РЕФЕРАТ

Выпускная квалификационная работа 103 с., 68 рис., 42 табл., 21 источник.

Ключевые слова: корреляция, корреляционный анализ, распределение, коэффициент корреляции, однородность распределения, фактор, однофакторный анализ.

Объектом исследования являются результаты сдачи выпускных экзаменов по русскому языку и математике (ОГЭ и ЕГЭ) выпускниками школ томской области..

Цель работы — исследования набора данных с применением различных статистических методов для выявления и описания различного рода зависимостей по результатам сдачи ОГЭ и ЕГЭ по русскому языку и математике при различном наборе факторов.

В процессе исследования проводились анализ литературы, обзор различных методов статистического анализа.

В результате исследования были выявлены зависимости изучаемых переменных от различных наборов факторов.

Работа имеет следующую структуру.

Первый раздел описывает предметную область, значимость исследований в данной предметной области, а также описывает исходные данные для анализа.

Второй раздел посвящен описанию методов статистического анализа, которые будут применены в дальнейшей работе.

Третий раздел содержит первичные исследования исходных данных.

Четвертый раздел включает проверку исходных данных на нормальность распределения для определения применимости отдельных методов статистического анализа.

Пятый раздел посвящен проведенному корреляционному анализу и его результатам.

Шестой раздел содержит информацию о результатах проведения однофакторного анализа исходных данных.

В седьмом разделе приводится анализ перспективности и технико-экономических показателей, а также ресурсоэффективность и актуальность данного исследования.

Восьмой раздел посвящен анализу аспектов социальной ответственности.

Область применения: использование результатов статистического анализа для повышения качества обучения учеников Томской области.

В будущем планируется дальнейшее исследование данных с применением различных нестандартных моделей.

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ, СОКРАЩЕНИЯ

ОГЭ – Основной государственный экзамен.

ЕГЭ – Единый государственный экзамен.

Дисперсия – математическое ожидание квадрата отклонения случайной величины от ее математического ожидания.

Квантиль – значение, которое заданная случайная величина не превышает с фиксированной вероятностью.

Корреляционная матрица – матрица коэффициентов корреляции нескольких случайных величин.

Нормальное распределение – распределение вероятностей, которое описывается плотностью $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-a)^2}{2\sigma^2}}$ (для непрерывной случайной величины)

Ранг – положение в упорядоченном ряду значений. Значения в некотором измерении сообщают только положение этого значения относительно всех других, однако ничего не говорит о расстоянии между значениями.

Оценка - это число, вычисляемое на основе наблюдений, предположительно близкое к оцениваемому параметру.

Случайная величина – величина, которая в результате испытания примет одно и только одно возможное значение, заранее неизвестное и зависящее от случайных факторов, которые предварительно не могут быть учтены.

Стандартное отклонение – показатель рассеивания значений случайной величины относительно её математического ожидания.

Уровень значимости – достаточно малая вероятность, при которой событие можно считать практически невозможным.

r – коэффициент корреляции.

p-level – уровень значимости.

m – размерность множественной регрессии.

Multiple R – коэффициент множественной корреляции.

R^2_{adj} – (скорректированый) коэффициент детерминации.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В современном мире информационных технологий статистический анализ играет существенную роль как в государственных, промышленных масштабах, так и в социальных. В результате анализа возможно выявление различных зависимостей от набора некоторых факторов, количественное описание данной зависимости, оценка различных параметров, таких как: распределение данных, математическое ожидание, дисперсия и др.

Целью выпускной квалификационной работы является исследование исходного набора данных с применением различных статистических методов для выявления и описания различного рода зависимостей результатов экзаменов учеников Томской области за 2013 (9 класс) и за 2015 (11 класс). Данные об экзаменах содержат информацию о количестве набранных учениками баллов по двум дисциплинам: русский язык и математика.

Для анализа использовался программный пакет StatSoft STATISTICA, который является мощным инструментом для анализа данных, визуализации результатов, подготовки прогнозов, нейросетевых вычислений, data mining’a, контроля уровня качества и др.

Для достижения цели были поставлены следующие задачи:

- изучение предметной области;
- выбор и освоение наиболее приемлемых для анализа методов;
- изучение вспомогательного программного обеспечения;
- анализ данных с использованием выбранных статистических методов;
- описание полученных результатов;
- анализ ресурсоэффективности и ресурсосбережения;
- анализ аспектов социальной ответственности.

1 ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ

Определение закономерностей, которым следуют массовые случайные явления, основано на изучении результатов наблюдений путем применения методов статистического анализа.

Первая задача математической статистики – указать и определить методы сбора и группировки статистических сведений, получаемых в результате наблюдений или в результате эмпирических исследований.

Вторая задача математической статистики – в зависимости от целей исследований разработать методы для анализа статистических данных:

- оценка неизвестной вероятности какого-либо события; оценка неизвестной функции распределения; оценка параметров неизвестного вида распределения; оценка зависимости одной случайной величины от другой или нескольких случайных величин и др.;
- проверка статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид которого известен[1].

Современная математическая статистика разрабатывает способы определения числа необходимых испытаний до начала исследования (планирование эксперимента), в ходе исследования (последовательный анализ) и решает многие другие задачи[1]. Современную математическую статистику определяют как науку о принятии решений в условиях неопределенности.

Задача математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов[1].

Для анализа были предоставлены данные по Томской области, включающие в себя результаты ОГЭ и ЕГЭ учеников школ.

2 МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА

Существуют различные методы статистического анализа, предназначенные для выявления определенных зависимостей и решения различных задач математической статистики.

В данной работе будут использоваться следующие наиболее распространенные методы статистического анализа:

- корреляционный анализ;
- факторный анализ.

2.1 Описание метода корреляционного анализа

Огромный интерес для большого количества задач представляет обнаружение и определение связей между двумя, тремя и более случайными величинами. В инженерных исследованиях подобные задачи чаще всего сводятся к обнаружению связи между некоторым предполагаемым возбуждением и наблюдаемым от него откликом в изучаемой физической системе. Применительно к нашей работы будем рассматривать взаимосвязи между различными факторами, влияющими на качество обучения, и результатами обучения (тестирования).

Относительную силу и существование таких взаимосвязей возможно измерить через коэффициент корреляции.

Основная задача корреляционного анализа состоит в том, чтобы выявить связи между случайными переменными через точечную и интервальную оценки различных коэффициентов корреляции.

Коэффициент корреляции r_{xy} определяется через корреляционный момент (ковариацию) K_{xy} по формуле:

$$\rho_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y} = \frac{M[(X - m_x)(Y - m_y)}{\sigma_x \sigma_y} . \quad (2.1)$$

Величина p_{xy} характеризует силу взаимосвязи между случайными переменными X и Y в генеральной совокупности. Исходя из свойств

коэффициент корреляции видно, что он может являться показателем силы взаимосвязи только если две переменные зависимы линейно. Данный коэффициент равен нулю, если случайные величины зависимы нелинейно. Данный коэффициент изменяется в пределах [-1;1]. Предельные значения говорят о полной линейной зависимости. Чем больше по модулю данный коэффициент отличен от нуля, тем сильнее теснота связи [2].

2.2 Описание метода факторного анализа

Факторный анализ можно трактовать как раздел многомерного статистического анализа, который объединяет методы оценки размерностей множеств наблюдаемых переменных путем исследования структур ковариационных или корреляционных матриц.

Благодаря данному виду анализа исследователь может решить две основные задачи: компактно, но при этом всесторонне, описать предмет измерения. С помощью факторного анализа можно выявлять факторы, что отвечают за присутствие статистических линейных корреляционных связей между наблюдаемыми переменными.

Факторный анализ представляет собой методику системного и комплексного измерения и изучения воздействия факторов на величину результата (результативного показателя).

Основные задачи факторного анализа:

1. Отбор факторов, определяющих исследуемые результативные показатели.

2. Классификация и систематизация факторов с целью обеспечения комплексного и системного подхода к исследованию их влияния на результаты деятельности.

3.Определение формы зависимости между факторами и результативным показателем.

4.Моделирование взаимосвязей между результативными и факторными показателями.

3 ПОДГОТОВКА И ПЕРВИЧНЫЙ АНАЛИЗ ДАННЫХ

3.1 Подготовка и первичный анализ исходных данных

Процесс подготовки данных к анализу заключался в удалении записей с нулевыми результатами по экзаменам у учеников, после чего объем данных принял значение 3632 записи из первоначальных 4861 (см. рис. 3.1).

	3 Муниц	4 ОО	5 Класс	6 ОГЭ2013рус	7 ОГЭ2013мат	8 ОО	9 Класс	10 ЕГЭ2015русПерв	11 ЕГЭ2015матпПерв	12 ЕГЭ2015матбПерв	13 ЕГЭ2015русТест	14 ЕГЭ2015матТест	15 Категории
1	г.Томск	Кадетски	9Б	27	16	ОГКОУ К	11А	31	6	14	55	27	1
2	г.Томск	МАОУ СС	9Б	27	19	МАОУ СС	11А	39	6	NULL	65	27	1
3	Томский	МБОУ Ра	9	32	15	МБОУ Ра	11	31	7	17	55	33	0
4	г.Стрежев	МОУ СОЦ	9"Б"	28	21	МОУ СОЦ	11А	36	6	NULL	61	27	0
5	г.Томск	МБОУ СС	9В	32	27	МБОУ ли	11Е	51	20	NULL	87	80	1
6	г.Северск	МБОУ СС	9А	30	27	МБОУ СС	11А	51	11	NULL	87	55	1
7	г.Томск	МАОУ ги	9А	33	32	МАОУ ги	11А	44	23	NULL	71	86	1
8	г.Томск	МАОУ ги	9В	31	19	МАОУ ги	11Б	49	7	16	82	33	1
9	г.Северск	МБОУ СС	9Б	32	22	МБОУ СС	11Б	55	12	NULL	98	59	1
10	г.Северск	МБОУ Се	9В	32	28	МБОУ Се	11А	35	9	NULL	60	45	1
11	г.Северск	ОГБОУ С	9В	33	14	ОГКОУ К	11Б	28	3	9	51	14	1
12	г.Томск	ОГБОУ Т	9Б	39	28	ОГБОУ Т	11Б	50	16	NULL	84	72	1
13	г.Томск	МАОУ СС	9Б	32	23	МАОУ СС	11Б	44	9	15	71	45	1
14	г.Томск	МАОУ ги	9г	26	25	МАОУ ги	11Г	39	9	NULL	65	45	1
15	Кожевник	МАОУ Ко	9	32	15	МАОУ Ко	11	43	5	15	70	23	0
16	г.Томск	МОУ СОЦ	9Б	37	25	МАОУ СС	11А	52	14	NULL	90	68	1
17	г.Томск	ОГБОУ Т	9Б	30	22	ОГБОУ Т	11Б	54	12	NULL	95	59	1
18	г.Томск	МОУ СОЦ	9	31	19	МБОУ СС	11А	39	8	NULL	65	39	1
19	Кожевник	МБОУ Ос	9	40	15	МБОУ Ос	11	37	6	15	62	27	0
20	г.Томск	МАОУ СС	9В	33	29	МАОУ СС	11Б	47	12	20	76	59	1
21	г.Томск	МАОУ ги	9д	34	21	МАОУ ги	11Г	42	4	NULL	69	18	1
22	г.Томск	МОУ СОЦ	9Б	34	23	МАОУ СС	11Б	47	5	NULL	76	23	1
23	г.Томск	МАОУ СС	9Б	35	24	МАОУ СС	11Б	45	8	NULL	72	39	1
24	Бакчарск	МКОУ Ва	9	38	22	МКОУ Ва	11А	48	12	NULL	79	59	0
25	Бакчарск	МБОУ Ба	9Б	29	17	МБОУ Ба	11Б	42	8	NULL	69	39	0
26	Асиновск	МОУ СОЦ	9А	35	29	МБОУ СС	11А	34	7	18	59	33	0

Рис.3.1 Таблица данных о результатах экзаменов

На рис.3.2 представлена диаграмма средних и стандартных отклонений выборок по русскому языку результатов ОГЭ. Такие же диаграммы были построены для других результатов по схожим выборкам.

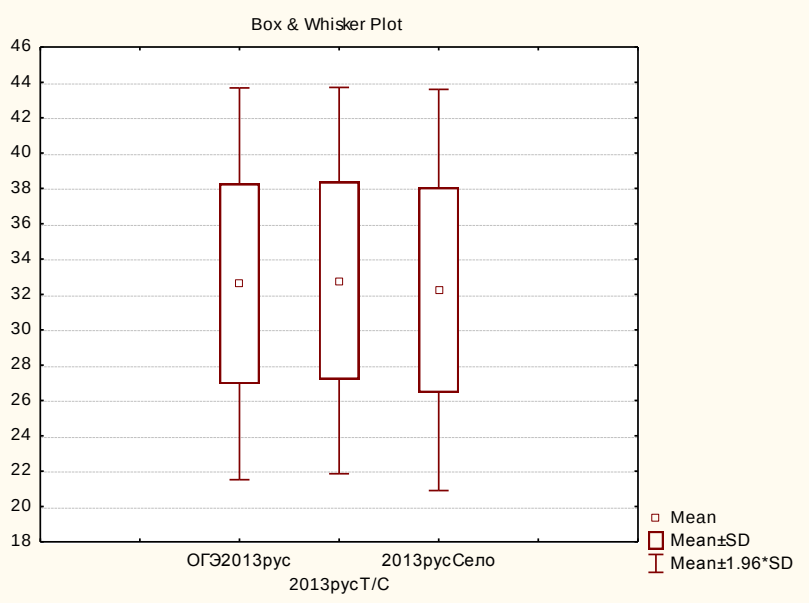


Рис. 3.2 Значения выборочных средних и стандартных отклонений за 2013 г. по русскому языку

	T-test for Independent Samples (9-11-Сайт-отредакт in Workbook1_9-11) Note: Variables were treated as independent samples									
Group 1 vs. Group 2	Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances
ОГЭ2013рус vs2013русТ/С	32,6027(	32,7856!	-1,2341(	6014	0,21719:	3632	2384	5,65538(	5,57591(	1,02870!	0,44980
ОГЭ2013рус vs. 2013русСело	32,6027(	32,2532:	1,87189	4878	0,06128:	3632	1248	5,65538(	5,79037(	1,04830!	0,30409(
2013русТ/С vs. 2013русСело	32,7856!	32,2532:	2,69698	363(	0,00702!	2384	1248	5,57591(	5,79037(	1,07840:	0,12407:

Рис.3.3 Результаты сравнения выборочных средних и выборочных дисперсий за 2013 г. по русскому языку

Обозначения применяемых на рис.3.3 обозначений:

Mean Group – среднее арифметическое выборки;

t-value – значение статистики (распределение Стьюдента);

df – число степеней свободы распределения Стьюдента;

p – вероятность того, что случайная величина примет значение большее чем t-value (двусторонний критерий);

Valid N – объем выборки;

Std.Dev. Group – стандартное отклонение выборки;

F-ratio Variances – значение отношений дисперсий двух выборок;

P variances – вероятность того, что случайная величина примет значение большее F.

При сравнении дисперсий выборок выдвигаются две гипотезы: нулевая – дисперсии выборок равны, альтернативная – дисперсии выборок различны. В случае с ОГЭ2013рус-2013русТ/С и ОГЭ2013рус-2013русСело принимается нулевая гипотеза о равенстве дисперсий (45%, 30% и 12% соответственно). Проверка t-критерий для переменных ОГЭ2013рус-2013русТ/С и ОГЭ2013рус-2013русСело свидетельствует, что с вероятность более 21% и 6% (что выше уровня значимости) принимается нулевая гипотеза об однородности выборочных средних, для ОГЭ2013русТ/С-2013русСело верна альтернативная гипотеза – средние оценки значимо различны.

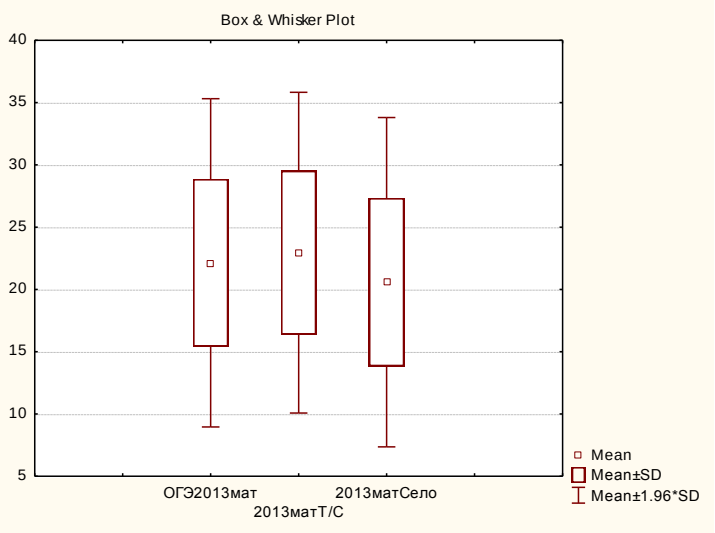


Рис. 3.4 Значения выборочных средних и стандартных отклонений за 2013 г. по математике

	T-test for Independent Samples (9-11-Сайт-отредакт in Workbook1_9-11) Note: Variables were treated as independent samples									
Group 1 vs. Group 2	Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances
ОГЭ2013мат vs. 2013матТ/С	22,1384	22,9517	-4,6315	6014	0,00000	3632	2384	6,72264	6,56762	1,04776	0,21249
ОГЭ2013мат vs. 2013матСело	22,1384	20,5849	7,0371	4878	0,00000	3632	1248	6,72264	6,74444	1,00649	0,88304
2013матТ/С vs. 2013матСело	22,9517	20,5849	10,2191	3630	0,00000	2384	1248	6,56762	6,74444	1,05457	0,27883

Рис.3.5 Результаты сравнения выборочных средних и выборочных дисперсий за 2013 г. математике

При рассмотрении результатов экзаменов ОГЭ за 2013 год по математике во всех случаях принимается гипотеза об однородности дисперсий, однако также во всех трех случаях отвергается нулевая гипотеза о равенстве выборочных средних в пользу альтернативной – выборочные средние различны.

