Министерство науки и образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ЛАБОРАТОРНЫМ РАБОТАМ ПО КУРСУ «МАТЕМАТИЧЕСКАЯ СТАТИСТИКА ДЛЯ АНАЛИЗА ДАННЫХ»

Авторы:
Доц. Минитаева А.М.
Студент гр. ИУ6-63Б Наконечный Р. Ю.

Москва, 2024

Оглавление

Предисловие

В ходе данных методических указаний в рамках курса “Математическая статистика для анализа данных” мы познакомимся со способами обработки статистической информации компьютерными методами, а конкретно, научимся работать со статистическими данными в программе Microsoft Excel в рамках нескольких следующих лабораторных работ.

Но почему именно Excel?

Эта программа достаточно проста в освоении и большинство из вас уже частично с ней знакомились еще в школе или самостоятельно. Работать со статистическими данными в Excel очень удобно, можно быстро строить большие таблицы и обрабатывать их. Также представлено много функций, которые сильно упрощают и ускоряют работу, беря на себя сложные математические формулы.

Если вам удобнее работать в других аналогичных программах для работы с таблицами, можете использовать их. Но все примеры, приведенные здесь, относятся именно к Excel, и не всегда реализация функций в других программах будет совпадать с той, которая представлена в Excel.

Каждой лабораторной работе соответствует задание по варианту, данные к каждому из вариантов можно найти в отдельных файлах. Рекомендуется сначала ознакомиться с пояснениями, указанными здесь, а затем приступать к выполнению задания соответствующего лабораторной работе и вашему варианту.

По каждой лабораторной работе необходимо написать отчет с подробным объяснением и демонстрацией выполнения задания, написать вывод и ответить на контрольные вопросы.

Лабораторная работа №1. Количественные и качественные данные.

Цель:

Знакомство с типами данных и простыми методами их обработки.

Задачи:

- Узнать как данные разделяются на различные типы;
- Узнать о методах обработки различных данных;
- Научиться основам мат. статистики в программе Excel.

Теоретическая часть

Статистика – наука, в которой излагаются общие вопросы сбора, измерения, мониторинга, анализа массовых статистических данных и их сравнение, или, если говорить проще, изучающая большие совокупности однородных объектов на основании выборочного исследования.

В статистике исследуемые объекты или явления называются генеральной совокупностью, а часть объектов такой совокупности, отобранных для её изучения, называют выборочной совокупностью (выборкой).

Типы данных.

В статистике все данные делят на 2 типа:

Качественные данные – данные, обладающие слабой степенью формализации, т. е. которые нельзя измерить числовыми методами.

Количественные данные – числовые данные, которые можно использовать в процессе математических и статистических исследований.

Например, если ученикам школы выдали анкету, в которой они должны составить свое мнение об обучении, то ответы на вопросы можно отнести к разным типам данных следующим образом:

	A	B	C	D
1	**Качественные данные**		**Количественные данные**	
2	Пол	М	Возраст	20
3	Оценка	Нравится	Средняя оценка	4,5

Рис. 1.

Но почему словесная оценка, которую выставляет ученик программе обучения, считается примером качественных данных?

Даже если будет всего несколько вариантов ответа, таких как "Нравится", "Не знаю", "Не нравится", мы не можем знать разницу между этими вариантами. Неизвестно, насколько сильно "Нравится" отличается от "Не знаю", "Не знаю" от "Не нравится" и тому подобное. В отличие от таких оценок, возраст или среднюю оценку ученика мы можем измерить, и знаем насколько численно отличается одно значение от другого.

Можете потренироваться и определить типы данных, приведенных на рисунке:

	A	B	C	D	E	F
1	Респондент	Макс. кол-во подтягиваний	Оценка вкуса батончика	Вес	Рост	Родной город
2	1	10	вкусно	65	170	Москва
3	2	5	невкусно	55	175	Анапа
4	3	6	не знаю	49	168	Самара
5	4	8	ниже среднего	61	159	Брянск
6	5	2	выше среднего	70	177	Орёл
7	...	...	...	...	...	...
8						

Рис. 2.

Также стоит сказать, что на практике, некоторые качественные данные могут рассматриваться как количественные, например если дать в соответствие каждому варианту ответа на вопрос об оценке обучения какое то число, то мы получим количественные данные:

A	B
Нравится	1
Не знаю	0
Не нравится	-1

Рис. 3.

Таким образом, одни и те же данные могут рассматриваться как качественные и как количественные, все зависит от того, рассматриваются они в теории или на практике.

Знакомство с количественными данными.

Количественные данные, в отличие от качественных – данные, поддающиеся математической и статистической обработке.

Рассмотрим, как их можно обрабатывать.

Например, у нас есть список лучших производителей докторской колбасы с названием и ценой за продукт в розницу. Составим таблицу (вместо названий производителей будем использовать порядковые номера):

	A	B
1	Производитель	Цена
2	1	200
3	2	300
4	3	350
5	4	260
6	5	400
7	6	370
8	7	410
9	8	500
10	9	230
11	10	540
12	11	610
13	12	220
14	13	690
15	14	580
16	15	340
17	16	390
18	17	430
19	18	640
20	19	660
21	20	290
22		

Рис. 4.

Мы можем разбить все продукты на категории в зависимости от их стоимости: 200-299, 300-399, 400-499, 500-599, 600-699.

Такое разделение в статистике называется распределением, а каждая категория - интервалом.

Также мы можем и обозначить среднюю цену в каждой категории: 249,5 , 349,5 , 449,5 , 549,5 , 649,5 . Каждое такое число будет называться серединой интервала. Это число получается путем сложения верхней и нижней границы диапазона и делением результата на два.

Число производителей в каждой категории называется абсолютной частотой. Этот термин обозначает, как часто определенное значение или категория встречается в наборе данных. Используя число всех элементов совокупности и число элементов определенного типа (абсолютную частоту), можно посчитать относительную частоту (выражается в процентах). В нашем примере число ресторанов с ценами от 200 до 299 равно 5, тогда относительная частота:

$$5/20 * 100\% = 25\%$$

Для того чтобы посчитать относительную частоту в Excel, можно ввести формулу в нужную ячейку. Для этого выберите ячейку, в которую необходимо

ввести формулу, и в строке формул напишите "=", а затем имена ячеек и действия, которые вы хотите с ними произвести. Это будет выглядеть примерно так:

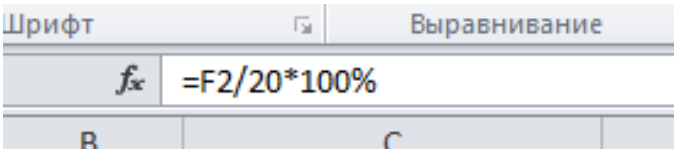


Рис. 5.

После этого вы можете выбрать ячейку, в которой введена формула, и потянув за ее правый нижний угол, распространить формулу на другие ячейки:

Рис. 6.

При этом ячейки, которые используются для вычислений в формуле, будут заменяться на следующие за ними в зависимости от направления, в котором вы распространяете формулу. Кроме этого, вы можете скопировать формулу и вставить ее в другую ячейку. Тогда также номера ячеек в формуле поменяются на соответствующие другому расположению.

Если вы хотите, чтобы адреса ячеек не менялись при копировании формул, то необходимо их закрепить с помощью знака $. Для полного закрепления адреса необходимо поставить этот символ перед номером столбца и перед номером строки. Если поставить $ только перед одним из этих адресов, то не будет меняться адрес столбца или номер строки соответственно. В программе применения этого символа выглядит вот так:

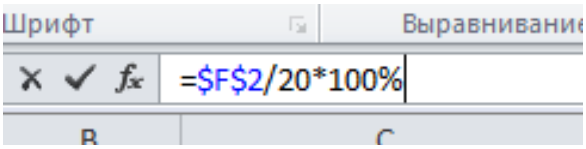


Рис. 7.

Посчитав относительную частоту для всех интервалов, получим:

D	E	F	G
Интервал	Середина интервала	Кол-во производителей (абс. частота)	Относительная частота
200-299	249,5	5	25,00%
300-399	349,5	5	25,00%
400-499	449,5	3	15,00%
500-599	549,5	3	15,00%
600-699	649,5	4	20,00%

Рис. 8.

Прим. Чтобы вычислить абсолютную частоту для большого массива данных, можно воспользоваться функцией ЧАСТОТА() для массива. Сначала создайте столбец с верхними границами интервалов, затем введите функцию, выберите в качестве аргументов столбец с ценами и столбец с верхними границами интервалов и нажмите Enter:

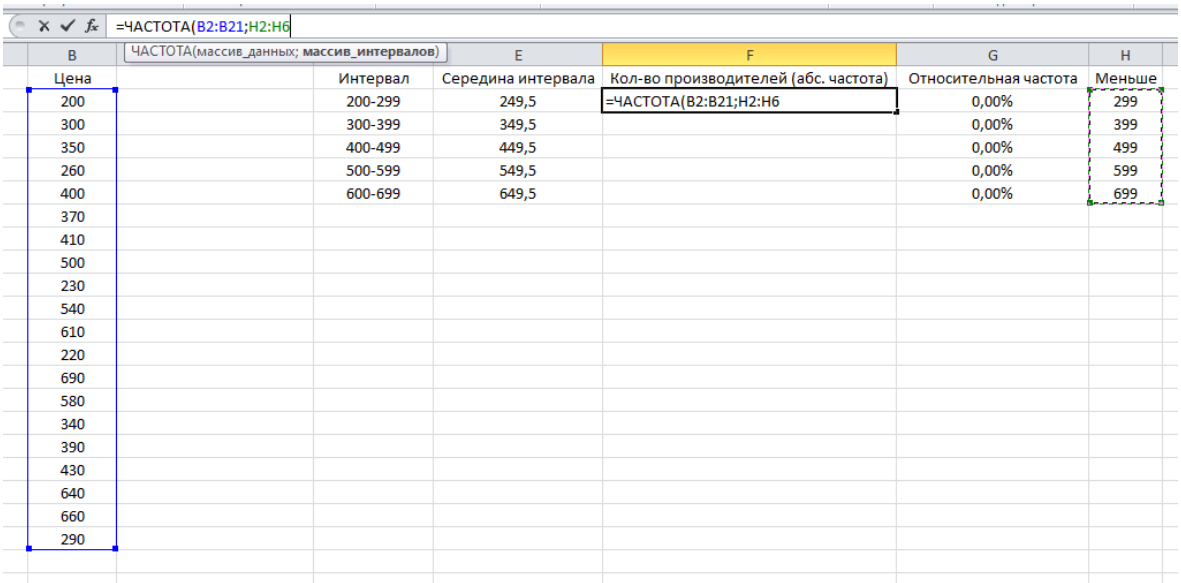


Рис. 9.

