УДК 532.696.1

Г.И. Манко, А.А. Довгополая

ОБ ИСПОЛЬЗОВАНИИ ИНФОРМАЦИОННОГО КРИТЕРИЯ ПРОВЕРКИ ГИПОТЕЗ О ЗАКОНЕ РАСПРЕДЕЛЕНИЯ ОШИБОК ИЗМЕРЕНИЯ

ГВУЗ «Украинский государственный химико-технологический университет», г. Днепропетровск

Рассмотрены вопросы использования информационных критериев согласия. Описаны способы определения критической области методом статистического моделирования. Предложены методы получения критических значений информационных критериев проверки гипотез о законах распределения ошибок измерений.

Essentially, all models are wrong, but some are useful.
George E. P. Box

Постановка проблемы

При исследовании метрологических характеристик средств измерения требуется знание законов распределения ошибок измерений. Во многих случаях распределение ошибок оказывается далёким от нормального. Приходится подбирать закон распределения, обеспечивающий более адекватные результаты исследований. Экспериментальным путем получается выборка значений ошибок, строится гистограмма и выполняется проверка на соответствие тому или иному теоретическому закону, что определяется с помощью критериев согласия. В качестве критериев согласия обычно используются статистики – функции от выборочных значений [1].

Широко используемые критерии согласия имеют ряд недостатков, налагающих определенные ограничения при их использовании. Общей характерной особенностью всех таких критериев является то, что они позволяют лишь установить, имеются ли основания отвергнуть проверяемую гипотезу. Если оснований нет, то математическая статистика рекомендует проводить дополнительные исследования.

Известно, что решение многих проблем статистического анализа возможно при использовании информационного подхода. Это отметили многие исследователи, начиная с Н. Винера и К. Шеннона.

Анализ известных исследований и публикаций

Попытки использования информационных подходов при решении задач проверки статистических гипотез делались неоднократно. Так, Кульбак [2] предложил рассматривать логарифм отношения плотностей вероятностей $f_1(x)$ и $f_2(x)$, соответствующих двум альтернативным гипотезам H_1 и H_2, как информацию в точке наблюдения *x* для различения в пользу H_1 против H_2. Тогда средняя для всего множества X={x} информация от наблюдения относительно вероятностной меры μ_2 (соответствующей гипотезе H_2) для различения в пользу H_1 против H_2, определяется формулой

$$I(2:1)=\int f_2(x)\log\frac{f_2(x)}{f_1(x)}d\lambda(x),$$

где λ – может быть мерой μ_1 или μ_2 или $(\mu_1+\mu_2)/2$.

В развитие идей Кульбака в работе [3] предложено определять на интервале U расхождение между плотностью вероятностей f(x) и её оценкой f*(x), вычисленной по выборке $\{x_i, i=1,2,...,N\}$, с помощью условной апостериорной энтропии H_N, которая остается после обработки выборки и построения оценки f*(x):

$$H_N=\int_U f^*(x)\log\frac{f^*(x)}{f(x)}dx.$$

Аналогично определяется условная априорная энтропия, вычисляемая с использованием априорной плотности $f_0(x)$:

$$H_0 = \int_U f_0(x) \log \frac{f_0(x)}{f(x)} dx.$$

Тогда разность априорной и апостериорной энтропий дает величину условной информации, извлекаемой из выборки путем вычисления оценки $f^*(x)$.

В работе [4] сделана попытка выработать общий формализованный подход к выбору статистики, позволяющий разрабатывать новые более совершенные критерии согласия на основе использования понятия неорганизованности Ю.М. Горского [5]. Неорганизованность – обобщенная мера различия какого-либо элемента x_j в отношении эталона порядка $x_{эт}$, которая стремится к нулю при $x_j \to x_{эт}$. Её можно рассматривать как взвешенную по вероятностям P_j некоторую функцию параметра P_j, оценивающего степень отличия x_j от $x_{эт}$. По Горскому, для большинства задач оценки неорганизованности достаточно использовать одну из четырех функциональных зависимостей: линейную, степенную, логарифмическую или экспоненциальную:

$$\overline{O} = k \sum_j P_j \Pi_j; \qquad (1)$$

$$\overline{O} = k \sum_j P_j \Pi_j^i; \qquad (2)$$

$$\overline{O} = k \sum_j P_j \log \Pi_j; \qquad (3)$$

$$\overline{O} = k \sum_j P_j \exp(\Pi_j - C). \qquad (4)$$

Тогда, приняв в (1) в качестве параметра P_j максимум абсолютной величины разности реальной и гипотетической функций распределения, получим при k=1 выражение критерия согласия Колмогорова. Положив в (2) значение P_j равным отношению квадрата разности реальных и гипотетических частот попадания значений случайной величины в j-й интервал диапазона её изменения к гипотетической частоте и приняв все $P_j=1/m$ (m – число интервалов), будем иметь при k=т, i=2 выражение для критерия согласия χ^2. Принимая в (3) за параметр P_j отношение реальных вероятностей распределения p_j к гипотетическим q_j, получаем формулу полезной информации Бонгарда [6]:

$$I_п = \sum_j p_j \log \frac{p_j}{q_j}.$$

Поскольку эта величина отрицательная, следует говорить о дезинформации, содержащейся в проверяемой гипотезе. Сравнивая количество дезинформации, вносимой разными гипотезам, можно выбрать оптимальный закон распределения, наилучшим образом аппроксимирующий результаты экспериментальных исследований.

Для практического применения такого информационного критерия проверки гипотез о законах распределения в [4] предлагается использовать следующую статистику:

$$\delta = \frac{1}{n} \sum_j \nu_j \log \frac{\nu_j}{n q_j}, \qquad (5)$$

где n_j – число реализаций выборки, попадающих в j-тый интервал.

Формулировка цели статьи

Целью данной работы является исследование информационного критерия согласия (5) и выработка практических рекомендаций по его использованию для проверки статистических гипотез о законах распределения ошибок измерений.

Изложение основного материала

Как показано в [3], при вычислении значения информационного критерия получается случайное выборочное значение, по которому нельзя сделать определенного вывода относительно качества анализа. Применение информационных характеристик сталкивается с проблемой отыскания распределения случайных значений критерия, что особенно затруднительно при теоретических исследованиях. Поэтому информационные критерии согласия до сих пор не нашли практического применения а математической статистике.

Тем не менее, вычислительные мощности современных компьютеров позволяют решать многие задачи путем статистического эксперимента. Нормативный документ Р 50.1.037-2002 [7] рекомендует следующее: если для описания выборки используется закон распределения вероятностей F(x, и) и найдены оценки его параметров $\hat{\theta}$, а для проверки сложной гипотезы H_0: $F(x) \in \{F(x, \hat{\theta}),\ \hat{\theta} \in Q\}$ исследователю неизвестно распределение статистики соответствующего критерия согласия, целесообразно воспользоваться методикой компьютерного анализа статистических закономерностей, хорошо зарекомендовавшей себя при моделировании распределений статистик критериев.

Для этого следует в соответствии с законом F(x,q) смоделировать N выборок того же объема n, что и выборка, для которой необходимо проверить гипотезу H_0: $F(x) \in \{F(x, \hat{\theta}),\ \hat{\theta} \in Q\}$. Далее для каждой из N выборок вычислить оценки тех же параметров закона, а затем значение статистики S соответствующего критерия согласия. В результате будет получена выборка значений статистики

$S_1,S_2,...,S_N$ с законом распределения $G(Sn^{½}H_0)$ для проверяемой гипотезы H_0. По этой выборке при достаточно большом N можно построить достаточно гладкую эмпирическую функцию распределения $G_N(Sn^{½}H_0)$, которой можно непосредственно воспользоваться для вывода о том, следует ли принимать гипотезу H_0. Отклонения смоделированного распределения от теоретического при N=2000 обычно имеют порядок $\approx\pm0{,}015$. При необходимости, можно построить приближенную аналитическую модель, аппроксимирующую $G_N(Sn^{½}H_0)$, и тогда уже, опираясь на эту модель, принимать решение относительно проверяемой гипотезы.

Широкие возможности для статистического моделирования обеспечивает программный пакет Statistics Toolbox системы MATLAB, который и был использован для исследования информационного критерия согласия (5).

Обрабатывалась выборка из 100 значений ошибок измерения цифровым вольтметром образцового переменного напряжения 10 В, генерируемого установкой поверки вольтметров В1-20.

Средствами языка системы MATLAB была создана программа, реализующая процесс статистического моделирования с использованием следующих законов распределения (в скобках указан префикс для имен функций пакета Statistics Toolbox):

а) нормальное распределение (norm);

б) распределение Релея (rayl);

в) экспоненциальное распределение (exp);

г) распределение Вейбулла (wbl);

д) обобщенное распределение Парето (gp).

