# Сравнение искаженных гистограмм вероятностными методами*

***А. Е. Лепский***
alex.lepskiy@gmail.com
Национальный исследовательский университет «Высшая школа экономики», Россия, Москва 101000, ул. Мясницкая, 20

В работе исследована задача об устойчивости вероятностных способов сравнения гистограмм относительно их искажений. Под сравнением понимается отношение полного предпорядка на множестве всех гистограмм, согласованное с условием упорядоченности аргументов гистограмм по возрастанию их важности. Под искажением понимаются интервальные поточечные изменения. Найдены необходимые и достаточные условия на уровень искажений гистограмм, при которых сравнение двух гистограмм не изменяется. Исследование проведено для трех популярных вероятностных методов сравнения: с помощью математического ожидания, с помощью стохастического доминирования, с помощью стохастического предшествования. Доказанные утверждения проиллюстрированы исследованиями устойчивости сравнений гистограмм результатов ЕГЭ абитуриентов, поступивших в вузы.

**Ключевые слова**: сравнение гистограмм; искажение гистограмм; устойчивость сравнения

# Comparison of distorted histograms by probability methods*

***A. E. Lepskiy***
Higher School of Economics, 20 Myasnitskaya Str., Moscow 101000, Russia

This paper is devoted to study of stability of comparison of histograms with help of different probability methods.

**Background**: The comparison of histograms is necessary in many applied problems of data processing. In this paper, the comparison of type “more–less” is considered. But the histograms may be distorted. The nature of these distortions can be different. Then, it is a problem to find the conditions on distortions under which the comparison of the two histograms is not changed.

**Methods**: There are many approaches to comparison of histograms. In this paper, the three popular probabilistic methods of comparison of histograms are considered: comparison of mathematical expectations, comparison with help of principle of stochastic dominance, and comparison with the help of stochastic precedence. In this paper, the interval distortions of histograms is considered.

**Results**: The necessary and sufficient conditions of preservation for comparison of distorted histograms are found with respect to different probability indices of comparison. The description of set of admissible distortions preserving the comparison of two histograms is found. The characteristics of stability of histograms to distortion are introduced. These characteristics are calculated for histograms of USE (Unified State Exam) of applicants admitted in 2012 in

*Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ в 2014 г. и при частичной поддержке РФФИ, проект № 14-07-00189.

Russian universities. It is shown that the stability of comparison of histograms to distortion cannot correspond to the values of difference index of comparison (margin).
**Concluding Remarks**: The found conditions invariability of comparing histograms can be used to estimate the reliability of results of different rankings, data processing, etc. in terms of different types of uncertainty: stochastic uncertainty, the uncertainty associated with the distortion of the data in filling data gaps, etc.

**Keywords**: comparison of histograms; distortions of histograms; stability of comparison

## Введение

Проблема сравнения гистограмм возникает при решении многих прикладных задач анализа данных. Под сравнением в данном случае будем понимать определение на множестве гистограмм вида $U = (x_i, u_i)_{i \in I}$, $x_i < x_{i+1}$, $i \in I$, отношения типа «больше-меньше». Примерами таких задач являются: сравнение результатов различных опытов (см., например, [6]), сравнение показателей функционирования каких-либо однородных (организационных, технических и пр.) систем [2], принятие решений в условиях нечеткостной неопределенности [17], моделирование нечетких предпочтений [5], сравнение распределений доходов в рамках социально-экономического анализа [1], [7], [3], ранжирование учащихся по результатам-гистограммам их оценок [19], [11] и т.д.

Для решения задач сравнения гистограмм применяются различные подходы. Одним из наиболее популярных является вероятностный подход, в котором сравниваются некоторые числовые характеристики случайных величин, связанных со сравниваемой парой гистограмм. Другой подход основан на применении методов ранжирования распределений доходов в теории коллективного выбора [7]. В этом случае сравниваются гистограммы доходов вида $U = (i, u_i)_{i=1}^{n_U} = (u_i)_{i=1}^{n_U}$, где $u_1 \leqslant u_2 \leqslant ... \leqslant u_{n_U}$ с помощью функций благосостояния $W(U)$, удовлетворяющих условиям симметричности, монотонности, вогнутости и др. Если размерности векторов-гистограмм одинаковы, то указанный подход равносилен ранжированию упорядоченных по возрастанию векторов. В этом случае можно использовать, например, методы теории важности критериев [9], методы некомпенсаторного выбора [23] и др.