Далее, чтобы применить функцию для массива, выделите столбец абсолютной частоты, нажмите на строку формул, зажмите Ctrl+Shift и нажмите Enter:

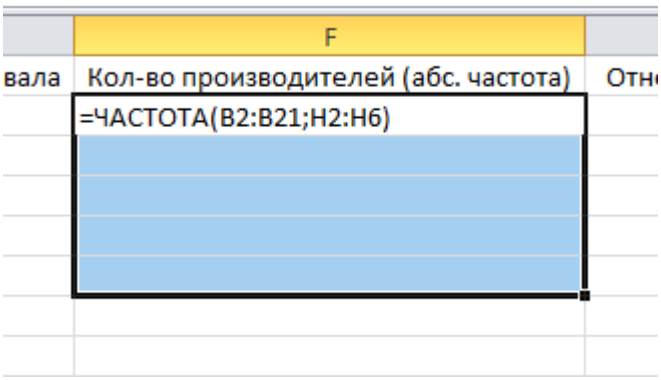


Рис. 10.

Формула применится для массива и будет выглядеть так:

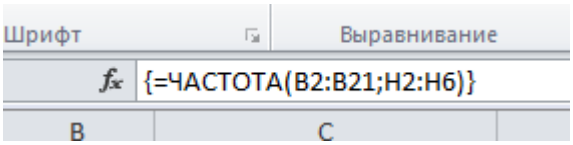


Рис. 11.

В нашем случае итоговая таблица выглядит вполне понятно, но бывает и так, что данные сложно воспринимаются в таком виде. Поэтому, чтобы лучше познакомиться с ними, используют графический метод. В нашем примере для этого подойдет столбчатая диаграмма – гистограмма. В зависимости от ситуации можно использовать и другие виды диаграмм.

Для построения гистограммы в программе Excel необходимо выделить нужные столбы таблиц, выбрать панель "Вставка"->" Гистограмма":

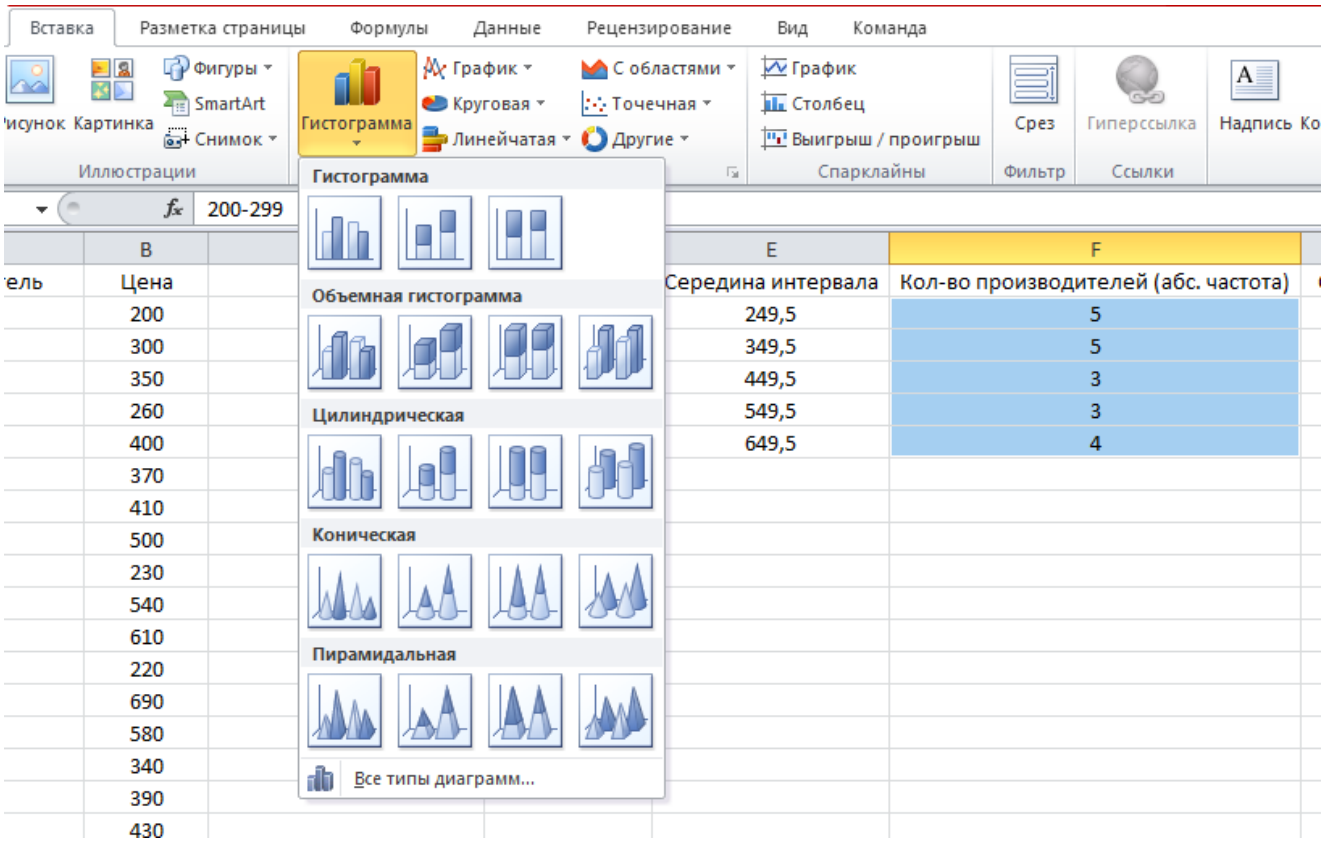


Рис. 12.

И тогда мы получим:

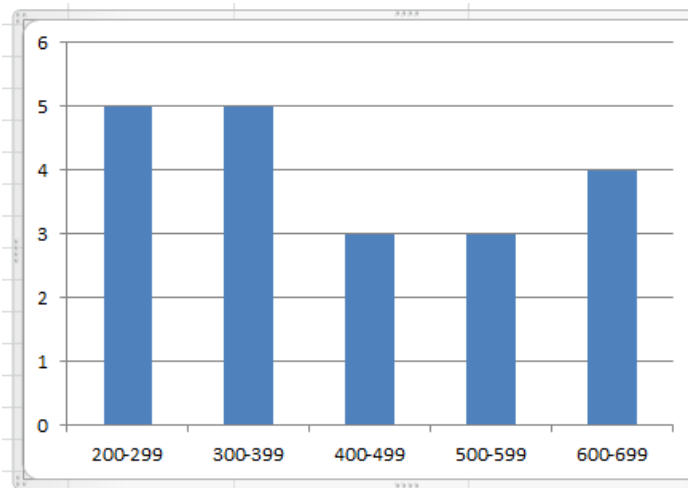


Рис. 13.

Остальные виды диаграмм строятся аналогично, через панель "Вставка". Стоит также сказать, что величину интервалов, на которые мы разбили все цены на продукт, в математике определяют специальными формулами.
Для определения количества интервалов используют формулу Стерджесса:

$$\text{Кол-во интервалов} = 1 + \frac{\log_{10} N}{\log_{10} 2}$$

Рис. 14.

Величину интервала определяют следующей формулой:

$$\frac{MAX - MIN}{\text{Кол-во интервалов}},$$

где MAX — максимальное значение в совокупности,
MIN — минимальное значение в совокупности.

Рис. 15.

В Excel нет встроенной функции для формулы Стерджесса, но ее можно легко вычислить с помощью обычной формулы:

```
=1 + 3.322 * LOG10(COUNT(A1:A10))
```

Величину интервала можно вычислить, используя результат предыдущей формулы (ячейка B1) и максимальное и минимальное значения в совокупности:

```
=(МАКС(A1:A10)-МИН(A1:A10))/B1
```

Однако часто бывает, что распределение менее понятно, если количество и величина интервалов определена таким способом. Поэтому этот метод использовать необязательно, и решение о выборе интервалов стоит за тем, кто проводит статистический анализ.

Знакомство с качественными данными.

Как и говорилось ранее, качественные данные нельзя измерить числовыми методами, однако их тоже можно обрабатывать. Рассмотрим на примере: Ученики школы отвечали на вопрос о качестве еды в столовой одним из вариантов: "Нравится" "Не очень" "Не нравится":

	F	G
1	Ученик	Ответ
2	1	Нравится
3	2	Не нравится
4	3	Нравится
5	4	Нравится
6	5	Нравится
7	6	Не очень
8	7	Не очень
9	8	Не нравится
10	9	Не нравится
11	10	Нравится
12	11	Не очень
13	12	Не нравится
14	13	Не очень
15	14	Не очень
16	15	Не нравится
17	16	Нравится
18	17	Не очень
19	18	Нравится
20	19	Не очень
21	20	Не нравится
22		

Рис. 16.

Используя эти данные, мы можем составить следующую таблицу:

I	J	K
Ответ	Количество	%
Нравится	7	35
Не очень	7	35
Не нравится	6	30

Рис. 17.

Для большей наглядности можно построить диаграмму:

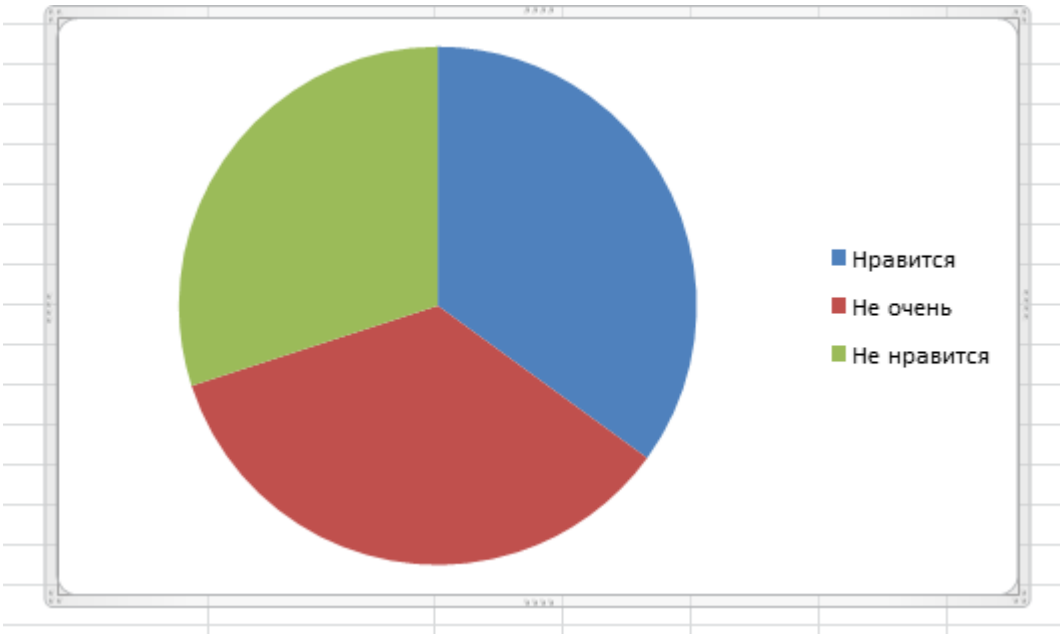


Рис. 18.