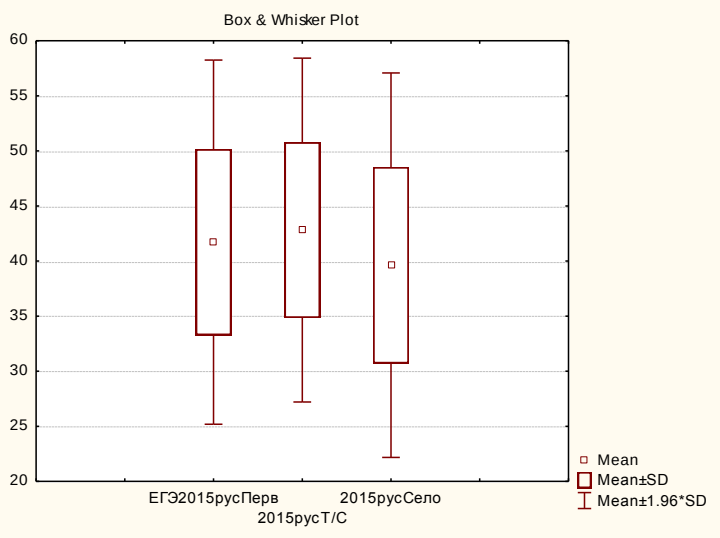


Рис. 3.6 Значения выборочных средних и стандартных отклонений за 2015 г. по русскому языку

	T-test for Independent Samples (9-11-Сайт-отредакт in Workbook1_9-11) Note: Variables were treated as independent samples										
Group 1 vs. Group 2	Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances
ЕГЭ2015русПерв vs. 2015русТ/С	41,7194	42,8154	-5,0399	6014	0,00000	3632	2384	8,43363	7,96221	1,12191	0,00214
ЕГЭ2015русПерв vs. 2015русСело	41,7194	39,6258	7,4574	4878	0,00000	3632	1248	8,43363	8,90339	1,11450	0,01812
2015русТ/С vs. 2015русСело	42,8154	39,6258	11,0021	3630	0,00000	2384	1248	7,96221	8,90339	1,25038	0,00000

Рис.3.7 Результаты сравнения выборочных средних и выборочных дисперсий за 2015 г. по русскому языку

Рассмотрев результаты сравнения выборочных средних и выборочных дисперсий за 2015 г. по русскому языку видно, что отвергаются нулевые гипотезы об однородности выборочных средних и дисперсий.

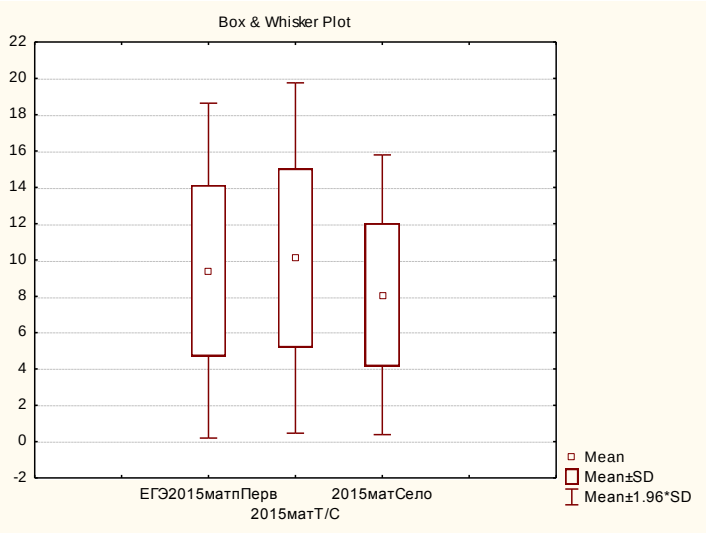


Рис.3.8 Значения выборочных средних и выборочных дисперсий за 2015 г. математике

T-test for Independent Samples (9-11-Саят-отредакт in Workbook1_9-11)
Note: Variables were treated as independent samples

Group 1 vs. Group 2	Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances
ЕГЭ2015матпПерв vs. 2015матТ/С	9,41382	10,1094	-5,5098	601	0,00000	363	238	4,70235	4,92036	1,09487	0,01465
ЕГЭ2015матпПерв vs. 2015матСел	9,41382	8,0849	8,9661	487	0,00000	363	124	4,70235	3,92787	1,43322	0,00000
2015матТ/С vs. 2015матСело	10,1094	8,0849	12,5868	363	0,00000	238	124	4,92036	3,92787	1,56920	0,00000

Рис.3.9 Результаты сравнения выборочных средних и выборочных дисперсий за 2015 г. математике

По рис. 3.9 видно, что результаты аналогичны результатам на рис. 3.8 и принимаются альтернативные гипотезы о разности дисперсий.

3.2 Построение рейтинга школ

Процесс создания рейтинга школ начинался с подготовки данных. Изначально все данные находились в одной таблице и для построения рейтинга учащиеся были отсортированы по школам. Школы с количеством учеников 4 и менее в расчет не брались. В итоге для построения рейтинга были отобраны 129 школ. Далее школы были упорядочены по среднему баллу ее учеников. Ниже приведены 15 лучших школ по всей области по русскому языку и математике среди 9 классов (2013г) и 11 классов (2015г).

Таблица 3.1 Лучшие школы по русскому языку (область)

Русс.яз 2013 г. 9 класс	Русс.яз 2015 г. 11 класс
НОУ гимназия Томь	МАОУ Малиновская СОШ Томского района
МАОУ Малиновская СОШ Томского района	МАОУ Сибирский лицей
МБОУ Пудовская СОШ (9кл-рус.яз)	МАОУ гимназия № 55
МБОУ "Чернореченская СОШ" Томского района	МАОУ гимназия № 24
МОУ СОШ №2 г.Асино Томской области.	МАОУ "СОШ № 80"
МАОУ гимназия №55 города Томска	МБОУ Парабельская СОШ
МБОУ Спасская СОШ Томского района	ОГБОУ Томский физико-технического лицей
ОГБОУ ТФТЛ	МАОУ СОШ № 65
МБОУ Северская гимназия	МБОУ Академический лицей
МОУ лицей №8 имени Н. Н. Рукавишникова	МБОУ Северский лицей
МБОУ Курлекская СОШ Томского района	МАОУ Гуманитарный лицей
МБОУ Северский лицей (9кл-рус.яз)	НОУ гимназия Томь
МБОУ Самусьский лицей	МБОУ Русская классическая гимназия № 2
МАОУ Сибирский лицей г. Томска	МБОУ Северская гимназия
МБОУ СОШ № 83	МБОУ Кожевниковская СОШ № 1

Таблица 3.2 Лучшие школы по математике (область)

Математика 2013 г. 9 класс	Математика 2015 г. 11 класс
МАОУ Малиновская СОШ Томского района	ОГБОУ Томский физико-технического лицей
МБОУ СОШ № 80	МАОУ гимназия № 55
ОГБОУ ТФТЛ	МАОУ Спасская СОШ Томского района
МБОУ Спасская СОШ Томского района	МАОУ "СОШ № 80"
МАОУ гимназия №55 города Томска	МБОУ СОШ № 49
МБОУ Северский лицей	МБОУ Академический лицей
НОУ гимназия Томь	МБОУ Северский лицей
МБОУ Северская гимназия	МАОУ Малиновская СОШ Томского района
МАОУ СОШ № 35 г. Томска	МАОУ гимназия № 29
МОУ лицей №8 имени Н. Н. Рукавишникова	МАОУ СФМЛ
МАОУ Северский физико-математический лицей	НОУ гимназия Томь
МАОУ Гуманитарный лицей	МБОУ Кожевниковская СОШ № 1
МБОУ Академический Лицей	МБОУ СОШ № 51
МАОУ гимназия №29	МБОУ Шегарская СОШ № 1
МБОУ СОШ № 49	МБОУ Курлекская СОШ Томского района

Исходя из приведенных выше результатов, можно выделить 5 школ, встречающихся во всех четырех списках: МАОУ Малиновская СОШ Томского района, ОГБОУ Томский физико-технический лицей, МБОУ Северский лицей, НОУ гимназия Томь и МАОУ гимназия №55.

Также подобные рейтинги были составлены для школ Томска/Северска и сельских школ области.

Таблица 3.3 Лучшие школы по русскому языку (Томск/Северск)

Русс.яз 2013 г. 9 класс	Русс.яз 2015 г. 11 класс
НОУ гимназия Томь	МАОУ Сибирский лицей

МАОУ гимназия №55 города Томска	МАОУ гимназия № 55
ОГБОУ ТФТЛ	МАОУ гимназия № 24
МБОУ Северская гимназия	МАОУ "СОШ № 80"
МОУ лицей №8 имени Н. Н. Рукавишникова	ОГБОУ ТФТЛ
МБОУ Северский лицей	МАОУ СОШ № 65
МБОУ Самусьский лицей	МБОУ Академический лицей
МАОУ Сибирский лицей г. Томска	МБОУ Северский лицей
МБОУ СОШ № 83	МАОУ Гуманитарный лицей
МАОУ СОШ № 37	НОУ гимназия Томь

Таблица 3.4 Лучшие школы по математике (Томск/Северск)

Математика 2013 г. 9 класс	Математика 2015 г. 11 класс
МБОУ СОШ № 80	ОГБОУ ТФТЛ
ОГБОУ ТФТЛ	МАОУ гимназия № 55
МАОУ гимназия №55 города Томска	МАОУ "СОШ № 80"
МБОУ Северский лицей	МБОУ СОШ № 49
НОУ гимназия Томь	МБОУ Академический лицей
МБОУ Северская гимназия	МБОУ Северский лицей
МАОУ СОШ № 35 г. Томска	МАОУ гимназия № 29
МОУ лицей №8 имени Н. Н. Рукавишникова	МАОУ СФМЛ
МАОУ СФМЛ	НОУ гимназия Томь
МАОУ Гуманитарный лицей	МБОУ СОШ № 51

Исходя из приведенных результатов, выделяются 4 школы, встречающиеся во всех четырех списках: МАОУ гимназия №55 города Томска, ОГБОУ Томский физико-технический лицей, МБОУ Северский лицей, НОУ гимназия Томь.

Таблица 3.5 Лучшие школы по русскому языку (сельские)

Русс.яз 2013 г. 9 класс	Русс.яз 2015 г. 11 класс
МАОУ Малиновская СОШ Томского района	МАОУ Малиновская СОШ
МБОУ Пудовская СОШ	МБОУ Парабельская СОШ
МБОУ "Чернореченская СОШ"	МБОУ Кожевниковская СОШ № 1
МОУ СОШ №2 г.Асино Томской области	МБОУ СОШ № 7 г. Колпашево
МБОУ Спасская СОШ Томского района	МБОУ Зональненская СОШ
МБОУ Курлекская СОШ Томского района	МБОУ Белоярская СОШ № 1
МБОУ Кисловская СОШ	МАОУ гимназия № 2 г. Асино

МБОУ Шегарская СОШ №1	МАОУ Молчановская СОШ № 2
МАОУ Кожевниковская СОШ №2	МБОУ Пудовская СОШ Кривошеинского р.
МАОУ Сергеевская СОШ Первомайского р.	МБОУ Парбигская СОШ Бакчарского р.

Таблица 3.6 Лучшие школы по математике (сельские)

Математика 2013 г. 9 класс	Математика 2015 г. 11 класс
МАОУ Малиновская СОШ Томского района	МАОУ Спасская СОШ Томского района
МБОУ Спасская СОШ Томского района	МАОУ Малиновская СОШ Томского р.
МБОУ Шегарская СОШ №1	МБОУ Кожевниковская СОШ № 1
МБОУ "Коломиногривская СОШ"	МБОУ Шегарская СОШ № 1
МБОУ "Нелюбинская СОШ" Томского района	МБОУ Курлекская СОШ Томского района
МБОУ Парабельская гимназия	МБОУ Парабельская гимназия
МАОУ СОШ №1 с Александровское	МБОУ Зональненская СОШ Томского р.
МБОУ Кожевниковская СОШ №1	МБОУ Молчановская СОШ № 1
МАОУ СОШ № 7	МАОУ СОШ № 2 г. Колпашево
МБОУ СОШ № 7 Колпашево	МБОУ Рассветовская СОШ Томского р.

Исходя из приведенных результатов, можно выделить 2 школы, встречающиеся во всех четырех списках: МАОУ Малиновская СОШ Томского района, МБОУ Кожевниковская СОШ №1.

3.3 Группирование данных по фактору смены места обучения и их анализ

Для проведения первичного анализа и однофакторного анализа была проведена работа с данными на выявление учеников сменивших и сохранивших место учебы после девятого класса, в результате чего были получены две выборки с 3046 и 535 наблюдениями. Количество учеников сохранивших место учебы после девятого класса равно 3046, а сменивших – 535.

Ниже представлена таблица, содержащая информацию о среднем балле по предмету учеников, сохранивших и сменивших место учебы, а также процентную составляющую этого балла от максимальной оценки.

Таблица 3.7 Средние баллы по категориям

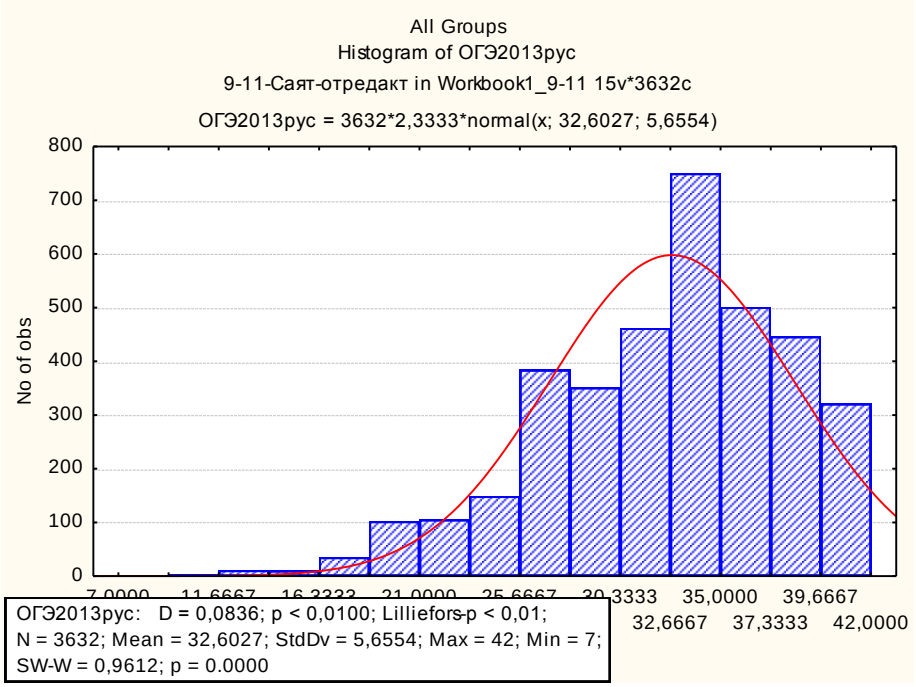
9 класс (2013 г.)		
Категория	Русс.яз.	Математика
Общее кол-во(3581)	32,62(77.67%)	22,37(58,87%)
Сохранившие место учебы(3046)	32,54(77,47%)	22,1(58,16%)
Изменившие место учебы(535)	33,06(78,71%)	23,89(62,87)
11 класс (2015 г.)		
Категория	Русс.яз.	Математика
Общее кол-во(3581)	41,75(74,55%)	9,43(31,43%)
Сохранившие место учебы(3046)	41,61(74,3%)	9,06(30,2%)
Изменившие место учебы(535)	42,53(75,95%)	11,56(38,53%)

В скобках указан процент среднего балла от максимального балла за экзамен. Максимальные баллы по экзаменам в девятом классе 2013 г. составляют 42 и 38 баллов по русскому языку и математике соответственно, а по экзаменам в одиннадцатом классе 2015 г. по русскому языку и математике 56 и 30 баллов соответственно.

4 ПРОВЕРКА ДАННЫХ НА НОРМАЛЬНОСТЬ РАСПРЕДЕЛЕНИЯ

4.1 Проверка исходных данных на нормальность распределения

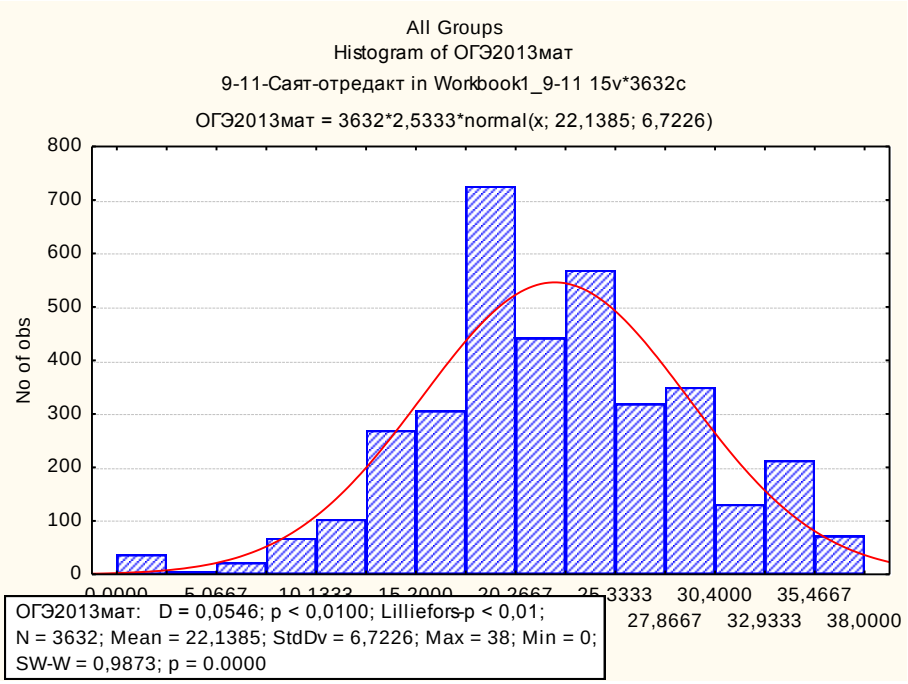
При проверке исходных данных на нормальность распределения проверяются каждая переменная на принадлежность ее наблюдений к нормальному распределению. Выдвигаются нулевая гипотеза о принадлежности переменной к нормальному распределению и альтернативная – переменная не относится к нормальному распределению. Результаты тестов представлены на рис. 4.1 – 4.4.



