СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ПРОГРАММЕ Statistica: ДИСПЕРСИОННЫЙ АНАЛИЗ С ПОВТОРНЫМИ ИЗМЕРЕНИЯМИ

Пример 1 – Известны величины прибыли 12 предприятий за три года (2012, 2013, 2014):

2012	300	240	380	180	320	340	280	170	300	320	300	210
2013	330	210	370	190	340	360	250	210	290	300	280	220
2014	270	210	300	210	320	300	270	200	270	300	260	190

Это значит, например, что прибыль первого предприятия в 2012, 2013 и 2014 годах составила 300, 330 и 270 ден.ед. соответственно.

Требуется определить, значимо ли различаются прибыли в разные годы.

В этой задаче анализируемая *переменная* – прибыль. Имеется один *фактор* – время. В данной задаче имеются три *зависимых* выборки. Под зависимыми выборками понимают значения *одной и той же* переменной (в данном случае – прибыли), полученные в *разные периоды* (в данном случае – в 2012, 2013, 2014 году). Такие задачи решаются методами *дисперсионного анализа с повторными измерениями*. В данной задаче – *три* повторных измерения (данные за три года).

Рассмотрим решение этой задачи в Statistica (в MS Excel нет средств для решения таких задач).

Примечание – Если имеются только две зависимые выборки (например, данные о прибылях за два года), то используется *t*-критерий Стьюдента для зависимых выборок: в MS Excel для решения таких задач применяется инструмент **Парный двухвыборочный t-тест для средних**, (см. лабораторную работу №1), в Statistica – элемент меню t-**test, dependent samples** (см. лабораторную работу №4).

Примечание – Дисперсионный анализ *с повторными измерениями* необходимо отличать от *дисперсионного анализа с повторениями*. Дисперсионный анализ *с повторными измерениями* применяется для решения задач, где анализируемый фактор (или один из анализируемых факторов) – время. Дисперсионный анализ *с повторениями* применяется в случаях, когда имеется несколько факторов (например, регион, отрасль промышленности и форма собственности), влияющих на анализируемую величину (например, прибыль), причем для каждой комбинации факторов известно *несколько* значений анализируемой величины (т.е., например, известна прибыль нескольких государственных предприятий нефтяной промышленности в Северном регионе; аналогично – для других комбинаций факторов). Пример такой задачи рассмотрен в лабораторной работе №3 (в MS Excel). Для решения таких задач в MS Excel применяется инструмент **Двухфакторный дисперсионный анализ с повторениями**, в Statistica - элемент меню **Statistics** – **ANOVA** – **Factorial ANOVA**.

1 Присвоить переменным имена, например, **prib2012, prib2013** и **prib2014**. Ввести данные (см. рисунок 1).

	1 prib2012	2 prib2013	3 prib2014
1	300	330	270
2	240	210	210
3	380	370	300
4	180	190	210
5	320	340	320
6	340	360	300
7	280	250	270
8	170	210	200
9	300	290	270
10	320	300	300
11	300	280	260
12	210	220	190

Рисунок 1 – Данные для задачи дисперсионного анализа с повторными измерениями в Statistica

2 Выбрать **Statistics** – **ANOVA** – **Repeated measures ANOVA**. Нажать **OK**.

3 В окне **ANOVA/MANOVA Repeated measures ANOVA** нажать кнопку **Variables** и выбрать переменные для анализа: в списке **Dependent variables** выбрать переменные **prib2012, prib2013** и **prib2014**, а в списке **Categorical factors** не выбирать ничего (так как в данной задаче нет категориальных переменных). Нажать **OK**.

4 В окне **ANOVA/MANOVA Repeated measures ANOVA** нажать кнопку **Within effects**. Убедиться, что в поле **No. of levels** установлено значение 3 (количество повторений). В поле **Factor name** ввести произвольное имя фактора, например, **God** (или любое другое). Нажать **OK**.

5 В окне **ANOVA/MANOVA Repeated measures ANOVA** еще раз нажать **OK**. Появляется окно **ANOVA Results** с несколькими вкладками.

6 Выбрать вкладку **Quick** и нажать кнопку **All effects**. Выводится окно результатов (см. рисунок 2).

