МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний аерокосмічний університет ім. М. Є. Жуковського

«Харківський авіаційний інститут»

факультет програмної інженерії та бізнесу

кафедра 603

**Лабораторна робота № 2**

з дисципліни «Емпіричні методи програмної інженерії»

# на тему:« Первинна обробка експерементальних данних »

Виконав: студент 2 курсу групи № 622п

напряму підготовки (спеціальності)

121 інженерія програмного забезпечення

(шифр і назва напряму підготовки (спеціальності))

Дудченко Р. О.

(прізвище й ініціали студента)

Прийняв:

ст.викл. Лучшева О.В

(посада, науковий ступінь, прізвище й ініціали)

Національна шкала:

Кількість балів:

Харків 2020

**Цель работы**

Освоение методики выявления грубых погрешностей наблюдений, а также представление экспериментальных данных в графическом виде и вычисление их основных числовых характеристик.

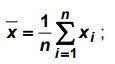
**Постановка задачи**

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 (согласно варианту 8 α1 = 0.08, α2 = 0.25) 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.
8. 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.

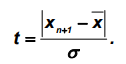
1. Находим число, подозрительное на выброс, и перемещаем его в конец выборки – на (n+1)-е место.
2. Рассчитываем такие характеристики выборки (без числа, подозрительного на выброс):

– эмпирическое среднее



– несмещенную эмпирическую дисперсию

– среднеквадратическое отклонение



1. Вычисляем значение.
2. Находим табличное значение критерия Стьюдента kP t , 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).
3. Если t ≥ tkP , то это выброс, его исключаем или (в зависимости от конкретных условий задачи) заменяем на своем месте величиной, интерполированной по соседним значениям: i-2 i-1 i i x Выброс 21 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 табл. 2.1 – это вероятность того, что подозрительное значение является выбросом, то уровень значимости α=1–Р – это вероятность ошибки 1-го рода, т. е. вероятность принятия решения о том, что xn+1 есть 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 процентах (%), т. е. вероятностям 0,95, 0,99, 0,999 соответствуют уровни 5; 1 и 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 табл. 2.1 можно получить, обращаясь к функции qt(0,975, k).

**Порядок выполнения работы**

1. Сформирован ряд наблюдений случайной величины (Рис.1):

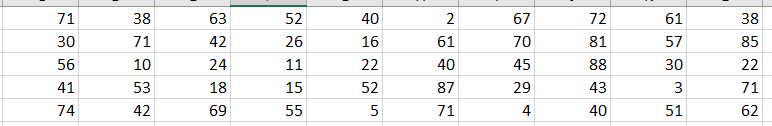


Рис.1

1. Внесено в одно из наблюдений грубую погрешность (max + номер\_варианта) (Рис.2):

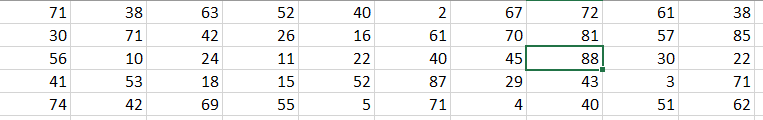


Рис.2

1. Продемонстрировано практическую возможность обнаружения аномальных значений наблюдений с использованием критерия Стьюдента средствами Excel (**Книга2.xlsx – Лист1**):
2. Оформлено полученные значения ряда наблюдений в виде таблицы (подозрительное на выброс число перемещено в конец выборки – на (n+1)-е место) (Рис.3):

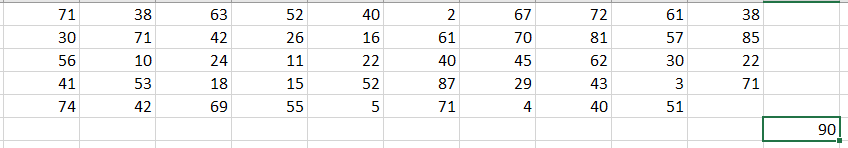


Рис.3

1. Рассчитано значение выборочного среднего для x1, x2, …, xn с использованием функции СРЗНАЧ (Рис.4): ( = 45,52):

