



УДК 004.2

**SEMI-AUTOMATIC SYSTEM FOR VERIFICATION OF STATISTICAL
HYPOTHESES ABOUT MEAN AND DISPERSIONS****ПОЛУАВТОМАТИЧЕСКАЯ СИСТЕМА ДЛЯ ПРОВЕРКИ СТАТИСТИЧЕСКИХ
ГИПОТЕЗ О СРЕДНИХ И ДИСПЕРСИЯХ****Solomin A.V. / Соломин А.В.***PhD, as.prof. / к.ф.-м.н., доц.*

ORCID: 0000-0002-5226-8813

Kornienko G.A. / Корниенко Г.А.**Getun G.V. / Гетун Г.В.***PhD, as.prof. / к.т.н., доц.*

ORCID: 0000-0002-3317-3456

Ostapenko G.F. / Остапенко Г.Ф.*National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv,**Prosp.Peremohy, 37, 03056**Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского», Киев, пр-т Победы, 37, 03056*

Аннотация. В работе предложен полуавтоматический программный инструмент для усовершенствования технологии проведения статистических исследований и уменьшения вероятности ошибок, связанных с некорректным использованием статистических методов из-за несоответствия требуемых для этого предпосылок.

Ключевые слова: статистика, статистическая гипотеза, доказательная медицина, NI LabVIEW

Вступление.

Статистические методы исследований исключительно важны во многих областях деятельности, особенно в тех из них, где закономерности выявляются и устанавливаются на основе эмпирически полученных данных. Это относится к экономике, социологии, предсказанию катаклизмов и др. Особо следует выделить биомедицину, где статистические методы являются основой так называемой доказательной медицины. Здесь важнейшее значение приобретает корректность их использования, т.к. цена ошибок очень велика. В действительности же, как оказывается, такие ошибки нередки и объясняются слабой математической подготовкой специалистов с медицинским образованием. Дело в том, что набор статистических методов очень широк, но применение каждого из них требует выполнения особых предпосылок, например, нормального распределения случайных величин или одинаковости дисперсий и т.п. Использование статистического критерия без проверки соответствующих условий его применимости может привести к очень негативным последствиям, например, к ошибочным утверждениям об эффективности воздействия нового лекарственного препарата.

Наличие огромного количества научной литературы с соответствующими рекомендациями не решает проблемы, т.к. выбор и оценка применимости статистических методов все равно требует фундаментальных математических знаний, не случайно этими вопросами занимались выдающиеся математики.

Целью данной работы было создание программного обеспечения, которое



позволяет в полуавтоматическом режиме обеспечить выбор и проверку применимости статистических методов исследований, не требуя от пользователя глубоких математических знаний.

Основной текст. Для решения поставленной задачи предложено построить древовидную структуру из подпрограмм, представляющих конкретные статистические методы, причем в узлах этого дерева расположены средства принятия решений – ветвлений на основе анализа условий их применимости.

На вход системы подаются два массива значений, представляющие собой две выборки случайных величин. Результатом работы системы будет утверждение о статистически значимом различии средних значений случайных величин в этих выборках или отсутствии такого различия. Например, это может быть свидетельством результата влияния какого-то лекарственного препарата на артериальное давление пациента. Причем выборками могут быть либо две группы пациентов (экспериментальная – под воздействием препарата и контрольная – без такового), либо одна группа (до воздействия препарата и после такового). Второй случай называется связанными выборками или выборками, которые естественным образом разбиваются на пары, и статистические критерии в этом случае также иные [1]. Применяемые статистические методы также отличаются для случаев нормального закона распределения случайных величин и закона распределения, отличного от нормального; для случаев выборок с одинаковыми и с различными дисперсиями. В случае закона распределения, отличного от нормального, используется особый статистический метод, когда анализируются значения не самой случайной величины, а ее рангов.

Такое разнообразие возможных ситуаций и используемых при этом статистических методов и критериев потребовало создания программной системы в виде древовидной структуры с узлами принятия решений о путях дальнейшего выполнения программы. В каждом узле такой структуры пользователь получает информацию о результатах проверки параметров условий применимости конкретных методов с соответствующими рекомендациями о ветвлении. Параллельно по пути следования оказывается доступной информация о равенстве или различии дисперсий, о разновидности закона распределения и др.

