О РЕШЕНИИ ЗАДАЧ СТАТИСТИЧЕСКОГО АНАЛИЗА ИНТЕРВАЛЬНЫХ НАБЛЮДЕНИЙ*

Б. Ю. Лемешко, С. Н. Постовалов
Новосибирский государственный технический университет
Россия

Рассмотрены модели порождения интервальных наблюдений. Предложена процедура проверки гипотез о согласии теоретического закона распределения с интервальной выборкой. Сформулирована и доказана теорема об асимптотических свойствах границ вероятности согласия.

Рассмотрим следующую модель порождения исходных данных. Пусть в результате эксперимента наблюдаются значения y_i одномерной непрерывной случайной величины ξ

$$y_i = x_i + z_i, \tag{1}$$

где x_i — точное значение, а z_i — погрешность наблюдения. Если погрешность z_i не превышает по модулю некоторого числа d_i, то об истинном значении x_i можно сказать, что оно принадлежит интервалу $[a_i, b_i]$, где

$$a_i = y_i - d_i \qquad \text{и} \qquad b_i = y_i + d_i.$$

Таким образом, интервал $[a_i, b_i]$ содержит всю информацию об i-й реализации случайной величины ξ.

Определение 1. *Интервальным наблюдением называется интервал, содержащий не известное точно значение реализации случайной величины.*

Определение 2. *Интервальной выборкой объема n называется множество из n интервальных наблюдений*:

$$\mathbf{X}_n = \{[a_i, b_i] \mid a_i \le x_i \le b_i,\ a_i \in \mathbb{R},\ b_i \in \mathbb{R},\ i = 1, \ldots, n\}. \tag{2}$$

Такие модели в [1] называются реалистическими.

Замечание 1. К подобной математической модели могут привести процедуры группирования и цензурирования данных, хорошо известные в классической статистике. Отличие заключается в том, что интервалы группирования задаются априори, а в модели (1) границы интервалов связаны с наблюдениями. Тем не менее, несмотря на различные порождающие механизмы, все выводы, полученные для интервальной выборки (2), можно перенести на случай группированных, цензурированных и частично группированных выборок [2, 3].

Замечание 2. Интервалы $[a_i, b_i]$ в модели (2) могут быть бесконечными. Эта ситуация может возникнуть, например, в случае, когда стрелка измерительного прибора зашкаливает, и поэтому установить точное значение границы не представляется возможным.

Основную информацию о распределении случайной величины ξ исследователь получает по эмпирической функции распределения или гистограмме, на которые опираются статистические методы анализа. Однако для интервальной выборки построение этих функций, в общем случае, неоднозначно. Действительно, для построения гистограммы область определения случайной величины разбивается на k непересекающихся интервалов точками $X_0 < X_1 < \ldots < X_k$ и подсчитывается количество наблюдений, попавших в интервалы $(X_j, X_{j+1}]$, $j = 0, \ldots, k-1$. Если интервальное наблюдение $[a_i, b_i]$ покрывает точку разбиения X_j, то точное значение наблюдения можно отнести как к интервалу $[X_{j-1}, X_j]$, так и к интервалу $[X_j, X_{j+1}]$. Множество всех допустимых гистограмм можно получить простым перебором. Мощность этого множества резко возрастает с ростом объема выборки, поэтому использование гистограммы для наглядного представления данных и статистического анализа затруднительно.

Более простым оказывается построение множества всех допустимых эмпирических функций распределения. Упорядочим граничные точки интервалов:

$$a_{(1)} \le a_{(2)} \le \ldots \le a_{(n)}, \qquad b_{(1)} \le b_{(2)} \le \ldots \le b_{(n)}.$$

Предположим, что все точные значения наблюдений x_i совпали с левыми границами интервалов. Тогда эмпирическая функция распределения будет иметь следующий вид:

$$\overline{F_n}(x) = \begin{cases} 0, & x < a_{(1)}, \\ \dfrac{i}{n}, & a_{(i)} \le x < a_{(i+1)},\ i = 1, \ldots, n-1, \\ 1, & x \ge a_{(n)}. \end{cases}$$

