УДК УДК 62–192: 519.2 + 629.7.017.1

© Лукин В. Л., Сухорученков Б. И., Кузнецов В. И.
Lukin V. , Sukhoruchenkov B. , Kuznetsov V.

СТАТИСТИЧЕСКИЙ МЕТОД ВЫЯВЛЕНИЯ АНОМАЛЬНЫХ РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙ ХАРАКТЕРИСТИК ТЕХНИЧЕСКИХ СИСТЕМ

STATISTICAL METHOD OF EXPOSUREOF ANOMALOUS RESULTS OF MEASURINGSOF DESCRIPTIONS OF TECHNICAL SYSTEMS

Аннотация. *Решается задача проверки на достоверность результатов измерений характеристик технических систем. Предлагается новый метод выявления аномальных данных на основе построения и анализа плотности вероятности оценок толерантных границ для нормально распределенной случайной величины.*

Annotation. *The task of checking for authenticity of results of measurings of descriptions of the technical systems decides. The new method of exposure of anomalous information is offered on the basis of construction and analysis of closeness of probability of estimations of tolerant scopes for the normally distributed casual size.*

Ключевые слова. *Измерения, нормальное распределение, аномальные результаты, статистические методы, толерантные границы.*

Key words. *Measurements, normal distribution, anomalous results, statistical methods, tolerant borders.*

1. Постановка задачи

При испытаниях или целевом применении технических систем (ТС) измеряются различные характеристики (параметры состояния) ТС. На основе обработки результатов измерений осуществляется оценивание характеристик ТС и контроль соответствия их предъявляемым требованиям. Практика показывает, что измерения характеристик ТС могут содержать отдельные резко выделяющиеся значения. Это может быть результатом естественного случайного отклонения характеристики ТС и (или) погрешностей измерений. Однако возможны и аномальные результаты (грубые ошибки, выбросы, промахи), обусловленные сбоями при измерениях и регистрации данных, резкими отклонениями условий наблюдений, нештатной работой ТС, ошибками операторов и др. Наличие аномальных результатов может привести к недостоверным результатам при оценивании и контроле соответствия характеристик ТС предъявляемым требованиям. Поэтому необходимо выявлять и устранять аномальные результаты измерений.

Рассмотрим результаты измерений (выборку) x_i, $i = 1,..., n$, случайной величины (СВ) X из нормальной генеральной совокупности (ГС) с неизвестными математическим ожиданием (МО) M и среднеквадратическим отклонением (СКО) σ. Элементы выборки $\{x_i\}$ могут формироваться различными способами. Это могут быть результаты измерений реализаций случайной характеристики ТС или измерений неслучайного параметра, искаженные нормально распределенными погрешностями. Если измеряемая характеристика ТС переменна, то выборка формируется из остаточных невязок после сглаживания экспериментальных данных. В качестве

Лукин Владимир Леонидович – доктор технических наук, профессор, академик-секретарь секции "Инженерные проблемы стабильности и конверсии" Российской инженерной академии, тел. 8-(495)543-36-77;
Сухорученков Борис Иванович – доктор технических наук, профессор, член-корреспондент Российской инженерной академии, профессор кафедры ракетного вооружения Военной академии РВСН имени Петра Великого, тел. 8-(495)696-06-48;
Кузнецов Валерий Иванович – кандидат технических наук, старший научный сотрудник, начальник отдела ОАО «Военно-инженерной корпорации», тел. 543-36-76.

Lukin Vladimir – doctor of engineerings sciences, professor, akademik-sekretar' sections the "Engineerings problems of stability and conversion" of the Russian engineering academy, tel. 8-(495)543-36-77;
Sukhoruchenkov Boris – doctor of engineerings sciences, professor, corresponding member of the Russian engineering academy, professor of department of rocket armament of the Military academy of RVSN of the name of Peter Great, tel. 8-(495)696-06-48;
Kuznetsov Valeriy – the candidate of technical sciences, the senior research scientist, the chief of division of OJSC "Militaryengineering corporation", tel. 8-495-5433676.

основного варианта будем рассматривать выборку данных измерений случайной характеристики ТС.

Для контроля характеристики ТС необходимо найти оценку МО и ее СКО. Оценивание производится одним из статистических методов. При этом возможны три варианта оценивания. Первый вариант – когда все элементы выборки являются (считаются, признаются) достоверными. Второй вариант – когда некоторые резко выделяющиеся (экстремальные) значения выборки предполагаются недостоверными и исключаются из выборки. Третий вариант – когда экстремальные значения выборки, хотя возможно и являются (предполагаются) недостоверными, но оставляются в выборке.

Ставятся следующие задачи: выбор статистического метода оценивания характеристик ТС по результатам измерений, обоснование метода выявления и устранения возможных недостоверных элементов выборки и определение варианта обработки данных измерений, обеспечивающего наибольшую точность при предположении, что в выборке могут присутствовать аномальные элементы. Решение указанных задач излагается далее.

2. Выбор статистического метода обработки данных измерений

Прежде чем проверять элементы выборки на достоверность, необходимо оценить МО и СКО СВ. Для оценивания параметров нормального распределения по выборке можно использовать различные статистические методы: метод моментов, метод максимального правдоподобия (ММП), метод наименьших квадратов [1, 3, 5, 7]. В работах [2, 8] обоснован метод несмещенных оценок (МНО) и показано, что он обеспечивает максимальную точность оценивания неизвестных параметров. В соответствии с МНО оценивание МО и СКО производится на основе построения плотности вероятности оценок МО и СКО в следующем порядке.

