Лабораторная работа № 2  
Тема: «ПЕРВИЧНАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ»

**Цель работы:** Освоить методику выявления грубых погрешностей наблюдений, научиться представлять экспериментальные данные в графическом виде и вычислять их основные числовые характеристики.

**Постановка задачи**

1) Сформировать ряд наблюдений случайной величины.

2) Внести в одно из наблюдений грубую погрешность.

3) Продемонстрировать практическую возможность обнаружения аномальных значений наблюдений с использованием критерия Стьюдента средствами Excel.

4) Продемонстрировать практическую возможность обнаружения аномальных значений наблюдений с использованием критерия Стьюдента средствами MathCad.

5) Для уровней значимости, указанных в таблице вариантов заданий, определить пределы, в которых значение выборки не будет считаться выбросом из нее.

6) Представить экспериментальные данные в виде гистограммы, полигона частот и кумуляты средствами Excel.

7) Вычислить основные числовые характеристики выборки средствами Excel и MathCad. Результаты сравнить.

6) Оформить отчет с результатами вычислений, выполненных в Excel и Mathcad, написать выводы к работе.

**Теоретические сведения**

Критерий Стьюдента чаще всего используется на практике для отсева грубых погрешностей (аномальных значений) экспериментальных данных.

Пусть имеется ряд наблюдений *x1, x2, …, xn+1* случайной величины.

1. Находим число, подозрительное на выброс, и перемещаем его в конец выборки - на (*n+1)*-е место.
2. Рассчитываем следующие характеристики выборки (без числа, подозрительного на выброс):

эмпирическое среднее ,

несмещенную эмпирическую дисперсию ,

среднеквадратическое отклонение 

1. Вычисляем значение



1. Находим табличное (см. Таблицу 1) значение критерия Стьюдента , где *k* – количество степеней свободы, в нашем случае равное *n–1*, а *P* – заданная вероятность, с которой *xn+1* считается выбросом (например, 0.95, 0.99 или 0.999).

**Таблица 1**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***P*** | **0,95** | **0,99** | **0,999** | ***P*** | **0,95** | **0,99** | **0,999** |
| ***k*** | ***k*** |
| 4 | 2,78 | 4,60 | 8,61 | 19 | 2,093 | 2,861 | 3,883 |
| 5 | 2,57 | 4,03 | 6,86 | 20 | 2,086 | 2,845 | 3,849 |
| 6 | 2,45 | 3,71 | 5,96 | 25 | 2,064 | 2,797 | 3,745 |
| 7 | 2,37 | 3,50 | 5,41 | 30 | 2,045 | 2,756 | 3,659 |
| 8 | 2,31 | 3,36 | 5,04 | 35 | 2,032 | 2,729 | 3,600 |
| 9 | 2,26 | 3,25 | 4,78 | 40 | 2,023 | 2,708 | 3,558 |
| 10 | 2,23 | 3,17 | 4,59 | 45 | 2,016 | 2,692 | 3,527 |
| 11 | 2,20 | 3,11 | 4,44 | 50 | 2,009 | 2,679 | 3,502 |
| 12 | 2,18 | 3,06 | 4,32 | 60 | 2,001 | 2,662 | 3,464 |
| 13 | 2,16 | 3,01 | 4,22 | 70 | 1,996 | 2,649 | 3,439 |
| 14 | 2,15 | 2,98 | 4,14 | 80 | 1,991 | 2,640 | 3,418 |
| 15 | 2,13 | 2,95 | 4,67 | 90 | 1,987 | 2,633 | 3,403 |
| 16 | 2,12 | 2,92 | 4,02 | 100 | 1,984 | 2,627 | 3,392 |
| 17 | 2,11 | 2,90 | 3,97 | 120 | 1,980 | 2,617 | 3,374 |
| 18 | 2,10 | 2,88 | 3,92 | ∞ | 1,960 | 2,576 | 3,291 |

Очевидно, что значение наблюдения может оказаться выбросом с вероятностью 0.95, но не быть выбросом с вероятностью, например, 0.99.

