Разработка генератора массивов данных с заданными статистическими характеристиками для обучения студентов по направлению подготовки «Биотехнические системы и технологии»

С. В. Стародубцев
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
stepanstaro1309@gmail.com

Г. А. Машевский
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

***Аннотация*. Рассмотрены результаты, полученные в процессе разработки генератора массивов данных с заданными статистическими характеристиками для обучения студентов по направлению подготовки «Биотехнические системы и технологии».**

Ключевые слова: генератор массива данных

I. Введение

При обучении специалистов в области обработки биомедицинских данных, необходимо иметь большое количество массивов данных в качестве обучающих заданий. Применение для этого, данных реально полученных от пациентов массивов данных, возможно далеко не всегда, прежде всего из-за сложности подбора достаточного количества обучающих заданий для ознакомления с применением конкретных статистических критериев и методов. Многие статистические критерии подходят только для обработки определённых типов данных, и часто имеют специфические условиях для своего применения [1]. Поэтому возникает потребность в инструменте для генерации обучающих массивов данных с заранее заданными характеристиками.

На данном этапе наши усилия были сосредоточены на решении двух основных проблем. Во-первых, необходимо иметь возможность генерации значений переменных, соответствующих определённому закону распределения случайной величины, а также модифицировать данные распределения (например, за счёт изменения их асимметрии или эксцесса). Второй задачей является генерация пар (или большого числа переменных) предназначенных для проверки различных статистических критериев, например – коэффициентов корреляции. Результаты, полученные в процессе разработки такой программы, рассмотрены в этой статье.

II. Реализация программы

A. Выбор языка программирования и программного комплекса

Для создания программы был выбран пакет прикладных программ для решения задач технических вычислений MATLAB, по причине наличия высокого уровня языка, ориентированного на решение вычислительных задач.

B. Выбор законов распределений данных

В качестве основных распределений были взяты:

- равномерное;
- нормальное;
- логнормальное.

Равномерное распределение было выбрано, т. к. оно часто встречается в измерительной практике при округлении до целых значений шкал. Ошибка будет равномерно распределена между соседними целыми делениями.

Нормальное распределение имеет множество проявлений в биологии и применяется для моделирования ситуаций действий большого числа различных факторов [2]. Также на его использовании основана оценка значимости большого количества статистических гипотез [3].

Логнормальное распределение необходимо для моделирования экспоненциальных величин с нормальным законом.

C. Реализованные функции на данный момент

Равномерное распределение строится между двумя пределами с указанием количества значений и выбросов. Генерация распределения осуществляется функцией rand. Выбросы генерируются в диапазоне от 1,5 до 3,5 диапазонов относительно центра распределения. При указании количества выбросов большим, чем количество значений, будет выведено предупреждение. На рис. 1 представлена генерация равномерного распределения в пределах [-100;100] тысячи значений. На рис. 2 построен массив с количеством выбросов равным 100.

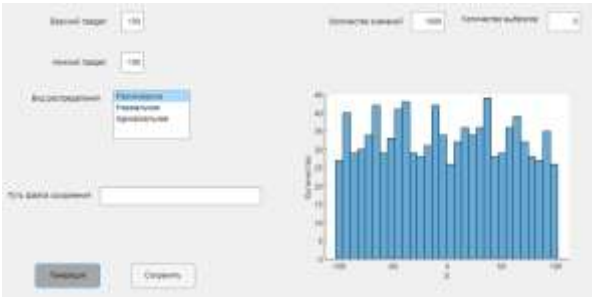

Рис. 1. Равномерное распределение, выполненное программой

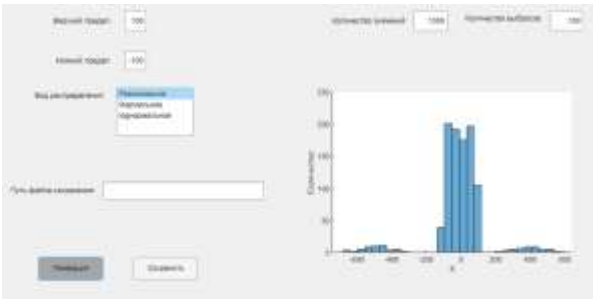

Рис. 2. Равномерно распределение с выбросами, выполненное программой

Для Гауссова распределения задаются дисперсия, математическое ожидание, количество значений и выбросов. Генерация распределения осуществляется функцией randn. Выбросы генерируются в диапазоне от 4 до 6 величин дисперсии относительно математического ожидания. По результатам генерации массива программа осуществляет проверку 5 % критерием Колмогорова и 5 % критерием Пирсона с помощью функций kstest и chi2gof соответственно. На рис. 3 приведена генерация тысячи значений, нормально распределенных с дисперсией 10 и математическим ожиданием 5. На рис. 4 построен массив с количеством выбросов равным 100.