Плотность вероятности случайной характеристики X с нормальным распределением записывается в виде функции от МО и СКО:

$$f(x/M,\sigma) = (2\pi)^{-0,5}\,\sigma^{-1}\exp\left[-\frac{1}{2\sigma^2}(x-M)^2\right]. \qquad (1)$$

При известных МО и СКО плотность вероятности совокупности возможных значений случайной характеристики $\{x_i\}$ представляется в виде

$$f(\{x_i\}/M,\sigma) = (2\pi)^{-0,5n}\times$$
$$\times\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-M)^2\right]; \qquad (2)$$

Так как параметры M и σ неизвестны, то для их оценивания на основе полученной выборки $\{x_i\}$ по МНО строится плотность вероятности $f(m, s)$ возможных оценок МО m и СКО s по следующим зависимостям [2, 8]:

$$g(m,s) = (2\pi)^{-0,5n}\times$$
$$\times s^{-n}\exp\left[-\frac{1}{2s^2}\sum_{i=1}^{n}(x_i-m)^2\right]; \qquad (3)$$

$$k = \left[\int_{-\infty}^{\infty}\int_{0}^{\infty} g(m,s)\,ds\,dm\right]^{-1}; \qquad (4)$$

$$f(m,s) = k\,g(m,s). \qquad (5)$$

На основе плотности (5) строятся автономные плотности вероятности оценок МО и СКО по зависимостям

$$f(m) = \int_{0}^{\infty} f(m,s)\,ds; \quad f(s) = \int_{-\infty}^{\infty} f(m,s)\,dm. \qquad (6)$$

Несмещенные точечные оценки параметров M и σ и их дисперсии определяются на основе автономных плотностей (6) по известным зависимостям для первых и вторых центральных моментов распределения

$$\bar{M} = \int_{-\infty}^{\infty} m\,f(m)\,dm;$$
$$\bar{\sigma}^2_{\bar{M}} = \int_{-\infty}^{\infty}(m-\bar{M})^2 f(m)\,dm; \qquad (7)$$

$$\bar{\sigma} = \int_{0}^{\infty} s\,f(s)\,ds;$$
$$\sigma^2_{\bar{\sigma}} = \int_{0}^{\infty}(s-\bar{\sigma})^2 f(s)\,ds. \qquad (8)$$

Сравнение оценок, получаемых по зависимостям (7) и (8), с оценками по ММП показало [2, 8], что по МНО получаются более точные оценки МО и СКО. Кроме того, как будет показано в п. 5, МНО позволяет построить плотности вероятности оценок границ толерантного интервала и использовать их для повышения достоверности выявления аномальных элементов выборки. Поэтому далее для анализа выборки будем использовать МНО.

3. Сравнение точности различных вариантов обработки экспериментальных данных

Проведем анализ точности оценивания МО характеристики ТС и ее СКО при вариантах обработки выборки, описанных в п. 1. Если все элементы выборки являются достоверными, как в первом варианте, то в результате обработки данных одним из статистической методов получаются оценки МО и СКО, а также СКО оценки МО. Обозначим их соответственно в виде $\bar{M}_1$, $\bar{\sigma}_1$ и $\sigma_{\bar{M}1}$. Если резко выделяющиеся (экстремальные) значения выборки $x_{эк}$ являются недостоверными и исключаются из выборки, как во втором варианте, то получаются оценки МО и ее СКО по сокращенной выборке. Обозначим их в виде $\bar{M}_2$ и $\sigma_{\bar{M}2}$. При третьем варианте, когда экстремальные значения хэк возможно являются недостоверными, но оставляются в выборке, определяются оценки МО и ее СКО, которые обозначим в виде $\bar{M}_3$ и $\sigma_{\bar{M}3}$.

Исследования были проведены следующим образом. Объем выборки варьировался в пределах $n\in[10; 60]$. При заданных МО и СКО методом статистического моделирования формировались достоверные элементы выборки из нормальной ГС. Вычислялись оценки МО $\bar{M}_1$ и ее СКО $\sigma_{\bar{M}1}$ по полученной выборке по МНО. Затем выделялся один экстремальный элемент выборки и его значения варьировались по степени удаления от оценки МО. При этом оценивались МО и ее СКО в соответствии со 2-м и 3-м вариантами обработки выборки. Для анализа погрешностей при этих вариантах использовался показатель в виде отношения СКО оценки МО к СКО, которое получается при исходной выборке с достоверными элементами

$$\delta_2 = \frac{\sigma_{\bar{M}2}}{\sigma_{\bar{M}1}}; \quad \delta_3 = \frac{\sigma_{\bar{M}3}}{\sigma_{\bar{M}1}}. \tag{9}$$

Результаты вычислений показали, что показатель δ_2 зависит только от объема выборки и определяется по зависимости

$$\delta_2 = \sqrt{n/(n-1)}\ . \tag{10}$$

Показатель δ_3 зависит от объема выборки и от степени отклонения недостоверного результата $x_{нд}$ от оценки МО, которую представим в относительном виде

$$\Delta_{нд} = \frac{\left| x_{нд} - \bar{M}_1 \right|}{\bar{\sigma}_1}. \tag{11}$$

Зависимости показателя δ_3 от n и $\Delta_{нд}$ показаны на рис. 1. Эти зависимости можно выразить в виде приближенной формулы

$$\delta_3 \cong \sqrt{1 + \Delta_{нд}^2}\ . \tag{12}$$

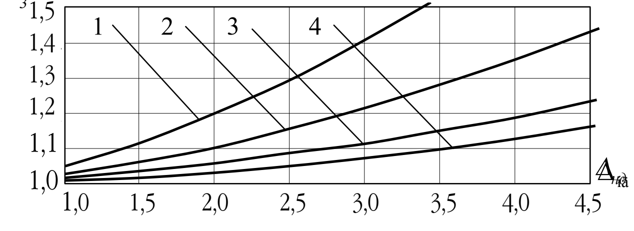

Рис. 1. Зависимости относительных погрешностей оценок МО δ_3 от объема выборки n и от степени отклонения $\Delta_{нд}$:
1 – при n = 10; 2 – при n = 20; 3 – при n = 40; 4 – при n = 60