1. Если , то это выброс, его исключаем или (в зависимости от конкретных условий задачи) заменяем на своем месте величиной, интерполированной по соседним значениям.

*i*

*i-1*

*i*

*i-2*

*x*

выброс

При отделении выбросов по критерию Стьюдента нулевой гипотезой является принадлежность подозрительного значения имеющейся выборке данных, т.е. что это значение – не выброс. Так как величина P в Таблице 1 – это вероятность того, что подозрительное значение является выбросом, то уровень значимости α=1–Р – это вероятность ошибки 1-го рода – вероятность принятия решения о том, что xn+1 выброс, если на самом деле это не выброс. Чем меньше уровень значимости α, тем меньше риск ошибочного отклонения проверяемой гипотезы. Уровни значимости в литературе часто указывают в %, т.е. вероятностям 0.95, 0.99, 0.999 соответствуют уровни 5%, 1%, 0.1%.

Вместо табличных значений распределения Стьюдента удобно использовать значения функции ***СТЬЮДРАСПОБР(*α*, k)*** в среде EXCEL или ***qt(P, k)*** в пакете MathCad. Следует отметить, что MathCad возвращает значения для *P* = 1–0.5α, т.е. значения первого столбца вышеприведенной таблицы можно получить, обращаясь к функции *qt(0.975, k)*.

**Получение наблюдений**

В качестве наблюдений ***x1, x2, …, xn+1*** необходимо использовать 50 целых случайных чисел в заданном по варианту диапазоне, которые были получены в лабораторной работе №1.

Грубую погрешность вносим программно, прибавлением № варианта к максимальному значению выборки – это и будет подозрительное на выброс значение, которое следует переставить в конец выборки – теперь оно будет ***xn+1***.

**Обнаружение аномальных значений наблюдений с использованием критерия Стьюдента средствами Excel**

1. Оформить полученные значения ряда наблюдений в виде таблицы.
2. Рассчитать значение выборочного среднего для ***x1, x2, …, xn*** с использованием функции **СРЗНАЧ**.
3. Рассчитать несмещенную оценку СКО для ***x1, x2, …, xn*** с помощью функции **СТАНДОТКЛОН**.
4. Вычислить значение *t* и сравнить его со значением критерия Стьюдента , полученным с использованием функции **СТЬЮДРАСПОБР**(α,k) для 5%-го уровня значимости.
5. Результат сравнения отобразить на листе Excel с помощью функции **ЕСЛИ**.

**Обнаружение аномальных значений наблюдений с использованием критерия Стьюдента средствами MathCad**

1. Представить полученные значения ряда наблюдений в виде массива.

Это можно сделать следующими способами:

* 1. через список значений:

Указать начальное значение индексов элементов массива:

*ORIGIN:=1*

Набрать

*n:=*

и после знака присваивания ввести количество значений ряда наблюдений без учета значения, проверяемого на выброс.

Задать диапазон индексов элементов массива:

*i:=1..n*

Набрать

*ai:=*

и после знака присваивания ввести *n* значений ряда наблюдений через запятую.

Показать массив наблюдений, введя *а=*

* 1. через команду ввода матриц:

Ввести название массиваи нажать на пиктограмму **Matrix or Vector** панели инструментов **Matrix** либо комбинацию клавиш **<Ctrl>+<M>**. Задать размеры массива.

**

Заполнить массив данными.

*Примечания*:

* + - знак присваивания **":="** можно вводить, выбрав соответствующую пиктограмму панели инструментов **Сalculator**, или нажатием комбинации клавиш **<Shift>+<:>**
    - знак диапазона **".."** можно вводить, нажав соответствующую пиктограмму панели инструментов **Сalculator**, или нажатием клавиши **<;>**
    - нижний индекс можно вводить, нажав соответствующую пиктограмму панели инструментов **Matrix** или на клавишу **<[>**

1. Рассчитать значение выборочного среднего с использованием функции **mean,** задав в качестве аргумента имя введенного массива

m:=mean(a)

Показать вычисленное значение, введя m=

1. Рассчитать несмещенную оценку СКО с помощью функции **Stdev,** задав в качестве аргумента имя введенного массива:

s:=Stdev(a)

Показать вычисленное значение

1. Вычислить значение ***t****,* введя формулу для его вычисления через знак присваивания.

