. Для данного примера $\overline{X}_1$=637,60, $\overline{X}_2$=641,17, $\overline{X}_3$=624,20, $\overline{X}$=634,75.

Метод дисперсионного анализа основан на сопоставлении *межгруппового и внутригруппового разброса*. Упрощенно говоря, определяется следующее: можно ли утверждать, что различие *между группами* значительно *больше*, чем различие значений исследуемой величины *внутри групп* (или, другими словами, можно ли утверждать, что значения исследуемой величины *из разных групп* отличаются друг от друга *намного больше*, чем значения *в одной группе*). Если это так, то группы отличаются значимо; значит, влияние фактора, по которому различаются группы, существенно.

Для данного примера определяется следующее: если зарплаты в разных отраслях отличаются друг от друга *значительно больше*, чем зарплаты в пределах одной отрасли, значит, различие между отраслями значимо, т.е. отрасль существенно влияет на зарплату.

Алгоритм решения задачи следующий.

Определяется показатель различия между группами – межгрупповой разброс (называемый также *дисперсией вариантов*):

$$S_v^2 = \frac{\sum_{k=1}^{m} n_k \cdot (\overline{X}_k - \overline{X})^2}{m-1}.$$

Для данного примера S_v^2=(5·(637,60 – 634,75)^2 + 6·(641,17 – 634,75)^2 + + 5·(624,20 – 634,75)^2)/(3–1) = 422,08.

Определяется показатель различия между отдельными значениями внутри групп – внутригрупповой разброс (называемый также *дисперсией ошибки*):

$$S^2 = \frac{\sum_{k=1}^{m} \sum_{i=1}^{n_k} (x_{ik} - \overline{X}_k)^2}{n-m}.$$

Указание – Здесь из *каждого значения* вычитается среднее *по его группе*, разности возводятся в квадрат, квадраты суммируются.

Для данного примера S^2 =((640–637,6)^2+(637–637,6)^2+…+(619–624,2)^2)/ / (16–3) = 91,45.

Для оценки статистической значимости различия между группами используется *F*-критерий Фишера:

$$F = \frac{S_v^2}{S^2}.$$

Для данного примера *F* = 422,08/91,45 = 4,62.

По значению *F*-критерия и числу степеней свободы *m*–1 и *n–m* из таблиц распределения Фишера (или с помощью программных средств) определяется расчетный уровень значимости *P*, который представляет собой (упрощенно говоря) вероятность того, что различие между группами *статистически незначимо*.

В MS Excel для определения *P* используется функция **F.РАСП.ПХ** со следующими параметрами: **X**: *F*; **Степени свободы 1**: *m*-1; **Степени свободы 2:** *n-m* (для данного примера *m*-1 = 2, *n-m* = 13). Получим *P*=0,0306.

Если $P < \alpha$ (обычно используют α=0,05), то различие между группами *статистически значимо*, т.е. влияние анализируемого фактора значимо.

В данном примере $P < \alpha$. Это означает, что группы различаются значимо (или, более точно, разброс средних значений групп существенно больше, чем разброс значений внутри групп). Существенное различие между группами означает, что фактор, по которому эти группы различаются (в данном примере – отрасль), значим.

1.2 Выявление групп, различающихся значимо

Если различие между средними значениями групп оказалось статистически значимым, то возникает следующая задача: выяснить, *какие именно* группы различаются статистически значимо. Для этого выполняется анализ на основе наименьшей существенной разности (НСР).

Основная идея метода анализа на основе НСР следующая. Для *каждой пары групп* вычисляется разность средних значений. Если эта разность превышает некоторую предельную величину – НСР, то различие между группами признается статистически значимым. Значение НСР вычисляется для каждой пары групп согласно методу, рассматриваемому ниже.

Выполним анализ на основе НСР для рассматриваемого примера. Здесь требуется выяснить, какие именно отрасли значимо отличаются друг от друга по зарплате. Задача решается в следующем порядке.

1 Для каждой пары групп вычисляется вспомогательная величина – статистическая ошибка разности средних:

$$S_{dij} = \sqrt{S^2 \cdot \frac{n_i + n_j}{n_i \cdot n_j}},$$

где i, j – номера сравниваемых групп;

n_i, n_j – количество данных в сравниваемых группах;

S^2 – внутригрупповой разброс (дисперсия ошибки), найденный выше.

