**Министерство образования Республики Беларусь**

Учреждение образования

БелорусскиЙ государственный университет

информатики и радиоэлектроники

Факультет компьютерных систем и сетей

Кафедра программного обеспечения информационных технологий

# оТЧЕТ

по расчетной работе

на тему:

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ

|  |  |  |
| --- | --- | --- |
| Выполнил  Студент гр. 651005 |  | И. Л. Савенок |
| Проверил |  | Н. В. Лапицкая |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Минск, 2018

**СОДЕРЖАНИЕ**

1 Постановка задачи..................................................................................... 2

2 Используемые формулы............................................................................ 3

2.1 Дискретные данные............................................................................ 3

2.2 Непрерывные данные……................................................................. 4

2.3 Общие формулы.................................................................................. 5

2.4 Критерии согласия.............................................................................. 6

3 Анализ данных........................................................................................... 7

3.1 Преобразование данных..................................................................... 7

2.2 Определение класса величины.......................................................... 7

2.3 Расчет числовых характеристик........................................................ 8

2.3 Анализ качества разбиения на группы............................................. 10

2.3 Выдвижение гипотезы........................................................................ 11

4 Выводы....................................................................................................... 12

**1 ПОСТАНОВКА ЗАДАЧИ**

Целью исследовательской работы ставлю нахождение данных с их последующими обработкой и анализом таким образом, чтобы выдвинуть и проверить гипотезу о виде распределения данных.

В обработку включаются следующие шаги:

* экспорт с организацией структуры для хранения;
* очистка данных;
* установление класса величины;
* подсчет основных характеристик;
* построение таблиц распределения;
* анализ качества разбиения на группы (для непрерывных данных);
* выдвижение о виде распределения;
* подсчет теоретической функции, частоты распределения;
* сравнение теоретических и практических значений.

В данной исследовательской работе рассматриваются следующие данные: минимальная и максимальная дневные температуры, дневное количество осадков в Минске за 1992 год.

**2 ИСПОЛЬЗУЕМЫЕ ФОРМУЛЫ**

**2.1 Дискретные данные**

Математическое ожидание:

где:

*n* - количество всех вариант, т.е. объём выборки;

*xi* – значение одной варианты.

Дисперсия:

где:

– математическое ожидание;

*n* – количество всех вариант, т.е. объём выборки;

*xi* – значение одной варианты.

Так как осуществляется проверка гипотезы на распределение Пуассона, используется соответствующая формула для нахождения теоретических частот:

где:

– число наблюдений;

*x* – значение варианты;

– средняя величина;

– основание натурального логарифма.

При этом теоретические частоты для интервалов определяются как:

где:

– объем выборки n;

– вероятности, соответствующие конкретному закону распределения, вычисленные по выборочным данным.

В качестве предварительной проверки на распределение Пуассона имеем:

где:

– выборочная дисперсия;

– несмещенная оценка генеральной средней.

При этом выборочная дисперсия определяется как

где:

*x* – значение варианты;

– средняя величина;

– объем выборки.

Несмещенная оценка генеральной средней вычисляется по формуле

где:

*x* – значение варианты;

– частота варианты;

– объем выборки.

**2.2 Непрерывные данные**

Для разбиения на интервалы используется формула Стерджесса:

где:

*xmax* – максимальное значение выборки;

*xmin* – минимальное значение выборки;

*N* – количество всех вариант, т.е. объём выборки.

Математическое ожидание:

где:

*ni* – сумма частот каждого интервала;

*n* – сумма частот всех интервалов;

*xi* – середина интервала.

Дисперсия:

где:

– математическое ожидание;

*ni* – сумма частот каждого интервала;

*n* – сумма частот всех интервалов;

*xi* – середина интервала.

После определения основных характеристик, необходимо осуществить анализ качества разбиения на группы. Для этого используется теорема сложений дисперсий:

где

- общая дисперсия;

– средняя из групповых дисперсий;

– межгрупповая дисперсия.