Frequency table: ОГЭ2013рус (9-11-Саят-отредакт in Workbook1_9
K-S d=,08362, p<,01 ; Lilliefors p<,01
Shapiro-Wilk W=,96122, p=0,0000

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
0,000000<x<=5,00000	0	0	0,0000(	0,000(	0,0000(	0,000(
5,000000<x<=10,0000	4	4	0,1101:	0,110:	0,1101:	0,110:
10,00000<x<=15,0000	17	21	0,4680(	0,578:	0,4680(	0,578:
15,00000<x<=20,0000	92	113	2,5330-	3,111:	2,5330-	3,111:
20,00000<x<=25,0000	304	417	8,3700-	11,481:	8,3700-	11,481:
25,00000<x<=30,0000	735	1152	20,2367(	31,718:	20,2367(	31,718:
30,00000<x<=35,0000	1212	2364	33,3700-	65,088:	33,3700-	65,088:
35,00000<x<=40,0000	1106	3470	30,4515-	95,539(	30,4515-	95,539(
40,00000<x<=45,0000	162	3632	4,4603:	100,000(	4,4603:	100,000(
Missing	0	3632	0,0000(		0,0000(	100,000(

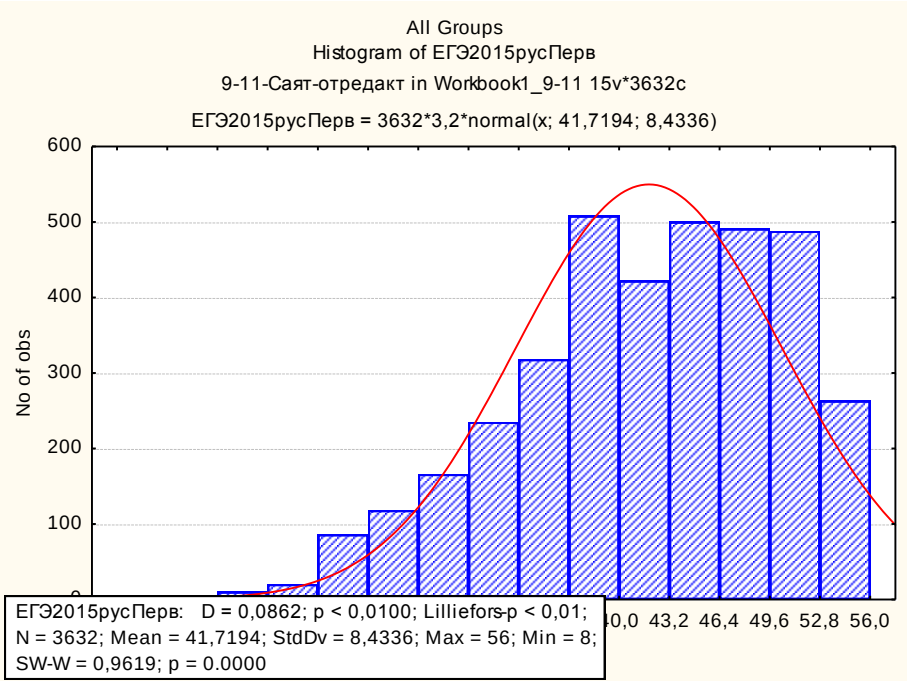
Рис. 4.1 Результаты проверки данных по русскому языку за 2013 г.



Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
Frequency table: ОГЭ2013мат (9-11-Саят-отредакт in Workbook1_9						
K-S d=,05457, p<,01 ; Lilliefors p<,01						
Shapiro-Wilk W=,98733, p=,00000						
-5,00000<x<=0,000000	37	37	1,0187	1,0187	1,0187	1,0187
0,000000<x<=5,00000	5	42	0,1376	1,1564	0,1376	1,1564
5,000000<x<=10,0000	90	132	2,4779	3,6344	2,4779	3,6344
10,00000<x<=15,0000	372	504	10,2422	13,876	10,2422	13,876
15,00000<x<=20,0000	1032	1536	28,4141	42,290	28,4141	42,290
20,00000<x<=25,0000	1011	2547	27,8359	70,126	27,8359	70,126
25,00000<x<=30,0000	669	3216	18,4196	88,546	18,4196	88,546
30,00000<x<=35,0000	344	3560	9,4713	98,017	9,4713	98,017
35,00000<x<=40,0000	72	3632	1,9823	100,000	1,9823	100,000
Missing	0	3632	0,0000		0,0000	100,000

Upper Boundary	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
Variable: ОГЭ2013мат, Distribution: Normal (9-11-Саят-отредакт in Workbook1_9-11)									
Chi-Square = 108.29143, df = 6 (adjusted) , p = 0.00000									
<= 0.00000	37	37	1,0187	1,0187	1,799	1,799	0,0495	0,0495	35,201
5.00000	5	42	0,1376	1,1564	17,799	19,598	0,4900	0,5396	-12,799
10.00000	90	132	2,4779	3,6344	109,30	128,89	3,0093	3,5489	-19,300
15.00000	372	504	10,2422	13,876	394,65	523,55	10,8660	14,415	-22,656
20.00000	1032	1536	28,4141	42,290	839,18	1362,73	23,1053	37,520	192,814
25.00000	1011	2547	27,8359	70,126	1051,88	2414,62	28,9615	66,481	-40,884
30.00000	669	3216	18,4196	88,546	777,46	3192,09	21,4060	87,888	-108,469
35.00000	344	3560	9,4713	98,017	338,70	3530,79	9,3256	97,213	5,294
40.00000	72	3632	1,9823	100,000	86,88	3617,67	2,3920	99,605	-14,880
< Infinity	0	3632	0,0000	100,000	14,321	3632,00	0,3943	100,000	-14,321

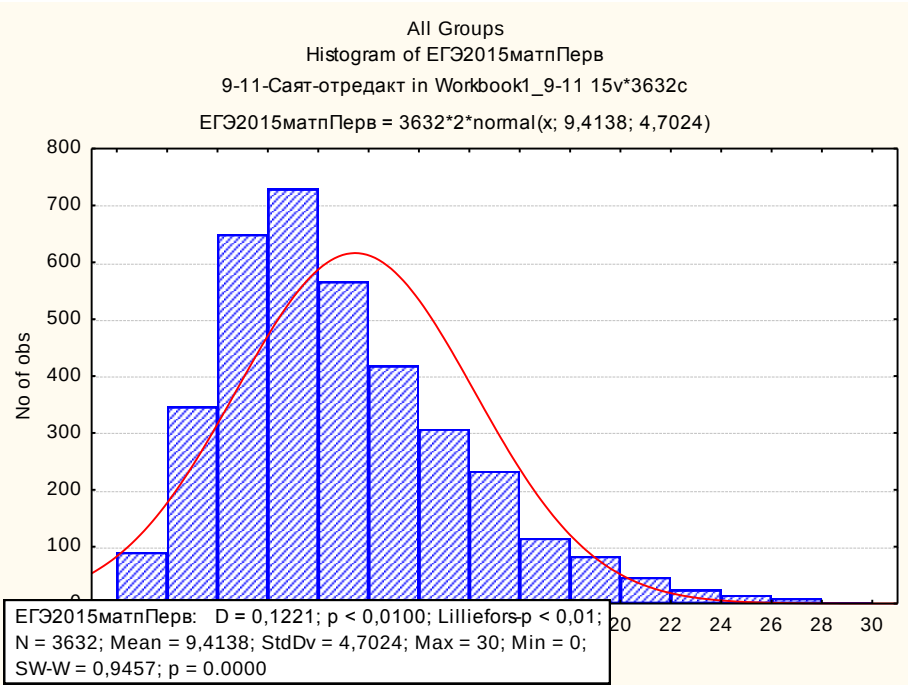
Рис. 4.2 Результаты проверки данных по математике за 2013 г.



Frequency table: ЕГЭ2015русПерв (9-11-Саят-отредакт in Workbook1_9
K-S d=,08620, p<,01 ; Lilliefors p<,01
Shapiro-Wilk W=,96186, p=0,0000

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
0,000000<x<=10,0000	4	4	0,1101	0,110	0,1101	0,110
10,00000<x<=20,0000	33	37	0,9085	1,018	0,9085	1,018
20,00000<x<=30,0000	370	407	10,1872	11,205	10,1872	11,205
30,00000<x<=40,0000	1061	1468	29,2125	40,418	29,2125	40,418
40,00000<x<=50,0000	1596	3064	43,9427	84,361	43,9427	84,361
50,00000<x<=60,0000	568	3632	15,6387	100,000	15,6387	100,000
Missing	0	3632	0,0000		0,0000	100,000

Рис. 4.3 Результаты проверки данных по русскому языку за 2015 г.



Variable: ЕГЭ2015матпПерв, Distribution: Normal (9-11-Саят-отредакт in Workbook1_9-11)
Chi-Square = 598.11895, df = 11 (adjusted) , p = 0.00000

Upper Boundary	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= -2.00000	0	0	0,0000	0,000	27,627	27,628	0,7606	0,760	-27,628
0.00000	4	4	0,1101	0,110	54,621	82,249	1,5038	2,264	-50,621
2.00000	86	90	2,3678	2,478	126,378	208,62	3,4795	5,744	-40,378
4.00000	346	436	9,5264	12,004	244,661	453,28	6,7362	12,480	101,339
6.00000	649	1085	17,8689	29,873	396,326	849,61	10,9120	23,392	252,673
8.00000	729	1814	20,0715	49,944	537,212	1386,82	14,7910	38,183	191,788
10.00000	566	2380	15,5837	65,528	609,327	1996,15	16,7766	54,960	-43,328
12.00000	418	2798	11,5088	77,037	578,322	2574,47	15,9229	70,883	-160,322
14.00000	307	3105	8,4526	85,490	459,306	3033,78	12,6461	83,529	-152,306
16.00000	233	3338	6,4152	91,905	305,241	3339,02	8,4042	91,933	-72,241
18.00000	115	3453	3,1663	95,071	169,740	3508,76	4,6734	96,607	-54,740
20.00000	83	3536	2,2852	97,356	78,979	3587,74	2,1745	98,781	4,021
22.00000	46	3582	1,2665	98,623	30,747	3618,49	0,8465	99,628	15,252

Frequency table: ЕГЭ2015матпПерв (9-11-Саят-отредакт in Workbook1_9
K-S d=,12207, p<,01 ; Lilliefors p<,01
Shapiro-Wilk W=,94572, p=0,0000

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
-5,00000<x<=0,00000	4	4	0,1101	0,1101	0,1101	0,1101
0,000000<x<=5,00000	744	748	20,4845	20,5947	20,4845	20,5947
5,000000<x<=10,0000	1632	2380	44,9339	65,5286	44,9339	65,5286
10,00000<x<=15,0000	854	3234	23,5132	89,0419	23,5132	89,0419
15,00000<x<=20,0000	302	3536	8,3149	97,3568	8,3149	97,3568
20,00000<x<=25,0000	77	3613	2,1200	99,4769	2,1200	99,4769
25,00000<x<=30,0000	19	3632	0,5231	100,000	0,5231	100,000
Missing	0	3632	0,0000		0,0000	100,000

Рис. 4.4 Результаты проверки данных по математике за 2015 г.

При рассмотрении результатов проверки исходных данных видно, что в каждом из рассматриваемых случаев нулевая гипотеза о принадлежности выборки к Гауссовому распределению может быть принята с вероятностью много меньше, чем 5%(уровень значимости), из чего следует ни одна из переменных (результатов экзаменов) не распределена по нормальному закону и соответственно принимается альтернативная гипотеза.

4.2 Проверка данных на принадлежность к одной генеральной совокупности

Гипотезы об однородности выборок – это гипотезы о том, что изучаемые выборки принадлежат одной и той же генеральной совокупности. В данной работе выборками являются результаты экзаменов учеников Томска/Северска и Томской области по двум дисциплинам (русский язык и математика) за 9 класс 2013 г. и за 11 класс 2015 г.

4.2.1 Тест Колмогорова-Смирнова

В критерий Колмогорова-Смирнова используется та же идея, что используется в критерий Колмогорова. Однако в критерии Колмогорова

эмпирическая функция распределения сравнивается с теоретической, а в критерии Колмогорова-Смирнова идет сравнение двух эмпирический функции распределения.

Статистика критерия Колмогорова-Смирнова имеет вид:

$$\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max \left| F_{n_1}(x) - F_{n_2}(x) \right|, \qquad (4.1)$$

где $F_{n_1}(x)$ и $F_{n_2}(x)$ – эмпирические функции распределения, построенные по двум выборкам с объемами n_1 и n_2.

Для теста Колмогорова-Смирнова (далее КС) был использовать модуль Statistics-Nonparametrics и процедура этого модуля *Comparing two independent samples* (*groups*). Для анализа исходные данные разбивались по фактору принадлежности школы к Томску/Северску или областным не городским школам. В итоге в количестве 2384 человек количество учащихся насчитывается в школах Томска/Северска, а 1248 – в областных не городских школ.

	Kolmogorov-Smirnov Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000								
variable	Max Neg Differnc	Max Pos Differnc	p-level	Mean Group 1	Mean Group 2	Std.Dev. Group 1	Std.Dev. Group 2	Valid N Group 1	Valid N Group 2
ОГЭ2013рус	-0,00087	0,04678	p < .10	32,7856	32,2532	5,57591	5,79037	2384	1248

Рис.4.5 Тест КС для результатов по русскому языку за 9 класс 2013г

	Kolmogorov-Smirnov Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000								
variable	Max Neg Differnc	Max Pos Differnc	p-level	Mean Group 1	Mean Group 2	Std.Dev. Group 1	Std.Dev. Group 2	Valid N Group 1	Valid N Group 2
ОГЭ2013мат	0,00	0,14381	p < .001	22,9517	20,5849	6,56762	6,74444	2384	1248

Рис.4.6 Тест КС для результатов по математике за 9 класс 2013г

	Kolmogorov-Smirnov Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000								
variable	Max Neg Differnc	Max Pos Differnc	p-level	Mean Group 1	Mean Group 2	Std.Dev. Group 1	Std.Dev. Group 2	Valid N Group 1	Valid N Group 2
ЕГЭ2015русПерв	0,00	0,14545	p < .001	42,8154	39,6258	7,96221	8,90339	2384	1248

Рис.4.7 Тест КС для результатов по русскому языку за 11 класс 2015г

variable	Kolmogorov-Smirnov Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000 Max Neg Differnc	Max Pos Differnc	p-level	Mean Group 1	Mean Group 2	Std.Dev. Group 1	Std.Dev. Group 2	Valid N Group 1	Valid N Group 2
ЕГЭ2015матпере	0,00	0,17662	p < .001	10,1094	8,08493	4,92036	3,92787	238	124

Рис.4.8 Тест КС для результатов по математике за 11 класс 2015г

Для теста КС использовались выборки из результатов экзаменов по предметам для школьников Томска/Северска (1) и сельских школ(0). Как видно из результатов тест КС показывает, что только в случае экзамена за 9 класс по русскому языку есть вероятности однородности выборок.

4.2.2 Тест Манна-Уитни

U-критерий Манна — Уитни (англ. *Mann — Whitney U-test*) — статистический критерий, который используемся при оценке различий между двумя независимыми выборками по уровню какого-либо признака, измеренного количественно[2]. Статистика критерия выглядит следующим образом:

$$U = W - \frac{1}{2}m(m+1) = \sum_{i-1}^{n} \sum_{j-1}^{m} \delta_{ij} \qquad (4.2)$$

где W – статистики Вилкоксона, предназначенная для проверки этой же гипотезы H_0.

Для теста Манна-Уитни (далее МУ) был использовать модуль Statistics-Nonparametrics и процедура этого модуля *Comparing two independent samples* (*groups*).

variable	Mann-Whitney U Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000 Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level	Z adjusted	p-level	Valid N Group 1	Valid N Group 2
ОГЭ2013рус	440663	219089	141151	2,53569	0,01122	2,53968	0,01109	238	124

Рис.4.9 Тест МУ для результатов по русскому языку за 9 класс 2013г

variable	Mann-Whitney U Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000 Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level	Z adjusted	p-level	Valid N Group 1	Valid N Group 2
ОГЭ2013мат	460908	198844	120906	9,28102	0,00000	9,29190	0,00000	238	124

Рис.4.10 Тест МУ для результатов по математике за 9 класс 2013г

variable	Mann-Whitney U Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000 Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level	Z adjusted	p-level	Valid N Group 1	Valid N Group 2
ЕГЭ2015русПерв	463838	195914	117976	10,2573	0,00000	10,2643	0,00000	2384	1248

Рис.4.11 Тест МУ для результатов по русскому языку за 11 класс 2015г

variable	Mann-Whitney U Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000 Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level	Z adjusted	p-level	Valid N Group 1	Valid N Group 2
ЕГЭ2015матпПерв	468582	191170	113232	11,8381	0,00	11,8700	0,00	2384	1248

Рис.4.12 Тест МУ для результатов по математике за 11 класс 2015г

Тест МУ отвергает гипотезу об однородности выборок в каждом из представленных результатов. Если рассмотреть результаты за 2013 год по русскому языку можно заметить, что вероятность принадлежности выборок к единой выше, чем в других случаях, но все же значительно меньше уровня значимости.

4.2.3 Тест Уальда-Вольфовица

При тесте Уальда-Вольфовица значения обеих групп выстраиваются по присуждаемым им рангам в единую последовательность. Далее происходит подсчёт числа смен группового признака, который поможет найти количество непрерывных последовательностей. При появлении одинаковых значений (ранговых связок) выделяются значения максимального и минимального числа возможных непрерывных последовательностей. По количеству непрерывных последовательностей находится вероятность ошибки p. В случае с переменными с маленьким числом категории тест становится не пригодным, т.к. сильно растет количество одинаковых значений, т.е. ранговых связок.

Для теста Уальда-Вольфовица (далее УВ) был использовать модуль Statistics-Nonparametrics и процедура этого модуля *Comparing two independent samples* (*groups*).