Далее префикс для имен функций будем в общем виде обозначать буквами pref.

С помощью функции preffit(x) были определены параметры распределения parmhat, с помощью функции prefpdf(x,muhat) сгенерирована выборка гипотетических значений ошибки и построена соответствующая теоретическая кривая. По выборкам экспериментальных и теоретических значений построены соответствующие гистограммы. По формуле (5) рассчитано значение информационного критерия согласия.

Исследование распределения вероятностей значений информационного критерия проводилось методом Монте-Карло. Генерировалась выборка случайных значений v_j, для которых определялось значение критерия d, и так повторялось 3000 раз. Это с запасом превышает рекомендованное в [7] число N=2000. Для полученной таким образом выборки значений критерия с помощью средств Statistics Toolbox была построена гистограмма с наложенной на нее кривой функции плотности распределения нормального закона (функция histfit), рассчитаны первые четыре момента распределения (математическое ожидание mean, среднеквадратичное отклонение std, несмещенная асимметрия skewness и несмещенный эксцесс kurtosis). Критическое значение было определено как квантиль уровня 0,01 с помощью функции prctile. Пример результатов расчетов для обобщенного распределения Парето (Generalized Pareto) показан на рис. 1, значения критерия согласия приведены в табл. 1.

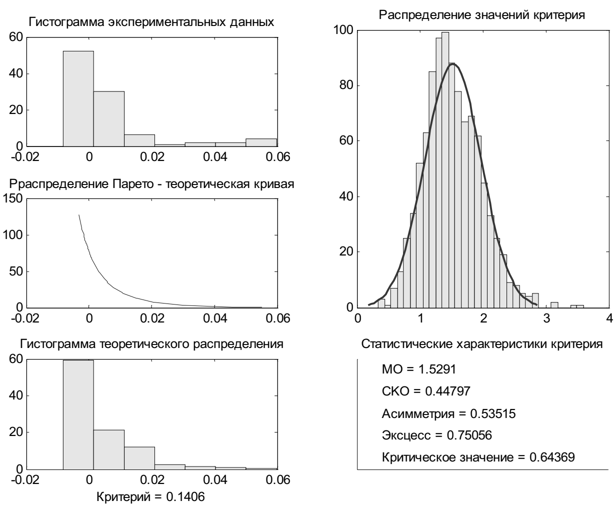

Рис. 1.

Таблица 1

Значения информационного критерия согласия

Закон распределения	Значение критерия согласия	Критическое значение
Релея	1,25	0,72
Нормальный	0,49	0,45
Экспоненциальный	0,17	0,57
Вейбулла	0,16	0,55
Generalized Pareto	0,14	0,64

Анализ результатов расчетов показывает, что гипотезы о распределении Релея и нормальном распределении ошибок измерения должны быть отвергнуты. Наименьшую дезинформацию вносит принятие гипотезы об обобщенном распределении Парето. И это можно было ожидать, поскольку форма гистограммы экспериментальных данных характерна для распределений с «тяжелыми хвостами» (heavy tails) – это распределение, хвост которого нельзя «отрезать», т.к. нельзя пренебречь существенной частью ошибок измерений. Такие распределения часто описываются обобщенным распределением Парето, имеющим следующую плотность:

$$f\ (x)=\frac{1}{\sigma}\left(1+\frac{k(x-\theta)}{\sigma}\right)^{\left(-1-\frac{1}{k}\right)},a>0.$$

Здесь σ – параметр масштаба, k – параметр формы, θ – порог. В нашем исследовании

σ=0,0078, k=0,2721, θ=−0,0033.

О характере распределения информационного критерия согласия (правый график на рис. 1) можно сделать следующие выводы. Распределение имеет выраженную асимметрию и небольшой положительный эксцесс. Такую картину могут дать логнормальное распределение, гамма-распределение, распределение Вейбулла, обобщенное распределение экстремального значения (Generalized extreme value). Все эти гипотезы были проверены с помощью информационного критерия. Результаты расчетов сведены в табл. 2. Пример результатов расчетов показан на рис. 2.

