ФГБОУ ВПО "Саратовский государственный университет имени Н.Г.Чернышевского"

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ В GNUMERIC

Учебное пособие

А.В. Букушева

Саратов
2015

Букушева А.В. Статистическая обработка данных в Gnumeric. Учебное пособие. Саратов, 2015. 70 с.

Пособие предназначено для проведения лабораторных занятий по дисциплине "Основы математической обработки информации". В учебном пособии приводятся примеры решения задач в программе Gnumeric. Учебное пособие адресовано студентам, обучающимся по направлению 44.03.01 - Педагогическое образование (Математическое образование). От читателя требуются знания основ математической обработки результатов педагогического эксперимента. Пособие может оказаться полезным для читателей, интересующихся применением информационных технологий в психолого-педагогических исследованиях.

СОДЕРЖАНИЕ

Введение

Для педагога важно уметь грамотно планировать и проводить психолого-педагогические эксперименты, обрабатывать их результаты. Неотъемлемой частью педагогических исследований являются статистические методы, так как без них при решении целого ряда исследовательских задач невозможно дать объективную интерпретацию результатов измерений. В педагогике и психологии статистические методы прочно утвердились тогда, когда эти науки стали активно использовать эксперимент в качестве метода научного исследования. Педагогические измерения сопровождаются некоторой ошибкой, которую вызывают несовершенство диагностического инструментария, различные обстоятельства, связанные с условиями проведения измерений. Поэтому результат педагогического исследования имеет вероятностный характер, следовательно, необходимо доказывать статическую достоверность (значимость) полученных результатов.

Знания, умения и навыки необходимые будущему учителю для решения исследовательских задач в области образования формируются, в частности, при изучении дисциплины „Основы математической обработки информации".

В данной работе показаны некоторые возможности использования программы Gnumeric при решении задач обработки экспериментальных данных методами математической статистики. Пособие предназначено для проведения лабораторных занятий по дисциплине „Основы математической обработки информации". Цель практикума состоит в освоении компьютерных методов решения задач обработки данных психолого-педагогического эксперимента.

1. Числовые характеристики распределений

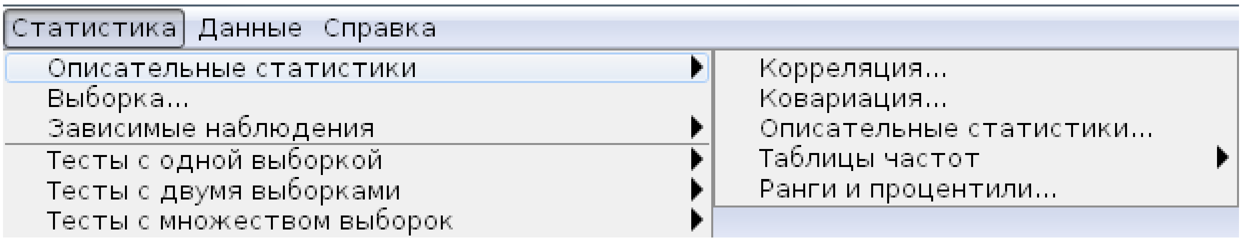
Для экпериментальных данных, полученных по выборке, можно вычислить ряд числовых характеристик: мода, медиана, среднее арифметическое, раброс выборки, дисперсия, стандартное отклонение.

Инструменты статистической обработки данных находятся в пункте главного меню „Статистика". Для получения основных статистических характеристик выборки используется команда „Статистика/Описательные статистики/Описательные статистики".

Пример 1. Для данных, представленных в таблице:

	A	B	C
1	Основные статистические характеристики		
2	Испытуемые	Выборка 1	Выборка 2
3	1	30	24
4	2	27	17
5	3	23	17
6	4	22	15
7	5	19	15
8	6	19	14
9	7	18	14
10	8	16	13
11	9	15	12
12	10	14	12
13	11	13	11
14	12	12	11
15	13	12	8
16	14	12	8
17	15	10	7
18	16	10	7
19	17	10	4
20	18	10	0

определим основные статистические характеристики. Выделяем диапазон ячеек B3:C20 и используем соответственно команды „Статистика/Описательные статистики/Описательные статистики".

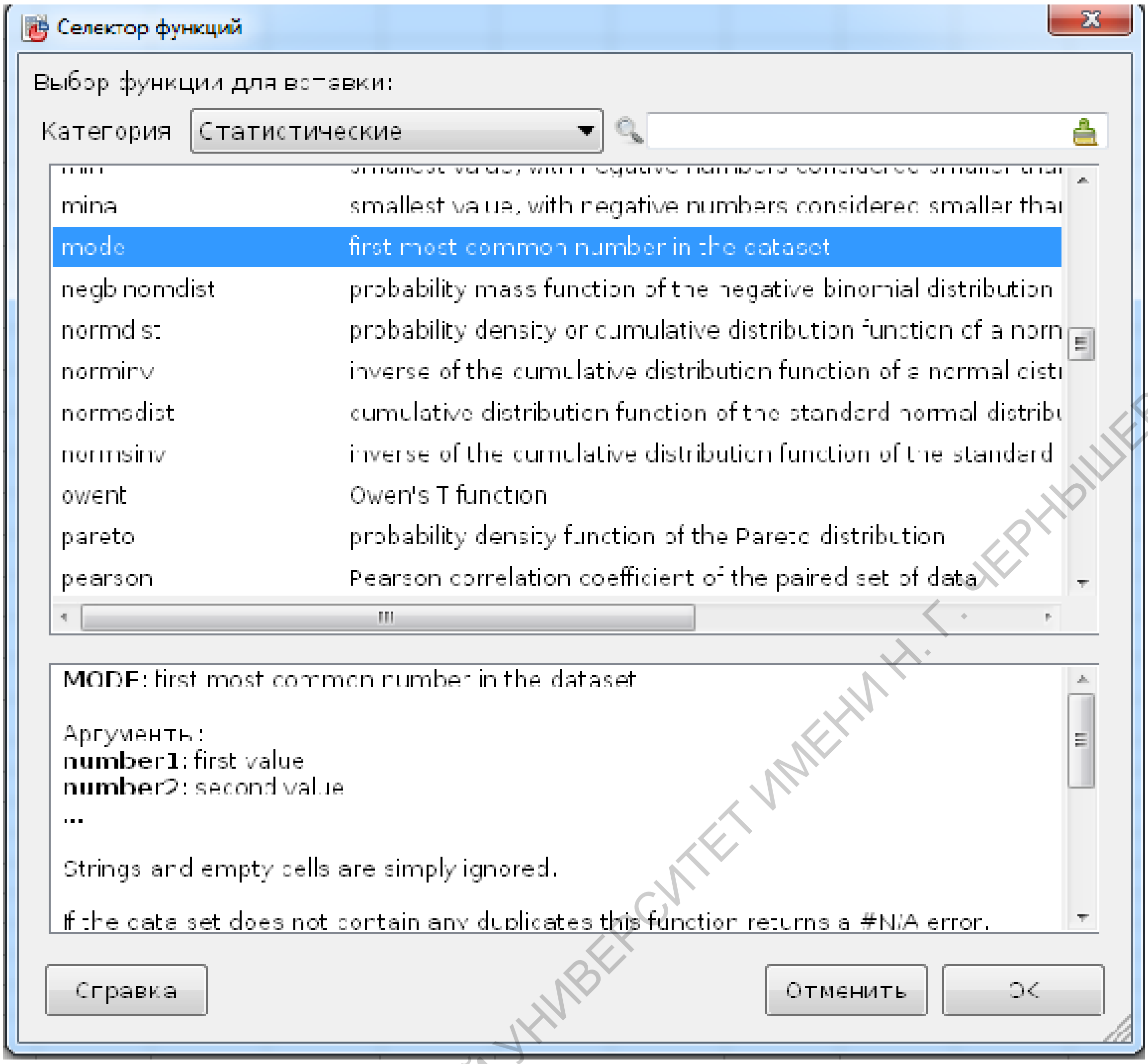


Результат представлен на рисунке:

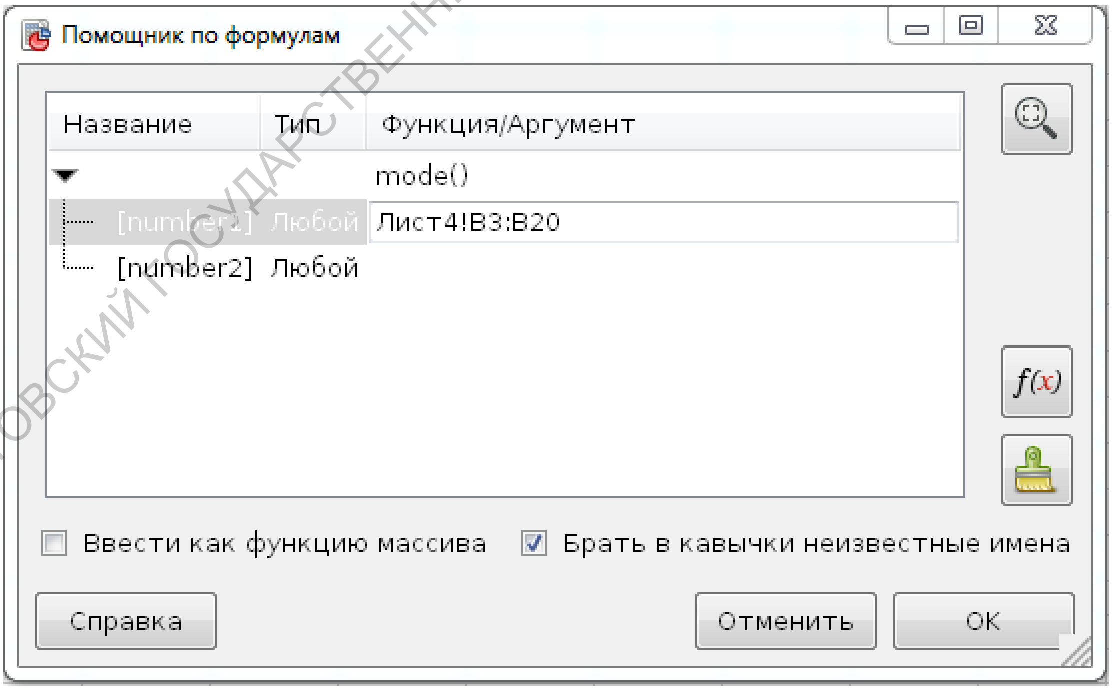
	A	B	C
1		*Столбец 1*	*Столбец 2*
2	*Среднее*	16,2222222222222	11,6111111111111
3	*Стандартная ошибка*	1,4338475036634	1,2913459788849
4	*Медиана*	14,5	12
5	*Мода*	10	17
6	*Стандартное отклонение*	6,0832997581668	5,4787169911651
7	*Выборочная дисперсия*	37,006535947712	30,016339869281
8	*Эксцесс*	0,0578439449148	0,8860679274894
9	*Асимметрия*	0,9360798387095	0,0102142172614
10	*Диапазон*	20	24
11	*Минимум*	10	0
12	*Максимум*	30	24
13	*Сумма*	292	209
14	*Количество*	18	18

Программа Gnumeric, как все электронные таблицы, имеют большое число встроенных функций. Для вызова селектора встроенных функций используется либо главное меню „Вставка/Функция“, либо кнопка f(x) в панели инструментов. После этого появляется диалог выбора функции, в которой есть категория „Статистические функции“. Все функции имеют название и описание на английском языке.

Например, вычислим моду первой выборки. Выделяем ячейку B22 и выбираем кнопку f(x) в панели инструментов. В категории „Статистические“ выбираем `mode`.



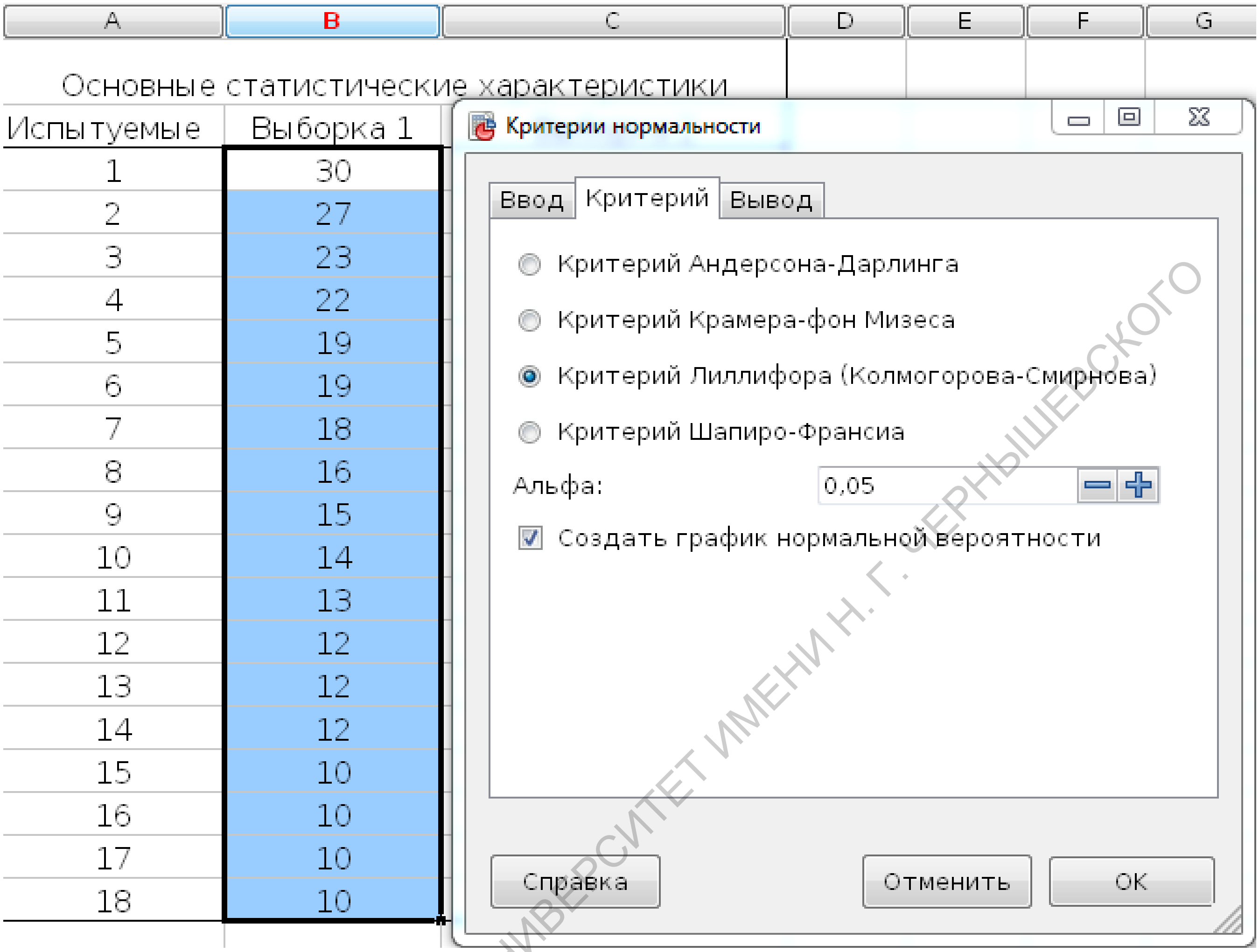
Выделяем диапазон B3:B20 и OK.



	A	B	C
1	Основные статистические характеристики		
2	Испытуемые	Выборка 1	Выборка 2
3	1	30	24
4	2	27	17
5	3	23	17
6	4	22	15
7	5	19	15
8	6	19	14
9	7	18	14
10	8	16	13
11	9	15	12
12	10	14	12
13	11	13	11
14	12	12	11
15	13	12	8
16	14	12	8
17	15	10	7
18	16	10	7
19	17	10	4
20	18	10	0
21			
22	мода	10	

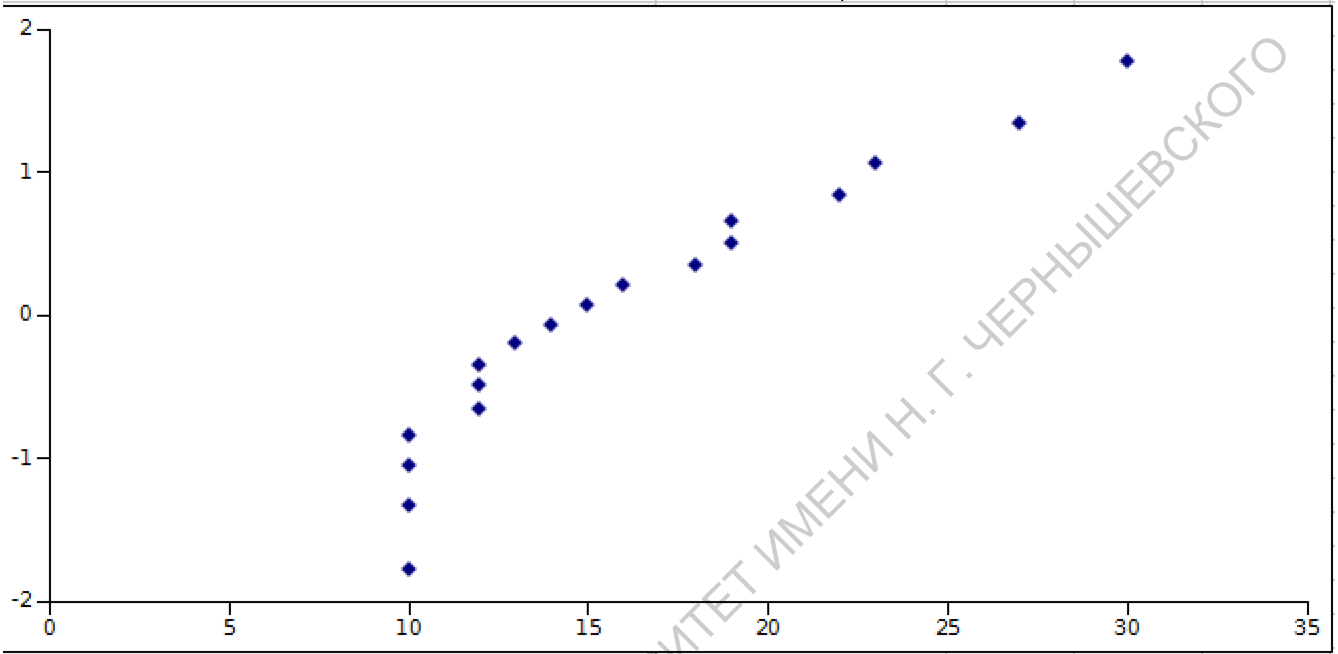
2. Нормальное распределение

Пример 2. Выясним является ли распределение переменной нормальным. Выделяем диапазон ячеек B3:C22. Используем соответственно команды "Статистика/Тесты с одной выборкой/ Критерии нормальности. В диалоговом окне будут четыре критерия, выбираем "Критерий Лиллифора (Колмогорова-Смирнова) далее "OK".



