

Министерство образования Республики Беларусь
Учреждение образования «Витебский государственный
университет имени П.М. Машерова»
Кафедра информатики и информационных технологий

А.А. Чиркина, Н.В. Булгакова

Использование статистических методов в педагогическом исследовании

Методические рекомендации

*Витебск
ВГУ имени П.М. Машерова
2015*

УДК 37.012.4(075.8)

ББК 74в631я73

Ч-65

Печатается по решению научно-методического совета учреждения образования «Витебский государственный университет имени П.М. Машерова». Протокол № 1 от 23.10.2015 г.

Авторы: доцент кафедры информатики и информационных технологий ВГУ имени П.М. Машерова, кандидат биологических наук **А.А. Чиркина**; старший преподаватель кафедры информатики и информационных технологий ВГУ имени П.М. Машерова **Н.В. Булгакова**

Р е ц е н з е н т :

доцент кафедры прикладной математики
ВГУ имени П.М. Машерова, кандидат физико-математических наук,
доцент *Л.В. Маркова*;
доцент кафедры технологии и оборудования машиностроительного
производства УО «ВГТУ» *Ю.Е. Махаринский*

Чиркина, А.А.

Ч-65 Использование статистических методов в педагогическом исследовании : методические рекомендации / А.А. Чиркина, Н.В. Булгакова. – Витебск : ВГУ имени П.М. Машерова, 2015. – 48 с.

Методические рекомендации написаны в соответствии с программой курса «Статистические методы в педагогическом исследовании» и предназначены для отработки практических навыков по статистическим методам анализа экспериментальных данных в педагогических исследованиях с помощью пакета программ MS Excel. Содержат теоретические сведения, необходимые при подготовке к лабораторным работам, варианты заданий, контрольные вопросы для самопроверки, примеры решения задач.

Предлагаемое учебное издание рекомендовано студентам специальности 1-02 05 01 «Математика и информатика» дневной и заочной форм обучения.

УДК 37.012.4(075.8)

ББК 74в631я73

© Чиркина А.А., Булгакова Н.В., 2015

© ВГУ имени П.М. Машерова, 2015

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	4
Общие указания по выполнению лабораторных работ	5
ЛАБОРАТОРНЫЕ РАБОТЫ	6
Лабораторная работа 1	
ТЕМА: «Получение статистических характеристик выборки. Меры центральной тенденции, вариации. Анализ вида распределения выборочной совокупности. Контроль данных статистического наблюдения, определение выбросов»	6
Лабораторная работа 2	
ТЕМА: «Анализ вида распределения экспериментальных данных»	12
Лабораторная работа 3	
ТЕМА: «Статистическое изучение связи. Корреляционный анализ»	20
Лабораторная работа 4	
ТЕМА: «Регрессионный анализ»	25
Лабораторная работа 5	
ТЕМА: «Интервальное оценивание. Определение доверительных интервалов для среднего и медианы. Сравнение группы с популяций»	31
Лабораторная работа 6	
ТЕМА: «Проверка гипотез. Выявление достоверности различий между двумя выборками с помощью t-критерия Стьюдента» ...	35
Лабораторная работа 7	
ТЕМА: «Построение таблиц сопряженности. Выявление достоверности различий между двумя выборками с помощью критерия хи-квадрат»	40
РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА	46
ПРИЛОЖЕНИЯ	47

ВВЕДЕНИЕ

Цель изучения дисциплины «Статистические методы в педагогическом исследовании» – формирование профессиональной компетентности у будущих специалистов высшей квалификации в обработке результатов экспериментов и измерений в педагогике и психологии с применением методов математической статистики. Изучение предлагаемого курса позволяет сформировать практические навыки обработки и анализа статистических данных, расчета важнейших статистических показателей с использованием программных средств статистической обработки данных. В результате изучения дисциплины студент должен

знать:

- принципы и методы сбора, обработки, анализа и наглядного представления информации, формы подготовки и представления экспериментальных данных, используемые в педагогических исследованиях;
- структуру и содержание педагогического эксперимента, его виды;
- методы математической статистики и алгоритмы решения типовых задач, используемые при планировании, проведении и обработке результатов экспериментов в педагогике и психологии;
- основы статистического моделирования и прогнозирования;
- инструментальные средства статистической обработки и анализа данных.

уметь:

- собирать и регистрировать статистическую информацию, проводить первичную обработку и контроль материалов наблюдения;
- выбирать и использовать математические методы статистической обработки данных; решать типовые для педагогики и психологии статистические задачи;
- планировать педагогический эксперимент, выполнять статистическую обработку и анализ результатов экспериментальных данных, анализировать, обобщать, обрабатывать и оформлять результаты экспериментального исследования;
- использовать инструментальные средства статистической обработки данных.

Выполнение студентами лабораторных работ и четкое следование методическим рекомендациям способствует закреплению и углублению теоретических знаний, интенсификации практической составляющей использования математической статистики в педагогическом исследовании.

Общие указания по выполнению лабораторных работ

Целью проведения лабораторных работ является закрепление теоретических основ учебной дисциплины «Статистические методы в педагогическом исследовании», овладение современными методами обработки результатов педагогических измерений с применением компьютерных средств, знакомство с особенностями применения статистических методов для решения профессиональных задач в педагогических исследованиях.

К выполнению лабораторных работ допускаются студенты, прошедшие инструктаж по правилам техники безопасности в компьютерной лаборатории.

Лабораторные работы снабжены необходимыми теоретическими сведениями, подробными указаниями по их выполнению, индивидуальными заданиями. Лабораторная работа состоит из следующих этапов:

- изучение лекционного материала, материалов по темам данных методических рекомендаций и дополнительной литературы перед выполнением каждой лабораторной работы;
- выполнение работы на компьютере в соответствии с заданием;
- оформление отчета;
- защита лабораторной работы преподавателю.

Отчет по каждой лабораторной работе должен содержать:

- заголовок лабораторной работы (название и цель работы);
- задание к лабораторной работе;
- постановку задачи, используемый метод решения;
- описание последовательности действий, произведенных при выполнении работы (ход работы) и комментарии к ним;
- результаты выполнения лабораторной работы и выводы по теме исследования.

Отчет о лабораторной работе должен быть защищен студентом в установленные сроки. При защите отчета студент обязан знать основные теоретические сведения, содержание, порядок выполнения работы и ответить на поставленные контрольные вопросы.

При выполнении лабораторной работы студент должен продемонстрировать: умение работать с учебной и учебно-методической литературой; знание теоретического материала и умение использовать его для решения конкретных практических задач; познавательные и аналитические способности, самостоятельность мышления, творческую активность.

ЛАБОРАТОРНЫЕ РАБОТЫ

Лабораторная работа 1

Тема: «Получение статистических характеристик выборки. Меры центральной тенденции, вариации. Анализ вида распределения выборочной совокупности. Контроль данных статистического наблюдения, определение выбросов»

Цель: изучить методику и освоить практические приемы статистической обработки вариационного ряда средствами пакета «Microsoft Excel». Провести первичную обработку данных, вычислить оценки числовых характеристик положения, вариации и формы выборочной совокупности. Проанализировать полученные результаты.

Основные теоретические сведения

Для объективной оценки полученных результатов исследования необходима их математическая обработка. Среди выборочных характеристик выделяют показатели, относящиеся к центру распределения (среднее арифметическое, мода, медиана), показатели рассеяния вариант (размах вариации, интерквартильный размах, дисперсия, стандартное отклонение) и меры формы распределения (асимметрия, эксцесс). В электронной таблице Excel имеется ряд специальных функций, предназначенных для вычисления выборочных характеристик.

Показатель	Функция Excel
Средним значением выборки называется величина $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, где $x_1, \dots, x_n$ – значения вариант, n – число единиц совокупности. Среднее значение – это центр выборки, вокруг которого группируются элементы выборки. При увеличении числа наблюдений среднее приближается к математическому ожиданию.	СРЗНАЧ() вычисляет среднее арифметическое из нескольких чисел.
Мода – элемент выборки с наиболее часто встречающимся значением (наиболее вероятная величина).	МОДА() вычисляет наиболее часто встречающееся значение.
Медиана – это число, которое является серединой выборки, т.е. половина чисел имеет значения большие, чем медиана, а половина чисел имеет значения меньшие, чем медиана. Если ряд распределения дискретный и имеет нечетное число членов, то медианой будет варианта, находящаяся в середине упорядоченного ряда, т.е.	МЕДИАНА() позволяет получить медиану заданной выборки.

Показатель	Функция Excel
элемент с номером $(n+1)/2$. Если число членов ряда четное, то медиана равна среднему членов ряда с номерами $n/2$ и $n/2+1$.	
Дисперсия – это средняя арифметическая квадратов отклонений каждого значения признака от среднего значения выборки $D = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. Дисперсия выборки – это параметр, характеризующий степень разброса элементов выборки относительно среднего значения.	ДИСП() позволяет оценить дисперсию по выборочным данным.
Среднее квадратичное отклонение (выборочное стандартное отклонение) представляет собой корень квадратный из дисперсии $s = \sqrt{D}$. Параметр аналогичен дисперсии и используется в тех случаях, когда необходимо, чтобы показатель разброса случайной величины выражался в тех же единицах, что и среднее значение этой случайной величины.	СТАНДОТКЛОН() вычисляет стандартное отклонение.
Эксцесс – характеризует степень концентрации случаев вокруг среднего значения и является своеобразной мерой крутости кривой. Функция плотности нормального распределения имеет эксцесс равный 0. Если эксцесс больше нуля, то кривая распределения характеризуется островершинностью, т.е. является более крутой по сравнению с нормальной, а случаи более густо группируются вокруг среднего. При отрицательном эксцессе кривая является более плосковершинной, т.е. более полой по сравнению с нормальным распределением.	ЭКССЕСС() вычисляет оценку эксцесса по выборочным данным. Можно использовать один массив или одну ссылку на массив вместо аргументов, разделяемых точкой с запятой.
Асимметрия – величина, характеризующая несимметричность распределения элементов выборки относительно среднего значения. Принимает значения от -1 до 1. В случае симметричного распределения асимметрия равна 0. Если правая ветвь кривой, начиная от вершины, больше левой (правосторонняя асимметрия), то коэффициент асимметрии больше нуля. Если левая ветвь кривой больше правой (левосторонняя асимметрия), то коэффициент асимметрии меньше нуля.	СКОС() позволяет оценить асимметрию выборочного распределения.

Показатель	Функция Excel
Размах вариации (амплитуда, интервал) – разность между наибольшим $x_{\max}$ и наименьшим $x_{\min}$ значениями вариант: $R = x_{\max} - x_{\min}$. Этот показатель улавливает только крайние отклонения и не отражает отклонений всех вариант в ряду.	МИН() возвращает наименьшее, а МАКС() возвращает наибольшее значение из набора значений.
Квартиль – значения, которые делят вариационный ряд на четыре равные части. Верхний квартиль – это такое значение случайной величины, больше которого по величине 25% случаев выборки. Нижний квартиль – это такое значение случайной величины, меньше которого по величине 25% случаев выборки. Между первым и вторым квартилем также расположено 25% данных. Второй квартиль совпадает с медианой.	КВАРТИЛЬ() возвращает квартиль множества данных. 0 - минимальное значение; 1 - первый квартиль; 2 - значение медианы; 3 – третий квартиль; 4 - максимальное значение.

Стандартная ошибка (ошибка среднего) находится из выражения $m = \frac{s}{\sqrt{n}}$. Стандартная ошибка – это параметр, характеризующий степень возможного отклонения среднего значения, полученного на исследуемой ограниченной выборке, от истинного среднего значения (математического ожидания), полученного на всей совокупности элементов. С помощью стандартной ошибки задается доверительный интервал для среднего. 95%-ный доверительный интервал, равный $\bar{x} \pm 2m$, обозначает диапазон, в который с вероятностью $p=0,95$ (при достаточно большом числе наблюдений $n>30$) попадает среднее генеральной совокупности.

Коэффициент вариации используют для сравнения рассеивания двух и более признаков, имеющих различные единицы измерения. Коэффициент вариации представляет собой относительную меру рассеивания, выраженную в процентах. Он вычисляется по формуле: $V = \frac{s}{\bar{x}} \times 100\%$, где V – искомый показатель, s – среднее квадратичное отклонение, $\bar{x}$ – средняя величина.

Задание и методические рекомендации к выполнению работы

Имеются следующие данные о времени, которое студенты затрачивают на выполнение контрольного задания (минут): 15, 16, 17, 18, 19, 19, 19, 19, 20, 20, 21, 21, 21, 21, 21, 22, 22, 22, 22, 23, 23, 24, 24,

25, 25, 26, 27, 29, 30, 45. Для анализа результатов выполнения задания требуется рассчитать статистические показатели: моду, медиану, среднее значение выборки, размах вариации, интерквартильный размах, дисперсию, среднее квадратичное отклонение, стандартную ошибку, коэффициент вариации.