Effect	Repeated Measures Analysis of Variance (дисп_анализ_повтс Sigma-restricted parameterization Effective hypothesis decomposition SS	Degr. of Freedom	MS	F	p
Intercept	2662336	1	2662336	283,1436	0,000000
Error	103431	11	9403		
GOD	3339	2	1669	4,1762	0,029008
Error	8794	22	400		

Рисунок 2 – Результаты решения задачи дисперсионного анализа с повторными измерениями

В этом окне отображается расчетный уровень значимости (**p**). В данном случае **p**<0,05. Это означает, что фактор «время» статистически значим, т.е. прибыли в разные годы различались значимо.

7 Вернуться в окно **ANOVA Results**. Нажать кнопку **All effects/Graphs,** выбрать фактор **God** и нажать **OK**. Строится график (см. рисунок 3), отражающий величины прибыли по годам. Видно, что прибыль в 2012 и 2013 году была почти одинаковой, а в 2014 – существенно ниже.

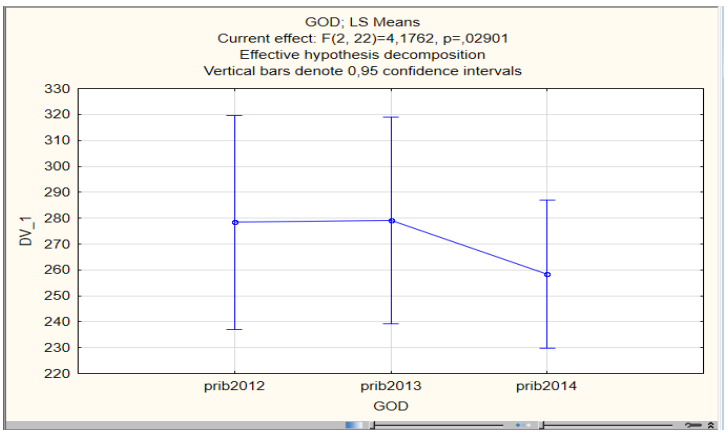


Рисунок 3 – Результаты решения задачи дисперсионного анализа с повторными измерениями: средние по годам

Выявление групп, различающихся значимо

8 Вернуться в окно **ANOVA Results**. Перейти на вкладку **Post-hoc** (если она отсутствует, нажать **More results**). Нажать кнопку **Fisher LSD**. Результаты приведены на рисунке 4.

LSD test; variable DV_1 (дисп_анализ_повторны
Probabilities for Post Hoc Tests
Error: Within MS = 399,75, df = 22,000

Cell No.	GOD	{1} 278,33	{2} 279,17	{3} 258,33
1	prib2012		0,919607	0,022695
2	prib2013	0,919607		0,018159
3	prib2014	0,022695	0,018159	

Рисунок 4 – Результаты решения задачи дисперсионного анализа с повторными измерениями: анализ на основе НСР

Для каждой пары групп (группы в данном случае соответствуют годам) вычисляется уровень значимости. Если он меньше заданного уровня α, то различие между группами статистически значимо. В данном примере видно, что группа 3 (т.е. прибыль в 2014 г.) значимо отличается от двух других групп. Из средних значений (278,33, 279,17 и 258,33), а также из рисунка 3 видно, что прибыли в 2014 году значительно **ниже**, чем в 2012 и 2013 годах.

Пример 2 – Пусть данные о прибылях за три года имеются для предприятий двух отраслей (нефтяная и металлургическая):

Отрасль	Н	М	Н	М	Н	Н	М	М	М	Н	М	Н
2012	300	240	380	180	320	340	280	170	300	320	300	210
2013	330	210	370	190	340	360	250	210	290	300	280	220
2014	270	210	300	210	320	300	270	200	270	300	260	190

Требуется определить факторы, существенно влияющие на прибыль.

В данной задаче анализируются *два фактора*: время и отрасль. Это задача дисперсионного анализа с повторными измерениями и с категориальной (группирующей) переменной. Здесь отрасль – категориальная переменная.

1 Ввести исходные данные, как показано на рисунке 5.

