СБОРНИК НАУЧНЫХ ТРУДОВ НГТУ. - 1995. - № 2. - 21 - 30
УДК 519.2 (075.8)

К ИСПОЛЬЗОВАНИЮ НЕПАРАМЕТРИЧЕСКИХ КРИТЕРИЕВ ПО ЧАСТИЧНО ГРУППИРОВАННЫМ ДАННЫМ

Б.Ю. ЛЕМЕШКО♣, С.Н. ПОСТОВАЛОВ♦

Рассмотрены вопросы применения непараметрических критериев согласия Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса в ситуациях, когда исходные наблюдения представляют собой группированные или частично группированные выборки. Приведены результаты программной реализации и моделирования.

Характер выборочных данных, по которым осуществляется анализ распределений, может быть различным. Наиболее общим случаем является частично группированная выборка [1]. Выборка является *негруппированной*, если выборочные значения представляют собой индивидуальные значения наблюдений из области определения случайной величины. Выборка является *группированной*, если область определения случайной величины разбита на k непересекающихся интервалов граничными точками

$$x_0 < x_1 < \ldots < x_{k-1} < x_k,$$

где x_0 - нижняя грань области определения случайной величины X; x_k - верхняя грань области определения случайной величины X, и зафиксированы количества наблюдений n_i, попавших в i -й интервал значений. Выборка является *частично группированной*, если имеющаяся в нашем распоряжении информация связана с множеством непересекающихся интервалов, которые делят область определения случайной величины так, что каждый интервал принадлежит к одному из двух типов:

а) i -й интервал принадлежит к первому типу, если число n_i известно, но индивидуальные значения x_{ij}, $j=\overline{1,n_i}$ неизвестны;

б) i -й интервал принадлежит ко второму типу, если известно не только число n_i, но и все индивидуальные значения x_{ij}, $j=\overline{1,n_i}$.

Область определения случайной величины в этом случае можно представить в виде $X = X_{(1)} \cup X_{(2)}$, где $X_{(1)}$ - множество интервалов первого типа, а $X_{(2)}$- множество интервалов второго типа.

В развиваемой на кафедре прикладной математики НГТУ программной системе "Статистический анализ одномерных наблюдений случайных величин" [2] проверка гипотез о согласии осуществляется по критериям χ^2- Пирсона,

---

♣ Доцент кафедры прикладной математики, канд. техн. наук
♦ Аспирант кафедры прикладной математики

отношения правдоподобия, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса. Использование совокупности критериев даёт возможность принимать более обоснованное решение, а при противоречивости выводов по отдельным критериям - формировать компромиссный критерий и делать окончательный вывод с его учетом.

При проверке гипотез о согласии для найденного значения соответствующей статистики S^* вычисляется вероятность

$$p = P\{S > S^*\} = \int_{S^*}^{\infty} g(s)ds,$$

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 тех случаях, когда исходные данные представляют собой группированную или частично группированную выборку, применение непараметрических критериев Колмогорова [3], Смирнова [4], ω^2 и Ω^2 Мизеса [5-7] *невозможно*, так как соответствующие статистики предусматривают, что известны все индивидуальные значения наблюдений.

В данной работе предлагается следующий подход к использованию непараметрических критериев в ситуации группированных и частично группированных данных. Для статистики соответствующего критерия находятся оценки сверху и снизу ($\underline{S}^*$ и $\overline{S}^*$) и на основании верхней и нижней границ вероятности согласия ($p_{\max} = P\{S > \underline{S}^*\}$ и $p_{\min} = P\{S > \overline{S}^*\}$), которые позволяют оценить степень согласия теоретического и эмпирического законов распределения, делаются статистические выводы.

1. КРИТЕРИЙ КОЛМОГОРОВА

Соответствующая статистика 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. Закон распределения этой статистики найден Колмогоровым [3].

