УДК 681.31

Коваленко И.И., Гавриш Т.С.

ОЦЕНИВАНИЕ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ВЕРОЯТНОСТНЫХ РАСПРЕДЕЛЕНИЙ МАЛЫХ ВЫБОРОК ДАННЫХ

Рассматривается проблема оценивания статистических характеристик выборок данных малых объемов. Доказано, что одним из способов решения этой проблемы есть использование процедур непараметрического интервального оценивания статистических моментов.
Ключевые слова: *маленькая выборка, статистическое оценивание, непараметрическая статистика.*

Розглядається проблема оцінювання статистичних характеристик виборок даних малого об'єму. Доведено, що одним із способів вирішення цієї проблеми є використання процедур непараметричного інтервального оцінювання статистичних моментів.
Ключові слова: *мала виборка, статистичне оцінювання, непараметрична статистика.*

The problem of evaluation of statistical characteristics of small size data samples is examined. It is well-proven that one of decision methods of this problem is the use procedures of non-parametric interval evaluation of statistical moments.
Key words: *statistical evaluation, nonparametric statistics.*

ВВЕДЕНИЕ

На практике достаточно часто приходится работать в условиях ограниченных объемов выборок. Особенно остро это ощущают различные службы (технологические, планово-экономические, контроля качества и др.) предприятий, имеющих мелкосерийное производство. Такое же положение существует в производстве и эксплуатации дорогостоящих и высоконадежных технических изделий.

При анализе статистического материала ограниченного объема задача оценивания функции распределения вероятностей и ее характеристик (оценок статистических моментов) принимает проблематичный характер особенно для выборок очень малого объема, содержащих $n \leq 20$ значений.

В работе [1] рассмотрены параметрические подходы, позволяющие решать отмеченные задачи оценивания, однако получаемая при этом функция распределения имеет вид ступенчатой кривой, а процедуры вычисления статистических моментов (математического ожидания и дисперсии) трудоемки и не удобны в практике инженерных вычислений.

Одним из путей оценивания при работе с малой выборкой является вычисление интервальных оценок, что позволяет узнать точность и надежность точечных оценок. В работе [4] рассмотрены простые процедуры вычисления указанных точечных и интервальных оценок характеристик распределения, основанные на непараметрическом подходе. Основная ее идея базируется на результатах исследований, говорящих о том, что распределения данных, как правило, отличны от нормальных [3]. Поэтому предлагается строить точечные оценки, используя выборочные аналоги их теоретических характеристик, а для получения интервальных оценок – использовать асимптотическую нормальность выборочных моментов, основанную на центральной предельной теореме и теореме о наследовании сходимости. Поскольку принимается, что функция распределения произвольна (с точностью до условий регулярности типа существования моментов), то задачи доверительного оценивания характеристик распределения являются непараметрическими [4].

При этом в отличие от случая, когда данные распределены нормально и доверительные границы точечных оценок определяются с использованием квантилей распределения Стьюдента, предлагается те же доверительные границы определять по значениям функции стандартного нормального распределения с заданной доверительной вероятностью.

Целью статьи является исследование процедур непараметрического интервального оценивания статистических моментов для случая малых выборок (объемом $n \le 20$ значений) данных о трудоемкости изготовления бортовых секций корпуса контейнеровоза (табл. 1).

Таблица 1

Значения трудоемкостей изготовления бортовых секций контейнеровоза (n = 16 секций)

Номер секции	410	411	420	421	430	431	440	441	450	451	460	461	470	471	480	481
Трудоемкость (чел/ч)	120	115	110	118	175	196	240	180	126	110	110	118	150	178	190	205
Ранжированный ряд	110	110	110	115	118	118	120	126	150	175	178	180	190	196	205	240
Вероятности значений ряда (p_i)	0,1875			0,0625	0,1250		0,0625	0,0625	0,0625	0,0625	0,0625	0,0625	0,0625	0,0625	0,0625	0,0625

ИЗЛОЖЕНИЕ ОСНОВНОГО МАТЕРИАЛА

Предварительно введем в рассмотрение ряд соотношений, определяющих статистические характеристики случайных величин, которые будут использоваться при дальнейшем изложении материала:

- выборочный центральный момент (μ) первого порядка (выборочное среднее арифметическое):

$$\mu_1 = \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}; \tag{1}$$

- выборочный центральный момент второго порядка (выборочная дисперсия или среднее квадратическое отклонение):

$$\mu_2 = S^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}; \tag{2}$$

$$S = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}};$$

- выборочный третий центральный момент:

$$\mu_3 = \frac{(x_1 - \bar{x})^3 + (x_2 - \bar{x})^3 + \ldots + (x_n - \bar{x})^3}{n}; \tag{3}$$

- выборочный четвертый центральный момент:

$$\mu_4 = \frac{(x_1 - \bar{x})^4 + (x_2 - \bar{x})^4 + \ldots + (x_n - \bar{x})^4}{n}; \tag{4}$$

- коэффициент асимметрии:

$$\sqrt{\beta_1} = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2]^{\frac{3}{2}}}; \qquad (5)$$

- коэффициент эксцесса:

$$\beta_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2]^2} - 3. \qquad (6)$$

Некоторые подходы к подбору эмпирического распределения вероятностей при малых объемах выборок.