Критерий Колмогорова-Смирнова позволяет оценить вероятность того, данная выборка принадлежит генеральной совокупности с нормальным распределением. Если эта вероятность $p \leq 0,05$, то данное эмпирическое распределение существенно отличается от нормального, если $p > 0,5$, то делают вывод о приблизительном соответствии данного эмпирического распределения нормальному. Из рисунка видно, что распределение переменной является нормальным.

A	B	C	D	E
Критерий Лиллифора (Колмогорова-Смирнова)	*Столбец 1*			
Альфа	0,05			
p-значение	0,3198352096998682			
Статистика	0,1531925323631472			
N	18			
Заключение	Возможно нормален			



3. Непараметрические критерии для зависимых выборок

3.1. Критерий T - Вилкоксона

T - Вилкоксона основан на ранжировании абсолютных величин разности между двумя рядами выборочных значений в первом и втором эксперименте. Далее подсчитывается сумма рангов для положительных разностей и сумма рангов для отрицательных разностей. Меньшая из сумм принимается в качестве эмпирического значения критерия, значение которого сравнивается с табличным значением.

Пример 3. Проверим гипотезу о различии значений показателя, измеренного дважды на одной и той же выборке.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	№	1	2	3	4	5	6	7	8	9	10	11	12
2	Условие 1	6	11	12	8	5	10	7	6	3	9	4	5
3	Условие 2	14	5	8	10	14	7	12	13	11	10	15	16

Используем соответственно команды: „Статистика/ Тесты с двумя выборками/ Claims About Two Medians/ Wilcoxon Signed Rank Test“. Появится диалоговое окно, выбираем соответствующие диапазоны: „Диапазон переменной 1“ - ячейки B2:M2, „Диапазон переменной 2“ - ячейки B3:M3.

B8 = =min(B7;B6)

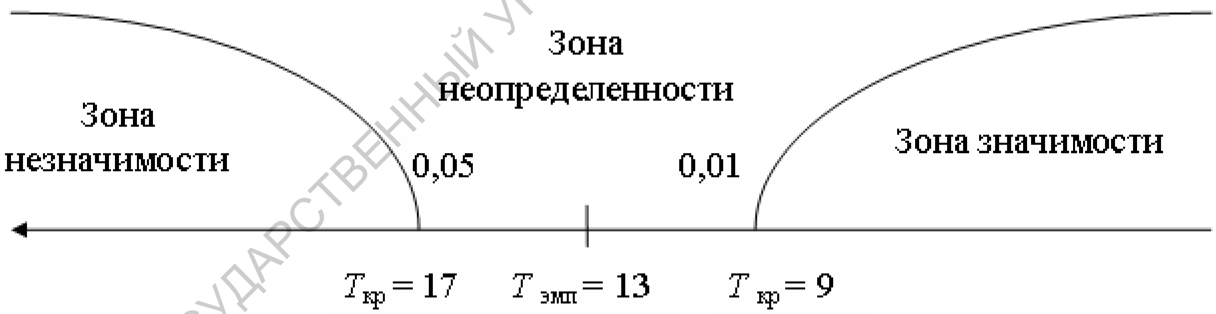
	A	B	C
1	Wilcoxon Signed Rank Test	Переменная 1	Переменная 2
2	Median	6,5	11,5
3	Observed Median Difference	−6	
4	Predicted Median Difference	0	
5	N	12	
6	S−	65	
7	S+	13	
8	Test Statistic	13	
9	α	0,05	
10	P(T≤t) one-tailed	0,0227293761719099	
11	P(T≤t) two-tailed	0,0454587523438197	

Суммы рангов для положительных и отрицательных разностей соответственно равны $T_1 = 13$ (ячейка B7), $T_2 = 65$ (ячейка B6). За эмпирическое значение критерия $T_{\text{эмп}}$ принимается меньшая сумма: $T_{\text{эмп}} = 13$.

Критические значения для $n = 12$ (приложение):

$$T_{\text{кр}} = \begin{cases} 17 \quad \text{для} \quad p \leq 0,05; \\ 9 \quad \text{для} \quad p \leq 0,01. \end{cases}$$

Строим ось значимости:



$T_{\text{кр}}$ попало в зону неопределенности. На уровне 5% принимается статистическая гипотеза H_1 о различии двух условий по уровню выраженности изучаемого признака. На уровне 1% принимается гипотеза H_0 о сходстве. Выбор уровня значимости (5% или 1%) определяется планом и задачами эксперимента.

3.2. Критерий знаков G

Критерий знаков дает возможность установить, насколько однонаправленно изменяются значения признака при повторном измерении связанной (зависимой), однородной выборки.

Критерий знаков G основан на учете знака разностей (сдвигов) значений признака в каждой паре его измерений. Далее вычисляется общее число нулевых, положительных и отрицательных сдвигов. Сумма сдвигов, получившаяся наименьшей, принимается за эмпирическое значение.

Пример 4. Проверим гипотезу о различии значений показателя, измеренного дважды на одной и той же выборке.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	№	1	2	3	4	5	6	7	8	9	10	11	12
2	Условие 1	6	11	12	8	5	10	7	6	3	9	4	5
3	Условие 2	14	5	8	10	14	7	12	13	11	10	15	16

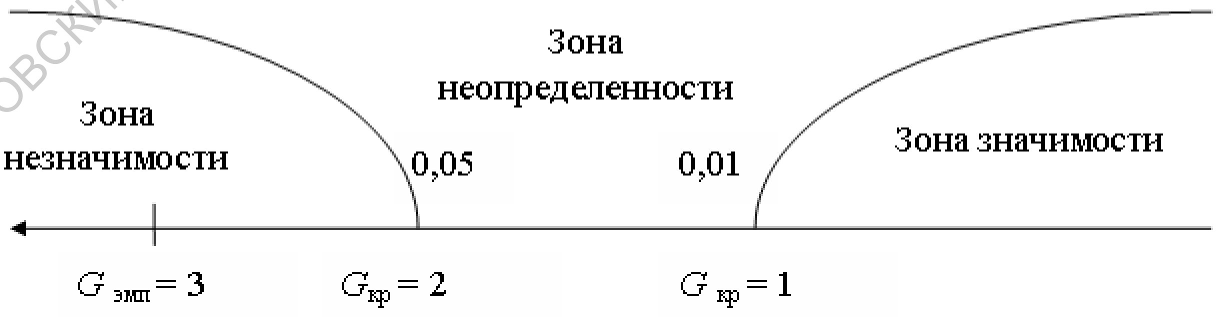
Используем соответственно команды: „Статистика/ Тесты с двумя выборками/ Claims About Two Medians/ Критерий знаков“. Появится диалоговое окно, выбираем соответствующие диапазоны: „Диапазон переменной 1“ - ячейки B2:M2, „Диапазон переменной 2“ - ячейки B3:M3.

	A	B	C
1	*Sign Test*	*Переменная 1*	*Переменная 2*
2	*Median*	6,5	11,5
3	*Predicted Difference*	0	
4	*Test Statistic*	3	
5	*N*	12	
6	*α*	0,05	
7	*P(T≤t) one-tailed*	0,072998046875	
8	*P(T≤t) two-tailed*	0,14599609375	

Эмпирическое значение $G_{эмп}$ равно 3 (Test Statistic). Критические значения $n = 12$ (приложение) равны:

$$G_{кр} = \begin{cases} 2 & \text{для} \quad p \leq 0,05; \\ 1 & \text{для} \quad p \leq 0,01. \end{cases}$$

Строим ось значимости:



$G_{\text{эмп}}$ попало в зону незначимости, принимается гипотеза H_0 об отсутствии различий.

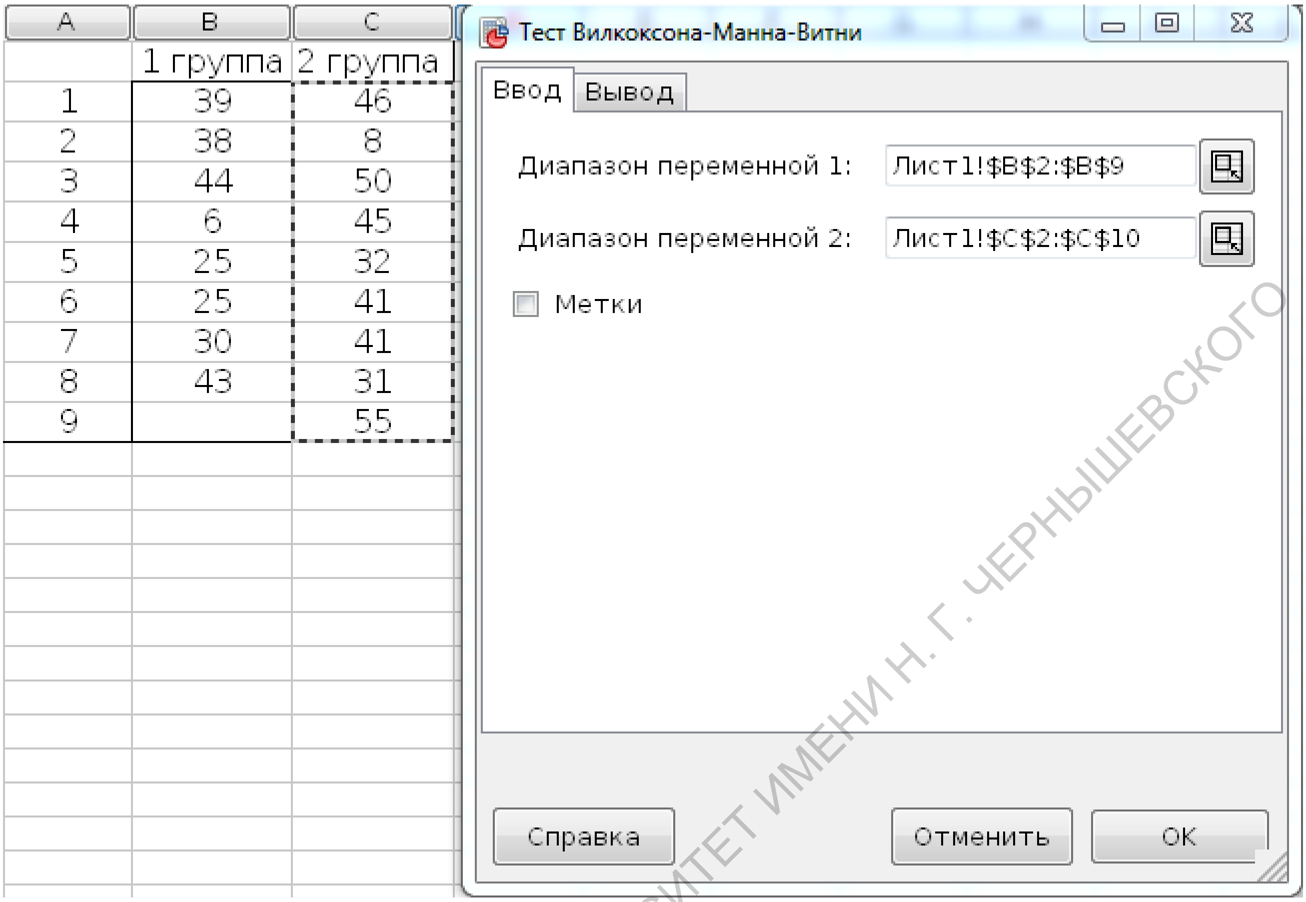
4. Непараметрический критерий для независимых выборок: U - Вилкоксона-Манна-Уитни

Эмпирическое значение критерия U - Вилкоксона-Манна-Уитни показывает, насколько совпадают (пересекаются) два ряда значений измеренного признака. Чем меньше совпадение, тем больше различаются эти два ряда. Основная идея критерия основана на представлении всех значений двух выборок в виде одной общей последовательности упорядоченных (ранжированных) значений. Нулевой статистической гипотезе будет соответствовать ситуация, когда значения одной выборки будут равномерно распределены среди значений другой выборки, то есть когда два ряда значений пересекаются в наибольшей возможной степени. Напротив, отклонение этой гипотезы будет соответствовать ситуация, когда значения одной из выборок будут преобладать на одном из концов объединенного ряда - пересечение двух рядов будет минимальным.

Пример 5. Две неравные по численности группы испытуемых решали техническую задачу. Показателем успешности служило время решения. В первой группе испытуемые получали дополнительную мотивацию в виде денежного вознаграждения. Психолога интересует вопрос - влияет ли вознаграждение на успешность решения задачи? Были получены следующие результаты времени решения задачи в секундах:

	A	B	C
1		1 группа	2 группа
2	1	39	46
3	2	38	8
4	3	44	50
5	4	6	45
6	5	25	32
7	6	25	41
8	7	30	41
9	8	43	31
10	9		55

Используем соответственно команды: „Статистика/ Тесты с двумя выборками/ Claims About Two Medians/ Тест Вилкоксона-Манна-Витни“. Появится диалоговое окно, выбираем соответствующие диапазоны:



На новом листе будет представлен результат команды:

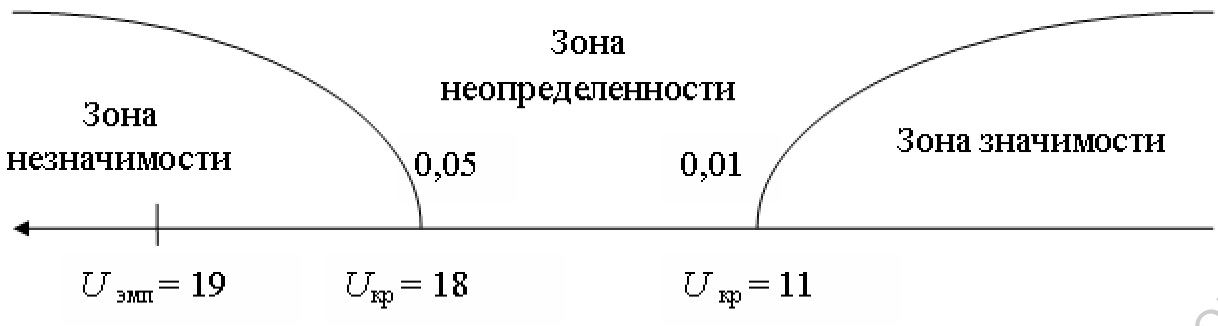
	A	B	C	D
1	*Тест Вилкоксона-Манна-Витни*			
2		*Переменная 1*	*Переменная 2*	*Итого*
3	*Rank-Sum*	55	98	153
4	*N*	8	9	17
5	*U*	19	53	72
6	*Ties*	2		
7	*Statistic*	55		
8	*U-Statistic*	19		
9	*p-Value*	0,10187605853790074		

Таким образом, $U_{\text{эмп}} = 19$ (U-Statistic ячейка B8).

Определяем критические значения по таблице, которая состоит из нескольких таблиц, рассчитанных отдельно для уровней p=0,05, p=0,01, а также для величин n_1 и n_2 (приложение). В нашем случае $n_1 = 8$, $n_2 = 9$.

$$U_{\text{кр}} = \begin{cases} 18 \quad \text{для} \quad p \le 0,05; \\ 11 \quad \text{для} \quad p \le 0,01. \end{cases}$$

Ось значимости имеет вид:



Ось значимости имеет направление справа налево. При этом числовые значения по оси абсцисс по мере увеличения уровней значимости убывают, чем меньше взаимопересечений в двух рядах, тем больше достоверность их различий.

Полученное значение $U_{\text{эмп}}$ попало в зону незначимости, следовательно, принимается гипотеза H_0 о сходстве, а гипотеза H_1 о наличии различий отклоняется. Таким образом, дополнительная мотивация не приводит к статистически значимому увеличению эффективности решения технической задачи.

5. Критерий хи-квадрат

Критерий хи-квадрат (χ^2) используется в двух вариантах:

1) как расчет согласия эмпирического распределения и предполагаемого теоретического; в этом случае проверяется гипотеза H_0 об отсутствии различий между теоретическим и эмпирическим распределениями;

2) как расчет однородности двух независимых экспериментальных выборок; в этом случае проверяется гипотеза H_0 об отсутствии различий между двумя эмпирическими (экспериментальными) распределениями.

Расчетная формула критерия хи-квадрат для сравнения теоретического и эмпирического распределений:

$$\chi^2_{\text{эмп}} = \sum_{i=1}^{k} \frac{(f_{\text{э}} - f_{\text{т}})^2}{f_{\text{т}}},$$

где $f_{\text{э}}$ - эмпирическая частота,

$f_{\text{т}}$ - теоретическая частота,

k - количество разрядов признака.

Расчетная формула критерия хи-квадрат для сравнения двух эмпирических распределений:

$$\chi^2_{\text{эмп}} = \frac{1}{n_1 n_2} \sum_{i=1}^{k} \frac{(n_1 \cdot m_{2i} - n_2 \cdot m_{1i})^2}{m_{1i} + m_{2i}},$$

где n_1, n_2 - соответственно число элементов в первой и во второй выборках.

Элементы каждой выборки распределяются на k категорий, соответствующих разрядам исследуемого признака. На основе полученных результатов составляется таблица вида:

выборки	разряды исследуемого признака				$\sum$
	1	2	...	k	
1	m_{11}	m_{12}	...	m_{1k}	n_1
2	m_{21}	m_{22}	...	m_{2k}	n_2
$\sum$	$m_{11}+m_{21}$	$m_{12}+m_{22}$	...	$m_{1k}+m_{2k}$	n_1+n_2

Для критерия хи-квадрат оценка уровней значимости определяется по числу степеней свободы, которое обозначается ν, и вычисляется по формуле: $\nu = c_1 - 1$, где c_1 определяется по выборочным данным и представляет собой число элементов в выборке. Если при расчете критерия используется таблица экспериментальных данных, то величина ν рассчитывается по формуле: $\nu = (c_1 - 1)(c_2 - 1)$, где c_1 - число строк, c_2 - число столбцов.