Определение общей дисперсии:

где:

– середина интервала;

– средняя величина;

*n* – количество всех интервалов.

Определение сриедней из групповых дисперсий:

где:

– среднее квадратичное отклонение;

*fi* – частота интервала.

Определение межгрупповой дисперсии:

где:

– середина интервала;

– средняя величина;

*fi* – частота интервала.

При этом для определения средней величины воспользуемся формулой:

где:

– середина интервала;

*fi* – частота интервала.

Так как осуществляется проверка гипотезы на нормальное распределение, используется соответствующая формула для нахождения частот:

где:

– число наблюдений;

– величина интервала;

– среднее квадратичное отклонение;

– табулированная величина, отыскиваемая в таблице по отклонениям.

**2.3 Общие формулы**

Среднее квадратичное отклонение:

**2.4 Критерии согласия**

Определяем практическое значение :

где:

– практическая частота интервала;

– теоретическая частота интервала.

Для определения теоретических частот воспользуемся формулой

По критерию Пирсона справедливо неравенство

*χ*2набл*> χ2*d,α

где

– табличное критическое значение, в котором d – число степеней свободы d=n-r-1, n – число интервалов, r – число параметров (нормальный r =2, пуассоновский r=1).

По критерию Романовского вычисляется величина R:

где:

– практическое значение;

– число степеней свободы.

Если R < 3, то гипотеза Н0 принимается.

По критерию Ястремского вычисляется величина J:

где:

n – число групп в вариационном ряду;

– число степеней свободы.

Если J < 3, то гипотеза Н0 принимается.

**3 АНАЛИЗ ДАННЫХ**

**3.1 Преобразование данных**

Изначально у нас имеется файл с неочищенными файлами, полученный с сайта «Геоинформационная система "Метео измерения онлайн"» в следующем формате (см. рисунок 1):

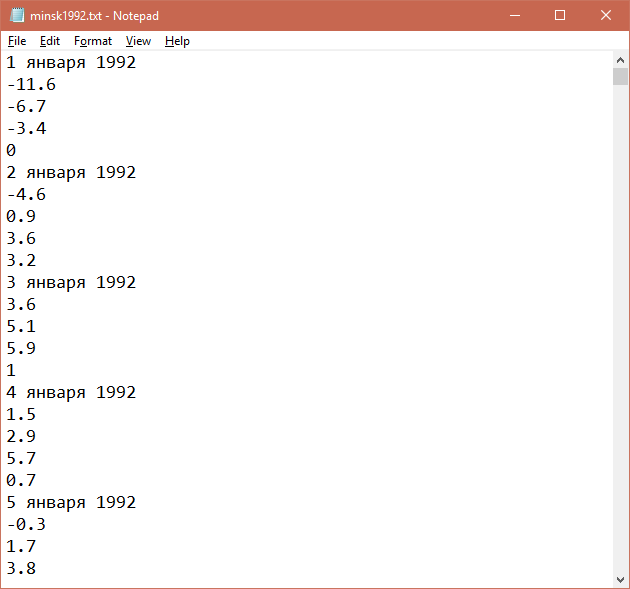


Рисунок 1. Файл с неочищенными данными

Для того, чтобы с данными можно было работать, был написан специальный парсер, осуществляющий их очистку, удаление не интересующих элементов. Впоследствии преобразованные данные помещаются в структуру для анализа, представляющую собой списки элементов. Результат преобразования на рисунке 2.

**3.2 Определение класса величины**

Следующим шагом была описан разработан анализатор, позволяющий определить, относятся данные к дискретным или непрерывным. В основе принципа лежит вычисление отношения уникальных элементов к количеству всех записей.

В итоге, **минимальные и максимальные значения температур** были отнесены к классу **непрерывных** данных, в то время как **количество осадков** причислено к классу **дискретных** данных.