Таким образом, мы смогли привести и качественные данные к адекватному виду, с которым можно вести дальнейшую работу.

Практическая часть

Для данной в файле таблицы обработать качественные и количественные данные, для количественных данных найти интервалы с помощью формулы Стерджесса, абсолютную частоту, относительную частоту и построить гистограмму.

Контрольные вопросы

1. **Что такое количественные и качественные данные?**
2. **Что отличает качественные и количественные данные?**
3. **Что такое относительная и абсолютная частота?**
4. **Зачем нужно строить диаграммы?**
5. **Как нужно находить величину интервалов?**

Лабораторная работа №2. Среднее, медиана, стандартное и нормированное отклонение.

Цель:

Знакомство с основными мерами оценки данных и их вычислением в Excel.

Задачи:

- Узнать более сложные статистические метрики;
- Узнать о более сложных функциях в Excel;
- Научиться применять новые функции при работе с таблицами.

Теоретическая часть

В статистике часто приходится сравнивать разные значения, но что делать, если результаты одинаковые, по каким признакам можно их сравнить?
В этом могут помочь следующие величины:

Среднее арифметическое – число, равное сумме всех чисел множества, деленной на их количество. Например, у нас есть таблица баллов учеников двух классов за экзамен по химии:

Класс А	Ученик	Балл	Класс Б	Ученик	Балл
	1	77		1	88
	2	65		2	61
	3	90		3	65
	4	80		4	74
	5	82		5	85
	6	61		6	89
	7	73		7	78
	8	89		8	60
	9	59		9	72
	10	91		10	81

Рис. 19.

У учеников 2 из класса А и 3 из класса Б результаты одинаковые, однако ученика 3 из класса Б можно оценить выше, если сравнить результаты со средним результатом по классу (76,7 для класса А и 75,3 для класса Б). Для вычисления среднего в программе Excel используется функция СРЗНАЧ(). Введите имя этой функции в строке формул, а затем выберите диапазон ячеек, для которых необходимо посчитать среднее. Можно выделить ячейки в таблице, либо самим просто написать имя первой и последней ячеек через двоеточие:

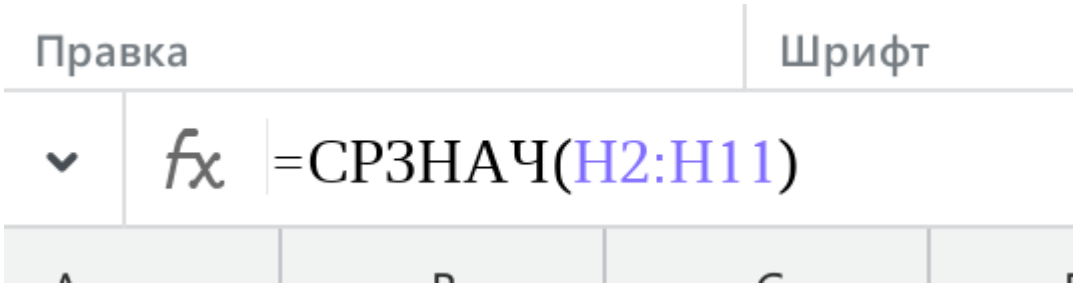


Рис. 20.

Медиана – число, которое находится в середине набора, если его упорядочить по возрастанию, то есть такое число, что половина из элементов

набора не меньше него, а другая половина не больше. Если число элементов чётное, то медианой будет среднее значение между двумя центральными элементами. Медиана используется, когда есть аномально большие или маленькие значения, которые будут сильно искажать сравнение с использованием среднего арифметического. Медиана может использоваться как одна из характеристик выборки или совокупности чисел.

Например:

Класс А	Ученик	Балл	Класс Б	Ученик	Балл
	1	20		1	88
	2	65		2	61
	3	90		3	65
	4	80		4	74
	5	82		5	85
	6	61		6	89
	7	73		7	78
	8	89		8	60
	9	59		9	72
	10	91		10	81

Рис. 21.

В классе А у ученика под номером 1 аномально низкий результат, который сильно уменьшит значение среднего по классу, поэтому для сравнения лучше использовать медиану. В программе Excel для вычисления медианы существует функция МЕДИАНА(). Аналогично прошлой функции введите имя формулы и выберите диапазон:

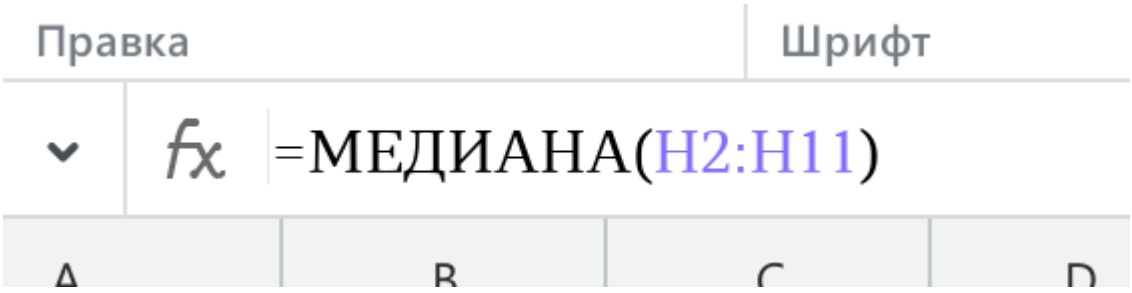


Рис. 21.

Среднеквадратическое отклонение (или стандартное отклонение) – показатель рассеивания значений случайной величины относительно ее математического ожидания. Например:

Класс А	Ученик	Балл	Класс Б	Ученик	Балл
	1	20		1	88
	2	65		2	61
	3	40		3	65
	4	80		4	74
	5	82		5	85
	6	61		6	89
	7	73		7	78
	8	89		8	60
	9	99		9	72
	10	100		10	81

Рис. 22.

В классе Б все результаты находятся в небольшом диапазоне значений, а в классе А результаты могут сильно отличаться. Для того чтобы описать разброс значений в каждом из классов и увидеть разницу между ними необходимо посчитать стандартное отклонение.

Оно считается по формуле:

простое

$$\sigma = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$$

взвешенное

$$\sigma = \sqrt{\frac{\sum (x_i - \overline{x})^2 \cdot f_i}{\sum f_i}}$$

Рис. 23.

В программе Excel стандартное отклонение можно посчитать с помощью функции СТАНДОТКЛОН().

У этой функции есть несколько вариантов подсчета:

СТАНДОТКЛОН.В() посчитает стандартное отклонение для аргументов, являющихся выборкой из генеральной совокупности, СТАНДОТКЛОН.Г() посчитает стандартное отклонение для всей генеральной совокупности, СТАНДОТКЛОНА() считает стандартное отклонение для выборки из генеральной совокупности, которая может включать текст и логические значения,

СТАНДОТКЛОНПА() считает стандартное отклонение всей генеральной совокупности, которая может включать текст и логические значения.

Мы предположим, что результаты, доступные нам в таблице являются выборкой и будем использовать СТАНДОТКЛОН.В():

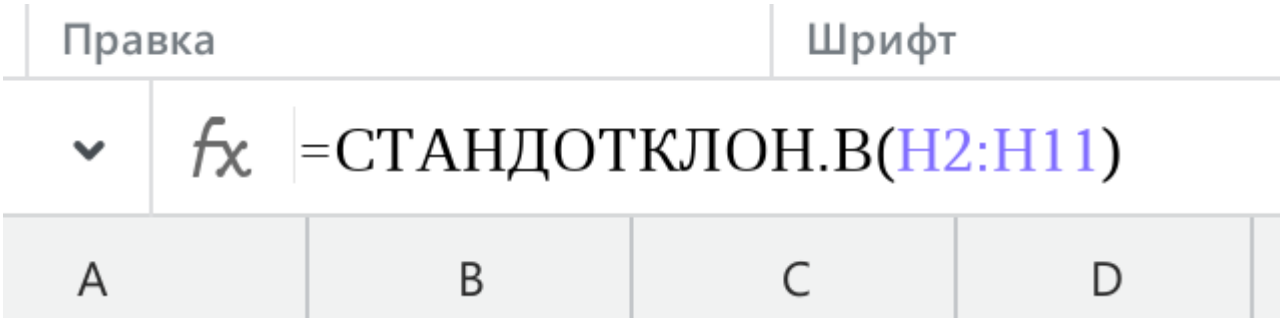


Рис. 24.

Нормированное отклонение.

Нормированное отклонение (или Z – показатель) – мера относительного разброса наблюдаемого или измеренного значения, которая показывает, сколько стандартных отклонений составляет его разброс относительно среднего значения. Иными словами, с помощью этой метрики можно сравнивать отдельные значения со средним. Для вычисления Z – оценки используется следующая формула:

$$\text{Нормированное отклонение (Z-показатель)} = \frac{\text{Значение} - \text{Среднее значение}}{\text{Стандартное отклонение}}$$

Рис. 25.

Вспомним пример с результатами учеников за экзамен по химии:

F	G	H	I	J	K	L
Класс А	Ученик	Балл		Класс Б	Ученик	Балл
	1	20			1	88
	2	65			2	61
	3	40			3	65
	4	80			4	74
	5	82			5	85
	6	61			6	89
	7	73			7	78
	8	89			8	60
	9	99			9	72
	10	100			10	81
	Среднее	70,9			Среднее	75,3
	Медиана	76,5			Медиана	76
	Отклонение	25,501416083			Отклонение	10,750193797

Рис. 26.

Мы уже научились считать среднее значение и стандартное отклонение, которые нужны для подсчета Z – показателя. Теперь необходимо использовать функцию НОРМАЛИЗАЦИЯ() для подсчета нормированного отклонения:

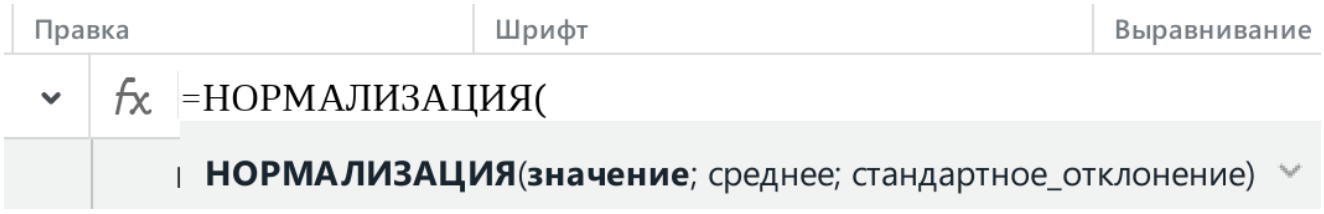


Рис. 27.

