УДК 517.18

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В СРЕДЕ WXMAXIMA

Т. М. Головко, А. О. Захарова, Б. А. Акишин

Донской государственный технический университет, г. Ростов-на-Дону, Российская Федерация

akiboralex@mail.ru

Рассматриваются возможности и целесообразность использования некоммерческой системы компьютерной математики Maxima при решении задач математической статистики, отмечается наличие множества встроенных функций, позволяющих рассчитывать основные показатели описательной статистики, функций распределения вероятностей, группирование вариационных рядов, осуществление графического сопровождения, проверку статистических гипотез и другие.

Ключевые слова: система компьютерной математики, случайная величина, математическая статистика, функция распределения, вариационный ряд, графический интерфейс, проверка статистических гипотез.

UDC 517.18

STATISTICAL DATA ANALYSIS IN WXMAXIMA ENVIRONMENT

T. M.Golovko, A. O. Zakharova, B. A. Akishin

Don State Technical University, Rostov-on-Don, Russian Federation

akiboralex@mail.ru

The article considers the possibilities and the feasibility of using non-commercial computer algebra system Maxima in solving tasks of mathematical statistics. The paper notes the presence of many built-in functions, which allows calculating the main indicators of descriptive statistics, probability distribution functions, variational series grouping, the implementation of graphics support, the verification of statistical hypotheses and others.

Keywords: system of computer mathematics, random variable, mathematical statistics, distribution function, variational series, graphic interface, verification of statistical hypotheses

Введение. Программа Maxima в графическом интерфейсе wxMaxima содержит ряд встроенных функций для решения задач теории вероятностей, математической статистики и статистического анализа [1]. Возможности данной программы в этих вопросах уступают таким специализированным пакетам, как, Statistika. Стандартные процедуры статистического оценивания Maxima выполняет достаточно просто и надежно. Открытость программного кода позволяет легко создавать новые, необходимые пользователю функции.

Статистический анализ в среде wxMaxima. Возможности программы Maxima можно проиллюстрировать на примере анализа выборки данных непрерывной случайной величины.

Проведем статистическую обработку выборки из N=100 значений случайной величины X, представляющей собой рост (в сантиметрах) посетителей магазина мужской одежды [2].[1] Исход-

[1] В пособии [2] аналогичные статистические расчеты были проведены в Excel и MathCAD

ные данные хранятся в текстовом файле.

1. Ввод-вывод матричных данных. При обработке больших объемов исходных данных, используемых в математической статистике, весьма важными процедурами являются организация ввода и хранения этих данных.

Для чтения и записи матричных или потоковых данных в составе Maxima предусмотрен пакет numericalio. Основные функции этого пакета:

— read_matrix (file_name, separator_flag) — считывание матрицы из файла с именем file_name;

— read_list (file_name, separator_flag) — считывание списка из файла;

— write_data (object, file_name, separator_flag) — осуществление вывода объекта (списка, матрицы, массива) в файл.

Список можно преобразовать в матрицу функцией matrix, а матрицу в список — функцией makelist [3].

Для разделения данных в потоке используются различные символы, определяемые параметром separator_flag: comma — запятая, pipe — вертикальная черта, simicolon — двоеточие, tab — символ табуляции, space — пробел и др. По умолчанию разделителем считается пробел.

Итак, подключается пакет numericalio и загружаются данные для рассматриваемого примера из текстового файла в список wxMaxima с помощью функции read_list :

```
(%i3) load(descriptive)$ load(numericalio)$
      A : read_list (file_search ("F:\\example_1.txt"));
(%o3) [154,164,180,179,169,168,177,150,177,166,158,182,157,159,197,
177,187,178,184,169,190,146,178,171,181,186,172,186,156,197,190,176,
176,159,162,174,171,186,156,182,172,175,181,163,170,176,171,179,158,
186,156,183,170,189,167,187,182,185,173,187,175,188,171,170,164,161,
183,178,182,178,176,144,159,174,159,171,172,167,158,180,182,176,161,
175,178,190,178,170,177,183,180,161,178,169,172,176,164,173,172,181]

(%i4) length (A);
(%o4) 100
```

Рис.1. Загрузка данных из внешнего файла

Здесь функция length вычисляет длину списка — 100 чисел.