Вышеупомянутое дерево решений, алгоритм и программный код можно условно разделить на 9 основных модулей, которые полуавтоматически переключаются во время работы программы на основе соответствующих решений в узлах дерева.

Система реализована в среде разработки NI LabVIEW, которая в настоящее время играет роль неформального стандарта для задач программного обеспечения в приборостроении и особенно в биомедицинской инженерии [2].

В начальном модуле пользователю предлагается выбрать один из двух вариантов: выборки содержат попарно связанные данные или нет. В первом случае происходит переход на седьмой модуль, иначе – на второй.

Во втором модуле проводится проверка вида распределения обоих



входных массивов (выборок) на соответствие нормальному закону распределения.

В зависимости от результата дальше предлагается переход на модуль 3 (проверка одинаковости дисперсии) или на модуль 6 (применение непараметрического критерия). Сравнение дисперсии в 3-ом модуле выполняется с помощью критерия Фишера.

В случае одинаковости дисперсий предлагается переход на модуль 4 (проверка гипотезы о равенстве средних двух выборок при одинаковых дисперсиях), иначе – на модуль 5 (проверка гипотезы о равенстве средних двух выборок при неодинаковых дисперсиях).

Шестой модуль проводит проверку гипотезы о равенстве средних значений двух выборок в случае закона распределения, отличного от нормального, путем ранжирования и применения двухвыборочного критерия Уилкоксона. Используется разработанная подпрограмма ранжирования выборок и проверки критерия.

Далее в дереве следуют статистические алгоритмы для связанных выборок, то есть когда исследуемая выборка естественным образом разбивается на пары, например, показатели одних и тех же больных до и после лечения (такие исследования можно назвать проверкой действенности обработки).

В седьмом модуле опять проводится тестирование на соответствие нормальному закону распределения, и в случае позитивного результата происходит переход на восьмой модуль, иначе – на девятый. В восьмом модуле применяется специальный t-критерий Стьюдента для связанных выборок, а в девятом – критерий знаков.

Интерфейс системы реализован в виде страниц-закладок, соответствующих выполняемым в конкретный момент модулям.

На рис.1 приведен интерфейс первого модуля, а на рис.2 – его блок-диаграмма, которая в программной среде NI LabVIEW соответствует коду (графический язык G).

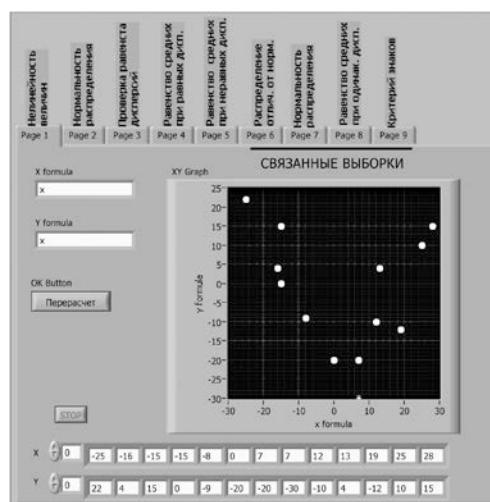


Рис.1. Вид интерфейса первого модуля

Дополнительное преимущество предлагаемой системы связано с возможностью анализа нелинейных статистических закономерностей в случаях,



когда ищутся связи не просто между статистическими величинами, а между какими-то функциями этих величин.