	1 prib2012	2 prib2013	3 prib2014	4 otrasl
1	300	330	270	n
2	240	210	210	m
3	380	370	300	n
4	180	190	210	m
5	320	340	320	n
6	340	360	300	n
7	280	250	270	m
8	170	210	200	m
9	300	290	270	m
10	320	300	300	n
11	300	280	260	m
12	210	220	190	n

Рисунок 5 – Данные для задачи дисперсионного анализа с повторными измерениями и категориальной переменной

2 Выбрать **Statistics – ANOVA – Repeated measures ANOVA**. Нажать **OK**.

3 В окне **ANOVA/MANOVA Repeated measures ANOVA** нажать кнопку **Variables** и выбрать переменные для анализа: в списке **Dependent variables** выбрать переменные **prib2012, prib2013** и **prib2014**, а в списке **Categorical factors** – **Otrasl**. Нажать **OK**.

4 В окне **ANOVA/MANOVA Repeated measures ANOVA** нажать кнопку **Within effects**. Убедиться, что в поле **No. of levels** установлено значение 3 (количество повторений). В поле **Factor name** ввести произвольное имя фактора, например, **God** (или любое другое). Нажать **OK**.

5 В окне **ANOVA/MANOVA Repeated measures ANOVA** еще раз нажать **OK**. Появляется окно **ANOVA Results** с несколькими вкладками.

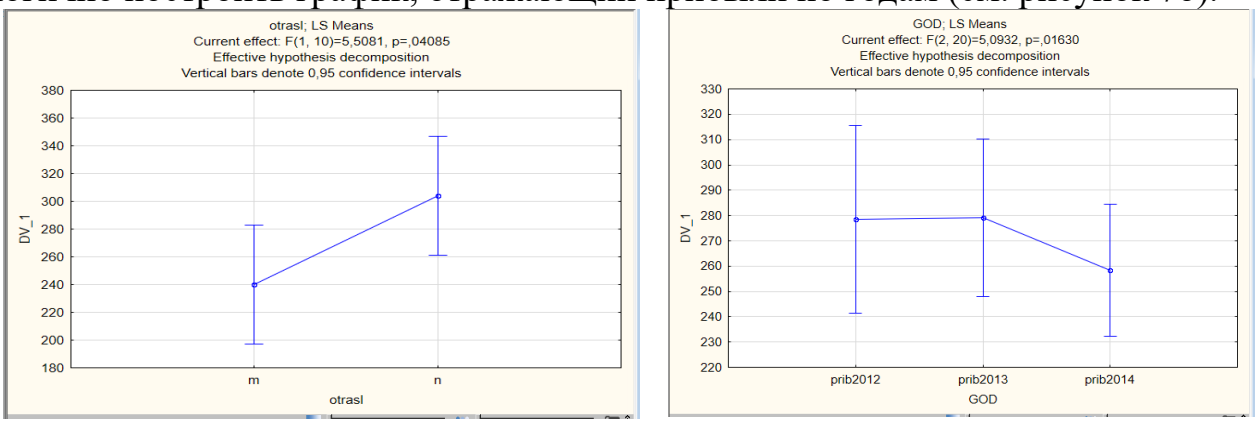
6 Выбрать вкладку **Quick** и нажать кнопку **All effects**. Выводится окно результатов (см. рисунок 6).

Effect	Repeated Measures Analysis of Variance (дисп_анализ_повт Sigma-restricted parameterization Effective hypothesis decomposition **SS**	Degr. of Freedom	MS	F	p
Intercept	2662336	1	2662336	399,1841	0,000000
otrasl	36736	1	36736	5,5081	0,040853
Error	66694	10	6669		
GOD	3339	2	1669	5,0932	0,016300
GOD*otrasl	2239	2	1119	3,4153	0,052968
Error	6556	20	328		

Рисунок 6 – Результаты решения задачи дисперсионного анализа с повторными измерениями и категориальной переменной

Появляется окно результатов с вычисленными уровнями значимости (**p**) для обоих факторов (отрасли и времени) и их взаимодействия (строка **GOD*otrasl**). Для обоих факторов **p**<0,05. Это означает, что оба фактора значимы, т.е. прибыль существенно различается как в разные годы, так и в разных отраслях. Для их взаимодействия уровень значимости **p** лишь немного превышает 0,05; это означает, что влияние одного фактора (например, отрасли) может зависеть от другого фактора (например, года). Чтобы исследовать влияние факторов и их взаимодействия, требуется дополнительный анализ (см. ниже).