Variable	Wald-Wolfowitz Runs Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000 Valid N Group 1	Valid N Group 2	Mean Group 1	Mean Group 2	Z	p-level	Z adjstd	p-level	No. of Runs	No. of ties
ОГЭ2013рус	2384	1248	32,7856	32,2532	-0,71166	0,47667	0,69327	0,48813	1620	1607

Рис.4.13 Тест УВ для результатов по русскому языку за 9 класс 2013г

Variable	Wald-Wolfowitz Runs Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000 Valid N Group 1	Valid N Group 2	Mean Group 1	Mean Group 2	Z	p-level	Z adjstd	p-level	No. of Runs	No. of ties
ОГЭ2013мат	2384	1248	22,9517	20,5849	-0,85883	0,39043	0,84043	0,40066	1616	1601

Рис.4.14 Тест УВ для результатов по математике за 9 класс 2013г

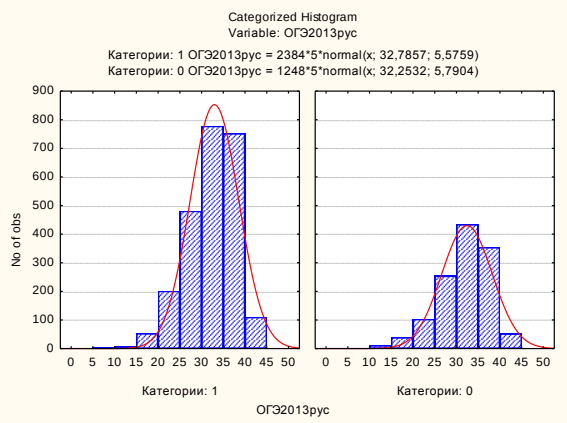
Variable	Wald-Wolfowitz Runs Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000 Valid N Group 1	Valid N Group 2	Mean Group 1	Mean Group 2	Z	p-level	Z adjstd	p-level	No. of Runs	No. of ties
ЕГЭ2015русПерв	2384	1248	42,8154	39,6258	-3,3606	0,00077	3,34222	0,00083	1548	1531

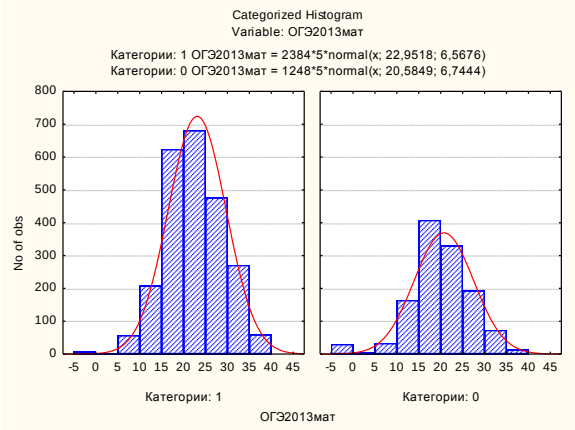
Рис.4.15 Тест УВ для результатов по русскому языку за 11 класс 2015г

Variable	Wald-Wolfowitz Runs Test (9-11-Саят-отредакт in Workbook1_9-11) By variable Категории Marked tests are significant at p <,05000 Valid N Group 1	Valid N Group 2	Mean Group 1	Mean Group 2	Z	p-level	Z adjstd	p-level	No. of Runs	No. of ties
ЕГЭ2015матпПерв	2384	1248	10,1094	8,08493	-2,8823	0,00394	2,86393	0,00418	1561	1555

Рис.4.16 Тест УВ для результатов по математике за 11 класс 2015г

В результатах теста Уальда-Вольфовица гипотеза об однородности принимается для баллов по русскому языку и математики за 2013 год с высокой вероятностью(рис. 4.13 и рис. 4.14).





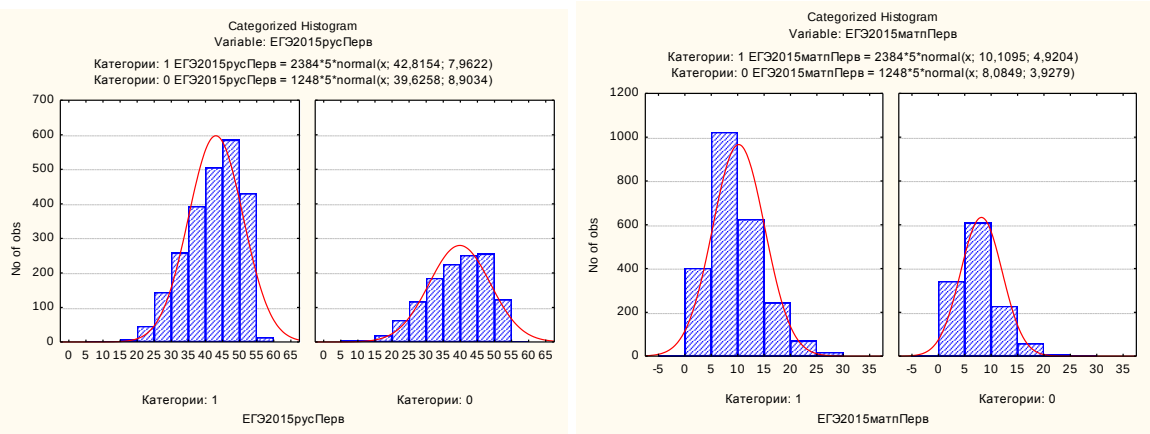


Рис.4.17 Гистограммы результатов проверки на однородность

На рисунке 4.17 представлены гистограммы проверки на нормальность выборок, проверенных выше на однородность распределения.

5 КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Корреляционный анализа вызывает колоссальный интерес, т.к. предоставляет возможность обнаружить взаимные связи между случайными величинами. К примеру, можно выявить наличие связи между увеличением продаж алкогольной продукции и увеличением дорожно-транспортных происшествий, между локацией учреждения и его работоспособностью и т.п.

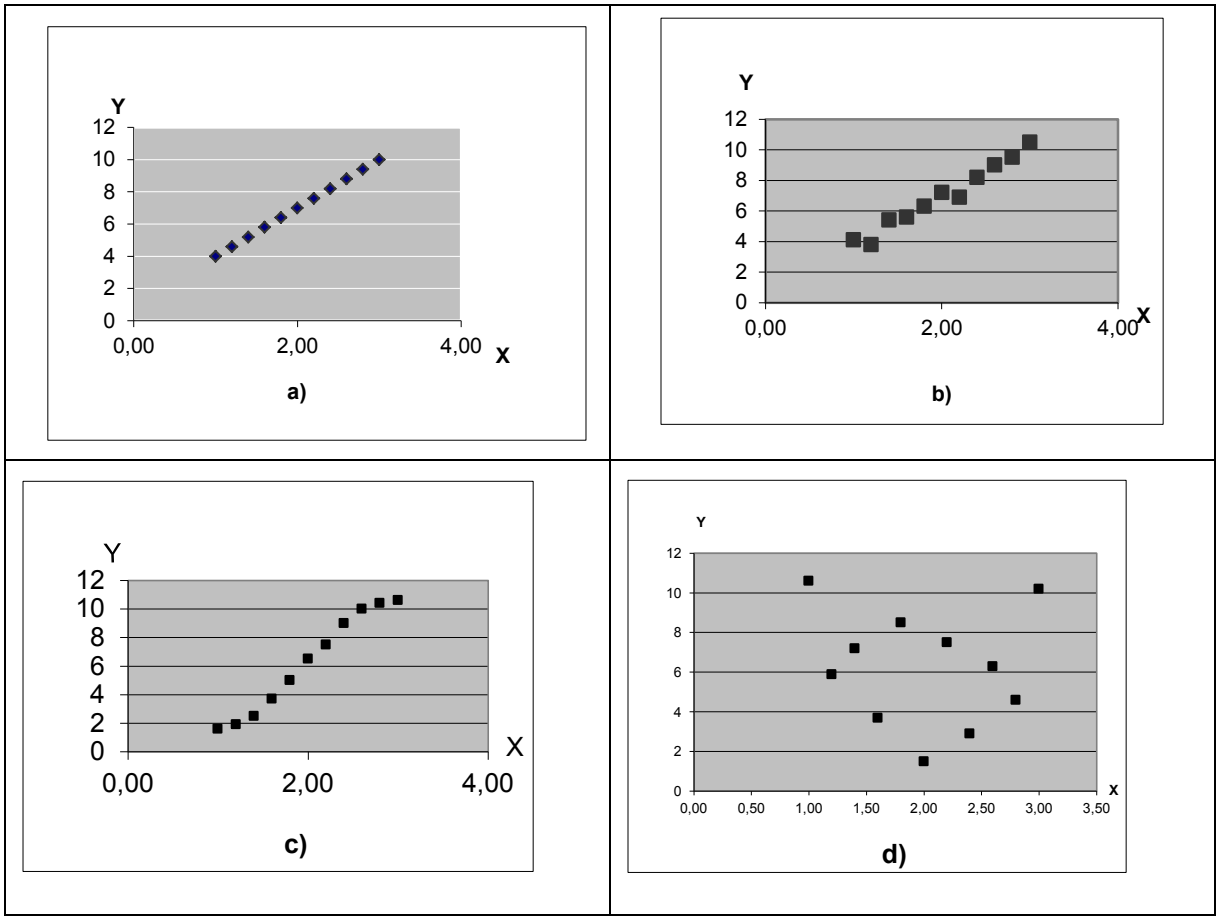


Рис. 5.1. Различные степени корреляции: **a)** – точная линейная корреляция; **b)** – умеренная линейная корреляция; **c)** – нелинейная корреляция; **d)** – отсутствие корреляции.

5.1 Коэффициент корреляции Пирсона

Для оценки наличия или отсутствия линейной связи между случайными величинами рассчитывается коэффициент корреляции Пирсона. Также коэффициент Пирсона очень хорошо определяет силу этой связи, поэтому он также носит название **коэффициент линейной корреляции Пирсона.**

Коэффициент корреляции можно оценить по выборочным данным следующим образом:

$$r = \rho_{xy} = \frac{S_{xy}}{S_x S_y}. \qquad (5.1)$$

где x и y случайные величины.

	Correlations (9-11-Саят-отредакт in Workbook1_9-11) Marked correlations are significant at p < ,05000 N=3632 (Casewise deletion of missing data)			
Variable	ОГЭ2013рус	ОГЭ2013мат	ЕГЭ2015русПерв	ЕГЭ2015матпПерв
ОГЭ2013рус	1,000(	,478(	,606(	,4357
	p= ---	p=0,0(	p=0,0(	p=0,0(
ОГЭ2013мат	,478(	1,000(	,537(	,641(
	p=0,0(	p= ---	p=0,0(	p=0,0(
ЕГЭ2015русПерв	,606(	,537(	1,000(	,563(
	p=0,0(	p=0,0(	p= ---	p=0,0(
ЕГЭ2015матпПерв	,4357	,641(	,563(	1,000(
	p=0,0(	p=0,0(	p=0,0(	p= ---

Рис.5.2 Линейный корреляционный анализ данных по школам области

	Correlations (9-11-Саят-отредакт in Workbook1_9- Marked correlations are significant at p < ,05000 N=2384 (Casewise deletion of missing data)			
Variable	2013 русТ/С	2013матТ/С	2015русТ/С	2015матТ/С
2013русТ/С	1,000(	,505(	,633(	,453(
	p= ---	p=0,0(	p=0,0(	p=0,0(
2013матТ/С	,505(	1,000(	,543(	,661(
	p=0,0(	p= ---	p=0,0(	p=0,0(
2015русТ/С	,633(	,543(	1,000(	,548(
	p=0,0(	p=0,0(	p= ---	p=0,0(
2015матТ/С	,453(	,661(	,548(	1,000(
	p=0,0(	p=0,0(	p=0,0(	p= ---

Рис.5.3 Линейный корреляционный анализ данных по школам Томска/Северска

	Correlations (9-11-Саят-отредакт in Workbook1_9-11) Marked correlations are significant at p < ,05000 N=1248 (Casewise deletion of missing data)			
Variable	2013русСело	2013матСело	2015русСело	2015матСело
2013русСело	1,000(	,428(	,570(	,407(
	p= ---	p=0,0(	p=0,0(	p=0,0(
2013матСело	,428(	1,000(	,491(	,568(
	p=0,0(	p= ---	p=0,0(	p=0,0(
2015русСело	,570(	,491(	1,000(	,564(
	p=0,0(	p=0,0(	p= ---	p=0,0(
2015матСело	,407(	,568(	,564(	1,000(
	p=0,0(	p=0,0(	p=0,0(	p= ---

Рис.5.4 Линейный корреляционный анализ данных по сельским школам

	КК Пирсона (НеПерешедшие3046) Уровень значимости p <,05000 N=3046			
	ОГЭ2013рус	ОГЭ2013мат	ЕГЭ2015русПерв	ЕГЭ2015матпПерв
ОГЭ2013рус	1,0000	,4884	,6105	,4443
	p= ---	p=0,00	p=0,00	p=0,00
ОГЭ2013мат	,4884	1,0000	,5552	,6825
	p=0,00	p= ---	p=0,00	p=0,00
ЕГЭ2015русПерв	,6105	,5552	1,0000	,5691
	p=0,00	p=0,00	p= ---	p=0,00
ЕГЭ2015матпПерв	,4443	,6825	,5691	1,0000
	p=0,00	p=0,00	p=0,00	p= ---

Рис. 5.5 Линейный корреляционный анализ данных результатов экзаменов учеников сохранивших место учебы

	КК Пирсона (Перешедшие535) Уровень значимости p <,05000 N=535			
	ОГЭ2013рус	ОГЭ2013мат	ЕГЭ2015русПерв	ЕГЭ2015матпПерв
ОГЭ2013рус	1,0000	,5692	,5770	,4474
	p= ---	p=0,00	p=0,00	p=0,00
ОГЭ2013мат	,5692	1,0000	,6052	,7019
	p=0,00	p= ---	p=0,00	p=0,00
ЕГЭ2015русПерв	,5770	,6052	1,0000	,5888
	p=0,00	p=0,00	p= ---	p=0,00
ЕГЭ2015матпПерв	,4474	,7019	,5888	1,0000
	p=0,00	p=0,00	p=0,00	p= ---

Рис. 5.6 Линейный корреляционный анализ данных результатов экзаменов учеников сменивших место учебы

5.2 Коэффициент корреляции Кенделла

Коэффициент ранговой корреляции Кендалла вычисляется по формуле:

$$\tau = 1 - \frac{4K}{n(n-1)}, \qquad (5.2)$$

где K – статистика Кендалла.

	Kendall Tau Correlations (9-11-Саят-отредакт in Workbook1_9-1 MD pairwise deleted Marked correlations are significant at p <,05000			
Variable	ОГЭ2013рус	ОГЭ2013мат	ЕГЭ2015русПерв	ЕГЭ2015матпПерв
ОГЭ2013рус	1,00000	0,36310	0,45192	0,33461
ОГЭ2013мат	0,36310	1,00000	0,41656	0,51320
ЕГЭ2015русПерв	0,45192	0,41656	1,00000	0,44016
ЕГЭ2015матпПерв	0,33461	0,51320	0,44016	1,00000

Рис.5.7 Ранговый корреляционный анализ Кенделла данных по школам области

Variable	Kendall Tau Correlations (9-11-Саят-отредакт in ' MD pairwise deleted Marked correlations are significant at p <,05000			
	2013русТ/С	2013матТ/С	2015русТ/С	2015матТ/С
2013русТ/С	1,00000(	0,37951(	0,47404(	0,35358:
2013матТ/С	0,37951(	1,00000(	0,41438:	0,52206.
2015русТ/С	0,47404(	0,41438:	1,00000(	0,42824(
2015матТ/С	0,35358:	0,52206.	0,42824(	1,00000(

Рис.5.8 Ранговый корреляционный анализ Кенделла данных по школам Томска/Северска

Variable	Kendall Tau Correlations (9-11-Саят-отредакт in Workboc MD pairwise deleted Marked correlations are significant at p <,05000			
	2013русСелс	2013матСелс	2015русСелс	2015матСелс
2013русСело	1,00000(	0,33043:	0,41178:	0,29788'
2013матСело	0,33043:	1,00000(	0,39221(	0,46513:
2015русСело	0,41178:	0,39221(	1,00000(	0,42816'
2015матСело	0,29788'	0,46513:	0,42816'	1,00000(

Рис.5.9 Ранговый корреляционный анализ Кенделла данных по сельским школам

	КК Тау Кендалла (НеПерешедшие3046) Уровень значимости p <,05000			
	ОГЭ2013рус	ОГЭ2013мат	ЕГЭ2015русПерв	ЕГЭ2015матпПерв
ОГЭ2013рус	1,000000	0,358469	0,454944	0,333226
ОГЭ2013мат	0,358469	1,000000	0,414145	0,516114
ЕГЭ2015русПерв	0,454944	0,414145	1,000000	0,434346
ЕГЭ2015матпПерв	0,333226	0,516114	0,434346	1,000000

Рис.5.10 Ранговый корреляционный анализ Кенделла данных результатов экзаменов учеников сохранивших место учебы

	КК Тау Кендалла (Перешедшие535) Уровень значимости p <,05000			
	ОГЭ2013рус	ОГЭ2013мат	ЕГЭ2015русПерв	ЕГЭ2015матпПерв
ОГЭ2013рус	1,000000	0,409754	0,427663	0,347256
ОГЭ2013мат	0,409754	1,000000	0,456462	0,548927
ЕГЭ2015русПерв	0,427663	0,456462	1,000000	0,466493
ЕГЭ2015матпПерв	0,347256	0,548927	0,466493	1,000000

Рис.5.11 Ранговый корреляционный анализ Кенделла данных результатов экзаменов учеников сменивших место учебы

5.3 Коэффициент корреляции Спирмена

Коэффициент ранговой корреляции Спирмена находится по формуле:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(r_i - s_i)^2}{n^3 - n} \qquad (5.4)$$

где r_i и s_i - ранги i-го объекта по переменным X и Y;

n – количество пар наблюдений.