Аналогично, если все точные значения совпали с правыми границами интервалов, эмпирическая функция распределения примет вид

$$\underline{F_n}(x) = \begin{cases} 0, & x < b_{(1)}, \\ \dfrac{i}{n}, & b_{(i)} \le x < b_{(i+1)},\ i = 1, \ldots, n-1, \\ 1, & x \ge b_{(n)}. \end{cases}$$

В общем случае эмпирическая функция распределения будет принадлежать множеству, ограниченному сверху $\overline{F_n}(x)$ и снизу $\underline{F_n}(x)$:

$$\underline{F_n}(x) \ \le \ F_n(x) \ \le \ \overline{F_n}(x) \qquad \forall x \in \mathbb{R}. \tag{3}$$

Следующий пример иллюстрирует вид $\overline{F_n}(x)$ и $\underline{F_n}(x)$ в зависимости от формы представления данных.

Пример 1. Была сгенерирована обычная выборка объемом 100 наблюдений. Ее эмпирическая функция распределения приведена на рис. 1, *а*. Рис. 1, *б* соответствует предположению, что наблюдения фиксировались с абсолютной погрешностью, а рис. 1, *в* – с относительной погрешностью в исходных данных. Наконец, в последнем случае (рис. 1, *г*) исходная выборка сгруппирована в 10 интервалов. На рис. 1, *б* – 1, *г* показаны графики функций $\underline{F_n}(x)$ и $\overline{F_n}(x)$.

Применение классических методов статистического анализа к интервальным выборкам в явном виде невозможно. Для адаптации известных методов обычным приемом может