1. **Параметрический подход** к оцениванию распределений, если заведомо известна ограниченность возможных значений случайной величины x с одной стороны, предполагает пользоваться семействами логарифмически нормальных или гамма-распределений; при ограничении x сверху и снизу – семейством бета-распределений. Если ввести в рассмотрение такие показатели формы распределения, как как квадрат коэффициента асимметрии $\sqrt{\beta_1}$ и коэффициент эксцесса β_2, выраженные через моменты третьего (3) и четвертого (4) порядка, то можно указать области значений (β1, β2), в которых распределения принадлежат к тому или иному типу. С этой целью можно использовать график Пирсона, заимствованный из работы [5] (рис. 1).

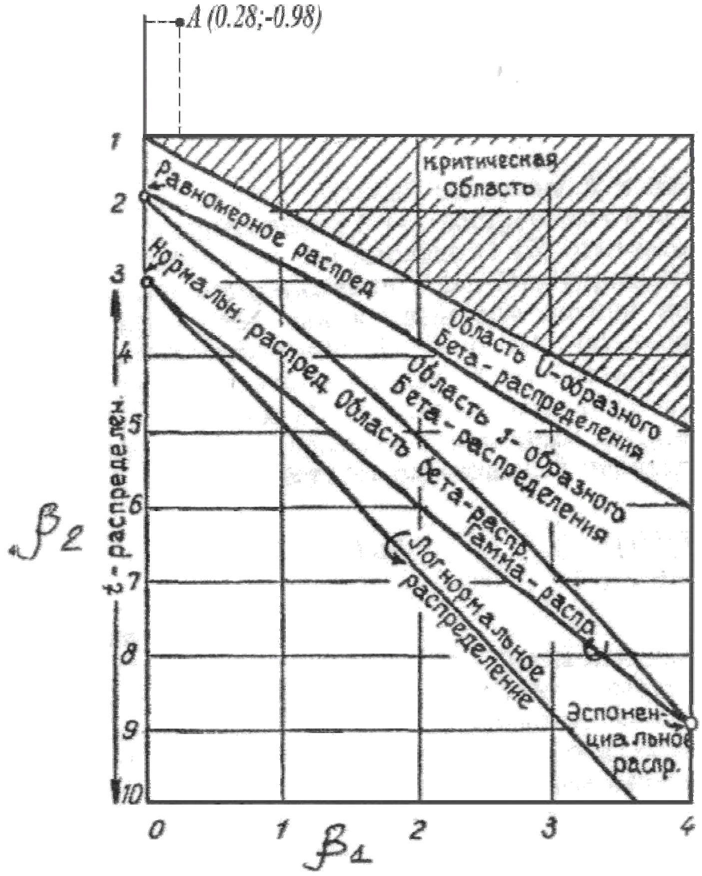

Рис. 1. Области для различных распределений в плоскости (β_1, β_2)

Для выбора модели распределения при данном подходе необходимо по выборке данных вычислить оценки β_1 и β_2 и отыскать соответствующую точку на рис. 1. Так, например, для нормального распределения β_1=0, β_2=3; для равномерного – β_1=0, β_2=1,8; для экспоненциального β_1=4, β_2=9; поэтому эти распределения отображаются на плоскости (β_1, β_2) каждое одной точкой. Другим распределениям, например, Стьюдента, логарифмически нормальному, гамма, соответствуют различные кривые, третьим – целые области.

2. В основу другого, **информационного подхода** к подбору распределения вероятностей может быть положена информация, как отражение случайной выборкой изучаемого явления. Для количественной оценки информации можно воспользоваться понятием энтропии как мерой неопределенности изучаемого явления. Энтропия H(X) является удобной мерой неопределенности законов распределения вероятностей, и между ними существует зависимость: величина энтропии, а следовательно количество информации определяются видом закона распределения. Более удобным информационным критерием является энтропийный коэффициент $K_Э$:

$$K_Э = \frac{\varDelta_Э}{S};$$

$$\varDelta_Э = \frac{1}{2} e^{H(X)};\ H(X) = -\sum_{i=1}^{n} p(x_i)\log(x_i). \tag{7}$$

Здесь $\varDelta_Э$ – энтропийная погрешность; $H(X)$ – энтропия; р – вероятности значений выборки x_i.