2. Расчёт описательной статистики. Функции программы Maxima для расчета так называемой описательной статистики объединены в пакет descriptive [4].

Таблица 1.

Основные функции пакета descriptive

mean(list), mean(matrix)[2]	Вычисление среднего
cor(matrix)	Вычисляет корреляционную матрицу
cov(matrix)	Вычисляет ковариационную матрицу
median(list), median(matrix)	Вычисляет медиану
var(list), var(matrix)	Вычисляет дисперсию случайной величины
std(list), std(matrix),	Вычисляет среднеквадратичное отклонение
skewness(list), skewness(matrix)	Вычисление асимметрии
kurtosis(list), kurtosis(matrix)	Вычисление эксцесса
quantile(list, p), quantile(matrix, p)	Вычисление p-квантиля
maxi(list), maxi(matrix), mini(list), mini(matrix)	Выбор наибольшего и наименьшего значения в выборке соответственно
range(list), range(matrix)	Размах вариации выборки

Простым способом расчета многих описательных статистик является использование панели инструментов — Maxima → Панели → Статистика.

Рассчитаем численно (с флагом numer [3]) некоторые показатели для выборки A рассматриваемого примера:

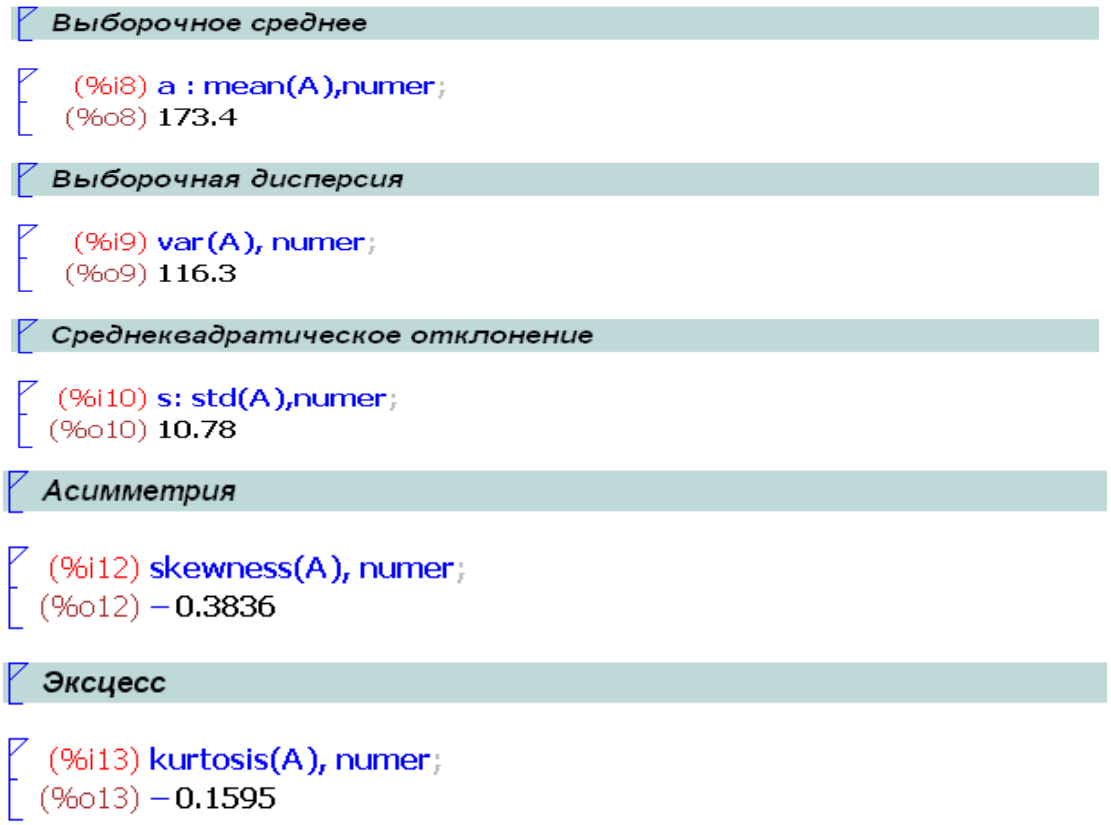

Рис.2. Расчет основных статистических показателей

3. Построение вариационного ряда. Для группировки данных в виде вариационных рядов в пакете descriptive заложены две функции [4]:

[2] Если в качестве входного параметра указана матрица, то вычисление соответствующей характеристики осуществляется отдельно по каждому столбцу

— discrete_freq (list) — создание дискретного вариационного ряда для выборки дискретной случайной величины, представленной списком list;

— continuous_freq (list, m) — создание интервального вариационного ряда для выборки непрерывной случайной величины. Функция делит диапазон от наименьшего до наибольшего значений в списке list на m интервалов и подсчитывает частоты. На выходе формируются два списка — список значений границ интервалов и список частот.