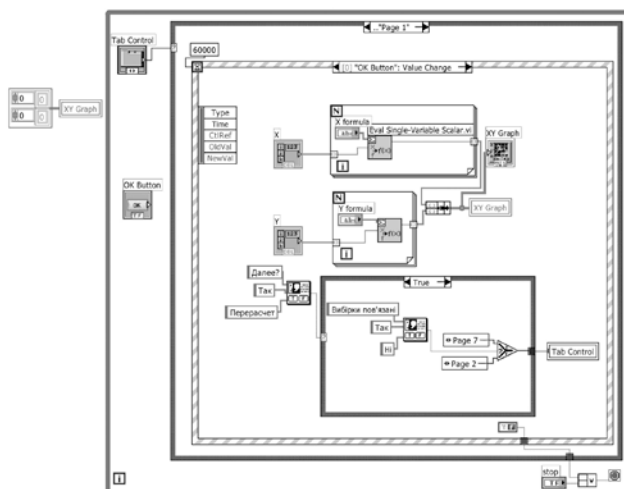


Рис.2. Блок-диаграмма первого модуля

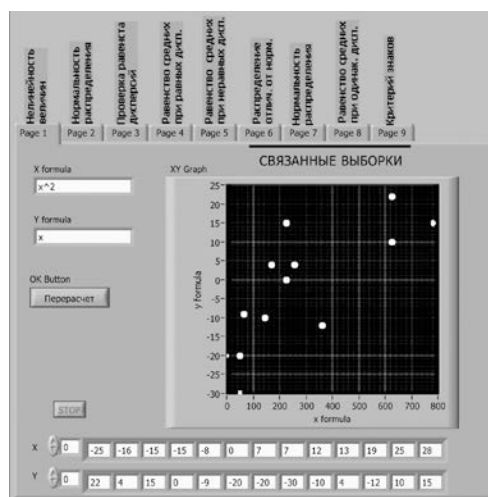


Рис.3. Вид интерфейса первого модуля после применения нелинейного преобразования

Пример такого нелинейного преобразования легко заметить, сравнивая рис.1 и рис.3. Здесь в графическом виде отображен характер связи случайных величин в первой и второй выборке (массивах x и y), которые отложены в качестве координат точек по соответствующим осям. Заметно, что на первом рисунке получившаяся из точек кривая несколько похожа на параболу. Напрашивается попытка сравнивать не величины y и x , а величины y и x^2 , что позволяет сделать предлагаемая программная система. Результат приведен на рис.3. Видно, что в этом случае получившаяся кривая близка к прямой линии, поэтому целесообразно проанализировать далее в системе статистические закономерности между величинами y и x^2 . В случаях периодически изменяющихся случайных величин можно применять в подобных случаях тригонометрические нелинейные преобразования случайных величин, в некоторых случаях – экспоненциальные или логарифмические и т.п.



Заключение и выводы.

Предложенная система позволяет проверять статистические гипотезы в полуавтоматическом режиме с гибкой адаптацией к конкретным ситуациям, в том числе с возможностью исследования статистических характеристик случайных величин, предварительно преобразованных посредством различных нелинейных функций.

Использование системы для проверки статистических гипотез в биомедицине кроме увеличения производительности и эффективности соответствующих исследований уменьшает вероятность ошибок, связанных с неправомерным применением статистических критериев.

Литература:

1. Лапач С.Н. Статистические методы в медико-биологических исследованиях с использованием Excel / С.Н. Лапач, А.В. Чубенко, П.Н. Бабич. – К. : МОРИОН, 2001. – 408 с.
2. Програмування в NI LabVIEW. Технологія розробки віртуальних приладів : навч. посіб. / О.Г. Кисельова, А.В. Соломін. – К. : НТУУ «КПІ», 2014. – 276 с.

References:

1. Lapach S.N. Statistichrskie metody v medico-biologicheskikh issledivanijah s ispolzovaniem Excel [Statistical methods in biomedical research using Excel] . – К. : МОРИОН, 2001. – 408 с.
2. Kiseleva O.G., Solomin A.V. Programuvannja v NI LabVIEW. Technologija rozrobki virtualnyh prykladiv [Programming in NI LabVIEW. Technology of virtual devices development] – К. : НТУУ «КПІ», 2014. – 276 с.

Abstract. *The semi-automatic programmatic instrument is offered for the improvement of statistical researches and diminishing of error's probability, related to the improper use of statistical methods by disparity of the pre-conditions required for this purpose.*

Key words: *statistics, statistical hypothesis, evidential medicine, NI LabVIEW.*

Статья отправлена: 16.03.2019 г.

© Соломин А.В.