Результаты проведенных исследований показали, что если в выборке имеется хотя бы один недостоверный элемент, то точность оценивания характеристик ТС снижается. При $\Delta_{нд}$ = 1 погрешности δ_3 совпадают с погрешностями δ_2. Если некоторый элемент хнд недостоверен, то для снижения погрешностей оценивания характеристик ТС при $\Delta_{нд} > 1$ целесообразно исключать значение $x_{нд}$ из выборки. Для этого необходимо выявить аномальные элементы выборки.

4. Анализ классических методов выявления аномальных элементов выборки

Задача выявления и устранения аномальных результатов наблюдений ставилась и решалась в инженерной практике и математической статистике на протяжении нескольких десятилетий [1, 3, 4, 6]. Выявление и устранение недостоверных данных наблюдений производится в несколько этапов [7]. Вначале бракуются все результаты, выходящие за границы интервала возможного варьирования характеристики ТС. Затем в зависимости от физического смысла характеристики могут исключаться значения, не соответствующие монотонному характеру изменения характеристики при последовательных наблюдениях, а также значения, приращения которых превышают предельно возможную скорость изменения характеристики. На заключительном этапе, когда остаются результаты, не сильно отличающиеся от остальных членов выборки, выявление аномальных данных наблюдений производится на основе статистических методов [3, 6].

В принципе аномальными могут быть любые элементы выборки. Однако даже недостоверные элементы, попадающие во внутреннюю область выборки, выявить практически невозможно, при этом они не сильно искажают получаемые оценки характеристик ТС. Поэтому анализу на достоверность подвергаются только элементы, попадающие в граничные области выборки. Наиболее известные критерии исключения грубых ошибок из выборок, извлекаемых из нормальных генеральных совокупностей, принадлежат Н. В. Смирнову и Ф. Груббсу [1, 3, 6].

Изложим часто используемый критерий согласно работам [3, 6]. Рассмотрим выборку x_i, $i = 1,..., n$, нормально распределенной СВ X. Допустим, что экстремальный элемент выборки $x_{эк}$ (минимальный или максимальный) вызывает подозрение и его необходимо проверить на достоверность. Для этого по выборке вычисляются оценки МО M и СКО σ по методу максимального правдоподобия по зависимостям

$$\bar{M} = \frac{1}{n}\sum_{i=1}^{n} x_i\,; \quad \bar{\sigma} = \left[\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - M\right)^2\right]^{0,5}. \tag{13}$$

На основе статистик (13) вводится безразмерная величина «типа Стьюдента»

$$V = \frac{\left| x_{эк} - \bar{M} \right|}{\bar{\sigma}}. \tag{14}$$

Распределение случайной величины (14) зависит не только от статистик (13), но и от объема выборки n. Квантили этого распределения в зависимости от n и уровня значимости α представляются в виде таблицы с двумя входами [3, 6]: n [3; 25] и α [0,01; 0,10]. Квантили при некоторых значениях n и α приведены в табл. 1.

Таблица 1

Квантили распределения величины $V = \frac{|\, x_{эк} - \bar{M} \,|}{\bar{\sigma}}$

Уровень значимости α	Объем выборки n				
	5	10	15	20	25
0,10	1,791	2,146	2,326	2,447	2,537
0,05	1,869	2,294	2,493	2,623	2,717
0,01	1,955	2,540	2,800	2,959	3,071

Проверка элементов выборки объема n осуществляется следующим образом. Выдвигается нулевая гипотеза H_0, что экстремальное значение $x_{эк}$ не является аномальным (принадлежит нормальной ГС). Для проверки гипотезы задается уровень значимости α и по таблице квантилей величины V определяется критический уровень $V_{кр}$. На основе выборки вычисляются оценки МО и СКО по зависимостям (13) и значение параметра V по формуле (14). Решение принимается по правилу

$$\left.\begin{array}{l}\text{гипотеза } H_0 \text{ принимается, если } V \le V_{кр};\\ \text{гипотеза } H_0 \text{ отклоняется, если } V > V_{кр}.\end{array}\right\} \qquad (15)$$

Если гипотеза отклонена, то экстремальный элемент выборки $x_{эк}$ признается аномальным и исключается из выборки.

Изложенный метод позволяет выявить аномальные данные наблюдений. Однако следует иметь в виду, что принятие нулевой гипотезы не подтверждает, что элемент $x_{эк}$ достоверен, а лишь означает, что экспериментальные данные не противоречат этому предположению. В действительности с высокой вероятностью Р = 1–α значение $x_{эк}$ может относиться к аномальному. Используемый показатель (14) недостаточно чувствителен к аномальным результатам, так как оценки МО и СКО определяются по всем элементам выборки, включая и возможные недостоверные значения. В результате по правилу (15) удается выявить только аномальные данные, сильно отклоняющиеся от совокупности остальных элементов выборки. Показатель (14) обоснован только для проверки одного экстремального значения, хотя в выборке могут быть несколько аномальных результатов, примерно одинаково удаленных от оценки МО. Рассмотренный критерий справедлив для нормального распределения и непригоден при других распределениях, например, при экспоненциальном распределении, которое применяется при оценивании показателей надежности ТС, а также при распределениях Релея и Максвелла, которые используются при оценивании показателей точности функционирования ТС. Приводимые таблицы квантилей показателя V [3, 6] ограничены объемом выборки $n \le 25$.