Зададим число интервалов $m = 8$ и рассчитаем интервальный вариационный ряд для рассматриваемого примера: [3]

```
Интервальный вариационный ряд

(%i14) X_n : continuous_freq  (A,8), numer;
(%o14) [[144,150.6,157.3,163.9,170.5,177.1,183.8,190.4,197.0],[3,5,12,14,
27,23,14,2]]
```

Рис.3. Создание вариационного ряда

В используемой версии wxMaxima 15.08.02 не удалось найти готовых встроенных функций или отдельных пакетов для расчета характеристик вариационного ряда, поэтому, используя встроенный макроязык программирования, необходимо создать собственные функции для расчета взвешенного среднего и среднеквадратического отклонения интервального вариационного ряда по формулам:

$$\overline{Xv} = \frac{1}{N}\sum_{i=1}^{m}\frac{x_i + x_{i-1}}{2}\cdot n_i \,,$$

$$\overline{Sv} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{m}\left(\frac{x_i + x_{i-1}}{2} - \overline{X}\right)^2 \cdot n_i} \,, \qquad (1)$$

где в качестве значения показателя на интервале принимается полусумма его значений на концах.

```
Среднее интервального вариационного ряда

(%i15) mean_vs(X,Y,m):=block( [N,i,s], N:0, s:0,
          for i:1 thru m do (N : N+ Y[i], s:s+(X[i]+X[i+1])/2.*Y[i]), s: s/N)$

(%i16) Xv:mean_vs(X_n[1], X_n[2], 8);
(%o16) 173.3

Среднеквадратическое отклонение

(%i17) std_vs(X,Y,m):=block( [N, i, s, s1], N:0, s:0, s1:0,
          for i:1 thru m do (N : N+ Y[i], s:s+(X[i]+X[i+1])/2.*Y[i]), s: s/N,
          for i:1 thru m do ( s1:s1+((X[i]+X[i+1])/2.-s)^2*Y[i]), s1:sqrt( s1/(N-1)))$

(%i18) Sv:std_vs(X_n[1], X_n[2], 8);
(%o18) 10.56
```

Рис.4. Расчет статистических показателей вариационного ряда

Рисунок 4 показывает, что среднее интервального вариационного ряда $Xv = 173{,}3$ и его среднеквадратическое отклонение $Sv = 10{,}56$ мало отличаются от выборочных среднего $a = 173{,}4$ и сред-

[3] Именно столько интервалов разбиения рекомендует формула Старджеса для N=100

неквадратического отклонения $s = 10{,}78$ соответственно.