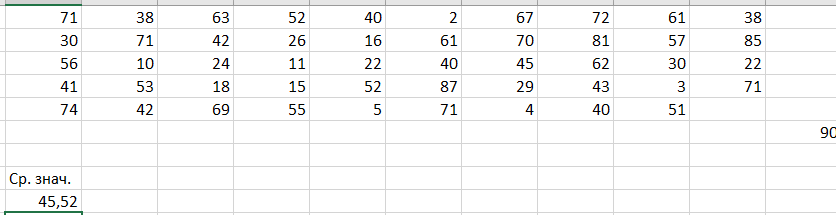


Рис.4

1. Рассчитано несмещенную оценку СКО для x1, x2, …, xn с помощью функции СТАНДОТКЛОН: (σ = 22,96169) (Рис.5):

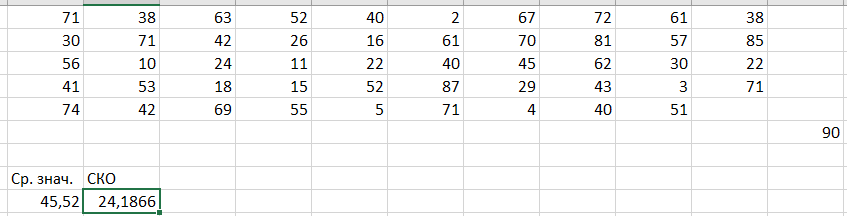
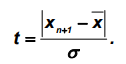


Рис.5

1. Вычислено значение t по формуле (Рис.6):

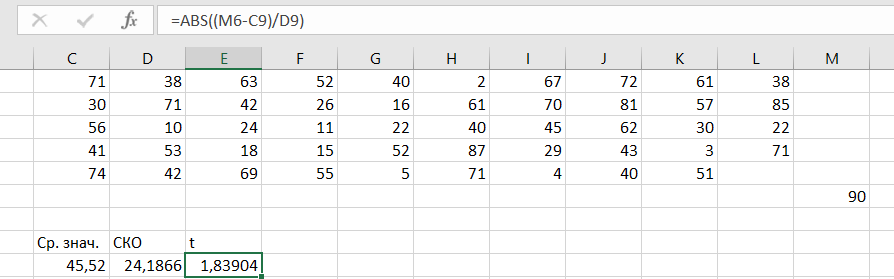


Рис.6

получено значение критерия Стьюдента  с использованием функции СТЬЮДРАСПОБР(α,k) для пятипроцентного (5 %) уровня значимости (k - количество степеней свободы, в нашем случае равное 49 – 1 = 48) (Рис.7):

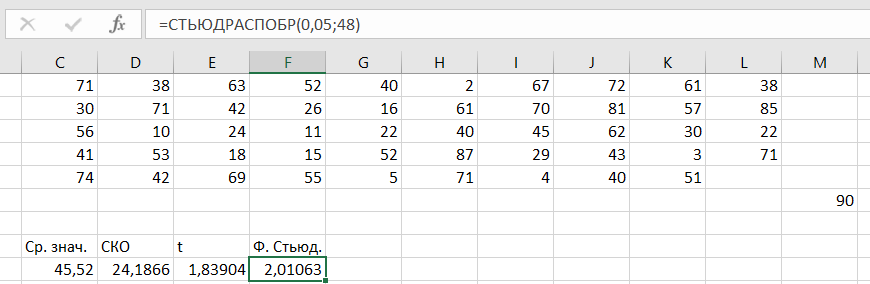


Рис.7

1. Результат сравнения (t и ) отображен на листе Excel с помощью функции ЕСЛИ (Рис.8):

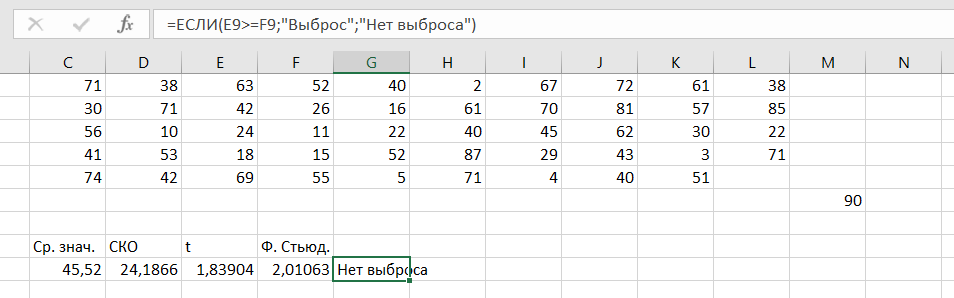


Рис.8

1. Продемонстрировано практическую возможность обнаружения аномальных значений наблюдений с использованием критерия Стьюдента средствами MathCad:
2. представлено полученные значения ряда наблюдений в виде массива (Рис.9):