Variable	Spearman Rank Order Correlations (9-11-Саят-отредакт in Workbook1_9 MD pairwise deleted Marked correlations are significant at p <,05000			
	ОГЭ2013рус	ОГЭ2013мат	ЕГЭ2015русПерв	ЕГЭ2015матпПерв
ОГЭ2013рус	1,00000	0,50177	0,61143	0,46041
ОГЭ2013мат	0,50177	1,00000	0,57540	0,67644
ЕГЭ2015русПерв	0,61143	0,57540	1,00000	0,59793
ЕГЭ2015матпПерв	0,46041	0,67644	0,59793	1,00000

Рис.5.12 Ранговый корреляционный анализ Спирмена данных по школам области

Variable	Spearman Rank Order Correlations (9-11-Саят-отр MD pairwise deleted Marked correlations are significant at p <,05000			
	2013русТ/С	2013матТ/С	2015русТ/С	2015матТ/С
2013русТ/С	1,00000	0,52296	0,63889	0,48612
2013матТ/С	0,52296	1,00000	0,57250	0,68896
2015русТ/С	0,63889	0,57250	1,00000	0,58403
2015матТ/С	0,48612	0,68896	0,58403	1,00000

Рис.5.13 Ранговый корреляционный анализ Спирмена данных по школам Томска/Северска

Variable	Spearman Rank Order Correlations (9-11-Саят-отредакт i MD pairwise deleted Marked correlations are significant at p <,05000			
	2013русСелс	2013матСелс	2015русСелс	2015матСелс
2013русСело	1,00000	0,45762	0,56037	0,40869
2013матСело	0,45762	1,00000	0,54303	0,61190
2015русСело	0,56037	0,54303	1,00000	0,58156
2015матСело	0,40869	0,61190	0,58156	1,00000

Рис.5.14 Ранговый корреляционный анализ Спирмена данных по сельским школам

	КК Спирмена (НеПерешедшие3046) Уровень значимости p <,05000			
	ОГЭ2013рус	ОГЭ2013мат	ЕГЭ2015русПерв	ЕГЭ2015матпПерв
ОГЭ2013рус	1,000000	0,495438	0,615108	0,457050
ОГЭ2013мат	0,495438	1,000000	0,572157	0,680781
ЕГЭ2015русПерв	0,615108	0,572157	1,000000	0,588677
ЕГЭ2015матпПерв	0,457050	0,680781	0,588677	1,000000

Рис.5.15 Ранговый корреляционный анализ Спирмена данных результатов экзаменов учеников сохранивших место учебы

	КК Спирмена (Перешедшие535) Уровень значимости p <,05000			
	ОГЭ2013рус	ОГЭ2013мат	ЕГЭ2015русПерв	ЕГЭ2015матпПерв
ОГЭ2013рус	1,000000	0,562159	0,578405	0,478999
ОГЭ2013мат	0,562159	1,000000	0,628469	0,722513
ЕГЭ2015русПерв	0,578405	0,628469	1,000000	0,636075
ЕГЭ2015матпПерв	0,478999	0,722513	0,636075	1,000000

Рис.5.16 Ранговый корреляционный анализ Спирмена данных результатов экзаменов учеников сменивших место учебы

Проанализировав результаты корреляционного анализа по всем случаям было замечено, что помимо естественной зависимости между результатами по одному предмету также коррелируют результаты экзаменов по русскому языку за 2015 год с математикой за 2015 год и результаты по математике за 2013 год с результатами по русскому языку за 2015 год. Также был проведен корреляционный анализ данных по ученикам сменивших и сохранивших место учебы после девятого класса, который показал более сильную корреляцию среди учеников, сменивших место учебы кроме случая с русским языком в 9ом и 11ом классах: здесь уровень коррелированности ниже. Исходя из данного анализа, можно предположить, что учащиеся, хорошо знающие математику, успешнее напишут экзамен и по русскому языку.

6 ОДНОФАКТОРНЫЙ АНАЛИЗ

Задачами однофакторного анализа являются задачи исследования зависимости при одном факторе, который влияет на конечный результат и он принимает только конечное число уровней.

Фактором или факторами называют то, что по нашему мнению должно оказывать влияние на конечный результат.

Уровень фактора или способ обработки – это конкретная реализация фактора (например, определённый школьный учебник, или выбранное лекарство)[2].

Отклик - значение измеряемого признака (т.е. величина результата).

6.1 Ранговый однофакторный анализ

Выдвигаем нулевую гипотезу H_0 о принадлежности откликов (баллов) к одному и тому же распределению. То есть влияние фактора смены или сохранения места учебы не существенно. Для проверки данных гипотез данные были отсортированы на учеников сменивших школу после 9ого класса и на учеников не сменивших школу. Количество сменивших школу учеников 535(1), не сменивших – 3046(0).

Критерий Краскела-Уоллиса

	Ранговый ДА Краскела-Уоллиса; ОГЭ2013ру Груп. (независ.) переменная: КатегПоПерехо Кр.Краскела-Уоллиса: H (1, N= 3581) =4,690			
ОГЭ2013рус	Код	N	Сумма Рангов	Средний Ранг
0	0	304€	540769	1775,34
1	1	535	100587	1880,14

Рис. 6.1 Результаты теста Краскела-Уоллиса

для русского языка 2013 г. 9 классы

	Ранговый ДА Краскела-Уоллиса; ОГЭ2013ма Груп. (независ.) переменная: КатегПоПерехо Кр.Краскела-Уоллиса: H (1, N= 3581) =35,47			
ОГЭ2013мат	Код	N	Сумма Рангов	Средний Ранг
0	0	304€	532419	1747,92
1	1	535	108938	2036,22

Рис. 6.2 Результаты теста Краскела-Уоллиса

для математики 2013 г. 9 классы

	Ранговый ДА Краскела-Уоллиса; ЕГЭ20 Груп. (независ.) переменная: КатегПоПе Кр.Краскела-Уоллиса: H (1, N= 3581) =7			
ЕГЭ2015рус.яз.	Код	N	Сумма Рангов	Средний Ранг
0	0	3046	539569	1771,40
1	1	535	101787	1902,57

Рис. 6.3 Результаты теста Краскела-Уоллиса

для русского языка 2015 г. 11 классы

	Ранговый ДА Краскела-Уоллиса; ЕГЭ2(Груп. (независ.) переменная: КатегПоП Кр.Краскела-Уоллиса: H (1, N= 3581) =			
ЕГЭ2015матем	Код	N	Сумма Рангов	Средний Ранг
0	0	3046	529131	1737,13
1	1	535	112226	2097,68

Рис. 6.4 Результаты теста Краскела-Уоллиса

для математики 2015 г. 11 классы

В приведенных результатах приняты следующие обозначения:

Код – уникальный код группы;

H – статистика Краскела-Уоллиса;

p – вероятность принятия гипотезы H_0.

Анализируя средние ранги, которые представлены на рисунках 5.1-5.4 можно говорить о влиянии уровня фактора на итоговые баллы. Из результатов видно, что перешедшие в другую школу ученики показывали лучшие результаты еще до перехода в другую школу, а по результатам экзаменов 2015 г. видно, что разница в уровне знании учеников стала только существеннее. Из этого можно судить, что изменение места учебы не только не оказывает отрицательного влияния на уровень знаний ученика, но способствует их улучшению.

В статистике Краскела – Уоллиса вычисляется сумма квадратов разностей средних рангов в группе и среднего ранга по всей выборке. Тогда, при верности гипотезы H_0 - и влияние фактора несущественно, то значение статистики мало. В случае с русским языком $H = 7.335$ и нулевая гипотеза может быть принята с вероятностью p = 0.0068. А в случае с математикой

$H = 55.642$ и нулевая гипотеза может быть принята с вероятностью $p = 0.0001$. Так как мы задали уровень значимости, равный $\alpha = 0.05$, то нулевая гипотеза должна быть отвергнута в обоих случаях в пользу альтернативной гипотезы H_1 о существенном влиянии фактора.

Медианный тест (критерий)

	Медианный тест, общ медиана = Груп. (независ.) переменная: Кате Хи-квадрат = 5,429065 сс = 1 р =		
ОГЭ2013рус	0	1	Всего
<= Медианы: наблюд.	1601,00(	252,000(	1853,00(
ожидаемые	1576,16:	276,837!	
набл.-ожид	24,837	-24,837!	
> Медианы: наблюд	1445,00(	283,000(	1728,00(
ожидаемые	1469,83'	258,162!	
набл.-ожид	-24,837	24,837!	
Сумма: наблюд	3046,00(	535,000(	3581,00(

Рис. 6.5 Результаты медианного теста

для русского языка 2013 г. 9 классы

	Медианный тест, общ медиана = Груп. (независ.) переменная: Кате Хи-квадрат = 33,77293 сс = 1 р =		
ОГЭ2013мат	0	1	Всего
<= Медианы: наблюд.	1706,00(	227,000(	1933,00(
ожидаемые	1644,21:	288,789-	
набл.-ожид	61,789	-61,789-	
> Медианы: наблюд	1340,00(	308,000(	1648,00(
ожидаемые	1401,78!	246,210(	
набл.-ожид	-61,789	61,789-	
Сумма: наблюд	3046,00(	535,000(	3581,00(

Рис. 6.6 Результаты медианного теста

для математики 2013 г. 9 классы

	Медианный тест, общ медиана = Груп. (независ.) переменная: Кате Хи-квадрат = 9,345870 сс = 1 р =		
ЕГЭ2015русПерв	0	1	Всего
<= Медианы: наблюд.	1613,00(	245,000(	1858,00(
ожидаемые	1580,41(	277,584!	
набл.-ожид	32,584	-32,584!	
> Медианы: наблюд	1433,00(	290,000(	1723,00(
ожидаемые	1465,58-	257,415!	
набл.-ожид	-32,584	32,584!	
Сумма: наблюд	3046,00(	535,000(	3581,00(

Рис. 6.7 Результаты медианного теста

для русского языка 2015 г. 11 классы

	Медианный тест, общ медиана = Груп. (независ.) переменная: Кате Хи-квадрат = 39,41953 сс = 1 р =		
ЕГЭ2015матем	0	1	Всего
<= Медианы: наблюд.	1848,00	247,000	2095,00
ожидаемые	1782,00	312,992	
набл.-ожид	65,992	-65,992	
> Медианы: наблюд	1198,00	288,000	1486,00
ожидаемые	1263,99	222,007	
набл.-ожид	-65,992	65,992	
Сумма: наблюд	3046,00	535,000	3581,00

Рис. 6.8 Результаты медианного теста

для математики 2015 г. 11 классы

Ранги в группах не превышавших медиану приведены в верхней части таблицы. В нижней части таблицы – аналогичные значения, превысившие значение медианы.

В случае с русским языком количественная оценка статистики $\chi^2 = 9.346$ свидетельствует о том, что нулевая гипотеза может быть принята с вероятностью $p = 0.0022$, что ниже уровня значимости, следовательно, принимается гипотеза H_1. А при рассмотрении результатов по математике $\chi^2 = 39.42$ говорит о том, что нулевая гипотеза может быть принята с вероятностью $p = 0.00001$, что также ниже уровня значимости, следовательно, принимаем гипотезу H_1.

Критерий Манна-Уитни

Для критерия Манна-Уитни формулируем нулевую гипотезу H_0 об однородности исходных выборок, соответственно гипотеза H_1 утверждает, что выборки не однородны, т. е. влияние фактора существенно.

Таблица 6.1 Результаты теста Манна-Уитни

Перем.	U критерий Манна-Уитни Отмеченные критерии значимы на уровне p <,05000							
	Сум.ранг Группа 1	Сум.ранг Группа 2	U	Z	p-уров.	Z скорр.	p-уров.	N
ЕГЭ2015рус.яз.	5395695	1017876	755114,0	-2,70640	0,006802	-2,70826	0,006764	3046-535
ЕГЭ2015матем	5291311	1122261	650729,5	-7,43924	0,000000	-7,45936	0,000000	3046-535

U – статистика Манна-Уитни для малых выборок;

Z – нормальная аппроксимация статистики Манна-Уитни для больших выборок;

p уров. – вероятность принятия гипотезы H_0;

Z скорр – скорректированная нормальная аппроксимация статистика Манн-Уитни.

Исходя из того, что в обоих случаях (предметах) статистика U достаточно велика и нулевая гипотеза может быть принята с вероятностями p=0,006802 и p=0,000001, что гораздо меньше уровня значимости, мы принимается альтернативная гипотеза - влияние фактора значительное.

Также для проверки гипотезы была искусственно сделана выборка из учеников, сохранивших место учебы в количество 485 наблюдений. Количество сменивших школу учеников 535(1), не сменивших – 485(0).

Критерий Краскела-Уоллиса

	Ранговый ДА Краскела-Уоллиса; ОГЭ2013ру Груп. (независ.) переменная: КатегПоПерехо Кр.Краскела-Уоллиса: H (1, N= 1020) =3,041			
ОГЭ2013рус	Код	N	Сумма Рангов	Средний ранг
0	0	485	255772,(	527,364
1	1	535	264938,(	495,211

Рис. 6.9 Результаты теста Краскела-Уоллиса

2013 г. по русскому языку

	Ранговый ДА Краскела-Уоллиса; ОГЭ2013ма Груп. (независ.) переменная: КатегПоПерехо Кр.Краскела-Уоллиса: H (1, N= 1020) =,0031			
ОГЭ2013мат	Код	N	Сумма Рангов	Средний Ранг
0	0	485	247331,(	509,960
1	1	535	273379,(	510,988

Рис. 6.10 Результаты теста Краскела-Уоллиса

2013 г. по математике

	Ранговый ДА Краскела-Уоллиса; ЕГЭ20 Груп. (независ.) переменная: КатегПоПе Кр.Краскела-Уоллиса: H (1, N= 1020) =2			
ЕГЭ2015рус	Код	N	Сумма Рангов	Средний Ранг
0	0	485	255623,!	527,058
1	1	535	265086,!	495,488

Рис. 6.11 Результаты теста Краскела-Уоллиса

2015 г. по русскому языку

	Ранговый ДА Краскела-Уоллиса; ЕГЭ2(Груп. (независ.) переменная: КатегПоП(Кр.Краскела-Уоллиса: H (1, N= 1020) =			
ЕГЭ2015матем	Код	N	Сумма Рангов	Средний Ранг
0	0	485	236118,(	486,841:
1	1	535	284592,(	531,947'

Рис. 6.12 Результаты теста Краскела-Уоллиса

2015 г. по математике

Из результатов видно, что наивысшие баллы обеспечиваются при изменении места учебы, а худшие – при сохранении места учебы. Такое расхождение с предыдущим анализом объясняется большей разницей в количестве наблюдений в выборках.

Если верна гипотеза H_0 и влияние фактора незначительно, то значение статистики маленькое. В случае с русским языком $H = 2.927$ и нулевая гипотеза может быть принята с вероятностью $p = 0.0871$. А в случае с математикой $H = 5.987$ и нулевая гипотеза может быть принята с вероятностью $p = 0.0144$. Так как заданный нами уровень значимости равен 0.05, то можно сделать вывод: в случае с русским языком влияние фактора не значимо и принимаем нулевую гипотезу, а в случае с математикой принимаем альтернативную гипотезу – влияние фактора значимо.

Медианный тест (критерий)

	Медианный тест, общ медиана = Груп. (независ.) переменная: Кате Хи-квадрат = 2,201149 сс = 1 р =		
ОГЭ2013рус	0	1	Всего
<= Медианы: наблюд.	244,000(	294,000(	538,00(
ожидаемые	255,813'	282,186:	
набл.-ожид	-11,813'	11,813'	
> Медианы: наблюд	241,000(	241,000(	482,00(
ожидаемые	229,186:	252,813'	
набл.-ожид	11,813'	-11,813'	
Сумма: наблюд	485,000(	535,000(	1020,00(

Рис. 6.13 Результаты медианного теста

2013 г. по русскому языку

	Медианный тест, общ медиана = Груп. (независ.) переменная: Кате Хи-квадрат = ,1136459 cc = 1 p =		
ОГЭ2013мат	0	1	Всего
<= Медианы: наблюд.	268,000	290,000	558,00
ожидаемые	265,323	292,676	
набл.-ожид	2,6765	-2,6765	
> Медианы: наблюд	217,000	245,000	462,00
ожидаемые	219,676	242,323	
набл.-ожид	-2,6765	2,6765	
Сумма: наблюд	485,000	535,000	1020,00

Рис. 6.14 Результаты медианного теста

2013 г. по математике

	Медианный тест, общ медиана = Груп. (независ.) переменная: Кате Хи-квадрат = ,7538534 cc = 1 p =		
ЕГЭ2015рус	0	1	Всего
<= Медианы: наблюд.	257,000	298,000	555,00
ожидаемые	263,897	291,102	
набл.-ожид	-6,897	6,897	
> Медианы: наблюд	228,000	237,000	465,00
ожидаемые	221,102	243,897	
набл.-ожид	6,897	-6,897	
Сумма: наблюд	485,000	535,000	1020,00

Рис. 6.15 Результаты медианного теста

2015 г. по русскому языку

	Медианный тест, общ медиана = Груп. (независ.) переменная: Кате Хи-квадрат = 4,635544 cc = 1 p =		
ЕГЭ2015матем	0	1	Всего
<= Медианы: наблюд.	281,000	274,000	555,00
ожидаемые	263,897	291,102	
набл.-ожид	17,102	-17,102	
> Медианы: наблюд	204,000	261,000	465,00
ожидаемые	221,102	243,897	
набл.-ожид	-17,102	17,102	
Сумма: наблюд	485,000	535,000	1020,00

Рис. 6.16 Результаты медианного теста

2015 г. по математике

В случае с русским языком $\chi = 0.754$ и нулевая гипотеза может быть принята с вероятностью p = 0.3853, что значительно превышает уровень значимости, следовательно, принимается гипотеза H_0. А при рассмотрении результатов по математики $\chi = 4.636$ и нулевая гипотеза может быть принята с вероятностью p = 0.0313, что меньше уровня значимости, следовательно, и принимается гипотеза H_1.

Критерий Манна-Уитни

Гипотезы H_0 и H_1 аналогичны тем, что выдвигались в случае с выборками в 3046 и 535 наблюдений.

Таблица 6.2 Результаты теста Манна-Уитни

Перем.	U критерий Манна-Уитни Отмеченные критерии значимы на уровне p <,05000							
	Сум.ранг Группа 1	Сум.ранг Группа 2	U	Z	p-уров.	Z скорр.	p-уров.	N
ЕГЭ2015рус.яз.	265086,5	255623,5	121706,5	-1,70912	0,087430	-1,71062	0,087152	535-485
ЕГЭ2015матем	284592,0	236118,0	118263,0	2,44200	0,014607	2,44667	0,014419	535-485

U – статистика Манна-Уитни для малых выборок;

Z – нормальная аппроксимация статистики Манна-Уитни для больших выборок;

p уров. – вероятность принятия гипотезы H_0;

Z скорр – скорректированная нормальная аппроксимация статистика Манн-Уитни.