Достоверность выявления аномальных результатов наблюдений можно повысить на основе оценок границ толерантного интервала.

5. Построение плотности вероятности оценок толерантных границ для нормального распределения

Толерантным интервалом (ТИ) называется такой интервал [T_1; T_2], в который с вероятностью γ попадает доля генеральной совокупности СВ не менее заданного значения P [1, 6]. В настоящее время толерантные границы (ТГ) T_1 и T_2 для нормальной СВ определяются на основе оценок МО и СКО по приближенным зависимостям, приведенным в работе [6]. Для более корректного оценивания границ T_1 и T_2 можно использовать МНО (см. п. 2), в соответствии с которым сначала строится плотность вероятности оценок ТГ следующим образом. Возможные оценки ТГ нормальной СВ, которые обозначим в виде t_1 и t_2, определяются на основе возможных оценок МО m и СКО s по зависимостям

$$t_1 = m - u_P s; \qquad t_2 = m + u_P s, \qquad (16)$$

где u_P – двусторонняя квантиль распределения, соответствующая вероятности (доле) P.

Как следует из (16), оценки ТГ являются функциями от случайных аргументов (оценок) m и s. Поэтому для построения плотности вероятности оценок ТГ необходимо построить плотности вероятности $f(m, s)$ оценок МО m и СКО s по полученной выборке. Если выборка $\{x_i\}$ взята из нормальной ГС, то плотность вероятности $f(m, s)$ строится по последовательным зависимостям (3)–(5). На основе этой плотности строятся плотности вероятности оценок ТГ $f(t_1)$ и $f(t_2)$ как функции от случайных аргументов [5]. Рассмотрим построение плотности $f(t_1)$. Для этого введем в рассмотрение дополнительные функции t_{11} и t_{12}:

$$t_{11} = \varphi_1(m, s) = m - u_P s; \quad t_{12} = \varphi_2(m, s) = s. \qquad (17)$$

На основе (17) найдем обратные функции

$$m = \psi_1(t_{11}, t_{12}) = t_{11} + u_P t_{12}; \quad s = \psi_2(t_{11}, t_{12}) = t_{12}. \qquad (18)$$

Определим производные от этих функций и якобиан [5]

$$\zeta(t_{11}, t_{12}) = \begin{vmatrix} \dfrac{\partial\psi_1(t_{11}, t_{12})}{\partial t_{11}} & \dfrac{\partial\psi_1(t_{11}, t_{12})}{\partial t_{12}} \\ \dfrac{\partial\psi_2(t_{11}, t_{12})}{\partial t_{11}} & \dfrac{\partial\psi_2(t_{11}, t_{12})}{\partial t_{12}} \end{vmatrix} = \begin{vmatrix} 1 & u_P \\ 0 & 1 \end{vmatrix} = 1. \qquad (19)$$

Плотность вероятности $f_1(t_{11}, t_{12})$ функций t_{11} и t_{12}

определяется на основе плотности вероятности оценок $f(m, s)$ по зависимости [5]

$$f_1(t_{11}, t_{12}) = f(t_{11} + u_P t_{12}; t_{12})\, \zeta(t_{11}, t_{12}). \qquad (20)$$

Плотность вероятности собственно оценок левой ТГ $t_1 = t_{11}$ вычисляется на основе (20) как для элемента случайного вектора с учетом ограничения оценки СКО $t_{12} = s > 0$

$$f(t_1) = \int_0^{\infty} f_1(t_{11}, t_{12})\, dt_{12}\,. \qquad (21)$$

Аналогично строится плотность вероятности оценок правой ТГ t_2. При этом знак перед квантилью u_P в зависимостях (17), (18) и (20) меняется на противоположный.

На основе построенных плотностей $f(t_1)$ и $f(t_2)$ определяются оценки толерантных границ $\overline{T}_1$ и $\overline{T}_2$ при заданной доверительной вероятности γ из соотношений

$$\int_{\overline{T}_1}^{\infty} f(t_1)dt_1 = \gamma_1; \quad \int_{-\infty}^{\overline{T}_2} f(t_2)dt_2 = \gamma_2, \qquad (22)$$

где γ_1 и γ_2 – доверительные вероятности, выбираемые из условия $\gamma_1 + \gamma_2 = 1 + \gamma$; обычно $\gamma_1 = \gamma_2 = 0{,}5(\gamma + 1)$.

Пример 1. При 10 независимых наблюдениях получены следующие реализации нормальной СВ X:

$$\{x_i\} = \{3{,}1;\ 2{,}7;\ 0{,}5;\ 3{,}5;\ 2{,}9;\ 4{,}2;\ 2{,}2;\ 5{,}1;\ 1{,}3;\ 4{,}5\}.$$

Необходимо построить плотность вероятности оценок МО и СКО.

Плотность вероятности оценок МО и СКО $f(m, s)$,

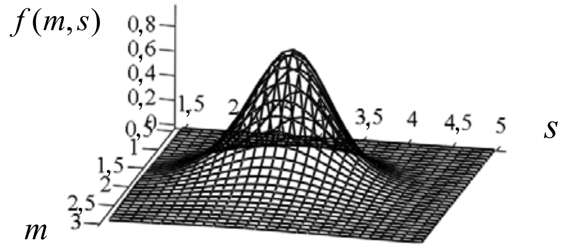

Рис. 2. Общий вид плотности вероятности оценок МО m и СКО s

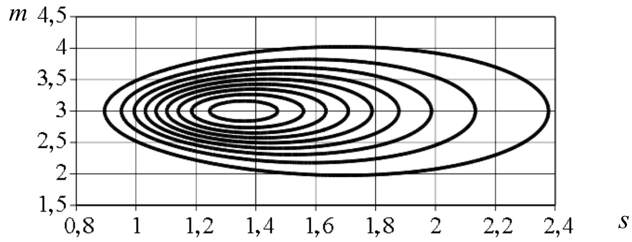

Рис. 3. Карта уровней плотности вероятности оценок МО m и СКО s

построенная в системе вычислений MathCAD по зависимостям (3)–(5) с учетом полученной выборки, показана на рис. 2 и 3.