Рис.9

1. Рассчитано значение выборочного среднего с использованием функции **mean** (Рис.10):

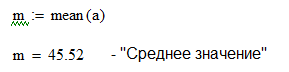


Рис.10

1. Рассчитано несмещенную оценку СКО с помощью функции **Stdev** (Рис.11):

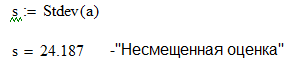
****

Рис.11

1. Вычислено значение **t** (Рис.12):

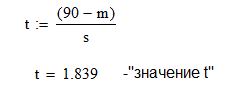
****

Рис.12

1. Вычислено значение критерия Стьюдента с использованием функции **qt** (Рис.13):

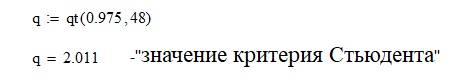
****

Рис.13

1. Сравнено значение t со значением критерия Стьюдента (Рис.14):

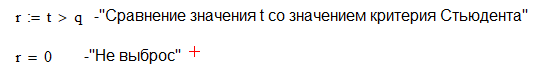
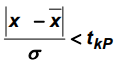
****

Рис.14

1. Определены пределы для заданного уровня значимости (согласно варианту 8 α1 = 0.08, α2 = 0.25) (**LB\_2.xlsx – Лист2**):
2. Найдены tkP1 и tkP2 для k = n–1 и уровней значимости α1 (Рис.15) и α2 (Рис.16):

|  |  |
| --- | --- |
| Рис.15 | Рис.16 |

1. Определены допустимые пределы случайной величины, решая неравенство  для tkP1 верхний предел (Рис.17) и нижний предел (Рис.18):

|  |  |
| --- | --- |
| Рис.17 | Рис.18 |

и для tkP2 верхний предел (Рис.19) и нижний предел (Рис.20):

|  |  |
| --- | --- |
| Рис.19 | Рис.20 |

1. Проиллюстрированы полученные результаты на числовой оси (Рис.21):

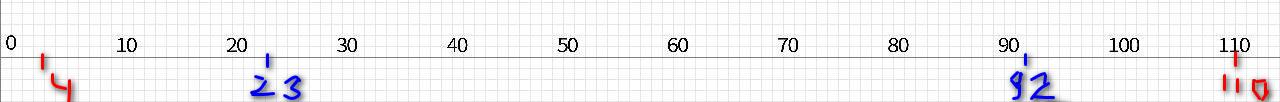


Рис.21

1. Сформулирован вывод о зависимости допустимых пределов от уровня значимости нулевой гипотезы: длина допустимого предела увеличивается при уменьшении уровня значимости нулевой гипотезы.
2. Представлены экспериментальные данные в виде гистограммы, полигона частот и кумуляты средствами Excel (**LB\_2.xlsx - Лист2**):
3. Определена длина шага по формуле , где  (Рис.22):

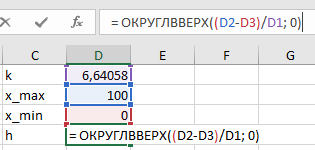
****

Рис.22

1. Определены нижние и верхние пределы для каждого интервала средствами Excel (Рис.23):

****

Рис.23

1. Вычислена частота используя функцию ЧАСТОТА() (Рис.24):



Рис.24

1. Вычислена относительная частота по формуле  (Рис.25):

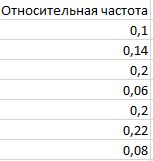


Рис.25

1. Вычислена накопленная частота по формуле  (Рис.26):