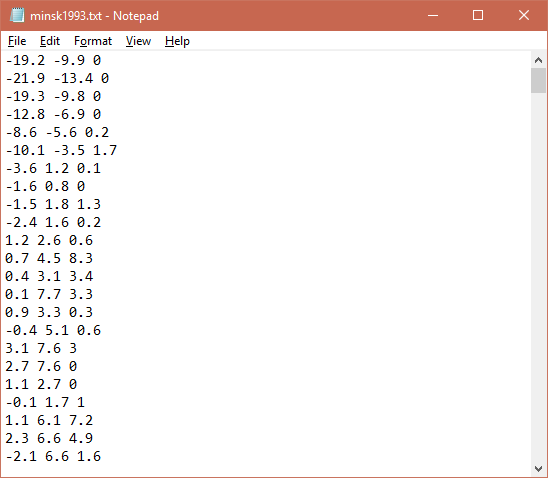


Рисунок 2. Преобразованные данные

**3.3 Расчет числовых характеристик**

Были разработаны специальные алгоритмы для определения математического ожидания, дисперсии, среднего квадратичного отклонения для каждой величины.

Установлены следующие значения:

* Минимальные температуры

Математическое ожидание: 3.41

Дисперсия: 63.00

Среднее квадратичное отклонение: 7.94

* Максимальные температуры

Математическое ожидание: 11.05

Дисперсия: 110.10

Среднее квадратичное отклонение: 10.49

* Количество осадков

Математическое ожидание: 1.99

Дисперсия: 27.51

Среднее квадратичное отклонение: 5.25

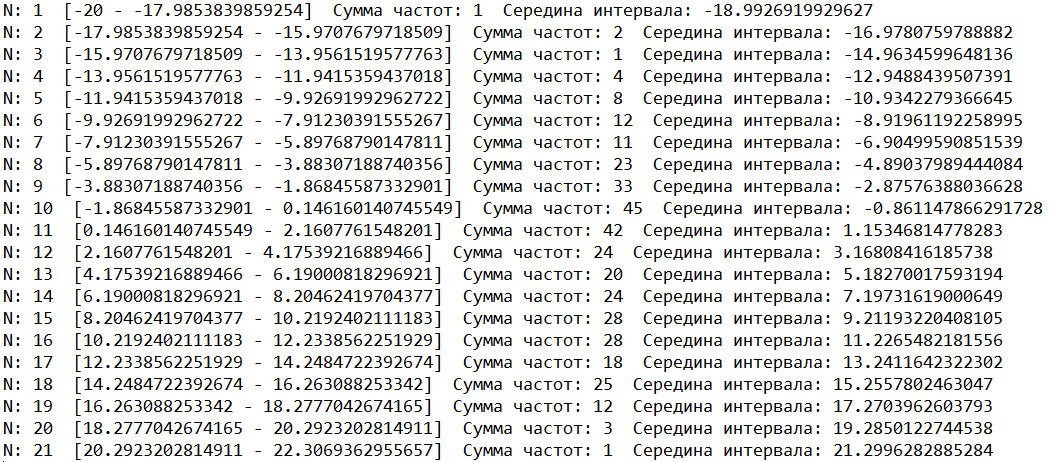


Рисунок 3. Экспортированная таблица распределения для НСВ (минимальные температуры)

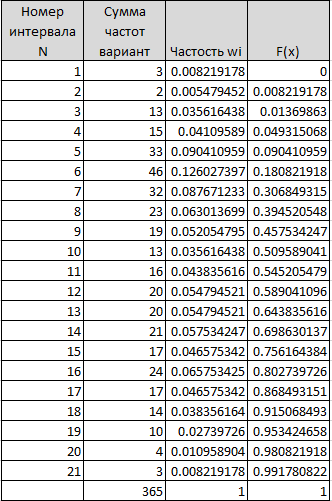


Рисунок 4. Значения частости, функции распределения для интервалов для НСВ (минимальные температуры)