7 Вернуться в окно **ANOVA Results**. Нажать кнопку **All effects/Graphs,** выбрать фактор **Otrasl** и нажать **OK**. Строится график (см. рисунок 7а), отражающий прибыли по отраслям. Снова вернуться в окно **ANOVA Results** и аналогично построить график, отражающий прибыли по годам (см. рисунок 7б).

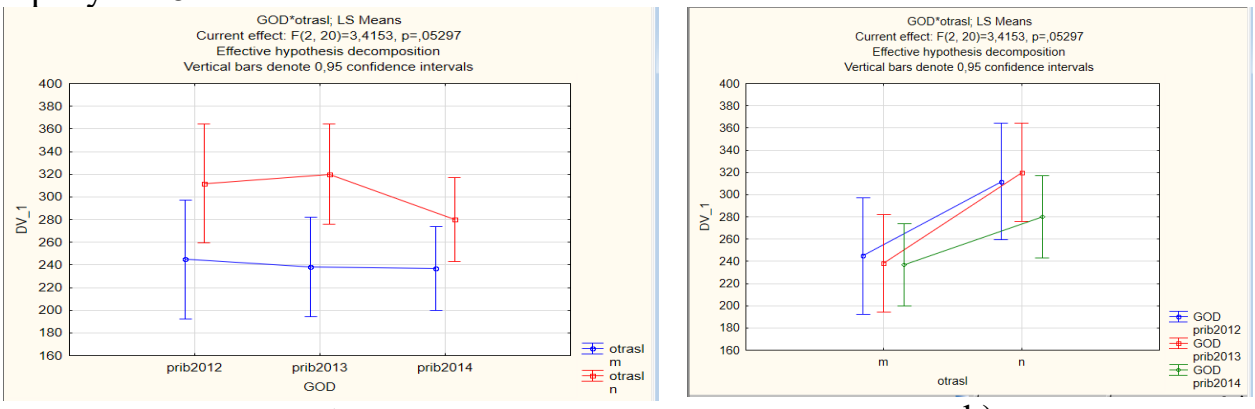


а) б)

Рисунок 7 – Результаты решения задачи дисперсионного анализа с повторными измерениями и категориальной переменной: а) средние по отраслям; б) средние по годам

Выводы, которые можно сделать из этих двух графиков, следующие: а) прибыли в нефтяной промышленности значительно выше, чем в металлургической; б) прибыли в 2012 и 2013 годах примерно одинаковы, а в 2014 году – значительно ниже.

8 Чтобы проанализировать возможные взаимодействия факторов, вернуться в окно **ANOVA Results**, нажать кнопку **All effects/Graphs**., затем выбрать **GOD * OTRASL** и нажать **OK**. В следующем окне выбрать настройки **X axis** – **GOD**, **Line pattern** – **Otrasl**. Нажать **OK**. Полученный график см. на рисунке 8а. Снова вернуться в окно **ANOVA Results** и выполнить те же действия, но выбрать **X axis** – **Otrasl**, **Line pattern** – **GOD**. Полученный график см. на рисунке 8b.



а) b)

Рисунок 8 – Результаты решения задачи дисперсионного анализа с повторными измерениями и категориальной переменной: средние для всех групп в двух формах

Из этих графиков, особенно из рисунка 8а, можно сделать еще один вывод: снижение прибыли в 2014 году наблюдалось только в нефтяной промышлен-

ность (но не в металлургии). Кроме того, оба графика подтверждают, что прибыли в нефтяной промышленности во все годы были значительно выше, чем в металлургии.

Выявление групп, различающихся значимо

9 Вернуться в окно **ANOVA Results**. Перейти на вкладку **Post-hoc** (если она отсутствует, нажать **More results**). В поле **Effect** выбрать **GOD * OTRASL**. Нажать кнопку **Fisher LSD**. Результаты расчетов на основе НСР приведены на рисунке 9.