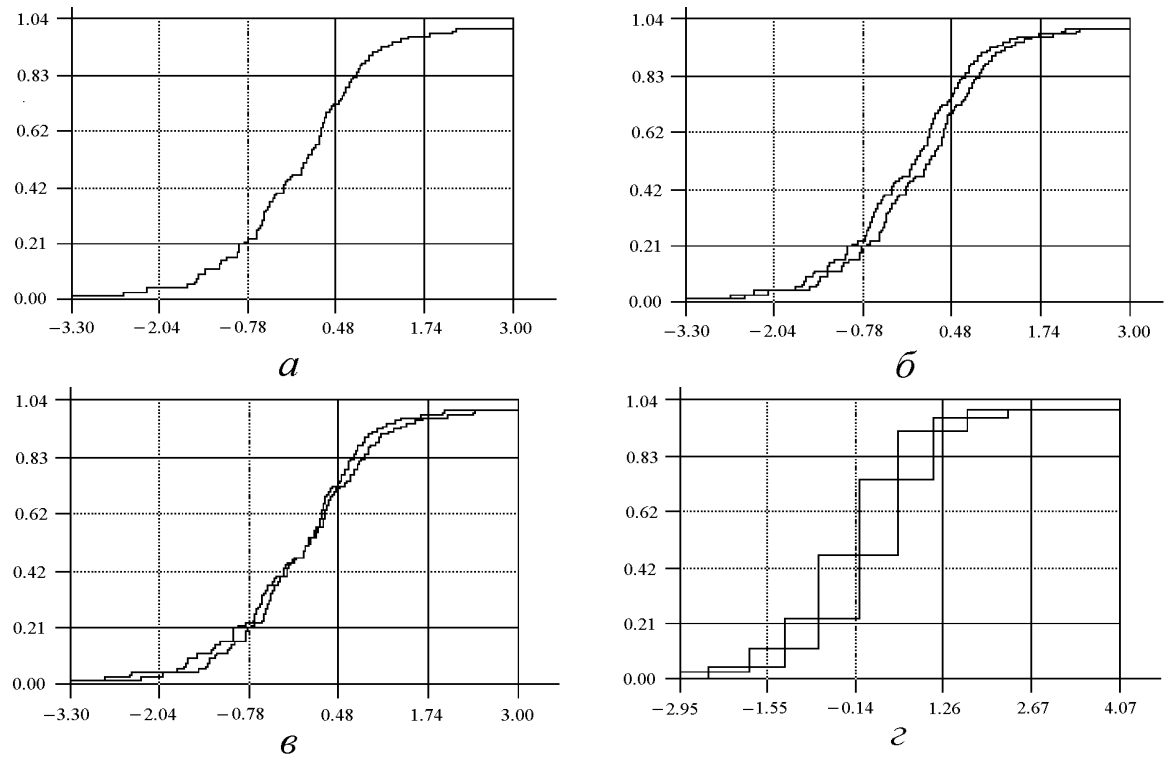

Рис. 1. Эмпирическая функция распределения обычной (*а*) и интервальных (*б*–*г*) выборок.

служить построение интервала неопределенности [4] интересующей исследователя статистики. В частности, множества допустимых гистограмм и эмпирических функций распределения, рассмотренные выше, построены в соответствии с этим принципом. В самом деле, если исходные данные известны с точностью до интервала, то естественным является описание статистики также с помощью интервала. При этом статистические выводы становятся менее определенными, но более надежными.

Далее рассмотрим процедуры проверки гипотез о согласии теоретического закона распределения случайной величины с интервальной выборкой. Аналогичные результаты были получены в [5, 6]. Gastaldi в [5] нашел верхнюю и нижнюю границы статистики Колмогорова в случае, когда выборка задана с пропусками данных, но при этом известно количество пропущенных наблюдений на интервалах между членами вариационного ряда (аналог частично группированной выборки). Орлов в [6] сформулировал общие подходы к проверке гипотез в случае интервального представления выборки и в качестве одного из примеров рассмотрел критерий Смирнова однородности двух выборок.

При проверке гипотез о согласии для найденного значения соответствующей статистики $S^\star$ вычисляется вероятность

$$p = P\{S > S^\star\} = \int\limits_{S^\star}^{\infty} g(s)ds,$$

где $g(s)$ — плотность распределения статистики при условии истинности нулевой гипотезы. При заданном уровне значимости α гипотеза о согласии не отвергается, если $p > \alpha$.

В дальнейшем вероятность $P\{S > S^\star\}$ будем называть вероятностью согласия. Когда выборка задана неточно, то статистика принадлежит интервалу $[\underline{S^\star}, \overline{S^\star}]$, где на основании (3) границы определяются следующим неравенством:

$$\underline{S^\star} = \inf_{\underline{F_n} \le F_n \le \overline{F_n}} S^\star(F_n, F) \le S^\star(F_n, F) \le \sup_{\underline{F_n} \le F_n \le \overline{F_n}} S^\star(F_n, F) = \overline{S^\star}. \tag{4}$$

Вероятность $P\{S > S^\star\}$ будет принадлежать интервалу $[p_{\min}, p_{\max}]$, где

$$p_{\min} = \int_{\overline{S^\star}}^{\infty} g(s)ds, \qquad p_{\max} = \int_{\underline{S^\star}}^{\infty} g(s)ds.$$

Тогда, при заданном уровне значимости α, гипотезу о согласии следует отклонить, если $p_{\max} \le \alpha$; гипотезу о согласии не следует отвергать, если $p_{\min} > \alpha$.

Рассмотрим использование этого подхода на примере **критерия Колмогорова**. Статистика критерия имеет вид

$$D_n = \sup_x |F_n(x) - F(x)|,$$

где $F_n(x)$ — эмпирическая функция распределения, $F(x)$ — теоретическая, согласие с которой проверяется, n — объем выборки. Преобразуем неравенство (3) к виду

$$\underline{F_n}(x) - F(x) \le F_n(x) - F(x) \le \overline{F_n}(x) - F(x),$$

$$F(x) - \overline{F_n}(x) \le F(x) - F_n(x) \le F(x) - \underline{F_n}(x).$$

Эти неравенства выполняются для всех , поэтому они сохраняются при взятии супремума:

$$\sup_x(\underline{F_n}(x) - F(x)) \le \sup_x(F_n(x) - F(x)) \le \sup_x(\overline{F_n}(x) - F(x)),$$
$$\sup_x(F(x) - \overline{F_n}(x)) \le \sup_x(F(x) - F_n(x)) \le \sup_x(F(x) - \underline{F_n}(x)).$$