Таким образом, задача подбора распределения состоит в определении $K_Э$ для значений исследуемой выборки данных и сравнения величин полученных коэффициентов со значениями таблицы 2 [2].

Непараметрические статические оценки интервального оценивания характеристик произвольной функции распределения рассмотрим в соответствии с работой [4].

1. *Интервальное оценивание математического ожидания*. Точечной оценкой для математического ожидания в силу закона больших чисел является выборочное среднее $\bar{x}$. Нижняя и верхняя границы доверительного интервала (I_H, I_B) для $\bar{x}$ имеют вид:

$$I_H = \bar{x} - \frac{U(p)S}{\sqrt{n}};$$

$$I_B = \bar{x} + \frac{U(p)S}{\sqrt{n}}, \tag{8}$$

где p – доверительная вероятность, $U(p)$ – число, заданное равенством $\Phi(U(p))=(1+p)/2$, где $\Phi(x)$ – функция стандартного нормального распределения.

Например, при $p = 95\%$ (т.е. при $p = 0{,}95$) имеем $U(p)$=1,96. функция $U(p)$ имеется в большинстве литературных источников по теории вероятностей и математической статистике. В отличие от параметрического подхода, когда для определения доверительных интервалов используются квантили распределения Стьюдента, в данном выражении присутствует величина $U(p)$.

2. *Интервальное оценивание дисперсии*. Точечной оценкой дисперсии является выборочная дисперсия S^2. Доверительные границы находятся с помощью величины

$$d^2 = \frac{\mu_4 - \left(\frac{n-1}{n}\right)^4 S^4}{n}.$$

Нижняя и верхняя доверительные границы для дисперсии имеют вид:

$$I_H = S^2 - U(p)d\ ;$$

$$I_D = S^2 + U(p)d\ . \qquad (9)$$

где $U(p)$ – квантиль нормального распределения порядка (1+р)/2; d – положительный квадратный корень из величины d_2.

3. *Интервальное оценивание среднего квадратического отклонения*. Точечной оценкой является выборочное среднее квадратическое отклонение, т.е. неотрицательный квадратный корень из выборочной дисперсии. Дисперсия случайной величины – выборочного среднего квадратического отклонения S – оценивается как дробь $\frac{d^2}{4S^2}$. Нижняя и верхняя доверительные границы для среднего квадратического отклонения имеют вид:

$$I_H = S - \frac{U(p)d}{2S}\ ,\ I_B = S + \frac{U(p)d}{2S}\ . \qquad (10)$$

4. *Интервальное оценивание коэффициента вариации*. Коэффициент вариации $V_n = \frac{S}{\bar{x}}$ является важной статистической характеристикой, так как с его помощью оценивается изменчивость исследуемых данных. Коэффициент вариации V_n оценивается с помощью вспомогательной величины

$$D^2 = \frac{V_n^4 - \frac{V_n^2}{4} + \frac{\mu_4}{4S^2\bar{x}^2} - \frac{\mu_3}{\bar{x}^3}}{n}\ . \qquad (11)$$

Нижняя и верхняя доверительные границы для V_n имеют вид:

$$IH = V_n - U(p)D;\ \ IB = V_n + U(p)D, \qquad (12)$$

где D – положительный квадратный корень из величины D2.

ПРИМЕР РЕЗУЛЬТАТОВ ВЫЧИСЛЕНИЙ

1. По данным второй строки табл. 1 были подсчитаны значения коэффициентов асимметрии и эксцесса:

$$\beta_1 = 0{,}28,\ \beta_2 = -0{,}98$$

Точка на графике (рис. 1), полученная на пересечении β_1и β_2, принадлежит области, которая не охватывается распределениями, перечисленными выше (попадает в критическую область).

Далее с использованием данных (четвертая строка табл. 1) и выражений (7) был подсчитан энтропийный коэффициент:

$$H(X) = 3{,}58,\ S = 42{,}11,\ \Delta_Э = 17{,}898,\ K_Э = 0{,}425.$$

Сравнение полученного $K_Э$ с данными табл. 2 показывают, что распределение исследуемой выборки является экспоненциальным.

Таким образом, проведенные проверки убедительно говорят о том, что закон распределения выборки анализируемых данных не является нормальным. Это подтверждает основной вывод работы [4]: при анализе реальных данных следует использовать непараметрические доверительные границы.