Пример 2. В условиях примера 1 построить плотности вероятности оценок ТГ для СВ X при заданной доле $P = 0{,}9$.

Плотности вероятности оценок ТГ строятся по изложенной выше методике по зависимостям (16)–(21). При точных значениях МО и СКО для вероятности P=0,9 двусторонняя квантиль $u_{0,9} = 1{,}645$. Плотности вероятности оценок $f(t_1)$ и $f(t_2)$, построенные при этом значении квантили, показаны на рис. 4.

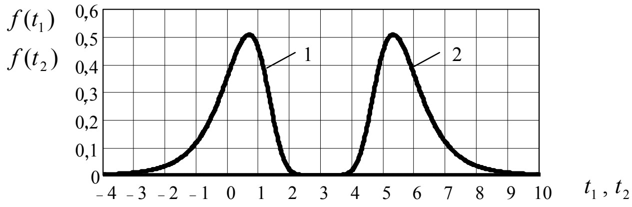

Рис. 4. Плотности вероятности оценок толерантных границ СВ: 1 – $f(t_1)$; 2 – $f(t_2)$

Аналогично можно построить плотности вероятности оценок ТГ других типов непрерывных распределений: экспоненциального, Релея, Максвелла и др.

На основе построенных плотностей можно проверить экстремальные элементы выборки на достоверность, что рассматривается далее.

6. Метод выявления аномальных результатов измерений

Толерантный интервал, построенный по методике п. 5 по зависимости (22), означает, что с вероятностью γ за его пределами может оказаться не более $(1 - P)$% элементов выборки. Поэтому использование толерантного интервала для выявления аномальных результатов измерений основано на следующих основных положениях. Если при построении ТИ задать вероятность (долю) P, близкую к единице, то за пределами ТИ может оказаться незначительная доля кондиционных элементов выборки. Границы ТИ расширяются при увеличении доверительной вероятности γ. При $P \to 1$ и $\gamma \to 1$ оценка ТИ расширяется настолько, что попадание в него кондиционных элементов выборки становится маловероятной. Если при этом экстремальные элементы выборки выходят за пределы ТИ, то их можно отнести к аномальным и исключить из выборки. Как следует из п. 2, при этом потери точности будут значительно ниже по сравнению с тем, если недостоверный элемент оставить в выборке.

Для проверки элементов выборки на достоверность с помощью ТИ можно использовать два способа, рассмотренные далее.

6.1. Общий метод выявления аномальных результатов измерений

При обработке результатов измерений необходимо определить, какие элементы выборки можно отнести к аномальным. При нормальном и близких к нему распределениях СВ это зависит от принятой вероятности возможных отклонений экстремальных значений от математического ожидания. Если МО M и СКО σ известны, то доля P элементов выборки попадает на отрезок R, который зависит от двусторонней квантили u

$$R = [M - u\sigma; M + u\sigma]. \qquad (23)$$

Доля элементов, выходящих за пределы отрезка, равна $(1 - P)$. При $P \to 1$ элементы выборки, не попадающие на отрезок R, можно считать недостоверными. Квантиль u зависит от вероятности P, см. табл. 2.

Таблица 2

Квантили нормального распределения

Вероятность P	0,95	0,98	0,99	0,997
Квантиль u_P	1,96	2,33	2,58	3,00
Ожидаемое число выбросов за пределы отрезка R	1 из 20	1 из 50	1 из 100	3 из 1000

Если МО и СКО неизвестны, то отрезок R определяется по зависимости (23) при оценках МО и СКО. В этом случае при распределениях, близких к нормальному, для выявления промахов в некоторых источниках [4] рекомендуется выбирать значения квантили в зависимости от объема выборки

$$\left.\begin{array}{l} \text{при } 6 < n \le 100 \; u = 4; \\ \text{при } 100 < n \le 1000 \; u = 4{,}5; \\ \text{при } 1000 < n \le 10\,000 \; u = 5. \end{array}\right\} \qquad (24)$$

Выбор величины квантили осуществляется в зависимости от качества измерительной информации и предполагаемой доли возможных аномальных элементов в полученной выборке, а также от особенностей и важности решаемой задачи. При этом необходимо учитывать возможные потери из-за снижения точности оценивания и достоверности контроля характеристик ТС при исключении кондиционных элементов выборки, с одной стороны, и из-за снижения достоверности контроля характеристик ТС при возможном наличии промахов в выборке, с другой стороны.

При выбранном значении квантили u крайние (экстремальные) элементы выборки $x_{эк}$ (их может быть несколько) предполагаются достоверными, если они находятся в пределах отрезка (23), и аномальными, если выходят за пределы отрезка R. Однако определить точные значения границ отрезка R невозможно, так как МО и СКО неизвестны. На основе полученных экспериментальных данных можно лишь оценить границы отрезка R (толерантного интервала) с некоторой погрешностью.