Выявить наличие среди исходных данных резко выделяющихся значений признаков (выбросов данных) с целью исключения из выборки аномальных единиц наблюдения. Построить гистограмму.

Выполнение задания

Ввести исходные данные в диапазон ячеек A1:A30 и заполнить таблицу формулами для расчета основных статистик. Для ввода формул можно воспользоваться мастером функций. Например, для ввода формулы, подсчитывающей количество чисел в диапазоне A1:A30, нужно установить курсор в ячейку, в которой будет находиться формула, нажать знак «=» и щелкнуть по кнопке **fx** «Вставить функцию» в строке формул. На экране появится диалоговое окно **Мастера функций**, в котором нужно выбрать в категории *Статистические* функцию СЧЁТ(). На следующем шаге нужно указать аргументы функции, в данном случае в поле *Значение1* нужно ввести диапазон ячеек A1:A30, либо выделить данный диапазон с помощью мыши и нажать ОК. Ввод других функций производится аналогично с учетом категории, к которой относится та или иная функция. В результате расчетов получаются следующие значения показателей:

Количество элементов	30	Минимальное значение	15
Мода	21	Максимальное значение	45
Медиана	21,5	Размах вариации	30
Среднее значение	22,53	1 квартиль	19,25
Дисперсия	30,26	3 квартиль	24
Стандартное отклонение	5,50	Интерквартильный размах	4,75
Стандартная ошибка	1,00	Асимметрия	2,44
Коэффициент вариации	24,41	Эксцесс	9,05

Далее нужно определить, имеются ли среди наблюдений резко выделяющиеся значения – выбросы, и при необходимости исключить их из выборки. Умеренные выбросы удалены ниже первого квартиля или выше третьего от 1,5 IQR, но не более 3 IQR, где IQR – интерквартильный размах. Экстремальные выбросы удалены ниже первого квартиля или выше третьего более чем на 3 IQR.

Расчетные значения границ для выбросов:

Умеренные выбросы		Экстремальные выбросы	
Нижняя граница	12,125	Нижняя граница	5
Верхняя граница	31,125	Верхняя граница	38,25

Нормированное отклонение $t = \frac{x - \bar{x}}{s}$ также позволяет решить, относится ли данное наблюдение к интересующей нас совокупности. Ответ будет положительным, когда $t < 3$.

Таким образом, в выборке имеется один экстремальный выброс, равный 45 ($45 > 38,25$; $t = 4,08$), который нужно исключить и повторить расчет статистических показателей для оставшихся данных. Повторный расчет дает следующие значения показателей:

Количество элементов	29	Минимальное значение	15
Мода	21	Максимальное значение	30
Медиана	21	Размах вариации	15
Среднее значение	21,76	1 квартиль	19
Дисперсия	12,69	3 квартиль	24
Стандартное отклонение	3,56	Интерквартильный размах	5
Стандартная ошибка	0,66	Асимметрия	0,42
Коэффициент вариации	16,37	Эксцесс	0,19

Те же самые характеристики можно получить, используя имеющийся в пакете Excel набор мощных инструментов для углубленного анализа данных. Для использования пакета анализа нужно выбрать во вкладке **Данные** команду **Анализ данных**. Для определения основных выборочных характеристик используется процедура **Описательная статистика**. В появившемся диалоговом окне необходимо выполнить следующие действия:

- указать входной диапазон, то есть ввести ссылку на ячейки, содержащие анализируемые данные;
- установить переключатель в положение *Выходной интервал*;
- указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа (указать адрес левой верхней ячейки выходного диапазона);
- в разделе *Группирование* установить переключатель *по столбцам*;
- установить флажок в поле *Итоговая статистика*;
- нажать кнопку ОК.

В результате анализа в указанном выходном диапазоне будут получены показатели описательной статистики.

При обработке данных часто возникает необходимость в их группировке. Группировка позволяет представить первичные данные в компактном виде, выявить закономерности варьирования изучаемого признака. Количество интервалов (классов) можно приблизительно наметить следующим образом: при количестве наблюдений 25-40 – 5-6 классов, при количестве наблюдений 40-60 – 6-8 интервалов, 60-100 – 7-10, 100-200 наблюдений – 8-12, более 200 наблюдений –

10-15 интервалов. Более точно определить оптимальное количество интервалов (k), которое зависит от объема выборки, позволяет формула Стерджесса: $k = 1 + 3,322 \lg n$, где n – число элементов выборки.

В задании $k = 1 + 3,322 \lg 30 = 1 + 3,322 * 1,477 = 5,9$. Так как число интервалов не может быть дробным, то его необходимо округлить до ближайшего целого числа, то есть до 6. Ширина интервала определяется по формуле: $h = R/k$, где R – размах вариации: $h = (30 - 15)/6 = 2,5$.

Для построения гистограммы можно использовать Пакет анализа. В диалоговом окне **Гистограмма** необходимо выполнить следующие действия:

- задать *Входной интервал*, то есть ввести ссылку на диапазон анализируемых данных;
- задать интервал карманов – набор граничных значений, определяющих интервалы, которые должны быть введены в возрастающем порядке. В Excel вычисляется число попаданий данных между текущим началом отрезка и соседним большим по порядку, если такой есть. При этом включаются значения на нижней границе отрезка и не включаются значения на верхней границе. Если диапазон карманов не был введен, то набор отрезков, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически;
- указать выходной интервал, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа (указать адрес левой верхней ячейки выходного диапазона);
- установить флажок в поле *Вывод графика*.

В результате анализа в выходном диапазоне начиная с указанной ячейки будут получены соответствующие результаты.

Оценив визуально гистограмму можно сделать предположение о виде распределения экспериментальной совокупности данных.

Задания для индивидуального выполнения

В столбцах таблицы 1 представлены данные для статистического анализа. Номер варианта указан в первой строке таблицы.

Таблица 1

Время выполнения контрольного задания (в минутах)

№	Вариант индивидуального задания														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1.	14	23	16	10	9	27	28	14	12	19	13	20	2	10	23
2.	6	6	20	14	13	9	5	1	18	28	23	15	14	23	7
3.	19	23	18	16	15	0	6	11	2	21	21	26	16	16	8
4.	21	18	15	14	18	18	10	3	7	10	16	35	2	22	1
5.	23	11	15	26	20	25	7	7	18	21	42	1	10	17	6
6.	13	2	15	11	15	25	21	19	7	7	19	27	16	21	14
7.	17	15	21	4	11	12	13	12	21	21	13	7	9	8	18

№	Вариант индивидуального задания														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
8.	27	14	4	17	17	13	10	5	16	17	15	10	22	13	11
9.	21	2	13	12	12	17	33	10	8	11	4	21	12	19	20
10.	12	1	16	10	21	11	38	12	4	14	18	11	13	15	12
11.	13	13	13	38	23	18	16	19	12	26	14	20	15	13	19
12.	18	12	18	5	8	12	15	28	14	6	9	14	9	6	12
13.	14	19	16	14	45	19	31	27	8	7	18	20	18	22	26
14.	15	13	3	10	13	16	24	18	13	22	16	12	26	14	13
15.	13	15	4	14	28	38	9	1	21	12	25	9	23	38	7
16.	26	12	12	15	15	17	18	8	8	14	13	16	16	13	18
17.	16	20	17	15	19	15	16	16	19	22	15	15	11	7	17
18.	39	25	40	22	25	17	18	45	20	11	17	22	24	9	56
19.	8	19	34	52	17	14	65	25	20	55	12	16	14	52	29
20.	25	10	8	9	13	19	17	19	19	20	26	15	6	13	23
21.	8	16	11	16	20	37	8	5	19	13	21	13	24	17	8
22.	10	55	13	18	24	24	2	12	13	9	22	25	15	5	51
23.	17	19	13	5	1	16	22	12	56	16	48	12	14	17	10
24.	46	11	19	13	37	14	18	52	10	13	19	48	23	6	20
25.	11	17	12	15	4	17	12	8	3	2	6	19	58	3	8
26.	14	2	22	19	21	18	7	22	41	9	22	16	18	22	15
27.	16	6	10	22	13	18	25	12	20	26	8	1	36	20	8
28.	16	4	21	23	14	13	20	12	24	21	12	12	27	20	19
29.	21	5	11	24	5	20	24	11	3	11	17	13	22	16	14
30.	15	17	19	26	17	12	18	14	14	9	9	23	9	24	2

Лабораторная работа 2

Тема: «Анализ вида распределения экспериментальных данных»

Цель: изучить методы проверки соответствия экспериментальных данных нормальному закону распределения с помощью пакета «Microsoft Excel».

Основные теоретические сведения

Применяемые статистические методы, как правило, ориентированы на обработку данных, полученных из генеральной совокупности с определенным законом распределения. Наиболее часто требуется установить, отобрана ли выборка из генеральной совокупности, в которой изучаемый признак имеет нормальное распределение. Нормальный закон распределения занимает среди других законов распределения особое положение и является наиболее часто встречающимся на практике. Главная его особенность состоит в том, что он является предельным законом, к которому приближаются другие законы распределения при часто встречающихся типичных условиях.

Для предварительной оценки близости эмпирического закона распределения к нормальному используют графический метод, показатели центра распределения и вариации, основные свойства и выборочные параметры формы распределения, среднее абсолютное отклонение. Совокупность этих показателей позволяет дать качественную оценку близости эмпирических данных к нормальной форме распределения.

Графический метод позволяет визуально оценить форму распределения по гистограмме. Анализируются количество вершин в гистограмме, ее асимметричность и выраженность «хвостов». Если на гистограмме четко прослеживаются два-три «горба» частот вариантов, это говорит о том, что выборка неоднородна и значения признака концентрируются сразу в нескольких интервалах, что не соответствует нормальному закону распределения. Если гистограмма имеет одновершинную форму, есть основания предполагать, что выборочная совокупность может иметь характер распределения, близкий к нормальному. Нормальное распределение является симметричным, среднее $\bar{x}$, мода Mo и медиана Me совпадают.

Коэффициент асимметрии As позволяет судить о симметричности выборочных данных. Если асимметрия правосторонняя ($As > 0$) то правая часть эмпирической кривой оказывается длиннее левой, т.е. имеет место неравенство $\bar{x} > Me > Mo$, что означает преимущественное появление в распределении более высоких значений признака. Если асимметрия левосторонняя ($As < 0$), то левая часть эмпирической кривой оказывается длиннее правой и выполняется неравенство $\bar{x} < Me < Mo$, означающее, что в распределении чаще встречаются более низкие значения признака. Чем больше величина $|As|$, тем более асимметрично распределение. Оценочная шкала асимметрии: если значение асимметрии $As \leq 0,25$, то имеется слабая асимметрия; $0,25 < As \leq 0,5$ – умеренная асимметрия; $As > 0,5$ – крайне асимметричное распределение.

Показатель эксцесса Ek характеризует крутизну кривой распределения – ее заостренность или пологость по сравнению с нормальной кривой. Как правило, коэффициент эксцесса вычисляется только для симметричных или близких к ним распределений. Если $Ek > 0$, то форма кривой является более островершинной, чем нормальная. Это говорит о скоплении значений признака в центральной зоне ряда распределения, то есть о преимущественном появлении в данных значений, близких к средней величине. Если $Ek < 0$, то форма кривой более пологая по сравнению с нормальной. Это означает, что значения признака не концентрируются в центральной части ряда, а рассеяны по всему диапазону от минимального до максимального значения. Если абсолютное значение эксцесса $|Ek| \leq 0,2$, то эксцесс практически отсут-

ствуется; $0,2 < |E| \leq 0,3$ – слабый эксцесс; $0,3 < |Ek| \leq 0,6$ – умеренный эксцесс; $0,6 < |Ek| \leq 1,0$ – сильный эксцесс; $|Ek| > 1$ – очень сильный эксцесс. Чем больше абсолютная величина $|Ek|$, тем существеннее распределение отличается от нормального.

Принято говорить, что предположение о нормальности распределения не противоречит имеющимся данным, если асимметрия лежит в диапазоне от -0,2 до 0,2, а эксцесс – от -1 до 1.

Для нормального закона распределения характерно следующее свойство: приблизительно 68% значений вариант находятся в интервале $\bar{x} \pm s$; 95% – в интервале $\bar{x} \pm 2s$; 99% – в интервале $\bar{x} \pm 3s$, где $\bar{x}$ – среднее значение, а s – стандартное отклонение выборки. Если полученная структура рассеяния вариант по трем диапазонам незначительно расходится с данным свойством, можно предположить, что распределение экспериментальных данных близко к нормальному.