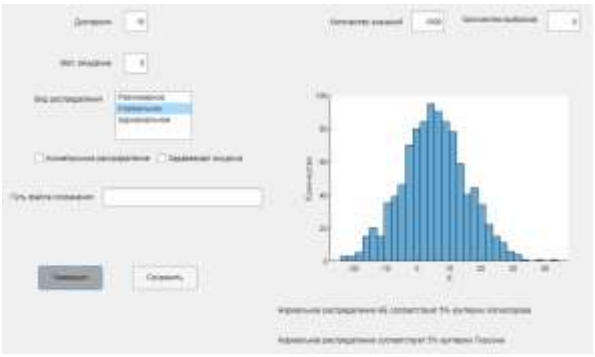

Рис. 3. Нормальное распределение, выполненное программой

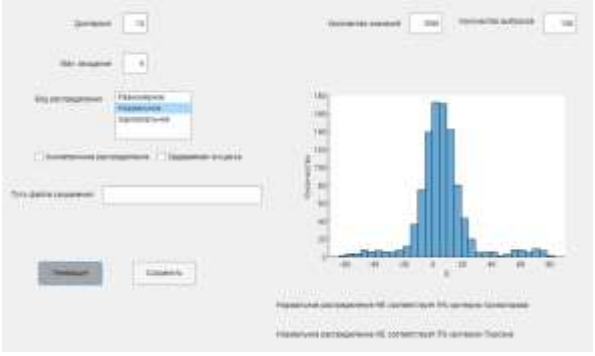

Рис. 4. Нормальное распределение с выбросами, выполненное программой

У логнормального закона значимыми параметрами являются, среднее значение и стандартное отклонение. Генерация распределения осуществляется функцией lognrnd. Для демонстрации генерации логнормального распределения ниже приведен рис. 5. На рисунке, сгенерирован массив со средним значением 0 и стандартным отклонением 0.7.

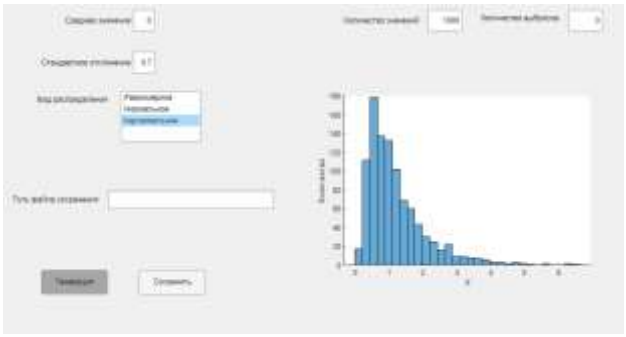

Рис. 5. Логнормальное распределение, выполненное программой

Таким образом, в ходе выполненных работ удалось создать генератор массивов заданного размера и содержащих в себе заданное количество выбросов.

III. Заключение

В ходе выполнения работы удалось создать инструмент, который может быть полезен при обучении студентов и аспирантов основам медицинской статистики. Полученные массивы могут быть сохранены в формате .xls и в дальнейшем обрабатываться в специализированном программном обеспечении. В дальнейшем при развитии данной программы предполагается добавление возможности модифицирования формы выбранного распределения, в частности, изменения величины её асимметрии и эксцесса, а также проверки соответствия, сгенерированного распределения теоретическому, при помощи критериев Колмогорова–Смирнова, χ^2-Пирсона и Шапиро–Уилкоксона, что облегчит преподавателю постановку оценки студенту при выполнении учебного задания.

Список литературы

[1] Генкин А.А. Новая информационная технология анализа медицинских данных (программный комплекс ОМИС). СПб.: Политехника, 1999. 191 с.: ил.

[2] Юнкеров В.И., Григорьев С.Г., Резванцев М.В. Математико-статистическая обработка данных медицинских исследований, 3-е изд., доп. СПб.: ВМедА, 2011. 318 с.

[3] Зайцев В.М., Савельев С.И. Практическая медицинская статистика: Учебное пособие / под редакцией академика РАМН, профессора д.м.н., заслуженного деятеля науки России А.И. Потапова и профессора, д.м.н. О.Г. Хурцилава. Тамбов: ООО «Цифра», 2013. 580 с.