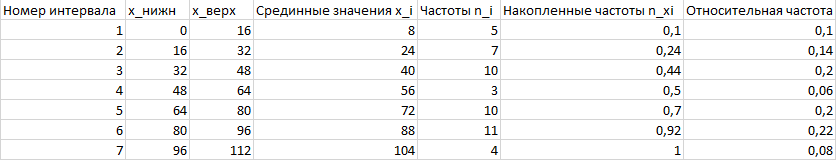


Рис.26

1. Построен полигон частот (образуется ломаной линией, соединяющей точки, соответствующие серединным значениям интервалов группировки и частотам этих интервалов) (Рис.27):



Рис.27

1. Построен полигон накопленных частот, кумулята (получается при соединении отрезками прямых точек, координаты которых соответствуют верхним границам интервала группировки и накопленным частотам) (Рис.28):



Рис.28

1. Построена гистограмма, где основание каждого прямоугольника – ширина интервала группировки, а высота такова, что площадь прямоугольника пропорциональна частоте попадания в данный интервал (Рис.29):

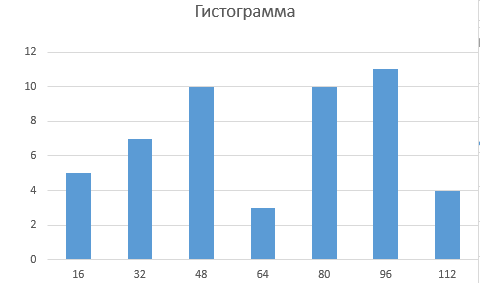


Рис.29

1. Вычислены основные числовые характеристики выборки средствами Excel и MathCad (**LB\_2.xlsx – Лист3**):
2. Вычислены среднее значение, медиана, мода, максимальное и минимальное значения выборки средствами MathCad (Рис.30):

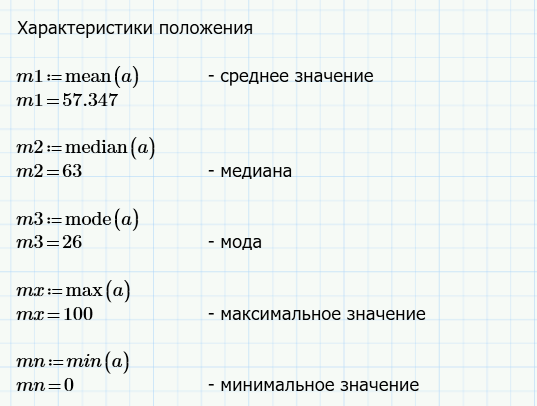


Рис.30

и средствами Excel среднее значение (Рис.31), медиана (Рис.32), мода (Рис.33), минимальное значение (Рис.34), максимальное значение (Рис.35):

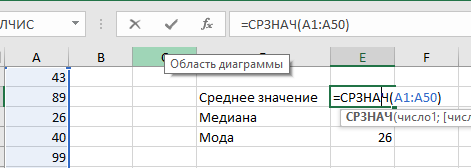


Рис.31

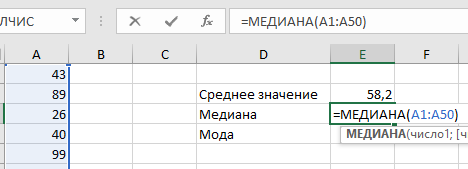


Рис.32

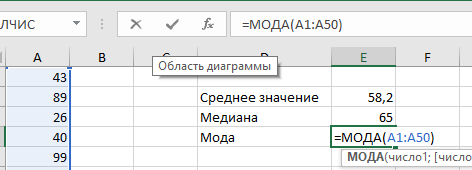


Рис.33

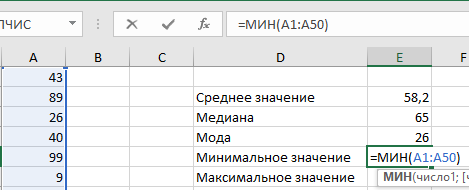


Рис.34

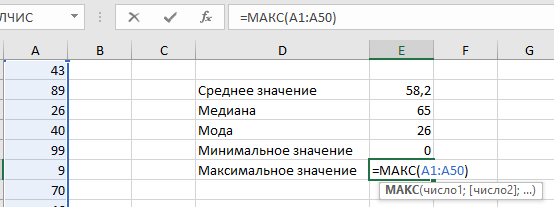


Рис.35

1. Вычислены размах вариации, дисперсия и среднее квадратическое отклонение выборки средствами MathCad (Рис.36):