Показать вычисленное значение.

*Примечани*я:

* + - Модуль в формуле можно получить, нажав на соответствующую пиктограмму панели инструментов **Сalculator** или клавишу **<|>**.
    - Деление получается нажатием **</>** или через панель инструментов **Сalculator**.

1. Вычислить значения критерия Стьюдента, с использованием функции ***qt(0.95,k)****,* ***qt(0.975,k)****,*вывести вычисленное значениеи сравнить его со значением ***t****.*
2. Результат сравнения отобразить на рабочем листе MathCad
3. Оформить документ MathCad с комментариями

*Примечание*. Вставить комментарий можно с помощью команды **Insert /Text Region** (**Вставка / Текстовая область**) или перед тем как ввести первый символ, нажать клавишу **<">**.

**Определение пределов для заданного уровня значимости**

1. Найти  для *k* = *n–1* и уровней значимости α, взятых из таблицы вариантов заданий (Таблица 2)

2. Определить допустимые пределы случайной величины, решив неравенство для каждого  из найденных в п. 1:

,

где  и  – уже найденные значения выборочного среднего и несмещенной оценки СКО

3. Проиллюстрировать решения п. 2 графически, например, на числовой оси:

Указать выборочное среднее, четыре найденных границы допустимых пределов и соответствующие уровни значимости

4. Сформулировать выводы о зависимости допустимых пределов от уровня значимости нулевой гипотезы

**Таблица 2.**

**Варианты заданий**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Уровень значимости,** α1 | **Уровень значимости,** α2 | **№ п/п** | **Уровень значимости,** α1 | **Уровень значимости,** α2 |
| 1 | 0.01 | 0.19 | 14 | 0.14 | 0.01 |
| 2 | 0.02 | 0.18 | 15 | 0.15 | 0.02 |
| 3 | 0.03 | 0.17 | 16 | 0.16 | 0.03 |
| 4 | 0.04 | 0.16 | 17 | 0.17 | 0.04 |
| 5 | 0.05 | 0.15 | 18 | 0.18 | 0.05 |
| 6 | 0.06 | 0.14 | 19 | 0.19 | 0.06 |
| 7 | 0.07 | 0.26 | 20 | 0.20 | 0.07 |
| 8 | 0.08 | 0.25 | 21 | 0.21 | 0.08 |
| 9 | 0.09 | 0.24 | 22 | 0.22 | 0.09 |
| 10 | 0.10 | 0.23 | 23 | 0.23 | 0.10 |
| 11 | 0.11 | 0.22 | 24 | 0.24 | 0.11 |
| 12 | 0.12 | 0.21 | 25 | 0.25 | 0.12 |
| 13 | 0.13 | 0.20 | 26 | 0.26 | 0.13 |

**Визуализация экспериментальных данные в виде гистограммы, полигона частот и кумуляты средствами Excel**

Первичные экспериментальные данные нуждаются в обработке, которая начинается с их группировки. Группировка представляет собой процесс систематизации, упорядочения первичных данных с целью извлечения содержащейся в них информации.

Выполняя группировку, необходимо разбить весь диапазон варьирования признака на интервалы. Предпочтительны интервалы одинаковой ширины. Приближенно число интервалов k можно оценить исходя только из объема выборки n, например по формуле Стерджеса:

k=1+3,32lg(n).

Ширина каждого интервала определяется следующим образом:

*,*

где h – ширина интервала, xmax и xmin – максимальная и минимальная варианты выборки. Для их нахождения используются функции Excel **МАКС()** и **МИН()** соответственно. Найденное значение шага интервала при необходимости округляют в большую сторону, чтобы не уменьшать диапазон варьирования признака. Для округления можно использовать функцию Excel **ОКРУГЛВВЕРХ()**.

Нижняя граница первого интервала выбирается как xmin.