Выберите последовательно ячейки со значением, для которого считается нормированное отклонение, средним арифметическим и стандартным отклонением. Затем, чтобы при копировании формулы ячейки со средним и стандартным отклонением не менялись, закрепите их с помощью $. Итоговая формула будет выглядеть так:

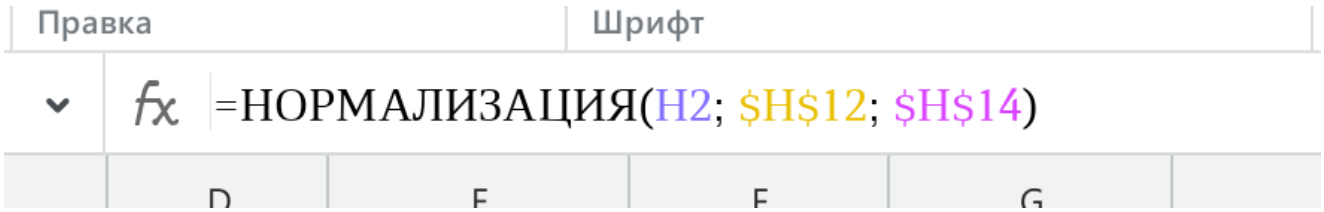


Рис. 28.

Вписав формулу во все необходимые ячейки, получим такой результат:

F	G	H	I	J	K	L	M
Класс А	Ученик	Балл	Норм. отклонение	Класс Б	Ученик	Балл	Норм. отклонение
	1	20	-2,00		1	88	1,18
	2	65	-0,23		2	61	-1,33
	3	40	-1,21		3	65	-0,96
	4	80	0,36		4	74	-0,12
	5	82	0,44		5	85	0,90
	6	61	-0,39		6	89	1,27
	7	73	0,08		7	78	0,25
	8	89	0,71		8	60	-1,42
	9	99	1,10		9	72	-0,31
	10	100	1,14		10	81	0,53
	Среднее	70,9			Среднее	75,3	
	Медиана	76,5			Медиана	76	
	Отклонение	25,501416083			Отклонение	10,750193797	

Рис. 29.

Отрицательные z-показатель означают, что значение меньше среднего, положительный, что больше среднего.

Среднее значение нормированного отклонения всегда равно 0, а стандартное отклонение нормированных отклонений всегда равно 1.

В чём бы ни измерялась переменная, среднее значение ее нормированных отклонений всегда равно 0, а стандартное отклонение нормированных отклонений всегда равно 1.

Благодаря этим свойствам z-показатель может использоваться для сравнения значений, имеющих разный размах (разность между максимальным и минимальным значениями) и разные единицы измерения.

Практическая часть

Для таблицы из лабораторной работы №1 найдите среднее, медиану, стандартное и нормированное отклонение.

Контрольные вопросы

1. **Что такое медиана?**
2. **Чем медиана отличается от среднего арифметического?**
3. **Что такое стандартное отклонение?**
4. **Что такое нормированное отклонение?**
5. **Зачем нужно стандартное и нормированное отклонение?**

Лабораторная работа №3. Распределения вероятностей.

Цель:

Узнать о реализации разных видов распределений в Excel.

Задачи:

- Изучить функции для работы с вероятностными распределениями в Excel;
- Узнать о полезных функциях из семейств функций распределения в Excel;
- Научиться применять функции распределений.

Теоретическая часть

В прошлых лабораторных работах были рассмотрены основы статистики и стандартные статистические функции в Excel. В третьей лабораторной работе мы кратко пройдемся по функциям, которые в Excel отвечают за распределения и понадобятся для выполнения задания.

Нормальное распределение.

Для расчета плотности или вероятности нормального распределения используется функция НОРМ.РАСП().

Аргументы:

x – значение, для которого рассчитывается плотность/вероятность или значение функции нормального распределения

среднее – математическое ожидание, используемое в качестве первого параметра модели нормального распределения

стандартное_отклонение – среднеквадратическое отклонение

интегральная – при 0 рассчитывается плотность, при 1 – вероятность

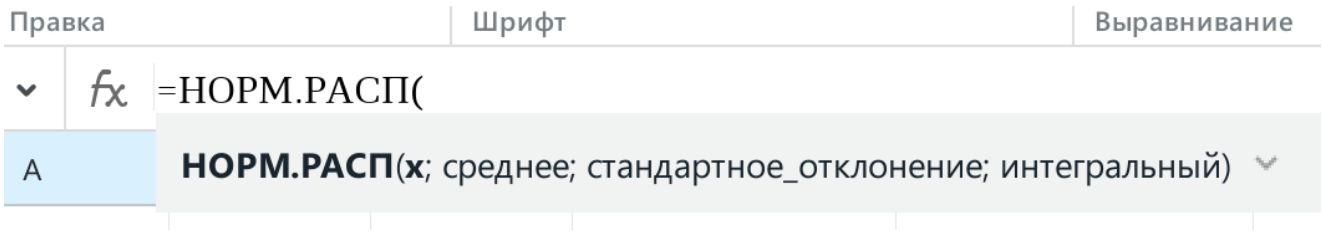


Рис. 30.

Функция обратная НОРМ.РАСП() – НОРМ.ОБР(). Считает значение x при известной вероятности.

Аргументы:

x – вероятность

среднее – математическое ожидание

стандартное отклонение – среднеквадратическое отклонение

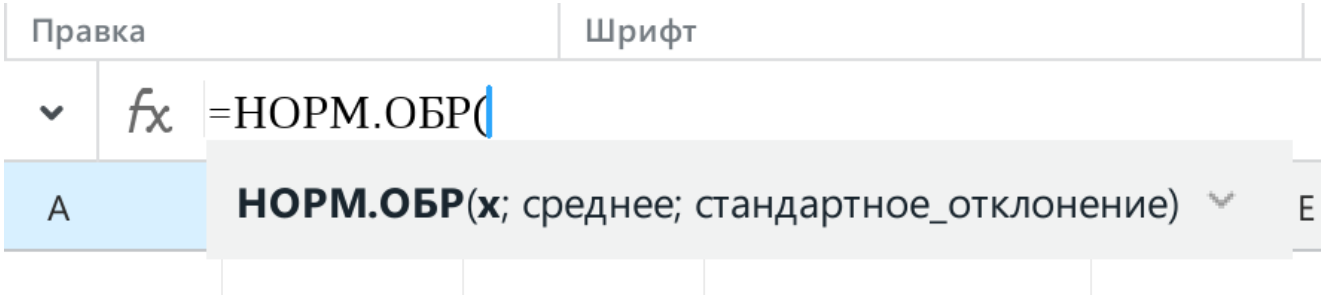


Рис. 31.

Пример:

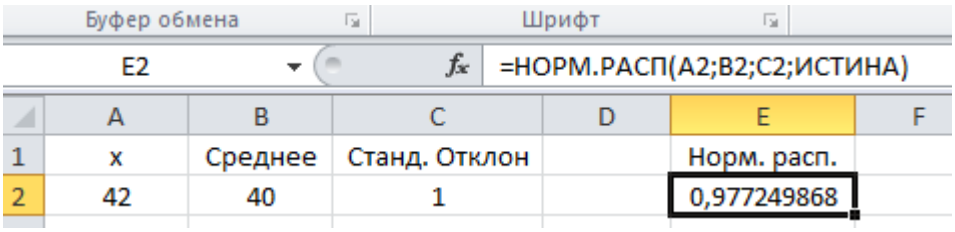


Рис. 32.

Стандартное нормальное распределение.

Для расчета плотности или вероятности стандартного нормального распределения используется функция НОРМ.СТ.РАСП().

Аргументы:

x – значение стандартизированной переменной

[интегральный] – при 0 рассчитывается плотность, при 1 - вероятность

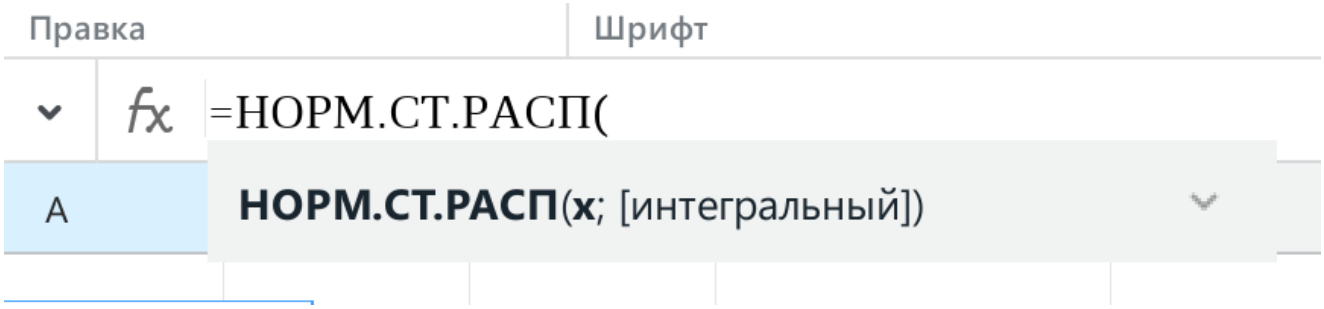


Рис. 33.

Функция обратная НОРМ.СТ.РАСП() – НОРМ.СТ.ОБР(). Считает значение x при известной вероятности.

Аргументы:

x – вероятность

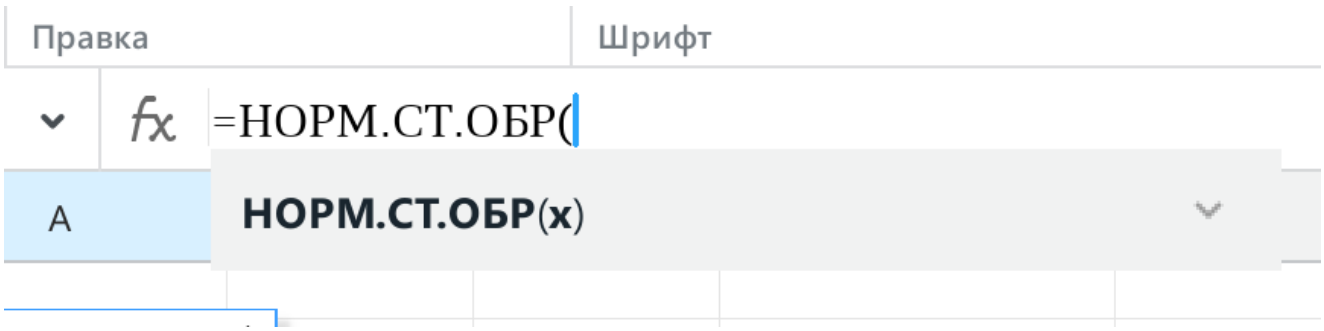


Рис. 34.

Пример:

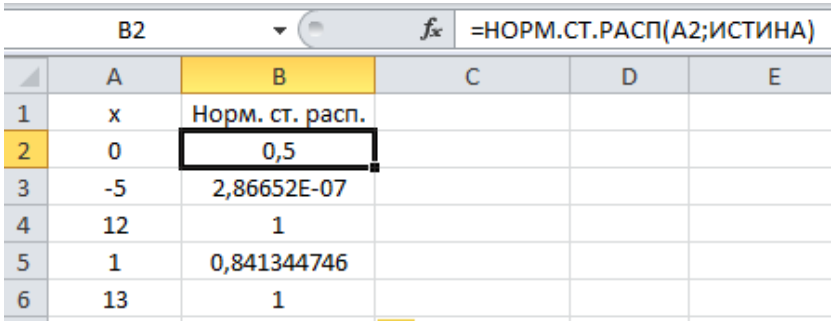


Рис. 35.