Использование среднего абсолютного отклонения позволяет проверить нормальность распределения для выборки небольшого объема. Среднее абсолютное отклонение находят следующим образом:

$$\Delta_{abs} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n},$$
 где x_i – значения вариант, $\bar{x}$ – среднее значение, n – объем выборки. Затем проверяют выполнение следующих условий:

- количество положительных и отрицательных отклонений от среднего приблизительно равны;
- половина (или чуть больше) отклонений от среднего по абсолютной величине меньше среднего абсолютного отклонения;
- ни одно из отклонений не превышает среднее абсолютное отклонение больше чем в 3-4 раза;
- выполняется следующее условие: $\left| \frac{\Delta_{abs}}{s} - 0,7979 \right| < \frac{0,4}{\sqrt{n}}.$

Если все эти условия выполняются, то можно считать, что гипотеза о нормальном распределении не противоречит имеющимся данным.

Следует отметить, что на основании выборочных параметров распределения сколько-нибудь уверенно о нормальности закона распределения можно говорить, если имеется не менее 100 наблюдений. Поэтому целесообразно использовать несколько методов оценки, которые дополняют друг друга, и на основе полученных оценок сделать вывод о степени близости распределения экспериментальных данных к нормальному распределению.

Наиболее убедительные результаты дает проверка статистических гипотез о виде распределения (критериев согласия). Среди кри-

то есть «карманы» для построения гистограммы. Гистограмму строится с помощью процедуры пакета анализа **Гистограмма**.

В результате анализа в выходном диапазоне будут получены следующие результаты (рисунок 1):

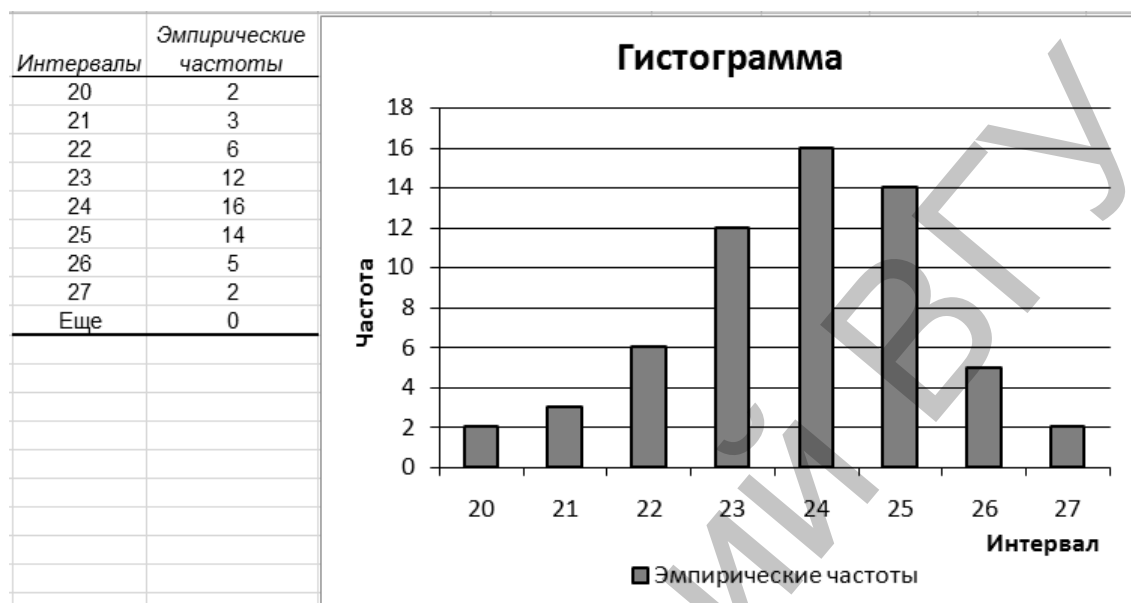


Рисунок 1 – Гистограмма веса детей в группе

Гистограмма имеет одну вершину и является достаточно симметричной. Для расчета выборочных параметров формы распределения эксцесса и асимметрии нужно ввести в ячейки D12 и D13 формулы =ЭКССЕСС(A2:A61) и =СКОС(A2:A61). Расчетное значение эксцесса равно 0,043, асимметрии – -0,368. Значения асимметрии и эксцесса показывают, что предположение о нормальности распределения не противоречит имеющимся данным.

Далее для проведения вычислений понадобятся такие данные как количество измерений, среднее значение и стандартное отклонение. Заполнить ячейки D14, D15, D16 формулами =СУММ(D2:D10), =СРЗНАЧ(A2:A61), =СТАНДОТКЛОН(A2:A61). Результаты: количество измерений – 60; среднее значение – 23,82; стандартное отклонение – 1,568.

Для подсчета количества вариантов, находящихся в диапазонах $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$, можно воспользоваться функцией СЧЁТЕСЛИМН(), которая используется для подсчета количества ячеек на основании нескольких критериев. Например, если в диапазоне A2:A61 находится исходная выборка, а в ячейках C30 и D30 находятся значения $\bar{x} - 3s$ и $\bar{x} + 3s$ соответственно, то для нахождения количества элементов выборки, попадающих в диапазон $\bar{x} \pm 3s$ можно использовать функцию

=СЧЁТЕСЛИМН(\$A\$2:\$A\$61;">"&C30;\$A\$2:\$A\$61;"<"&D30).

Количество попаданий в остальные диапазоны находится аналогично.

В задании 42 элемента (70%) выборки находятся в интервале $\bar{x} \pm s$, 56 элементов (93%) – в интервале $\bar{x} \pm 2s$ и 100% данных попадают в интервал $\bar{x} \pm 3s$, что говорит о близости эмпирического распределения к нормальному.

Для расчета среднего абсолютного отклонения нужно добавить в рабочую таблицу два столбца, в которых рассчитать отклонение и абсолютное значение отклонения каждой варианты от среднего значения. Далее находят значение среднего абсолютного отклонения и проверяют условия нормальности. В задании среднее абсолютное отклонение равно 1,226.

Для того, чтобы найти количество положительных и отрицательных отклонений от среднего, можно воспользоваться функцией СЧЁТЕСЛИ(диапазон, условие), которая подсчитывает количество ячеек, отвечающих определенному условию. Например, если отклонения были вычислены в столбце G, то количество положительных отклонений рассчитывается по формуле =СЧЁТЕСЛИ(G2:G61;">0") и равно 37. Аналогичным образом определяют количество отрицательных отклонений, равное 23 и количество отклонений от среднего, меньших по абсолютной величине среднего абсолютного отклонения. В примере меньших отклонений 42. Максимальное отклонение по модулю равно 3,82, что также укладывается в заданные требования. Проверка последнего условия дает следующие результаты:

$$\left| \frac{1,226}{1,568} - 0,7979 \right| = 0,0157; \quad \frac{0,4}{\sqrt{60}} = 0,0516; \quad 0,0157 < 0,0516,$$

следовательно, условие выполняется. Таким образом, предварительная оценка вида распределения по совокупности проведенных тестов позволяет сделать вывод о том, что экспериментальные данные не противоречат нормальному закону распределения.

Использование критерия согласия хи-квадрат дает возможность с заданной вероятностью утверждать, что наблюдаемые значения принадлежат к генеральной совокупности с нормальным законом распределения. Для применения критерия хи-квадрат нужно получить теоретические частоты для нормального распределения с теми же параметрами, как у экспериментальных данных. В таблицу вводятся:

	Е	Ф
1	Теоретические частоты	Теоретические частоты
2	=НОРМРАСП(C2;\$D\$15;\$D\$16;0)	=E2*\$D\$14

Далее формулы из ячеек E2:F2 необходимо размножить на диапазон E3:F9. В результате в столбце F будут получены теоретические частоты.

Для визуализации данных целесообразно добавить полученные теоретические частоты к гистограмме: щелкнуть правой кнопкой мы-

ши по гистограмме, выбрать пункт контекстного меню *Исходные данные*. В появившемся окне перейти на вкладку *Вид*, щелчком по кнопке *Добавить* добавить еще один ряд, в поле *Имя* указать ячейку F1, в поле *Значения* указать диапазон ячеек F2:F9. В результате будут получены данные и диаграмма следующего вида (рисунок 2):

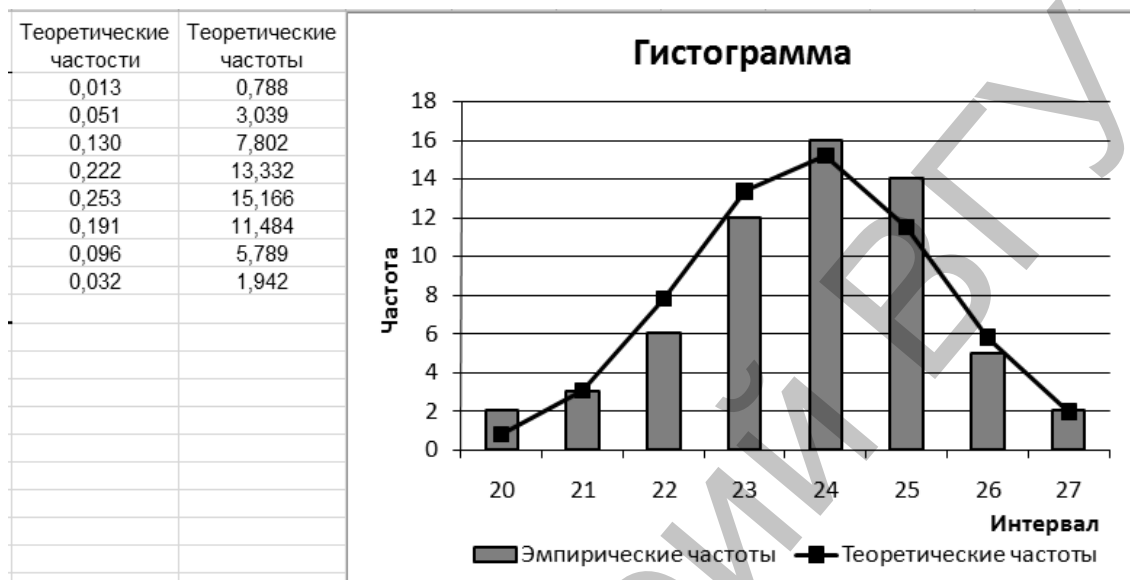


Рисунок 2 – Теоретические и эмпирические частоты

Для окончательного решения о виде распределения нужно ввести в ячейку C17 «Хи2-тест», а в ячейку D17 формулу =ХИ2ТЕСТ(D2:D9;F2:F9), где диапазон D2:D9 задает фактический интервал (интервал данных, которые содержат наблюдения, подлежащие сравнению с ожидаемыми значениями), а диапазон F2:F9 – ожидаемый интервал (интервал данных, который содержит теоретические значения для соответствующих наблюдаемых).

В результате будет вычислена вероятность соответствия экспериментальных данных нормальному закону распределения, равная 0,874. Поскольку полученная вероятность близка к единице, можно говорить, что данные не противоречат нормальному закону распределения.

Задания для индивидуального выполнения

В столбцах таблицы 2 представлены экспериментальные данные. Определить соответствие данных нормальному закону распределения. Номер варианта указан в первой строке таблицы.