Третий подход к ранжированию гистограмм связан с применением инструментария сравнения нечетких чисел. В этом случае каждой гистограмме $U = (x_i, u_i)_{i \in I}$ можно поставить в соответствие нечеткое множество или, в частном случае, нечеткое число [12] с функцией принадлежности $U = (u_i)_{i \in I}$, определенной на универсальном множестве $X = (x_i)_{i \in I}$. После чего можно использовать методы сравнения нечетких чисел [20], [14], [8]. Обзор и анализ основных подходов сравнения гистограмм дан в [11].

При сравнении гистограмм необходимо учитывать, что сами гистограммы могут быть заданы с той или иной степенью неточности. Характер этих неточностей может быть различным. Например, при сравнении гистограмм как результатов опытов неточность может носить вероятностный характер. А при сравнении распределений доходов в теории коллективного выбора неточность может быть результатом намеренного искажения данных. В ходе выборных кампаний обсуждается анализ (в том числе и сравнение) гистограмм распределений избирательных участков или голосов избирателей по явке и по процентам голосования за различных кандидатов (партии). Такие гистограммы также могут быть искажены, и в том числе в результате фальсификаций или, в более общем случае, в результате манипулирования данными. Еще один часто встречающийся тип искажений – заполнение пробелов в данных.

Во всех этих и других ситуациях возникает задача определения, может ли данное искажение гистограмм изменить их сравнение определенным методом на противоположное. Или, для каких искажений результат сравнения не изменится?

Цель данной работы получить ответы на эти вопросы. В работе будет проанализирована устойчивость к искажениям некоторых наиболее популярных вероятностных методов сравнения гистограмм. Некоторые результаты этой работы были анонсированы в [21].

## Основные определения и обозначения

Под гистограммой в этой работе будем понимать пару двух упорядоченных наборов чисел $U = (x_i, u_i)_{i \in I}$, где $(x_i)_{i \in I}$ – упорядоченный по возрастанию вектор различных аргументов гистограммы (т.е. $x_i < x_{i+1}$, $i \in I$), $(u_i)_{i \in I}$ – вектор неотрицательных значений гистограммы, $I$ – некоторое индексное множество.

На множестве гистограмм $\mathcal{U} = \{U\}$ необходимо построить отношение полного предпорядка (рефлексивного, полного и транзитивного отношения) $R$. Если гистограммы $U$ и $V$ находятся в отношении $R$ (т.е. $(U, V) \in R$)), то будем обозначать это так: $U \succcurlyeq V$ и говорить, что «$U$ больше $V$». Если же $U \succcurlyeq V$ и $V \succcurlyeq U$, то будем называть эти гистограммы «равными» и обозначать это так: $U \sim V$.

Будем также предполагать, что отношение $R$ должно быть согласовано с условием упорядоченности аргументов гистограмм по возрастанию их важности. Под согласованностью будем понимать следующее условие: если $U' = (x_i, u'_i)$, $U'' = (x_i, u''_i)$ две такие гистограммы, что $u'_i = u''_i$ для всех $i \neq k, l$ и $u'_l - u''_l = u''_k - u'_k \geqslant 0$, то $U'' \succcurlyeq U'$ при $k > l$ и $U' \succcurlyeq U''$ при $k < l$.

