ЛИТЕРАТУРА

1. Krivtsov A.M. Molecular Dynamics Simulation of Impact Fracture in Polycrystalline Materials // Mechanica. – 2003. – № 38. – С. 61–70.
2. Allen M. Introduction to Molecular Dynamics Simulation [Electronic resource]. – 2004. – P. 1-28. – Mode of access: http://www.fz-juelich.de/nic-series/volume23
3. Kristiansen K., Wouterse A., Philipse A. Simulation of random packing of binary sphere mixtures by mechanical contraction / Physica. – 2005. – № 358. – С. 249-262.
4. Волошин В.П., Медведев Н.Н. Исследование препика структурного фактора. Анализ неоднородных упаковок Леннард-Джонсовских атомов // Журнал структурной химии. – 2005. – Том 46, №1. – С. 96-100

УДК 519.233.3

Дмитриева А.В.
Научный руководитель: доц. Дереченник С.С.

КРИТЕРИЙ СОГЛАСИЯ ПИРСОНА И МЕТОДЫ ГРУППИРОВАНИЯ ДАННЫХ

Цель данной работы – исследовать эффективность критериев согласия типа хи-квадрат (χ^2) Пирсона при разбиении области определения случайной величины на интервалы равной длины.

При использовании критериев согласия типа χ^2 область определения разбивается на k интервалов граничными точками: $x_0 < x_1 < ... < x_{k-1} < x_k$.

Статистика X_n^2 Пирсона вычисляется в соответствии с соотношением [1]:

$$\chi^2 = N \sum_{i=1}^{k} \frac{\left(n_i / N - P_i(\theta)\right)^2}{P_i(\theta)}, \qquad (1)$$

где n_i – количество наблюдений, попавших в интервал;

$P_i(\theta) = \int_{x_{i-1}}^{x_i} f(x,\theta)\, dx$ – вероятность попадания наблюдения в i-й интервал;

$n = \sum_{i=1}^{k} n_i$ – количество всех наблюдений;

$\sum_{i=1}^{k} P_i(\theta) = 1$ – суммарная вероятность.

Статистические свойства критериев типа χ^2 зависят от того, каким именно образом область определения случайной величины разбивается на интервалы, а также от числа интервалов группирования. При практическом использовании критериев выбирают либо интервалы равной длины, либо интервалы равной вероятности (равной частоты), либо асимптотически оптимальные интервалы (в этом случае минимизируются потери в информации Фишера).

Рекомендуемое в различных источниках количество интервалов группирования, используемое при вычислении оценок параметров, построении гистограмм, а также при проверке статистических гипотез с помощью критерия Пирсона, колеблется в очень широких пределах. Большинство из рекомендуемых формул для оценки числа интервало

носит эмпирический характер и обычно дает завышенные величины. Практически все рекомендации по выбору числа интервалов исходят из того, чтобы при данном объеме выборки как можно лучше приблизить плотность распределения ее непараметрической оценкой (гистограммой).

Определение количества интервалов *k* при использовании интервалов равной длины традиционно связывается с объемом выборки *n*. Существующие рекомендации, однако, весьма противоречивы. Так, например, известны формулы Старджесса $k = 3.3 \lg n + 1$, Брукса и Краузера $k = 5 \lg n$, соотношение $k = \sqrt{n}$ и т.д. [2-4]. На практике же обычно руководствуются требованием, чтобы в интервалы попадало не менее 5-10 наблюдений.