Таблица 2

Экспериментальные данные

№	Вариант индивидуального задания														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	49	25	70	50	96	55	35	78	31	43	84	28	46	39	75
2	61	36	86	42	88	57	28	68	36	51	94	31	44	44	66
3	40	30	70	50	93	53	38	66	34	48	95	32	35	39	70

№	Вариант индивидуального задания														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4	56	32	87	50	91	48	34	54	34	53	99	30	47	39	66
5	60	34	76	43	80	45	32	72	38	45	89	35	55	40	55
6	50	34	76	59	73	53	28	73	34	53	89	36	43	31	56
7	54	34	99	57	83	58	39	71	39	55	95	33	57	38	55
8	48	32	82	48	93	49	32	70	32	46	96	30	43	39	57
9	56	27	96	46	84	45	34	89	40	44	95	35	54	41	52
10	52	24	80	41	90	45	37	71	36	50	83	41	55	41	68
11	50	24	85	44	82	43	37	82	26	47	88	34	52	33	57
12	53	24	92	47	98	60	34	69	35	52	99	30	45	39	59
13	41	33	76	48	95	59	37	65	41	50	94	38	47	38	74
14	55	32	91	58	91	53	40	61	45	53	90	36	56	46	51
15	53	35	89	59	85	45	33	71	30	42	92	32	60	43	62
16	41	30	72	47	89	55	31	67	36	62	83	36	50	50	55
17	45	22	89	50	83	45	34	72	39	50	86	30	46	31	62
18	51	34	78	45	99	55	34	65	27	52	92	34	49	43	55
19	48	40	81	53	93	49	29	66	31	44	81	34	47	45	61
20	45	32	77	53	79	48	30	75	36	57	99	36	47	42	55
21	48	22	94	45	83	51	31	60	31	45	84	36	52	42	58
22	42	35	73	40	99	49	35	69	36	61	99	35	52	35	57
23	55	32	86	51	74	53	36	65	33	43	76	34	48	39	51
24	48	29	62	43	91	51	39	72	42	49	97	34	56	35	62
25	51	27	84	63	87	41	42	72	37	48	90	34	42	40	55
26	53	39	71	40	93	42	31	68	35	50	79	30	55	39	51
27	46	29	76	53	96	50	39	76	42	58	86	32	57	39	64
28	47	33	78	54	89	64	37	66	36	57	73	39	44	44	62
29	51	31	78	56	88	44	33	75	40	41	90	38	51	35	56
30	57	23	85	55	85	45	37	59	38	44	99	35	54	35	62
31	56	35	79	56	95	52	34	74	36	53	98	38	52	42	58
32	48	29	71	50	90	58	32	82	31	58	80	35	41	41	61
33	42	32	83	49	88	46	37	75	38	50	94	36	52	37	62
34	56	26	70	52	97	47	36	74	35	51	97	31	48	31	63
35	53	27	71	53	98	51	35	67	39	57	76	30	48	43	49
36	47	31	79	56	99	62	34	67	33	60	95	35	48	36	59
37	44	31	56	57	95	49	28	76	28	49	95	31	50	40	75
38	47	38	81	44	89	56	32	70	36	47	87	28	50	41	62
39	52	34	91	36	80	56	38	77	31	54	83	35	51	40	61
40	53	24	74	49	91	57	39	66	36	54	99	38	52	39	54
41	49	33	86	60	91	44	32	73	27	46	76	28	59	43	70
42	48	31	95	57	86	52	34	68	35	53	90	41	55	37	60
43	51	29	80	48	87	47	40	71	37	48	87	30	53	45	63
44	44	28	81	52	86	47	34	76	31	52	91	33	50	40	58
45	61	35	79	45	96	58	32	71	35	56	93	38	47	41	54
46	57	23	87	53	92	51	35	67	33	52	96	39	54	49	72

№	Вариант индивидуального задания														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
47	45	29	82	47	95	45	43	71	36	53	82	39	54	42	59
48	46	31	55	48	96	50	34	58	38	53	88	41	40	45	62
49	51	26	91	48	84	52	28	83	31	55	96	29	47	37	72
50	45	29	68	44	99	51	35	67	38	49	87	37	46	43	61
51	54	26	87	41	97	38	32	65	40	52	91	30	49	43	48
52	51	29	78	44	91	44	32	69	36	51	99	39	57	44	64
53	42	29	83	51	88	44	39	78	37	49	96	36	53	37	62
54	53	32	63	52	75	37	35	75	35	60	83	29	48	36	72
55	47	31	81	48	92	52	35	73	36	45	90	32	52	46	52
56	43	26	92	43	83	52	40	63	40	46	98	43	53	43	57
57	42	38	78	44	95	46	42	74	36	50	80	33	47	44	65
58	42	25	76	47	88	61	26	76	37	48	91	35	48	36	62
59	49	29	81	51	96	46	34	65	37	52	81	32	45	43	66
60	47	32	91	56	93	53	30	64	34	43	90	32	51	45	58

Лабораторная работа 3

Тема: «Статистическое изучение связи. Корреляционный анализ»

Цель: изучить методы оценки величины коэффициента корреляции и проверки его значимости для выявления статистической связи между признаками.

Основные теоретические сведения

Корреляционный анализ служит для выявления взаимосвязей между рядами наблюдений. В качестве меры связи используется коэффициент корреляции R – параметр, характеризующий степень взаимосвязи между выборками. Наибольшее распространение получил коэффициент линейной корреляции (Пирсона), предполагающий нормальный закон распределения наблюдений. Пределы изменения коэффициента корреляции от -1 (обратная зависимость) до 1 (прямая зависимость). При значении равно нулю зависимость между двумя выборками отсутствует. Для оценки степени взаимосвязи существуют следующие эмпирические правила. При $|R| > 0,95$ (без учета знака), то принято считать, что между выборками существует практически полная линейная корреляционная связь; $0,7 < |R| < 0,95$ – сильная корреляционная связь; $0,3 < |R| < 0,7$ – связь средней силы; $0 < |R| < 0,3$ – слабая корреляционная связь.

В ходе корреляционного анализа решаются следующие задачи: установление направления (прямая или обратная) корреляционная связь и оценка тесноты (силы, плотности) связи. Расчет коэффициента корреляции (Пирсона) производится по следующей формуле:

$$R = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}, \text{ где } x \text{ и } y - \text{ряды наблюдений}$$

ния, между которыми нужно установить связь, а n – число наблюдений в каждом ряду.

Коэффициент корреляции Пирсона применяется для данных, распределенных нормально. Если распределение не является нормальным или закон распределения неизвестен, то коэффициент связи для двух переменных вычисляется как коэффициент линейной корреляции для их рангов. Это коэффициент ранговой корреляции Спирмена. Для получения рангов можно использовать процедуру **Ранг и Персентиль** Пакета Анализа.

Задания и методические рекомендации к выполнению работы

Задание 1. Требуется определить, имеется ли связь между количеством ошибок, допускаемых при работе с компьютерным текстом, от длительности работы на компьютере.

Выполнение задания

Для выполнения задания необходимо ввести в таблицу исходные данные: в столбец А – число ошибок (x), В – время работы (y), в столбцах С, D и Е выполнить вычисления промежуточных величин x^2 , y^2 , xy . Исходные данные и формулы представлены на рисунке 3:

	А	В	С	Д	Е
1	х	у	х ²	у ²	х*у
2	6	1	=А2^2	=В2^2	=А2*В2
3	8	1	=А3^2	=В3^2	=А3*В3
4	7	2	=А4^2	=В4^2	=А4*В4
5	6	3	=А5^2	=В5^2	=А5*В5
6	6	3	=А6^2	=В6^2	=А6*В6
7	5	4	=А7^2	=В7^2	=А7*В7
8	8	5	=А8^2	=В8^2	=А8*В8
9	6	5	=А9^2	=В9^2	=А9*В9
10	9	6	=А10^2	=В10^2	=А10*В10
11	9	6	=А11^2	=В11^2	=А11*В11
12	=СУММ(А2:А11)	=СУММ(В2:В11)	=СУММ(С2:С11)	=СУММ(Д2:Д11)	=СУММ(Е2:Е11)

Рисунок 3 – Заполнение таблицы данными и формулами для задания 1

Для расчета коэффициента корреляции необходимо ввести в одну из ячеек таблицы, например, в ячейку В14 следующую формулу:

=(Е12-А12*В12/10)/КОРЕНЬ((С12-А12^2/10)*(Д12-В12^2/10)).

В результате расчетов будет получено значение коэффициента корреляции, равное 0,4141. Как видно из полученного результата, между числом ошибок и длительностью работы есть прямая корреляционная связь средней силы. В данном случае, по-видимому, можно говорить о причинно-следственной связи между этими факторами. Од-

нако утверждать, что утомляемость связана только с длительностью работы нельзя. С помощью коэффициента детерминации R^2 можно определить долю влияния анализируемого признака на результативный признак. Если учесть, что длительность рабочего дня является не единственным фактором, способствующим развитию усталости, то R^2 показывает долю тех изменений, которые обусловлены анализируемым фактором. В данном случае $R^2 = 0,4141^2 \approx 0,17$, то есть доля влияния продолжительности дня на усталость – 17 %. Следовательно, на долю других факторов развития усталости приходится 83 % влияния.

Расчет коэффициента корреляции можно также выполнить с помощью статистической функции КОРРЕЛ(), в данном задании в ячейку вводится следующая формула: =КОРРЕЛ(A2:A11;B2:B11).

Для оценки достоверности корреляционной связи вычисляется ошибка коэффициента корреляции m_R по формуле: $m_R = \frac{1 - R^2}{\sqrt{n - 2}}$, где n

– число наблюдений. Для того, чтобы результат можно было признать статистически достоверным, необходимо, чтобы абсолютное значение коэффициента корреляции R превышало ошибку m_R не менее чем в два раза. В данном примере $m_R = 0,29293$ (провести вычисления самостоятельно). Если рядов наблюдений больше двух, получаемые коэффициенты сводят в корреляционную матрицу. В Excel для получения корреляционной матрицы используется процедура Пакета Анализа **Корреляция**.

Задание 2. Имеются результаты ежемесячных наблюдений за состоянием погоды и посещаемостью кинотеатра, парка и бассейна. Необходимо определить, имеется ли связь между состоянием погоды и посещаемостью кинотеатра, парка и бассейна. Данные представлены в таблице 3:

Таблица 3

Число ясных дней	Посетителей кинотеатров	Посетителей парков	Посетителей бассейнов
8	495	132	164
14	503	348	182
20	380	643	153
25	305	865	203
20	348	743	169
15	465	541	199

Выполнение задания

Ввести данные в таблицу в диапазон ячеек A1:D7 (в первой строке расположить заголовки столбцов данных). В Пакете Анализа для проведения корреляционного анализа используется процедура

Корреляция. В появившемся диалоговом окне необходимо выполнить следующие действия:

- задать *Входной интервал*, то есть ввести ссылку на диапазон анализируемых данных;
- указать *выходной диапазон*, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа (указать адрес левой верхней ячейки выходного диапазона).

После нажатия кнопки ОК будут получены следующие результаты (рисунок 4):

F	G	H	I	J
	<i>ясн.дни</i>	<i>кинотеатры</i>	<i>парки</i>	<i>бассейны</i>
ясн.дни	1			
кинотеатры	-0,9218543	1		
парки	0,9745756	-0,919375244	1	
бассейны	0,2877364	-0,120007005	0,3244702	1

Рисунок 4 – Корреляционная матрица

Интерпретация результатов: из таблицы видно, что корреляция между состоянием погоды и посещаемостью кинотеатра равна -0,922 (сильная отрицательная связь); между состоянием погоды и посещаемостью парка 0,975 (практически полная положительная корреляционная связь); между состоянием погоды и посещаемостью бассейна 0,288 (слабая положительная корреляционная связь). Также можно определить, что между посещаемостью кинотеатра и парка имеется сильная отрицательная связь ($R = -0,919$), а связь между посещаемостью кинотеатра и бассейна практически отсутствует ($R = -0,12$).

Задания для индивидуального выполнения

В таблице 4 представлены показатели, полученные в результате обследования группы учащихся. Определить, имеется ли связь между заданными показателями.

Решить задачу с помощью расчета коэффициента корреляции Пирсона и с использованием процедуры Пакета Анализа **Корреляция**.

Между первыми двумя показателями найти коэффициент корреляции Спирмена.

Вариант 1. Рост, вес, возраст, средний балл сессии.

Вариант 2. Возраст, время решения контрольной работы, результаты тестирования, время, затрачиваемое на спорт.

Вариант 3. Время решения контрольной, средний балл сессии, количество пропусков, количество друзей.

Вариант 4. Результаты тестирования, IQ, время, затрачиваемое на спорт, средний балл сессии.

Вариант 5. Время решения контрольной работы, время, затрачиваемое на спорт, время, проведенное в социальных сетях, количество друзей.

Вариант 6. Возраст, IQ, средний балл сессии, время решения контрольной работы.

Вариант 7. Возраст, время, затрачиваемое на спорт, время, проведенное в социальных сетях, количество друзей,

Вариант 8. Рост, результаты тестирования, количество пропусков, время, проведенное в социальных сетях.

Вариант 9. Возраст, результаты тестирования, средний балл сессии, количество пропусков.

Вариант 10. Возраст, время решения контрольной работы, IQ, количество друзей.

Вариант 11. Рост, вес, средний балл сессии, время, затрачиваемое на спорт.

Вариант 12. Рост, средний балл сессии, результаты тестирования, время, проведенное в социальных сетях.

Вариант 13. Возраст, количество пропусков, IQ, время, проведенное в социальных сетях.

Вариант 14. Время решения контрольной работы, результаты тестирования, количество пропусков, время, затрачиваемое на спорт.

Вариант 15. Возраст, время решения контрольной работы, средний балл сессии, IQ.