Анализируя результаты по русскому языку за 2015 г. можно сделать вывод, что изменение откликов незначимо и две выборки можно признать однородными. А по результатам по математике за 2015 г. 11 классы мы отвергаем нулевую гипотезу в пользу альтернативной, т.к. вероятность принятия нулевой гипотезы 0.0146, что меньше уровня значимости.

6.2 Дисперсионный однофакторный анализ

Так как предварительный ранговый однофакторный анализ подтвердил гипотезу о существенности влияния фактора в случае с выборками в 3046 и 535 наблюдений, попробуем оценить это влияние количественно в рамках дисперсионного анализа. Выдвигаем и проверяем нулевую гипотезу H_0 – влияние фактора на распределение данных не значительно.

Таблица 6.3 Результаты дисперсионного анализа

Переменная	Дисперсионный анализ Отмечены эффекты,значимые на уров. p < ,05000							
	Сум.квад эффект	Ст.св. эффект	Ср.квад. эффект	Сум.квад ошибки	Ст.св. ошибки	Ср.квад. ошибки	F	p
ЕГЭ2015рус.яз.	380,221	1	380,221	253438,6	3579	70,81268	5,3694	0,020549
ЕГЭ2015матем	2832,512	1	2832,512	75970,4	3579	21,22671	133,44	0,0001

Сум.квад эффект – сумма квадратов факторов;

Ст.св.эффект – число степеней свободы фактора;

Ср.квад.эффект – средний квадрат фактора;

Сум.квад.ошибки – сумма квадратов

Ст.св.ошибки – число степенй свободы равная $N - k$;

Ср.квад.ошибки – оценка дисперсии

F – значение статистики Фишера;

p – вероятность принятия гипотезы H_0.

Статистика Фишера в случае с русским языком равна F=5.3694 и отличается от единицы с вероятностью p=0.0205. В случае с математикой F=133.44 незначительно отличается от единицы с вероятностью p=0.0000001. Исходя из данных результатов отвергаем нулевую гипотезу в обоих случаях в пользу альтернативной – влияние фактора существенно.

Таблица 6.4 Влияние уровня фактора на отклики

Итоговая таблица средних N=3581 (Нет пропусков в завис. перем.)						
Категория	2015 рус.яз. Среднее	Доверит. -95,000%	Доверит. +95,000%	2015 рус.яз. N	2015 рус.яз. Сумма	2015 рус.яз Ст.откл.
0	41,61490	41,31676	41,91305	3046	126759,0	8,392140
1	42,52897	41,80331	43,25464	535	22753,0	8,544365
Всего	41,75147	41,47559	42,02734	3581	149512,0	8,420161
Итоговая таблица средних N=3581 (Нет пропусков в завис. перем.)						
Категория	2015 рус.яз. Дисперсия	2015 рус.яз. Минимум	2015 рус.яз. Максимум	2015 рус.яз. 25%	2015 рус.яз. Медиана	2015 рус.яз. 75%
0	70,42801	10,00000	56,00000	36,00000	43,00000	48,00000
1	73,00618	9,00000	56,00000	37,00000	44,00000	49,00000
Всего	70,89911	9,00000	56,00000	36,00000	43,00000	49,00000
Итоговая таблица средних N=3581 (Нет пропусков в завис. перем.)						
Категория	2015 матем Среднее	Доверит.- 95,000%	Доверит. +95,000%	2015 матем N	2015 матем Сумма	2015 матем Ст.откл.
0	9,06402	8,91363	9,21440	3046	27609,00	4,232974
1	11,55888	11,02111	12,09664	535	6184,00	6,331938
Всего	9,43675	9,28303	9,59047	3581	33793,00	4,691694
Итоговая таблица средних N=3581 (Нет пропусков в завис. перем.)						
Категория	2015 матем Дисперсия	2015 матем Минимум	2015 матем Максимум	2015 матем 25%	2015 матем Медиана	2015 матем 75%

Итоговая таблица средних N=3581 (Нет пропусков в завис. перем.)						
Категория	2015 рус.яз. Среднее	Доверит. -95,000%	Доверит. +95,000%	2015 рус.яз. N	2015 рус.яз. Сумма	2015 рус.яз Ст.откл.
0	17,91807	1,000000	28,00000	6,000000	8,00000	12,00000
1	40,09344	1,000000	30,00000	6,000000	10,00000	16,00000
Всего	22,01199	1,000000	30,00000	6,000000	9,00000	12,00000

Теперь проведем анализ с выборками в 485(0) и 535(1) наблюдений.

Таблица 6.5 Результаты дисперсионного анализа

Переменная	Дисперсионный анализ Отмечены эффекты,значимые на уров. p < ,05000							
	Сум.квад эффект	Ст.св. эффект	Ср.квад. эффект	Сум.квад ошибки	Ст.св. ошибки	Ср.квад. ошибки	F	p
ЕГЭ2015рус.яз.	361,8607	1	361,8607	65518,72	1018	64,36024	5,62243	0,017917
ЕГЭ2015матем	573,2467	1	573,2467	30702,28	1018	30,15941	19,00722	0,000014

Статистика Фишера в случае с русским языком равна F=5.6224 и отличается от единицы с вероятностью p=0.017917. В случае с математикой F=19.01 и отличается от единицы с вероятностью p=0.000014. Исходя из данных результатов отвергаем нулевую гипотезу в обоих случаях в пользу альтернативной – влияние фактора существенно.

Таблица 6.6 Влияние уровня фактора на отклики

Итоговая таблица средних N=3581 (Нет пропусков в завис. перем.)						
Категория	2015 рус.яз. Среднее	Доверит. -95,000%	Доверит. +95,000%	2015 рус.яз. N	2015 рус.яз. Сумма	2015 рус.яз Ст.откл.
0	43,72165	43,06105	44,38225	485	21205,00	7,404129
1	42,52897	41,80331	43,25464	535	22753,00	8,544365
Всего	43,09608	42,60205	43,59011	1020	43958,00	8,040659

Итоговая таблица средних N=3581 (Нет пропусков в завис. перем.)						
Категория	2015 рус.яз. Дисперсия	2015 рус.яз. Минимум	2015 рус.яз. Максимум	2015 рус.яз. 25%	2015 рус.яз. Медиана	2015 рус.яз. 75%
0	54,82112	18,00000	56,00000	39,00000	45,00000	,00000
1	73,00618	9,00000	56,00000	37,00000	44,00000	49,00000
Всего	64,65219	9,00000	56,00000	38,00000	45,00000	49,00000

Итоговая таблица средних N=3581 (Нет пропусков в завис. перем.)						
Категория	2015 матем Среднее	Доверит. -95,000%	Доверит. +95,000%	2015 матем N	2015 матем Сумма	2015 матем Ст.откл.
0	10,05773	9,66680	10,44867	485	4878,00	4,381682
1	11,55888	11,02111	12,09664	535	6184,00	6,331938

Итоговая таблица средних N=3581 (Нет пропусков в завис. перем.)						
Категория	2015 рус.яз. Среднее	Доверит. -95,000%	Доверит. +95,000%	2015 рус.яз. N	2015 рус.яз. Сумма	2015 рус.яз Ст.откл.
Всего	10,84510	10,50471	11,18549	1020	11062,00	5,540070

Итоговая таблица средних N=3581 (Нет пропусков в завис. перем.)						
Категория	2015 матем Дисперсия	2015 матем Минимум	2015 матем Максимум	2015 матем 25%	2015 матем Медиана	2015 матем 75%
0	19,19914	1,000000	25,00000	7,000000	10,00000	13,00000
1	40,09344	1,000000	30,00000	6,000000	10,00000	16,00000
Всего	30,69237	1,000000	30,00000	7,000000	10,00000	14,00000

Как видно из представленных далее диаграмм рассеяния, разница между результатами учеников сменивших и сохранивших место учебы существенна в случае с математикой.

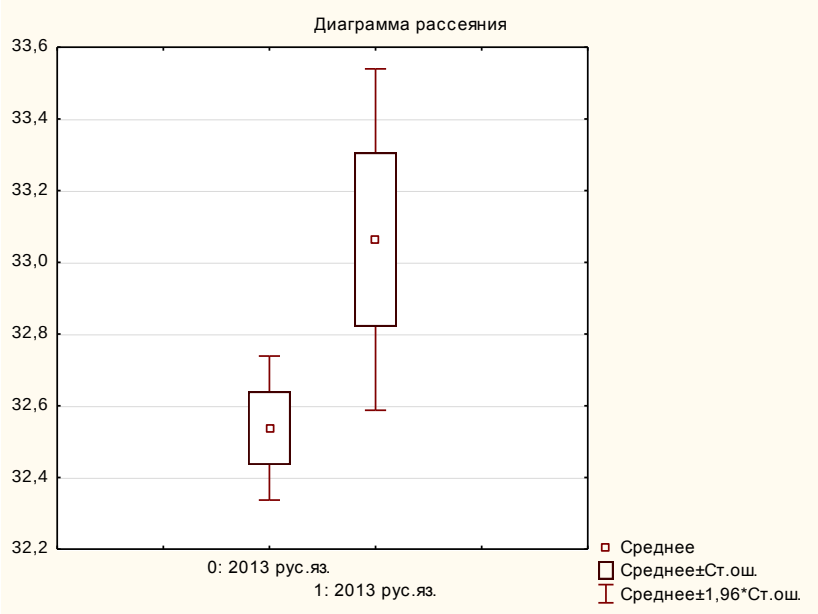


Рис. 6.17 Значения выборочных средних и стандартной ошибки за 2013 г. по русскому языку

На рис.6.17 видно, что результаты экзаменов по русскому языку между группами за девятый класс отличаются незначимо.

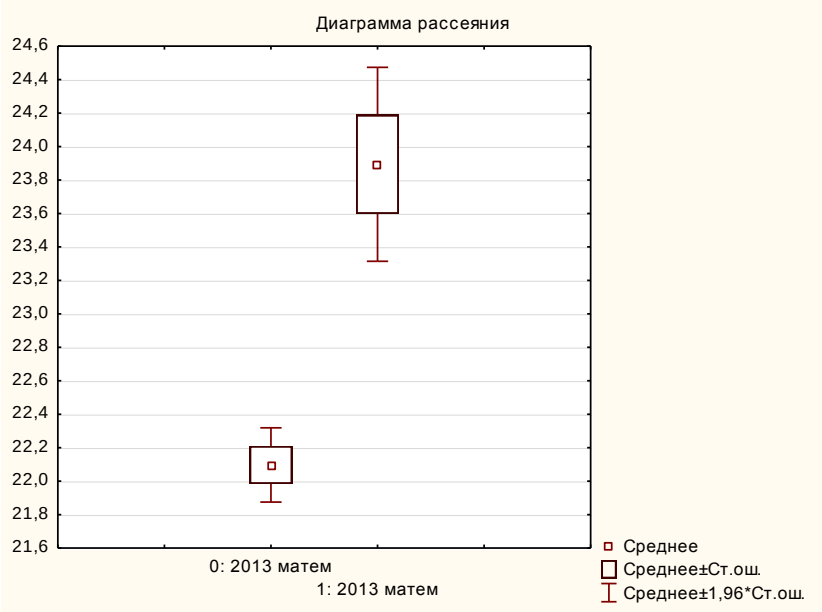


Рис. 6.18 Значения выборочных средних и стандартной ошибки за 2013 г. по математике

На рис.6.18 видно, что результаты экзаменов по математике между группами за девятый класс значимо отличаются.

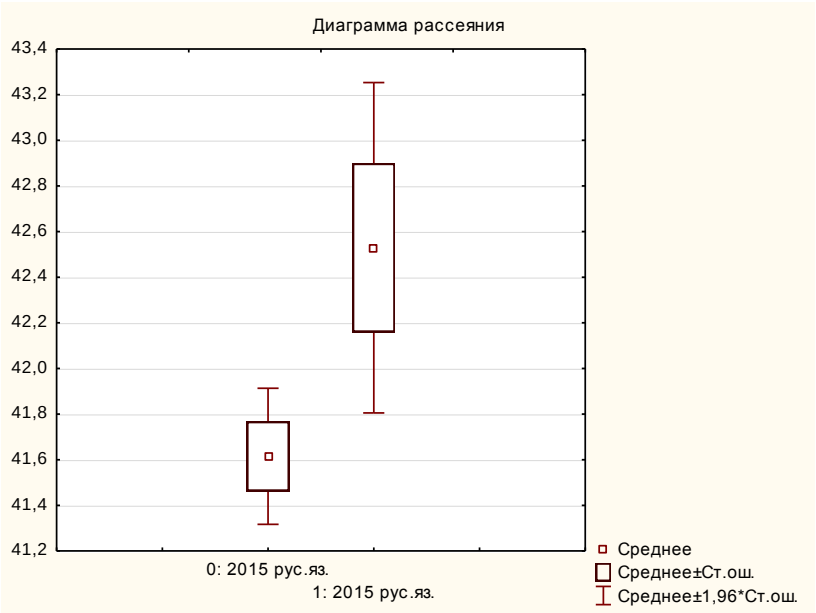


Рис. 6.19 Значения выборочных средних и стандартной ошибки за 2015 г. по русскому языку

На рис.6.19 видно, что результаты экзаменов по русскому языку между группами за одиннадцатый класс отличаются незначимо.

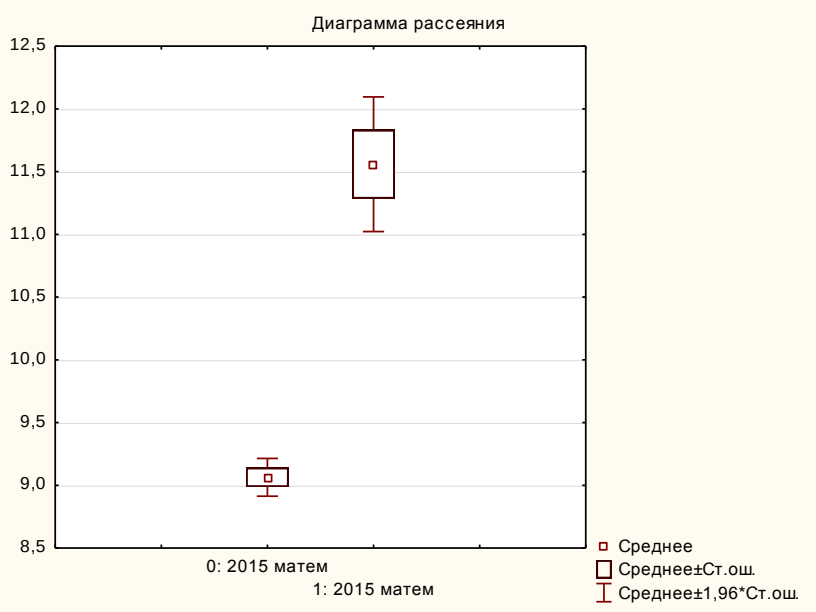


Рис. 6.20 Значения выборочных средних и стандартной ошибки за 2015 г. по математике

На рис.6.20 видно, что результаты экзаменов по математике между группами за одиннадцатый класс значимо отличаются.

ЗАКЛЮЧЕНИЕ

В результате выполнения выпускной квалификационной работы было проведено исследования зависимостей оценок ОГЭ и ЕГЭ(2013 и 2015г соответственно) по русскому языку и математике от фактора сохранения или смены места обучения учениками после 9ого класса, а также на наличие корреляции между результатами экзаменов.

Получены точечные и интервальные оценки качества обучения в школах, что позволило построить рейтинг школ для различных МО. Методами проверки непараметрических гипотез исследованы распределения оценок учащихся (ОГЭ и ЕГЭ) по русскому языку и математике. Для выявления связи между различными оценками использовалась корреляция Пирсона и ранговые корреляции (Кендалла и Спирмена).

В процессе выполнения ВКР были решены следующие задачи:

- изучение предметной области;
- анализ различных статистических методов, выбор наиболее подходящих для исследования и их освоения;
- изучение и освоение вспомогательного программного обеспечения;
- корреляционный анализ;
- однофакторный анализ;
- анализ ресурсоэффективности и ресурсосбережения;
- анализ аспектов социальной ответственности.

Корреляционный анализ дал возможность установить связь между переменными. Проанализировав результаты корреляционного анализа было выявлено, что помимо естественной связь между результатами ОГЭ и ЕГЭ по одному предмету имеет место связь между результатами ЕГЭ по русскому языку и ЕГЭ по математике, а также между результатами ОГЭ по математике и ЕГЭ по русскому языку. Из данных наблюдений можно сделать вывод, что учащиеся, хорошо знающие математику, успешнее напишут экзамен и по русскому языку. Также был проведен корреляционный анализ между

учениками, сохранившими и сменившими место учебы после девятого класса. Анализ данных групп показал более сильную связь среди учеников, сменивших место учебы после девятого класса, кроме случая корреляции между ОГЭ по русскому языку и ЕГЭ по русскому языку: здесь уровень коррелированности ниже.

В ходе проведение однофакторного анализа было выявлено, что изменение места учебы после девятого класса не оказывает отрицательного эффекта на результаты экзаменов. А при проведении однофакторного анализа над искусственной выборкой было выявлено, что влияние фактора существенно на результаты экзаменов по математике, в то время как эффект на результаты экзаменов по русскому языку незначителен.

Планируется продолжить исследование на основе полученных результатов с данными экзаменов ОГЭ и ЕГЭ за другие годы.

По результатам исследований опубликованы тезисы доклада [7], материалы исследования рейтинга школ для различных типов МО и построение однофакторных моделей приняты к публикации [8] и доложены на конференции «Information Technologies in science, management, social sphere and medicine».