4. Построение графических иллюстраций. Графические иллюстрации по статистической обработке данных можно произвести в Maxima при помощи нескольких функций: гистограмма, диаграмма рассеяния, круговая и секторная диаграммы, диаграмма Бокса-Вискера. [1].

В интерфейсе wxMaxima удобнее использовать wx- аналоги функций, осуществляющие вывод графика непосредственно на экран в активном документе [3].

Рассмотрим некоторые графики более подробно.

Гистограмма вариационного ряда по данным рассматриваемого примера.

```
(%i19) wxhistogram (A, nclasses=8, frequency=density, grid=true,
       title="Плотность распределения роста",
       xlabel="см",  fill_color=blue, fill_density=0.2)$
(%t19)
```

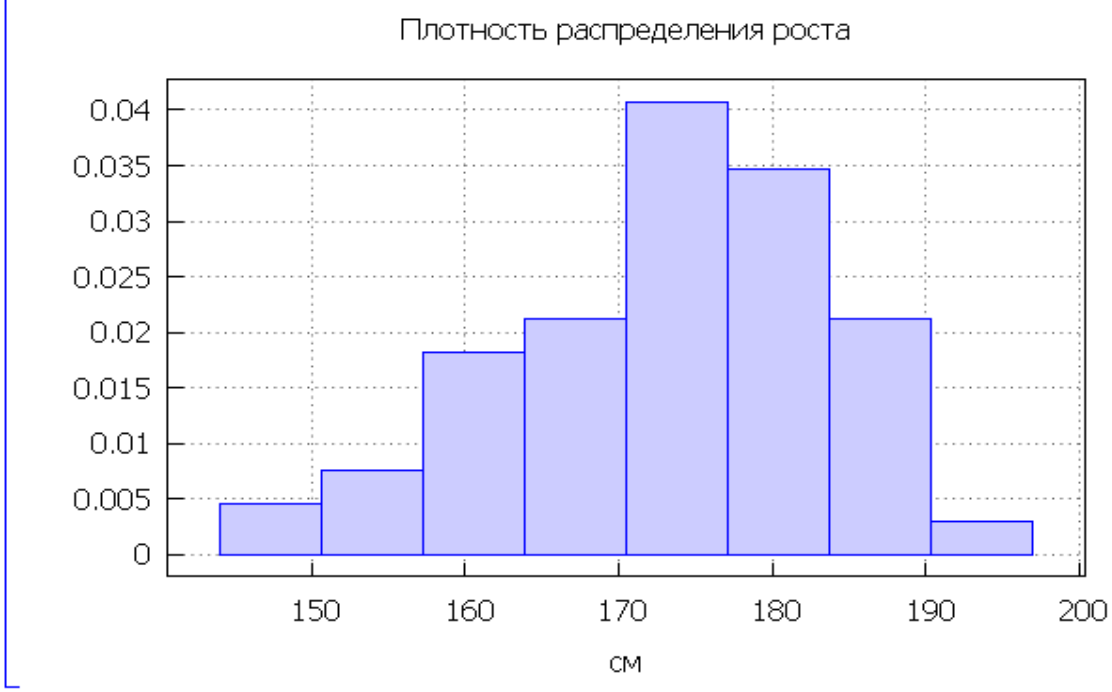

Рис.5. Гистограмма вариационного ряда

Использовались следующие опции функции wxhistogram:

— если задать значение опции nclasses, то для выборки A автоматически создается интервальный вариационный ряд;

— значение опции frequency=density указывает на то, что будет построена гистограмма плотности относительных частот вариационного ряда.[4]

Диаграмма Бокса–Вискера («ящик с усами») отображает одновременно несколько величин, которые характеризуют вариационный ряд снизу-вверх:

— наименьшее значение — 144 см;

— 0,25–квантиль, называемый первым (нижним) квартилем — 167 см;

— 0,5–квантиль, называемый медианой или вторым квартилем — 175 см;

[4] По внешнему виду гистограммы можно предположить нормальный закон распределения непрерывной случайной величины X, представляющей собой рост покупателей

— 0,75–квантиль, называемый третьим (верхним) квартилем — 181 см;
— наибольшее значение — 197 см.