Объединим эти неравенства в одно и, учитывая, что статистика D_n не может быть отрицательной, получим:

$$\begin{aligned} \underline{D_n} &= \max\{\sup_x(\underline{F_n}(x) - F(x)), \sup_x(F(x) - \overline{F_n}(x)), 0\} \le \\ &\le D_n = \max\{\sup_x(F_n(x) - F(x)), \sup_x(F(x) - F_n(x))\} \le \\ &\le \overline{D_n} = \max\{\sup_x(\overline{F_n}(x) - F(x)), \sup_x(F(x) - \underline{F_n}(x))\}. \end{aligned} \tag{5}$$

Аналогичные оценки верхней и нижней границ получены для статистик критериев согласия Смирнова, ω^2 и Ω^2 Мизеса [7]. Следующий пример иллюстрирует применение рассмотренного подхода.

Пример 2. Была сгенерирована выборка из 100 наблюдений с абсолютной погрешностью $\varepsilon = 0.05$, и проверено согласие с нормальным распределением с параметрами $\mu = -0.0786$ и $\sigma = 0.9916$ (рис. 2).

На диаграмме в правом верхнем углу цифрами обозначена вероятность согласия по критериям: 1 — отношения правдоподобия, 2 — χ^2 Пирсона, 3 — Колмогорова, 4 — Смирнова, 5, 6 — ω^2 и Ω^2 Мизеса. Заштрихованные области показывают интервалы неопределенности вероятности согласия.

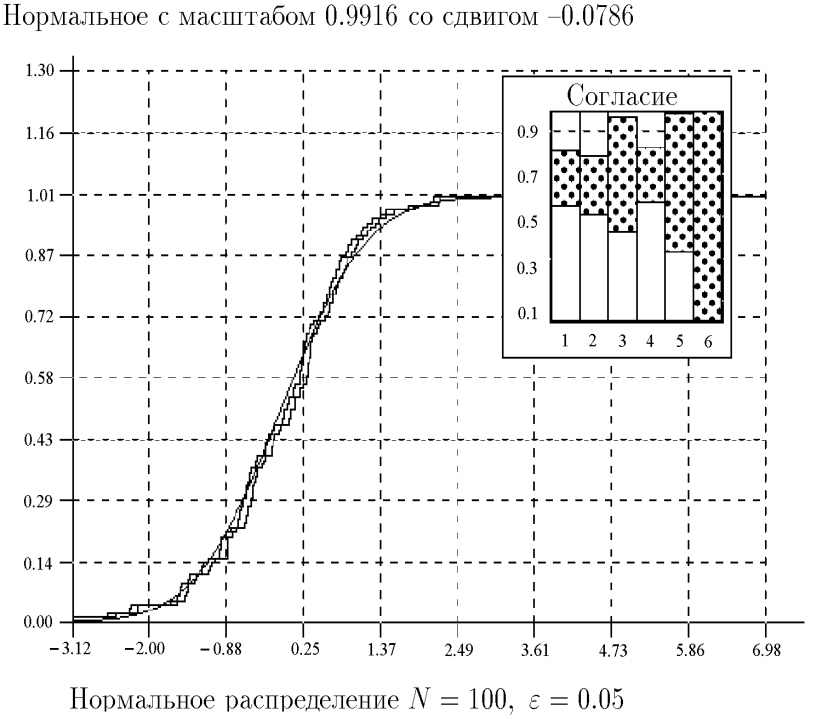

Рис. 2. Проверка согласия интервальной выборки с нормальным распределением.

На основании проверки гипотез можно сделать следующие выводы.

При уровне значимости $\alpha = 0.3$ гипотеза о согласии *не отвергается* по критериям отношения правдоподобия, χ^2 Пирсона, Колмогорова, Смирнова, ω^2.

При уровне значимости $\alpha = 0.5$ гипотеза о согласии *не отвергается* по критериям отношения правдоподобия, χ^2 Пирсона, Смирнова.