Для выявления недостоверных элементов на основе оценок ТГ, см. п. 5, используется следующий общий метод. Для повышения чувствительности метода к аномальным результатам при построении ТИ целесообразно предварительно исключить из выборки наиболее подозрительные резко выделяющиеся элементы, так как в случае их недостоверности они могут исказить оценки толерантных границ. Выбирается доля (вероятность) P и соответствующее значение квантили u. По методу, изложенному в п. 5, строятся плотности вероятности оценок ТГ $f(t_1)$ и $f(t_2)$ по сокращенной выборке. На основе плотностей строятся функции распределения (ФР) оценок ТГ по зависимостям

$$F(t_1) = \int_{-\infty}^{t_1} f(t_1)\,dt_1; \quad F(t_2) = \int_{-\infty}^{t_2} f(t_2)\,dt_2. \qquad (25)$$

ФР (25) зависят от условий решаемой задачи (от параметров n, P и u). Их удобно представлять в зависимости от безразмерного параметра V, аналогичного показателю (14):

$$V = \frac{|t - \bar{M}|}{\bar{\sigma}}, \qquad (26)$$

где t – оценка ТГ: $t = t_1$ или $t = t_2$; $\bar{M}$ и $\bar{\sigma}$ – оценки МО и СКО, определяемые по МНО.

Виды ФР $F(V)$, построенные в системе вычислений MathCAD по методике п. 5 и зависимости (25) с учетом (26) при отдельных значениях вероятности P и квантили u из табл. 2, представлены на рис. 5 –7. Здесь также показаны принятые критические значения квантили $u_{кр}$.

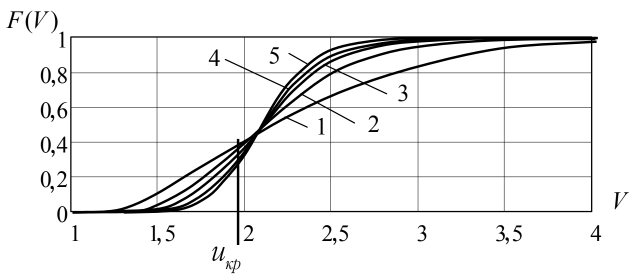

Рис. 5. Функции распределения оценок толерантной границы при P = 0,95 (при квантили $u_{кр}$ = 1,96) при различных объемах выборки: 1 – n = 10; 2 – n = 20; 3 – n = 30; 4 – n =40; 5 – n = 60

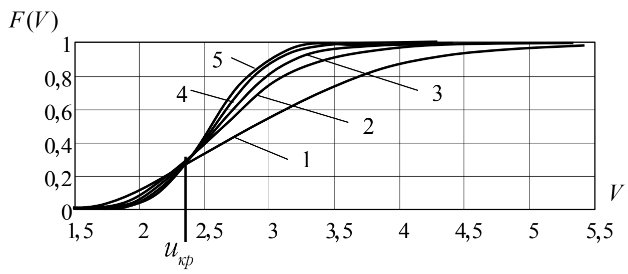

Рис. 6. Функции распределения оценок толерантной границы при P = 0,98 (при квантили $u_{кр}$ = 2,33) при различных объемах выборки: 1 – n = 10; 2 – n = 20; 3 – n = 30; 4 – n = 40; 5 – n = 60

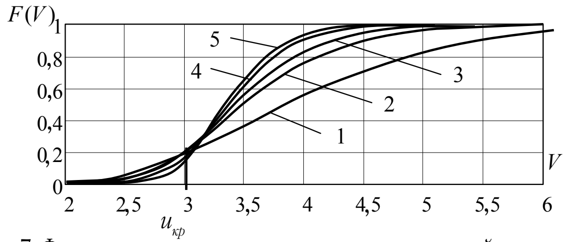

Рис. 7. Функции распределения оценок толерантной границы при $P = 0{,}98$
(при квантили $u_{кр} = 3$) при различных объемах выборки:
1 – $n = 10$; 2 – $n = 20$; 3 – $n = 30$; 4 – $n = 40$; 5 – $n = 60$

На основе построенных ФР можно найти оценки ТГ при заданной доверительной вероятности γ с учетом (22), (25), (26)

$$\bar{T}_1 = \bar{M} - V_1\,\bar{\sigma};\quad \bar{T}_2 = \bar{M} + V_2\,\bar{\sigma}, \tag{27}$$

где V_1, V_2 – показатели, определяемые по ФР при $F(V_1) = \gamma_1$ и $F(V_2) = \gamma_2$; γ_1 и γ_2 – как в зависимости (22).

ФР $F(V)$ характеризует вероятность, что кондиционные результаты наблюдений не выходят за пределы оценок границ ТИ (27). Поэтому их можно использовать для проверки нулевой гипотезы H_0 о том, что экстремальные элементы $x_{эк}$ не являются аномальными. Для проверки гипотезы задается уровень значимости $\alpha=1-\gamma$. Для каждого значения $x_{эк}$ вычисляется показатель $V_{эк}$ по зависимости (26) при $t = x_{эк}$:

$$V_{эк} = \frac{\left|x_{эк} - \bar{M}\right|}{\bar{\sigma}}. \tag{28}$$

Определяется ФР $F(V)$ при аргументе $V_{эк}$. Решение принимается отдельно по каждому значению $x_{эк}$ по правилу, аналогичному (15)

$$\left.\begin{array}{l}\text{гипотеза } H_0 \text{ принимается, если } F(V_{эк}) \le 1 - \alpha;\\ \text{гипотеза } H_0 \text{ отклоняется, если } F(V_{эк}) > 1 - \alpha.\end{array}\right\} \tag{29}$$