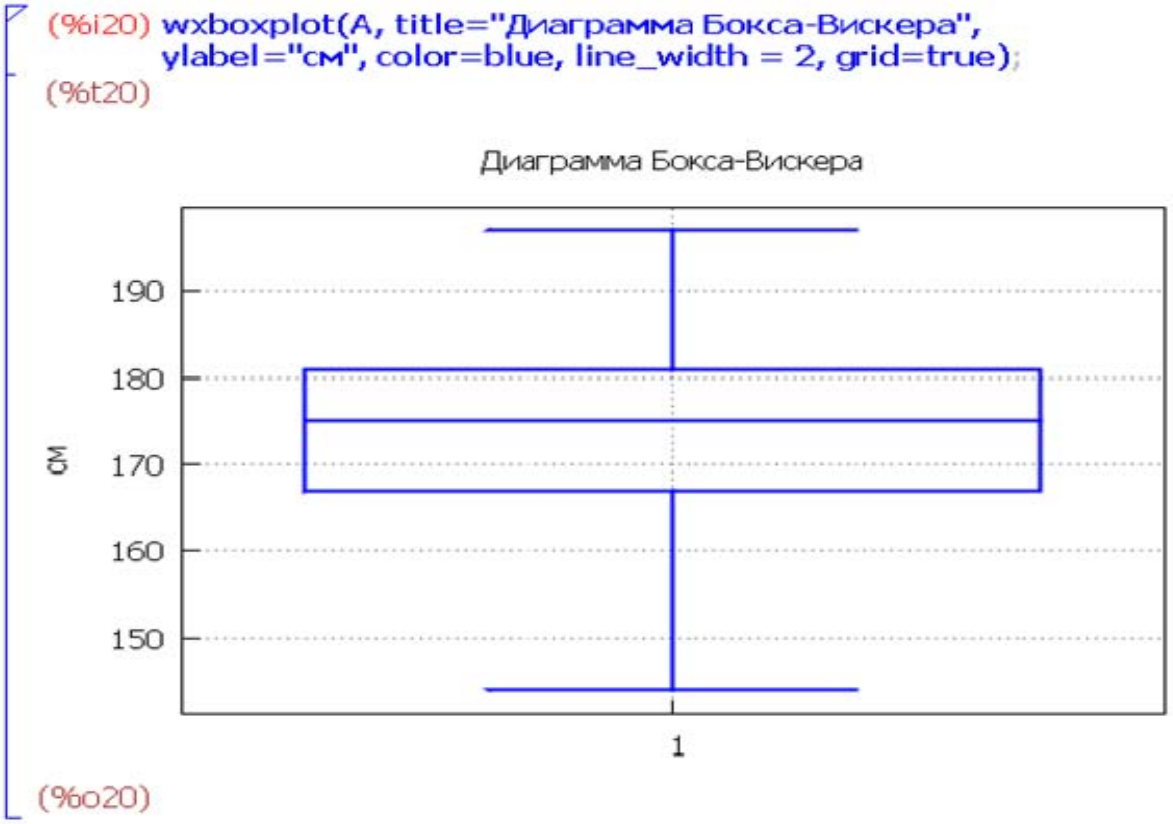

Рис.6. Диаграмма Бокса-Вискера для вариационного ряда

Разность между третьим и первым квартилями называется интерквартильным размахом (181–167=14 см.), который является характеристикой разброса значений случайной величины и вместе с медианой может быть использован в случае распределений с большими выбросами, либо при невозможности вычисления математического ожидания и среднеквадратического отклонения.

Если A является матрицей, то функция boxplot строит «ящики с усами» для каждого столбца.