При уровне значимости $\alpha = 0.9$ гипотеза о согласии *отвергается* по критериям отношения правдоподобия, χ^2 Пирсона, Смирнова.

По остальным критериям однозначного вывода сделать невозможно.

Очевидно, что чем меньше интервал неопределенности $[p_{\min}, p_{\max}]$, тем более определенные выводы можно сделать. На длину интервала неопределенности $\Delta p = p_{\max} - p_{\min}$ влияют неопределенность в задании исходных данных, выбранная модель, критерий согласия и количество наблюдений. О том, как увеличение объема выборки влияет на Δp, говорит следующая теорема об асимптотических свойствах оценок границ статистики критерия Колмогорова по интервальной выборке.

Теорема. *Пусть задана последовательность интервальных выборок* $\mathbf{X}_n$, *для которых нижняя и верхняя границы эмпирической функции распределения* $\underline{F_n}(x)$ *и* $\overline{F_n}(x)$ *сходятся в равномерной метрике соответственно к* $\underline{F}(x)$ *и* $\overline{F}(x)$ *со скоростью* $O(1/n)$, *и* $\sup\limits_x(\overline{F}(x) - \underline{F}(x)) \geq c > 0$.

Пусть также $\mathcal{F}$ *— это множество всех функций распределения, непрерывных справа,* $p_{max}(F, \mathbf{X}_n)$ *и* $p_{min}(F, \mathbf{X}_n)$ *— соответственно верхняя и нижняя границы вероятности согласия по критерию Колмогорова.*

Тогда при $n \to \infty$:

1. $\forall F \in \mathcal{F}$, *таких что* $\forall x \left(\underline{F}(x) \leq F(x) \leq \overline{F}(x)\right)$,

$$\text{а)}\ \ p_{max}(F, \mathbf{X}_n) \to 1, \qquad \text{б)}\ \ p_{min}(F, \mathbf{X}_n) \to 0;$$

2. $\forall F \in \mathcal{F}$, *таких что* $\exists x \left(\left(F(x) < \underline{F}(x) \right) \vee \left(F(x) > \overline{F}(x) \right) \right)$,

$$\text{а)}\ \ p_{max}(F, \mathbf{X}_n) \to 0, \qquad \text{б)}\ \ p_{min}(F, \mathbf{X}_n) \to 0.$$

Доказательство. Статистика $S = \dfrac{(6nD_n + 1)^2}{18n}$ при достаточно большом n имеет распределение

$$P\{S > S^\star\} = 1 - K\left(\sqrt{\frac{S^\star}{2}} \right),$$

где $K(y) = \sum\limits_{k=-\infty}^{\infty} (-1)^k e^{-2k^2y^2}$ — функция распределения Колмогорова [8].

Для оценок границ $\underline{D_n}$ и $\overline{D_n}$ статистики D_n, определенных в (5), при $n \to \infty$ имеем:

$$p_{\min} = 1 - K\left(\frac{6n\overline{D_n} + 1}{6\sqrt{n}} \right) \to 0, \text{ если } \exists \lambda > 0 : \overline{D_n} \geq \lambda; \tag{6}$$

$$p_{\max} = 1 - K\left(\frac{6n\underline{D_n} + 1}{6\sqrt{n}} \right) \to \begin{cases} 1, \text{ если } (\underline{D_n} = 0) \vee (\underline{D_n} = O(1/n)); \\ 0, \text{ если } \exists \lambda > 0 : \underline{D_n} \geq \lambda. \end{cases} \tag{7}$$

Теперь для доказательства теоремы достаточно исследовать асимптотическое поведение оценок границ $\underline{D_n}$ и $\overline{D_n}$.

1. Пусть $F(x)$ — произвольная функция распределения, проходящая между $\underline{F}(x)$ и $\overline{F}(x)$.