Таблица 4

№	Рост, см	Вес, кг	Возраст, лет	Время решения контрольной работы, часов	Результаты тестирования, баллов	Средний балл сессии	Количество пропусков в месяц	IQ (тест Айзенка)	Время, затрачиваемое на спорт, часов в месяц	Время, проведенное в соц.сетях, часов в день	Количество друзей, человек
1	159	49	17	2,3	225	7,4	3	104	24	2,1	9
2	160	51	18	2,2	220	6,7	11	110	32	1,4	15
3	161	49	20	2,1	211	5,2	4	105	24	1,8	10
4	162	54	21	2,4	300	4,8	0	148	24	1,7	8
5	162	53	22	3,2	322	8,6	1	160	60	1,5	21
6	164	53	22	2,8	277	8,3	0	125	16	1,4	3
7	166	53	19	3	300	8,6	0	143	90	1,1	24
8	169	52	21	2,3	225	7	3	111	24	1,4	10
9	170	54	20	3,3	303	9,5	0	134	48	1,2	18

№	Рост, см	Вес, кг	Возраст, лет	Время решения контрольной работы, часов	Результаты тестирования, баллов	Средний балл сессии	Количество пропусков в месяц	IQ (тест Айзенка)	Время, затрачиваемое на спорт, часов в месяц	Время, проведенное в соц.сетях, часов в день	Количество друзей, человек
10	170	52	20	2,3	231	5,3	0	101	24	1,8	8
11	171	55	19	2,7	272	7,5	3	117	16	1,7	6
12	171	53	20	3,5	353	8,7	0	164	48	1,8	19
13	172	55	18	2	199	6,2	2	94	32	2,1	13
14	174	58	22	2,2	221	5,5	2	99	48	1,7	23
15	176	57	21	2	202	4,9	3	86	64	2,1	17
16	176	54	21	2,1	207	5,2	7	103	48	1,8	21
17	178	57	21	2,2	218	5	0	105	90	1,8	22
18	181	57	20	2	203	5,9	0	85	48	2,1	20
19	183	61	21	2	202	5,1	2	100	32	2,1	16
20	192	62	20	2	198	5,6	3	88	60	2	4

Лабораторная работа 4

Тема: «Регрессионный анализ»

Цель: изучить методы построения регрессионных моделей связи изучаемых признаков и оценки значимости регрессионного уравнения.

Основные теоретические сведения

В регрессионном анализе рассматривается связь между одной переменной, называемой зависимой переменной, и несколькими другими, называемыми независимыми переменными. Эта связь представляется с помощью математической модели, т.е. уравнения, которое связывает зависимую переменную с независимыми с помощью функции регрессии.

В линейном регрессионном анализе связь между случайными величинами предполагается линейной. В этом случае в линейной регрессионной модели имеются две переменные x и y . Требуется по n парам наблюдений $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ построить прямую линию, называемую линией регрессии, которая наилучшим образом приближает наблюдаемые значения. Уравнение этой линии $y = ax + b$. С помощью уравнения регрессии можно предсказать ожидаемое значение за-

висимой величины y , соответствующее заданному значению независимой переменной x .

Математически решение уравнения линейной регрессии сводится к вычислению параметров a и b таким образом, чтобы точки исходных данных как можно ближе лежали к линии регрессии. Для этого вычисляют параметры по формулам, которые обеспечивают наименьший квадрат отклонений этих точек от линии регрессии (метод наименьших квадратов):

$$a = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}; \quad b = \frac{\sum x^2 \sum y - \sum x \sum xy}{n \sum x^2 - (\sum x)^2}.$$

Задание и методические рекомендации к выполнению работы

Задание 1. Найти уравнение регрессии по данным пяти наблюдений ($n=5$) зависимой и независимой переменных y и x .

Выполнение задания

Для выполнения задания заполнить расчетную таблицу следующим образом (рисунок 5):

	A	B	C	D	E	F	G
1	i	x	y	x^2	x*y	a	=(5*E7-B7*C7)/(5*D7-B7^2)
2	1	2	2,6	=B2^2	=B2*C2	b	=(D7*C7-B7*E7)/(5*D7-B7^2)
3	2	3	2,4	=B3^2	=B3*C3		
4	3	4	1,8	=B4^2	=B4*C4		
5	4	5	1,1	=B5^2	=B5*C5		
6	5	7	1,3	=B6^2	=B6*C6		
7		=СУММ(B2:B6)	=СУММ(C2:C6)	=СУММ(D2:D6)	=СУММ(E2:E6)		

Рисунок 5 – Заполнение расчетной таблицы данными и формулами

В итоге получаются следующие результаты: $a = -0,3$ и $b = 3,1$. Искомое уравнение регрессии $y = -0,3x + 3,1$.

Одним из самых удобных и практичных способов регрессионного анализа в Excel является использование мастера диаграмм. В окне мастера диаграмм необходимо выбрать точечный тип диаграммы и ее вид; на следующем шаге указать диапазон ячеек с исходными данными B1:C6 и отметить, что ряды данных находятся в столбцах. Для завершения построения диаграммы необходимо указать ее название и место размещения (текущий или новый лист).

Для получения линии регрессии на появившейся диаграмме нужно выделить щелчком левой клавиши мыши ряд полученных точек, а затем щелчком правой клавиши мыши вызвать контекстное меню и в списке команд выбрать **Добавить линию тренда**. В окне «Линия тренда» нужно установить тип «Линейная», затем перейти во вкладку **Параметры** и выделить поля «показывать уравнение на диаграмме» и «поместить на диаграмму величину достоверности аппроксимации».

На диаграмме появится окно с автоматически вычисленным уравнением линейной регрессии, а также значение коэффициента детерминации R -квадрат, который является мерой эффективности регрессионной модели. Коэффициент детерминации определяет, с какой степенью точности полученное регрессионное уравнение описывает (аппроксимирует) исходные данные. Максимально возможное значение R -квадрат равно 1. Если R -квадрат больше 0,95, говорят о высокой степени аппроксимации. Если R -квадрат лежит в диапазоне от 0,8 до 0,95, говорят об удовлетворительной аппроксимации. Если R -квадрат меньше 0,6, принято считать, что точность аппроксимации недостаточна и модель требует улучшения. С помощью этого коэффициента можно путем подбора выбрать уравнение регрессии наиболее полно аппроксимирующее ту или иную взаимосвязь. В данном примере полиномиальное уравнение линии тренда лучше приближает исходные данные, нежели линейное.

В случае если исследуется зависимость между одной зависимой переменной и несколькими независимыми, строят уравнение множественной линейной регрессии: $y = a_0 + a_1x_1 + \dots + a_nx_n$. Более глубоко регрессионный анализ позволяет проводить процедура **Регрессия** надстройки Пакет Анализа.

Задание 2.

Имеются следующие данные:

x_1 – количество набранных баллов по результатам тестирования по физике;

x_2 – количество набранных баллов по результатам тестирования по математике;

x_3 – количество набранных баллов по результатам тестирования по русскому языку;

x_4 – средний балл аттестата;

y – средний балл по результатам первой сессии.

Требуется построить регрессионное уравнение вида $y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4$ определив значения коэффициентов a_0, a_1, a_2, a_3, a_4 .

Выполнение задания

Ввести данные в таблицу в диапазон ячеек A1:E22 (рисунки 6).

	A	B	C	D	E
1	X1	X2	X3	X4	Y
2	11	63	51	5,3	6,1
3	10	55	58	6,2	6,6
4	8	57	68	5,5	6,7
5	10	41	53	4,3	5,0
6	23	68	77	7,4	8,1
7	23	73	69	7,1	7,7
8	21	52	52	5,0	5,8
9	15	51	49	5,4	5,8
10	26	71	71	6,2	7,5
11	56	65	56	5,9	6,7
12	56	76	78	7,6	8,2
13	42	70	75	7,0	8,6
14	42	77	74	7,2	8,4
15	50	68	62	6,4	7,3
16	50	80	84	7,9	8,9
17	47	78	80	7,8	9,1
18	72	68	70	7,0	7,9
19	93	81	91	8,2	9,6
20	93	76	85	7,6	9,0
21	92	88	92	8,6	9,8
22	92	82	87	8,4	9,7

Рисунок 6 – Исходные данные

Для проведения регрессионного анализа используется процедура **Регрессия** Пакета Анализа. В диалоговом окне необходимо выполнить следующие действия:

- задать *Входной интервал Y*, то есть ввести ссылку на диапазон анализируемых зависимых данных, содержащих один столбец данных;
- задать *Входной интервал X*, то есть ввести ссылку на диапазон анализируемых независимых данных, содержащих один или несколько столбцов данных;
- указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа (указать адрес левой верхней ячейки выходного диапазона);
- если необходимо визуально проверить отличие экспериментальных точек от предсказанных по регрессионной модели, следует установить флажок в поле *График подбора*;
- нажать кнопку ОК.

Выходной диапазон будет включать в себя результаты дисперсионного анализа, коэффициенты регрессии, стандартную погрешность вычисления y , среднеквадратичные отклонения, число наблюдений, стандартные погрешности для коэффициентов.

Интерпретация результатов. В таблице *Регрессионная статистика* приводится значение коэффициента детерминации R -квадрат, которое определяет, с какой точностью полученное регрессионное уравнение аппроксимирует исходные данные. Поскольку R -квадрат больше 0,95, можно говорить о высокой точности аппроксимации (рисунок 7).

G	H	I	J	K	L
ВЫВОД ИТОГОВ					
<i>Регрессионная статистика</i>					
Множественный R	0,990966096				
R-квадрат	0,982013804				
Нормир. R-квадрат	0,977517255				
Стандартная ошибка	0,210751247				
Наблюдения	21				
<i>Дисперсионный анализ</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>
Регрессия	4	38,80061021	9,700152554	218,3927715	9,69968E-14
Остаток	16	0,710657408	0,044416088		
Итого	20	39,51126762			
	<i>Коэффициенты</i>	<i>Станд. ошибка</i>	<i>t-статистика</i>	<i>P-Значение</i>	<i>Нижние 95%</i>
Y-пересечение	0,150391456	0,358654569	0,419321176	0,680557165	-0,609922261
X1	0,003365378	0,002548106	1,320737178	0,205169185	-0,002036365
X2	0,026276954	0,011677284	2,250262492	0,038853627	0,001522218
X3	0,034952427	0,009618686	3,633804934	0,002233847	0,014561725
X4	0,468305907	0,146075298	3,205921287	0,005508724	0,158640111

Рисунок 7 – Результаты регрессионного анализа

В таблице *Дисперсионный анализ* оценивается общее качество полученной модели. В данном случае, основной результат дисперсионного анализа состоит в том, что уравнение линейной регрессии является значимым, полученная значимость определяется с помощью критерия Фишера – p , который должен быть меньше, чем 0,05 (строка *Регрессия*, столбец *Значимость F*, в данном примере – $9,69968 \cdot 10^{-14}$). Значения коэффициентов регрессии находятся в столбце *Коэффициенты* и соответствуют: Y -пересечение – a_0 ; переменная X_1 – a_1 ; X_2 – a_2 и т.д.

В столбце *P-значение* приводится достоверность отличия соответствующих коэффициентов от нуля. В случаях, когда $P > 0,05$, коэффициент может считаться нулевым. Это означает, что соответствующая независимая переменная практически не влияет на зависимую переменную. В данном примере для переменной x_1 значение $P > 0,05$, следовательно, эту переменную нужно исключить и повторить расчеты для оставшихся переменных x_2, x_3, x_4 . Результаты повторных расчетов представлены на рисунке 8:

G	H	I	J	K	L
ВЫВОД ИТОГОВ					
<i>Регрессионная статистика</i>					
Множественный R	0,989976221				
R-квадрат	0,980052919				
Нормир. R-квадрат	0,976532846				
Стандартная ошибка	0,2153157				
Наблюдения	21				
<i>Дисперсионный анализ</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>
Регрессия	3	38,72313316	12,90771105	278,4183384	1,20412E-14
Остаток	17	0,788134464	0,046360851		
Итого	20	39,51126762			
<i>Коэффициенты стандартная ошибка t-статистика P-Значение Нижние 95%</i>					
Y -пересечение	-0,150536244	0,282979463	-0,531968796	0,601631237	-0,747570717
X_2	0,028584096	0,011795945	2,423213617	0,026834547	0,003696826
X_3	0,036854077	0,009716283	3,793022162	0,001452687	0,016354512
X_4	0,49162071	0,148145234	3,318505061	0,004064399	0,17906159

Рисунок 8 – Результаты регрессионного анализа после удаления незначимой переменной

Окончательное уравнение множественной линейной регрессии имеет вид:

$$y = -0,1505 + 0,02858 x_2 + 0,0369 x_3 + 0,4916 x_4.$$

Имея данное уравнение, можно определить значение зависимой переменной для любых значений независимых. Например, для $x_2=70$; $x_3=70$ и $x_4=7$ предполагаемое значение среднего балла будет:

$$y = -0,1505 + 0,02858 \cdot 70 + 0,0369 \cdot 70 + 0,4916 \cdot 7 = 7,8715.$$

Задания для индивидуального выполнения

Построить линейное регрессионное уравнение по экспериментальным данным. Найти значение зависимой переменной y для заданного значения независимой переменной x . Построить точечный график по заданным экспериментальным данным. Найти уравнение линии тренда, наиболее точно отражающее экспериментальную зависимость (проверить различные типы линии тренда).