Проверку гипотезы H_0 можно осуществить также непосредственно на основе показателя V. Для этого определяется критическое значение показателя $V_{кр}$, при котором ФР $F(V_{кр}) = 1 - \alpha$. Решение по каждому значению $x_{эк}$ принимается по правилу:

$$\left.\begin{array}{l}\text{гипотеза } H_0 \text{ принимается, если } V_{эк} \le V_{кр};\\ \text{гипотеза } H_0 \text{ отклоняется, если } V_{эк} > V_{кр}.\end{array}\right\} \tag{30}$$

Оценки ТГ (27) зависят от значений доли P и доверительной вероятности γ, которые часто задаются субъективно. При увеличении P оценки ТГ могут изменяться в широком диапазоне, особенно при ограниченном объеме выборки. Рассмотрим, например, объем выборки $n\in[10; 20]$ и $\gamma = 0{,}8$ и $\gamma_1 = \gamma_2 = 0{,}9$. При этом ФР $F(V) = 0{,}9$ и на основе рис. 5-7 получаются следующие диапазоны критических значений показателя V:

$$\begin{array}{l}\text{при } P = 0{,}950\,|\,V_{кр}|\,[2{,}7;\ 3{,}3];\\ \text{при } P = 0{,}980\,|\,V_{кр}|\,[3{,}5;\ 4{,}5];\\ \text{при } P = 0{,}997\,|\,V_{кр}|\,[4{,}5;\ 5{,}5].\end{array} \tag{31}$$

С учетом (31) получается соответствующий значительный интервал для оценок границ ТИ (27). Критические значения (31) превышают квантили показателя V при $\alpha = 1 - \gamma_1 = 0{,}1$, который используется в классическом методе (см. табл. 1), и отличаются от рекомендуемых значений квантилей (24).

При использовании изложенного метода выявления аномальных результатов наблюдений следует иметь в виду, что значения P и γ нельзя занижать, так как при этом растет риск, что $(1-P)$% нормальных элементов выборки с вероятностью γ могут быть признаны недостоверными. Нецелесообразно также завышать значения P и γ, ибо при этом растет вероятность принять недостоверные данные за нормальные, что ведет к потере точности оценивания характеристик ТС, см. п. 3. Одну и ту же оценку ТИ можно получить при различных сочетаниях параметров Р и γ (если одно значение увеличивать, а другое снижать). Анализ показывает, что при выявлении недостоверных результатов наблюдений целесообразно выбирать значения Р и γ, примерно соответствующие друг другу, а уровень значимости $\alpha=1-\gamma$ выбирать в пределах $\alpha\in[0{,}01; 0{,}10]$, как это принято при использовании классических методов [3, 6].

Изложенный метод проверки достоверности результатов измерений характеристик ТС отличается высокой трудоемкостью. Далее рассматривается более простой метод анализа экстремальных элементов выборки.

6.2. Инженерный метод выявления аномальных результатов измерений

При выявлении аномальных данных по аналогии с классическим методом [3, 6], изложенным в п. 4, представляется целесообразным предположить, что в выборке только один элемент может выходить за границы ТИ. В соответствии с этим проанализируем возможности контроля достоверности элементов выборки из нормального распределения объема n на основе построения оценок толерантных границ для доли совокупности, равной $P=(n-1)/n$. Функции распределения оценок ТГ F(V) при этой величине P в зависимости от показателя (26) представлены на рис. 8.

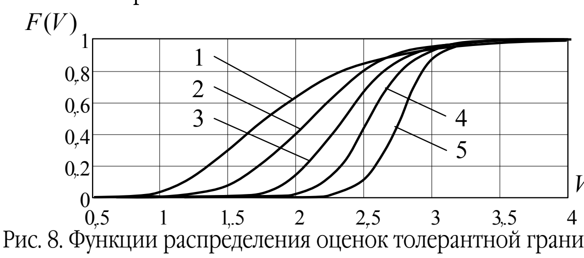

Рис. 8. Функции распределения оценок толерантной границы при значении доли $P=(n-1)/n$ при некоторых объемах выборки:
1–$n = 10$; 2–$n = 20$; 3–$n = 40$; 4–$n = 60$; 5–$n = 100$

Из рис. 8 видно, что с ростом объема выборки ТИ расширяется, а диапазон оценок его границ сокращается. Однако даже при значительном объеме выборки (при n=100) достоверные элементы выборки могут принадлежать множеству от $x_{\min} = \bar{M} - V\,\bar{\sigma}$ до $x_{\max} = \bar{M} + V\,\bar{\sigma}$, при этом показатель V может варьироваться в широких пределах $V \in [2{,}2;\ 3{,}5]$.

Для сокращения диапазона вариации показателя V учтем, что при проверке элементов выборки на достоверность классическими методами [3, 6] изначально предполагается, что экстремальные значения выборки являются достоверными. При этом выдвигается и проверяется нулевая гипотеза, что значения $x_{эк}$ относятся к генеральной совокупности, и соответственно задается невысокий уровень значимости $\alpha \in [0{,}05;\ 0{,}2]$. Функции распределения оценок ТГ, которые соответствуют этому диапазону значений α, показаны на рис. 9 (данные рис. 8 в увеличенном масштабе).

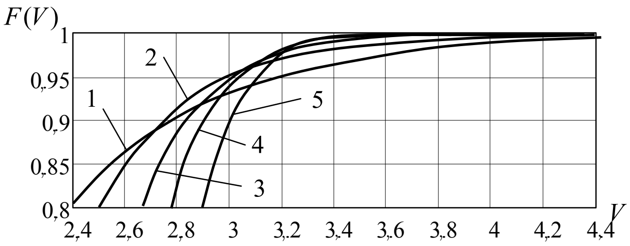

Рис. 6. Функции распределения оценок толерантной границы при значении доли $P=(n-1)/n$ при некоторых объемах выборки:
1–n =10; 2–n =20; 3–n =40; 4 –n =60; 5–n =100

Из рис. 9 следует, что при ограниченном уровне значимости α границы возможного выхода экстремальных значений выборки за пределы нормальной совокупности значительно сужаются. Критические значения показателя V при некоторых значениях α приведены в табл. 2. Эти значения определены по данным рис. 9 при $F(V_{кр})=1-\alpha_1$ и уровне значимости $\alpha_1 = 0{,}5\alpha$ для каждой ТГ.