а) Согласно (5) оценка снизу для нижней границы D_n имеет вид

$$\underline{D_n} = \max\left\{ \sup_x(\underline{F_n}(x) - F(x)), \sup_x(F(x) - \overline{F_n}(x)), 0 \right\}.$$

Если неравенство строгое: $\forall x\ \underline{F}(x) < F(x) < \overline{F}(x)$, то первые две величины в фигурных скобках будут отрицательными и $\underline{D_n} = 0$. Если $F(x)$ совпадает с $\underline{F}(x)$ на множестве $A \subseteq \mathbb{R}$ и с $\overline{F}(x)$ на множестве $B \subseteq \mathbb{R}$, то

$$\underline{D_n} = \max\left\{ \sup_{x \in A}(\underline{F_n}(x) - \underline{F}(x)), \sup_{x \in B}(\overline{F}(x) - \overline{F_n}(x)) \right\} \leq$$

$$\leq \max\left\{ \sup_{x \in A} |\underline{F_n}(x) - \underline{F}(x)|, \sup_{x \in B} |\overline{F}(x) - \overline{F_n}(x)| \right\} \ = \ O(1/n).$$

б) Пусть x_0 — точка, в которой

$$\sup_x(\overline{F}(x) - \underline{F}(x)) \geq \overline{F}(x_0) - \underline{F}(x_0) \geq c > 0.$$

Обозначим $a = \overline{F}(x_0) - F(x_0) \geq 0$ и $b = F(x_0) - \underline{F}(x_0) \geq 0$. Тогда $a + b = \overline{F}(x_0) - \underline{F}(x_0) \geq c > 0$ и $\max\{a, b\} \geq c/2$. Используя оценку сверху для верхней границы D_n и введенные обозначения, получим:

$$\overline{D_n} = \max\left\{ \sup_x(\overline{F_n}(x) - F(x)), \sup_x(F(x) - \underline{F_n}(x)) \right\} =$$

$$= \max\left\{\sup_x(\overline{F_n}(x) - \overline{F}(x)) + \sup_x(\overline{F}(x) - F(x)), \sup_x(F(x) - \underline{F}(x)) + \sup_x(\underline{F}(x) - \underline{F_n}(x))\right\} \geq$$
$$\geq \max\left\{O(1/n) + a\ ,\ b + O(1/n)\right\}.$$

Тогда $\exists\lambda > 0$ и $\exists n_0 : \forall n > n_0$

$$\overline{D_n} \geq \max\{a, b\} + O(1/n) \geq c/2 + O(1/n) \geq \lambda > 0.$$

2. Так как $p_{\max} \geq p_{\min}$, то (а) $\Rightarrow$ (б), и достаточно показать, что $p_{\max} \to 0$.

Пусть x_0 — точка, в которой $F(x) > \overline{F}(x)$ (аналогично рассматривается случай, когда $F(x) < \underline{F}(x)$). Обозначим $d = F(x_0) - \overline{F}(x_0) > 0$.

Тогда $\exists\lambda > 0$ и $\exists n_0 : \forall n > n_0$

$$\underline{D_n} = \max\left\{\sup_x(\underline{F_n}(x) - F(x)), \sup_x(F(x) - \overline{F_n}(x)), 0\right\} \geq$$
$$\geq F(x_0) - \overline{F_n}(x_0) = F(x_0) - \overline{F}(x_0) + \overline{F}(x_0) - \overline{F_n}(x_0) \geq$$
$$\geq d + O(1/n) \geq \lambda > 0.$$

Теорема доказана.

Поведение $p_{\max}$ и $p_{\min}$ иллюстрирует следующий пример.

Пример 3. Были сгенерированы три интервальные выборки с абсолютной погрешностью $\varepsilon = 0.05$, подчиненные одному и тому же закону распределения, объемом 100, 500 и 1000 наблюдений. Затем исследовано поведение $p_{\min}$ и $p_{\max}$ при проверке согласия по критерию Колмогорова с нормальным распределением, у которого параметр μ зафиксирован, а параметр σ изменялся от 0.5 до 1.5 (рис. 3). Хорошо видно, что с ростом количества наблюдений верхняя кривая согласия ($p_{\max}$) становится более крутой, а нижняя ($p_{\min}$) становится ближе к нулю. Это означает, что множество распределений, не отвергаемых по критерию согласия, уменьшается при одном и том же уровне значимости, но неопределенность при принятии решений о согласии для этих распределений увеличивается.