Вычислим статистическую ошибку разности средних, например, для первой и второй группы. В данном примере $S^2 = 91{,}45$ (см. выше), $n_1=5$, $n_2=6$:

$$S_{d12} = \sqrt{91{,}45 \cdot \frac{5+6}{5 \cdot 6}} = 5{,}79.$$

Аналогично вычисляются статистические ошибки разности средних для других пар групп: $S_{d13} = 6{,}05$, $S_{d23} = 5{,}79$.

2 Для каждой пары групп вычисляется НСР:

$$\text{НСР}_{ij} = t_{\alpha;n-m} \cdot S_{dij},$$

где $t_{\alpha;n-m}$ – квантиль распределения Стьюдента для заданного уровня значимости α и числа степеней свободы $n–m$. В MS Excel для определения $t_{\alpha;n-m}$ используется функция **СТЬЮДЕНТ.ОБР.2Х** с параметрами **Вероятность:** 0,05, **Степени свободы**: 13. В данном примере $t_{\alpha;n-m}=t_{0.05;13} = 2{,}1604$.

Вычислим НСР:

$\text{НСР}_{12} = 2{,}1604 \cdot 5{,}79 = 12{,}51$;

$\text{НСР}_{13} = 2{,}1604 \cdot 6{,}05 = 13{,}07$;

$\text{НСР}_{23} = 2{,}1604 \cdot 5{,}79 = 12{,}51$.

3 Для каждой пары групп вычисляется разность средних значений. Если эта разность (по абсолютной величине) превышает НСР, то средние значения в группах различаются значимо. Результаты рассматриваемого примера приведены ниже:

Группы	Разность средних	НСР	Различие
П1 – П2	3,57	12,51	незначимо
П1 – П3	13,40	13,07	значимо
П2 – П3	16,97	12,51	значимо

Из полученных результатов видно, что по средним значениям зарплат значимо различаются отрасли П1 и П3, а также П2 и П3. Различие между отраслями П1 и П2 незначимо. Таким образом, можно предположить, что зарплаты в отрасли П3 значимо отличаются от зарплат в других отраслях. При этом средняя зарплата в отрасли П3 меньше, чем в других отраслях (средние значения по группам вычислены выше). Таким образом, можно сделать вывод, что зарплаты в отраслях П1 и П2 значимо выше, чем в отрасли П3.

1.3 Решение задачи в MS Excel

Исходные данные и результаты решения задачи однофакторного дисперсионного анализа в MS Excel показаны на рисунке 1.1.

1 Перейти на свободный рабочий лист. В ячейках A1, B1, C1 ввести заголовки «П1», «П2», «П3». В ячейках A2:A6, B2:B7 и C2:C6 ввести значения зарплат.

2 Из меню **Данные – Анализ данных** выбрать инструмент **Однофакторный дисперсионный анализ**. Указать **Входной интервал:** A1:C7, **Альфа**: 0,05. Установить переключатель **Группирование: по столбцам**. Установить флажок **Метки**. В области **Параметры вывода** выбрать место для вывода результатов (в данном примере установлен переключатель **Выходной интервал** и указана ячейка A10). Для получения результатов нажать **OK**.

	A	B	C	D	E	F	G
1	П1	П2	П3				
2	640	638	625				
3	637	625	617				
4	648	662	631				
5	624	645	629				
6	639	640	619				
7		637					
8							
9							
10	Однофакторный дисперсионный анализ						
11							
12	ИТОГИ						
13	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
14	П1	5	3188	637,6	75,3		
15	П2	6	3847	641,1667	147,766667		
16	П3	5	3121	624,2	37,2		
17							
18							
19	Дисперсионный анализ						
20	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
21	Между группами	844,1667	2	422,0833	4,61551942	0,030576894	3,805567417
22	Внутри групп	1188,833	13	91,44872			
23							
24	Итого	2033	15				

Рисунок 1.1 – Решение задачи однофакторного дисперсионного анализа

Расчетный уровень значимости P, используемый для оценки значимости различия между группами, выводится как **P-значение**. Выводятся также некоторые другие величины, в частности, средние по группам (**Среднее**), межгрупповой разброс S_v^2 (**MS между группами**), внутригрупповой разброс S^2 (**MS внутри групп**), F-критерий Фишера (**F**) и некоторые другие.