5. Функции распределения случайных величин. Пакет distrib содержит набор функций для вычисления вероятностных характеристик многих дискретных и непрерывных распределений [4]. Существует соглашение об именах функций в пакете distrib. Каждое имя функции состоит из двух частей: первая часть указывает на параметр, который необходимо вычислить, вторая часть имени — это прямая ссылка на вероятностную модель.

Например, pdf_normal вычисляет плотность вероятности нормального закона распределения, а cdf_binomial — функцию распределения биномиального закона и т.д.

По данным интервального вариационного ряда примера рассчитаем плотность вероятности нормального закона распределения с математическим ожиданием $a = 173{,}4$ и среднеквадратическим отклонением $s = 10{,}78$ на границах интервалов:

```
(%i22) load (distrib)$

(%i23) Dn: pdf_normal (X_n[1], Xv, Sv), numer;
(%o23) [8.064 10^-4,0.003776,0.01193,0.02541,0.0365,0.03537,
0.02311,0.01019,0.003029]
```

Рис.7. Расчет плотности вероятности нормального закона распределения

Строим кривую нормального распределения — по форме она похожа на построенную ранее гистограмму вариационного ряда.

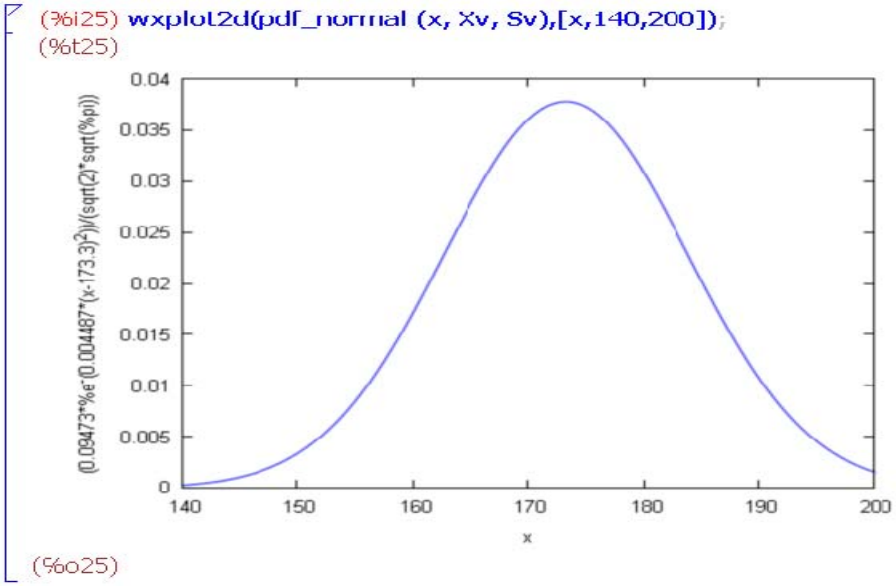

Рис.8. График плотности вероятности

6. Проверка статистических гипотез**.** Для проверки статистических гипотез в Maxima подключается пакет stats [4], который включает следующие функции:

— `test_mean` — функция позволяет оценить среднее значение и его доверительный интервал по выборке. Функция использует проверку по критерию Стьюдента;

— `test_means_difference` – функция позволяет проверить принадлежность двух выборок к одной генеральной совокупности;

— `test_variance` — оценка доверительного интервала для дисперсии выборки. Данная функция использует тест χ^2. Предполагается, что распределение выборки нормальное;

— test_normality — проверка нормальности распределения по критерию Шапиро-Уилка;

— другие стандартные тесты.

Функции пакета stats возвращают данные типа inference_result. Объекты этого типа содержат ряд необходимых результатов для анализа статистических распределений и проверки гипотез. Часть из них по умолчанию выводится на экран, а другие могут быть вызваны отдельно.

В отличие от классического подхода к проверке гипотез через критические значения распределения соответствующих статистик, в функциях пакета stats используется процедура проверки гипотез с помощью, так называемого P-value (P-значения), которое фактически является вероятностью «ошибки 1-го рода». Если p(t) меньше заданного уровня значимости, то нулевая гипоте-

за отвергается в пользу альтернативной. В противном случае она не отвергается. Преимуществом данного подхода является возможность определения уровня значимости, на котором нулевая гипотеза будет отвергнута или принята.