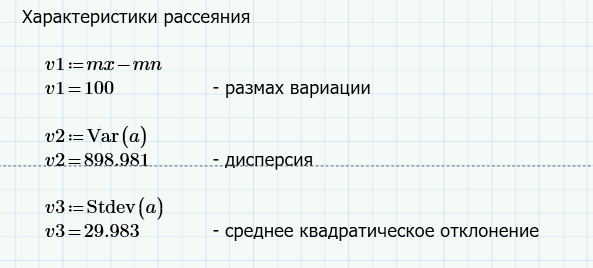
****

Рис.36

и средствами Excel размах вариации (Рис.37), дисперсия (Рис.38), среднее квадратическое отклонение (Рис.39):

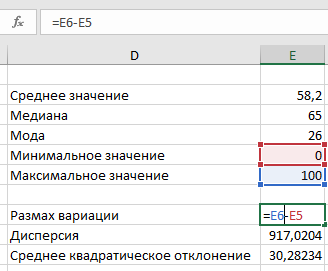
****

Рис.37

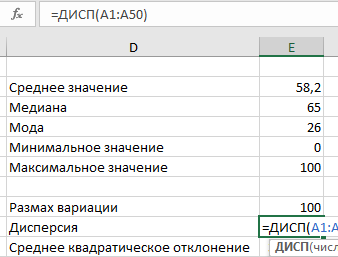
****

Рис.38

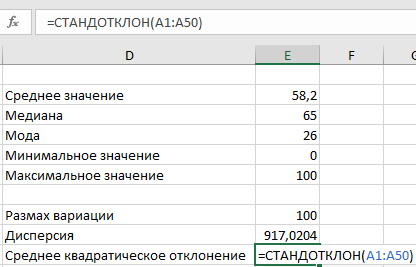


Рис.39

1. Вычислены коэффициент асимметрии, эксцесс, стандартную ошибку выборки средствами MathCad (Рис.40):

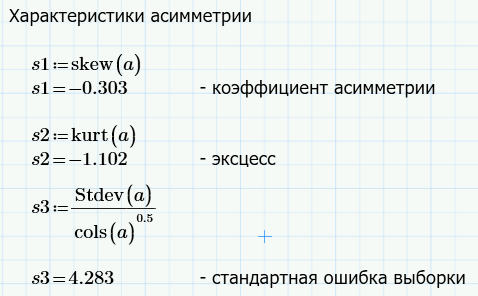
****

Рис.40

и средствами Excel коэффициент асимметрии (Рис.41), эксцесс (Рис.42), стандартная ошибка выборки (Рис.43):

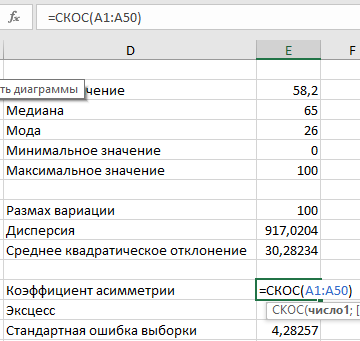
****

Рис.41

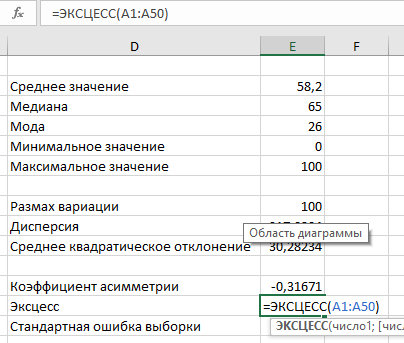


Рис.42

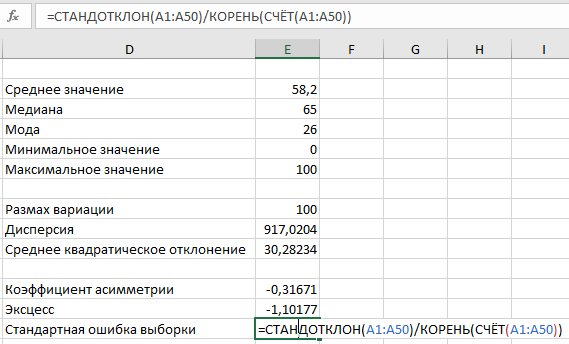


Рис.43

1. Указанные числовые характеристики вычислены также, используя статистическую процедуру «Описательная статистика» пакета Анализ данных приложения Excel (Рис.44):

