СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ПРОГРАММЕ Statistica: КОРРЕЛЯЦИОННЫЙ, РЕГРЕССИОННЫЙ, ДИСПЕРСИОННЫЙ АНАЛИЗ

1 Корреляционный анализ

1.1 Коэффициент парной корреляции

Пример 1 – Исследуются факторы, влияющие на текучесть кадров на предприятиях. Анализируется зависимость текучести кадров от заработной платы и размера предприятия (численности работающих). Получены данные по девяти предприятиям.

Средняя заработная плата, ден.ед.	135	170	115	160	230	180	90	210	150
Численность работающих	200	320	290	200	170	230	280	300	260
Текучесть кадров, %	6,2	8,7	11,2	4,9	3,1	4,5	9,7	4,2	5,1

Найти коэффициенты парной корреляции: текучести кадров с заработной платой и с численностью работающих. Проанализировать полученные результаты.

1 Присвоить свободным столбцам имена ZARP, CHISL и TEK. Ввести в эти столбцы исходные данные.

2 Из меню **Statistics** выбрать команду **Basic Statistics and Tables**, затем – команду **Correlation Matrices**. На экран выводится окно **Product Moment and Partial Correlations**.

3 Перейти на *вкладку* **Options** (не путать с кнопкой **Options**). Установить переключатель **Display r, p–levels and N's**, чтобы на экран выводились коэффициенты корреляции и расчетные уровни значимости (а также объемы выборок). Убедиться, что в поле **p–level for highlighting** указано **.05**; это означает, что коэффициенты корреляции с расчетным уровнем значимости менее 0,05 (т.е. статистически значимые) будут выделяться на экране красным цветом.

4 Для выбора анализируемых переменных нажать кнопку **Two lists (rect. matrix)**. В появившемся окне **Select one or two variable lists** в первом списке выбрать переменную TEK, во втором – переменные ZARP и CHISL. Нажать **OK**. Выполняется возврат в окно **Product Moment and Partial Correlations**.

5 Для получения результатов нажать кнопку **Summary**. На экран выводятся коэффициенты корреляции, расчетные уровни значимости (**p**), а также объем выборки (**N**).

6 Выполнить анализ результатов и указать выводы в окне отчета.

1.2 Коэффициенты частной корреляции

Пример 2 – Исследуется связь прочности меха некоторых животных, оцениваемой по 100-балльной шкале (чем прочнее мех, тем выше балл), с длиной шерсти и с весом животного. Имеются данные для 12 животных (см. таблицу).

Длина шерсти, мм	14	25	11	17	29	10	15	27	19	11	18	28
Вес животного, кг	6,7	5,8	3,9	4,7	5,2	3,5	7,1	6,6	7,1	4,2	7,4	5,1
Показатель прочности меха, баллы	48	30	38	42	37	42	49	41	53	52	43	37

Вычислить коэффициенты частной корреляции между: а) показателем прочности меха и длиной шерсти; б) показателем прочности меха и весом животного.

Коэффициенты частной корреляции, как и коэффициенты парной корреляции, вычисляются с помощью команды **Statistics – Basic Statistics/Tables – Correlation Matrices**.

1 Присвоить переменным, в которых будут вводиться исходные данные, имена DLINA, VES и PROCHN. Ввести исходные данные. Для ввода исходных данных потребуется 12 строк и три столбца.

2 Найти (для последующего анализа) коэффициенты ***парной*** корреляции показателя прочности с длиной шерсти и весом животного, как показано в разделе 1.1.

3 Чтобы приступить к вычислению коэффициентов ***частной*** корреляции, вернуться в окно **Product-Moment and Partial Correlations** (нажатием на его кнопку в нижней части экрана). В этом окне перейти на вкладку **Options**. Для выбора анализируемых переменных нажать кнопку **Two lists (rect. matrix)**. В первом списке (**First variable list**) выбираются *две* переменные, *для которых определяется коэффициент частной корреляции*. Во втором списке (**Second variable list**) выбираются *все* переменные, *влияние которых требуется исключить*. Например, чтобы найти коэффициент частной корреляции между *прочностью меха* и *длиной шерсти* (при исключенном влиянии *веса животного*), необходимо в списке **First variable list** выбрать переменные DLINA и PROCHN (нажать клавишу **Ctrl** и, не отпуская ее, выбрать эти переменные с помощью мыши). В списке **Second variable list** выбрать переменную VES. Нажать **OK**. Выполняется возврат в окно **Product-Moment and Partial Correlations.**

4 Перейти на вкладку **Advanced/Plot**. Для получения результатов нажать кнопку **Partial Correlations**. В файлы для результатов (рабочую книгу и отчет) выводятся два окна. В основном окне результатов указывается коэффициент частной корреляции и расчетный уровень значимости (**p**); эти величины указываются в окне *дважды*. Указывается также объем выборки (**N**). Во втором окне указывается переменная, влияние которой исключается (в данном примере – переменная VES).