Пример 6 (сравнение эмпирического распределения с теоретическим). Психолог решает задачу: будет ли удовлетворенность работой на данном предприятии распределена равномерно по следующим альтернативам:

1 - работой вполне доволен;

2 - скорее доволен, чем не доволен;

3 - трудно сказать, не знаю, безразлично;

4 - скорее недоволен, чем доволен;

5 - совершенно недоволен работой.

Для решения этой задачи производится опрос случайной выборки из 65 респондентов (испытуемых) об удовлетворенности работой: „В какой степени Вас устраивает Ваша работа“, причем ответы должны даваться согласно указанным альтернативам.

Решение. Данные представлены в таблице (эмпирические частоты):

	A	B	C	D
1	Альтернативы	fэ		
2	1	8		
3	2	22		
4	3	14		
5	4	9		
6	5	12		

B7 =sum(B2:B6)

	A	B	C
1	Альтернативы	fэ	fт
2	1	8	
3	2	22	
4	3	14	
5	4	9	
6	5	12	
7	Сумма	=sum(B2:B6)	

sum(▶[values]◀;...)
values: список складываемых значений

Ctrl-F4 закрывает подсказку

Теоретические частоты равны $\frac{65}{5} = 13$.

C2 =65/5

	A	B	C	D	E
1	Альтернативы	fэ	fт	(fэ-fт)^2	(fэ-fт)^2/fт
2	1	8	=65/5		
3	2	22			
4	3	14			
5	4	9			
6	5	12			
7	Сумма	65			

Для каждой альтернативы находим разность между эмпирической и теоретической частотами и возводим ее в квадрат. Вводим формулу в ячейку D2 и копируем формулу в ячейки D3-D6.

D2 =(B2-C2)^2

	A	B	C	D	E
1	Альтернативы	fэ	fт	(fэ-fт)^2	(fэ-fт)^2/fт
2	1	8	13	=(B2-C2)^2	
3	2	22	13		
4	3	14	13		
5	4	9	13		
6	5	12	13		
7	Сумма	65			

E2 =D2/$C$2

	A	B	C	D	E
1	Альтернативы	fэ	fт	(fэ-fт)^2	(fэ-fт)^2/fт
2	1	8	13	25	=D2/$C$2
3	2	22	13	81	
4	3	14	13	1	
5	4	9	13	16	
6	5	12	13	1	
7	Сумма	65			

В ячейку E2 вводим формулу $\frac{(f_э - f_т)^2}{f_т}$ и копируем эту формулу в ячейки E3-E6.

E2 =D2/$C$2

	A	B	C	D	E	F
1	Альтернативы	fэ	fт	(fэ-fт)^2	(fэ-fт)^2/fт	
2	1	8	13	25	1,92307692	D6/$C$2
3	2	22	13	81		
4	3	14	13	1		
5	4	9	13	16		
6	5	12	13	1		
7	Сумма	65				

Находим $\chi^2_{эмп}$.

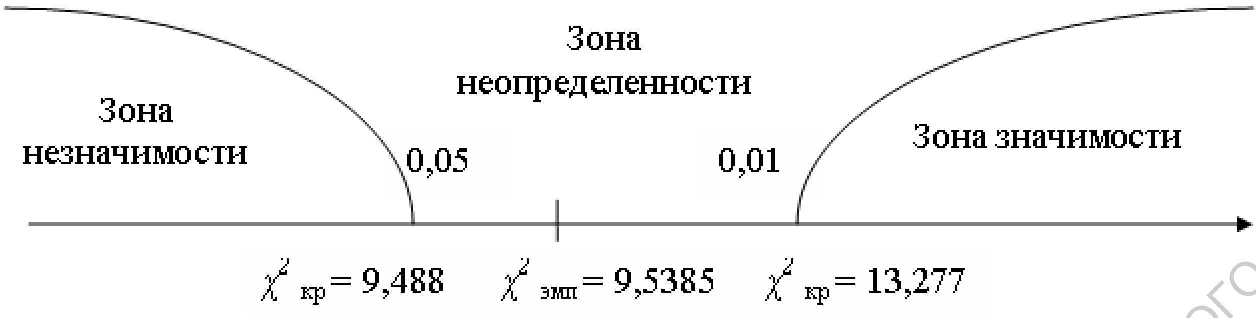
E8 =sum(E2:E6)

	A	B	C	D	E
1	Альтернативы	fэ	fт	(fэ-fт)^2	(fэ-fт)^2/fт
2	1	8	13	25	1,9231
3	2	22	13	81	6,2308
4	3	14	13	1	0,0769
5	4	9	13	16	1,2308
6	5	12	13	1	0,0769
7	Сумма	65			
8	χ^2 эмп				9,5385

Число степеней свободы определяем по формуле: $\nu = c_1 - 1$, где c_1 - количество альтернатив (строк). В нашем случае $c_1 = 5$, $\nu = 5 - 1 = 4$. Находим критические значения (приложения):

$$\chi^2_{кр} = \begin{cases} 9,488 & \text{для} \quad p \leq 0,05; \\ 13,277 & \text{для} \quad p \leq 0,01. \end{cases}$$

Строим ось значимости:



Выличина $\chi^2_{\text{эмп}}$ попала в зону неопределенности. Можно считать, что полученные различия значимы на уровне 5% и принять гипотезу H_1 о различии теоретического и эмпирического распределений. Эмпирическое распределение выбора альтернатив значимо отличается от теоретически предположенного равномерного выбора альтернатив.

Пример 7 (сравнение двух экспериментальных распределений). Была проведена выборка абитуриентов. Для каждого респондента выборки определены: а) пол; б) одна из 4 предпочитаемых специальностей. Результаты исследования представлены в таблице.

	A	B	C	D	E
1		факультеты			
2	пол	1	2	3	4
3	Ж (выборка 1)	4	21	17	9
4	М (выборка 2)	6	25	11	5

Проверяется гипотеза H_0: предпочтения у юношей и девушек в выборе специальностей совпадают. Альтернативная гипотеза H_1: предпочтения у юношей и девушек в выборе специальностей не совпадают.

Решение. Выделяем диапазон B3:E4. Используем команды: „Статистика/Тесты с множеством выборок/ Contingency Table/ Критерий однородности“, далее „OK“.

	A	B
1	*Критерий однородности*	
2	*Test Statistic*	3,0181603816364655
3	*Degrees of Freedom*	3
4	*p-Value*	0,3888336688171279
5	*Critical Value*	7,81472790325118

В ячейке B2 указано эмпирическое значение $\chi^2_{\text{эмп}} = 3,018$ (Test Statistic), степень свободы равна 3 - ячейка B3 (Degrees of Freedom), критическое значение $\chi^2_{\text{кр}} = 7,815$ для $p \leq 0,05$ указано в ячейке B5 (Critical Value).

Если вычислять $\chi^2_{\text{эмп}}$ с использованием формулы

$$\chi^2_{\text{эмп}} = \frac{1}{n_1 n_2} \sum_{i=1}^{k} \frac{(n_1 \cdot m_{2i} - n_2 \cdot m_{1i})}{m_{1i} + m_{2i}},$$

где n_1, n_2 - соответственно число элементов в первой и во второй выборках, то получим:

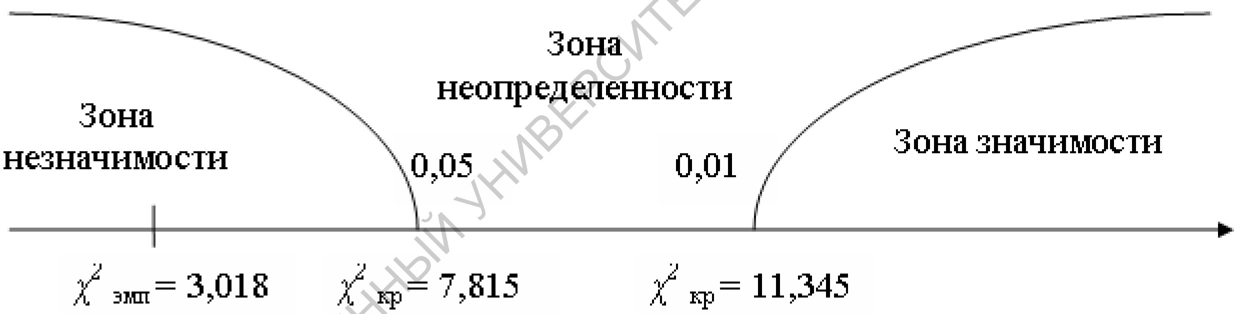
$$\chi^2_{\text{эмп}} = \frac{1}{51 \cdot 47} \sum_{i=1}^{4} \frac{(51 \cdot m_{2i} - 47 \cdot m_{1i})^2}{m_{1i} + m_{2i}} =$$

$$\frac{1}{51 \cdot 47}\left(\frac{(51 \cdot 6 + 47 \cdot 4)^2}{4+6} + \frac{(51 \cdot 25 + 47 \cdot 21)^2}{21+25} + \frac{(51 \cdot 11 + 47 \cdot 17)^2}{11+17} + \frac{(51 \cdot 5 + 47 \cdot 9)^2}{9+5}\right)$$

$= 3,018.$

Находим критическое значения $\nu = 3$ для $p \leq 0,01$: $\chi^2_{\text{кр}} = 11,345$ (приложение).

Строим ось значимости:



Полученная величина $\chi^2_{\text{эмп}}$ попала в зону незначимости. Принимается гипотеза H_0 об отсутствии различий между двумя эмпирическими распределениями. Поэтому предпочтения у юношей и девушек в выборе специальностей совпадают.

6. t-критерий Стьюдента

Критерий t Стьюдента относится к параметрическим критериям и направлен на оценку различий величин средних $\bar{X}$ и $\bar{Y}$ двух выборок X и Y, которые распределены по нормальному закону.

Случай независимых выборок

Формула для расчета по t-критерию Стьюдента:

$$t_{\text{эмп}} = \frac{|\bar{x} - \bar{y}|}{Sd}$$

где $Sd = \sqrt{S_x^2 + S_y^2}$.

Если $n_1 = n_2 = n$, то

$$Sd = \sqrt{S_x^2 + S_y^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2}{n^2 - n}}.$$

В случае неравночисленных выборок $n_1 \neq n_2$:

$$Sd = \sqrt{S_x^2 + S_y^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2}{n_1 + n_2 - 2} \cdot \frac{(n_1 + n_2)}{n_1 \cdot n_2}}.$$

Число степеней свободы вычисляется по формуле

$$k = n_1 + n_2 - 2,$$

где n_1 и n_2 соответственно величины первой и второй выборки.

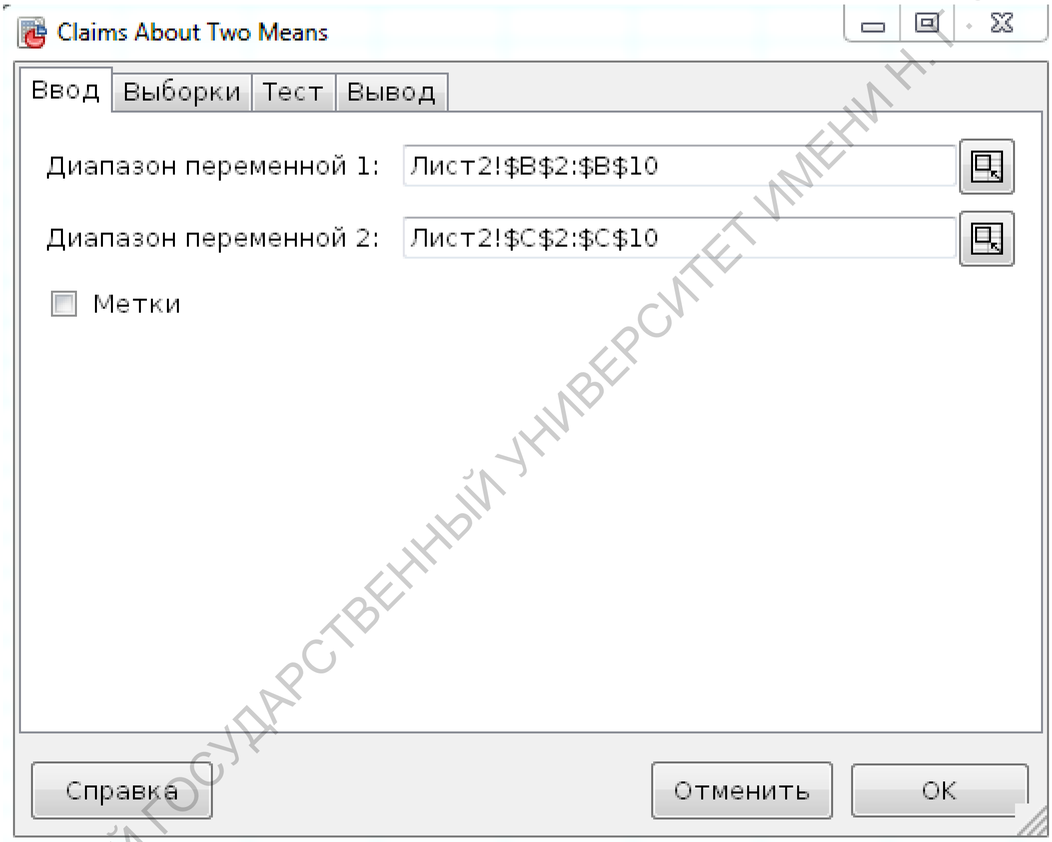
Пример 8. Пусть даны две группы: экспериментальная X и контрольная Y. Проверим гипотезу о том, что средние значения двух генеральных совокупностей, из которых извлечены сравниваемые независимые выборки, отличаются друг от друга.

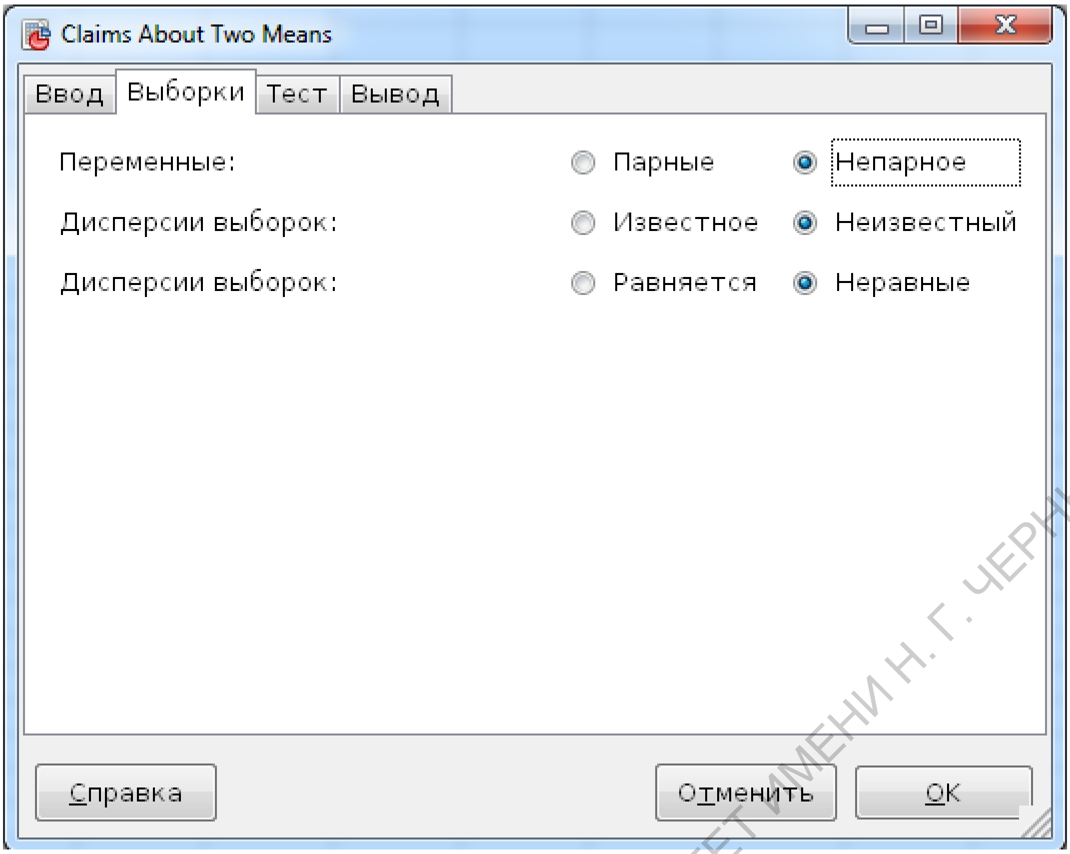
	A	B	C
1	№	X	Y
2	1	504	580
3	2	560	692
4	3	420	700
5	4	600	621
6	5	580	640
7	6	530	561
8	7	490	680
9	8	580	630
10	9	470	

Выясним являются ли распределения X и Y нормальными. Выделяем диапазон ячеек B3:C10. Используем соответственно команды "Статистика/Тесты с одной выборкой/ Критерии нормальности. В диалоговом окне будут четыре критерия, выбираем "Критерий Лиллифора (Колмогорова-Смирнова) далее "OK".

A	B	C
Критерий Лиллифора (Колмогорова-Смирнова)	Столбец 1	Столбец 2
Альфа	0,05	0,05
p-значение	0,7317837417691865	0,7093793858862563
Статистика	0,159537871873613	0,1701088361016358
N	9	8
Заключение	Возможно нормален	Возможно нормален

Из рисунка видно, что распределения X и Y являются нормальными. Находим эмпирическое значение $t_{\text{эмп}}$. Используем соответственно команды „Статистика/ Тесты с двумя выборками/ Claims About Two Means/ Unpaired Samples, Unequal Variances“. Появится диалоговое окно, выбираем соответствующие диапазоны.





	A	B	C
1		*Переменная 1*	*Переменная 2*
2	*Среднее*	526	638
3	*Дисперсия*	3579	2596,285714285714
4	*Наблюдения*	9	8
5	*Гипотетическое среднее отклонение*	0	
6	*df*	-112	
7	*t Stat*	14,981989966234208	
8	*P (T<=t) одностороннее*	-4,167624330177547	
9	*t критическое одностороннее*	0,0004135771475543091	
10	*P (T<=t) двухстороннее*	1,75318875102738	
11	*t критическое двухстороннее*	0,0008271542951086182	
12		2,1316727308742935	

Получим $t_{\text{эмп}} = 4,17$ (ячейка B8, значение по модулю). В ячейке B12 указано критическое значение 2,13 для p=0,05. По таблице находим критическое значение для p=0,01: $t_{\text{кр}} = 2,95$ (приложение).