Таблица 2

Законы распределений и соответствующие энтропийные коэффициент

Закон распределения	k	Закон распределения	k
1. Экспоненциальный $(\alpha = 1/4)$ $p(x)=\frac{1}{48}e^{-\sqrt[4]{\lvert x\rvert}}$	0,085	8. Лапласа $(\alpha = 1)$ $p(x)=\frac{1}{2}e^{-\lvert x\rvert}$	1,92
2. Экспоненциальный $(\alpha = 1/3)$ $p(x)=\frac{1}{12}e^{-\sqrt[3]{\lvert x\rvert}}$	0,424	9. t-распределение $n=6,\quad \upsilon=5$	1,97
3. Арксинусоидальный $p(x)=\begin{cases}0, & x<-a\\ \frac{1}{\pi\sqrt{a^2-x^2}}, & -a<x<a\\ 0, & x>a\end{cases}$	1,11	10. t-распределение $n=7,\quad \upsilon=6$	2,0
4. Экспоненциальный $(\alpha = 1/4)$ $p(x)=\frac{1}{48}e^{-\sqrt[4]{\lvert x\rvert}}$	1,35	11. t-распределение $n=8,\quad \upsilon=7$	2,013

Закінчення табл. 2

5. Равномерный (прямоугольный) $p(x)=\begin{cases}\frac{1}{b-a}, & a<x<b\\ 0, & x<a,\ x>b\end{cases}$	1,73	12. Симпсона (треугольный) $p(x)=\begin{cases}0, & x<a\\ \frac{4(x-a)}{(b-a)^2}, & a<x<\frac{b+a}{2}\\ \frac{4(b-x)}{(b-a)^2}, & \frac{a+b}{2}<x<b\quad x>b\end{cases}$	2,02
6. Экспоненциальный $(\alpha = 7)$ $p(x)=\frac{7}{2\Gamma(1/7)}e^{-\lvert x\rvert^7}$	1,87	13. t-распределение $n=11,\quad \upsilon=10$	2,047
7. Стьюдента (t-распределение) $p(x)=\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\upsilon\pi}\,\Gamma\left(\frac{\upsilon}{2}\right)\left(1+\frac{x^2}{\upsilon}\right)^{\frac{\upsilon+1}{\upsilon}}}$ $\upsilon=4,\ n=5$	1,90	14. Гауссов (нормальный) $(\alpha = 2)$ $p(x)=\frac{1}{\sqrt{\pi}}e^{-\frac{x^2}{2}}$	2,066

Таблица 3

Результаты вычислений статистических характеристик

Статистические характеристики	**Длина доверительного интервала**		**Разница**	**%**
	При нормальном распределении	**Непараметрический подход**		
$\bar{x}$	43,22	39,95	3,47	7,98
S^2	2457,15	1736,52	720,63	29,33
S	49,57	41,67	7,9	15,98
V	Нет методов нахождения [4]	0,0098		

2. На следующем этапе вычислений с использованием выражений (8)÷(12) были подсчитаны доверительные границы и доверительные интервалы для параметров $\bar{x}$, S_2, S, V, значения которых сведены в табл. 3. Анализ полученных результатов показывает, что длина доверительного интервала математического ожидания уменьшилась на 7,98 %, дисперсии – на 29,33 %, среднего квадратического отклонения – на 15,98 %.

ВЫВОДЫ

Таким образом, при использовании процедур непараметрического интервального оценивания статистических моментов для оценки трудоемкости изготовления бортовых секций корпуса контейнеровоза может быть увеличена точность точечных оценок, что, в свою очередь, позволяет вносить коррективы в календарные и сетевые графики выполнения работ.

Оценивание статистических характеристик вероятностных распределений малых выборок данных с использованием методов непараметрической статистики позволяет избежать заметно искаженных выводов, которые могут быть получены в предположении о нормальности распределения в ситуации, когда гипотеза нормальности не выполнена.

ЛИТЕРАТУРА

1. Гаскаров Д.В., Шаповалов В.И. Малая выборка. – М.: Статистика, 1978. – 248 с.
2. Коваленко И.И. Информационное описание согласованности экспертных оценок проектов // Сб. науч. трудов НУК, 2003. – № 6. – С. 141-149.
3. Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? // Журнал «Заводская лаборатория», 1991. – Т. 57. – № 7/ – С. 64-66.
4. Орлов А.И. Непараметрическое точечное и интервальное оценивание характеристик распределений // Журнал «Заводская лаборатория», 2004. – Т. 70. – № 5. – С. 65-70.
5. Тарасенко Ф.П. Непараметрическая статистика. – Томск: ТГУ, 1976. – 290 с.

Рецензенти: д.т.н., проф. Фісун М.Т.
д.т.н., проф. Данілов В.Я.

Стаття надійшла до редколегії 12.02.09