Средств для анализа на основе НСР в MS Excel нет, и расчет требуется выполнять вручную. Для расчета удобно использовать результаты решения задачи с использованием инструмента **Однофакторный дисперсионный анализ**. (средние по группам, внутригрупповой разброс).

Задание 1 – Имеются показатели прочности образцов четырех материалов:
- материал М1: 38; 35; 41; 32; 28; 36; 35;
- материал М2: 29; 31; 34; 25; 28; 25;
- материал М3: 37; 42; 46; 38; 35; 32; 39;
- материал М4: 32; 27; 35; 29; 26.

Определить, имеется ли статистически значимое различие между материалами по прочности. Если оно имеется, определить пары материалов, которые различаются статистически значимо. Определить наиболее и наименее прочный материал.

Решить задачу: а) вручную, используя MS Excel для вычислений; б) в MS Excel, используя инструмент **Однофакторный дисперсионный анализ**.

2 Многофакторный (двухфакторный) дисперсионный анализ

2.1 Решение задачи двухфакторного дисперсионного анализа в MS Excel

Пример 2 – Имеются данные о средней заработной плате на 24 предприятиях различного размера и формы собственности:
- государственные крупные: 180, 200, 215, 260;
- государственные средние: 215, 180, 130, 200;
- государственные мелкие: 200, 160, 130, 170;
- негосударственные крупные: 180, 190, 180, 250;
- негосударственные средние: 240, 230, 190, 220;
- негосударственные мелкие: 240, 260, 200, 250.

Требуется выяснить, какие факторы (или их комбинации) оказывают значимое влияние на заработную плату.

В данном примере исследуемая величина – зарплата. Анализируются два *фактора*: форма собственности и размер. Для каждой комбинации факторов имеются четыре значения исследуемой величины (т.е. четыре предприятия). Поэтому для анализа используются методы *многофакторного* (*двухфакторного*) *дисперсионного анализа с повторениями*.

Фактор «форма собственности» имеет два уровня (государственные и негосударственные): a=2. Фактор «размер» имеет три уровня (крупные, средние, мелкие): b=3. Имеется *шесть* групп данных. Размер каждой группы n_g=4. Полный размер выборки n=24.

Для решения этой задачи требуется большой объем расчетов, поэтому в данном пособии они не рассматриваются. Рассмотрим решение задачи с использованием MS Excel. Вид исходных данных показан на рисунке 2.1.

	A	B	C	D
1		Крупные	Средние	Мелкие
2	Государственные	180	215	200
3		200	180	160
4		215	130	130
5		260	200	170
6	Негосударственные	180	240	240
7		190	230	260
8		180	190	200
9		250	220	250

Рисунок 2.1 – Данные для двухфакторного дисперсионного анализа с повторениями в MS Excel

Примечание – Задачи двухфакторного дисперсионного анализа с повторениями могут решаться в MS Excel только при условии, что размеры всех групп одинаковы. В данном примере это так: размер всех групп n_g=4. Если размеры групп разные, то требуется использовать специализированные программные средства, например, Statistica.

Для решения задачи выбрать инструмент **Двухфакторный дисперсионный анализ с повторениями**. Указать параметры: **Входной интервал**: A1:D9; флажок **Метки** – установить; в поле **Число строк для выборки** ввести 4 (так как данные для каждой группы введены в четыре строки). В области **Параметры вывода** указать, куда требуется вывести результаты. Для получения результатов нажать **OK**. Результаты будет иметь примерно такой вид, как показано на рисунке 2.2.