Исследование влияния числа интервалов равной длины на проверку гипотез проводилось на примере анализа характеристик моделей однородно неупорядоченных дисперсных систем, полученных в результате вычислительных экспериментов [5]. Исследуемой случайной величиной являлся свободный от частицы нормированный объем ячейки Вороного (фактический объем, деленный на среднее по выборке и уменьшенный затем на минимальное полученное значение). Гипотеза о соответствии эмпирических данных двухпараметрическому распределению Вейбулла:

$$f(x) = \theta_0 \, \theta_1 \, x^{\theta_1 - 1} \cdot \exp\left\{ -\theta_0 \, x^{\theta_1} \right\} \qquad (2)$$

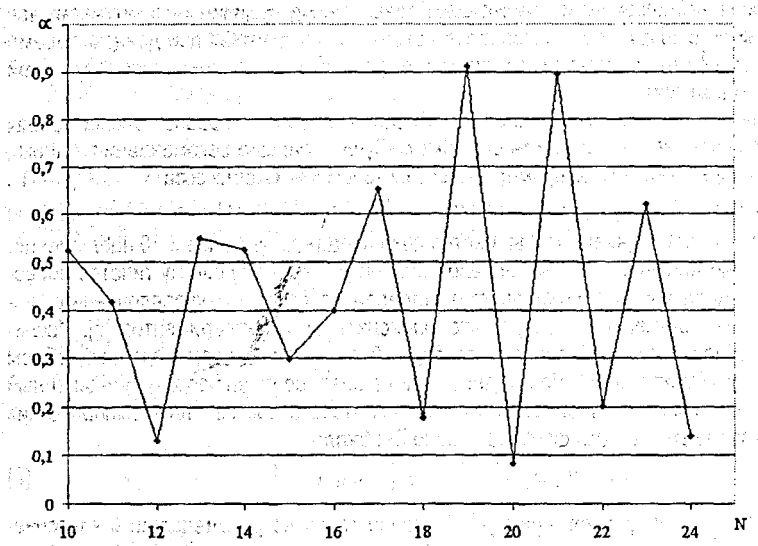
проверялась с помощью критерия χ^2 Пирсона согласно рекомендациям классических руководств по прикладной статистике. Область определения случайной величины разбивали на интервалы группирования равной величины, их число, при первоначальном разбиении, варьировалось в некотором диапазоне. Интервалы, в которые попадало менее 10 наблюдений, объединяли с соседними, в результате чего фактическое количество интервалов уменьшалось. Параметры распределения θ_0 и θ_1 оценивали методом минимизации расстояния χ^2, после чего вычисляли статистику Пирсона X_n^2. Принимая критическое значение $S_\alpha = X_n^2$, находили соответствующий полученной статистике, а также числу степеней свободы уровень значимости α, считая его степенью согласия, или вероятностью истинности проверяемой гипотезы. Результаты анализа для двух образцов (объем выборок - 388 и 397 наблюдений) на рисунке 1. Уровень значимости α определяли согласно соотношению:

$$\alpha = 1 - \int_0^{X_n^2} \frac{t^{\frac{n-3}{2}-1}}{2^{\frac{n-3}{2}} \cdot \Gamma\left(\frac{n-3}{2}\right)} \cdot \exp\left\{ -\frac{t}{2} \right\}, \qquad (3)$$

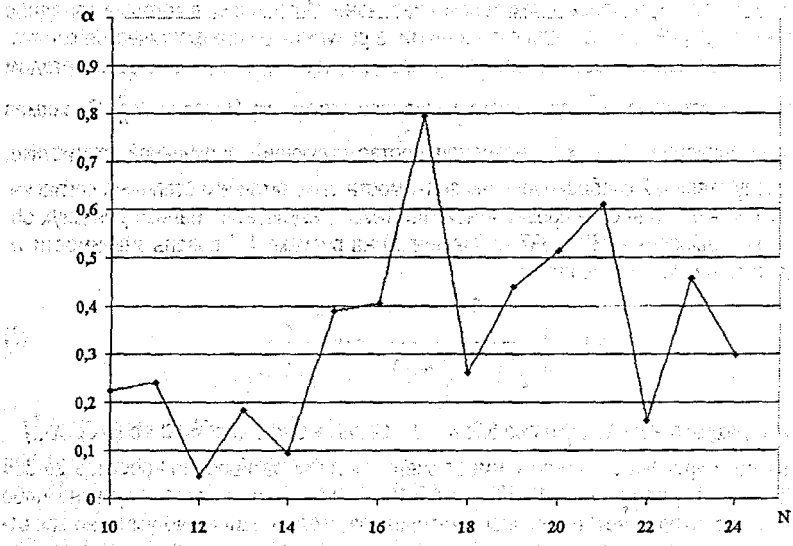
используя функцию пакета Mathematica 5.0: $\alpha = 1 - Quantile[\, ChiSquareDistribution[n-3],\ X_n^2 \,]$.