$t_{\text{эмп}} = 4,17 > 2,95$, $t_{\text{эмп}}$ попало в зону значимости. Гипотеза H_0 о сходстве отклоняется и на 1% уровне значимости принимается гипотеза H_1 - о различии между экспериментальными и контрольными группами.

Случай зависимых выборок

Вычисление значения осуществляется по формуле:

$$t_{\text{эмп}} = \frac{\bar{d}}{Sd}$$

где

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{\sum_{i=1}^{n}(x_i - y_i)}{n},$$

$d_i = x_i - y_i$ - разности между соответствующими значениями переменной X и переменной Y, а $\bar{d}$ среднее этих разностей,

$$Sd = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n^2 - n}}.$$

Число степеней свободы определяется по формуле $k = n - 1$.

Пример 9. В ходе проверки эффективности тренинга каждому из 8 членов группы задавался вопрос „Насколько часто твое мнение совпадает с мнением группы?“ - дважды, до и после тренинга. Для ответов использовалась 10-балльная шкала: 1 - никогда, ..., 5 - в половине случаев, ..., 10 - всегда. Проверялась гипотеза о том, что в результате тренинга самооценка конформизма участников возрастет.

	A	B	C
1		X1	X2
2	1	3	4
3	2	6	6
4	3	5	6
5	4	2	4
6	5	7	6
7	6	3	4
8	7	4	5
9	8	5	6

Используем соответственно команды „Статистика/ Тесты с двумя выборками/ Claims About Two Means/ Paired Samples“. Появится диалоговое окно, выбираем соответствующие диапазоны, далее „OK“.

	A	B	C
1		Переменная 1	Переменная 2
2	Среднее	4,375	5,125
3	Известная дисперсия	2,8392857142857144	0,9821428571428571
4	Наблюдения	8	8
5	Корреляция Пирсона	0,908947959029629	
6	Гипотетическое среднее отклонение	0	
7	Наблюдаемое среднее отклонение	-0,75	
8	Дисперсия отклонений	0,7857142857142857	
9	df	7	
10	t Stat	-2,393172105652397	
11	P (T<=t) одностороннее	0,023972386089811712	
12	t критическое одностороннее	1,8945786050900078	
13	P (T<=t) двухстороннее	0,047944772179623424	
14	t критическое двухстороннее	2,3646242515927853	

Получим $t_{эмп} = 2,39$ (ячейка B10, значение по модулю). В ячейке B14 указано критическое значение 2,36 для p=0,05. По таблице находим критическое значение для p=0,01: $t_{кр} = 3,5$.

$2,39 < t_{эмп} = 4,17 < 3,5$, $t_{эмп}$ попало в зону неопределенности. Таким образом, на 5% уровне значимости гипотеза H_0 отклоняется и принимается гипотеза H_1 - о различиях: показатель самооценки конформизма участников после тренинга увеличился статистически достоверно.

7. Корреляционный анализ

Для решения задач корреляционного анализа используются соответственно команды „Статистика/Описательные статистики/Корреляция“.

Коэффициент корреляции - это количественная мера силы и направления вероятностной взаимосвязи двух переменных; принимает значения в диапазоне от -1 до +1.

Сила связи достигает максимума при условии взаимно однозначного соответствия: когда каждому значению одной переменной соответствует только одно значение другой переменной (и наоборот), эмпирическая взаимосвязь при этом совпадает с функциональной линейной связью. Показателем силы связи является абсолютная величина коэффициента корреляции.

Направление связи определяется прямым или обратным соотношением значений двух переменных: если возрастанию значений одной переменной соответствует возрастание значений другой переменной, то взаимосвязь является обратной (отрицательной). Показателем направления связи является знак коэффициента корреляции.

7.1. Коэффициент корреляции Пирсона

Корреляция Пирсона есть мера линейной связи между двумя переменными. Она позволяет определить, насколько пропорциональна изменчивость двух переменных. Если переменные пропорциональны друг другу, то графически связь между ними можно представить в виде прямой.

В общем виде формула для подсчета коэффициента корреляции такова:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

где x_i - значения, принимаемые переменной X;

y_i - значения, принимаемые переменной Y;

$\bar{x}$ - средняя по X;

$\bar{y}$ - средняя по Y;

n - объем выборки.

Для применения коэффициента корреляции Пирсона необходимо соблюдать следующие условия:

1. Сравниваемые переменные должны быть получены в интервальной шкале или шкале отношений.

2. Распределения переменных X и Y должны быть близки к нормальному.

3. По каждой переменной должно быть представлено не менее 5 наблюдений. Верхняя граница выборки определяется имеющимися таблицами критических значений. Оценка уровня значимости по таблицам осуществляется при числе степеней свободы $k = n - 2$.

Пример 10. 20 школьникам были даны тесты на наглядно-образное и вербальное мышление. Измерялось среднее время решения заданий теста в секундах. Психолога интересует вопрос: существует ли взаимосвязь между временем решения этих задач? Переменная X обозначает среднее время решения наглядно-образных, а переменная Y - среднее время решения вербальных заданий тестов. Исходные данные представлены на рисунке 1.

	A	B	C
1		X	Y
2	№ испытуемых	среднее время решения наглядно-образных заданий	среднее время решения вербальных заданий
3	1	19	17
4	2	32	7
5	3	33	17
6	4	44	28
7	5	28	27
8	6	35	31
9	7	39	20
10	8	39	17
11	9	44	35
12	10	44	43
13	11	24	10
14	12	37	28
15	13	29	13
16	14	40	43
17	15	42	45
18	16	32	24
19	17	48	45
20	18	42	26
21	19	33	16
22	20	47	26

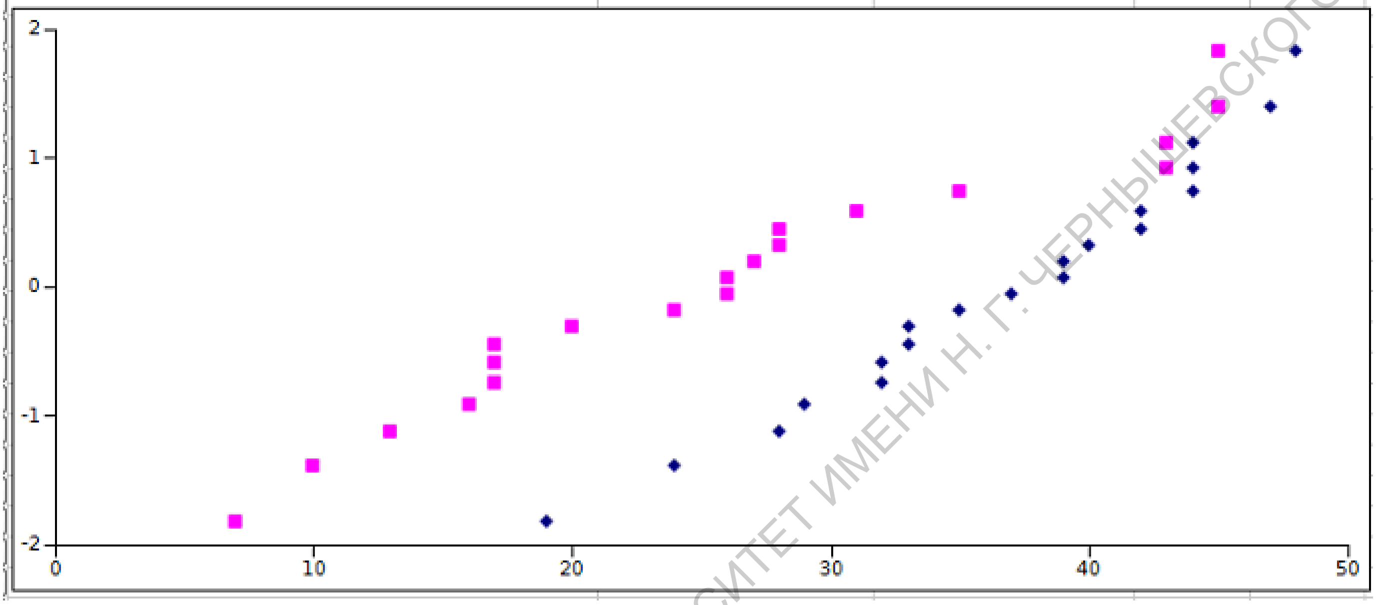
Решение. Для применение коэффициента корреляции Пирсона распределения переменных X и Y должны быть близки к нормальному. Выясним являются ли распределения переменных X и Y из условия задачи нормальными.

Выделяем диапазон ячеек B3:C22. Используем соответственно команды "Статистика/Тесты с одной выборкой/ Критерии нормальности. В диалоговом окне будут четыре критерия, выбираем "Критерий Лиллифора (Колмогорова-Смирнова) далее "ОК".

Критерий Колмогорова-Смирнова позволяет оценить вероятность того, данная выборка принадлежит генеральной совокупности с нормальным распределением. Если эта вероятность $p \leq 0,05$, то данное эмпирическое распределение существенно отличается от нормального, а если $p > 0,5$, то делают вывод о приблизительном соответствии данного эмпирического распределения нормальному. Из рисунка видно, что распределения переменных X и Y являются нормальными.

=if(B2>=B3;"Ненормален";"Возможно нормален")

	A	B	C	D	E
	Критерий Лиллифора (Колмогорова-Смирнова)	Столбец 1	Столбец 2		
	Альфа	0,05	0,05		
	p-значение	0,5896819806364546	0,5202119599897201		
	Статистика	0,12333057019827465	0,1287261953046744		
	N	20	20		
	Заключение	Возможно нормален	Возможно нормален		



Вычислим эмпирическую величину коэффициента корреляции. Выделяем диапазон ячеек B3:C22. Используем соответственно команды "Статистика/Описательные статистики/Корреляция". Эмпирическая величина коэффициента корреляции равна 0,669 (рис. 4).

	A	B	C	D	E
1	Корреляции	Столбец 1	Столбец 2		
2	Столбец 1	1			
3	Столбец 2	0,6692438094742086	1		
4					
5					

При нахождении критических значений для вычисленного коэффициента линейной корреляции Пирсона число степеней свободы рассчитывается как $k = n - 2$. В рассматриваемой задаче $n = 20$, тогда критические значения равны 0,44 при $p \leq 0,05$ и 0,56 при $p \leq 0,01$ (Приложение, таблица 1).

Ответ. Величина расчетного коэффициента корреляции попала в зону значимости $(0,669 > 0,56)$ - гипотеза H_0 отвергается и принимается гипотеза H_1. Значит, связь между временем решения наглядно-образных и вербальных задач статистически значима на 1% уровне и положительна.

7.2. Коэффициент корреляции рангов К. Спирмена

Коэффициент корреляциии рангов, предложенный К. Спирменом, относится к непараметрическим показателям связи между переменными.

Расчет коэффициента ранговой корреляции Спирмена

1. Определить два признака (две иерархии признаков), участвующие в сопоставлении (переменные X и Y).

$$X \quad x_1, x_2, ..., x_n$$

$$Y \quad y_1, y_2, ..., y_n$$

По каждой переменной должно быть представлено не менее 5 наблюдений. Верхняя граница выборки определяется имеющимися таблицами критических значений.

2. Проранжировать значения переменной X, начисляя ранг 1 наименьшему значению, в соответствии с правилами ранжирования. Занести ранги в первый столбец таблицы по порядку номеров испытуемых или признаков.

3. Проранжировать значения переменной Y, в соответствии с правилами ранжирования. Занести ранги во второй столбец таблицы по порядку номеров испытуемых или признаков.

4. Определить разности рангов каждой пары сопоставляемых значений: $d_i = x_i - y_i$.

5. Возвести каждую разность в квадрат d_i^2 и суммировать полученные результаты $\sum_{i=1}^{n} d_i^2$.

6. При наличии одинаковых рангов рассчитать поправки: $T_a = \frac{\sum(a^3-a)}{12}$, $T_b = \frac{\sum(b^3-b)}{12}$, где a - объем каждой группы одинаковых рангов в ранговом ряду X; b - объем каждой группы одинаковых рангов в ранговом ряду Y.

8. Рассчитать коэффициент ранговой корреляции ρ по формуле:

а) при отсутствии одинаковых рангов:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)},$$

б) при наличии одинаковых рангов:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2 + T_a + T_b}{n \cdot (n^2 - 1)},$$

где n - количество ранжируемых признаков (показателей, испытуемых);

d_i - разность между рангами по двум переменным для каждого испытуемого;

T_a, T_b - поправки на одинаковые ранги.

9. Определить по Таблице критические значения для данного n. Если ρ, превышает критическое значение или равен ему, корреляция достоверно отличается от 0.

Пример 11 (пример разных рангов). Психолог выясняет, как связаны между собой индивидуальные показатели готовности к школе, полученные до начала обучения в школе у 11 первоклассников и их средняя успеваемость в конце учебного года. Ранги показателей школьной готовности, полученные при поступлении в школу, и итоговые показатели успеваемости в конце года у учащихся представлены в таблице:

	A	B	C	D	E	F	G	H	I	J	K	L
1	№ учащихся	1	2	3	4	5	6	7	8	9	10	11
2	Ранги показателей школьной готовности	3	5	6	1	4	11	9	2	8	7	10
3	Ранги среднегодовой успеваемости	2	7	8	3	4	6	11	1	10	5	9

Решение.

Вычислим разности рангов каждой пары сопоставляемых значений: $d_i = x_i - y_i$:

B4 = =B2-B3

	A	B	C	D	E	F	G	H	I	J	K	L
1	№ учащихся	1	2	3	4	5	6	7	8	9	10	11
2	Ранги показателей школьной готовности	3	5	6	1	4	11	9	2	8	7	10
3	Ранги среднегодовой успеваемости	2	7	8	3	4	6	11	1	10	5	9
4	d	1	−2	−2	−2	0	5	−2	1	−2	2	1
5	d^2											

Вычислим квадрат каждой разности d_i^2:

B5 = =B4^2

	A	B	C	D	E	F	G	H	I	J	K	L
1	№ учащихся	1	2	3	4	5	6	7	8	9	10	11
2	Ранги показателей школьной готовности	3	5	6	1	4	11	9	2	8	7	10
3	Ранги среднегодовой успеваемости	2	7	8	3	4	6	11	1	10	5	9
4	d	1	−2	−2	−2	0	5	−2	1	−2	2	1
5	d^2	1	4	4	4	0	25	4	1	4	4	1

Найдем сумму квадратов раностей:

B6 = =sum(B5:L5)

	A	B	C	D	E	F	G	H	I	J	K	L
1	№ учащихся	1	2	3	4	5	6	7	8	9	10	11
2	Ранги показателей школьной готовности	3	5	6	1	4	11	9	2	8	7	10
3	Ранги среднегодовой успеваемости	2	7	8	3	4	6	11	1	10	5	9
4	d	1	−2	−2	−2	0	5	−2	1	−2	2	1
5	d^2	1	4	4	4	0	25	4	1	4	4	1
6	Сумма квадратов разностей рангов	52										

Вычислим коэффициент ранговой корреляции ρ по формуле:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}, \qquad n = 11.$$

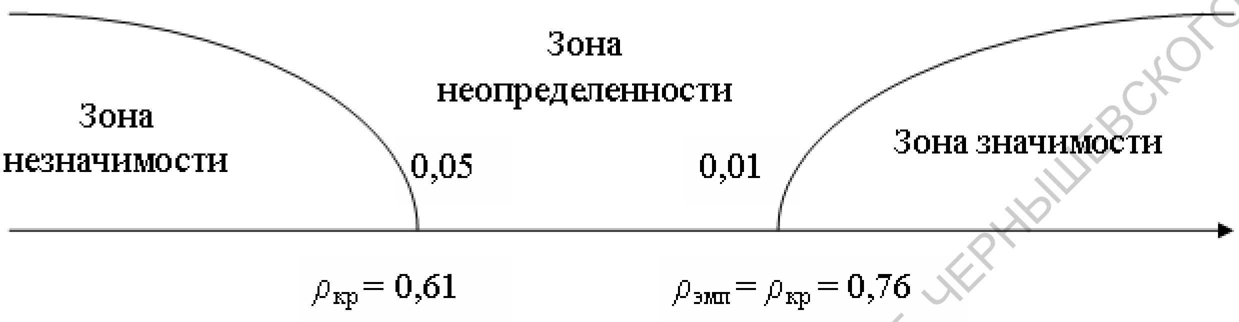
B7 = =1-6*B6/(11*(11^2-1))

	A	B	C	D	E	F	G	H	I	J	K	L
1	№ учащихся	1	2	3	4	5	6	7	8	9	10	11
2	Ранги показателей школьной готовности	3	5	6	1	4	11	9	2	8	7	10
3	Ранги среднегодовой успеваемости	2	7	8	3	4	6	11	1	10	5	9
4	d	1	−2	−2	−2	0	5	−2	1	−2	2	1
5	d^2	1	4	4	4	0	25	4	1	4	4	1
6	Сумма квадратов разностей рангов	52										
7	p	0,76										

Находим критические значения при $n = 11$:

$$\rho_{\text{кр}} = \begin{cases} 0,61 & \text{для} \quad p \le 0,05; \\ 0,76 & \text{для} \quad p \le 0,01. \end{cases}$$

Строим ось значимости:



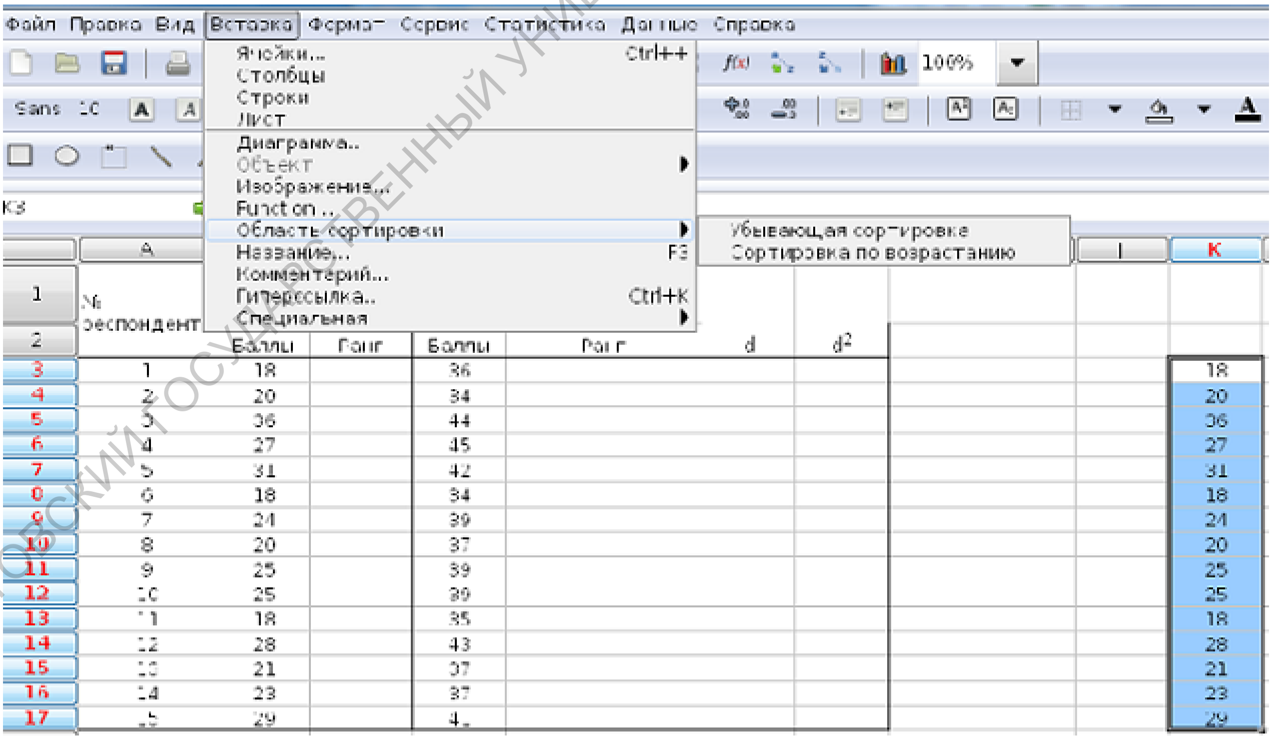
Ответ. Полученный коэффициент корреляции совпал с критическим значением для уровня значимости 1%. Отклоняется нулевая H_0 гипотеза о сходстве и принимается альтернативная H_1 о наличии различий. Следовательно, можно утверждать, что показатели школьной готовности и итоговые оценки первоклассников связаны положительной корреляционной зависимостью: чем выше показатель школьной готовности, тем лучше учится первоклассник.