Пусть задана частично группированная выборка. Введем обозначения:

$$N_i = \sum_{j=1}^{i} n_j; \quad N_{-1} = 0; \quad N_0 = n_0; \quad N_{k-1} = n; \quad N_{ij} = N_i + j$$

Эмпирическая функция распределения $F_n(x)$ полностью определена для интервалов второго типа:

$$F_n(x) = N_{i-1,j}/n; \quad \forall x \in \left[x_{ij}, x_{i,j+1}\right) \subseteq \left[x_i, x_{i+1}\right) \subseteq X_{(2)};$$
$$j = 1, \ldots, n_i; \quad (x_{i,n_i+1} \equiv x_{i+1}),$$

а также во всех граничных точках x_i, $i = 0, \ldots, k$:

$$F_n(x_i) = N_{i-1}/n.$$

На интервалах первого типа нам известно только, что

$$G_k^-(x) = N_{i-1}/n \le F_n(x) \le N_i/n = G_k^+(x),\ x \in \left[x_i, x_{i+1}\right) \subseteq X_{(1)}.$$

Вычислить непосредственное значение статистики D_n по частично группированной выборке невозможно, так как известны не все индивидуальные значения наблюдений. Поэтому найдем для неё оценки снизу и сверху. Мы можем ограничить D_n снизу следующим образом:

$$D_n = \sup_x |F_n(x) - F(x)| = \max\left\{\sup_{X_{(1)}} |F_n(x) - F(x)|, \sup_{X_{(2)}} |F_n(x) - F(x)|\right\};$$

$$D_n \ge \max\left\{\max_{i=0,\ldots,k} |N_{i-1}/n - F(x_i)|, \sup_{X_{(2)}} |F_n(x) - F(x)|\right\} = \underline{D_{nk}}.$$

Найдем теперь оценку сверху. Функции $G_k^+(x)$ и $G_k^-(x)$ построены так, что $\forall x \in X_{(1)}$: $G_k^-(x) \le F_n(x) \le G_k^+(x)$. Тогда

$$G_k^-(x) - F(x) \le F_n(x) - F(x) \le G_k^+(x) - F(x);$$
$$F(x) - G_k^+(x) \le F(x) - F_n(x) \le F(x) - G_k^-(x).$$

Обозначив через $A = \left\{x \in X_{(1)}: F_n(x) \ge F(x)\right\}$ и $B = \left\{x \in X_{(1)}: F_n(x) < F(x)\right\}$, найдем, что

$$D_n = \max\left\{\sup_{A \subseteq X_{(1)}} (F_n(x) - F(x)), \sup_{B \subseteq X_{(1)}} (F(x) - F_n(x)), \sup_{X_{(2)}} |F_n(x) - F(x)|\right\};$$

$$D_n \le \max\left\{\sup_{A \subseteq X_{(1)}} (G_k^+(x) - F(x)), \sup_{B \subseteq X_{(1)}} (F(x) - G_k^-(x)), \sup_{X_{(2)}} |F_n(x) - F(x)|\right\};$$

$$D_n \le \max\left\{\sup_{X_{(1)}} |G_k^+(x) - F(x)|, \sup_{X_{(1)}} |F(x) - G_k^-(x)|, \sup_{X_{(2)}} |F_n(x) - F(x)|\right\} = \overline{D_{nk}}.$$

Таким образом, $\underline{D_{nk}} \le D_n \le \overline{D_{nk}}$ и так как функция распределения является монотонно возрастающей, то $p_{\min} \le p \le p_{\max}$, где $p = 1 - K(g(D_n))$,

$p_{\min} = 1 - K\left(g(\overline{D_{nk}})\right)$, $p_{\max} = 1 - K\left(g(\underline{D_{nk}})\right)$, $K(\lambda)$ - функция распределения Колмогорова, а $g(y) = \sqrt{(6ny+1)^2 / 36n}$ [7].

2. КРИТЕРИЙ СМИРНОВА

Статистика Смирнова имеет вид

$$D_n^+ = \sup_x \left(F_n(x) - F(x)\right).$$

Эта статистика ограничивается в случае частично группированной выборки аналогично статистике Колмогорова:

$$D_n^+ \geq \max\left\{ \max_{i=0,\ldots,k} \left(N_{i-1}/n - F(x_i)\right), \sup_{X_{(2)}} \left(F_n(x) - F(x)\right) \right\} = \underline{D_{nk}^+};$$

$$D_n^+ \leq \max\left\{ \sup_{X_{(1)}} \left(G_k^+(x) - F(x)\right), \sup_{X_{(1)}} \left(F(x) - G_k^-(x)\right), \sup_{X_{(2)}} \left(F_n(x) - F(x)\right) \right\} = \overline{D_{nk}^+}.$$

Таким образом, $\underline{D_{nk}^+} \leq D_n^+ \leq \overline{D_{nk}^+}$ и так как функция распределения является монотонно возрастающей, то $p_{\min} \leq p \leq p_{\max}$, где $p = 1 - e^{-h(D_n^+)/2}$, $p_{\min} = 1 - e^{-h(\overline{D_{nk}^+})/2}$, $p_{\max} = 1 - e^{-h(\underline{D_{nk}^+})/2}$, где $h(y) = \sqrt{(6ny+1)^2 / 9n}$ [7].