Вариант 1.

x	2	4	5	7	9	12	15	18	19	23
y	0,11	0,09	0,12	1,1	0,15	0,18	0,24	0,51	0,64	0,92

$x=10$

Вариант 2.

x	11	13	16	23	28	29	33	35	39	43
y	1	2	8	15	15	18	18	19	20	21

$x=25$

Вариант 3.

x	25	28	29	31	33	34	35	38	39	45
y	101	110	127	128	162	160	177	187	192	192

$x=30$

Вариант 4.

x	1	2	5	6	7	9	11	14	18	19
y	0,11	0,14	0,12	0,14	0,15	0,18	0,29	0,3	0,64	0,92

$x=15$

Вариант 5.

x	10	15	19	22	35	38	45	50	56	60
y	1	9	11	17	18	20	21	19	22	21

$x=40$

Вариант 6.

x	22	28	29	31	33	34	35	38	39	45
y	45	43	45	50	56	61	68	70	85	133

$x=25$

Вариант 7.

x	10	12	15	22	28	30	33	37	39	42
y	1	2	11	17	18	18	21	19	22	21

$x=20$

Вариант 8.

x	10	15	19	22	35	38	45	50	56	60
y	0,21	0,19	0,21	0,18	0,29	0,35	0,35	0,48	0,61	0,9

$x=30$

Вариант 9.

x	45	46	48	49	51	52	55	58	60	65
y	15	43	58	69	68	75	82	99	101	103

$x=50$

Вариант 10.

x	1	4	6	7	9	10	12	14	15	17
y	2,1	6,3	4,3	5,2	5,8	9,3	14,2	15,1	18,6	28,9

$x=5$

Вариант 11.

x	3	4	6	8	10	12	16	18	20	23
y	10	9	13	12	16	17	24	52	65	89

$x=15$

Вариант 12.

x	12	13	17	28	29	30	33	36	38	42
y	1	2	7	14	15	17	18	19	21	23

 $x=25$ **Вариант 13.**

x	21	27	28	32	34	35	36	38	40	43
y	99	108	125	128	160	168	178	187	190	192

 $x=30$ **Вариант 14.**

x	1	3	4	5	7	8	10	14	17	19
y	0,12	0,13	0,12	0,15	0,16	0,18	0,28	0,31	0,65	0,90

 $x=15$ **Вариант 15.**

x	12	16	19	20	33	37	45	53	56	61
y	2	8	12	16	17	20	21	22	22	25

 $x=40$ **Лабораторная работа 5**

Тема: «Интервальное оценивание. Определение доверительных интервалов для среднего и медианы. Сравнение группы с популяцией»

Цель: изучить построение доверительных интервалов для математического ожидания и медианы. Научиться применять построение доверительных интервалов для решения практических задач.

Основные теоретические сведения

При проведении исследований изучаются, как правило, выборочные, а не генеральные совокупности. Возникает вопрос, в какой степени выборка отражает свойства генеральной совокупности, можно ли утверждать, что полученные путем исследования выборки статистические характеристики (средние величины, дисперсия и т.д.) равны характеристикам генеральной совокупности? Практика показывает, что рассчитанные выборочным путем числовые значения параметров являются лишь результатом приближенного статистического оценивания этих параметров генеральной совокупности.

Уровень значимости – максимальное значение вероятности события, при котором событие считается практически невозможным. Как правило, рассматривают уровни значимости $\alpha=0,05$ (обычные требования надежности) и $\alpha=0,01$ (повышенные требования надежности). Величину $P=1-\alpha$ называют доверительной вероятностью или уровнем надежности. Эта величина используется, чтобы судить о статистической значимости принятого решения. Таким образом, статистическая значимость выборочных характеристик представляет собой меру уверенности в их «истинности». Уровень значимости находится в обратной зависимости от надежности результата. Более высокая ста-

статистическая значимость соответствует более низкому уровню доверия к полученному значению.

Доверительный интервал – это вычисленный на основе выборки интервал значений признака, который с заданной доверительной вероятностью содержит неизвестный параметр генеральной совокупности. Если мы увеличиваем количество измерений, то оценка параметра становится более точной и доверительный интервал уменьшается (если в измерении нет систематических ошибок). Доверительный интервал означает не вероятность попадания значения оцениваемого параметра в пределы определенных границ, а то, что если мы возьмем достаточное число выборок, то в $P\%$ случаев параметр будет находиться в заданном интервале.

Доверительный интервал используется для сравнения выборочного среднего арифметического со средним значением генеральной совокупности, из которой взята выборка. Ошибку утверждения о том, что выборочное среднее равно генеральному среднему, характеризует параметр m – стандартная ошибка среднего. Для принятой в большинстве исследований доверительной вероятности 0,95, доверительный интервал для среднего равен $\pm 2m$. При доверительной вероятности 0,99 доверительный интервал принимают $\pm 3m$.

Для более точного определения границ доверительного интервала можно воспользоваться формулой: $\bar{x} \pm t_{n-1,p} s / \sqrt{n}$, где $\bar{x}$ – среднее значение; n – количество элементов в выборке; s – стандартное отклонение; $t_{n,p}$ – табличное значение распределения Стьюдента с числом степеней свободы $n-1$ и доверительной вероятностью P . Табличное значение распределения Стьюдента можно определить с помощью функции ЭТ Excel СТЬЮДРАСПОБР(альфа; количество степеней свободы); функция для определения доверительного интервала – ДОВЕРИТ(альфа; стандартное отклонение; размер выборки). Еще проще воспользоваться описательной статистикой Пакета Анализа. В строке **Уровень надежности** (95,0%) показана полуширина доверительного интервала для среднего.

Если распределение признака не является нормальным, либо вид распределения неизвестен, для описания выборки используют не среднее значение, а медиану и, соответственно, требуется определить доверительный интервал для медианы. Для этого нужно выполнить следующее: по таблице (см. приложение 1) с учетом объема выборки и доверительной вероятности найти ранги, значения которых соответствуют доверительному интервалу. Далее ранжировать значения признака и определить, которые из значений признака соответствуют табличным рангам.

На практике построение доверительных интервалов часто используется для сравнения выборки с генеральной совокупностью. Пусть известно среднее значение изучаемого признака генеральной совокупности. Необходимо сопоставить с ним среднее значение выборки. Одним из способов решения данной задачи является оценка изучаемого признака с помощью 95% доверительного интервала для среднего значения (в случае нормального закона распределения изучаемого признака) или медианы (если вид распределения не является нормальным или неизвестен). Если рассчитанный доверительный интервал для среднего не включает среднее значение генеральной совокупности, то с заданной доверительной вероятностью можно утверждать, что выборка статистически отличается от генеральной совокупности.

Задание и методические рекомендации к выполнению работы

Пусть имеются данные по росту детей в возрасте 10 лет (в см): 138, 139, 134, 128, 130, 145, 128, 129, 132, 133. Известно из исследований, что средний рост детей такого возраста составляет 140 см. Определить, не отстают ли в росте дети исследуемой группы. Решить задачу в предположении, что выборка имеет нормальное распределение, и что закон распределения неизвестен или не является нормальным.

Выполнение задания

Для решения задачи необходимо ввести в таблицу данные о росте детей. Предположим, выборка имеет нормальное распределение. Далее необходимо определить среднее значение и доверительный интервал для среднего выборки; рассчитать значения верхней и нижней границ доверительного интервала. Получаются следующие результаты: среднее значение выборки равно 133,6; полуширина доверительного интервала – 3,4; нижняя граница доверительного интервала – 130,2; верхняя граница доверительного интервала – 137.

В данном случае, так как известное среднее генеральной совокупности 140 см не входит в доверительный интервал от 130,2 до 137 см, можно утверждать с доверительной вероятностью 95%, что в исследуемой группе дети отстают по росту от сверстников.

Если выборка имеет распределение, отличное от нормального, необходимо найти медиану выборки и доверительный интервал для медианы. Медиану можно найти с помощью функции МЕДИАНА() или используя процедуру *Описательная статистика* Пакета Анализа. Для данной выборки медиана равна 132,5 см.

По таблице (см. приложение 1) определяем, что для заданного количества наблюдений, равного десяти и уровня значимости $\alpha=0,05$ доверительный интервал определяется 2 и 9 рангами. Для определения рангов значений признака применяется Пакет Анализа: процедура

Ранг и персентиль. В появившемся диалоговом окне необходимо выполнить следующие действия:

- задать *Входной интервал*, то есть ввести ссылку на диапазон анализируемых данных;
- указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа (указать адрес левой верхней ячейки выходного диапазона).

После нажатия кнопки ОК будут получены следующие результаты: в столбце Точка находятся порядковые номера наблюдений; Столбец1 – значения наблюдений, отсортированные в порядке убывания; Ранг – порядковые ранги наблюдений; Процент – процентные ранги.

Второй и девятый ранги определяют нижнюю и верхнюю границы доверительного интервала для медианы. Этим рангам соответствуют значения наблюдений 139 и 128 см. Таким образом, доверительный интервал для медианы находится в диапазоне от 128 до 139 см.

Известное среднее генеральной совокупности 140 см не входит в доверительный интервал для данной выборки от 128 до 139 см, следовательно, можно утверждать с доверительной вероятностью 95%, что в исследуемой группе дети отстают по росту от сверстников.

Задания для индивидуального выполнения

Пусть имеются данные по результатам измерений в выборке и значение среднего (медианы) в генеральной совокупности. Определить с уровнем значимости $\alpha=0,05$, отличается ли выборка от генеральной совокупности. Рассмотреть два случая: данные распределены по нормальному закону и закон распределения данных неизвестен.

Вариант 1. Данные: 12, 9, 10, 10, 2, 7, 16, 11, 14, 8, 4, 18, 12, 9, 13, 2, 17, 0, 8, 27, 12, 5. Среднее (медиана) = 14.

Вариант 2. Данные: 16, 12, 20, 12, 13, 9, 6, 13, 15, 14, 10, 6, 10, 16, 12, 14, 11, 13, 12, 14. Среднее (медиана) = 12.

Вариант 3. Данные: 4, 11, 1, 12, 10, 13, 1, 13, 3, 7, 1, 1, 3, 2, 6, 17, 3, 1, 11, 3, 9, 12, 14, 8. Среднее (медиана) = 11.

Вариант 4. Данные: 9, 15, 1, 3, 9, 15, 14, 17, 8, 16, 13, 15, 6, 16, 9, 7, 11, 7, 11, 10, 9, 12, 1. Среднее (медиана) = 12

Вариант 5. Данные: 18, 14, 7, 22, 12, 17, 19, 15, 12, 12, 12, 15, 15, 8, 9, 15, 16, 16, 14, 9. Среднее (медиана) = 15.

Вариант 6. Данные: 2, 9, 7, 22, 13, 5, 7, 10, 9, 13, 7, 6, 12, 7, 10, 10, 21, 15, 11, 10, 4, 5, 8. Среднее (медиана) = 10.

Вариант 7. Данные: 19, 11, 10, 12, 13, 17, 23, 14, 14, 10, 7, 24, 11, 15, 12, 5, 4, 0, 18, 11, 1. Среднее (медиана) = 17.

Вариант 8. Данные: 8, 6, 14, 12, 3, 10, 10, 6, 10, 19, 14, 7, 10, 16, 12, 12, 13, 12, 17, 13, 5. Среднее (медиана) = 11.

Вариант 9. Данные: 12, 11, 5, 20, 11, 12, 16, 15, 9, 16, 9, 15, 14, 9, 13, 7, 7, 11, 10, 14, 5. Среднее (медиана) = 10.

Вариант 10. Данные: 14, 8, 12, 9, 7, 18, 5, 3, 9, 14, 16, 6, 21, 17, 8, 11, 12, 8, 9, 13, 10, 11. Среднее (медиана) = 12.

Вариант 11. Данные: 13, 9, 11, 10, 4, 7, 17, 11, 13, 8, 4, 19, 13, 8, 13, 2, 17, 5, 8, 21, 13, 6. Среднее (медиана) = 14.

Вариант 12. Данные: 17, 12, 22, 12, 14, 9, 6, 14, 15, 15, 11, 6, 17, 19, 13, 14, 11, 16, 12, 15. Среднее (медиана) = 13.

Вариант 13. Данные: 4, 12, 4, 13, 10, 14, 5, 17, 3, 9, 3, 4, 3, 9, 3, 18, 3, 7, 11, 3, 8, 13, 12, 7. Среднее (медиана) = 12.

Вариант 14. Данные: 9, 18, 2, 3, 8, 13, 14, 16, 8, 16, 14, 15, 8, 13, 9, 8, 11, 7, 12, 10, 6, 15, 2. Среднее (медиана) = 12.

Вариант 15. Данные: 16, 15, 7, 21, 12, 18, 19, 19, 12, 15, 12, 15, 13, 8, 8, 15, 16, 17, 18, 7. Среднее (медиана) = 14.

Лабораторная работа 6

Тема: «Проверка гипотез. Выявление достоверности различий между двумя выборками с помощью t -критерия Стьюдента»

Цель: изучить параметрический t -критерий Стьюдента для проверки гипотезы о равенстве средних значений изучаемого признака.