СПИСОК ПУБЛИКАЦИЙ

1. Статистический анализ качества образования учеников Томской области / Темирбаев С.К. – Материалы 54-ой международной научной студенческой конференции(МНСК-2016): Новосибирск 2016

2. Statistical analysis of education quality in schools of Tomsk oblast by assessing grades of graduates from the 9^{th} and 11^{th} classes / Yu.Ya.Katsman, S.K.Temirbaev – Materials from the III International Scientific Conference «Information Technologies in science, management, social sphere and medicine»: Tomsk 2016.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Statistica. Искусство анализа данных на компьютере / В. Боровиков – Санкт-Петербург: Питер, 2003 – 688 с.
2. Статистическая обработка экспериментальных данных. Методические указания к лабораторным работам./ Ю.Я. Кацман - Издательство ТПУ.2008 37стр.
3. Теория вероятностей и математическая статистика / В. Гмурман – Москва: Юрайт, 2012 – 479 стр.
4. Statistica: Обзор методов анализа и и руководство пользователя [Руководство пользователя] / Санкт-Петербург: StatSoft, 2011 – 222 стр.
5. Ю.Я. Кацман Лекции по дисциплине «Статистическая обработка экспериментальных данных» Издательство ТПУ.2008
6. Образовательный математический сайт Exponenta [Электронный ресурс]: Раздел Statistica. URL: http://exponenta.ru/soft/Statist/Statist.asp (дата обращения: 09.03.2016)
7. Статистический анализ качества образования учеников Томской области / Темирбаев С.К. – Материалы 54-ой международной научной студенческой конференции(МНСК-2016): Новосибирск 2016
8. Statistical analysis of education quality in schools of Tomsk oblast by assessing grades of graduates from the 9^{th} and 11^{th} classes / Yu.Ya.Katsman, S.K.Temirbaev – Materials from the III International Scientific Conference «Information Technologies in science, management, social sphere and medicine»: Tomsk 2016.
9. ГОСТ 12.0.003-74 ССБТ. Опасные и вредные производственные факторы. Классификация. – М.: Информационно-издательский центр Минздрава России, 1974.
10. СанПиН 2.2.2/2.4.1340-03 Гигиенические требования к персональным электронно-вычислительным машинам и организации работы.

11. СанПиН 2.2.2.542-96 Гигиенические требования к видеодисплейным терминалам, персональным электронно-вычислительным машинам и организации работы.

12. ГОСТ 12.1.019-79 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты (с Изменением №1) Постановление Госстандарта СССР от 17.07.1979 N 2582. ГОСТ от 17.07.1979 N 12.1.019-79.

13. ГОСТ 12.4.124–83 ССБТ. Средства защиты от статического электричества. Общие технические требования. - М.: Изд-во стандартов, 1986.

14. ГОСТ 12.0.002-80 ССБТ. Термины и определения (с изменениями №1) Текст. - Введ. 1980-09-30. - М.: ИПК Изд-во стандартов, 2002.

15. СН 245-71. Санитарные нормы проектирования промышленных предприятий.

16. ГОСТ 12.1.005-88 ССБТ. Общие санитарно-гигиенические требования к воздуху рабочей зоны (с Изменением N 1) Постановление Госстандарта СССР от 29.09.1988 N 3388. ГОСТ от 29.09.1988 N 12.1.005-88.

17. СанПиН 2.2.2.542 – 96. Гигиенические требования к микроклимату производственных помещений: Санитарные правила и нормы. - М.: Информационно-издательский центр Минздрава России, 1997.

18. СНиП 23–05–95. Естественное и искусственное освещение. Введ. 01-01-96. М.:Информационно-издательский центр Минстроя России, 1996. –35 с.

19. СН 2.2.4/2.1.8.562-96. Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки .

20. СанПиН 2.2.4/2.1.8.055-96. Электромагнитные излучения радиочастотного диапазона (ЭМИ РЧ).

21. ГОСТ 12.1.004-85 ССБТ. Пожарная безопасность. Общие требования. - М.: Изд-во стандартов, 1986.- 94 с.

INTRODUCTION

In today's world of information technologies statistical analysis plays an important role at the level of industry, state and society as a whole. The analysis may let us identify different dependences on the set of some factors, a quantitative description of such dependences, evaluation of various parameters such a data distribution, expectation, variance and others.

The goal of final qualifying work is research of the input set of data using different statistical techniques for detection and description of different kinds of dependences of the Tomsk region pupil's exam results for 2013(9^{th} grade) and 2015(11^{th} grade). Exams data contains information about number of gained points by pupils in two disciplines: Russian language and Mathematics.

To achieve this goal following tasks were confirmed:

- Studying subject area;
- Selection and learning of the most reasonable methods for the analysis;
- Learning supporting software;
- Data analysis with selected statistical methods;
- A description of obtained results;
- Resource efficiency and saving analysis;
- Social responsibility aspects analysis.

1 SUBJECT AREA DESCRIPTION

Establishment of laws, to which are random mass phenomena obeyed, is based on the studying methods of experimental data (observation results) statistical analysis.

The mathematical statistics first problem is to specify picking and grouping methods for statistical data, which were received as a result of observations or specially set up experiments.

The mathematical statistics second problem is to develop statistical data analysis methods depending on the purpose of research. These include:

- Evaluation of unknown probability of event; evaluation of unknown distribution function; evaluation of distribution parameters, which form is unknown; evaluation of random variable depending on one or more random variables etc.;
- Statistical hypothesis verification about an unknown distribution form or distribution parameters value, which form is unknown.

Modern mathematical statistic develops necessary tests number determining methods before starting a research (experiment plan), during the research (sequential analysis) and solves many other problems. Modern mathematical statistics if defines as a decision-making science under uncertainty.

The mathematical statistics problem is to create picking and processing statistical activities to obtain scientific and practical conclusions.

Data about Tomsk region pupil’s exam results on BSE and UFE examinations was provided for analysis.

2 STATISTICAL ANALYSIS METHODS

There are different methods of statistical analysis designated to identify specific dependencies and solve various mathematical statistics problems.

In this work employs the most prevalent statistical analysis methods:

- Correlation analysis
- Factor analysis

2.1 Correlation analysis method description

For a wide class of problems there is an exceptional interest in detecting relation between two or more random variables. The engineering research of such problems usually reduces to finding links between some estimated excitation and observed yield of physical system being studied.

The existence of such relations and their comparative force can be measured by the correlation coefficient.

The main problem of correlation analysis is detection of relation between random variables through pointed and interval evaluation of different correlation coefficient's(paired, multiple, partial).

Correlation coefficient r_{xy} determines through correlation moment(covariance) K_{xy} by the formula:

$$\rho_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y} = \frac{M[(X - m_x)(Y - m_y)}{\sigma_x \sigma_y}. \qquad \text{(A.1)}$$

The p_{xy} value characterizes relation tightness between random variable X and Y in the population. From the correlation coefficient's properties is known that it is significative of relation's tightness only at linear dependence between two variables. For linearly independent variables, it equals null (zero). This coefficient varies in range [-1; 1]. Utmost values indicate entire linear dependence. The higher difference of this coefficient's modulus, the greater tightness of the relation.

2.2 Factor analysis method description

All economical enterprise business phenomena and processes are in relations and interdependence. They are interconnected, other indirectly. Hence an important

methodological problem in economic analysis is studying and measuring factor's influence on researched economic indicators values.

Factor analysis can be regarded as a branch of multivariate statistical analysis which combining observed dimension evaluation method by researching the covariance's structure and correlation matrices.

This type of analysis allow to researcher to solve two main problems: describe compact object of measurement and at the same time thoroughly. Factor analysis makes possible to detect factors, which are responsible for the presence of linear statistical correlations relations between observed variables.

Factor analysis is a complex and systematic method of studying and measuring factor's influence on resultative significant value.

The main objectives of the factor analysis:

1. Selection of the factors that determine studied resultative values.
2. Factor's classification and systematization in order to ensure complex and systematic method for researching their influence on activity results.
3. Determination of dependence between factors and resultative values.
4. Interrelation modeling between resultative and factorial values.

6 UNIVARIABLE ANALYSIS

The univariable analysis problem

One of the simplest situation in the researching of dependences is when it is possible to specify only one factor influencing on final result and this factor can take only finite number of values (levels). Such problems which are called problems of univariable analysis are often encountered in practice. A typical example of univariable analysis problems is comparison of results achieved by several different action methods which are aimed to achieve the same goal, for example, several school books or medicines.

The factor (or factors) is what we believe should influence the final result. In the examples above the factors are the conceptions of "school books" or "medicine".

Factor level or treatment method is a concrete realization of a factor (for example certain school book or selected medication).

Measured charachteristic's values are often called the yield. (that is the value of the result).

To compare effects of factor(s) on the result a certain statistics is required. Typically, it is collected in the following way: each of the K number processing methods are applied several times (not necessary the same number of times) to the test object and then record obtained yields. The result of these tests are K number of samples of different sizes. Depending on the number of influencing factors (in our case one factor) one says that the data is reduced to a table with one, two or more inputs.

6.1 Rank univariable analysis

If nothing is known about distribution of the unknown then use quantitative observation Xij for testing hypothesis H_0 becomes difficult. In these cases, it is easier to base findings on relationship "more-less" between observations, because they are independent of the type of distribution. All information contained in those ranks that number of Xij receives in sorting of the entire summation (direction of streamlining min→max or max→min not necessary).

Let us propose a null hypothesis H_0 – yields (marks) belong to the same distribution. That is, the factor's influence (change of school after ninth grade) is not essential. For testing this hypothesis, the data were sorted into pupils who changed school after the ninth grade and the pupils who did not change school. Number of pupils who changed school is 535 (sample "1") and there are 3046 of those who did not change school (sample "0").

Kruskal-Wallis test

	Kruskal-Wallis ANOVA by Ranks; BSE2013rus Independent (grouping) variable Kruskal-Wallis test: H(1, N=3581)=4,690567 p=,0303			
BSE2013rus	Code	N	Sum of Ranks	Mean Rank
0	0	3046	5407695	1775,343
1	1	535	1005877	1880,143

Fig. A.1 Results of Kruskal-Wallis test for Russian language 2013y 9th grades

	Kruskal-Wallis ANOVA by Ranks; BSE2014math Independen (grouping) variable Kruskal-Wallis test: H(1, N=3581)=35,47012 p=,00001			
BSE2014math	Code	N	Sum of Ranks	Mean Rank
0	0	3046	5324190	1747,928
1	1	535	1089382	2036,227

Fig. A.2 Results of Kruskal-Wallis test for Mathematics 2013y 9th grades

	Kruskal-Wallis ANOVA by Ranks;UFE2015rus Independen (grouping) variable Kruskal-Wallis test: H(1, N=3581)= 7,334785 p=,0068			
UFE2015rus	Code	N	Sum of Ranks	Mean Rank
0	0	3046	5395695	1771,403
1	1	535	1017876	1902,572

Fig. A.3 Results of Kruskal-Wallis test for Russian language 2015y 11th grades

	Kruskal-Wallis ANOVA by Ranks;UFE2015math Independen (grouping) variable Kruskal-Wallis test: H(1, N=3581)=55,64238 p=,00001			
UFE2015math	Code	N	Sum of Ranks	Mean Rank
0	0	3046	5291311	1737,134
1	1	535	1122261	2097,683

Fig. A.4 Results of Kruskal-Wallis test for Mathematics, 2015, 11th grades

In the adduced results following designations are used:

Code – unique group code;

H – Kruskall-Wallis statistics;

P –probability of accepting H_0 hypothesis.

Analyzing mean ranks shown in the resulting report one can be talk about influence of factor level on the final marks. From the results it is evident that pupils who changed school got better results before changing school, and according to results of examination in 2015y it is seen that the difference in the knowledge's level of pupils became more significant. From this it could be accepted that changing place of studying does not negatively impact on the pupil's knowledge level, even more, it contributes to their improvement.

Recall that in the Kruskal-Wallis test the sum of mean ranks difference's squares in group and mean rank across all sample are calculated. Then, if H hypothesis is true and factor's influence is insignificant the statistical value is small. In case of Russian language H=7.335 and the null hypothesis can be accepted with probability p=0.0068. In the case of Mathematics H=55.642 and null hypothesis can be accepted with probability p=0.0001.

Since given significance level (p-level) is bigger than α = 0.05, the null hypothesis should be rejected in both cases in favor of alternative hypothesis H1 – factor's influence is significant.

The median test

	Median Test, Overall Median=33,00 BSE2013rus Independent (grouping) variable: CategoryForTrasf Chi-Square=5,429065 df=4 p=,0198		
BSE2013rus	0	1	Всего
<= Median: observed	1601,000	252,0000	1853,000
expected	1576,163	276,8375	
obs.-exp.	24,837	-24,8375	
> Median: observed	1445,000	283,0000	1728,000
expected	1469,837	258,1625	
obs.-exp.	-24,837	24,8375	
Total: observed	3046,000	535,0000	3581,000

Fig. A.5 Results of Median Test for Russian language 2013y 9th grades

BSE2013math	Median Test, Overall Median=22,00 BSE2013math Independent (grouping) variable: CategoryForTrasf Chi-Square=33,77293 df=1 p=,00 0	1	Всего
<= Median: observed	1706,000	227,0000	1933,000
expected	1644,211	288,7894	
obs.-exp.	61,789	-61,7894	
> Median: observed	1340,000	308,0000	1648,000
expected	1401,789	246,2106	
obs.-exp.	-61,789	61,7894	
Total: observed	3046,000	535,0000	3581,000

Fig. A.6 Results of Median Test for Mathematics

2013y 9th grades

UFE2015rus	Median Test, Overall Median= 43,00 UFE2015rus Independent (grouping) variable: CategoryForTrasf Chi-Square=9,34587 df=1 p=,0022 0	1	Всего
<= Median: observed	1613,000	245,0000	1858,000
expected	1580,416	277,5845	
obs.-exp.	32,584	-32,5845	
> Median: observed	1433,000	290,0000	1723,000
expected	1465,584	257,4155	
obs.-exp.	-32,584	32,5845	
Total: observed	3046,000	535,0000	3581,000

Fig. A.7 Results of Median Test for Russian language

2015y 11th grades

UFE2015math	Median Test, Overall Median=9,00 UFE2015math Independent (grouping) variable: CategoryForTrasf Chi-Square=39,41953 df=1 p=,00 0	1	Всего
<= Median: observed	1848,000	247,0000	2095,000
expected	1782,008	312,9922	
obs.-exp.	65,992	-65,9922	
> Median: observed	1198,000	288,0000	1486,000
expected	1263,992	222,0078	
obs.-exp.	-65,992	65,9922	
Total: observed	3046,000	535,0000	3581,000

Fig. A.8 Results of Median Test for Mathematics

2015y 11th grades

In the upper part of the table number of ranks in groups are given which were less or equal to the median. At the bottom part of the table – similar values which were greater than the median value.

In the case of Russian language quantitative evaluation of statistics $\chi^2 = 9.346$ indicates that the null hypothesis can be accepted with probability p=0,0022 which is less than significance level(p-level), therefore, the alternative hypothesis should be accepted. Results for Mathematics $\chi^2 = 39.42$ indicate that the null hypothesis can be accepted with probability p=0.00001 which is less than significance level (p-level), therefore, the alternative hypothesis should be accepted.

Mann-Whitney test

For Mann-Whitney test it is necessary to formulate null hypothesis H0 – initial two samples are homogeneous, accordingly hypothesis H1 asserts (states) that the samples are not homogeneous, i.e., the factor's influence is significant.

Table A.1 Mann-Whitney Test results

Variable	Mann-Whitney U Test (Accuracy) By variable CategoryForTrasf Marked tests are significant at p<.05000							
	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-lvalue	Z adj	p-lvalue	N
UFE2015rus	5395695	1017876	755114,0	-2,70640	0,006802	-2,70826	0,006764	3046-535
UFE2015math	5291311	1122261	650729,5	-7,43924	0,000000	-7,45936	0,000000	3046-535

Legend:

U – Mann-Whitney statistics;

Z – normal approximation of Mann-Whitney statistics for large samples;

p – probability of H_0 hypothesis acceptance;

Zadjusted – adjusted normal approximation of Mann-Whitney statistics.

Assuming that in both cases (subjects) U statistics is large enough and null hypothesis can be accepted with probabilities p=0,006802 and p=0,000001which is much less than significance level (p-level) we accept alternative hypothesis – factor's influence is significant.

In addition, to test the hypothesis an artificial sample was made from pupils who kept their study place, it amounted to 485 observations. The number of pupils who changed school 535(1), who did not – 485(0).

Kruskall-Wallis test

	Kruskal-Wallis ANOVA by Ranks; BSE2013rus Independen (grouping) variable Kruskal-Wallis test: H(1, N=3581)= 3,041647 p=,0812			
BSE2013rus	Code	N	Sum of Ranks	Mean Rank
0	0	485	255772,0	527,3649
1	1	535	264938,0	495,2112

Fig. A.9 Results of Kruskal-Wallis test for Russian language 2013y 9th grades

	Kruskal-Wallis ANOVA by Ranks; BSE2013math Independen (grouping) variable Kruskal-Wallis test: H(1, N=3581)= ,0031047 p=,9556			
BSE2013math	Code	N	Sum of Ranks	Mean Rank
0	0	485	247331,0	509,9608
1	1	535	273379,0	510,9888

Fig. A.10 Results of Kruskal-Wallis test for Mathematics 2013y 9th grades

	Kruskal-Wallis ANOVA by Ranks; UFE2015rus Independen (grouping) variable Kruskal-Wallis test: H(1, N=3581)= 2,926594 p=,0871			
UFE2015rus	Code	N	Sum of Ranks	Mean Rank
0	0	485	255623,5	527,0588
1	1	535	265086,5	495,4888

Fig. A.11 Results of Kruskal-Wallis test for Russian language 2015y 11th grades

	Kruskal-Wallis ANOVA by Ranks; UFE2015math Independen (grouping) variable Kruskal-Wallis test: H(1, N=3581)=5,986708 p=,0144			
UFE2015math	Code	N	Sum of Ranks	Mean Rank
0	0	485	236118,0	486,8412
1	1	535	284592,0	531,9477

Fig. A.12 Results of Kruskal-Wallis test for Mathematics 2015y 11th grades

The results show that the highest scores are attained in case of changing place of studying and the worst – in case of keeping the same place of studying. This discrepancy with previous analysis is explained by big difference in the number of observations in samples.

If H_0 hypothesis is true and factor's influence is insignificant, the statistical value is small. In case of Russian language H=2,927 and null hypothesis can be accepted with probability p=0,0871. In the case of Mathematics H=5,987 and null hypothesis can be accepted with probability p=0.0144. Since the given significance level (p-level) is bigger than α = 0.05, then it can be concluded: in case of Russian language, factor's influence is insignificant, and in case of Mathematics we accept the alternative hypothesis – factor's influence is significant.