Хи-квадрат (критерий согласия Пирсона).

Функция ХИ2.РАСП() возвращает вероятность либо плотность распределения хи-квадрат.

Аргументы:

x – значение, для которого требуется вычислить плотность/вероятность

степени_свободы – число степеней свободы

интегральный – при 0 рассчитывается плотность, при 1 – вероятность

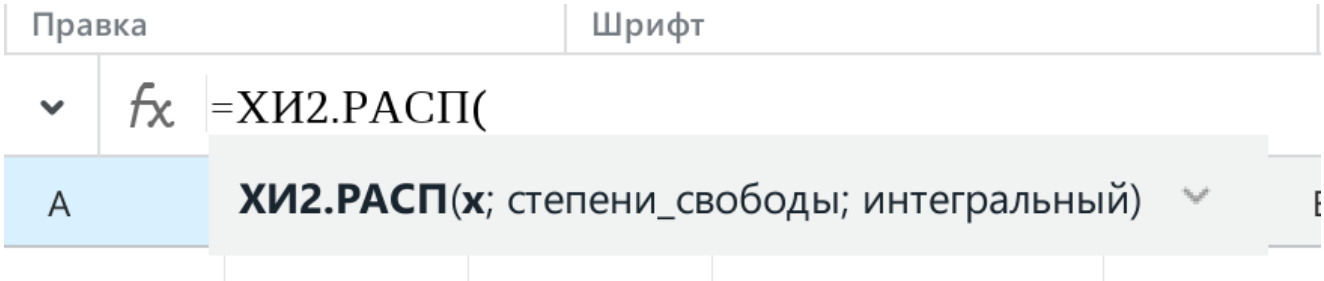


Рис. 36.

Функция обратная ХИ2.РАСП() – ХИ2.ОБР(). Считает значение x при

известной вероятности.

Аргументы:

вероятность – вероятность

степени_свободы – число степеней свободы

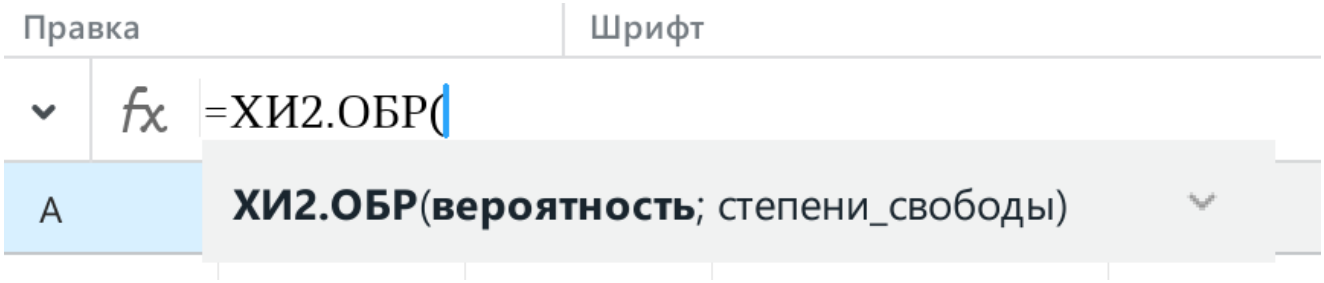


Рис. 37.

ХИ2.РАСП.ПХ() – считает правостороннюю вероятность распределения хи-квадрат.

Аргументы:

x – значение, для которого требуется вычислить вероятность

степени_свободы – число степеней свободы

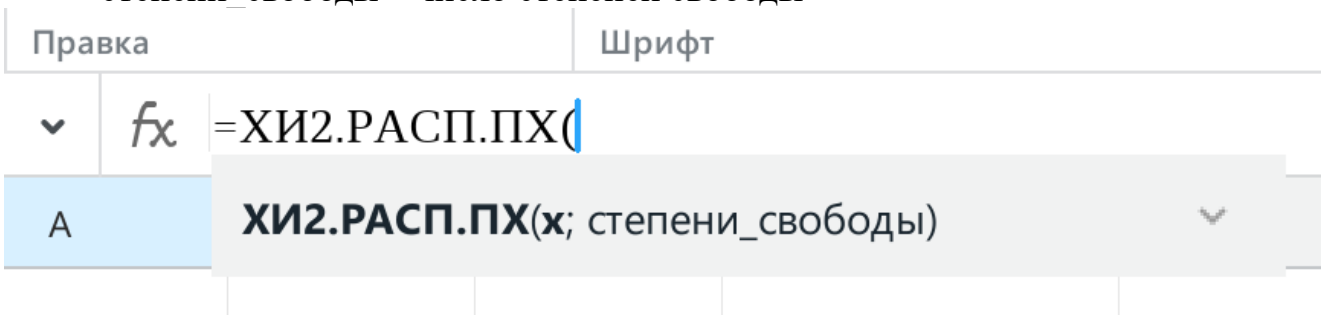


Рис. 38.

Функция обратная ХИ2.РАСП.ПХ() – ХИ2.ОБР.ПХ(). Считает значение x при известной правосторонней вероятности хи-квадрат.

Аргументы:

вероятность – вероятность

степени_свободы – число степеней свободы

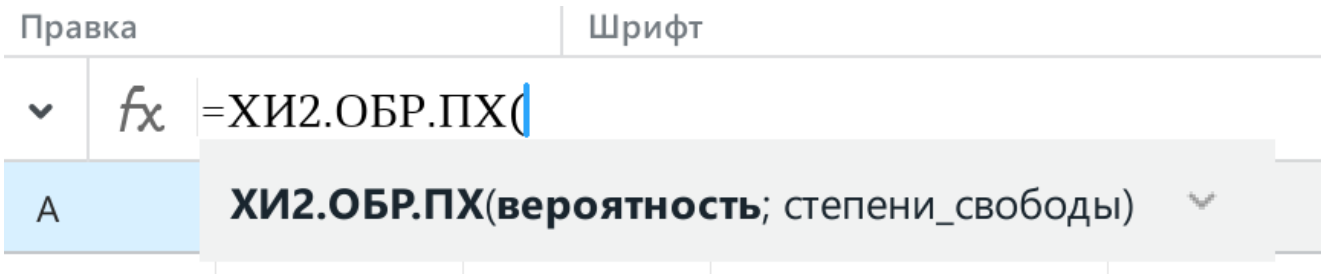
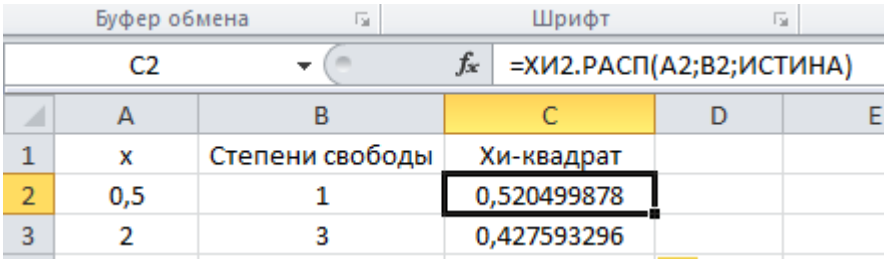


Рис. 39.

Пример:



	A	B	C	D	E
1	x	Степени свободы	Хи-квадрат		
2	0,5	1	0,520499878		
3	2	3	0,427593296		

Рис. 40.

Т - критерий Стьюдента.

Для работы с критерием Стьюдента в программе Excel имеется семь функций:

СТЬЮДЕНТ.РАСП() – левостороннее t-распределение Стьюдента, возвращает левостороннее p-значение при поданном значении t критерия, количества степеней свободы и 0 либо 1 для вычисления плотности либо вероятности соответственно

СТЬЮДЕНТ.РАСП.2Х() – двухстороннее распределение

СТЬЮДЕНТ.РАСП.ПХ() – правостороннее распределение

СТЬЮДЕНТ.ОБР() – обратная функция, возвращает t-значение, при поданной вероятности и количестве степеней свободы

СТЬЮДЕНТ.ОБР.2Х() – обратная функция для двухстороннего распределения

СТЬЮДЕНТ.ТЕСТ() - функция для проверки гипотезы о равенстве математических ожиданий в двух выборках

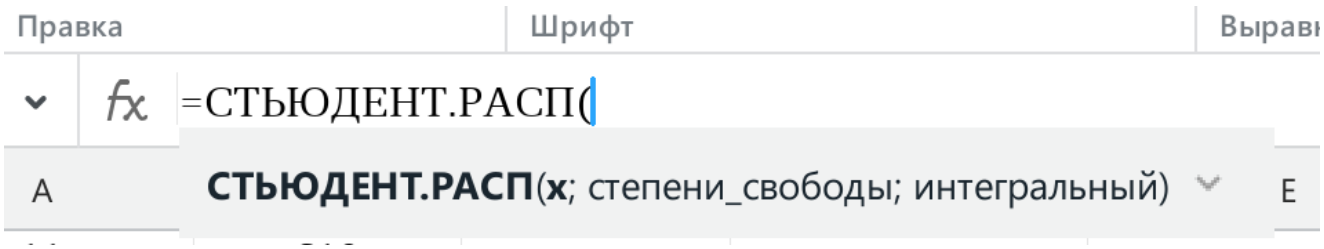


Рис. 41.

Пример:

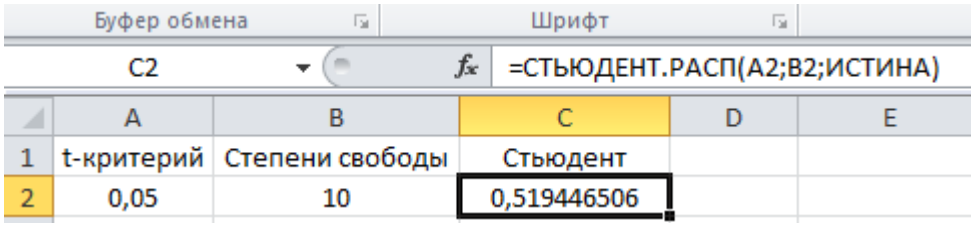


Рис. 42.

Критерий Фишера (F – тест).

Для осуществления преобразования Фишера в программе Excel используется функция ФИШЕР().