	A	B	C	D	E	F	G
12	Двухфакторный дисперсионный анализ с повторениями						
13							
14	ИТОГИ	Крупные	Средние	Мелкие	Итого		
15	*Государственные*						
16	Счет	4	4	4	12		
17	Сумма	855	725	660	2240		
18	Среднее	213,75	181,25	165	186,6667		
19	Дисперсия	1156,25	1372,917	833,3333	1365,152		
20							
21	*Негосударственные*						
22	Счет	4	4	4	12		
23	Сумма	800	880	950	2630		
24	Среднее	200	220	237,5	219,1667		
25	Дисперсия	1133,333	466,6667	691,6667	881,0606		
26							
27	*Итого*						
28	Счет	8	8	8			
29	Сумма	1655	1605	1610			
30	Среднее	206,875	200,625	201,25			
31	Дисперсия	1035,268	1217,411	2155,357			
32							
33							
34	Дисперсионный анализ						
35	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
36	Выборка	6337,5	1	6337,5	6,725129	0,01835659	4,413863053
37	Столбцы	189,5833	2	94,79167	0,10059	0,90480886	3,55456109
38	Взаимодействие	7556,25	2	3778,125	4,009211	0,03630144	3,55456109
39	Внутри	16962,5	18	942,3611			
40							
41	Итого	31045,83	23				

Рисунок 2.2 – Результаты двухфакторного дисперсионного анализа с повторениями в MS Excel

2.2 Анализ результатов

Расчетные уровни значимости (*P*) выводятся в колонке **P value**.

Для фактора, обозначенного как **Выборка** (в данном примере – форма собственности), *P*=0,018 < 0,05. Это означает, что форма собственности – значимый фактор, т.е. *зарплата существенно отличается для разных форм собственности*. Средняя зарплата для государственных и негосударственных предприятий составляет 186,67 и 219,17 ден.ед. соответственно, т.е. зарплата на негосударственных предприятиях существенно выше, чем на негосударственных.

Для фактора, обозначенного как **Столбцы** (в данном примере – размер предприятия), *P*=0,9 >0,05. Это означает, что размер предприятия – незначимый фактор, т.е. *на предприятиях, разных по размеру, зарплата отличается несущественно*. Средняя зарплата для крупных, средних и мелких предприятий составляет 206,88, 200,63 и 201,25 ден.ед. соответственно, т.е. они достаточно близки.

Кроме того, для фактора **Взаимодействие** *P*=0,036 <0,05. Это означает, что значимым является *взаимодействие факторов*. Другими словами, фактор (например, форма собственности) может влиять на исследуемую величину (зарплату) *по-разному, в зависимости от значения другого фактора* (например, размера предприятия). Чтобы проанализировать эти взаимодействия, составим таблицу средних значений по группам:

	Крупные	Средние	Мелкие
Государственные	213,75	181,25	165
Негосударственные	200	220	237,5

Построим графики (тип - **График**), отражающие эти средние значения (см. рисунок 2.3).

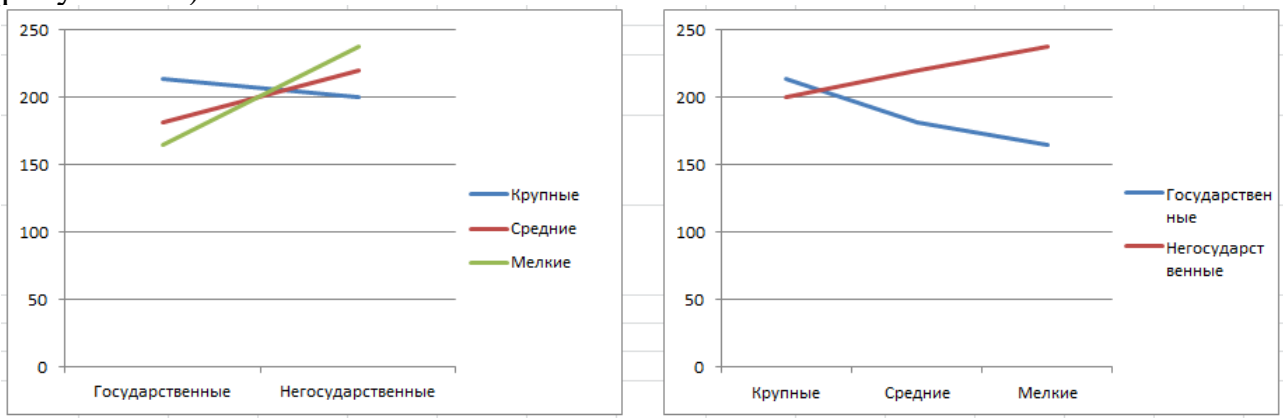

Рисунок 2.3 – Двухфакторный дисперсионный анализ с повторениями: средние по группам

Из первого графика видно, что зарплаты на негосударственных предприятиях выше, чем на государственных, для средних и мелких предприятий; для крупных предприятий, наоборот, зарплата на государственных предприятиях выше, чем на негосударственных. Из второго графика видно, что для государственных предприятий зарплата тем выше, чем больше размер предприятия; для негосударственных предприятий наблюдается обратное.