****

Рис.44

**Выполнение работы**

1. Практическая возможность обнаружения аномальных значений наблюдений с использованием критерия Стьюдента средствами Excel (**LB\_2.xlsx – Лист1**) (Рис.45):

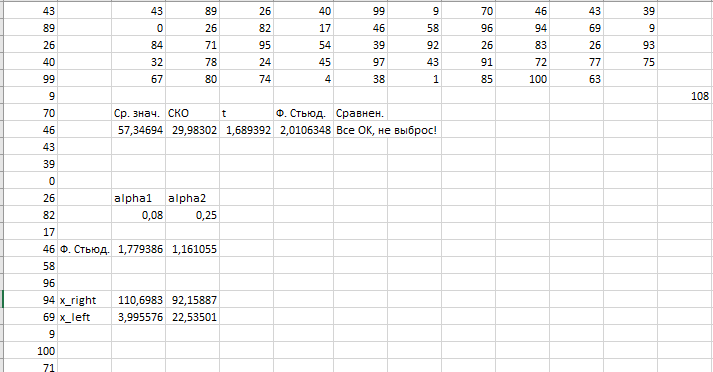
****

Рис.45

1. Практическая возможность обнаружения аномальных значений наблюдений с использованием критерия Стьюдента средствами MathCad (Рис.46):

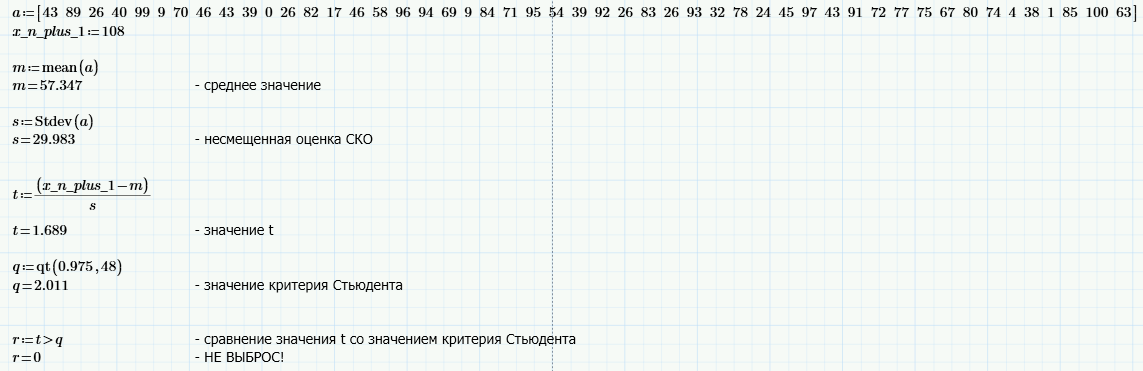
****

Рис.46

1. Для уровней значимости (согласно варианту 8 α1 = 0.08, α2 = 0.25) определены пределы, в которых значение выборки не будет считаться выбросом из нее (**LB\_2.xlsx – Лист1**) (Рис.47):

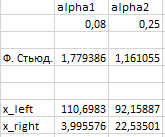
****

Рис.47

1. Экспериментальные данные в виде гистограммы, полигона частот и кумуляты средствами Excel (**LB\_2.xlsx – Лист2**) (Рис.48):