Аргументы:

значение – числовое значение, для которого необходимо получить преобразование

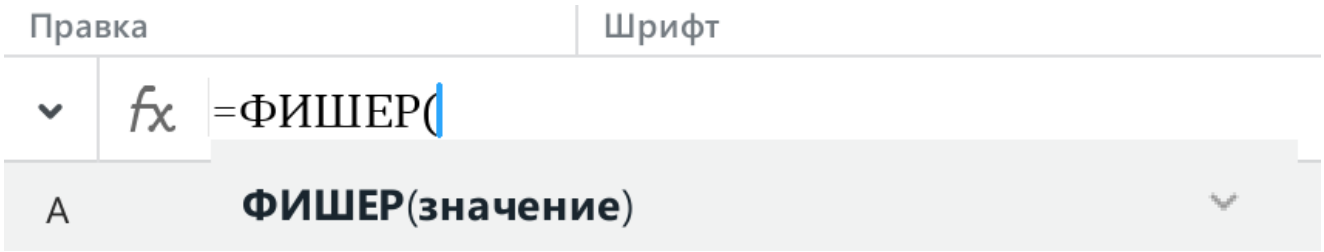


Рис. 43.

Пример:

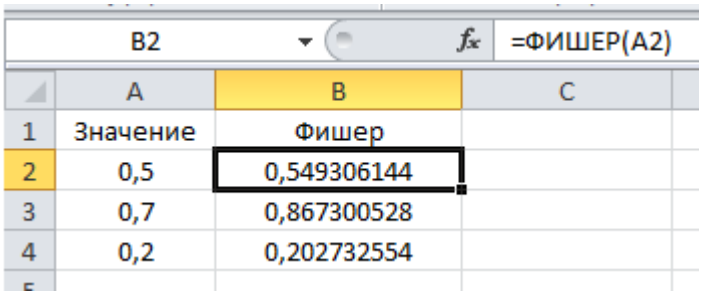


Рис. 44.

Практическая часть

Для данной в файле таблицы найти вероятности распределений, посчитать критерии Стьюдента и Фишера, уровень значимости принимать равным 0,05 и количество степеней свободы равное 5.

Контрольные вопросы

1. **Для чего нужен интегральный параметр в функциях Excel?**
2. **С помощью какой функции можно рассчитать вероятность стандартного нормального распределения?**
3. **Как рассчитать двустороннее распределение Стьюдента?**

Лабораторная работа №4. Регрессионный, дисперсионный и корреляционный анализ.

Цель:

Узнать, как можно выполнять разные виды статистического анализа в Excel.

Задачи:

- Научиться пользоваться пакетом анализа;
- Научиться пользоваться новыми функциями Excel.

Теоретическая часть

В данной лабораторной работе мы рассмотрим какие существует способы и функции для осуществления корреляционного, регрессионного и дисперсионного анализа в программе Excel.

Дисперсионный анализ.

Для выполнения дисперсионного анализа в Excel потребуется пакет инструментов анализа. Если во вкладке “Данные” в разделе “Анализ” у вас нету опции “Анализ данных”, то проделайте следующие действия:

1. Зайдите на вкладку “Файл” и выберите раздел “Параметры”:

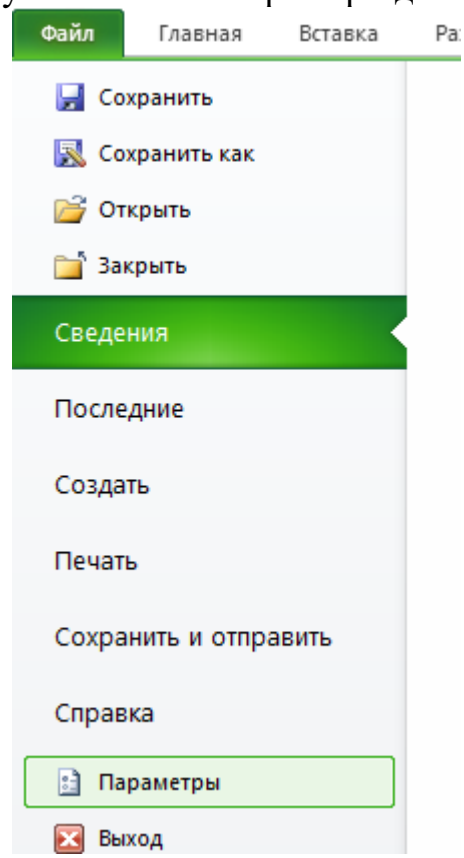


Рис. 45.

2. Выберите раздел “Надстройки”

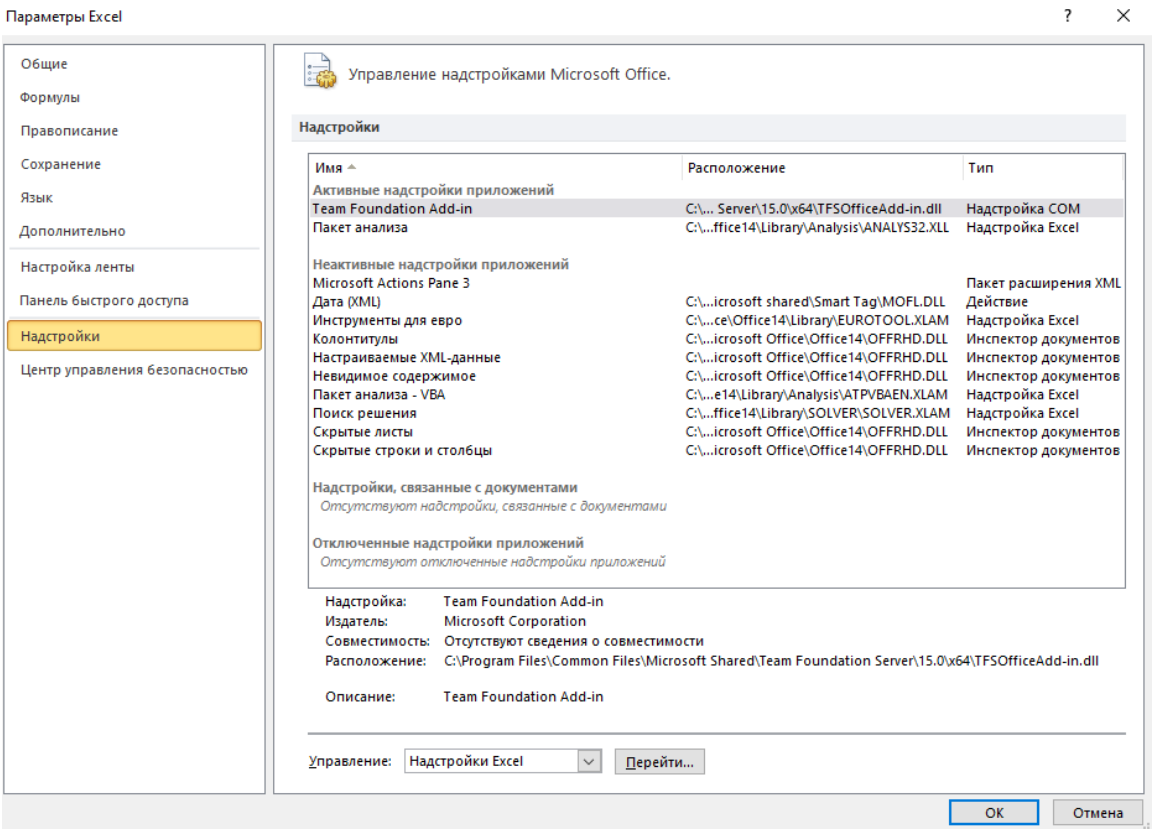


Рис. 46.

3. Выберите “Пакет анализа”, нажмите “Перейти”, во всплывающем окне выберите “Пакет анализа” и нажмите “Ок”.

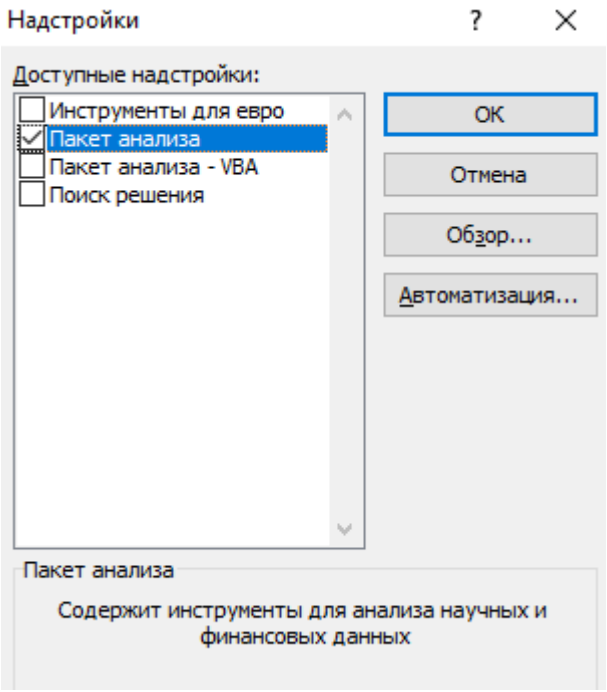


Рис. 47.

Нажав на опцию “Анализ данных” вы сможете увидеть множество

полезных функций, связанных с дисперсионным анализом и не только:

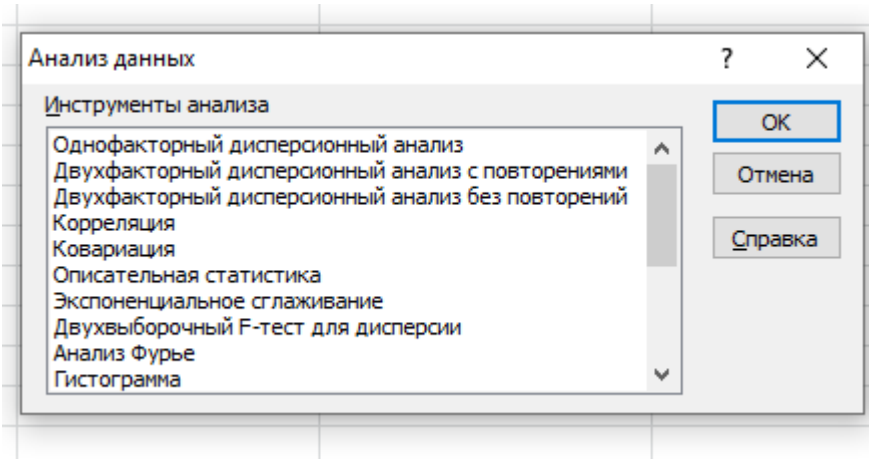


Рис. 48.

Для примера выполним однофакторный дисперсионный анализ. Предположим у нас есть таблица с некоторыми данными результатов тестов учащихся:

K	L	M
Тест 1	Тест 2	Тест 3
80	67	55
90	89	88
91	71	73
76	74	82
87	84	94
73	81	81
88	89	68
92	59	63
65	68	85
69	63	99

Рис. 49.