3. КРИТЕРИЙ ω^2 МИЗЕСА

Статистика критерия имеет вид

$$n\omega_n^2 = n\int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 dF(x) = \frac{1}{12n} + \sum_{i=0}^{k-1}\sum_{j=1}^{n_i}\left[F(x_{ij}) - \frac{2N_{i-1,j} - 1}{2n}\right]^2.$$

Предельный закон распределения для этой статистики был найден Н.В. Смирновым [5,6]. Обозначим через $S_i = \sum_{j=1}^{n_i}\left[F(x_{ij}) - \frac{2N_{i-1,j} - 1}{2n}\right]^2$. Для частично группированной выборки статистика имеет вид

$$n\omega_n^2 = \frac{1}{12n} + \sum_{(1)} S_i + \sum_{(2)} S_i,$$

где (1) и (2) означает суммирование по интервалам первого или второго типов. Найдем для статистики оценки сверху и снизу. Нетрудно показать, что из монотонности функции распределения следуют неравенства:

$$S_i \le \max\left\{\sum_{j=1}^{n_i}\left[F(x_i)-\frac{2N_{i-1,j}-1}{2n}\right]^2, \sum_{j=1}^{n_i}\left[F(x_{i+1})-\frac{2N_{i-1,j}-1}{2n}\right]^2\right\};$$

$$S_i \ge \sum_{j\in A_i}\left[F(x_i)-\frac{2N_{i-1,j}-1}{2n}\right]^2 + \sum_{j\in B_i}\left[F(x_{i+1})-\frac{2N_{i-1,j}-1}{2n}\right]^2,$$

где $A_i = \left\{j=1,\ldots,n_i: F(x_i) > \frac{2N_{i-1,j}-1}{2n}\right\}$ и $B_i = \left\{j=1,\ldots,n_i: F(x_{i+1}) < \frac{2N_{i-1,j}-1}{2n}\right\}$, причем $A_i \cap B_i = \varnothing$.

Тогда

$$n\omega_n^2 \ge \frac{1}{12n} + \sum_{(1)}\left\{\sum_{j\in A_i}\left[F(x_i)-\frac{2N_{i-1,j}-1}{2n}\right]^2 + \sum_{j\in B_i}\left[F(x_{i+1})-\frac{2N_{i-1,j}-1}{2n}\right]^2\right\} +$$

$$+\sum_{(2)}\sum_{j=1}^{n_i}\left[F(x_{ij})-\frac{2N_{i-1,j}-1}{2n}\right]^2 = \underline{\omega_{nk}^2},$$

$$n\omega_n^2 \le \frac{1}{12n} + \sum_{(1)}\max\left\{\sum_{j=1}^{n_i}\left[F(x_i)-\frac{2N_{i-1,j}-1}{2n}\right]^2, \sum_{j=1}^{n_i}\left[F(x_{i+1})-\frac{2N_{i-1,j}-1}{2n}\right]^2\right\} +$$

$$+\sum_{(2)}\sum_{j=1}^{n_i}\left[F(x_{ij})-\frac{2N_{i-1,j}-1}{2n}\right]^2 = \overline{\omega_{nk}^2}.$$

В результате получаем интервал, в который попадает неизвестная точно статистика $n\omega_n^2$: $p_{\min} = 1 - a1\left(\overline{\omega_{nk}^2}\right)$, $p_{\max} = 1 - a1\left(\underline{\omega_{nk}^2}\right)$, где $a1(\lambda)$ - предельная функция распределения статистики $n\omega_n^2$ [7].