Проверим некоторые статистические гипотезы для данных выборки рассматриваемого примера.

Разобьем выборку A на две части и проверим гипотезу о равенстве выборочных средних $H_0 : a_1 = a_2$ с надежностью $\alpha = 0{,}95$.

```
(%i24) load("stats")$

(%i26) A1: makelist(A[i], i, 1, 50)$A2: makelist(A[i], i, 51, 100)$

(%i28) a1 : mean(A1),numer; a2 : mean(A2),numer;
(%o27) 173.2
(%o28) 173.7
```

Рис.9. Средние частей выборки

Фактическая разность этих выборочных средних равна 0,5. Используем функцию test_means_difference :

```
--> test_means_difference(A1, A2, );
                    DIFFERENCE OF MEANS TEST
                    diff_estimate = − 0.54
                    conf_level = 0.95
                    conf_interval = [ − 4.864, 3.784 ]
(%o29)              method = Exact t−test. Welch approx.
                    hypotheses = H0: mean1 = mean2 , H1: mean1 # mean2
                    statistic = 0.248
                    distribution = [ student_t , 94.12 ]
                    p_value = 0.8047
```

Рис.10. Проверка гипотезы о равенстве выборочных средних

Значения опций приняты по умолчанию, в частности, дисперсии выборок считались не заданными. Анализ inference_result показывает, что нулевая гипотеза H_0 о равенстве средних может быть принята, т.к. величина p_value=0,8047 достаточно велика, чтобы ее отвергнуть.

Проверим гипотезу о нормальности закона распределения случайной величины X, представленной выборкой A:

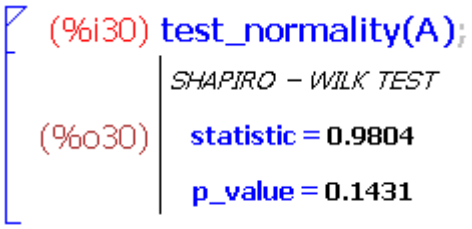

Рис.11. Проверка гипотезы о нормальном распределении

Проверка нормальности распределения осуществляется функцией test_normality. В этой функции реализован критерий Шапиро-Уилка, который считается наиболее эффективным для малых выборок. Функция возвращает два значения: величина W-статистики statistic=0,9804, характеризующая близость выборочного распределения к нормальному, и величина вероятности p_value. Так как p_value=0,1431 больше принятого уровня значимости 0,05, то нулевая гипотеза о нормальности распределения выборки A не отвергается.

Заключение. Система компьютерной алгебры Maxima позволяет решать большинство задач математической статистики. Программа Maxima является доступным и компактным программным продуктом, что обеспечивает удобство использования этой системы в научных исследованиях и учебном процессе.

Библиографический список

1. Чичкарев, Е. А. Компьютерная математика с Maxima: руководство для школьников и студентов / Е. А. Чичкарев. — Москва : ALT Linux, 2012. — 384 с.

2. Акишин, Б. А. Экономико-математические расчеты на персональном компьютере: учебно-методическое пособие / Б. А. Акишин, А. В. Галабурдин // Статистическая обработка данных. — Ч. 1. – Ростов-на-Дону : РАС ЮРГУЭС, 2007. — 52 с.

3. Акишин, Б. А. Решение математических задач с помощью пакета Maxima: учебное пособие / Б. А. Акишин, Л. В. Черкесова, А. В. Галабурдин — Ростов–на–Дону : Издательский центр ДГТУ, 2015. — 100 с.

4. Документация по текущей версии пакета Maxima [Электронный ресурс] — Режим доступа : http://maxima.sourceforge.net/docs/manual/en/maxima.html (дата обращения: 10.07.2016).