Для выполнения однофакторного дисперсионного анализа выберем его среди инструментов анализа и введем либо вручную, либо с помощью выделения ячеек диапазон ячеек. Так же следует выбрать и выходную ячейку:

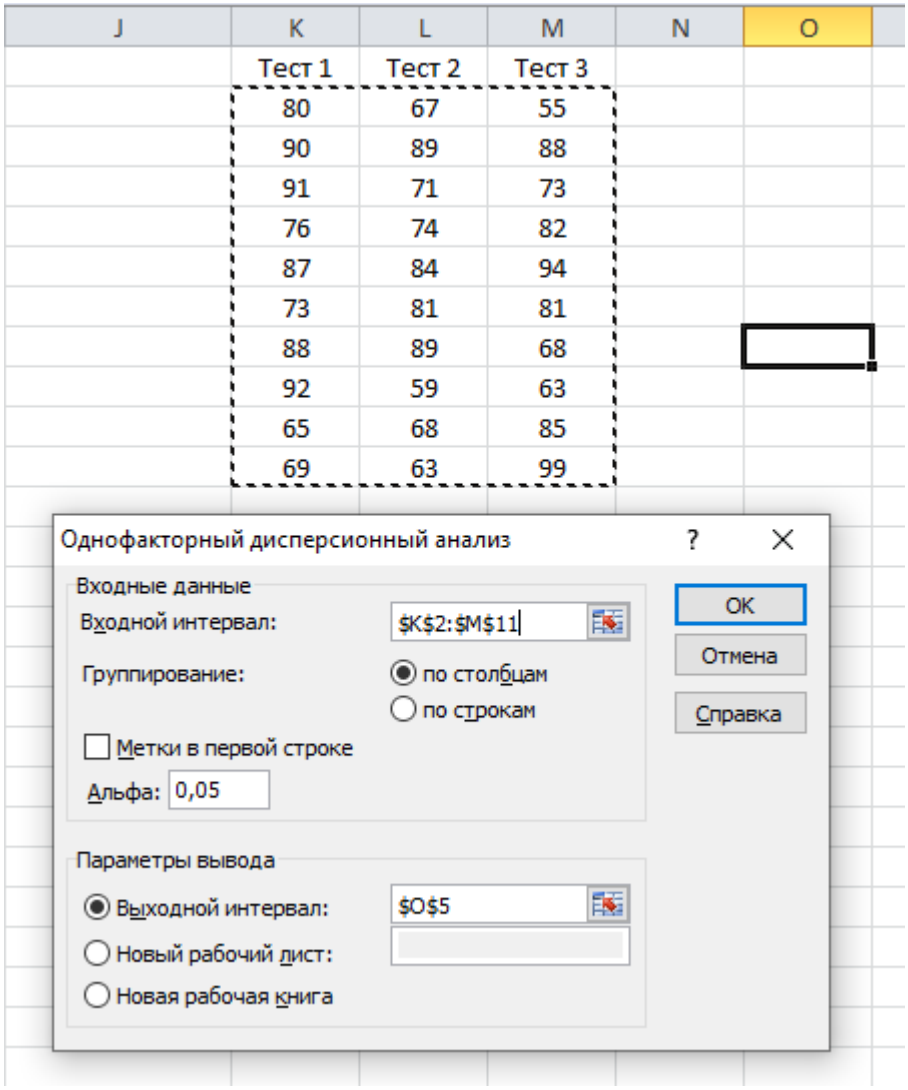


Рис. 50.

На выходе получим большую таблицу со всеми данными анализа:

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
Столбец 1	10,000	811,000	81,100	97,433
Столбец 2	10,000	745,000	74,500	115,167
Столбец 3	10,000	788,000	78,800	193,733

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	224,467	2,000	112,233	0,829	0,447	3,354
Внутри групп	3657,000	27,000	135,444			
Итого	3881,467	29,000				

Рис. 51.

Регрессионный анализ.

Для выполнения регрессионного анализа в Excel также отлично подойдет функция из пакета инструментов анализа. Предположим у нас есть таблица с количеством часов обучения и баллом, полученным за тест студентами:

	G	H
1	Часы	Балл
2	10	50
3	50	75
4	30	65
5	60	79
6	20	56
7	40	61
8	70	74
9	30	62
10	90	93
11	40	62
12	20	57
13	60	81
14	30	60
15	50	73
16	60	77
17	10	51
18	90	97
19	40	59
20	20	55
21	30	66

Рис. 52.

Для проведения регрессионного анализа выберем его в пакете анализа:

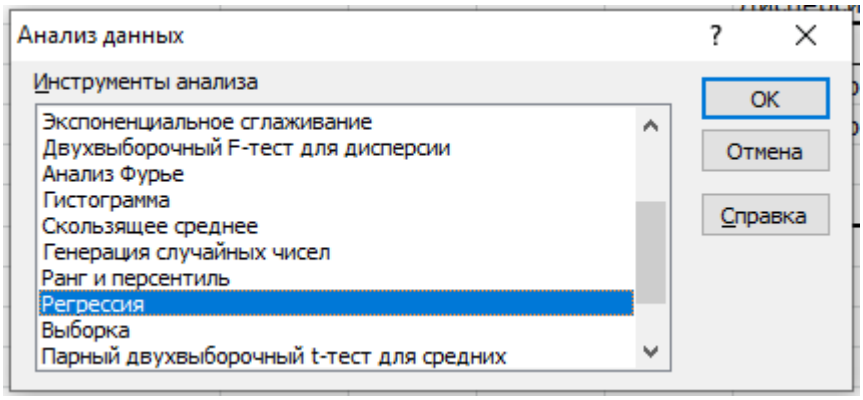


Рис. 53.

Далее введем параметры и ячейку вывода результата:

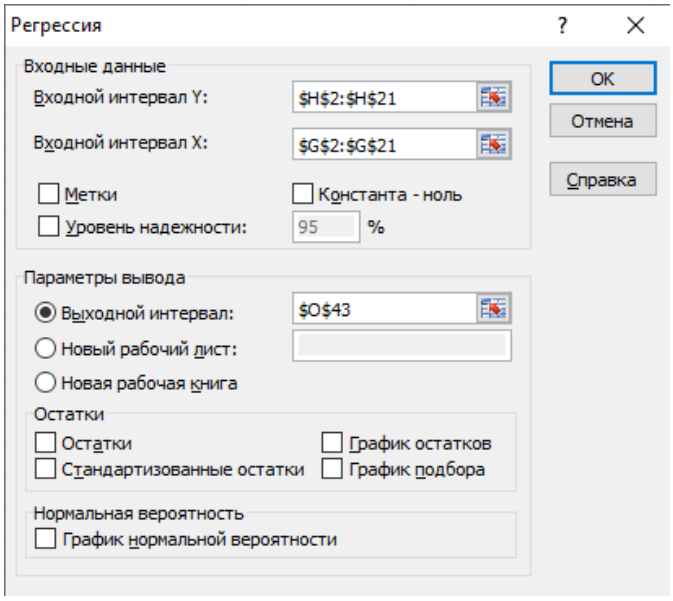


Рис. 54.

На выходе получим таблицу:

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,95859679
R-квадрат	0,918907806
Нормированный R-квадрат	0,914402684
Стандартная ошибка	3,83851878
Наблюдения	20

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	3005,333924	3005,333924	203,9695765	2,92114E-11
Остаток	18	265,2160757	14,73422643		
Итого	19	3270,55			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	44,99338061	1,803710877	24,94489621	2,06139E-15	41,20392468	48,78283655	41,20392468	48,78283655
Переменная X 1	0,533096927	0,037327034	14,28179178	2,92114E-11	0,454675738	0,611518115	0,454675738	0,611518115

Рис. 55.

Корреляционный анализ.

Для расчета коэффициента корреляции в программе Excel существует функция КОРРЕЛ().

Аргументы:

данные_y – первый диапазон значений

данные_x – второй диапазон значений

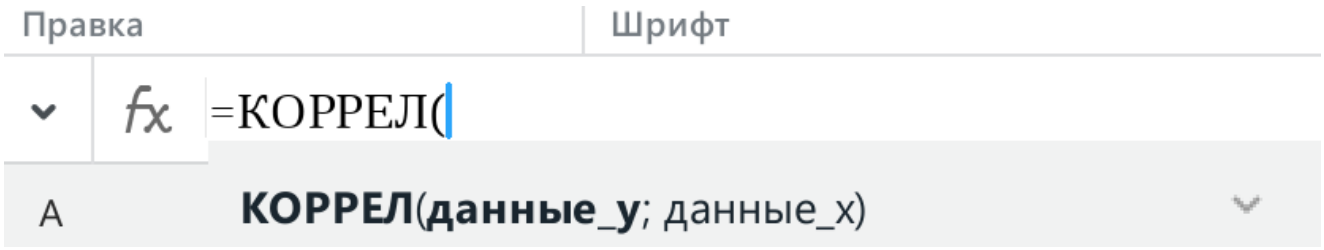
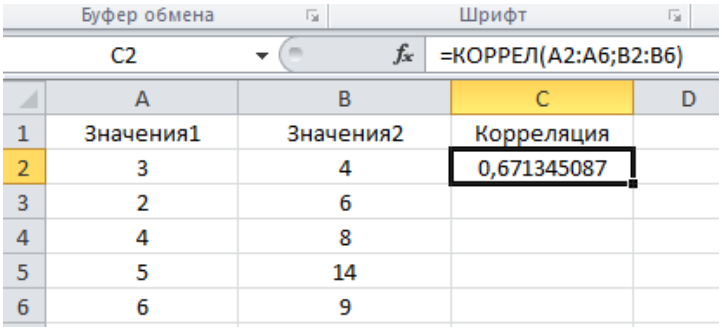


Рис. 56.

Пример:



	A	B	C	D
1	Значения1	Значения2	Корреляция	
2	3	4	0,671345087	
3	2	6		
4	4	8		
5	5	14		
6	6	9		

Рис. 57.

Практическая часть

Дисперсионный анализ:

Произведено по восемь испытаний на каждом из шести уровней фактора. Методом дисперсионного анализа при уровне значимости 0,01 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Принять $y_{ij} = x_{ij} - 100$.

Корреляционный анализ:

Тринадцать цветных полос расположены в порядке убывания окраски от темной к светлой и каждой полосе
присвоен ранг — порядковый номер. В итоге получена последовательность рангов
xi … (есть в файле)

При проверке способности различать оттенки цветов испытуемый расположил полосы в следующем порядке:
yi … (есть в файле)

Найти коэффициент ранговой корреляции Спирмена между «правильными» рангами x; и рангами y, которые присвоены полосам испытуемым.

Регрессионный анализ:

Найти выборочные уравнения прямых линий регрессии У на Х и Х на У по данным, приведенным в корреляционной таблице.

Контрольные вопросы

1. **Для чего нужен пакет инструментов анализа?**
2. **Как выполнять дисперсионный анализ в Excel?**
3. **Как выполнять корреляционный анализ в Excel?**
4. **Как выполнять регрессионный анализ в Excel?**

Лабораторная работа №5. Проверка гипотез в Excel.

Цель:

Применить полученные ранее знания и умения для проверки статистических гипотез в Excel.