4. КРИТЕРИЙ Ω^2 МИЗЕСА

Статистика имеет вид

$$N\Omega_n^2 = -N - 2\sum_{i=0}^{k-1}\sum_{j=1}^{n_i}\left\{\frac{2N_{i-1,j}-1}{2n}\ln F(x_{ij}) + \left(1-\frac{2N_{i-1,j}-1}{2n}\right)\ln(1-F(x_{ij}))\right\}.$$

Преобразуем её к следующему виду:

$$N\Omega_n^2 = -N - 2\ln\prod_{(1)+(2)}\prod_{j=1}^{n_i}\left(F(x_{ij})\right)^{\frac{2N_{i-1,j}-1}{2n}}\left(1-F(x_{ij})\right)^{1-\frac{2N_{i-1,j}-1}{2n}}.$$

и найдем для нее оценки сверху и снизу по частично группированной выборке. Из свойства монотонности функции распределения следует, что

$$\left(F(x_{ij})\right)^{\frac{2N_{i-1,j}-1}{2n}}\left(1-F(x_{ij})\right)^{1-\frac{2N_{i-1,j}-1}{2n}} \le \left(F(x_{i+1})\right)^{\frac{2N_{i-1,j}-1}{2n}}\left(1-F(x_i)\right)^{1-\frac{2N_{i-1,j}-1}{2n}};$$

$$\left(F(x_{ij})\right)^{\frac{2N_{i-1,j}-1}{2n}}\left(1-F(x_{ij})\right)^{1-\frac{2N_{i-1,j}-1}{2n}} \ge \left(F(x_i)\right)^{\frac{2N_{i-1,j}-1}{2n}}\left(1-F(x_{i+1})\right)^{1-\frac{2N_{i-1,j}-1}{2n}}.$$

Отсюда

$$N\Omega_n^2 \le -N - 2\ln\prod_{(1)}\prod_{j=1}^{n_i}\left(F(x_i)\right)^{\frac{2N_{i-1,j}-1}{2n}}\left(1-F(x_{i+1})\right)^{1-\frac{2N_{i-1,j}-1}{2n}} -$$

$$-2\ln\prod_{(2)}\prod_{j=1}^{n_i}\left(F(x_{ij})\right)^{\frac{2N_{i-1,j}-1}{2n}}\left(1-F(x_{ij})\right)^{1-\frac{2N_{i-1,j}-1}{2n}} = \underline{\Omega_{nk}^2};$$

$$N\Omega_n^2 \ge -N - 2\ln\prod_{(1)}\prod_{j=1}^{n_i}\left(F(x_{i+1})\right)^{\frac{2N_{i-1,j}-1}{2n}}\left(1-F(x_i)\right)^{1-\frac{2N_{i-1,j}-1}{2n}} -$$

$$-2\ln\prod_{(2)}\prod_{j=1}^{n_i}\left(F(x_{ij})\right)^{\frac{2N_{i-1,j}-1}{2n}}\left(1-F(x_{ij})\right)^{1-\frac{2N_{i-1,j}-1}{2n}} = \overline{\Omega_{nk}^2}.$$

В результате получаем интервал, в который попадает неизвестная точно статистика $n\Omega_n^2$: $p_{\min} = 1 - a2\left(\overline{\Omega_{nk}^2}\right)$, $p_{\max} = 1 - a2\left(\underline{\Omega_{nk}^2}\right)$, где $a2(\lambda)$ - предельная функция распределения статистики $n\Omega_n^2$ [7].

Следовательно, для рассмотренных критериев, при заданном уровне значимости α, *возможны следующие выводы: гипотезу о согласии следует отклонить, если* $p_{\max} \le \alpha$; *гипотезу о согласии не следует отвергать, если* $p_{\min} > \alpha$.

Частным случаем частично-группированной выборки является *цензурированная выборка*, и поэтому все сделанные выводы справедливы и для этого класса выборок [8].

Очевидно, что интервал неопределенности $\Delta p = p_{\max} - p_{\min}$ зависит от объема выборки, от числа интервалов, от метода группирования, от используемого критерия и от степени согласия выборки с гипотетическим распределением. Неопределенность в статистических выводах минимальна, когда в каждом интервале содержится по одному наблюдению. С уменьшением числа интервалов группирования точность статистических выводов будет падать. Из результатов тестирования на разных выборках было отмечено, что верхняя граница интервала $p_{\max}$ более чувствительна к отклонению эмпирической функции распределения от теоретической. При *плохом согласии*

верхняя граница вероятности согласия по критериям Колмогорова и Смирнова может оказаться ниже соответствующей вероятности по критериям χ^2 - Пирсона и отношения правдоподобия (рис.1). Это означает, что если по критериям χ^2 - Пирсона и отношения правдоподобия при уровне значимости $\alpha = 0.05$ гипотеза о согласии будет принята, то по критериям Колмогорова и Смирнова она будет отвергаться. На рис. 1, 3 и 4 полученные значения $p_{\min}$ и $p_{\max}$ для соответствующих статистик отображены двумя колонками в правом верхнем углу.