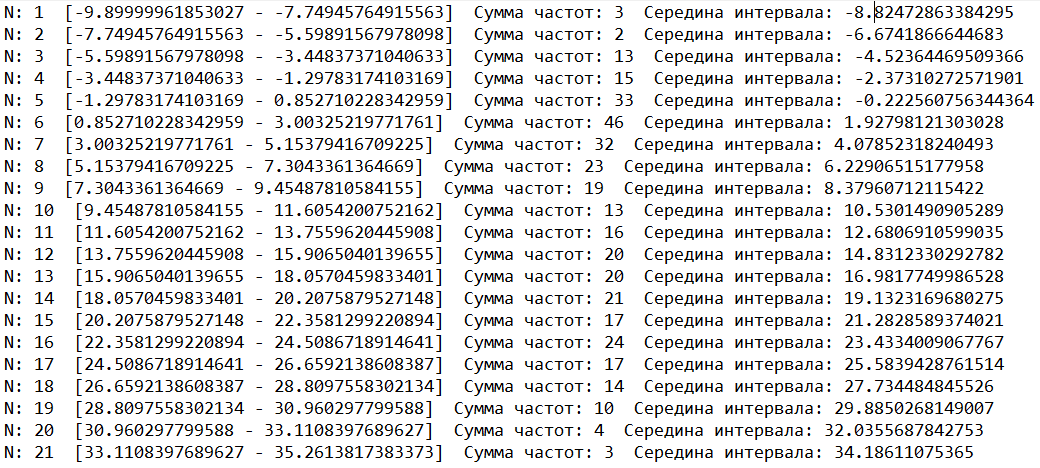


Рисунок 5. Экспортированная таблица распределения для НСВ (максимальные температуры)

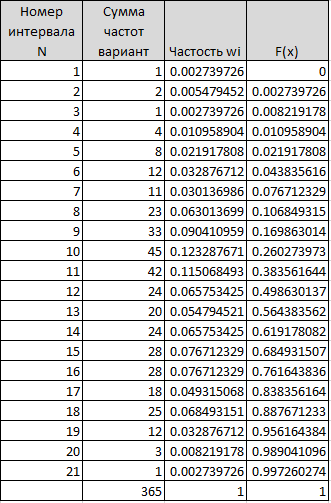


Рисунок 6. Значения частости, функции распределения для интервалов для НСВ (максимальные температуры)

Для дневного количества осадков за год были получены следующие результаты:

Значение: 0 Частота: 0.4986301

Значение: 0.1 Частота: 0.02191781

Значение: 0.2 Частота: 0.01917808

Значение: 0.3 Частота: 0.03013699

Значение: 0.4 Частота: 0.02191781

Значение: 0.5 Частота: 0.03013699

Значение: 0.6 Частота: 0.01643836

Значение: 0.7 Частота: 0.005479452

Значение: 0.8 Частота: 0.0109589

Значение: 0.9 Частота: 0.008219178

Значение: 1 Частота: 0.0109589

Значение: 1.1 Частота: 0.01643836

Значение: 1.2 Частота: 0.0109589

Значение: 1.3 Частота: 0.005479452

Значение: 1.4 Частота: 0.008219178

Значение: 1.5 Частота: 0.005479452

Значение: 1.6 Частота: 0.0109589

Значение: 1.7 Частота: 0.008219178

Значение: 1.8 Частота: 0.008219178

Значение: 2 Частота: 0.008219178

Значение: 2.1 Частота: 0.002739726

Значение: 2.2 Частота: 0.005479452

Значение: 2.3 Частота: 0.002739726

Значение: 2.4 Частота: 0.008219178

Значение: 2.6 Частота: 0.005479452

Значение: 2.7 Частота: 0.002739726

Значение: 2.8 Частота: 0.008219178

Значение: 2.9 Частота: 0.002739726

Значение: 3 Частота: 0.005479452

Значение: 3.1 Частота: 0.002739726

Значение: 3.3 Частота: 0.002739726

Значение: 3.4 Частота: 0.005479452

Значение: 3.5 Частота: 0.008219178

Значение: 3.6 Частота: 0.002739726

Значение: 3.7 Частота: 0.005479452

Значение: 3.9 Частота: 0.002739726

Значение: 4 Частота: 0.005479452

Значение: 4.1 Частота: 0.002739726

Значение: 4.3 Частота: 0.0109589

Значение: 4.5 Частота: 0.005479452

Значение: 4.6 Частота: 0.002739726

Значение: 4.7 Частота: 0.002739726

Значение: 4.8 Частота: 0.002739726

Значение: 4.9 Частота: 0.002739726

Значение: 5.1 Частота: 0.002739726

Значение: 5.4 Частота: 0.002739726

Значение: 5.5 Частота: 0.002739726

Значение: 5.8 Частота: 0.005479452

Значение: 5.9 Частота: 0.005479452

Значение: 6 Частота: 0.002739726

Значение: 6.1 Частота: 0.005479452

Значение: 6.2 Частота: 0.002739726

Значение: 6.3 Частота: 0.002739726

Значение: 6.4 Частота: 0.002739726

Значение: 6.5 Частота: 0.002739726

Значение: 6.6 Частота: 0.005479452

Значение: 6.7 Частота: 0.002739726

Значение: 6.8 Частота: 0.005479452

Значение: 7 Частота: 0.008219178

Значение: 7.2 Частота: 0.005479452

Значение: 7.4 Частота: 0.002739726

Значение: 7.6 Частота: 0.005479452

Значение: 7.8 Частота: 0.008219178

Значение: 7.9 Частота: 0.002739726

Значение: 8.1 Частота: 0.005479452

Значение: 8.3 Частота: 0.002739726

Значение: 9.1 Частота: 0.002739726

Значение: 9.5 Частота: 0.002739726

Значение: 9.8 Частота: 0.002739726

Значение: 10 Частота: 0.002739726

Значение: 10.4 Частота: 0.002739726

Значение: 11 Частота: 0.002739726

Значение: 11.5 Частота: 0.002739726

Значение: 12.8 Частота: 0.005479452

Значение: 14.4 Частота: 0.002739726

Значение: 14.6 Частота: 0.002739726

Значение: 17.5 Частота: 0.002739726

Значение: 23 Частота: 0.002739726

Значение: 57.9 Частота: 0.002739726

Значение: 60.4 Частота: 0.002739726

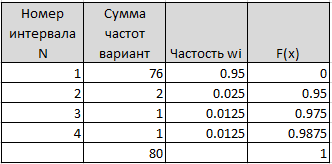


Рисунок 7. Значения частости, функции распределения для интервалов для ДСВ (количество осадков)

**3.4 Анализ качества разбиения на группы (для непрерывных данных)**

Следует отметить, что для случая с максимальными температурами отношение межгрупповой дисперсии к общей дисперсии **больше 60%** (правило сложения дисперсий соблюдается), что свидетельствует о высоком качестве разбиения на группы: 232.302682812196 / 330.377053698249 = 0.703144120367301 (> 0.6).

Для случая с минимальными температурами отношение межгрупповой дисперсии к общей дисперсии оказалось **меньше 60%**, что говорит о недостаточно высоком качестве разбиения на группы: 74.6001924136937 / 150.14867052069 = 0.496842177523071 (< 0.6).

**3.5 Выдвижение гипотезы**

Проверки на приблизительное равенство с нулем значений As и Ex, равные -0.036 и -0.62 соответственно для минимальных температур, а также 0.29 и -1.09 для минимальных температур, не дали веских оснований считать распределение близким к нормальному. Тем не менее, в результате анализа гистограмм частот, графиков функций распределения минимальных и максимальных температур, было решено выдвинуть гипотезу о **нормальном** распределении. Кроме того, вид графиков дает основания выдвинуть гипотезу о **равномерном** распределении.