Основные теоретические сведения

Статистическая гипотеза – это предположение о виде или отдельных параметрах распределения вероятностей, которое подлежит проверке на основе анализа имеющихся данных. Проверка статистических гипотез позволяет проверить или отвергнуть некоторое утверждение.

Нулевая гипотеза (H_0) – предположение о том, что события произошли случайно; альтернативная (H_1) – предположение о закономерности событий, о том, что имеет место воздействие какого-либо фактора. Нулевая гипотеза формируется таким образом, чтобы на основании статистических данных ее можно было отвергнуть с заранее заданной вероятностью ошибки α , которая называется уровнем значимости.

Выдвинутая гипотеза может быть правильной или неправильной, поэтому она проверяется статистическими методами. Как правило, для совместного анализа двух выборок проводят проверку статистических гипотез о принадлежности обеих выборок одной генеральной совокупности или о равенстве двух средних. Для решения задач такого типа используются так называемые критерии согласия. Критерии согласия могут быть параметрические, которые служат для про-

верки гипотез о параметрах нормально распределенной генеральной совокупности, и непараметрические, которые не требуют знания параметров распределения. Из параметрических критериев для проверки гипотез о равенстве генеральных средних наиболее часто используется t -критерий Стьюдента.

При использовании этого критерия нулевая гипотеза H_0 формулируется так: «Средние двух выборок относятся к одной и той же совокупности» или «Различия между средними выборок несущественны». Соответственно, альтернативная гипотеза H_1 – «различия между средними двух выборок неслучайны, статистически значимы».

Существует два варианта использования t -критерия для проверки гипотезы об отсутствии различий между выборками: если выборки независимы, несвязанны (двухвыборочный t -критерий) и парный t -критерий для связанных (зависимых) групп.

Задания и методические рекомендации к выполнению работы

Задание 1. Имеются данные об итоговой успеваемости (по 10-бальной системе) студентов, посещавших факультативные занятия по данному предмету и не посещавших. Определить, имеются ли различия между этими двумя группами студентов.

Посещавшие	8	8	7	9	8	6	7	8	6	5	8	10	6	8	9
Не посещавшие	5	6	7	6	8	7	7	5	8	6	5	9	6	7	5

Выполнение задания

Для оценки достоверности различий можно использовать функцию ТТЕСТ и процедуры Пакета Анализа. ТТЕСТ возвращает вероятность, соответствующую критерию Стьюдента и имеет следующий синтаксис:

ТТЕСТ(массив1; массив2; хвосты; тип)

Массив1 – первое множество данных.

Массив2 – второе множество данных.

Хвосты – число хвостов распределения. Если хвосты = 1, то функция ТТЕСТ использует одностороннее распределение. Если хвосты = 2, то функция ТТЕСТ использует двустороннее распределение. Обычно число хвостов – 2.

Тип — вид исполняемого t -теста. Возможны три варианта выбора: 1 – парный тест; 2 – двухвыборочный тест с равными дисперсиями; 3 – двухвыборочный тест с неравными дисперсиями.

Для решения задачи нужно ввести исходные данные в диапазон ячеек A1:B15. Далее установить курсор в свободную ячейку B16 и ввести в нее формулу =ТТЕСТ(A1:A15;B1:B15;2;3). В данном случае массив 1 задается в диапазоне ячеек A1:A15; массив 2 – в диапазоне B1:B15; число хвостов распределения 2; а в поле тип – 3, так как имеются две несвязанные выборки и говорить о равенстве дисперсий за-

труднительно. Результат: t -критерий равен 0,033011. Поскольку величина вероятности случайного различия между выборками меньше уровня значимости $\alpha=0,05$, то нулевая гипотеза отвергается. Следовательно, различия между выборками неслучайны и выборки считаются достоверно отличающимися друг от друга. Для сравнения двух независимых групп используется процедура **Двухвыборочный t -тест с различными дисперсиями** Пакета Анализа.

В появившемся диалоговом окне необходимо выполнить следующие действия:

- задать *Интервал переменной 1*, то есть ввести ссылку на первый диапазон анализируемых данных;
- задать *Интервал переменной 2*, то есть ввести ссылку на второй диапазон анализируемых данных;
- указать уровень значимости $\alpha=0,05$;
- указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа (указать адрес левой верхней ячейки выходного диапазона);
- нажать кнопку ОК.

Результаты анализа указаны на рисунке 9:

D	E	F
Двухвыборочный t-тест с различными дисперсиями		
	Переменная 1	Переменная 2
Среднее	7,533333333	6,466666667
Дисперсия	1,838095238	1,552380952
Наблюдения	15	15
Гипотетическая разность средних	0	
df	28	
t-статистика	2,243592625	
P(T<=t) одностороннее	0,016475744	
t критическое одностороннее	1,701130259	
P(T<=t) двухстороннее	0,032951487	
t критическое двухстороннее	2,048409442	

Рисунок 9 – Результаты применения t -критерия Стьюдента для независимых выборок

Если величина вероятности случайного появления анализируемых выборок $P(t)$ -двухстороннее меньше уровня значимости $\alpha=0,05$, считается что различия между выборками неслучайные. То есть, несмотря на малую разницу между средними баллами в группах (всего в один балл), различия достоверные.

Задание 2. Пусть имеются данные о результатах группы студентов по скоростному чтению до и после специального курса по быст-

рому чтению. Определить, произошли ли статистически значимые изменения скорости чтения у студентов.

Выполнение задания

Задача решается аналогично заданию 1, но при вводе функции ТТЕСТ указывается тип 1, так как выборки зависимые (опыт проводился с одной и той же группой студентов). При использовании Пакета Анализа выбирается процедура **Парный двухвыборочный t-тест для средних**. Результирующая таблица будет иметь следующий вид (рисунок 10):

	А	В	С	Д	Е	Ф
1	До курса	После курса		Парный двухвыборочный t-тест для средних		
2	86	82				
3	83	79			До курса	После курса
4	86	91		Среднее	79,9	83,3
5	70	77		Дисперсия	71,87777778	58,67777778
6	66	68		Наблюдения	10	10
7	90	86		Корреляция Пирсона	0,761861515	
8	70	81		Гипотетическая разность средних	0	
9	85	90		df	9	
10	77	85		t-статистика	-1,912649432	
11	86	94		P(T<=t) одностороннее	0,044043629	
12				t критическое одностороннее	1,833113856	
13	t-критерий =	0,088087258		P(T<=t) двухстороннее	0,088087258	
14				t критическое двухстороннее	2,262158887	

Рисунок 10 – Результаты применения t-критерия Стьюдента для зависимых выборок

Интерпретация результатов: средняя скорость чтения в группах различается значительно, но $P(t)$ -двухстороннее больше уровня значимости $\alpha=0,05$, следовательно, нулевую гипотезу отвергнуть нельзя. Различия между выборками могут быть случайными и средние выборок не считаются достоверно отличающимися друг от друга.

Задание для индивидуального выполнения

В результате измерений изучаемого признака получены две выборки x и y . Определить, есть ли статистические различия между выборками, или они принадлежат к одной генеральной совокупности. Рассмотреть два случая: выборки независимы; выборки зависимы. Для выполнения задания использовать соответствующую функцию и процедуру **t-тест** Пакета Анализа.

Вариант 1.

x	14	11	15	4	7	15	12	12	13	8	13	18	9	13	6
y	7	14	22	16	16	17	13	17	16	12	16	19	2	18	12

Вариант 2.

x	3	6	17	11	9	5	12	6	7	13	8	9	4	19	11
y	2	9	16	2	17	16	9	19	12	11	17	22	3	25	19

Вариант 3.

x	8	1	4	11	6	22	11	11	16	1	19	1	17	5	14
y	21	14	14	20	17	22	17	21	15	7	22	10	11	6	1

Вариант 4.

x	5	9	4	12	10	12	10	4	5	12	9	11	24	13	5
y	19	17	13	10	15	12	12	6	11	12	17	11	10	18	13

Вариант 5.

x	8	11	16	1	15	7	1	6	7	11	8	17	9	20	22
y	3	14	16	8	4	2	4	1	7	7	7	9	6	12	6

Вариант 6.

x	11	22	16	10	8	12	10	7	12	16	0	10	11	13	4
y	6	5	8	3	7	5	6	7	15	15	3	3	3	21	2

Вариант 7.

x	12	14	17	13	10	12	13	8	7	11	11	12	6	1	4
y	5	6	10	7	6	7	4	10	9	12	6	6	7	9	7

Вариант 8.

x	12	5	24	8	13	1	13	7	5	13	11	5	7	22	20
y	4	7	9	10	4	11	10	10	5	1	10	8	6	8	5

Вариант 9.

x	10	8	7	15	13	8	16	9	9	13	13	11	14	1	18
y	19	4	2	4	12	6	11	7	13	9	11	4	7	1	6

Вариант 10.

x	5	8	7	6	11	3	2	5	4	11	21	8	9	12	9
y	13	16	15	14	8	10	16	14	7	11	13	7	5	14	7

Вариант 11.

x	13	11	16	3	8	15	13	12	14	9	13	19	8	13	7
y	8	13	22	15	17	17	14	18	16	14	17	18	5	19	12

Вариант 12.

x	3	7	16	12	8	5	13	7	8	14	8	8	5	18	11
y	3	8	17	4	16	16	9	18	11	12	17	21	4	25	18

Вариант 13.

x	7	2	4	12	6	23	11	12	16	2	19	3	17	6	13
y	21	15	14	21	18	21	18	21	17	7	21	13	12	6	2

Вариант 14.

x	6	7	4	12	11	13	10	5	5	13	8	11	26	15	3
y	18	16	13	10	14	12	13	7	11	13	17	10	10	17	15

Вариант 15.

x	7	12	16	2	15	6	3	5	7	12	8	18	8	21	23
y	4	16	17	8	5	2	5	2	7	8	7	10	7	12	7

Лабораторная работа 7

Тема: «Построение таблиц сопряженности.

Выявление достоверности различий между двумя выборками с помощью критерия хи-квадрат»

Цель: изучить применение критерия хи-квадрат для проверки статистических гипотез о независимости и однородности выборок.

Основные теоретические сведения

Таблицы для двух переменных называются таблицами сопряженности. Они показывают связи или отношения между переменными. Если изучается связь между независимой и зависимой переменной, то зависимая чаще размещается в строках, а независимая в столбцах. Для статистической обработки таблиц сопряженности часто используется критерий хи-квадрат, который относится к непараметрическим критериям проверки гипотез. Этот критерий можно использовать с различными формами распределений совокупностей. Функция ХИ2ТЕСТ позволяет вычислить совпадения наблюдаемых (фактических) значений и теоретических (гипотетических) значений. Если вычисленная вероятность ниже уровня значимости $\alpha=0,05$, то нулевая гипотеза о принадлежности наблюдаемых и ожидаемых значений к одной генеральной совокупности отвергается, и утверждается, что наблюдаемые значения не соответствуют теоретическим значениям. Условием применения этого метода является наличие не менее 5 наблюдений в каждой группе вариант.

Задания и методические рекомендации к выполнению работы

Задание 1. В ходе социологического опроса среди взрослых сотрудников на вопрос о перенесенном в детстве заболевании ответы распределились следующим образом:

	Да	Нет	Не помню
Мужчины	58	11	10
Женщины	35	25	23

Определить, имеются ли достоверные отличия в ответах женщин и мужчин?

Выполнение задания

Выдвигается нулевая гипотеза H_0 : влияния пола на наличие перенесенных в детстве заболеваний нет. Альтернативная гипотеза – имеется влияние пола на наличие перенесенных заболеваний. Далее необходимо вычислить ожидаемые (теоретические) численности наблюдений по каждой ячейке. Для этого сначала находятся суммы по строкам и столбцам, например, в ячейку E3 нужно ввести формулу =СУММ(B3:D3), а в ячейку B5 – формулу =СУММ(B3:B4).