Пример 12. С помощью тестов оценивались уровни развития уверенности в себе и самоконтроля. По выборке из 15 человек необходимо определить наличие связи между данными признаками. Полученные данные представлены в таблице.

	A	B	C	D	E
1	№ респондента	уровень развития самостоятельности X		степень сформированности креативных способностей Y	
2		Баллы	Ранг	Баллы	Ранг
3	1	18		36	
4	2	20		34	
5	3	36		44	
6	4	27		45	
7	5	31		42	
8	6	18		34	
9	7	24		39	
10	8	20		37	
11	9	25		39	
12	10	25		39	
13	11	18		35	
14	12	28		43	
15	13	21		37	
16	14	23		37	
17	15	29		41	

Решение.

1. Для ранжирования значений переменной X (уровень развития самостоятельности), копируем диапазон B3:B17 в ячейки K3:K17. Сортируем значения переменной X по возрастанию: выделяем диапазон K3:K17 и используем команды „Вставка/ Область сортировки/ Сортировка по возрастанию“.



Проранжируем значения переменной X, начисляя ранг 1 наименьшему значению.

K	L
x	
18	1
18	2
18	3
20	4
20	5
21	**6**
23	**7**
24	**8**
25	9
25	10
27	**11**
28	**12**
29	**13**
31	**14**
36	**15**

Число 18 повторяется 3 раза, этому числу принадлежит 3 ранга – 1, 2, 3. Общий ранг вычисляем следующим образом: $\frac{1+2+3}{3} = 2$. Для вычисления среднего арифметического значения используем команду `average` (f(x)/ Статистические...). Аналогично вычисляем ранг для 20 и 25.

M3 =average(L3:L5)

K	L	M
x		
18	1	2
18	2	
18	3	
20	4	**4,5**
20	5	
21	**6**	
23	**7**	
24	**8**	
25	9	**9,5**
25	10	
27	**11**	
28	**12**	
29	**13**	
31	**14**	
36	**15**	

Вводим полученные ранги в столбец С.

	A	B	C	D	E	F	G
1	№ респондента	уровень развития самостоятельности X		степень сформированности креативных способностей Y			
2		Баллы	Ранг	Баллы	Ранг	d	d^2
3	1	18	2	36			
4	2	20	4,5	34			
5	3	36	15	44			
6	4	27	11	45			
7	5	31	14	42			
8	6	18	2	34			
9	7	24	8	39			
10	8	20	4,5	37			
11	9	25	9,5	39			
12	10	25	9,5	39			
13	11	18	2	35			
14	12	28	12	43			
15	13	21	6	37			
16	14	23	7	37			
17	15	29	13	41			

2. Аналогично проранжируем значения переменной Y (степень сформированности креативных способностей).

O	P	Q
Y		
34	1	**1,5**
34	2	
35	**3**	
36	**4**	
37	5	**6**
37	6	
37	7	
39	8	**9**
39	9	
39	10	
41	**11**	
42	**12**	
43	**13**	
44	**14**	
45	**15**	

Вычислим разности рангов каждой пары сопоставляемых значений: $d_i = x_i - y_i$:

	A	B	C	D	E	F	G
1	№ респондента	уровень развития самостоятельности X		степень сформированности креативных способностей Y			
2		Баллы	Ранг	Баллы	Ранг	d	d^2
3	1	18	2	36	4		
4	2	20	4,5	34	1,5		
5	3	36	15	44	14		
6	4	27	11	45	15		
7	5	31	14	42	12		
8	6	18	2	34	1,5		
9	7	24	8	39	9		
10	8	20	4,5	37	6		
11	9	25	9,5	39	9		
12	10	25	9,5	39	9		
13	11	18	2	35	3		
14	12	28	12	43	13		
15	13	21	6	37	6		
16	14	23	7	37	6		
17	15	29	13	41	11		

F3 = =C3-E3

	A	B	C	D	E	F	G
1	№ респондента	уровень развития самостоятельности X		степень сформированности креативных способностей Y			
2		Баллы	Ранг	Баллы	Ранг	d	d^2
3	1	18	2	36	4	=C3-E3	
4	2	20	4,5	34	1,5		
5	3	36	15	44	14		
6	4	27	11	45	15		
7	5	31	14	42	12		
8	6	18	2	34	1,5		
9	7	24	8	39	9		
10	8	20	4,5	37	6		
11	9	25	9,5	39	9		
12	10	25	9,5	39	9		
13	11	18	2	35	3		
14	12	28	12	43	13		
15	13	21	6	37	6		
16	14	23	7	37	6		
17	15	29	13	41	11		

Вычислим квадрат каждой разности d_i^2:

G3 =F3^2

	A	B	C	D	E	F	G
1	№ респондента	уровень развития самостоятельности X		степень сформированности креативных способностей Y			
2		Баллы	Ранг	Баллы	Ранг	d	d^2
3	1	18	2	36	4	−2	=F3^2
4	2	20	4,5	34	1,5	3	
5	3	36	15	44	14	1	
6	4	27	11	45	15	−4	
7	5	31	14	42	12	2	
8	6	18	2	34	1,5	0,5	
9	7	24	8	39	9	−1	
10	8	20	4,5	37	6	−1,5	
11	9	25	9,5	39	9	0,5	
12	10	25	9,5	39	9	0,5	
13	11	18	2	35	3	−1	
14	12	28	12	43	13	−1	
15	13	21	6	37	6	0	
16	14	23	7	37	6	1	
17	15	29	13	41	11	2	

Найдем сумму квадратов раностей:

G19 =sum(G3:G17)

sum(►[values]◄;...)
values: список складываемых значений
Ctrl-F4 закрывает подсказку

	A	B	C	D	E	F	G
1	№ респондента	уровень развития самостоятельности X					
2		Баллы	Ранг	Баллы	Ранг	d	d^2
3	1	18	2	36	4	−2	4
4	2	20	4,5	34	1,5	3	9
5	3	36	15	44	14	1	1
6	4	27	11	45	15	−4	16
7	5	31	14	42	12	2	4
8	6	18	2	34	1,5	0,5	0,25
9	7	24	8	39	9	−1	1
10	8	20	4,5	37	6	−1,5	2,25
11	9	25	9,5	39	9	0,5	0,25
12	10	25	9,5	39	9	0,5	0,25
13	11	18	2	35	3	−1	1
14	12	28	12	43	13	−1	1
15	13	21	6	37	6	0	0
16	14	23	7	37	6	1	1
17	15	29	13	41	11	2	4
18							
19	сумма квадратов разностей рангов						=sum(G3:G17

Вычислим поправки на одинаковые ранги: T_a и T_b по формулам: $T_a = \frac{\sum(a^3-a)}{12}$, $T_b = \frac{\sum(b^3-b)}{12}$.

Число 18 повторяется три раза. Поэтому в ячейке N3 вводим формулу

„$= 3^3 - 3$“:

K	L	M	N
X			Ta
18	1	2	=3^3-3
18	2		
18	3		
20	4	4,5	
20	5		
21	6		
23	7		
24	8		
25	9	9,5	
25	10		
27	11		
28	12		
29	13		
31	14		
36	15		

Число 20 повторяется 2 раза. Поэтому в ячейке N6 вводим формулу „$= 2^3 - 2$“. Аналогично для числа 25 в ячейке N11. Находим T_a:

K	L	M	N	O	P	Q
X			Ta	Y		
18	1	2	24	34	1	1,5
18	2			34	2	
18	3			35	3	
20	4	4,5	6	36	4	
20	5			37	5	6
21	6			37	6	
23	7			37	7	
24	8			39	8	9
25	9	9,5	6	39	9	
25	10			39	10	
27	11			41	11	
28	12			42	12	
29	13			43	13	
31	14			44	14	
36	15			45	15	
			=(N3+N6+N11)/12			

Аналогично вычисляем T_b.

K	L	M	N	O	P	Q	R
X			Ta	Y			Tb
18	1	**2**	24	**34**	1	**1,5**	6
18	2			34	2		
18	3			**35**	**3**		
20	4	**4,5**	6	**36**	**4**		
20	5			**37**	5	**6**	24
21	**6**			37	6		
23	**7**			37	7		
24	**8**			**39**	8	**9**	24
25	9	**9,5**	6	39	9		
25	10			39	10		
27	**11**			**41**	**11**		
28	**12**			**42**	**12**		
29	**13**			**43**	**13**		
31	**14**			**44**	**14**		
36	**15**			**45**	**15**		
			3				**4,5**

Вычислим коэффициент ранговой корреляции К. Спирмена по формуле:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2 + T_a + T_b}{n \cdot (n^2 - 1)}, \quad n = 15,$$

d_i^2 - ячейка G19, T_a - ячейка N18, T_b - ячейка R18.

B21 =1-(6*G19+N18+R18)/(15*(15^2-1))

	A	B	C	D	E	F	G
1	№ респондента	уровень развития самостоятельности X		степень сформированности креативных способностей Y			
2		Баллы	Ранг	Баллы	Ранг	d	d^2
3	1	18	2	36	4	−2	4
4	2	20	4,5	34	1,5	3	9
5	3	36	15	44	14	1	1
6	4	27	11	45	15	−4	16
7	5	31	14	42	12	2	4
8	6	18	2	34	1,5	0,5	0,25
9	7	24	8	39	9	−1	1
10	8	20	4,5	37	6	−1,5	2,25
11	9	25	9,5	39	9	0,5	0,25
12	10	25	9,5	39	9	0,5	0,25
13	11	18	2	35	3	−1	1
14	12	28	12	43	13	−1	1
15	13	21	6	37	6	0	0
16	14	23	7	37	6	1	1
17	15	29	13	41	11	2	4
18							
19	сумма квадратов разностей рангов						45
20							
21	коэффициент ранговой корреляции К. Спирмена	=1-(6*G19+N18+R18)/(15*(15^2-1))					

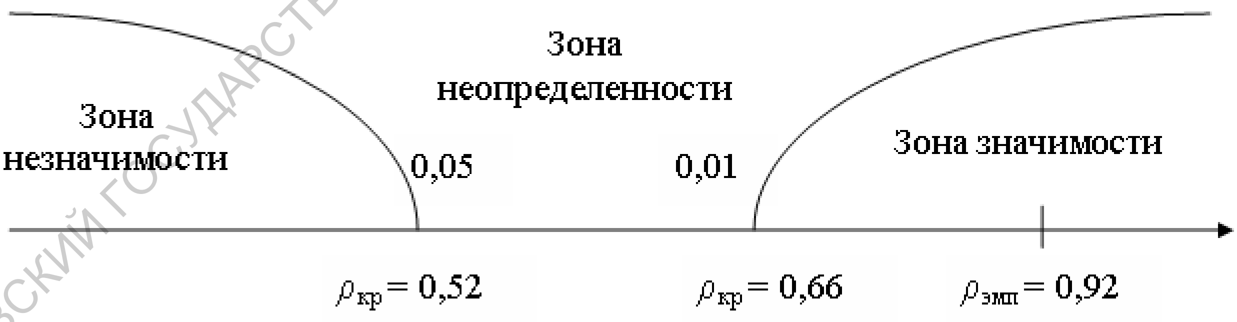
B21 =1-(6*G19+N18+R18)/(15*(15^2-1))

	A	B	C	D	E	F	G
1	№ респондента	уровень развития самостоятельности X		степень сформированности креативных способностей Y			
2		Баллы	Ранг	Баллы	Ранг	d	d^2
3	1	18	2	36	4	−2	4
4	2	20	4,5	34	1,5	3	9
5	3	36	15	44	14	1	1
6	4	27	11	45	15	−4	16
7	5	31	14	42	12	2	4
8	6	18	2	34	1,5	0,5	0,25
9	7	24	8	39	9	−1	1
10	8	20	4,5	37	6	−1,5	2,25
11	9	25	9,5	39	9	0,5	0,25
12	10	25	9,5	39	9	0,5	0,25
13	11	18	2	35	3	−1	1
14	12	28	12	43	13	−1	1
15	13	21	6	37	6	0	0
16	14	23	7	37	6	1	1
17	15	29	13	41	11	2	4
18							
19	сумма квадратов разностей рангов						45
20							
21	коэффициент ранговой корреляции К. Спирмена	0.9174107142857143					

Находим критические значения при $n = 15$:

$$\rho_{\text{кр}} = \begin{cases} 0,52 & \text{для} \quad p \le 0,05; \\ 0,66 & \text{для} \quad p \le 0,01. \end{cases}$$

Строим ось значимости:



Ответ. Коэффициент ранговой корреляции находится в зоне значимости. Нулевая гипотеза о сходстве коэффициента корреляции с нулем отклоняется. Принимается альтернативная гипотеза о значимости отличия коэффициента корреляции от нуля, т.е. принимается гипотеза о связи между уровнем развития самостоятельности и степенью сформированности креативных способностей.

7.3. Коэффициент корреляции φ

При сравнении двух переменных, измеренных в дихотомической шкале, мерой корреляционной связи служит коэффициент φ или коэффициент ассоциации.

Формула вычисления коэффициента корреляции $\varphi_{\text{эмп}}$:

$$\varphi_{\text{эмп}} = \frac{pxy - px \cdot py}{\sqrt{px \cdot (1 - px) \cdot py \cdot (1 - py)}}$$

где px - частота или доля признака, имеющего 1 по X, py - частота или доля признака, имеющего 1 по Y, pxy - частота или доля признака, имеющего 1 одновременно по X и по Y.

Частоты вычисляются следующим образом: подсчитывается количество 1 в переменной X и полученная величина делится на общее число элементов этой переменной - n. Аналогично подсчитываются частоты для переменной Y. Обозначение pxy - соответствует частоте или доле признаков, имеющих единицу как по X, так и по Y.

Пример 13. Влияет ли семейное положение на успешность учебы студентов-мужчин? Данные представлены в таблице: X - семейное положение (1 - женат, 0 - холост), Y - успешность обучения (0 - успешно, 1 - неуспешно).

	A	B	C
1		X 0-холост 1-женат	Y 0-успешно 1-неуспешно
2	1	0	0
3	2	1	1
4	3	0	1
5	4	0	0
6	5	1	1
7	6	1	0
8	7	0	0
9	8	1	1
10	9	0	0
11	10	0	1
12	11	0	0
13	12	1	1

Решение.

Первый способ. Вычислим количество единиц по X: в ячейке B14 находим сумму элементов ячеек B2:B13. Аналогично - C14 („=sum(C2:C13)“).

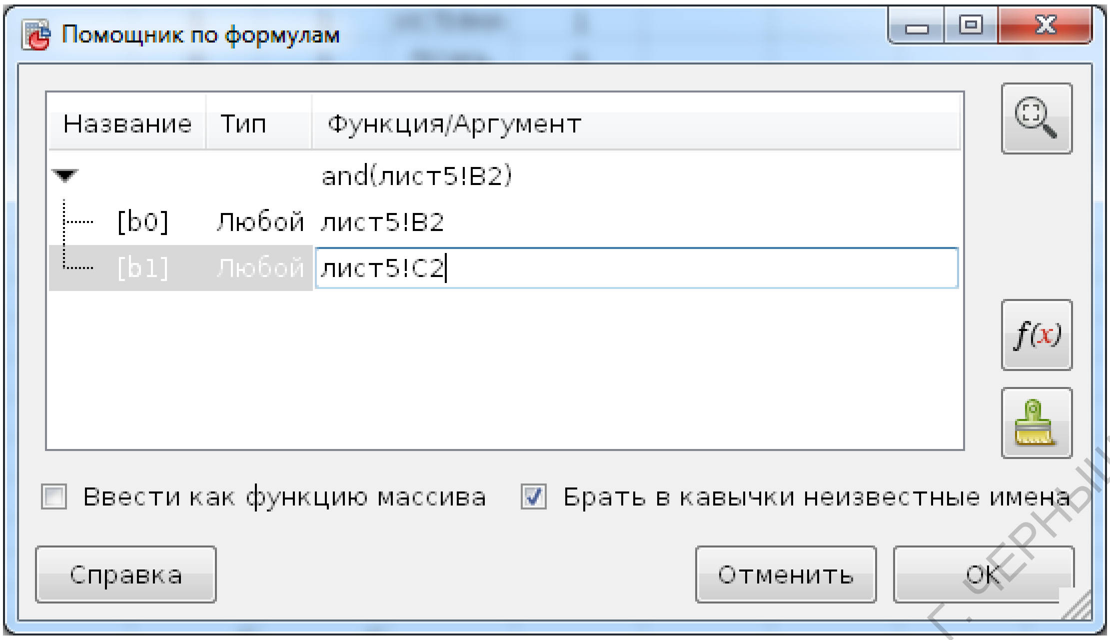
Тогда px (ячейка B15) - пять единиц (ячейка B14), деленное на общее число студентов, принявших участие в эксперименте. Аналогично py (ячейка C15) - общее число единиц по Y (ячейка C14), деленое на 12.