Прибавив к этой величине ширину интервала, получим нижнюю границу второго интервала, что и будет верхней границей первого интервала и т.д.

Срединные значения интервалов группировки xi определяется следующим образом:

,

где xнi – нижняя граница i-того интервала.

Количество нижних границ интервалов равно k+1.

Далее заполняется таблица следующего вида:

Таблица 3

**Табличное представление экспериментальных данных**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Номер интервала, i | Границы интервалов | | Срединные значения, xi | Частоты, ni | Накопленные частоты, nxi |
| хнi | хві |
|  |  |  |  |  |  |

При вычислении значений ni можно использовать функцию Excel **ЧАСТОТА()**. Общая сумма всех частот равна объему выборки n, что можно проверить, просуммировав значения указанного столбца с помощью функции Excel **СУММ()**.

Накопленная частота интервала – число, полученное последовательным суммированием частот в направлении от первого интервала к последнему, до того интервала включительно, для которого определяется накопленная частота.

Для вычисления накопленной частоты в таблицу 3 необходимо добавить столбец «Относительная частота» nотнi, которая рассчитывается как отношение соответствующей частоты к объему выборки:

Таким образом, накопленная частота определяется как:

nx1 = nотн1;

nxi+1 = nxi+ nотнi+1, i = 1..k+1.

Для повышения наглядности эмпирических распределений используется их графическое представление. Наиболее распространенными способами графического представления являются гистограмма, полигон частот, полигон накопленных частот (кумулята).

**Гистограмма** используется для графического представления распределений непрерывно варьирующих признаков и состоит из примыкающих друг к другу прямоугольников. Основание каждого прямоугольника – ширина интервала группировки, а высота такова, что площадь прямоугольника пропорциональна частоте попадания в данный интервал. Когда ширина всех интервалов одинакова, по оси ординат можно откладывать частоты интервалов ni.

**Полигон частот** образуется ломаной линией, соединяющей точки, соответствующие серединным значениям интервалов группировки и частотам этих интервалов.

**Полигон накопленных частот (кумулята)** получается при соединении отрезками прямых точек, координаты которых соответствуют верхним границам интервала группировки и накопленным частотам.

Для построения диаграммы в Excel необходимо выделить исходные данные для построения диаграммы и, с помощью вкладки *Вставка* на *Ленте* в группе *Диаграммы*, выбрать нужный вид диаграммы. Используя вкладки *Конструктор* и *Макет* на *Работа с диаграммами* выполнить оформление диаграммы.

**Основные числовые характеристики выборки**

Числовые характеристики выборки дают количественное представление об эмпирических данных и позволяют сравнивать их между собой. Наибольшее практическое значение имеют характеристики положения, рассеяния и асимметрии эмпирических распределений.

**Характеристики положения**

Выборочной характеристикой положения называется найденный по выборке числовой параметр, определяющий положение центра распределения наблюдаемых значений исследуемой случайной величины.

Чаще всего употребляются такие характеристики положения, как среднее арифметическое, медиана и мода.

Среднее арифметическое, или просто среднее, — одна из основных характеристик выборки. Оно представляет собой такое значение признака, сумма отклонений от которого выборочных значений признака равна нулю (с учетом знака отклонения).

Если воспользоваться геометрической интерпретацией, то среднее арифметическое можно определить как точку на оси х, которая является абсциссой центра масс гистограммы.

Для вычисления среднего арифметического в Excel используется функция **СРЗНАЧ()**, а в MathCad – **mean()**.

Медианой называется такое значение признака *x*, когда одна половина значений экспериментальных данных меньше ее, а вторая половина — больше.

Собственно, этим и ограничивается смысловое значение медианы. Широкое использование этой характеристики на практике объясняется простотой ее вычисления и независимостью от формы распределения эмпирических данных. Для тех случаев, когда эмпирическое распределение оказывается сильно асимметричным, среднее арифметическое теряет свою практическую ценность, поскольку при этом значительно большая часть значений признака оказывается выше или ниже среднего арифметического. В этой ситуации медиана представляет собой лучшую характеристику центра распределения.