Задачи:

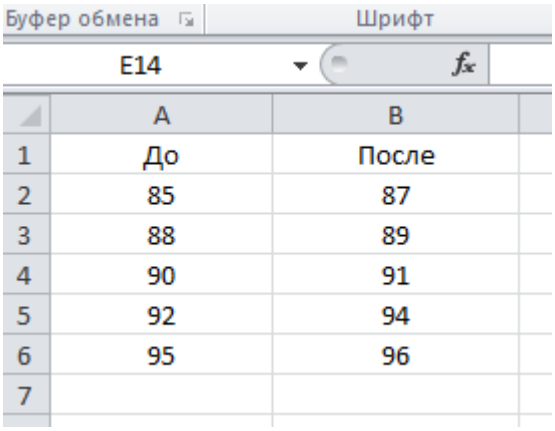
- Научиться решать более сложные статистические задачи на Excel;
- Научиться использовать известные функции для новых задач.

Теоретическая часть

Для проверки статистических гипотез в Excel может использоваться большое количество функций (многие были показаны в предыдущих лабораторных работах), и их использование конечно же зависит от проверяемой гипотезы, поэтому мы ограничимся несколькими простыми примерами для общего понимания того, как производить проверку гипотез в Excel.

Пример №1: проверить, есть ли изменения в результатах до и после эксперимента.

Данные:



	A	B
1	До	После
2	85	87
3	88	89
4	90	91
5	92	94
6	95	96
7		

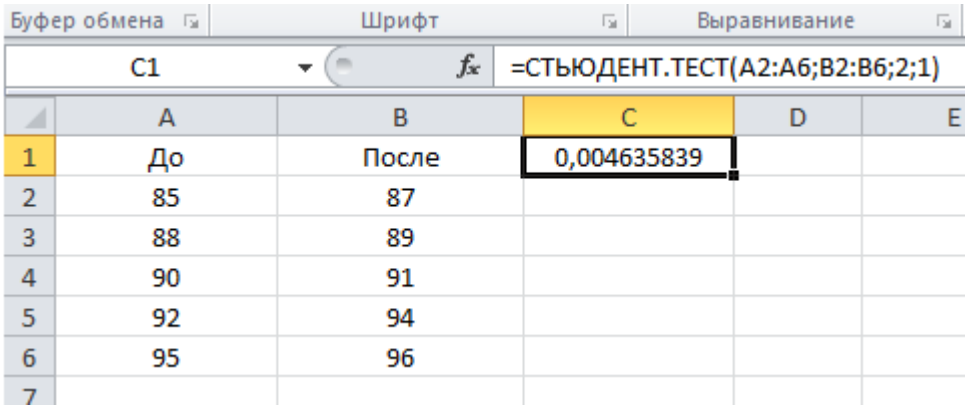
Рис. 58.

Нулевая гипотеза – изменений в результатах нет.

Альтернативная гипотеза – изменения есть.

Чтобы понять, есть ли изменения, мы можем сравнить p значение со стандартным уровнем значимости 0,05.

Для нахождения p значения используем функцию СТЬЮДЕНТ.ТЕСТ():



	A	B	C	D	E
1	До	После	0,004635839		
2	85	87			
3	88	89			
4	90	91			
5	92	94			
6	95	96			
7					

Рис. 59.

Цифра 2 в функции указывает на двусторонний, а 1 на парный тест.

В результате мы получаем число, которое сильно меньше порога значимости, соответственно изменения в результатах есть, нулевая гипотеза

отвергается, альтернативная подтверждается.

Пример №2: проверить, отличаются ли средние значения нескольких групп.

Данные:

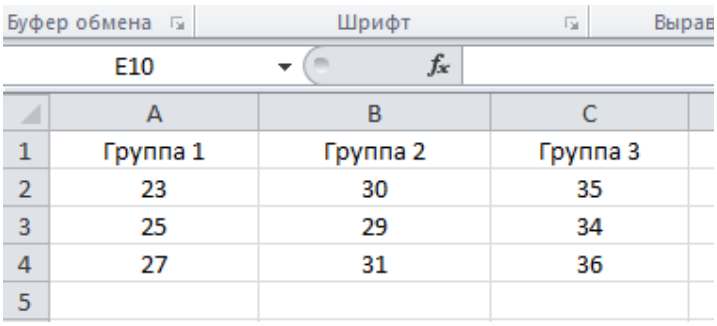


Рис. 60.

Нулевая гипотеза – средние значения равны.

Альтернативная гипотеза – средние значения отличаются.

Будем также сравнивать p значение с уровнем значимости 0,05, но в этот раз нам необходимо провести однофакторный дисперсионный анализ. Его можно провести уже с помощью известного вам пакета инструментов анализа:

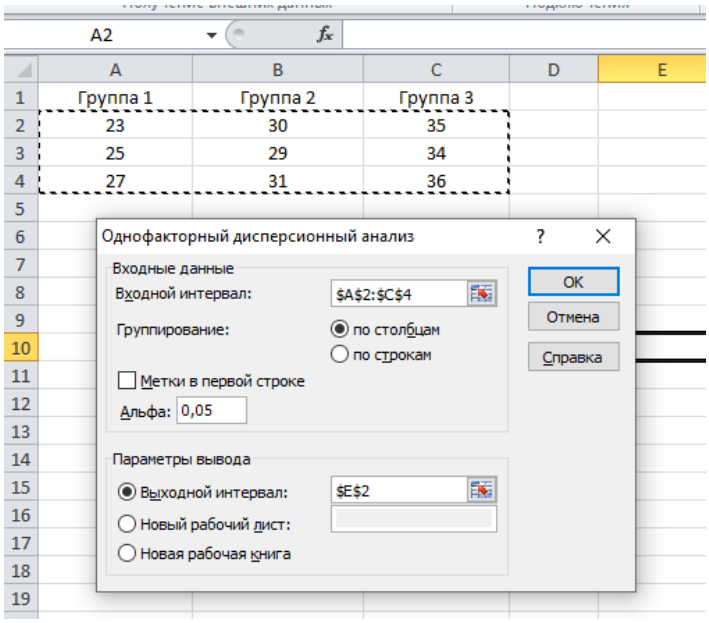


Рис. 61.

В результате мы получим p значение равное ~0,000406, значит, средние значения отличаются, нулевая гипотеза отвергается, альтернативная подтверждается.

Практическая часть

Решите задачу на проверку гипотезы в Excel.

Контрольные вопросы

1. **Для чего нужен пакет инструментов анализа?**
2. **Как пользоваться пакетом инструментов анализа?**
3. **Есть ли встроенные функции в Excel для какого либо вида анализа?**

Примечание

В предыдущих лабораторных работах вы научились работать с данными в Excel, но это не единственный инструмент для анализа. Существует множество и других вариантов.

Так как раньше вы уже знакомились с языками программирования для работы с данными, в том числе с Python, то в рамках следующей лабораторной работы вам потребуется, используя знание языка Python и методов математической статистики, решить задачи на регрессионный, дисперсионный, корреляционный анализ и задачи на проверку гипотез.

Лабораторная работа разделена на несколько частей, отчет можно по желанию объединить в один файл.

Лабораторная работа №6. Решение статистических задач в Python.

Цель:

Научиться применять знания математической статистики и языков программирования для работы с данными для решения задач.

Задачи:

- Повторить язык Python;
- Закрепить знания математической статистики;
- Узнать, как можно обрабатывать статистические данные и применять статистические функции в Python.

Теоретическая часть

Для анализа данных в Python нам потребуются следующие библиотеки: Pandas, Numpy, SciPy, Statsmodels, Scikit-learn. Можно также установить Matplotlib и Seaborn, если вы хотите визуализировать какие то данные в будущем.

Установить все эти библиотеки можно с помощью пакетного менеджера, например pip, который обычно устанавливается вместе с Python. Для установки библиотек запустите командную строку от имени администратора и введите:

```
pip install pandas numpy scipy statsmodels scikit-learn
```

После установки потребуется только подключить необходимые библиотеки в самом файле.
Разберем один пример решения статистических задач, чтобы вам было проще выполнять задания из следующих лабораторных работ.
Задание:

Точность работы станка-автомата проверяется по дисперсии контролируемого размера изделий, которая не должна превышать $\sigma_0^2 = 0,1$. Взята проба из 25 случайных отобранных изделий, причем получены следующие результаты измерений:

контролируемый размер изделий пробы	x_i	3,0	3,5	3,8	4,4	4,5
частота	n_i	2	6	9	7	1

Требуется при уровне значимости 0,05 проверить, обеспечивает ли станок требуемую точность.

Рис. 62.

(Правильный ответ: Станок не обеспечивает необходимую точность. Хи-квадрат, равное 48, больше критической точки, равной 36,4.)

Программа с комментариями:

```
import numpy as np
from scipy.stats import chi2

# Данные
sizes = np.array([3, 3.5, 3.8, 4.4, 4.5])
frequencies = np.array([2, 6, 9, 7, 1])

# Создание выборки на основе частот
sample = np.repeat(sizes, frequencies)

# Размер выборки
n = len(sample)

# Среднее значение
mean_size = np.mean(sample)

# Выборочная дисперсия
sample_variance = np.var(sample, ddof=1)

# Заданная дисперсия
sigma_0_squared = 0.1

# Вычисление статистики хи-квадрат
chi_square_statistic = (n - 1) * sample_variance / sigma_0_squared

# Уровень значимости
alpha = 0.05

# Критическая точка
critical_value = chi2.ppf(1 - alpha, n - 1)

# Результат
if chi_square_statistic < critical_value:
    print("Станок обеспечивает требуемую точность.")
else:
    print("Станок не обеспечивает требуемую точность.")

print(f"Статистика хи-квадрат: {chi_square_statistic:.2f}")
print(f"Критическое значение: {critical_value:.2f}")
```

Практическая часть

Лабораторная работа №6.1. Проверка статистических гипотез в Python.

Решите задачу на проверку статистической гипотезы из лабораторной работы №5 с помощью Python.

Лабораторная работа №6.2. Дисперсионный анализ в Python.

Выполните задание из лабораторной работы №4 по дисперсионному анализу на Python.

Лабораторная работа №6.3. Корреляционный анализ в Python.

Выполните задание из лабораторной работы №4 по корреляционному анализу на Python.

Лабораторная работа №6.4. Регрессионный анализ в Python.

Выполните задание из лабораторной работы №4 по регрессионному анализу на Python.

Контрольные вопросы

1. **Какие библиотеки понадобятся для работы с данными в Python?**
2. **Какие функции используются при разных видах анализа в Python (чаще всего)?**
3. **Какие преимущества, по вашему мнению, у языка Python по сравнению с другими средствами работы с данными?**

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Маккинни У. Python и анализ данных. Первичная обработка данных с применением pandas, NumPy и Jupiter, 3-е изд. – 2023
2. Гатман А. Дж., Голдмейер Дж.. Разберись в Data Science: как освоить науку о данных и научиться думать как эксперт. – Эксмо, 2023
3. Такахаси Син. Образовательная манга: занимательная статистика. – 2010
4. Такахаси Син. Образовательная манга: регрессионный анализ. – 2016