5 Сохранить рабочую книгу и файл-отчет с результатами.

6 Вернуться в окно **Product-Moment and Partial Correlations** (нажатием на его кнопку в нижней части экрана).

7 Найти коэффициент частной корреляции между прочностью меха и весом животного при исключенном влиянии длины шерсти (аналогично показанному на шагах 4 и 5).

8 Проанализировать полученные результаты. Сравнить результаты анализа коэффициентов парной и частной корреляции (указание – более достоверны коэффициенты ***частной*** корреляции).

Задание 1 – Анализируются факторы, влияющие на заработную плату на предприятиях по добыче некоторого стройматериала. Имеются данные по 11 предприятиям:

Производительность, т/день	115	96	124	88	117	106	92	105	120	92	125
Глубина добычи, м	4,3	5,6	3,4	9,7	6,5	3,1	4,1	7,1	3,7	2,8	5,1
Доля материала высшего сорта, %	46	34	48	32	46	36	34	35	47	32	47
Средняя зарплата, ден.ед.	1200	1000	1300	900	1250	1080	950	1000	1240	950	1300

Найти: а) коэффициенты парной корреляции заработной платы с каждым из анализируемых факторов (т.е. с производительностью, с глубиной добычи и с долей материала высшего сорта); б) коэффициенты частной корреляции заработной платы с каждым из анализируемых факторов. Для всех найденных коэффициентов корреляции выполнить проверку статистической значимости и сделать выводы.

2 Регрессионный анализ

Пример 3 – По данным примера 1 построить и проанализировать линейную модель связи текучести кадров со средней зарплатой и численностью работающих. Найти прогноз средней текучести кадров и ее отдельных значений при заработной плате 200 ден.ед. и численности работающих 300 чел.

1 Убедиться, что на рабочем листе (Spreadsheet) имеются исходные данные, введенные в примере 1 (переменные ZARP, CHISL и TEK).

2 Из меню **Statistics** выбрать команду **Multiple Regression**. Появляется окно **Multiple Linear Regression**.

3 В окне **Multiple Linear Regression** нажать кнопку **Variables** для выбора анализируемых переменных. В списке **Dependent var.** выбрать переменную TEK, в списке **Independent variable list** – переменные ZARP и CHISL. Нажать **OK**. Выполняется возврат в окно **Multiple Regression**. В нем также нажать **OK**.

4 На экран выводится окно **Multiple Regression Results** с несколькими кнопками для получения результатов. Для построения линейной модели, а также для проверки ее значимости перейти на вкладку **Quick** (или **Advanced**) и нажать кнопку **Summary: Regression results**. Отображаются *два* окна результатов. Основные результаты приводятся в окне **Data: Regression Summary for Dependent Variable** (см. рисунок 1).

N=9	Regression Summary for Dependent Variable: TEK (Spreadsheet1) R= ,88707071 R?= ,78689444 Adjusted R?= ,71585925 F(2,6)=11,078 p<,00968 Std.Error of estimate: 1,4904 b*	Std.Err. of b*	b	Std.Err. of b	t(6)	p-value
Intercept			7,723900	3,804490	2,03021	0,088638
ZARP	-0,672264	0,197655	-0,042634	0,012535	-3,40120	0,014474
CHISL	0,410558	0,197655	0,021990	0,010587	2,07715	0,083062

Рисунок 1 – Построение линейной модели связи между величинами *X* и *Y*

Коэффициенты линейной модели, полученные методом наименьших квадратов, указываются в столбце **B**. Здесь в строке **Intercept** указан коэффициент a_0, а в последующих строках – коэффициенты при соответствующих входных переменных.

Проверка адекватности модели выполняется по расчетному уровню значимости **p**. Если **p**<0,05, то модель адекватна (достаточно точна).

Проверка статистической значимости коэффициентов модели выполняется по расчетным уровням значимости, указанным в столбце **p–value**. Если **p–value**<0,05, то соответствующий коэффициент модели статистически значим.

Кроме того, в окне результатов выводится коэффициент детерминации (**R?**) и еще некоторые величины.

5 Выполнить анализ результатов и указать выводы в окне отчета. Привести уравнение модели зависимости текучести кадров от средней зарплаты и численности работающих, вывод об адекватности модели, выводы о статистической значимости (или незначимости) ее коэффициентов.