B15 = =B14/12

	A	B	C	D
1		X 0-холост 1-женат	Y 0-успешно 1-неуспешно	
2	1	0	0	
3	2	1	1	
4	3	0	1	
5	4	0	0	
6	5	1	1	
7	6	1	0	
8	7	0	0	
9	8	1	1	
10	9	0	0	
11	10	0	1	
12	11	0	0	
13	12	1	1	
14	сумма	5	6	
15	частота	=B14/12		

Подсчитаем pxy - долю студентов, имеющих единицу как по X, так и по Y. В нашем случае объем выборки мальнекий, можно количество единиц одновременно по X и по Y вычислить вручную. Если вычислить с использованием электронных таблиц, то можно 1) вычислить логическое умножение (конъюнкцию) переменных X и Y; 2) найти количество единиц в 1).

Выделяем ячейку D2 и выбираем кнопку f(x) в панели инструментов. В категории „Логические“ выбираем `and logical conjunction`. Далее появится диалоговое окно: выбираем B2 и C2 в сооответствующих строках:

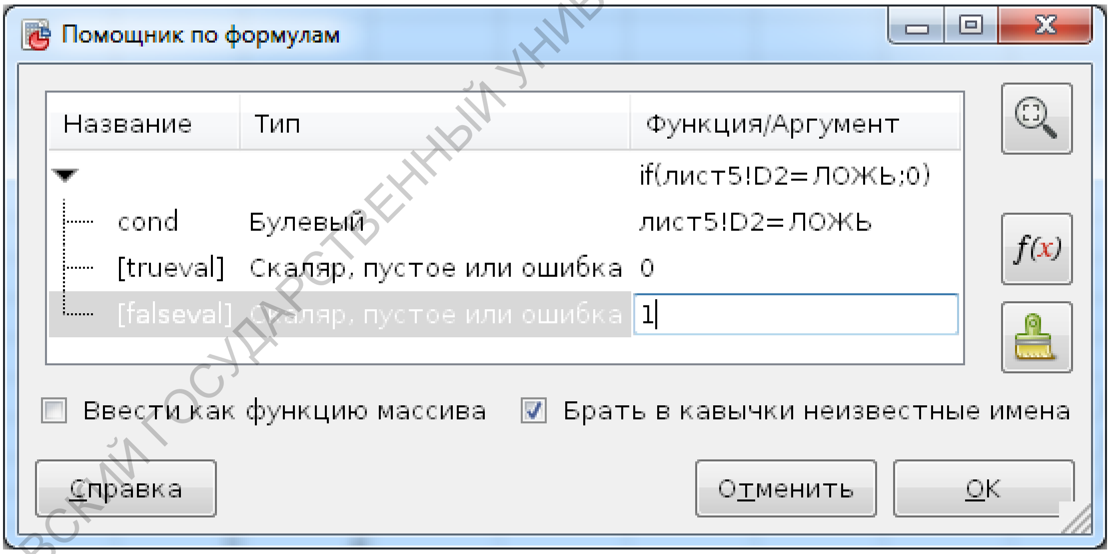


Копируем формулу из ячейки D2 до D13. Выделяем ячейку E2 и выбираем кнопку f(x) в панели инструментов. В категории „Логические“ выбираем `if` `выражение условия`. Далее появится диалоговое окно:

`cond` - выделяем ячейку D2 и набираем „=ЛОЖЬ“;

`trueval` - набираем „0“;

`falseval` - набираем „1“.



Копируем формулу из ячейки E2 до E13. В ячейке E14 находим сумму элементов ячеек E2:E13 (количество единиц одновременно по X и по Y). В ячейке E15 вычисляем частоту pxy - долю студентов, имеющих единицу как по X, так и по Y.

E15 = =E14/12

	A	B	C	D	E
1		X 0-холост 1-женат	Y 0-успешно 1-неуспешно	X&Y	
2	1	0	0	ЛОЖЬ	0
3	2	1	1	ИСТИНА	1
4	3	0	1	ЛОЖЬ	0
5	4	0	0	ЛОЖЬ	0
6	5	1	1	ИСТИНА	1
7	6	1	0	ЛОЖЬ	0
8	7	0	0	ЛОЖЬ	0
9	8	1	1	ИСТИНА	1
10	9	0	0	ЛОЖЬ	0
11	10	0	1	ЛОЖЬ	0
12	11	0	0	ЛОЖЬ	0
13	12	1	1	ИСТИНА	1
14	сумма	5	6		4
15	частота	0,4166667	0,5		0,3333333

Используя формулу (в скобках указаны соответствующие ячейки)

$$\varphi_{\text{эмп}} = \frac{pxy(E15) - px(B15) \cdot py(C15)}{\sqrt{px \cdot (1 - px) \cdot py \cdot (1 - py)}},$$

находим $\varphi_{\text{эмп}}$ (ячейка B19).

B18 = =sqrt(B15*(1-B15)*C15*(1-C15))

sqrt(▶x◀)
x: non-negative number

Ctrl-F4 закрывает подсказку

	A	B	C	D	E
1		X 0-холост 1-женат	0-ус 1-неуспешно	X&Y	
2	1	0	0	ЛОЖЬ	0
3	2	1	1	ИСТИНА	1
4	3	0	1	ЛОЖЬ	0
5	4	0	0	ЛОЖЬ	0
6	5	1	1	ИСТИНА	1
7	6	1	0	ЛОЖЬ	0
8	7	0	0	ЛОЖЬ	0
9	8	1	1	ИСТИНА	1
10	9	0	0	ЛОЖЬ	0
11	10	0	1	ЛОЖЬ	0
12	11	0	0	ЛОЖЬ	0
13	12	1	1	ИСТИНА	1
14	сумма	5	6		4
15	частота	0,4166667	0,5		0,3333333
16					
17	числитель	0,1250000			
18	знаменатель	=sqrt(B15*(1-B15)*C15*(1-C15))			
19	**эмп.значение**				

B19 = =B17/B18

	A	B	C	D	E
1		X 0-холост 1-женат	Y 0-успешно 1-неуспешно	X&Y	
2	1	0	0	ЛОЖЬ	0
3	2	1	1	ИСТИНА	1
4	3	0	1	ЛОЖЬ	0
5	4	0	0	ЛОЖЬ	0
6	5	1	1	ИСТИНА	1
7	6	1	0	ЛОЖЬ	0
8	7	0	0	ЛОЖЬ	0
9	8	1	1	ИСТИНА	1
10	9	0	0	ЛОЖЬ	0
11	10	0	1	ЛОЖЬ	0
12	11	0	0	ЛОЖЬ	0
13	12	1	1	ИСТИНА	1
14	сумма	5	6		4
15	частота	0,4166667	0,5		0,3333333
16					
17	числитель	0,1250000			
18	знаменатель	0,2465033			
19	**эмп.значение**	0,5070926			

Коэффициент корреляции φ не имеет стандартных таблиц для нахождения критических значений. Поиск критических значений осуществляется с помощью t-критерия Стьюдента по формуле:

$$T\phi = |\varphi_{\text{эмп}}| \cdot \sqrt{\frac{n-2}{1-\varphi^2_{\text{эмп}}}},$$

где n - число коррелируемых признаков, величина $T\phi$ проверяется на уровень значимости по таблице для t-критерия Стьюдента. Число степеней свободы в этом случае будет равно $k = n - 2$.

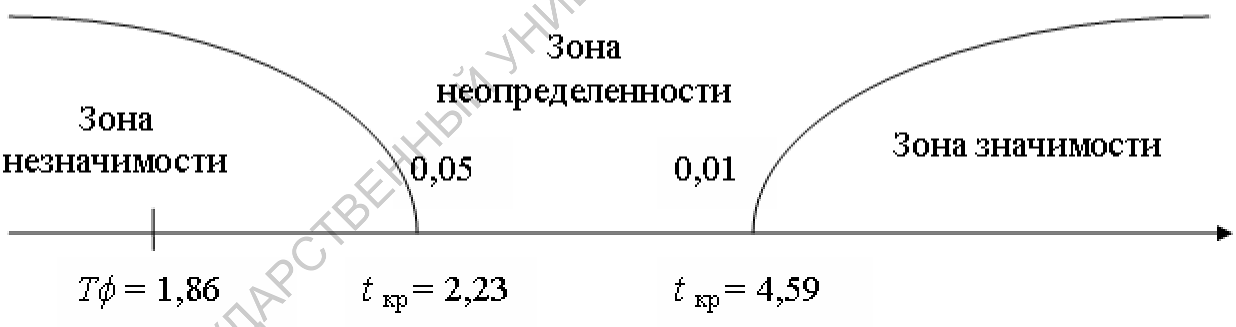
В ячейке B20 вводим формулу для вычисления $T\phi$:

19	эмп.значение	0,5070926
20	T	=abs(B19)*sqrt(10/(1-B19^2))

Число степеней свободы равно $k = n - 2$=12-2=10. Для $k = 10$ находим критические значения критерия Стьюдента:

$$t_{\text{кр}} = \begin{cases} 2,23 & \text{для} \quad p \leq 0,05; \\ 4,59 & \text{для} \quad p \leq 0,01. \end{cases}$$

Строим ось значимости:



Значение $T\phi$ попало в зону незначимости, т.е. нет никакой связи между успешностью обучения и семейным положением студентов. Гипотеза H_1 отклоняется и принимается гипотеза H_0 о сходстве коэффициента корреляции φ с нулем.

Второй способ. Представим данные этой задачи в виде таблицы (называется таблица сопряженности):

Значение признаков	Семейное положение	
	Холостые	Женатые
Плохо учится	$a = 2$	$b = 4$
Учится хорошо	$c = 5$	$d = 1$

Формула расчета коэффициента φ по таблице сопряженности:

$$\varphi = \frac{bc - ad}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}.$$

Далее вычисляется $T\phi$ по формуле из первого способа.

7.4. Бисериальный коэффициент корреляции

В тех случаях, когда одна переменная измеряется в дихотомической шкале (переменная X), а другая в шкале интервалов или отношений (переменная Y), используется бисериальный коэффициент корреляции. Переменная X, полученная в дихотомической шкале, принимает только два значения 0 и 1.

Расчет коэффициента производится по формуле

$$R_{\text{эмп}} = \frac{\bar{x}_1 - \bar{x}_0}{S_y} \cdot \sqrt{\frac{n_1 n_0}{n^2 - n}},$$

где

$\bar{x}_1$ - среднее по тем элементам переменной Y, которым соответствует 0 в переменной X,

n_0 - количество нулей в переменной X,

$n = n_1 + n_0$ - общее количество элементов в переменной X,

S_y - стандартное отклонение переменной Y, вычисляемое по формуле:

$$S_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Бисериальный коэффициент корреляции не имеет стандартных таблиц для нахождения критических значений. Поиск критических значений осуществляется с помощью t-критерия Стьюдента по формуле:

$$T\phi = |R_{\text{эмп}}| \cdot \sqrt{\frac{n-2}{1 - R^2_{\text{эмп}}}},$$

величина $T\phi$ проверяется на уровень значимости по таблице для t-критерия Стьюдента. Число степеней свободы в этом случае будет равно $k = n - 2$.

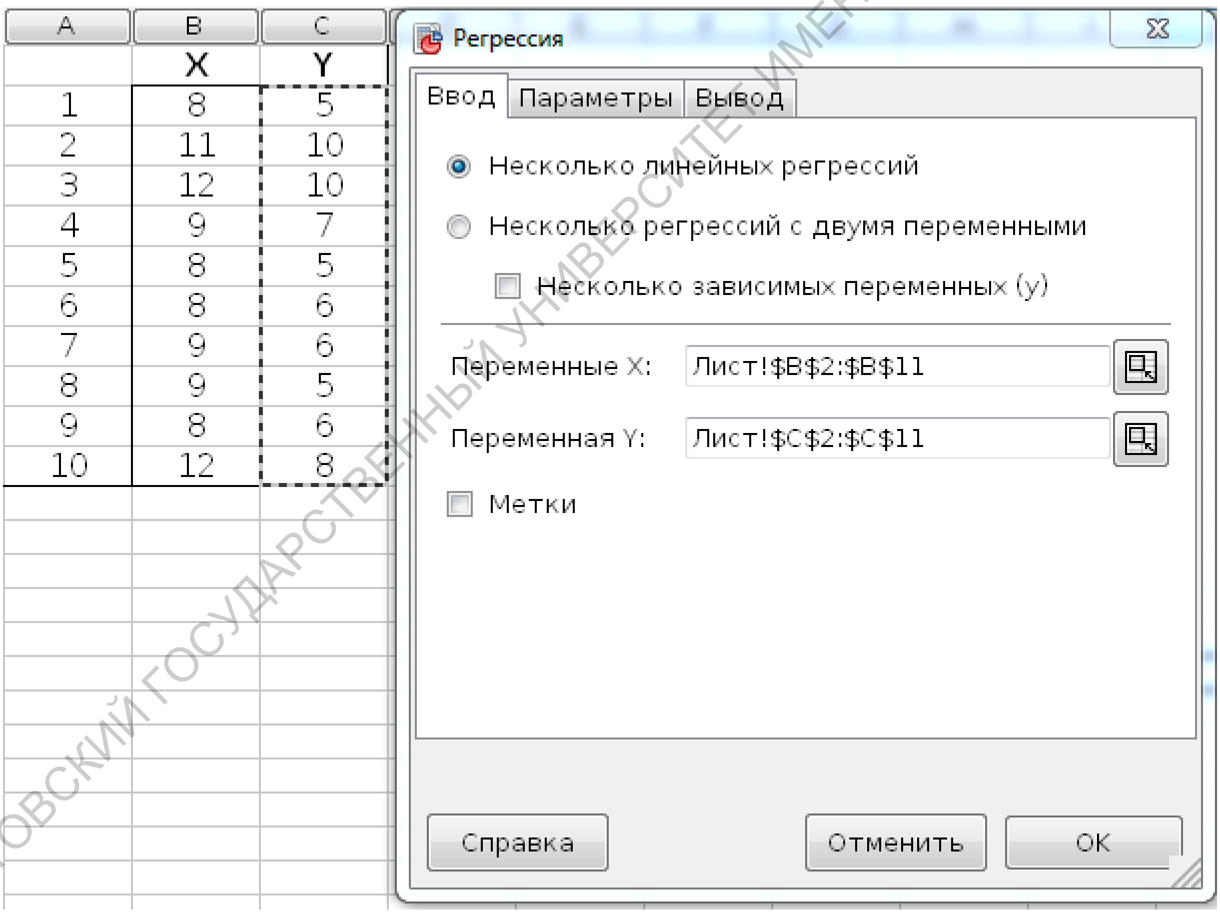
8. Регрессионный анализ

Для решения задач регрессионного анализа используются команды „Статистика/Зависимые наблюдения/ Регрессия“.

8.1. Линейная регрессия

Исследуем уравнения линейной регрессии вида $\hat{y} = f(x) = a + bx$.

Пример 14. Пусть даны значения независимой переменной X и зависимой переменной Y (вводим данные в диапазон A1:C11). Используем команды: „Статистика/Зависимые наблюдения/ Регрессия“, появится диалоговое окно. Выделяем соответствующие ячейки для переменных X и Y:



На отдельном листе появится таблица (результат представлен на следующих двух рисунках):

	A	B	C	D
1	*Итоговый вывод*		*Response Variable*	*Столбец 3*
2				
3	*Регрессионные статистики*			
4	*Множественная R*	0,86614		
5	*R^2*	0,75020		
6	*Стандартная ошибка*	1,02430		
7	*Вычисленная R*	0,71897		
8	*Наблюдения*	10		
9				
10	*Дисперсионный анализ*			
11		*степень свободы*	*сумма квадратов*	*Квадрат среднего*
12	*Регрессия*	1	25,20656	25,20656
13	*Остатки*	8	8,39344	1,04918
14	*Всего*	9	33,6	
15				
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*
17	*Пересечение*	−2,75410	1,97594	−1,39382
18	*Столбец 2*	1,01639	0,20736	4,90153
19				
20	*Ограничение*	*Столбец 2*	*Prediction*	*Столбец 3*
21	1	8	5,37705	5
22	1	11	8,42623	10
23	1	12	9,44262	10
24	1	9	6,39344	7
25	1	8	5,37705	5
26	1	8	5,37705	6
27	1	9	6,39344	6
28	1	9	6,39344	5
29	1	8	5,37705	6
30	1	12	9,44262	8

E	F	G	H	I
F	*Значимость F*			
24,02500	0,00119			
p-значение	*Ниже 95%*	*Выше 95%*		
0,20087	−7,31062	1,80242		
0,00119	0,53821	1,49457		
Residual	*Leverages*	*Internally studentized*	*Externally studentized*	*p-Value*
−0,37705	0,18033	−0,40659	−0,38359	71,27%
1,57377	0,20492	1,72310	1,91975	9,64%
0,55738	0,37705	0,68944	0,65719	53,21%
0,60656	0,10656	0,62649	0,59931	56,78%
−0,37705	0,18033	−0,40659	−0,38359	71,27%
0,62295	0,18033	0,67175	0,64342	54,05%
−0,39344	0,10656	−0,40637	−0,38368	71,26%
−1,39344	0,10656	−1,43923	−1,53556	16,85%
0,62295	0,18033	0,67175	0,64342	54,05%
−1,44262	0,37705	−1,78444	−1,92478	9,57%

Значение коэффициента $a = -2,7541$ (ячейка B17).