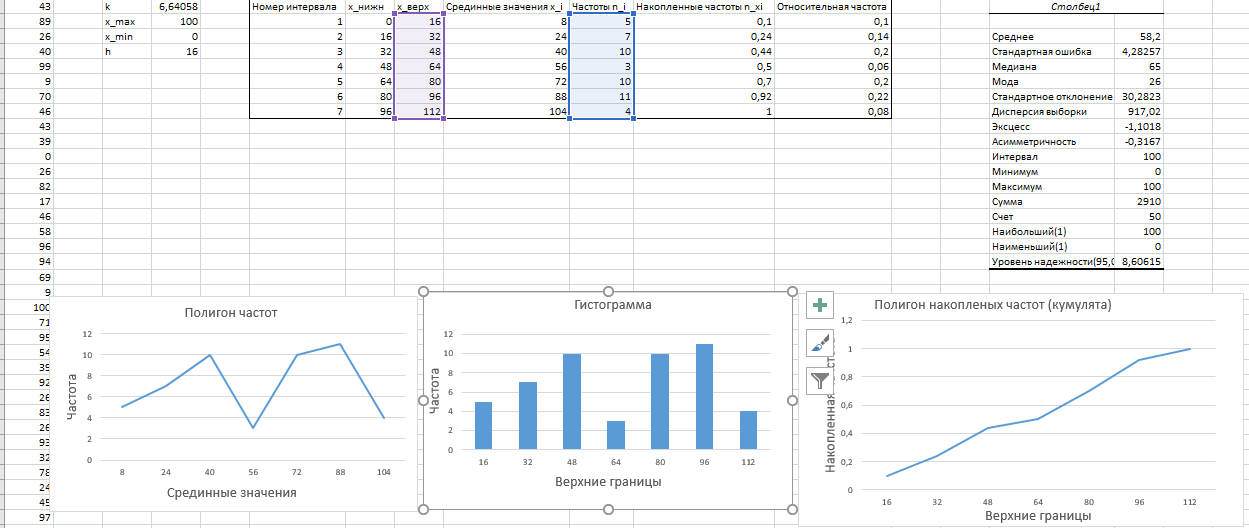
****

Рис.48

1. Основные числовые характеристики выборки средствами Excel (**LB\_2.xlsx – Лист3**) и MathCad (Таблица 1)**:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Характеристика функции | Excel | | MathCad | | Excel (Опсательная статистика) |
| Используемая функция | Полученное значение | Используемая  функция | Полученное значение | Полученное значение |
| Среднее значение | СРЗНАЧ(выборка) | 57.88 | mean(выборка) | 57.88 | 57.88 |
| Медиана | МЕДИАНА  (выборка) | 65 | Median  (выборка) | 65 | 65 |
| Мода | МОДА(выборка) | 26 | mode(выборка) | 26 | 26 |
| Минимальное значение | МИН(выборка) | 0 | min(выборка) | 100 | 0 |
| Максимальное значение | МАКС(выборка) | 100 | max(выборка) | 0 | 100 |
| Размах вариации | МАКС(выборка) – МИН(выборка) | 100 | max(выборка) – min(выборка) | 100 | 100 |
| Дисперсия | ДИСП(выборка) | 894.8424 | Var(выборка) | 894.842 | 894.8424 |
| СКО | СТАНДОТКЛОН  (выборка) | 29.9139 | Stdev(выборка) | 29.914 | 29.9139 |
| Коэффициент  асимметрии | СКОС(выборка) | -0.33949 | skew(выборка) | -0.339 | -0.33949 |
| Эксцесс | ЭКСЦЕСС(выборка) | -1.08856 | kurt(выборка) | -1.089 | -1.08856 |
| Стандартная ошибка выборки | СТАНДОТКЛОН  (выборка)/  КОРЕНЬ(СЧЁТ  (выборка)) | 4.230467 | Stdev(выборка)/  (cols(выборка)^ 0.5) | 4.23 | 4.230467 |

Таблица 1

**Вывод**

Была освоена методика выявления грубых погрешностей наблюдений, а также представление экспериментальных данных в графическом виде и вычисление их основных числовых характеристик.

Продемонстрирована практическую возможность обнаружения аномальных значений наблюдений с использованием критерия Стьюдента средствами MathCad и Excel, внесенная грубая погрешность выбросом не оказалась.

Были вычислены основные числовые характеристики выборки средствами Excel и MathCad, результаты оказались одинаковыми.

Был сформулирован вывод о зависимости допустимых пределов от уровня значимости нулевой гипотезы: длина допустимого предела увеличивается при уменьшении уровня значимости нулевой гипотезы.