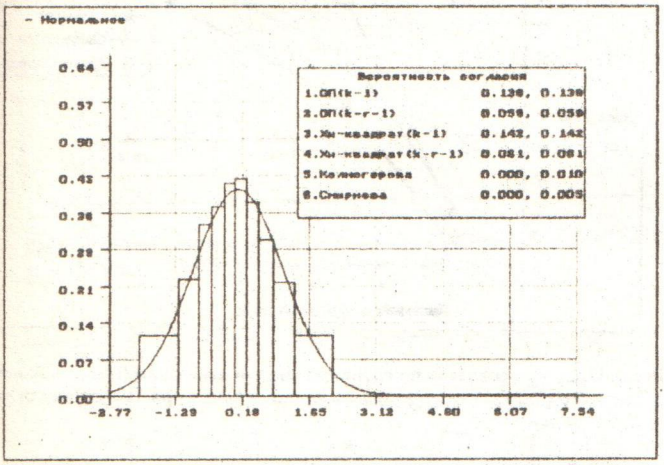

Рис.1. Результаты проверки согласия по группированной выборке (объем 4000 наблюдений), разбитых на 10 интервалов группирования

В случае, когда результаты проверки гипотез, полученные по параметрическим критериям, дают *очень хорошее* согласие, интервал $[p_{\min}, p_{\max}]$ в основном зависит от среднего числа наблюдений в одном интервале группирования, как это хорошо видно из графика на рис.2, где показана зависимость Δp от числа наблюдений в одном интервале (группированные выборки получены разбиением на интервалы равной частоты). Следует отметить, что более мощными при проверке соответствующих гипотез оказываются критерии Колмогорова и Смирнова по сравнению с критериями ω^2 и Ω^2 Мизеса.

Относительно критерия Ω^2 Мизеса необходимо дополнительно сделать следующее замечание. Многочисленные вычислительные эксперименты на реальных и имитируемых выборках позволяют утверждать, что по мощности он уступает остальным используемым критериям. Нередки случаи, когда гипотеза о

согласии в соответствии с этим критерием принимается, но отвергается по другим.

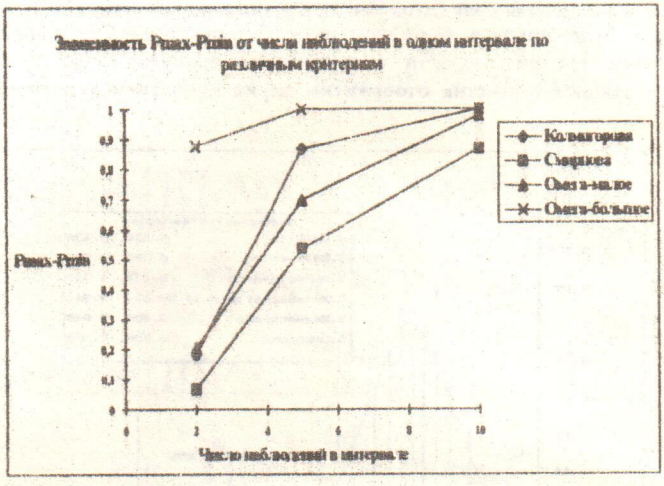

Рис.2. Проверка согласия по группированным выборкам (объём 100 наблюдений) с 10, 20 и 50 интервалами группирования при "хорошем согласии"

На рис. 3 и 4 приведены примеры проверки согласия по группированным данным нормального распределения с разными параметрами. Ступенчатые функции представляют собой верхнюю и нижнюю предельные границы для неизвестной эмпирической функции распределения, а плавная кривая - функцию распределения, согласие с которой проверяется.

При заданном уровне значимости $\alpha = 0.15$, гипотеза о согласии выборки с нормальным распределением с параметрами $\mu = 0.10$ и $\sigma = 1.10$ проходит по критериям χ^2- Пирсона, отношения правдоподобия, ω^2 и Ω^2 Мизеса и отвергается по критериям Колмогорова и Смирнова. Гипотеза о согласии выборки с нормальным законом с параметрами $\mu = 0.0$ и $\sigma = 1.0$ при том же уровне значимости не отвергается ни одним из критериев.

