Министерство науки и высшего образования Российской Федерации
Севастопольский государственный университет
Кафедра ИС

Отчет
по лабораторной работе №3
«Задача дисперсионного анализа. Методы дисперсионного анализа. Однофакторный дисперсионный анализ»
по дисциплине
«ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ»

Выполнил студент группы ИС/б-17-2-о
Горбенко К. Н.
Проверил
Сырых О.А.

Севастополь
2020

1 ЦЕЛЬ РАБОТЫ

• приобрести практические навыки в проведении дисперсионного анализа по экспериментальным данным;

• исследовать возможности языка R для проведения дисперсионного анализа.

2 ЗАДАНИЕ НА РАБОТУ

1. Создать набор данных согласно варианту;
2. Провести однофакторный дисперсионный анализ в среде `Rcmdr`.
3. По результатам дисперсионного анализа сформулировать выводы.
4. Построить диаграмму, отображающую средние значения и их доверительные интервалы для каждой группы.

Задача: в процессе исследования влияния цены за единицу продукции на объем продаж (шт.) в месяц были получены следующие результаты:

Таблица 1 – Данные по варианту

Номер наблюдения	Цена за единицу продукции			
	1000-1100	1100-1200	1200-1300	1300-1500
1	215	218	214	211
2	221	214	217	210
3	222	220	210	208
4	219	221		209
5		213		

3 ДИСПЕРСИОННЫЙ АНАЛИЗ ДАННЫХ ПО ВАРИАНТУ

3.1 Создание набора данных

Создадим .csv файл следующего содержания:

```
num,sales1000-1100,sales1100-1200,sales1200-1300,sales1300-1500
1,215,218,214,211
2,221,214,217,210
3,222,220,210,208
4,219,221,,209
5,,213,,
```

3.2 Дисперсионный анализ в MS Excel

Откроем полученный .csv файл в MS Excel, с помощью пакета «Пакет анализа» проведем однофакторный дисперсионный анализ данных. Результат анализа приведен на рисунке 1.

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
sales1000-1100	4	877	219.25	9.583333333
sales1100-1200	5	1086	217.2	12.7
sales1200-1300	3	641	213.666667	12.33333333
sales1300-1500	4	838	209.5	1.666666667

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	222.5333333	3	74.1777778	8.150160232	0.003161993	3.490294819
Внутри групп	109.2166667	12	9.10138889			
Итого	331.75	15				

Рисунок 1 – Результат дисперсионного анализа в Excel

По среднему значению количества продаж в зависимости от цены товаров отличаются незначительно: лучший результат при цене 1000-1100, худший - при цене 1300-1500. При этом сохраняется тенденция к уменьшению количества продаж при увеличении цены.

Для цены 1300-1500 количества продаж по дням практически не отличаются (208, 209, 210, 211), соответственно дисперсия принимает малые значения.

Т.к $F > F_{\text{крит}}$, значит отвергнута нулевая гипотеза и принята первая гипотеза с вероятностью ошибки $\alpha = 0,05$ можно утверждать, что влияние фактора (цены) на результирующий признак (количество продаж) существенно.

3.3 Дисперсионный анализ средствами языка R

Выполним дисперсионный анализ данных по варианту с помощью средств языка R. Результат изображен на рисунке 2.

```
Rcmdr>  AnovaModel.1 <- aov(X215 ~ sales1000.1100, data=Dataset)

Rcmdr>  summary(AnovaModel.1)
               Df Sum Sq Mean Sq F value  Pr(>F)
sales1000.1100  3 246.60   82.20   10.62 0.00141 **
Residuals      11  85.13    7.74
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Rcmdr>  with(Dataset, numSummary(X215, groups=sales1000.1100, statistics=c("mean",
Rcmdr+    "sd")))
                     mean       sd data:n
sales1000-1100 220.6667 1.527525      3
sales1100-1200 217.2000 3.563706      5
sales1200-1300 213.6667 3.511885      3
sales1300-1500 209.5000 1.290994      4
```

Рисунок 2 – Результат дисперсионного анализа средствами языка R

Диаграмма средних значений и их доверительных интервалов изображена на рисунке 3.

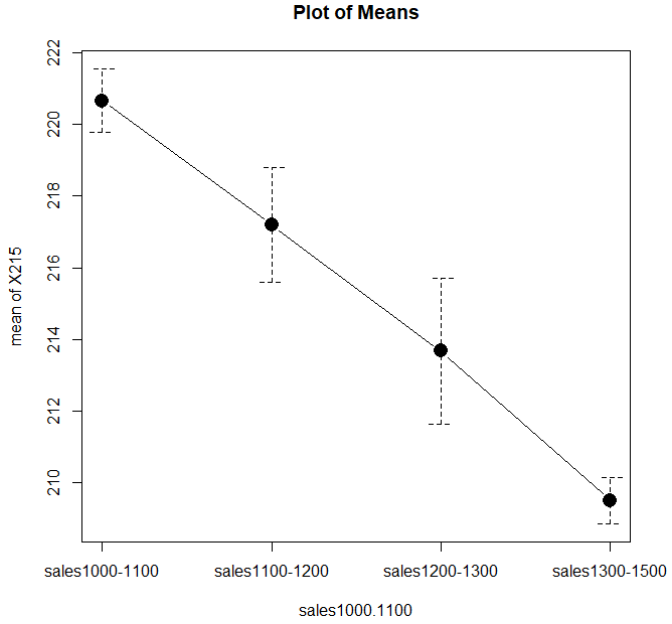

Рисунок 3 – Диаграмма средних значений и их доверительных интревалов для данных по варианту

Сравнение средних значений показало, что количество продаж наивысшее у товаров с ценой 1000-1100, наинизшее у товаров 1300-1500. Т.к. $Pr(> F)$ составляет 0.14%, мы отвергаем нулевую гипотезу. Значение F-критерия составило 10.62.