Ожидаемые численности наблюдений вычисляются исходя из следующего соотношения:

$$\text{Ожидаемое число наблюдений} = \frac{\text{Сумма по столбцу} \times \text{Сумма по строке}}{\text{Общий итог}}$$

Например, ожидаемое количество мужчин, которые дали ответ «Да» составляет $93 \times 79 / 162 \approx 45$. Для расчета ожидаемых значений следует заполнить диапазон A8:D9 формулами следующим образом (рисунок 11):

	A	B	C	D	E
1	<i>Фактическое распределение</i>				
2		Да	Нет	Не помню	Сумма
3	Мужчины	58	11	10	=СУММ(B3:D3)
4	Женщины	35	25	23	=СУММ(B4:D4)
5	Всего	=СУММ(B3:B4)	=СУММ(C3:C4)	=СУММ(D3:D4)	=СУММ(B5:D5)
6					
7	<i>Ожидаемое распределение</i>				
8	Мужчины	=B5*\$E\$3/\$E\$5	=C5*\$E\$3/\$E\$5	=D5*\$E\$3/\$E\$5	
9	Женщины	=B5*\$E\$4/\$E\$5	=C5*\$E\$4/\$E\$5	=D5*\$E\$4/\$E\$5	

Рисунок 11 – Заполнение рабочей таблицы данными и формулами
Результирующая таблица будет иметь вид (рисунок 12):

	A	B	C	D	E
1	<i>Фактическое распределение</i>				
2		Да	Нет	Не помню	Сумма
3	Мужчины	58	11	10	79
4	Женщины	35	25	23	83
5	Всего	93	36	33	162
6					
7	<i>Ожидаемое распределение</i>				
8	Мужчины	45,3519	17,556	16,092593	
9	Женщины	47,6481	18,444	16,907407	

Рисунок 12 – Результаты вычислений

Для вычисления значения критерия хи-квадрат нужно установить курсор в свободную ячейку таблицы, например, в ячейку A11 и ввести формулу: =ХИ2ТЕСТ(B3:D4;B8:D9). В результате будет получена вероятность нулевой гипотезы, равная 0,00031, что значительно меньше уровня значимости $\alpha=0,05$. Следовательно, нулевая гипотеза отвергается, то есть можно утверждать, что имеются достоверные отличия в ответах мужчин и женщин.

Задание 2. Пусть имеются по возрастные данные о заболеваемости группы работников школ. Определить, имеются ли достоверные различия в числе случаев заболевания в разных возрастных группах (рисунок 13):

	А	В	С	Д	Е
1	<i>Зависимость числа заболеваний от возраста (лет)</i>				
2	Вычисляемые данные	до 30	30-39	40-49	50 и старше
3	Число работников	55	60	120	47
4	Фактическое число случаев заболеваний	11	18	48	6

Рисунок 13 – Заполнение рабочей таблицы данными и формулами

Выполнение задания

Нулевая гипотеза H_0 : выборки принадлежат к одной генеральной совокупности. Альтернативная – имеются существенные различия между выборками.

Для использования критерия хи-квадрат необходимо определить ожидаемое значение рассматриваемого показателя. В данном случае будем исходить из предположения, что если бы заболеваемость в отдельной возрастной группе была такая, как в целом по всем обследованным, то ожидаемая численность заболевших составила бы:

$$\text{Ожидаемая численность заболевших} = \frac{\text{Численность работников в данной возрастной группе} \times \text{Общая численность заболевших}}{\text{Общая численность работников}}$$

Далее необходимо дополнить таблицу формулами для расчета ожидаемой численности.

В ячейки F3:F4 вводятся формулы для вычисления общего количества числа работников и общего числа случаев заболеваний; в ячейку F5 вводится формула для расчета процентного отношения общего количества заболевших к общему числу работников. В диапазон ячеек B5:E5 вводятся формулы для расчета ожидаемого числа случаев заболеваний. В результате расчетов будут получены следующие данные (рисунок 14):

	А	В	С	Д	Е	Ф
1	<i>Зависимость числа заболеваний от возраста (лет)</i>					
2	Вычисляемые данные	до 30	30-39	40-49	50 и старше	Сумма
3	Число работников	55	60	120	47	282
4	Фактическое число случаев заболеваний	11	18	48	6	83
5	Ожидаемое число случаев заболеваний	16	18	35	14	0,294

Рисунок 14 – Результаты вычислений

Затем вычисляется значение вероятности того, что события (число фактически заболевших работников в разных возрастных группах) произошли случайным образом. Для этого в свободную ячейку таблицы нужно ввести формулу: =ХИ2ТЕСТ(B4:E4;B5:E5). В

рассматриваемом примере полученное значение равно 0,01373. Поскольку величина вероятности случайного появления анализируемых выборок меньше уровня значимости $\alpha=0,05$, то нулевая гипотеза отвергается. Следовательно, на основании применения критерия хи-квадрат можно сделать вывод о том, что в возрастных группах выявлены статистически значимые отличия по количеству заболеваний.

Задания для индивидуального выполнения

Вариант 1. Имеются данные о результатах оценки слушателями курсов повышения квалификации метода подачи теоретического материала. Определить, зависит ли оценка от метода подачи материала.

Оценка	Метод подачи нового материала		
	I	II	III
Отрицательная	12	18	17
Положительная	21	25	20

Вариант 2. Имеются данные об отношении школьников к проведенному мероприятию. Есть ли отличия в отношении мальчиков и девочек к мероприятию?

Пол	Количество школьников, высказавших свое мнение о проведенном мероприятии		
	Положительное	Безразличное	Отрицательное
Мальчики	15	25	6
Девочки	30	35	30

Вариант 3. Имеются данные о трудоустройстве по специальности выпускников трех вузов. Определить, зависит ли от вуза успешность трудоустройства выпускника?

Данные	Вуз 1	Вуз 2	Вуз 3
Общее количество выпускников	200	100	300
Трудоустроившихся по специальности	175	62	220

Вариант 4. Имеется распределение школьников по кратности заболеваний за текущий период. Определить, зависит ли заболеваемость обследованных школьников от пола?

Пол	Кратность заболеваний		
	Не болели	1-2 раза	3 раза и более
Мальчики	29	36	15
Девочки	14	24	12

Вариант 5. В школе были проведены некоторые мероприятия, предположительно влияющие на успеваемость. Влияет ли вид мероприятия на успеваемость?

Вид мероприятия	Успеваемость		
	Увеличилась	Не изменилась	Уменьшилась
Мероприятие А	14	22	20
Мероприятие В	47	37	25

Вариант 6. Имеются данные об отношении студентов к проведенному мероприятию. Есть ли отличия в отношении студентов младших и старших курсов к мероприятию?

Курс	Количество школьников, высказавших свое мнение о проведенном мероприятии		
	Нравится	Безразлично	Не нравится
Студенты младших курсов	15	30	25
Студенты старших курсов	40	33	20

Вариант 7. Имеются данные по количеству сыгранных матчей и количеству забитых голов для трех школьных футбольных команд. Определить, есть ли различия между командами по уровню мастерства.

Данные	Команда 1	Команда 2	Команда 3
Количество матчей	15	10	12
Количество голов	33	9	25

Вариант 8. Для определения зависимости количества школьников, занимающихся в спортивных секциях от их местожительства были обследованы две группы школьников из районов А и В. Свидетельствуют ли полученные данные о зависимости возможности заниматься спортом от местожительства?

Район	Занятия школьников спортом		
	Не занимаются	Занимаются в одной секции	Занимаются в двух и более секциях
А	82	21	87
В	115	45	150

Вариант 9. Имеются данные об общем количестве и количестве переболевших простудными заболеваниями детей в трех дошкольных учреждениях. Определить, имеются ли различия по уровню заболеваемости между данными дошкольными учреждениями?

Данные	Д/с №1	Д/с №2	Д/с №3
Всего детей	150	75	120
Количество переболевших	65	19	47

Вариант 10. Имеется три вида витаминных комплексов для детей, уменьшающих заболеваемость простудными заболеваниями. В таблице указано количество случаев проявления ожидаемого эффекта от применения витаминов. Есть ли различия в результате применения данных витаминных комплексов?

Наличие ожидаемого эффекта	Витаминный комплекс		
	А	В	С
Есть эффект	50	74	48
Нет эффекта	12	25	6

Вариант 11. Имеются данные о результатах оценки студентами условий проживания в общежитиях. Определить, зависит ли оценка от общежития.

Оценка	Общежитие		
	№1	№2	№3
Отрицательная	10	11	12
Положительная	22	25	25

Вариант 12. Имеются данные по количеству тестовых заданий различного типа в трех итоговых контрольных. Определить, имеются ли существенные различия между контрольными.

Количество	Номер контрольного задания		
	1	2	3
Задания открытого типа	18	17	15
Задания закрытого типа	24	43	37

Вариант 13. Имеется распределение студентов по количеству пропусков за семестр. Определить, зависит ли количество пропусков от пола?

Пол	Количество пропусков		
	Меньше 10	10-30	30 и более
Юноши	120	40	10
Девушки	100	35	5

Вариант 14. Имеются данные по количеству соревнований и количеству полученных наград для трех студенческих команд. Определить, есть ли различия между командами по количеству наград.

Данные	Команда 1	Команда 2	Команда 3
Количество соревнований	19	12	15
Количество медалей	43	19	28

Вариант 15. В школе функционируют кружки. Определить, имеются ли различия в оценке заинтересованности школьников в работе кружков?

Вид кружка	Оценка заинтересованности		
	Кружок 1	Кружок 2	Кружок 3
Заинтересованы	15	27	24
Безразличны	48	35	22

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Гмурман В. Е. Теория вероятностей и математическая статистика. М., Высш. шк., 2003. – 479 с.
2. Грабарь М.И., Краснянская К.А. Применение математической статистики в педагогических исследованиях: Непараметрические методы. М.: Педагогика, 1977. –136с.
3. Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. – 495 с.
4. Сидоренко Е.В. Методы математической обработки в психологии. СПб.: Речь, 2000. – 350 с.
5. Ю.К. Бабанский, В. И. Журавлев, В. К. Розов и др. Введение в научное исследование по педагогике: Учеб. пособие для студентов пед.институтов. М.: Просвещение, 1988. – 239 с.
6. Кремер Н.Ш. Теория вероятностей и математическая статистика: Учебник для вузов. М.: ЮНИТИ-ДАНА, 2004. – 573 с.
7. Орлов А.И. Эконометрика. М.: Экзамен, 2003. – 576 с.

ПРИЛОЖЕНИЯ

Приложение 1

**Ранги, с помощью которых вычисляется доверительный
интервал для медианы распределения**

Объем выборки	Уровень значимости (доверительная вероятность)		
	$\alpha=0,1$ (90%)	$\alpha=0,05$ (95%)	$\alpha=0,01$ (99%)
10	2; 9	2; 9	1; 10
11	3; 9	2; 10	1; 11
12	3; 10	3; 10	2; 11
13	4; 10	3; 11	2; 12
14	4; 11	3; 12	2; 13
15	4; 12	4; 12	3; 13
16	5; 12	4; 13	3; 14
17	5; 13	5; 13	3; 15
18	6; 13	5; 14	4; 15
19	6; 14	5; 15	4; 16
20	6; 15	6; 15	4; 17
21	7; 15	6; 16	5; 17
22	7; 16	6; 17	5; 18
23	8; 16	7; 17	5; 19
24	8; 17	7; 18	6; 19
25	8; 18	8; 18	6; 20
26	9; 18	8; 19	7; 20
27	9; 19	8; 20	7; 21
28	10; 19	9; 20	7; 22
29	10; 20	9; 21	8; 22
30	11; 20	10; 21	8; 23
35	13; 23	12; 24	10; 26
40	15; 26	14; 27	12; 29
45	17; 29	16; 30	14; 32
50	19; 32	18; 33	16; 35
55	21; 35	20; 36	18; 38
60	24; 37	22; 39	20; 41
65	26; 40	25; 41	22; 44
70	28; 43	27; 44	24; 47
75	30; 46	29; 47	26; 50
80	33; 58	31; 50	29; 52
85	35; 51	33; 53	31; 55
90	37; 54	36; 55	33; 58
95	39; 57	38; 58	35; 61
100	42; 59	40; 61	37; 64

Приложение 2

Значения t-критерия Стьюдента при уровне значимости α : 0,10, 0,05, 0,01

Число степеней свободы	α			Число степеней свободы	α		
	0,1	0,05	0,01		0,1	0,05	0,01
1	6,314	12,706	63,66	18	1,734	2,101	2,878
2	2,92	4,3027	9,925	19	1,729	2,093	2,861
3	2,353	3,1825	5,841	20	1,725	2,086	2,845
4	2,132	2,7764	4,604	21	1,721	2,08	2,831
5	2,015	2,5706	4,032	22	1,717	2,074	2,819
6	1,943	2,4469	3,707	23	1,714	2,069	2,807
7	1,895	2,3646	3,5	24	1,711	2,064	2,797
8	1,86	2,306	3,355	25	1,708	2,06	2,787
9	1,833	2,2622	3,25	26	1,706	2,056	2,779
10	1,813	2,2281	3,169	27	1,703	2,052	2,771
11	1,796	2,201	3,106	28	1,701	2,048	2,763
12	1,782	2,1788	3,055	29	1,699	2,045	2,756
13	1,771	2,1604	3,012	30	1,697	2,042	2,75
14	1,761	2,1448	2,977	40	1,684	2,021	2,705
15	1,753	2,1315	2,947	60	1,671	2	2,66
16	1,746	2,1199	2,921	120	1,658	1,98	2,617
17	1,74	2,1098	2,898	∞	1,645	1,96	2,576