Для нормального распределения в случае минимальных температур было получено:

Pirson: гипотеза отклонена

X2obs = 43.5856778648129; X2da = 28.9

Romanovsky: гипотеза отклонена

R = 4.26427964413548

Yastremsky: гипотеза отклонена

J = 3.83977204595655

Для нормального распределения в случае максимальных температур было получено:

Pirson: гипотеза отклонена

X2obs = 89.1715064899343; X2da = 28.9

Romanovsky: гипотеза отклонена

R = 11.8619177483224

Yastremsky: гипотеза отклонена

J = 10.6810678432132

Для равномерного распределения в случае минимальных температур было получено:

Pirson: гипотеза отклонена

X2obs = 6806.5214500704; X2da = 28.9

Romanovsky: гипотеза отклонена

R = 1131.4202416784

Yastremsky: гипотеза отклонена

J = 1018.78773879214

Для равномерного распределения в случае максимальных температур было получено:

Pirson: гипотеза отклонена

X2obs = 6806.5217730316; X2da = 28.9

Romanovsky: гипотеза отклонена

R = 1131.42029550527

Yastremsky: гипотеза отклонена

J = 1018.78778726056



Рисунок 8. Гистограмма частот (минимальные температуры)

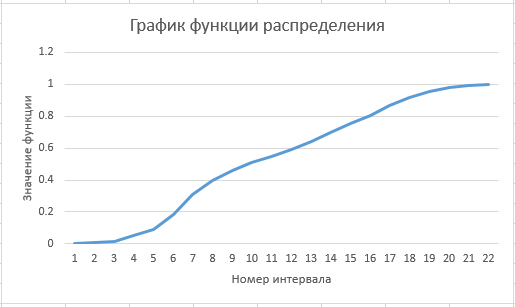


Рисунок 9. График функции распределения НСВ (минимальные температуры)



Рисунок 10. Гистограмма частот (максимальные температуры)



Рисунок 11. График функции распределения НСВ (максимальные температуры)

Предварительный анализ значений математического ожидания и дисперсии не дал веских оснований считать распределение близким к геометрическому. Тем не менее, в результате анализа графиков частот количества осадков, было решено выдвинуть гипотезу о **пуассоновском** распределении.

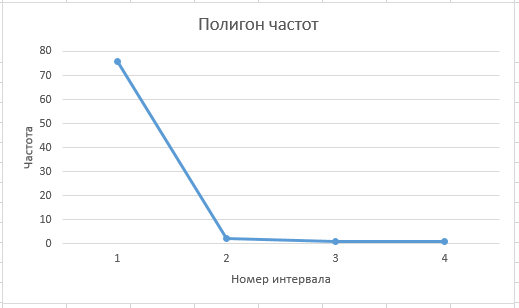


Рисунок 12. Полигон частот (количество осадков)

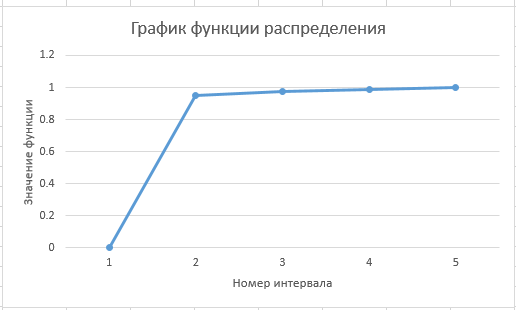


Рисунок 13. График функции распределения ДСВ (количество осадков)

Для пуассоновского распределения были получены следующие результаты:

Pirson: гипотеза отклонена

X2obs = 9.06124517421361; X2da = 6

Romanovsky: гипотеза отклонена

R = 3.53062258710681

Yastremsky: гипотеза принята

J = 0.554100163727708

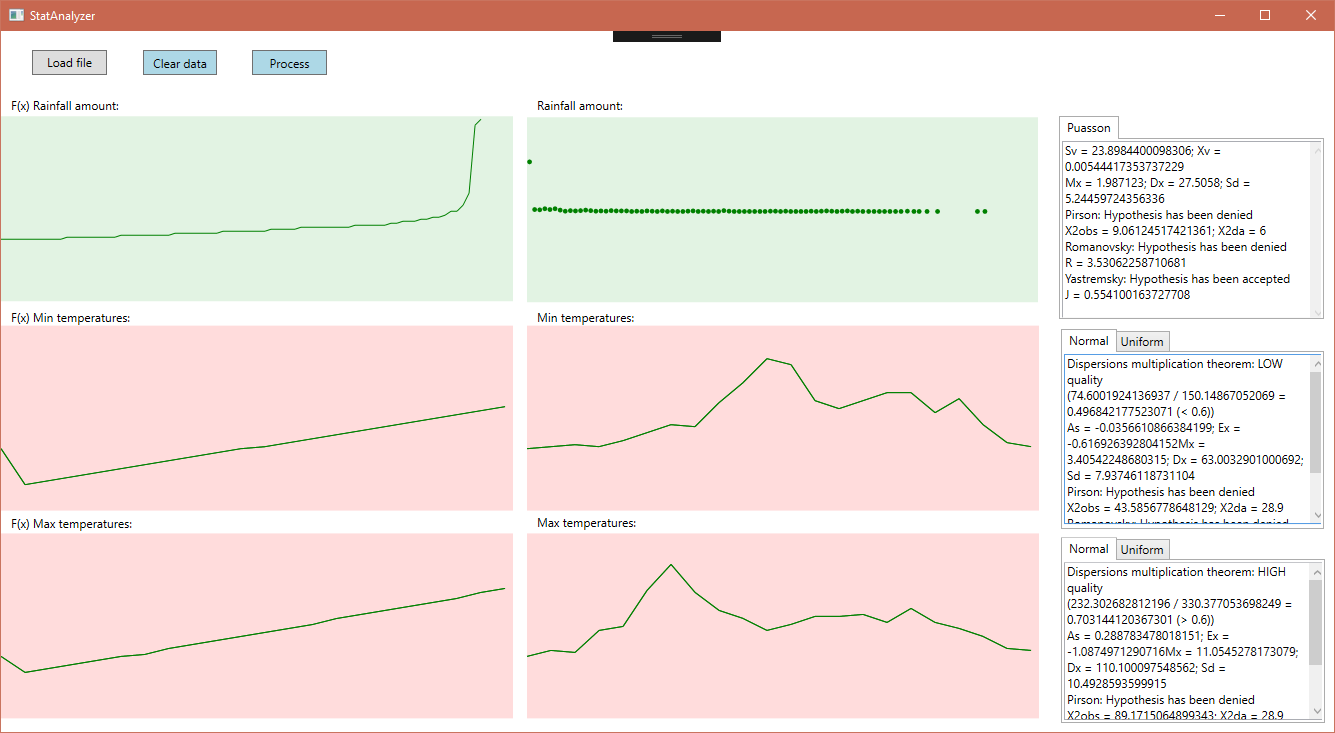


Рисунок 12. Результат работы программы

**4 ВЫВОДЫ**

По итогу вычислений теоретических частот для интервалов минимальных и максимальных значений температур, критериев Пирсона, Романовского, Ястремского были получены результаты, сообщающие, что гипотеза Н0 о том, что этот ряд подчиняется закону нормального распределения, отклоняется (χ2набл> χ2d,a для критерия Пирсона). Причиной этому являются сами данные, отражающие параметры реальных погодных характеристик, которые несмотря на то, что были очищены, все равно не подчиняются идеальному закону и имеют существенные отклонения в значениях частот интервалов.

В результате вычисления критериев Пирсона, Романовского, Ястремского для дискретной величины (количество осадков) и получения верных равенств соблюдения данных критериев, я пришел к выводу, что гипотеза Н0 о том, что ряд подчиняется закону геометрического распределения, принимается.