4 ДИСПЕРСИОННЫЙ АНАЛИЗ СВОИХ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ В MS EXCEL

4.1 Создание набора данных

Для анализа выберем зависимость количества заражений от индекса строгости. Преобразуем экспериментальные данные к следующему виду:

0-75	75-85	85-90	90-100
8250	3442	16003	7145
8989	2648	7945	8243
7791	2462	4816	3963
1758	13328	12698	3828
902	6681	6454	
2923	6459	1711	
4670	2538	1655	
6149	9942	8309	
	2343	7265	
	4654	5152	
	1701	4693	
		6072	
		2592	

Рисунок 4 – Преобразованные данные

4.2 Дисперсионный анализ в Excel

Выполним дисперсионный анализ в MS Excel. Результат выполнения представлен на рисунке 5.

Однофакторный дисперсионный анализ						
ИТОГИ						
Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
0-75	8	41432	5179	9579310.286		
75-85	11	56198	5108.909091	13712783.89		
85-90	13	85365	6566.538462	17032908.94		
90-100	4	23179	5794.75	5013505.583		
Дисперсионный анализ						
Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	15785915	3	5261971.666	0.397487643	0.755695399	2.901119584
Внутри групп	423618434.9	32	13238076.09			
Итого	439404349.9	35				

Рисунок 5 – Результат дисперсионного анализа в MS Excel

Для интервала индекса строгости 85-90 наблюдается наибольшее среднее значение количества заражений. При этом не наблюдается тенденции к уменьшению количества заражений при увеличении индекса строгости. $F < F_{\text{крит}}$, нулевую гипотезу отвергнуть нельзя, связь между переменными не доказана.

4.3 Дисперсионный анализ средствами языка R

Выполним дисперсионный анализ данных по варианту с помощью средств языка R. Результат изображен на рисунке 6.

```
Rcmdr>  AnovaModel.2 <- aov(ï..total_cases ~ stringency, data=Dataset)

Rcmdr>  summary(AnovaModel.2)
            Df    Sum Sq  Mean Sq F value Pr(>F)
stringency   3  15785915  5261972   0.397  0.756
Residuals   32 423618435 13238076

Rcmdr>  with(Dataset, numSummary(ï..total_cases, groups=stringency,
Rcmdr+    statistics=c("mean", "sd")))
           mean       sd data:n
0-75   5179.000 3095.046      8
75-85  5108.909 3703.078     11
85-90  6566.538 4127.094     13
90-100 5794.750 2239.086      4
```

Рисунок 6 – Результат дисперсионного анализа средствами языка R

Диаграмма средних значений и их доверительных интервалов для своих данных приведена на рисунке 7.

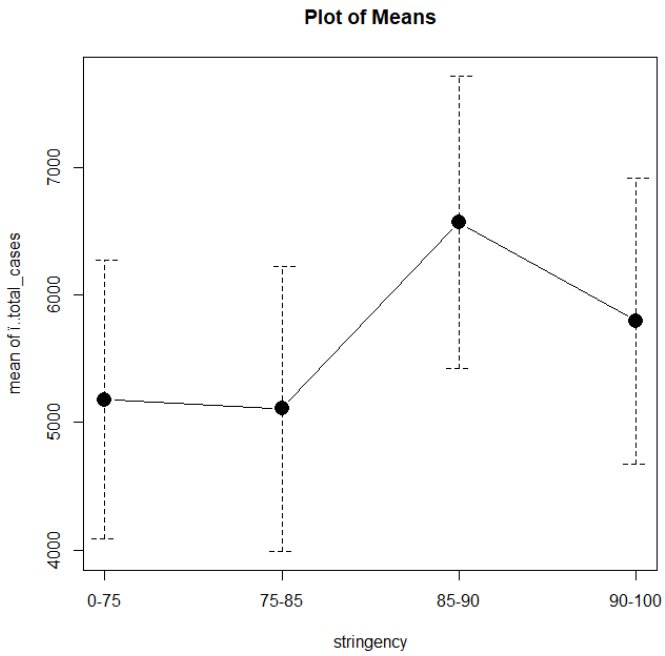

Рисунок 7 – Диаграмма средних значений и их доверительных интервалов для своих данных

Значение отношения межгрупповой изменчивости к внутригрупповой равно 0.397, то есть мы отклонились от левой части распределения Фишера. Уровень значимости говорит о том, что получить такое распределение мы можем с вероятностью 75%. Значит, влияние рассматриваемого фактора на результативный признак несущественно.

ВЫВОДЫ

В ходе выполнения лабораторной работы были приобретены практические навыки в проведении дисперсионного анализа по экспериментальным данным. Также был получен опыт анализа данных с помощью однофакторного дисперсионного анализа.