2.2 Выявление групп, различающихся значимо

Чтобы выявить группы (т.е. комбинации формы собственности и размера), существенно различающиеся по зарплате, воспользуемся методом HCP.

1 Для *каждой пары групп* вычислить *статистическую ошибку разности средних значений*:

$$S_{dij} = \sqrt{S_e^2 \cdot \frac{n_i + n_j}{n_i \cdot n_j}}.$$

где S_e^2 – дисперсия ошибки (или дисперсия внутри групп);

n_i и n_j – размеры групп.

Значение S_e^2 имеется в результатах дисперсионного анализа (см. рисунок 2.2) и обозначено как **MS** – **внутри**. В данном примере $S_e^2 = 942{,}36$.

В данном примере размеры всех групп одинаковы: n_g=4. Поэтому значение S_{dij} также одинаково для *всех пар групп* (будем обозначать его как S_d):

$$S_d = \sqrt{942{,}36 \cdot \frac{4+4}{4 \cdot 4}} = 21{,}71.$$

2 Для *каждой пары групп* вычислить HCP:

$НСР_{ij} = t_{\alpha;k} \cdot S_{dij}$,

где $t_{\alpha;k}$ – квантиль распределения Стьюдента;

k=(n-1) - (a-1) - (b-1) - (a-1)·(b-1) – число степеней свободы;

S_{dij} – статистическая ошибка разности средних.

Для данного примера значение S_{dij} одинаково для всех пар групп, поэтому НСР также одинаковы:

$НСР = t_{\alpha;k} \cdot S_d$.

В данном примере n=24, a=2, b=3, поэтому число степеней свободы составит k=(24-1) - (2-1) - (3-1) - (2-1)·(3-1) = 18 (оно также содержится в результатах дисперсионного анализа как **df – Внутри**).

Используя функцию **СТЬЮДЕНТ.ОБР.2Х** (где **Вероятность** = 0,05, **Df** = 18), найдем $t_{\alpha;k} = t_{0,05;18}$ = 2,1009.

Таким образом, НСР = 2,1009·21,71 = 45,6.

3 Для *каждой пары групп* вычислить абсолютную величину разности средних. Если она превышает НСР, то соответствующие группы различаются значимо. Средние значения по группам и их разности приведены в таблице ниже. Значимые разности (превышающие НСР) выделены жирным шрифтом.

Факторы		Гос., крупные	Гос., средние	Гос., мелкие	Негос., крупные	Негос., средние	Негос., мелкие
		213,75	181,25	165	200	220	237,5
Гос., крупные	213,75	–	32,5	**48,75**	13,75	6,25	23,75
Гос., средние	181,25		–	16,25	18,75	38,75	**56,25**
Гос., мелкие	165			–	35	**55**	**72,5**
Негос., крупные	200				–	20	37,5
Негос., средние	220					–	17,5
Негос., мелкие	237,5						–

Таким образом, существенно различаются зарплаты на государственных крупных от государственных мелких предприятий (на крупных – значительно выше). Кроме того, значимо отличаются зарплаты на негосударственных мелких предприятиях от государственных средних и мелких (на негосударственных – значительно выше).

Задание 1 – Известны величины средней зарплаты на 24 предприятиях двух отраслей промышленности (металлургическая и химическая), расположенных в четырех регионах:

		Регион			
		Север	Юг	Восток	Запад
Отрасль	Металлургическая	420 370 520	370 480 390	480 550 520	480 390 510
	Химическая	620 540 580	620 540 580	390 480 490	490 570 520

Выполнить анализ этих данных методами дисперсионного анализа, как показано выше. Определить, зависит ли зарплата от отрасли промышленности и региона. Если различие значимо, определить отрасли промышленности и регионы, для которых зарплата различается значимо.

Решить задачу в MS Excel.