LSD test; variable DV_1 (дисп_анализ_повторные_измерения)
Probabilities for Post Hoc Tests
Error: Between; Within; Pooled MS = 2441,7, df = 12,004

Cell No.	otrasl	GOD	{1} 245,00	{2} 238,33	{3} 236,67	{4} 311,67	{5} 320,00	{6} 280,00
1	m	prib2012		0,530844	0,434674	0,037594	0,022013	0,243404
2	m	prib2013	0,530844		0,874915	0,024517	0,014280	0,169822
3	m	prib2014	0,434674	0,874915		0,022013	0,012811	0,154668
4	n	prib2012	0,037594	0,024517	0,022013		0,434674	0,006621
5	n	prib2013	0,022013	0,014280	0,012811	0,434674		0,001055
6	n	prib2014	0,243404	0,169822	0,154668	0,006621	0,001055	

Рисунок 9 – Результаты анализа на основе НСР для всех факторов

Из этих результатов видно, что прибыли в нефтяной промышленности в 2012 и 2013 году были существенно *выше*, чем все остальные прибыли (т.е. прибыли в нефтяной промышленности в 2014 году или прибыли в металлургии в любой год). Ни в одном из годов не было значимых различий между прибылями в металлургии. В 2014 году не было значимого различия между отраслями.

10 Чтобы выполнить анализ на основе НСР для каждого фактора отдельно, снова вернуться в окно **ANOVA Results**. Выбрать вкладку **Post-hoc**. В поле **Effect** выбрать **GOD**, затем нажать **Fisher LSD**. Результаты анализа на основе НСР для прибылей по годам см. на рисунке 10а. Затем аналогично выполнить анализ прибылей по отраслям. Результаты см. на рисунке 10б.

LSD test; variable DV_1 (дисп_анализ_повторны
Probabilities for Post Hoc Tests
Error: Within MS = 327,78, df = 20,000

Cell No.	GOD	{1} 278,33	{2} 279,17	{3} 258,33
1	prib2012		0,911355	0,013600
2	prib2013	0,911355		0,010609
3	prib2014	0,013600	0,010609	

а)

LSD test; variable DV_1 (дисп_ана
Probabilities for Post Hoc Tests
Error: Between MS = 6669,4, df =

Cell No.	otrasl	{1} 240,00	{2} 303,89
1	m		0,040853
2	n	0,040853	

б)

Рисунок 10 – Анализ на основе НСР для каждого из факторов: а) время; b) отрасль

Эти результаты подтверждают выводы, сделанные ранее: а) прибыли в нефтяной промышленности значительно выше, чем в металлургической; б) прибыли в 2014 году были значительно ниже, чем в 2012 и 2013 годах.

Задание 1 – Известны средние зарплаты на шести предприятиях за четыре года (2015, 2016, 2017, 2018):

2015	420	370	480	480	620	620
2016	370	480	430	390	540	540
2017	520	460	570	510	600	580
2018	490	410	480	500	570	520

Найти, значимо ли различались зарплаты в разные годы. Если различие по годам значимо, то определить конкретные годы, когда зарплаты различались значимо.

Задание 2 – Известны средние зарплаты для десяти предприятий трех отраслей (металлургическая, химическая, нефтяная) за два года (2015, 2016):

	металлургическая			химическая				нефтяная		
2015	170	300	380	180	320	340	280	420	350	210
2016	240	370	430	220	340	360	260	330	270	200

Это значит, например, что на предприятии 1 (металлургическом) средняя зарплата составляла 170 ден.ед. в 2015 и 240 ден.ед. в 2016 году. На предприятии 6 (химическом) средняя зарплата составляла 280 ден.ед. в 2015 и 260 ден.ед. в 2016 году.

Определить факторы, значимо влиявшие на зарплату, т.е. найти, значимо ли отличались зарплаты по годам и по отраслям. Найти комбинации факторов, при которых зарплата различалась значимо.

Задание 3 – Известны средние зарплаты за два года (2015, 2016) для 22 предприятий трех отраслей (металлургическая, химическая, нефтяная), расположенных в двух регионах:

Северный регион:

	металлургическая			химическая				нефтяная		
2015	170	300	380	180	320	340	280	420	350	210
2016	240	370	430	220	340	360	260	330	270	200

Южный регион:

	металлургическая				химическая					нефтяная		
2015	270	320	350	480	420	310	480	420	380	510	400	470
2016	240	410	380	500	450	330	460	470	420	380	400	430

Определить факторы, значимо влиявшие на зарплату. Найти комбинации факторов, при которых зарплата различалась значимо.