Из доказанной теоремы и рассмотренного примера вытекают два следующих практических соображения. С одной стороны, очевидно, что, опираясь на критерий Колмогорова, в случае интервальной выборки можно отсеять определенное множество законов распределения, не согласующихся с выборкой. С другой стороны, в этой же ситуации невозможно с точностью до параметров идентифицировать закон распределения, наиболее хорошо согласующийся с выборкой, если, например, для двух различных оценок параметров $p_{\min} = 0$ и $p_{\max} = 1$.

Таким образом, очевидно, что получение точечных оценок параметров распределений является процедурой, в значительной степени зависящей от степени оптимизма исследователя относительно соответствия выбранной модели исходным интервальным данным [9]. Действительно, нижнюю границу вероятности согласия можно рассматривать как случай

наихудшего расположения точных значений наблюдений в интервалах (“крайний пессимизм"), а верхнюю — как случай наилучшего расположения точных значений наблюдений (“крайний оптимизм").

Если исследователем априорно задана некоторая параметрическая модель $F(x, \theta)$, то верхняя и нижняя границы искомой функции распределения также должны принадлежать этой модели:

$$\underline{F}(x) = F(x, \theta_1), \qquad \overline{F}(x) = F(x, \theta_2).$$

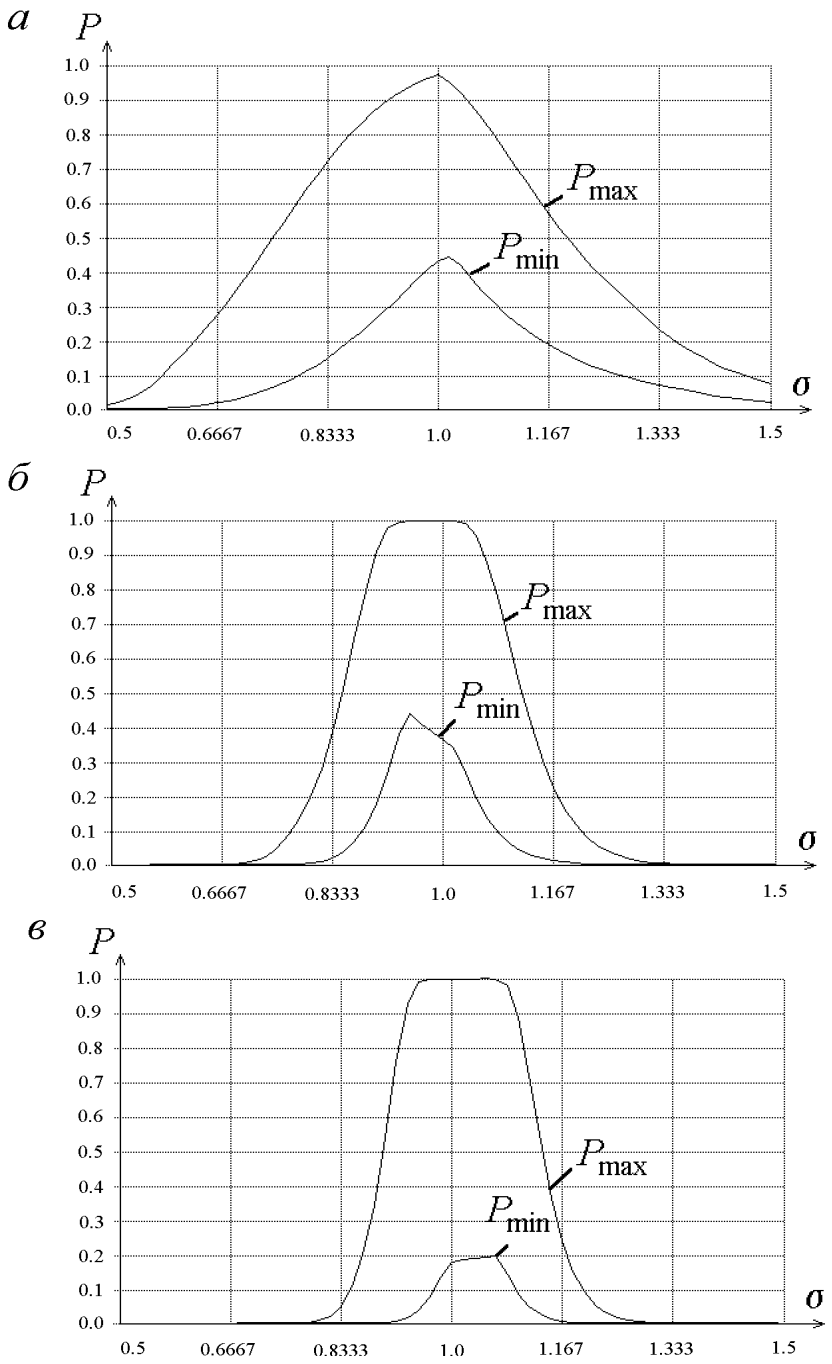

Рис. 3. Согласие интервальных выборок разного объема с нормальным распределением: *а* — 100 наблюдений, *б* — 500 наблюдений, *в* — 1000 наблюдений.

В случае скалярного параметра θ, используя $\underline{F_n}(x,\theta)$ и $\overline{F_n}(x,\theta)$, мы можем естественным образом получить интервальную оценку параметра, а в случае векторного параметра — оценить область его допустимых значений $T = \{\theta \in \Omega \mid \forall x\ F(x,\theta_1) \le F(x,\theta) \le F(x,\theta_2)\}$.

Заметим, что использование параметрической модели для описания интервальной выборки может оказаться не всегда приемлемым, так как верхняя и нижняя границы эмпирической функции распределения могут сходиться в общем случае к законам из разных параметрических семейств.