Таким образом, с одной стороны применение непараметрических критериев в случае группированных или частично-группированных выборок является оправданным и дает исследователю дополнительную информацию для размышления в ситуации, когда близость эмпирического и теоретического законов сомнительна. С другой - при близких эмпирическом и теоретическом законах статистические выводы по непараметрическим критериям малоинфор-

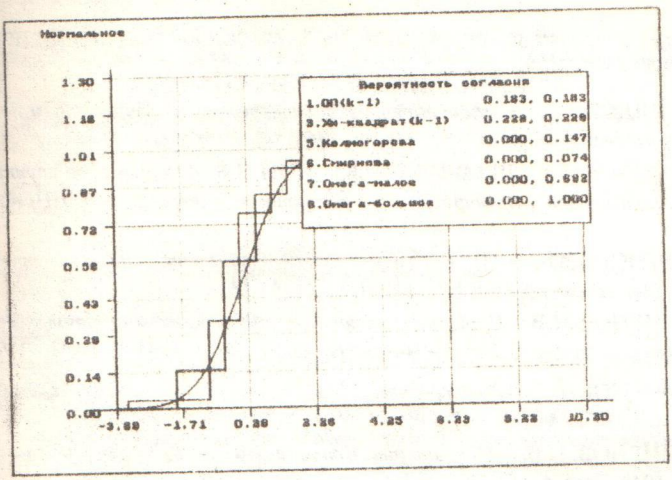

Рис. 3. Проверка согласия нормального распределения с параметрами $\mu = 0.10$ и $\sigma = 1.10$ по группированным данным

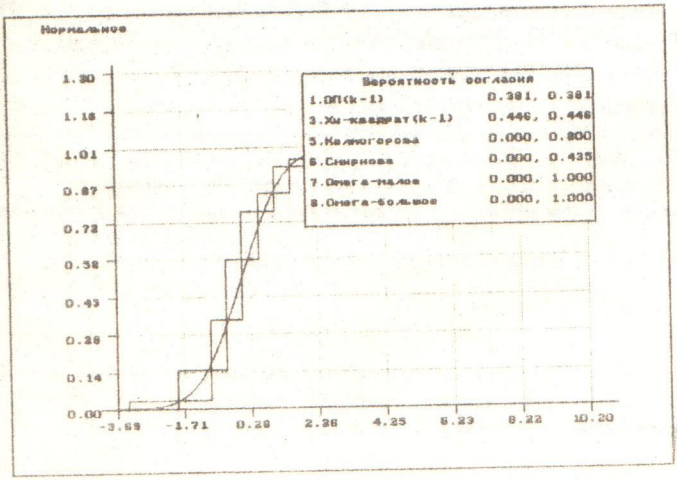

Рис. 4. Проверка согласия нормального распределения с параметрами $\mu = 0.0$ и $\sigma = 1.0$ по группированным данным

мативны, что требует ипользовать их в совокупности с параметрическими критериями согласия.

[1] КУЛЛДОРФ Г. *Введение в теорию оценивания по группированным и частично группированным выборкам.* - М.: Наука, 1966.

[2] ДЕНИСОВ В.И., ЛЕМЕШКО Б.Ю., ЦОЙ Е.Б. *Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов.* - Новосибирск, 1993.

[3] КОЛМОГОРОВ А.Н. *Об эмпирическом определении закона распределения* // Избранные труды по ТВ и МС. - 1967. - С. 134 - 141.

[4] СМИРНОВ Н.В. *Приближение законов распределений случайных величин по эмпирическим данным* // Успехи математических наук. - 1944. - С. 179 - 206.

[5] СМИРНОВ Н.В. *О распределении* ω^2 *критерия Мизеса* // Математический сборник. - 1937. - 44. - С. 973 - 994.

[6] СМИРНОВ Н.В. *О критерии Крамера-Мизеса* // Успехи математических наук. - 1949. Вып. 4. - С. 196 - 197.

[7] БОЛЬШЕВ Н.Л., СМИРНОВ Н.В. *Таблицы математической статистики.* - М.: Наука, 1983.

[8] GASTALDI TOMMASO. *A Kolmogorov-Smirnov test procedure involving apossibility censored or truncated sample* // Communications in statistics. Theory and methods. - 1993. - 22., № 1. - P. 31-39.