Median test

	Median Test, Overall Median=34,00; BSE2013rus Independent (grouping) variable: CategoryForTrans Chi-Square=2,201149 df=1 p=,01379		
BSE2013rus	0	1	Всего
<= Median: observed	244,0000	294,0000	538,000
expected	255,8137	282,1863	
obs.-exp.	-11,8137	11,8137	
> Median: observed	241,0000	241,0000	482,000
expected	229,1863	252,8137	
obs.-exp.	11,8137	-11,8137	
Total: observed	485,0000	535,0000	1020,000

Fig. A.13 Results of Median Test for Russian language 2013y 9th grades

	Median Test, Overall Median=34,00; BSE2013math Independent (grouping) variable: CategoryForTrans Chi-Square=,1136459 df=1 p=,7360		
BSE2013math	0	1	Всего
<= Median: observed	268,0000	290,0000	558,000
expected	265,3235	292,6765	
obs.-exp.	2,6765	-2,6765	
> Median: observed	217,0000	245,0000	462,000
expected	219,6765	242,3235	
obs.-exp.	-2,6765	2,6765	
Total: observed	485,0000	535,0000	1020,000

Fig. A.14 Results of Median Test for Mathematics 2013y 9th grades

	Median Test, Overall Median=34,00;UFE2015rus Independent (grouping) variable: CategoryForTrans Chi-Square=,7538534 df=1 p=,3853		
UFE2015rus	0	1	Всего
<= Median: observed	257,0000	298,0000	555,000
expected	263,8971	291,1029	
obs.-exp.	-6,8971	6,8971	
> Median: observed	228,0000	237,0000	465,000
expected	221,1029	243,8971	
obs.-exp.	6,8971	-6,8971	
Total: observed	485,0000	535,0000	1020,000

Fig. A.15 Results of Median Test for Russian language

2015y 11th grades

UFE2015math	Median Test, Overall Median=34,00;UFE2015math Independent (grouping) variable: CategoryForTrans Chi-Square=		
	0	1	Всего
<= Median: observed	281,0000	274,0000	555,000
expected	263,8971	291,1029	
obs.-exp.	17,1029	-17,1029	
> Median: observed	204,0000	261,0000	465,000
expected	221,1029	243,8971	
obs.-exp.	-17,1029	17,1029	
Total: observed	485,0000	535,0000	1020,000

Fig. A.16 Results of Median Test for Mathematics

2015y 11th grades

In the case of Russian language quantitative evaluation of statistics $\chi^2 = 0.754$ indicates that the null hypothesis can be accepted with probability p=0,3853 that is much bigger than significance level (p-level), therefore, null hypothesis H_0 is accepted. At considering results for Mathematics $\chi^2 = 4.636$ indicates that the null hypothesis can be accepted with probability p=0.0313 that less than significance level(p-level), therefore, alternative hypothesis H_1 is accepted.

Mann-Whitney test

Let us formulate null hypothesis H0 for Mann-Whitney test– initial two samples are homogeneous, accordingly hypothesis H1 asserts(states) that the samples are not homogeneous, i.e., the factor's influence is significant.

Table A.2 Mann-Whitney Test results

Variable	Mann-Whitney U Test (Accuracy) By variable CategoryForTrasf Marked tests are significant at p<.05000							
	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-lvalue	Z adj	p-lvalue	N
UFE2015rus	265086,5	255623,5	121706,5	-1,70912	0,087430	-1,71062	0,087152	535-485
UFE2015math	284592,0	236118,0	118263,0	2,44200	0,014607	2,44667	0,014419	535-485

Legend:

U – Mann-Whitney statistics;

Z – normal approximation of Mann-Whitney statistics for large samples;

p – probability of H_0 hypothesis acceptance;

Zadjusted – adjusted normal approximation of Mann-Whitney statistics.

Analysis of the results for the Russian language for 2015 leads to conclusions that the change of yields is insignificant and two samples can be accepted as homogeneous. And according to results in Mathematics for 11th grades of 2015 the null hypothesis is rejected in favor of its alternative, because null hypothesis acceptance probability is 0,0146, which is less than significance level (p-level).

6.2 One-way ANOVA test

In most cases an additive model is acceptable for data description. It assumes that x_{ij} yield's value can be represented as sum of factor's contribution (influence) and random value which is independent from factors contribution. In other words, each x_{ij} observation is the sum in form of:

$$x_{ij} = a_j + \varepsilon_{ij}; \quad j = 1,\ldots,k; \quad i = 1,\ldots,n. \tag{A.1}$$

Here $a_1, a_2, \ldots, a_k$ – are unknown non-random values that are results of relevant handling; ε_{ij} – independent identically distributed random values, which reflect internal variability inherent to observations.

If in the model in question is known that values $\varepsilon_{ij} \sim N(0, \sigma^2)$, it lets us to use stronger methods in univariable analysis model both for testing hypothesis and for parameters evaluating. The combination of this methods is called one-way ANOVA test.

This name is associated with the fact that analysis of the model is based on the comparison of two dispersions evaluations σ^2. One of them functions regardless of trueness of the hypothesis $H_0 : a_1 = a_2 = \ldots = a_k$. Another evaluation essentially uses this assumption, it affords close to σ^2 result only if the hypothesis is true. From comparison of these two evaluations it can be concluded that null hypothesis should be rejected if dispersions are considerably (significantly) different.

Since pre-rank univariable analysis confirmed hypothesis about significant factor's influence in case with samples in 3046 and 535 observations, we will try to

evaluate this influence quantitatively in ANOVA test frameworks. Check null hypothesis – factor's influence on data distribution is insignificant.

Table A.3 Results of ANOVA test(Analysis of Variance).

Variable	Analysis of Variances(ANOVA test) Marked effects are significant at. p < ,05000							
	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
UFE2015rus	380,221	1	380,221	253438,6	3579	70,81268	5,3694	0,020549
UFE2015math	2832,512	1	2832,512	75970,4	3579	21,22671	133,4409	0,000000

Legend:

SS(Sum of squares) effect – factor's sum of squares;

df effect – factor's number degrees of freedom;

MS(mean square) effect – factor's mean squares;

SS error – sum of squares;

df error – number of degrees of freedom equal N-k;

MS error – variance evaluation;

F – value of Fisher statistics;

P – probability of acceptance for null hypothesis(H_0).

In case of Russian language the Fisher statistics is F=5,3694 which is insignificantly different from one with probability p=0,0205. In case of mathematics F=133,44 insignificantly different from one with probability p=0,0000001. On the assumption of this results, we reject null hypothesis in both cases in favor of its alternative– factor's influence is significant.

Table A.4 Factor's level impact on yield.

Breakdown Table of Descriptive Statistics N=3581 (No missing data in dep.var.list)						
Group	UFE2015rus Means	Confidence -95,000%	Confidence +95,000%	UFE2015rus N	UFE2015rus Sum	UFE2015rus Std.Dev.
0	41,61490	41,31676	41,91305	3046	126759,0	8,392140
1	42,52897	41,80331	43,25464	535	22753,0	8,544365
All grps	41,75147	41,47559	42,02734	3581	149512,0	8,420161
Breakdown Table of Descriptive Statistics N=3581 (No missing data in dep.var.list)						

Group	UFE2015rus Variance	UFE2015rus Minimum	UFE2015rus Maximum	UFE2015rus 25%	UFE2015rus Median	UFE2015rus 75%
0	70,42801	10,00000	56,00000	36,00000	43,00000	48,00000
1	73,00618	9,00000	56,00000	37,00000	44,00000	49,00000
All grps	70,89911	9,00000	56,00000	36,00000	43,00000	49,00000

Breakdown Table of Descriptive Statistics
N=3581 (No missing data in dep.var.list)

Group	UFE2015mat Means	Confidence -95,000%	Confidence +95,000%	UFE2015mat N	UFE2015math Sum	UFE2015math Std.Dev.
0	9,06402	8,91363	9,21440	3046	27609,00	4,232974
1	11,55888	11,02111	12,09664	535	6184,00	6,331938
All grps	9,43675	9,28303	9,59047	3581	33793,00	4,691694

Breakdown Table of Descriptive Statistics
N=3581 (No missing data in dep.var.list)

Group	UFE2015math Variance	UFE2015math Minimum	UFE2015mat Maximum	UFE2015mat 25%	UFE2015math Median	UFE2015math 75%
0	17,91807	1,000000	28,00000	6,000000	8,00000	12,00000
1	40,09344	1,000000	30,00000	6,000000	10,00000	16,00000
All grps	22,01199	1,000000	30,00000	6,000000	9,00000	12,00000

Now let us analyze the samples of 485(0) and 535(1) observations and try to evaluate that influence quantitatively within the framework of variance analysis.

Let us check null hypothesis – factor influence on data distribution is insignificant.

Table A.5 Results of ANOVA test(Analysis of Variance).

Variable	Analysis of Variances(ANOVA test) Marked effects are significant at. p < ,05000							
	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
UFE2015rus	361,8607	1	361,8607	65518,72	1018	64,36024	5,62243	0,01791
UFE2015math	573,2467	1	573,2467	30702,28	1018	30,15941	19,00722	0,00001

In case of Russian language the Fisher statistics is F=5,6224 which is insignificantly different from one with probability p=0,017917. In case of mathematics F=19,01 is insignificantly different from one with probability p=0,000014. On the assumption of this results we reject null hypothesis in both cases in favor of its alternative– factor's influence is significant.

Table 5.6 Factor's level impact on yield.

Breakdown Table of Descriptive Statistics N=3581 (No missing data in dep.var.list)						
Group	UFE2015rus Means	Confidence -95,000%	Confidence +95,000%	UFE2015rus N	UFE2015rus Sum	UFE2015rus Std.Dev.
0	43,72165	43,06105	44,38225	485	21205,00	7,404129
1	42,52897	41,80331	43,25464	535	22753,00	8,544365
All grps	43,09608	42,60205	43,59011	1020	43958,00	8,040659

Breakdown Table of Descriptive Statistics N=3581 (No missing data in dep.var.list)						
Group	UFE2015rus Variance	UFE2015rus Minimum	UFE2015rus Maximum	UFE2015rus 25%	UFE2015rus Median	UFE2015rus 75%
0	54,82112	18,00000	56,00000	39,00000	45,00000	,00000
1	73,00618	9,00000	56,00000	37,00000	44,00000	49,00000
All grps	64,65219	9,00000	56,00000	38,00000	45,00000	49,00000

Breakdown Table of Descriptive Statistics N=3581 (No missing data in dep.var.list)						
Group	UFE2015mat Means	Confidence -95,000%	Confidence +95,000%	UFE2015mat N	UFE2015math Sum	UFE2015math Std.Dev.
0	10,05773	9,66680	10,44867	485	4878,00	4,381682
1	11,55888	11,02111	12,09664	535	6184,00	6,331938
All grps	10,84510	10,50471	11,18549	1020	11062,00	5,540070

Breakdown Table of Descriptive Statistics N=3581 (No missing data in dep.var.list)						
Group	UFE2015math Variance	UFE2015math Minimum	UFE2015mat Maximum	UFE2015mat 25%	UFE2015math Median	UFE2015math 75%
0	19,19914	1,000000	25,00000	7,000000	10,00000	13,00000
1	40,09344	1,000000	30,00000	6,000000	10,00000	16,00000
All grps	30,69237	1,000000	30,00000	7,000000	10,00000	14,00000

As seen from scatterplots below difference between pupils who changed school after 9th grade and who did not is significant in case of mathematics.

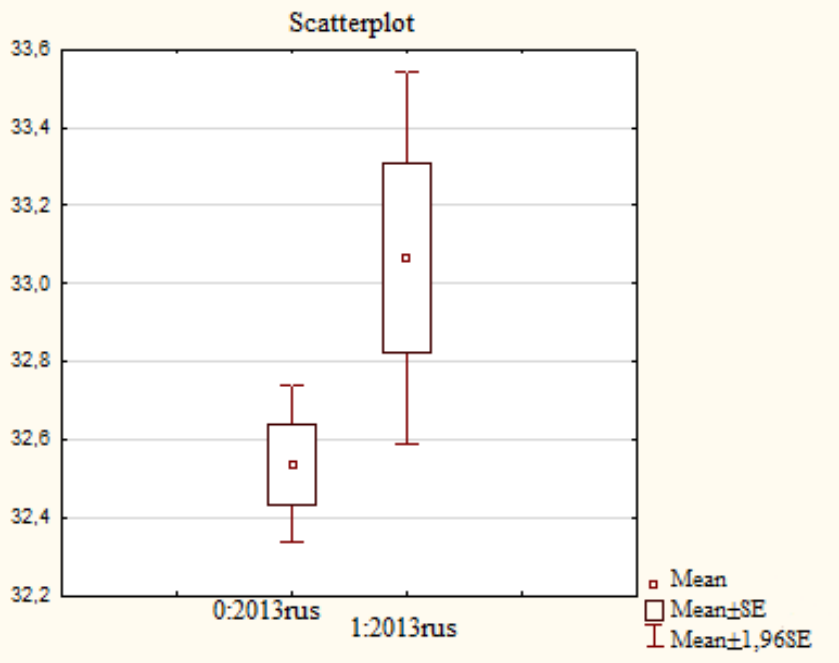


Fig. A.17 Values of sample means and standard error for 2013y, Russian language

Figure A.17 shows that exam results for Russian language for 9th grades differ insignificantly.

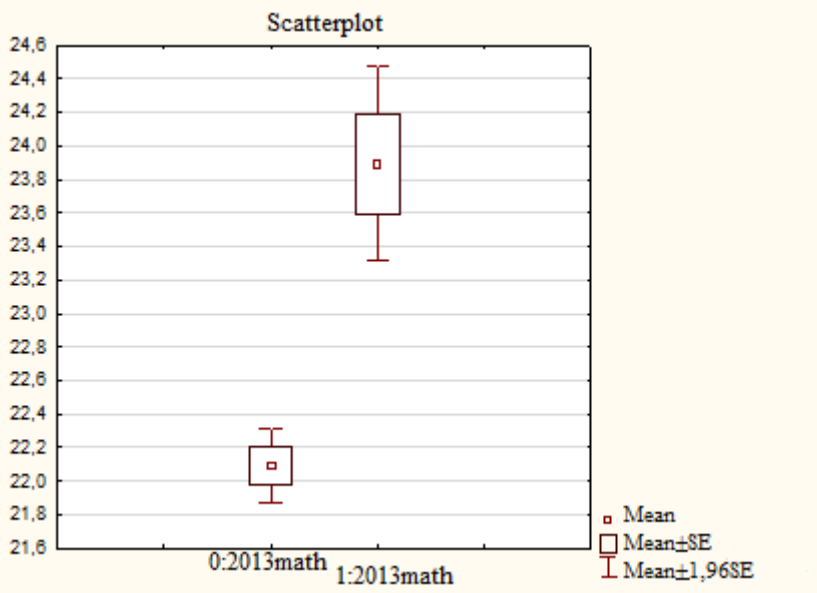


Fig. 5.18 Values of sample means and standard error for 2013y on mathematics

Figure A.18 shows that exam results for mathematics for 9th grades differ significantly.

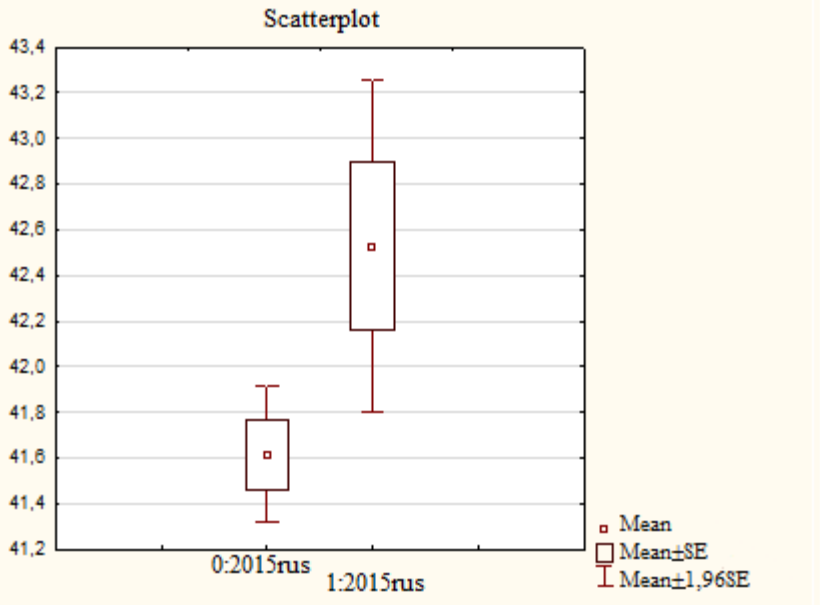


Fig. A.19 Values of sample means and standard error for 2015y on Russian language

Figure A.19 shows that exam results for Russian language for 11th grades differ insignificantly.

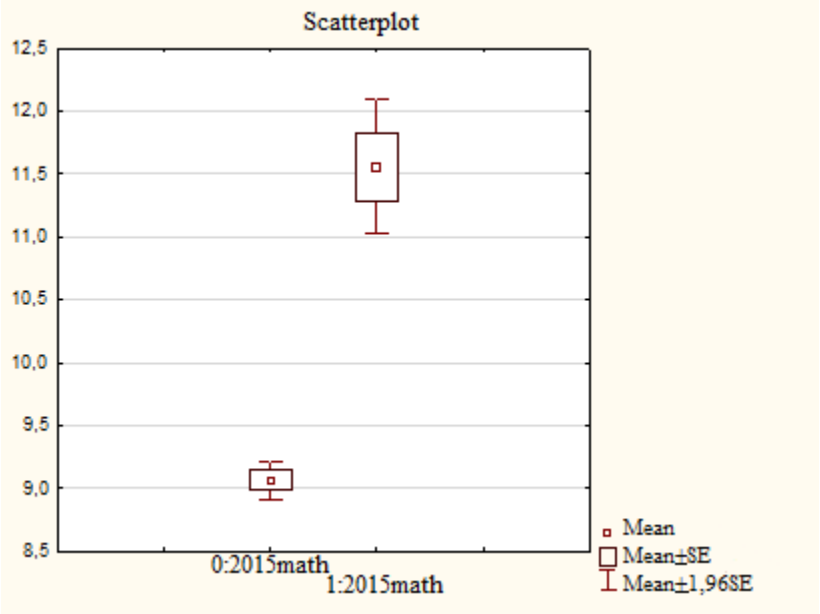


Fig. A.20 Values of sample means and standard error for 2015y on mathematics

Figure A.20 shows that exam results for mathematics for 11th grades differ significantly.