Значение коэффициента $b = 1,01639$ (ячейка B18).

Уравнение регрессии примет вид $\hat{y} = f(x) = -2,7541 + 1,01639x$.

Выборочный коэффициент корреляции $r_{xy} = 0,86614$ (ячейка B4).

Коэффициент детерминации R^2 равен 0,7502 (ячейка B5).

Среднеквадратическое отклонение y равно 1,0243 (ячейка B6).

Число степеней свободы равно 8 (ячейка B13).

Регрессионная сумма квадратов равна 25,20656 (ячейка C12).

Остаточная сумма квадратов равна 8,39344 (ячейка C13).

F-статистика 24,025 (ячейка E12).

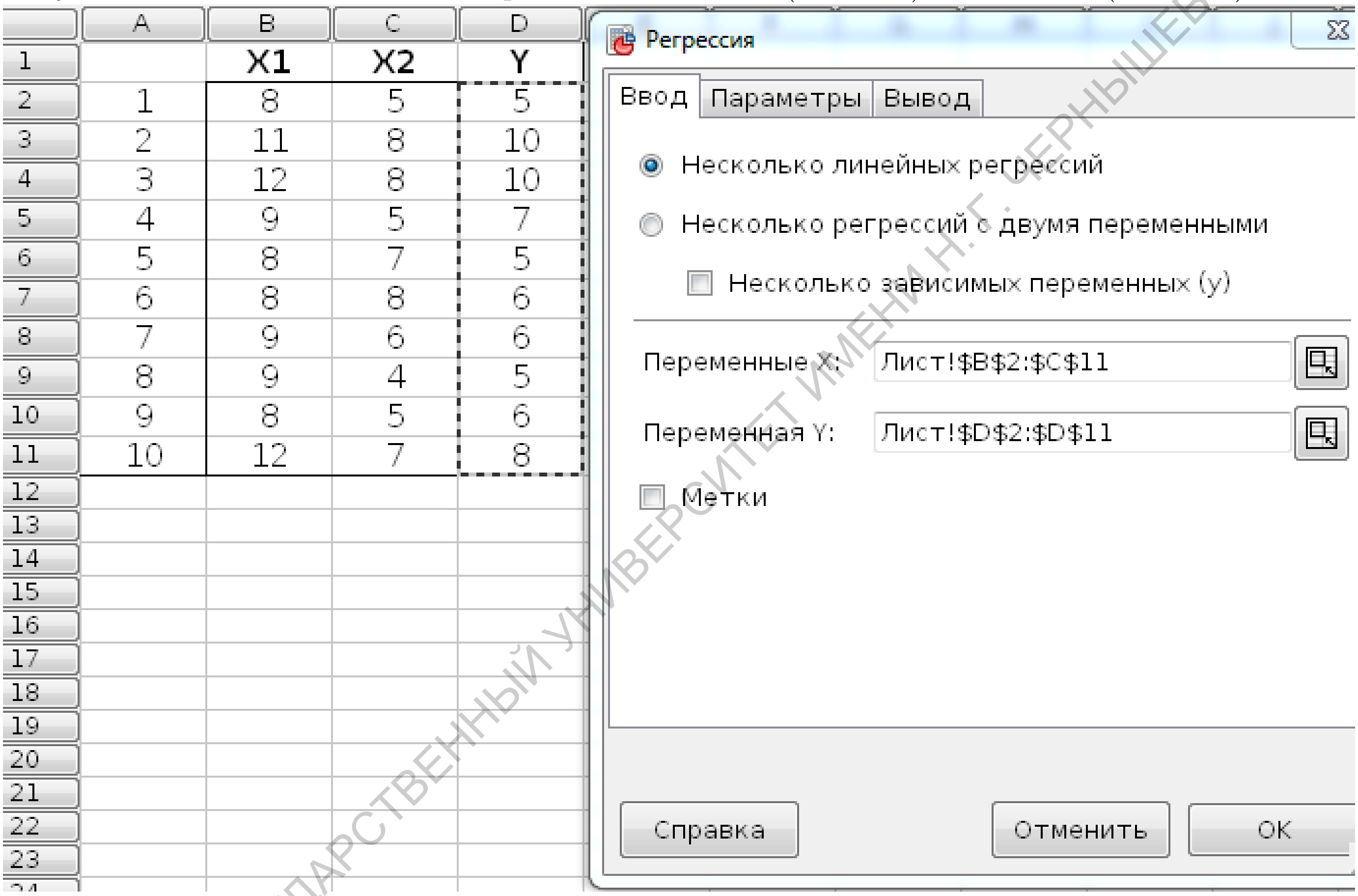
Среднеквадратическое отклонение a равно 1,97594 (ячейка C17).

Среднеквадратическое отклонение b равно 0,20736 (ячейка C18).

8.2. Множественная линейная регрессия

Исследуем уравнения линейной регрессии вида $\hat{y}(x_1, x_2) = b_0 + b_1 x_1 + b_2 x_2$.

Пример 15. Пусть даны значения независимой переменной X_1, X_2 и зависимой переменной Y (вводим данные в диапазон A1:D11). Используем команды: „Статистика/Зависимые наблюдения/ Регрессия“, появится диалоговое окно. Выделяем соответствующие ячейки для переменных X (B2:C11) и Y (D2:D11):



Результаты работы приведены на следующих двух рисунках:

	A	B	C	D
1	*Итоговый вывод*		*Response Variable*	*Столбец 4*
2				
3	*Регрессионные статистики*			
4	*Множественная R*	0,90090		
5	*R^2*	0,81162		
6	*Стандартная ошибка*	0,95091		
7	*Вычисленная R*	0,75780		
8	*Наблюдения*	10		
9				
10	*Дисперсионный анализ*			
11		*степень свободы*	*сумма квадратов*	*Квадрат среднего*
12	*Регрессия*	2	27,27041	13,63521
13	*Остатки*	7	6,32959	0,90423
14	*Всего*	9	33,6	
15				
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*
17	*Пересечение*	−3,53933	1,90658	−1,85637
18	*Столбец 2*	0,85393	0,22050	3,87263
19	*Столбец 3*	0,36704	0,24295	1,51078
20				
21	*Ограничение*	*Столбец 2*	*Столбец 3*	*Prediction*
22	1	8	5	5,12734
23	1	11	8	8,79026
24	1	12	8	9,64419
25	1	9	5	5,98127
26	1	8	7	5,86142
27	1	8	8	6,22846
28	1	9	6	6,34831
29	1	9	4	5,61423
30	1	8	5	5,12734
31	1	12	7	9,27715

E	F	G	H	I	J
F	*Значимость F*				
15,07941	0,00290				
p-значение	*Ниже 95%*	*Выше 95%*			
0,10577	−8,04767	0,96902			
0,00611	0,33252	1,37534			
0,17460	−0,20744	0,94152			
Столбец 4	*Residual*	*Leverages*	*Internally studentized*	*Externally studentized*	*p-Value*
5	−0,12734	0,21054	−0,1507	−0,1397	89,35%
10	1,20974	0,26913	1,4881	1,5713	16,72%
10	0,35581	0,39674	0,4817	0,4505	66,81%
7	1,01873	0,18887	1,1895	1,2044	27,38%
5	−0,86142	0,29401	−1,0781	−1,0624	32,89%
6	−0,22846	0,53157	−0,3510	−0,3263	75,52%
6	−0,34831	0,10754	−0,3877	−0,3625	72,94%
5	−0,61423	0,40075	−0,8344	−0,7966	45,60%
6	0,87266	0,21054	1,0329	1,0195	34,73%
8	−1,27715	0,39032	−1,7201	−1,8484	11,40%

Дадим краткую интерпретацию получившимся показателям.

Значение коэффициента $b_0 = -3,53933$ (ячейка B17).

Значение коэффициента $b_1 = 0,85393$ (ячейка B18).

Значение коэффициента $b_2 = 0,36704$ (ячейка B19).

Уравнение регрессии примет вид: $\hat{y}(x_1, x_2) = -3,53933 + 0,85393x_1 + 0,36704x_2$.

Множественный R - корень квадратный из коэффициента детерминации (ячейка B4). R-квадрат - коэффициент детерминации R^2 (ячейка B5).

Стандартная ошибка - оценка s для среднеквадратического отклонения σ (ячейка B6).

Нормированный R - квадрат - приведенный коэффициент детерминации (ячейка B7).

Наблюдения - число наблюдений n.

9. Дисперсионный анализ

Для решения задач дисперсионного анализа используются команды „Статистика/Тесты с множеством выборок/Дисперсионный анализ...“.

Ограничения метода однофакторного дисперсионного анализа для несвязанных выборок

1. Однофакторный дисперсионный анализ требует не менее трех градаций фактора и не менее двух испытуемых в каждой градации.

2. Результативный признак должен быть нормально распределен в исследуемой выборке.

Пример 16. Три различные группы из шести испытуемых получили списки из десяти слов. Первой группе слова предъявлялись с низкой скоростью - 1 слово в 5 секунд, второй группе со средней скоростью - 1 слово в 2 секунды, и третьей группе с большой скоростью - 1 слово в секунду. Было предсказано, что показатели воспроизведения будут зависеть от скорости предъявления слов. Результаты представлены в таблице

	A	B	C	D
1	№ испытуемого	Группа 1: низкая скорость	Группа 2: средняя скорость	Группа 3: высокая скорость
2	1	8	7	4
3	2	7	8	5
4	3	9	5	3
5	4	5	4	6
6	5	6	6	2
7	6	8	7	4

H_0: Различия в объеме воспроизведения слов между группами являются не более выраженными, чем случайные различия внутри каждой группы.

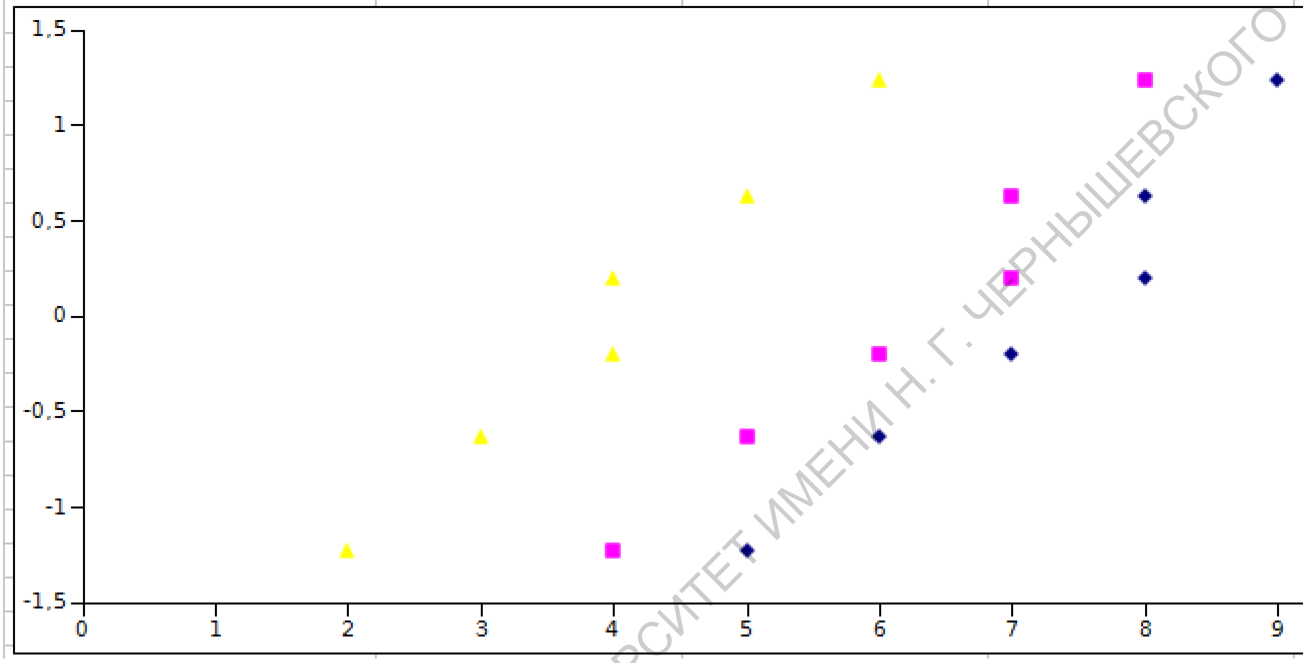
H_1: Различия в объеме воспроизведения слов между группами являются более выраженными, чем случайные различия внутри каждой группы.

Выясним являются ли распределения переменных из условия задачи нормальными.

Выделяем диапазон ячеек B2:D7. Используем соответственно команды „Статистика/Тесты с одной выборкой/ Критерии нормальности“. В диалоговом окне будут четыре критерия, выбираем „Критерий Лиллифора (Колмогорова-Смирнова)“, далее „OK“.

Из рисунка видно, что распределения переменных являются нормальными.

A	B	C	D
Критерий Лиллифора (Колмогорова-Смирнова)	Столбец 1	Столбец 2	Столбец 3
Альфа	0,05	0,05	0,05
p-значение	0,5245158368876769	0,5245158368876769	0,8667553798128483
Статистика	0,21435018925027557	0,21435018925027557	0,16666666666666666
N	6	6	6
Заключение	Возможно нормален	Возможно нормален	Возможно нормален



Выделяем диапазон B2:D7 и используем команды „Статистика/Тесты с множеством выборок/Дисперсионный анализ/Однофакторный...“. Результат представлен на следующих двух рисунках.

	A	B	C
1	Дисперсионный анализ: однофакторный		
2			
3	ИТОГО		
4	Группы	Количество	Сумма
5	Столбец 1	6	43
6	Столбец 2	6	37
7	Столбец 3	6	24
8			
9			
10	Дисперсионный анализ		
11	Источник дисперсии	Сумма квадратов	степень свободы
12	Между группами	31,444444444444446	2
13	В группах	31,666666666666668	15
14	Всего	63,111111111111114	17

D	E	F	G
Среднее	*Дисперсия*		
7,16666666666667	2,16666666666667		
6,16666666666667	2,16666666666667		
4	2		
Квадрат среднего	*F*	*Значение P*	*F критическое*
15,7222222222222	7,447368421052632	0,00567183949237680З	3,68232034367324l
2,11111111111111			

Из таблицы видно, что $F_{\text{эмп}} = 7,45$ (ячейка E12), $F_{\text{кр}} = 3,68$ (ячейка G12; в строке формул ячейки G12 указано, что $F_{\text{кр}}$ соответствует $p \leq 0,05$). Критические значения можно найти в таблице „Критические значения F Фишера для уровней статистической значимости $p \leq 0,05$ и $p \leq 0,01$: df_1 - число степеней свободы в числителе, df_2 - число степеней свободы в знаменателе“:

степень свободы 2 (ячейка C12) - $F_{\text{кр}} = 3,68$ (ячейка G12), $p \leq 0,05$;

степень свободы 15 (ячейка C13) - $F_{\text{кр}} = 6,36$, $p \leq 0,01$;

Вывод: H_0 отклоняется, т.к. $F_{\text{эмп}} > F_{\text{кр}}$. Принимается H_1. Различия в объеме воспроизведения слов между группами являются более выраженными, чем случайные различия внутри каждой группы. Итак, скорость предъявления слов влияет на объем их воспроизведения.

10. Задачи

1. Для приведенных данных определить числовые характеристики и дать заключение об отклонении данного распределения от нормального 18, 15, 16, 11, 14,15, 16,16, 16, 22, 17, 12, 11, 12, 18, 19, 20.

2. Две группы испытуемых дали сведения о своем весе. Можно ли на основании полученных данных утверждать, что вес в одной группе распределен более однородно, чем в другой?

70 69 68 64 78 65 69 71 68 74 70 67 74 74 81 68 72 69 61 75 56

73 69 73 65 68 74 71 80 64 60 67 70 74 53 70 56 67 65 72 72 69

3. Две группы испытуемых оценивались по 100-балльной шкале. Определить, значимо ли отличаются средние показатели первой группы от аналогичных показателей второй.

1 гр. 60 78 65 95 69 77 49 62 49 76 70 77 72 57 70 64 61 57

2 гр. 96 64 74 69 40 61 97 63 70 83 76 86 73 52 55 87 49 85

4. Перед сдачей экзаменов в конце семестра у 20 студентов университета был проведён опрос о том, какую оценку по сдаваемым в сессию курсам они ожидают получить. После сессии средние полученные оценки были сопоставлены со средними ожидаемыми. Результаты приведены в таблице:

№	1	2	3	4	5	6	7	8	9	10
Ожидаемая	3,4	3,1	3	2,8	3,7	3,5	2,9	3,7	3,5	3,2
Полученная	4,1	3,4	3,3	3	4,7	4,6	3	4,6	4,6	3,6

№	11	12	13	14	15	16	17	18	19	20
Ожидаемая	3	3,5	3,3	3,1	3,3	3,9	2,9	3,2	3,4	3,4
Полученная	3,5	4	3,6	3,1	3,3	4,5	2,8	3,7	3,8	3,9

Рассчитайте линейный коэффициент корреляции Пирсона, оцените его значимость при $p = 0{,}05$.

5. По приведенным значениям IQ (по Векслеру) у родителей и детей определить коэффициент корреляции Пирсона между уровнем интеллекта родителей и детей. На уровне $p = 0{,}05$. проверить значимость полученного коэффициента корреляции.

Родители: 117 108 121 106 117 105 118 128 116 122 98 128 99 126 103

Дети: 109 119 110 123 109 122 102 90 111 92 111 111 116 98 121

6. Знания десяти студентов проверены по двум тестам: A и B. Оценки по стобальной системе оказались следующими:

A	95	90	86	84	75	70	62	60	57	50
B	92	93	83	80	55	60	45	72	62	70

Найти выборочный коэффициент ранговой корреляции Спирмена между оценками по двум тестам.

Указание. Присвоим ранги x_i оценкам по тесту A. Эти оценки расположены в убывающем порядке, поэтому их ранги x_i равны порядковым номерам:

ранги x_i	1	2	3	4	5	6	7	8	9	10
оценки по тесту А	95	90	86	84	75	70	62	60	57	50

Присвоим ранги y_i оценкам по тесту B, для чего сначала расположим эти оценки в убывающем порядке:

оценки по тесту В	93	92	83	80	72	70	62	60	55	45

Индекс i при y должен быть равен порядковому номеру оценки студента по тесту A.