Для вычисления медианы в Excel используется функция **МЕДИАНА()**, а в MathCad – **median()**.

Мода представляет собой значение признака, встречающееся в выборке наиболее часто.

Интервал группировки с наибольшей частотой называется модальным. Для определения моды используется функция Excel МОДА(), а MathCad – **mode()**. Если выборка не содержит одинаковых данных, то функция МОДА возвращает значение ошибки #Н/Д.

Вообще, среднее, медиана и мода совпадают только в том случае, если распределение унимодальное (с одним максимумом) и симметричное. Чем больше распределе­ние отличается от симметричного, тем сильнее различие между этими характеристиками.

Максимальное и минимальное значения выборки можно подсчитать с помощью функций **МАКС()** и **МИН()** в Excel, а также – **max()** и **min()** в MathCad соответственно.

**Характеристики рассеяния**

Средние значения не дают полной информации о варьирующем признаке. Поэтому наряду со средними значениями вычисляют и характеристики рассеяния выборки.

Размах вариации вычисляется как разность между максимальной и минимальной вариантами выборки:

Размах вариации используется в практических исследованиях при малых (не более 10) объемах выборки. Например, по размаху вариации легко оценить, насколько различаются лучший и худший результаты в выборке. При больших объемах выборки к его использованию надо относиться с осторожностью.

Дисперсия и стандартное отклонение являются важнейшими характеристиками рассеяния.

Дисперсией называется средний квадрат отклонения значений признака от среднего арифметического. Дисперсия, вычисляемая по выборочным данным, называется выборочной дисперсией и обознача­ется S2.

Для вычисления дисперсии в Excel используется функция **ДИСП()**, а в MathCad – **Var()**.

Стандартным отклонением (или средним квдратическим отклонением) называется положительный корень квадратный из дисперсии.

Для вычисления стандартного отклонениея в Excel используется функция **СТАНДОТКЛ()**, а в MathCad – **stdev()**.

Размерность стандартного отклонения в отличие от размерности дисперсии совпадает с единицами измерения варьирующего признака, поэтому в практической статис­тике для характеристики рассеяния используют обычно стандартное отклонение, а не дисперсию.

Асимметрия характеризует степень несимметричности распределения относительно его среднего. Положительная асимметрия указывает на отклонение распределения в сторону положительных значений. Отрицательная асимметрия указывает на отклонение распределения в сторону отрицательных значений.

Для вычисления коэффициента асимметрии в Excel используется функция **СКОС()**, а в MathCad – **skew()**. Если имеется менее трех точек данных или стандартное отклонение равно нулю, функция **СКОС** возвращает значение ошибки #ДЕЛ/0!.

Эксцесс характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением. Положительный эксцесс обозначает относительно остроконечное распределение. Отрицательный эксцесс обозначает относительно сглаженное распределение.

Для вычисления коэффициента эксцесса в Excel используется функция **ЭКСЦЕСС()**, а в MathCad – **kurt()**.

Если задано менее четырех точек данных или если стандартное отклонение выборки равно 0, то функция **ЭКСЦЕСС** возвращает значение ошибки #ДЕЛ/0!.

Определить объем исследуемой выборки с помощью средств Excel можно, используя функцию **СЧЕТ**().

Стандартную ошибку выборки можно вычислить в Excel по следующей формуле:

**=СТАНДОТКЛОН(«выборка»)/КОРЕНЬ(«объем выборки»)**

Сумму значений выборки можно найти, используя функцию Excel **СУММ().**

С помощью приведенных выше функций Excel и MathCad вычислить средствами Excel и MathCad основные числовые характеристики сформированной выборки, заполнить таблицу:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Характеристика функции | Excel | | MathCad | |
| Используемая функция | Полученное значение | Используемая функция | Полученное значение |
|  |  |  |  |  |

Указанные числовые характеристики вычислить также, используя статистическую процедуру «Описательная статистика» пакета *Анализ* *данных* приложения Excel.

Результаты сравнить.

Приложение А