С учетом всего вышеизложенного можно сделать следующие выводы. При увеличении объема интервальной выборки для целого множества априори допустимых для описания данной случайной величины распределений длина интервала неопределенности вероятности согласия растет и стремится к единице. Это значит, что функцию распределения случайной величины, наблюдения которой фиксируются с неустранимой погрешностью, невозможно определить точно, даже при очень большом числе экспериментов. Для описания такой случайной величины лучше либо использовать интервальные оценки параметров функции распределения, либо по отдельности аппроксимировать верхнюю и нижнюю границы эмпирической функции распределения.

Список литературы

[1] Орлов А. И. О развитии реалистической статистики. В *"Стат. методы оценивания и проверки гипотез"*. Межвуз. сб. науч. трудов, Пермский ун-т, Пермь, 1990, 89–99.

[2] Лемешко Б. Ю., Постовалов С. Н. Статистический анализ одномерных наблюдений по частично группированным данным. *Изв. высших учебных заведений. Физика* **38**, №9, 1995, 39–45.

[3] Лемешко Б. Ю., Постовалов С. Н. К использованию непараметрических критериев по частично группированным данным. В *"Сб. науч. трудов НГТУ"*. Новосибирск, №2, 1995, 21–30.

[4] Канторович Л. В. О некоторых новых подходах к вычислительным методам и обработке наблюдений. *Сиб. мат. журн.* **3**, №5, 1962, 701–709.

[5] Gastaldi T. A Kolmogorov-Smirnov test procedure involving a possibility censored or truncated sample. *Communications in statistics. Theory and methods* **22**, №1, 1993, 31–39.

[6] Орлов А. И. Некоторые алгоритмы реалистической статистики. В *"Стат. методы оценивания и проверки гипотез. Межвуз. сб. науч. трудов"*. Пермский ун-т, Пермь, 1991, 77–86.

[7] Лемешко Б. Ю., Постовалов С. Н. Статистический анализ наблюдений, имеющих интервальное представление. В *"Сб. науч. трудов НГТУ"*. Новосибирск, №1, 1996, 3–12.

[8] Большев Л. Н., Смирнов Н. В. *Таблицы математической статистики.* Наука, М., 1965.

[9] Кузнецов В. П. *Интервальные статистические модели.* Радио и связь, М., 1991.