Найдем ранг y_i. Индекс $i = 1$ указывает, что рассматривается оценка студента, который занимает по тесту A первое место (эта оценка равна 95); из условия видно, что по тесту B студент получил оценку 92, которая расположена на втором месте. Таким образом, ранг $y_1 = 2$.

Найдем ранг y_2. Индекс $i = 2$ указывает, что рассматривается оценка студента, который занимает по тесту А второе место; из условия видно, что студент получил по тесту B оценку 93, которая расположена на первом месте. Таким образом, ранг $y_2 = 1$.

Аналогично найдем остальные ранги: $y_3 = 3$, $y_4 = 4$, $y_5 = 9$, $y_6 = 8$, $y_7 = 10$, $y_8 = 5$, $y_9 = 7$, $y_{10} = 6$.

Выпишем последовательности рангов x_i и y_i:

x_i	1	2	3	4	5	6	7	8	9	10
y_i	2	1	3	4	9	8	10	5	7	6

Далее найти коэффициент корреляции рангов К. Спирмена для случая разных рангов.

7. Два преподавателя оценили знания 12 учащихся по стобалльной системе и выставили им следующие оценки (в первой строке указано количество баллов, выставленных первым преподавателем, а во второй - вторым):

98	94	88	80	76	70	63	61	60	58	56	51
99	91	93	74	78	65	64	66	52	53	48	62

Найти коэффициент ранговой корреляции Спирмена между оценками двух преподавателей.

8. Группа людей была опрошена по поводу наличия у них дачи (0 - нет дачи, 1 - есть дача) и автомашины (0 - нет машины, 1 - есть машина). Определить значение коэффициента корреляции между наличием/ отсутствием дачи и наличием/ отсутствием автомашины.

Дача 0 1 1 1 1 0 0 1 1 0 0 1 1 0 0 0 1 0 1 1 1 0 0 0 1 0 1

Машина 0 0 1 1 1 0 0 0 0 1 0 0 0 1 0 0 1 1 0 1 1 0 0 0 1 1 0

9. У группы людей было определено значение IQ по Векслеру. Определить значения бисериального коэффициента между полом и уровнем интеллекта (0 - ж., 1 - м.)

Пол	1	0	1	1	1	0	0	1	0	0	0	0	0	1	1	1
IQ	103	130	116	112	106	112	84	124	105	99	138	110	121	93	113	98

10. Определялось значение IQ по Векслеру у родителей и их детей. Результаты приведены в таблице (0 - IQ ниже среднего, 1 - IQ выше среднего). Определить коэффициент корреляции между уровнем интеллекта у родителей и их детей.

		Родители	
		0	1
Дети	0	36	8
	1	16	72

11. Опрос 10 студентов университета позволяет выявить зависимость между средним баллом по результатам предыдущей сессии и числом часов в неделю затраченных студентом на самостоятельную подготовку.

Число часов	25	22	9	15	15	30	20	30	10	17
Средний балл	4,6	4,3	3,8	3,8	4,2	4,3	3,8	4	3,1	3,9

Постройте график исходных данных и определите по нему характер зависимости. Рассчитайте выборочный коэффициент линейной корреляции Пирсона, проверьте его значимость при p = 0,05. Постройте уравнение регрессии и дайте интерпретацию полученных результатов. Если студент занимается самостоятельно по 12 часов в неделю, то каков прогноз успеваемости?

12. Из студентов 3-го курса отобраны случайным образом 10 человек и подсчитаны средние оценки, полученные ими на 1-ом (X) и 3-м (Y) курсе.

X	3,5	4	3,8	4,6	3,9	3	3,5	3,9	4,5	4,1
Y	4,2	3,9	3,8	4,5	4,2	3,4	3,8	3,9	4,6	3

Полагая, что между X и Y имеет место линейная зависимость, определите выборочное уравнение линейной регрессии и объясните смысл полученных коэффициентов.

Список литературы

1. Гласс Дж., Стенли Дж. Статистические методы в педагогике и психологии. - М.: Прогресс. 1976 г. - 496 с.

2. Гмурман В.Е. Теория вероятностей и математическая статистика [Текст] : учебное пособие для бакалавров / В.Е. Гмурман. - 12-е изд. - Москва : Юрайт, 2013. - 478 с.

3. Ермолаев О.Ю. Математическая статистика для психологов: Учебник. - 3-е изд., испр. - М.: Московский психолого-социальный институт: Флинта, 2004. - 336 с.

4. Загвязинский В.И., Атаханов Р.А. Методология и методы психолого-педагогического исследования : учеб. пос. для студ. высш. пед. учеб. зав. - 2-е изд., стереотип. - М. : Академия, 2007. - 208 с.

5. Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. Учебное пособие. - СПб.: Речь, 2004. - 392 с.

6. Сидоренко Е.В. Метод математическо обработк в психологии. - СПб.: Речь, 2000. - 350 с.

7. Хахаев И.А. Gnumeric: Электронная таблица для всех. - М. : ALT Linux, 2011. 192 с.

Приложение

Таблица 1.

Критические значения критерия T - Вилкоксона

n	p		n	p	
	0,05	0,01		0,05	0,01
5	0	–	28	130	101
6	2	–	29	140	110
7	3	0	30	151	120
8	5	1	31	163	130
9	8	3	32	175	140
10	10	5	33	187	151
11	13	7	34	200	162
12	17	9	35	213	173
13	21	12	36	227	185
14	25	15	37	241	198
15	30	19	38	256	211
16	35	23	39	271	224
17	41	27	40	286	238
18	47	32	41	302	252
19	53	37	42	319	266
20	60	43	43	336	281
21	67	49	44	353	296
22	75	55	45	371	312
23	83	62	46	389	328
24	92	69	47	407	345
25	100	76	48	426	362
26	110	84	49	446	379
27	119	92	50	466	397

Таблица 2. Критические значения критерия G - знаков

n	p		*n*	p		*n*	p	
	0,05	0,01		0,05	0,01		0,05	0,01
5	0	-	37	13	10	69	26	25
6	0	-	38	13	11	70	27	25
7	0	0	39	13	11	71	27	26
8	1	0	40	14	12	72	28	26
9	1	0	41	14	12	73	28	27
10	1	0	42	15	13	74	29	27
11	2	1	43	15	13	75	29	27
12	2	1	44	16	13	76	29	28
13	3	1	45	16	14	77	30	28
14	3	2	46	16	14	78	30	29
15	3	2	47	17	15	79	31	29
16	4	2	48	17	15	80	31	30
17	4	3	49	18	15	81	32	30
18	5	3	50	18	16	82	32	31
19	5	4	51	19	17	83	33	31
20	5	4	52	19	17	84	33	31
21	6	4	53	20	18	85	33	32
22	6	5	54	20	19	86	34	32
23	7	5	55	20	19	87	34	33
24	7	5	56	21	19	88	35	33
25	7	6	57	21	20	89	35	34
26	8	6	58	22	20	90	36	34
27	8	7	59	22	21	91	36	34
28	8	7	60	22	21	92	37	35
29	9	7	61	23	21	93	37	35
30	10	8	62	23	22	94	38	36
31	10	8	63	24	22	95	38	36
32	10	8	64	24	23	96	38	37
33	11	9	65	25	23	97	39	37
34	11	9	66	25	24	98	39	38
35	12	10	67	26	24	99	40	38
36	12	10	68	26	24	100	40	38

Таблица 3. Критические значения критерия U Вилкоксона-Манна-Уитни

n_2 \ n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	$p = 0{,}05$																												
3	-	0																											
4	-	0	1																										
5	0	1	2	4																									
6	0	2	3	5	7																								
7	0	2	4	6	8	11																							
8	1	3	5	8	10	13	15																						
9	1	4	6	9	12	15	18	21																					
10	1	4	7	11	14	17	20	24	27																				
11	1	5	8	12	16	19	23	27	31	34																			
12	2	5	9	13	17	21	26	30	34	38	42																		
13	2	6	10	15	19	24	28	33	37	42	47	51																	
14	3	7	11	16	21	26	31	36	41	46	51	56	61																
15	3	7	12	18	23	28	33	39	44	50	55	61	66	72															
16	3	8	14	19	25	30	36	42	48	54	60	65	71	77	83														
17	3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96													
18	4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109												
19	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123											
20	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138										
21	-	-	19	26	34	41	49	57	65	73	81	89	97	105	113	121	130	138	146	154									
22	-	-	20	28	36	44	52	60	69	77	85	94	102	111	119	128	136	145	154	162	171								
23	-	-	21	29	37	46	55	63	72	81	90	99	107	116	125	134	143	152	161	170	180	189							
24	-	-	22	31	39	48	57	66	75	85	94	103	113	122	131	141	150	160	169	179	188	198	207						
25	-	-	23	32	41	50	60	69	79	89	98	108	118	128	137	147	157	167	177	187	197	207	217	227					
26	-	-	24	33	43	53	62	72	82	93	103	113	123	133	143	154	164	174	185	195	206	216	266	237	247				
27	-	-	25	35	45	55	65	75	86	96	107	118	128	139	150	160	171	182	193	203	213	225	236	247	258	268			
28	-	-	26	36	47	57	68	79	89	100	111	122	133	144	156	167	178	189	200	212	223	234	245	257	268	279	291		
29	-	-	27	38	48	59	70	82	93	104	116	127	139	150	162	173	185	196	208	220	232	243	255	267	278	290	302	314	
30	-	-	28	39	50	62	73	85	96	108	120	132	144	156	168	180	192	204	216	228	240	252	265	277	289	301	313	326	338

n_2 \ n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	$p = 0{,}01$																												
5	-	-	0	1																									
6	-	-	1	2	3																								
7	-	0	1	3	4	6																							
8	-	0	2	4	6	7	9																						
9	-	1	3	5	7	9	11	14																					
10	-	1	3	6	8	11	13	16	19																				
11	-	1	4	7	9	12	15	18	22	25																			
12	-	2	5	8	11	14	17	21	24	28	31																		
13	0	2	5	9	12	16	20	23	27	31	35	39																	
14	0	2	6	10	13	17	22	26	30	34	38	43	47																
15	0	3	7	11	15	19	24	28	33	37	42	47	51	56															
16	0	3	7	12	16	21	26	31	36	41	46	51	56	61	66														
17	0	4	8	13	18	23	28	33	38	44	49	55	60	66	71	77													
18	0	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88												
19	1	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101											
20	1	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114										
21	-	-	10	16	22	29	35	42	49	56	63	70	77	84	91	98	105	113	120	127									
22	-	-	10	17	23	30	37	45	52	59	66	74	81	89	96	104	111	119	127	134	142								
23	-	-	11	18	25	32	39	47	55	62	70	78	86	94	102	109	117	125	133	141	150	158							
24	-	-	12	19	26	34	42	49	57	66	74	82	90	98	107	115	123	132	140	149	154	166	174						
25	-	-	12	20	27	35	44	52	60	69	77	86	95	103	112	121	130	138	147	156	165	174	183	192					
26	-	-	13	21	29	37	46	54	63	72	81	90	99	108	117	126	136	145	154	163	173	182	191	201	210				
27	-	-	14	22	30	39	48	57	66	75	85	94	103	113	122	132	142	151	161	171	180	190	200	209	219	229			
28	-	-	14	23	32	41	50	59	69	78	88	98	108	118	128	138	148	158	168	178	188	198	208	218	229	239	249		
29	-	-	15	24	33	42	52	62	72	82	92	102	112	123	133	143	154	164	175	185	196	206	217	227	238	249	259	270	
30	-	-	15	25	34	44	54	64	75	85	95	106	117	127	138	169	160	171	182	192	203	214	225	236	247	258	270	281	292

Таблица 4.

Критические значения критерия χ^2 для уровня статистической значимости $p \leq 0,05$ и $p \leq 0,01$ при различном числе степеней свободы ν

ν	p		ν	p		ν	p	
	0,05	0,01		0,05	0,01		0,05	0,01
1	3,841	6,635	35	49,802	57,342	69	89,391	99,227
2	5,991	9,210	36	50,998	58,619	70	90,631	100,425
3	7,815	11,345	37	52,192	59,892	71	91,670	101,621
4	9,488	13,277	38	53,384	61,162	72	92,808	102,816
5	11,070	15,086	39	54,572	62,428	73	93,945	104,010
6	12,592	16,812	40	55,758	63,691	74	95,081	105,202
7	14,067	18,475	41	56,942	64,950	75	96,217	106,393
8	15,507	20,090	42	58,124	66,206	76	97,351	107,582
9	16,919	21,666	43	59,304	67,459	77	98,484	108,771
10	18,307	23,209	44	60,481	68,709	78	99,617	109,958
11	19,675	24,725	45	61,656	69,957	79	100,749	111,144
12	21,026	26,217	46	62,830	71,201	80	101,879	112,329
13	22,362	27,688	47	64,001	72,443	81	103,010	113,512
14	23,685	29,141	48	65,171	73,683	82	104,139	114,695
15	24,996	30,578	49	66,339	74,919	83	105,267	115,876
16	26,296	32,000	50	67,505	76,154	84	106,395	117,057
17	27,587	33,409	51	68,669	77,386	85	107,522	118,236
18	28,869	34,805	52	69,832	78,616	86	108,648	119,414
19	30,144	36,191	53	70,993	79,843	87	109,773	120,591
20	31,410	37,566	54	72,153	81,069	88	110,898	121,767
21	32,671	38,932	55	73,311	82,292	89	112,022	122,942
22	33,924	40,289	56	74,468	83,513	90	113,145	124,116
23	35,172	41,638	57	75,624	84,733	91	114,268	125,289
24	36,415	42,980	58	76,778	85,950	92	115,390	126,462
25	37,652	44,314	59	77,931	87,166	93	116,511	127,633
26	38,885	45,642	60	79,082	88,379	94	117,632	128,803
27	40,113	46,963	61	80,232	89,591	95	118,752	129,973
28	41,337	48,278	62	81,381	90,802	96	119,871	131.141
29	42,557	49,588	63	82,529	92,010	97	120,990	132,309
30	43,773	50,892	64	83,675	93,217	98	122,108	133,476
31	44,985	52,191	65	84,821	94,422	99	123,225	134,642
32	46,194	53,486	66	85,965	95,626	100	124,342	135,807
33	47,400	54,776	67	87,108	96,828			
34	48,602	56,061	68	88,250	98,028			

Таблица 5.

Критические значения t-критерия Стьюдента

Число степеней свободы k	p			Число степеней свободы k	p		
	0,05	0,01	0,001		0,05	0,01	0,001
1	12,71	63,66	64,6	18	2,10	2,88	3,92
2	4,30	9,92	31,6	19	2,09	2,86	3,88
3	3,18	5,84	12,92	20	2,09	2,85	3,85
4	2,78	4,6	8,61	21	2,08	2,83	3,82
5	2,57	4,03	6,87	22	2,07	2,82	3,79
6	2,45	3,71	5,96	23	2,07	2,81	3,77
7	2,37	3,50	5,41	24	2,06	2,80	3,75
8	2,31	3,36	5,04	25	2,06	2,79	3,73
9	2,26	3,25	4,78	26	2,06	2,78	3,71
10	2,23	3,17	4,59	27	2,05	2,77	3,69
11	2,20	3,11	4,44	28	2,05	2,76	3,67
12	2,18	3,05	4,32	29	2,05	2,76	3,66
13	2,16	3,01	4,22	30	2,04	2,75	3,65
14	2,14	2,98	4,14	40	2,02	2,70	3,55
15	2,13	2,95	4,07	60	2,00	2,66	3,46
16	2,12	2,92	4,02	120	1,98	2,62	3,37
17	2,11	2,90	3,97	∞	1,96	2,58	3,29

Таблица 6.

Критические значения коэффициента линейной корреляции r_{xy} Пирсона

$k = n-2$	p		$k = n-2$	p	
	0,05	**0,01**		**0,05**	**0,01**
5	0,75	0,87	37	0,32	0,41
6	0,71	0,83	38	0,31	0,40
7	0,67	0,80	39	0,31	0,40
8	0,63	0,77	40	0,30	0,39
9	0,60	0,74	41	0,30	0,39
10	0,58	0,71	42	0,30	0,38
11	0,55	0,68	43	0,29	0,38
12	0,53	0,66	44	0,29	0,37
13	0,51	0,64	45	0,29	0,37
14	0,50	0,62	46	0,29	0,37
15	0,48	0,61	47	0,28	0,27
16	0,47	0,59	48	0,28	0,36
17	0,46	0,58	49	0,28	0,35
18	0,44	0,56	50	0,27	0,35
19	0,43	0,55	55	0,26	0,34
20	0,42	0,54	60	0,25	0,33
21	0,41	0,53	65	0,24	0,31
22	0,40	0,52	70	0,23	0,30
23	0,40	0,51	80	0,22	0,28
24	0,39	0,50	90	0,21	0,27
25	0,38	0,49	100	0,20	0,25
26	0,37	0,48	125	0,17	0,23
27	0,37	0,47	150	0,16	0,21
28	0,36	0,46	200	0,14	0,18
29	0,36	0,46	300	0,11	0,15
30	0,35	0,45	400	0,10	0,13
31	0,34	0,44	500	0,09	0,12
32	0,34	0,44	700	0,07	0,10
33	0,33	0,43	900	0,06	0,09
34	0,33	0,42	1000	0,06	0,09
35	0,33	0,42	–	–	–
36	0,32	0,41	–	–	–

Таблица 7.

Критические значения коэффициента корреляции рангов Спирмена

n	p		*n*	p	
	0,05	0,01		0,05	0,01
5	0,94	–	23	0,42	0,53
6	0,85	–	24	0,41	0,52
7	0,78	0,94	25	0,49	0,51
8	0,72	0,88	26	0,39	0,50
9	0,68	0,83	27	0,38	0,49
10	0,64	0,79	28	0,38	0,48
11	0,61	0,76	29	0,37	0,48
12	0,58	0,73	30	0,36	0,47
13	0,56	0,70	31	0,36	0,46
14	0,54	0,68	32	0,36	0,45
15	0,52	0,66	33	0,34	0,45
16	0,50	0,64	34	0,34	0,44
17	0,48	0,62	35	0,33	0,43
18	0,47	0,60	36	0,33	0,43
19	0,46	0,58	37	0,33	0,43
20	0,45	0,57	38	0,32	0,41
21	0,44	0,56	39	0,32	0,41
22	0,43	0,54	40	0,31	0,40