Прогнозирование выходной величины

1 Вернуться в окно **Multiple Regression Results**. Если оно уже закрыто, заново выполнить шаги 1–4, описанные при построении модели.

2 Перейти на вкладку **Residuals/assumptions/predictions**.

3 Для получения прогноза *среднего* значения выходной величины (текучести кадров) с заданной точностью выполнить следующее:

- в области **Predict values** установить переключатель **Compute confidence limits**;
- в поле **Alpha** указать желаемую точность прогноза. Обычно используется стандартное значение 0,05, что соответствует 95-процентной точности;
- нажать кнопку **Predict dependent variable**;
- в появившемся окне указать заданные значения входных величин: ZARP: 200; CHISL: 300. Нажать **OK**.

На экран выводится окно результатов (см. рисунок 2). Границы диапазона, в котором с вероятностью 95% будет находиться среднее значение выходной величины при заданных значениях входных величин, указываются в полях **+/-95,0%CL**. В данном примере видно, что при зарплате 200 ден.ед. и числен-

ности работающих 300 чел. среднее значение текучести кадров с вероятностью 95% будет находиться в диапазоне от 3,42 до 8,17%.

4 Вернуться в окно **Multiple Regression Results**. Для прогнозирования *отдельных значений* выходной величины установить переключатель **Compute prediction limits** (вместо **Compute confidence limits**). Затем выполнить те же действия, что и при получении прогноза для среднего значения. Границы диапазона, в котором с вероятностью 95% будут находиться отдельные значения выходной величины при заданных значениях входных величин, указываются в полях **+/–95,0%PL**. Результаты для данного примера приведены на рисунке 3. В данном случае при зарплате 200 ден.ед. и численности работающих 300 чел. отдельные значения текучести с вероятностью 95% будут находиться в диапазоне от 1,45 до 10,14%.

Variable	Predicting Values for (Spreadsheet1) variable: TEK		
	b-Weight	Value	b-Weight * Value
ZARP	-0,042634	200,0000	-8,52676
CHISL	0,021990	300,0000	6,59701
Intercept			7,72390
Predicted			5,79415
-95,0%CL			3,42379
+95,0%CL			8,16451

Рисунок 2 – Прогноз среднего значения выходной величины

Variable	Predicting Values for (Spreadsheet1) variable: TEK		
	b-Weight	Value	b-Weight * Value
ZARP	-0,042634	200,0000	-8,52676
CHISL	0,021990	300,0000	6,59701
Intercept			7,72390
Predicted			5,79415
-95,0%PL			1,44465
+95,0%PL			10,14365

Рисунок 3 – Прогноз отдельных значений выходной величины

Задание 2 – По данным задания 1 построить (в системе Statistica) линейную модель зависимости средней зарплаты от анализируемых факторов (производительности, глубины добычи и доли материала высшего сорта). Выполнить анализ и проверку модели. Найти прогноз средней зарплаты и ее отдельных значений для предприятия, где производительность составляет 90 т/день, глубина добычи – 5 м, доля материала высшего сорта – 30%.

3 Однофакторный дисперсионный анализ

Пример 4 – 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;
– отрасль П2: 638, 625, 662, 645, 640, 637;
– отрасль П3: 625, 617, 631, 629, 619.

Требуется выяснить, имеются ли значимые различия в заработной плате между отраслями. Если такие различия есть, то выяснить, какие именно отрасли значимо различаются по зарплате.

а) Оценка значимости различия между группами

1 Присвоить переменным имена, например, OTRASL и ZARPLATA. Убедиться, что в рабочем листе имеется не менее 16 строк. Ввести данные, как показано на рисунке 4. Здесь OTRASL – группирующая переменная.

2 Выбрать **Statistics** – **ANOVA** – **One-way ANOVA**. Нажать **OK**.

3 В окне **ANOVA/MANOVA One-way ANOVA** нажать кнопку **Variables** для выбора анализируемых переменных. В списке **Categorical factor** выбрать переменную OTRASL, в списке **Dependent variables** – ZARPLATA. Нажать **OK**.

4 В окне **ANOVA/MANOVA One-way ANOVA** нажать **OK**. Появляется окно **ANOVA Results** с несколькими вкладками.

5 Перейти на вкладку **Quick** и нажать кнопку **All effects**. Окно результатов показано на рисунке 5.

Расчетный уровень значимости обозначается как **p** (в строке переменной OTRASL). Значение **p** <0,05 указывает, что между группами (отраслями) имеется значимое различие по величинам зарплат.

Кроме того, в окне результатов выводятся некоторые другие величины, например, в столбце **MS** – межгрупповой разброс S_v^2 (в строке OTRASL) и внутригрупповой разброс S^2 (в строке **Error**).