Достигаемый уровень значимости имеет очень широкий разброс: для образца 28-388 – от 0.08 до 0.91, для образца 46-397 – от 0.10 до 0.80. В результате, случайно либо преднамеренно выбрав некоторое число интервалов, можно одинаково успешно как отклонить, так и принять гипотезу, т.е. прийти к неверным статистическим выводам.

Следует заметить, однако, что оценки параметров распределения гораздо более устойчивы к числу интервалов и мало зависят от достигнутого уровня значимости.



а



б

Рис. 1. Зависимость уровня значимости α принятия гипотезы от числа интервалов N для образцов 28-388 (а) и 46-397 (б)

На основании анализа ряда публикаций (например, [6]), в которых и ранее обращалось внимание на неэффективность критерия χ^2 Пирсона с группированием данных наблюдений в интервалы равной длины, можно выделить следующие основные причины такой неэффективности:

1.Потеря информации при группировании данных наблюдений, которая, при разбиении на интервалы равной длины, является абсолютно непредсказуемой (неуправляемой). «Хорошее» разбиение, по-видимому, сильно зависит от распределения случайной величины.

2.Большинство результатов математической статистики имеют асимптотический характер, а на практике всегда имеют дело с ограниченными выборками.

3.Классические предположения математической статистики опираются, как правило, на нормальный закон распределения (например, ошибок наблюдений). Это не всегда справедливо в приложениях, а теоретическое обоснование подобных ситуаций связано с аналитически трудноразрешимыми задачами и потому отсутствует.

Таким образом, разбиение области определения случайной величины на интервалы равной длины применимо лишь для построения гистограмм или вычисления оценок параметров распределения методом минимизации расстояния χ^2, для проверки же гипотез с помощью критерия Пирсона таковое разбиение следует признать неэффективным.

ЛИТЕРАТУРА

1. Бочаров П.П., Печинкин А.В. Теория вероятностей. Математическая статистика. – М.: Гардарика, 1998. – 328 с.

2. Sturgess H.A. The choice of classic intervals // J. Am. Statist. Assoc. – March 1926. – P. 47.

3. Шторм Р. Теория вероятностей. Математическая статистика. Статистический контроль качества. – М.: Мир, 1970. – 368 с.

4. Heinhold I., Gaede K.W. Ingeniur statistic. – München; Wien, Springer Verlag, 1964. – 352 s.

5. Ртищева М.В., Разумейчик В.С., Дереченник С.С. Анализ топологических характеристик неплотных неупорядоченных монодисперсных структур / Проблемы проектирования и производства радиоэлектронных средств: Сб. материалов IV Международной НТК (25-26 мая 2006, г. Новополоцк). – Т.2. – Новополоцк: ПГУ, 2006. – С. 214-217

6. Лемешко Ю.Б., Чимитова Е.В. Об ошибках и неверных действиях, совершаемых при использовании критериев согласия типа χ^2 //Измерительная техника. – 2002. – №6. – С. 5-11.

УДК 693.22.004.18

Крапивин Ю.Б.

Научный руководитель: д.т.н. проф. Головко В.А.

НЕЙРОННЫЕ СЕТИ И ИХ ПРИМЕНЕНИЕ ДЛЯ ПОИСКА ИНФОРМАЦИИ В СЕТИ INTERNET

ВВЕДЕНИЕ

Наиболее быстро развивающейся технологией является Интернет, которая объединяет миллионы компьютеров, около сотни миллионов пользователей, число которых увеличивается на 50% ежегодно.

Основная часть информации в Интернете - это неструктурированное хранилище информации огромного объёма, которая характеризуется высокой динамичностью. В