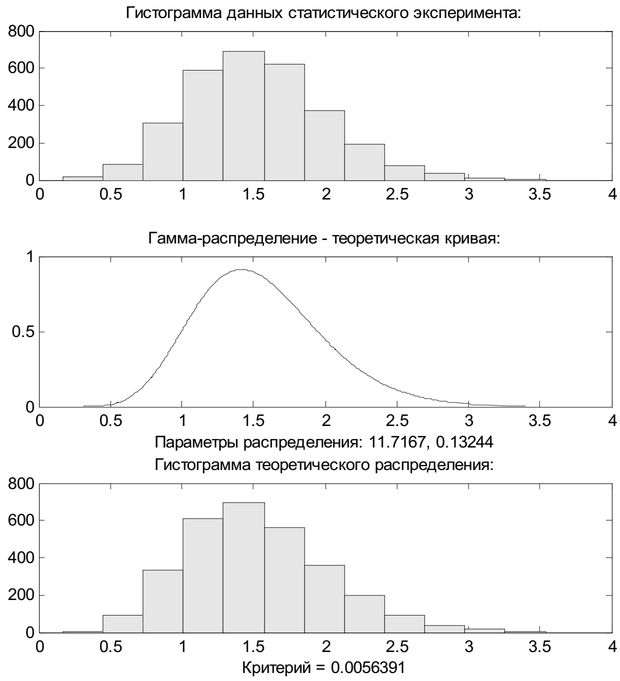

Рис. 2

Таблица 2

Результаты проверки гипотез о распределении критерия (5)

Закон распределения	Значение критерия согласия	Квантиль уровня 0,05
Вейбулла	0,0222	0,1331
Логнормальный	0,0196	0,1298
Нормальный	0,0117	0,1302
Гамма-распределение	0,0056	0,1331
Generalized extreme value	0,0033	0,1374

Как видим, наилучшую аппроксимацию обеспечивают гамма-распределение и обобщенное распределение экстремального значения. Однако, как показали дополнительные исследования, в области малых значений критерия (5) кривые плотности этих распределений мало отличаются от кривой нормального распределения (рис. 3), а квантили уровня 0,05 практически равны для этих распределений (см. последнюю колонку табл. 2). Это дает основание использовать таблицы значений функции нормального распределения для определения критических значений информационного критерия согласия.

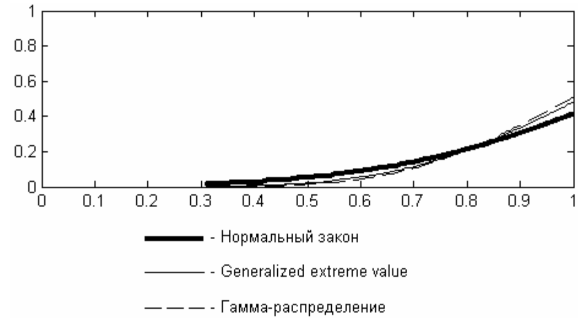

Рис. 3.

Вывод

Для проверки гипотез о законах распределения ошибок измерений целесообразно использование информационного критерия согласия на основе статистики (5). Это позволяет сравнивать аппроксимирующие распределения по степени их адекватности экспериментальным данным. Определение критической области возможно как методом проведения статистического эксперимента, так и с использованием таблиц значений функции нормального распределения.

СПИСОК ЛИТЕРАТУРЫ

1. *Справочник* по теории вероятностей и математической статистике / В.С. Королюк, Н.И. Портенко, А.В. Скороход, А.Ф. Турбин. − К.: Наук. думка, 1978. − 582 с.

2. *Кульбак С.* Теория информации и статистика: пер.с англ. − М.: Наука, 1967. − 408 с.

3. *Гаскаров Д.В., Шаповалов В.И.* Малая выборка. − М.: Статистика, 1968. − 248 с.

4. *Кафаров В.В., Мешалкин В.П., Манко Г.И.* Информационный критерий проверки гипотез о законах распределения характеристик надежности // Доклады АН СССР. − 1979. − Т.9. − № 4. − С.923-927.

5. *Горский Ю.М.* Информационные аспекты управления и моделирования. − М.: Наука, 1978. − 223 с.

6. *Бонгард М. М.* Проблемы узнавания. − М.: Наука, 1967. − 320 с.

7. *Р 50.1.037-2002* Рекомендации по стандартизации. Правила проверки согласия опытного распределения с теоретическим : методические рекомендации. − Новосибирск: Изд-во НГТУ, 2002. − 153 с.

Поступила в редакцию 5.12.2012