	1 OTRASL	2 ZARPLATA
1	p1	640
2	p1	637
3	p1	648
4	p1	624
5	p1	639
6	p2	638
7	p2	625
8	p2	662
9	p2	645
10	p2	640

Рисунок 4 – Исходные данные для задачи однофакторного дисперсионного анализа

Univariate Tests of Significance for ZARPLATA (Spreadsheet16)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	6390498	1	6390498	69880,67	0,000000
OTRASL	844	2	422	4,62	0,030577
Error	1189	13	91		

Рисунок 5 – Результаты решения задачи однофакторного дисперсионного анализа

б) Получение результатов по группам

1 Вернуться в окно **ANOVA Results**. Нажать кнопку **All effects/Graphs**, затем – кнопку **OK**. Строится графическое представление групп (отображаются средние по группам и доверительные интервалы для средних), как показано на рисунке 6. Из графика видно, что максимальные (и достаточно близкие) величины зарплаты имеются в отраслях П1 и П2, существенно меньшие зарплаты – в отрасли П3.

2 Снова вернуться в окно **ANOVA Results**. Перейти на вкладку **Summary**. Нажать кнопку **Cell statistics**. На экран выводятся статистические показатели

групп и выборки в целом (см. рисунок 7). Здесь видно, например, что средняя зарплата по выборке составляет 634,75 ден.ед., в том числе по отрасли П1 – 637,60 ден.ед., П2 – 641,17 ден.ед., П3 – 624,20 ден.ед.

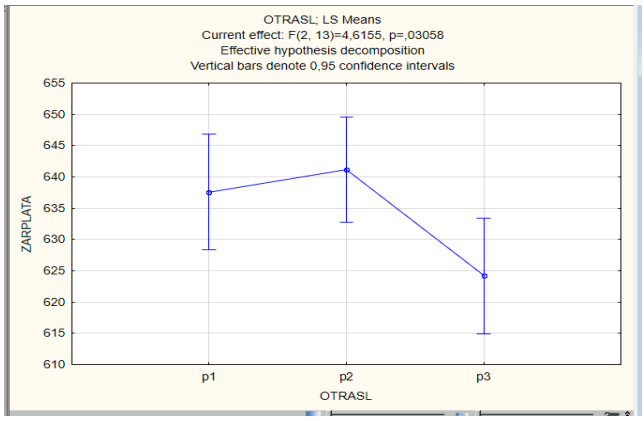

Рисунок 6 – Графическое представление групп

Effect	Descriptive Statistics (Spreadsheet16) Level of Factor	N	ZARPLATA Mean	ZARPLATA Std.Dev.	ZARPLATA Std.Err	ZARPLATA -95,00%	ZARPLATA +95,00%
Total		16	634,7500	11,64188	2,910470	628,5465	640,9535
OTRASL	p1	5	637,6000	8,67756	3,880722	626,8254	648,3746
OTRASL	p2	6	641,1667	12,15593	4,962638	628,4098	653,9235
OTRASL	p3	5	624,2000	6,09918	2,727636	616,6269	631,7731

Рисунок 7 – Статистические показатели по группам

в) Выявление групп, различающихся значимо

Снова вернуться в окно **ANOVA Results**. Перейти на вкладку **Post-hoc** (если она не отображается, нажать кнопку **More results**). Чтобы выполнить анализ на основе НСР, нажать кнопку **Fisher LSD**. Результаты приведены на рисунке 8.

Для каждой пары групп (отраслей) вычисляется расчетный уровень значимости (вероятность того, что различие между группами статистически *незначимо*). Если он меньше заданной величины α (обычно α=0,05), то различие между группами статистически значимо.

Cell No.	LSD test; variable ZARPLATA (Spreadsheet16) Probabilities for Post Hoc Tests Error: Between MS = 91,449, df = 13,000 OTRASL	{1} 637,60	{2} 641,17	{3} 624,20
1	p1		0,548568	0,045185
2	p2	0,548568		0,011713
3	p3	0,045185	0,011713	

Рисунок 8 – Результаты анализа на основе НСР

В данном примере значимо различие по величинам зарплат между отраслями П1 и П3, между П2 и П3 (для этих пар отраслей расчетные уровни значимости меньше, чем 0,05). Различие между зарплатами в отраслях П1 и П2 статистически незначимо (уровень значимости больше 0,05). С учетом средних значений в группах (см. рисунки 6, 7) можно сделать вывод, что зарплаты в отраслях П1 и П2 *значимо выше*, чем в отрасли П3.

Задание 3 – 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. Указать наиболее прочные и наименее прочные материалы.