Таблица 3

Критические значения показателя V

Уровни значимости α	0,20	0,10	0,05
Диапазоны критических значений квантили $V_{кр}$ при $n \in [10;\ 100]$	2,75-3,0	2,95-3,15	3,1-3,2 – при n >20; 3,2-3, при $n \in [10;\ 20]$

На основе полученных результатов можно рекомендовать следующий инженерный способ проверки экстремальных значений выборки на достоверность. Назначается уровень значимости α. По табл. 2 или по данным рис. 9 определяется критическое значение квантили $V_{кр}$. Из выборки временно устраняются экстремальные элементы, достоверность которых вызывает сомнение. На основе сокращенной выборки оцениваются МО и СКО по МНО или другими статистическими методами. Вычисляются показатели $V_{эк}$ для всех экстремальных значений по зависимости (28).

Решение по каждому экстремальному значению $x_{эк}$ принимается по правилу (30). Если гипотеза H_0 принимается, то значение $x_{эк}$ признается нормальным и возвращается в выборку.

Пример 3. При измерениях получены значения характеристики ТС: $\{x_i\}$ ={11,3; 9,2; 12,7; 15,6; 6,3; 10,6; 20,7; 12,4; 2,1; 9,8; 16,9; 10,9; 14,4; 8,1; 5,7}.

Предполагается, что характеристика имеет нормальное распределение. Необходимо проверить полученные результаты на достоверность при уровне значимости α=0,10.

Выборка имеет объем n=15. Подозрительными являются экстремальные элементы x_1 = 2,1 и x_2 = 20,7. Проверим эти значения на достоверность классическим методом (п. 4) и инженерным способом (п. 6.2).

В соответствии с классическим методом при n=15 и α=0,10 из табл. 1 находим критическое значение показателя $V_{кр}$=2,326. По зависимостям (13) по выборке определяем реализации оценок МО и СКО: $\hat{M} = 11{,}11$; $\hat{\sigma} = 4{,}70$. По зависимости (14) вычисляем показатель V для экстремальных значений: V_1=1,92; V_2=2,04. Эти значения меньше критической величины показателя. Поэтому в соответствии с правилом (15) оба экстремальных значения выборки следует признать достоверными.

В соответствии с инженерным методом при α=0,10 по табл.2 принимаем критическое значение показателя $V_{кр}$=3,1. Из выборки временно удаляем экстремальные значения и по сокращенной выборке получаем реализации оценок МО и СКО: $\hat{M} = 11{,}07$; $\hat{\sigma} = 1{,}94$. По зависимости (28) вычисляем показатель V для экстремальных значений: V_1=4,62; V_2=4,96. Полученные значения существенно выше критического показателя, поэтому в соответствии с правилом (30) оба экстремальных значения следует признать недостоверными и исключить из выборки. Если проанализировать элементы выборки, то такое решение в отличие от классического метода представляется более предпочтительным.

Выводы

На основе проведенных исследований получены следующие основные результаты.

Изложен метод несмещенных оценок (МНО), обеспечивающий наибольшую точность оценивания характеристик технических систем (ТС) по результатам измерений (п. 2).

Исследована точность возможных вариантов обработки результатов измерений и показана целесообразность исключения из выборки возможных аномальных элементов (п. 3).

Проведен анализ классических методов проверки результатов измерений на достоверность и показано, что они не всегда обеспечивают достаточную надежность выявления недостоверных результатов наблюдений (п. 4).

Предложен метод построения плотности вероятности оценок толерантных границ для нормальной случайной величины с использованием МНО (п. 5). На основе плотности вероятности оценок ТГ разработаны общий и инженерный методы проверки элементов выборки на достоверность (п. 6).

Разработанные методы могут использоваться для выявления аномальных результатов измерений характеристик ТС. Благодаря этому можно повысить точность оценивания характеристик ТС по экспериментальным данным, искаженным аномальными результатами измерений.

Литература

1. Вероятность и математическая статистика: Энциклопедия / под ред. Ю. В. Прохорова. М.: Большая Российская энциклопедия, 2003. – 911 с.

2. Волков Л. И., Лукин В. Л., Сухорученков Б. И. Методы статистического контроля надежности технических систем. Юбилейный: ЗАО «ПСТМ», 2008. – 332 с.

3. Линник Ю. В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений. М.: Физматгиз, 1962. – 352 с.

4. Новицкий П. В., Зограф И. А. Оценка погрешностей результатов измерений. Ленинград: Энергоатомиздат, 1985. – 248 с.

5. Пугачев В. С. Теория вероятностей и математическая статистика. М.: Физматлит, 2002. – 496 с.

6. Смирнов Н. В., Дунин-Барковский И. В. Курс теории вероятностей и математической статистики для технических приложений. М.: Наука, 1965. – 552 с.

7. Сухорученков Б. И., Меньшиков В. А. Методы анализа характеристик летательных аппаратов. М.: Машиностроение, 1995. – 475 с.

8. Сухорученков Б. И. Методы оценивания показателей безотказности по ограниченной выборке. // Сборник трудов. СИП РИА, 2006. Вып. 14. – С. 101-123.

Материал поступил в редакцию 06. 04. 2010 г.