Без ограничения общности можно считать, что сравниваемые гистограммы «выровнены по числу столбцов», т.е. если $U = (x_i^U, u_i)_{i \in I_U}$ и $V = (x_i^V, u_i)_{i \in I_V}$ две гистограммы, то $I_U = I_V$ и $\{x_i^U\}_{i \in I} = \{x_i^V\}_{i \in I}$. Действительно, для выравнивания гистограмм надо объединить множества аргументов гистограмм $X^U = \{x_i^U\}_{i \in I_U}$ и $X^V = \{x_i^V\}_{i \in I_V}$: $X = X^{(U)} \cup X^{(V)} = \{x_i\}$ и применить какую-либо процедуру заполнения пробелов в данных. Например, можно использовать следующее правило: $\tilde{u}_i = u_k$, если $x_k^U \leqslant x_i < x_{k+1}^U$. В результате вместо гистограмм $U$ и $V$ мы получим две новые гистограммы $U = (x_i, \tilde{u}_i)_{i \in I} = (\tilde{u}_i)_{i \in I}$ и $V = (x_i, \tilde{v}_i)_{i \in I} = (\tilde{v}_i)_{i \in I}$.

Таким образом, ниже будем считать, что все гистограммы имеют вид $U = (x_i, u_i)_{i \in I} = (u_i)_{i \in I}$.

## Вероятностные индексы парного сравнения

Пусть $U = (x_i, u_i)_{i \in I}$ и $V = (x_j, v_j)_{j \in I}$ – две гистограммы, $u_i \geqslant 0$, $v_j \geqslant 0$ для всех $i, j \in I$, $I$ – некоторое индексное множество.

Рассмотрим некоторый числовой индекс $r(U, V)$ парного сравнения гистограмм $U$ и $V$ на $\mathcal{U}^2$, от которого потребуем только, чтобы он был согласован с условием упорядоченности аргументов гистограмм по возрастанию их важности: если $U = (x_i, u_i)$, $V = (x_i, v_i)$ две такие гистограммы, что $u_i = v_i$ для всех $i \neq k, l$ и $u_l - v_l = v_k - u_k \geqslant 0$, то $r(U, V) \geqslant 0$ при $k > l$ и $r(U, V) \leqslant 0$ при $k < l$. Отсюда, в частности, следует, что $r(U, U) = 0$.

Если $r(U, V)$ можно задать с помощью некоторой функции $F(U)$ как $r(U, V) = F(U) - F(V)$, то введенное с помощью такого индекса $r(U, V)$ отношение $U \succcurlyeq V \Leftrightarrow r(U, V) \geqslant r(V, U) \Leftrightarrow \Delta_r(U, V) = r(U, V) - r(V, U) \geqslant 0$ будет отношением полного предпорядка. В общем случае знак разностного индекса сравнения $\Delta_r(U, V) = r(U, V) - r(V, U)$ может и не задавать транзитивное отношение.

Приведем примеры индексов парного сравнения гистограмм – вероятностных распределений. В этом случае считаем, что $U = (x_i, u_i)_{i\in I}$ и $V = (x_j, v_j)_{j\in I}$ – случайные величины, принимающие значения из множества $\{x_i\}_{i\in I}$ с вероятностями $\{u_i\}_{i\in I}$ и $(v_j)_{j\in I}$ соответственно, $\sum_{i\in I} u_i = 1$, $\sum_{i\in I} v_i = 1$.

**1. Сравнение по среднему значению**: $U \succcurlyeq V$, если $E[U] \geqslant E[V]$ . Обобщением этого способа является: $U \succcurlyeq V$, если $E[f(U)] \geqslant E[f(V)]$, где $f$ – некоторая функция (функция полезности). Для того чтобы индекс сравнения принимал значения из промежутка [0,1], нормируем его: $E_0[U] = \frac{1}{\Delta x}(E[U] - x_{\min})$, где $\Delta x = x_{\max} - x_{\min}$. Заметим, что $E_0[U] = E[U_0]$, где $U_0 = (x_i^0, u_i)_{i\in I}$, $x_i^0 = \frac{1}{\Delta x}(x_i - x_{\min}) \in [0,1]$ для всех $i \in I$. Соответствующий разностный индекс сравнения будем обозначать через $\Delta_E(U,V) = E_0[U] - E_0[V] = \frac{1}{\Delta x}(E[U] - E[V])$.

**2. Стохастическое доминирование**: $U \succcurlyeq V$, если $F_U(x) \leqslant F_V(x)$ для всех $x \in \mathbb{R}$, где $F_U(x) = \sum_{i:x_i<x} u_i$ – функция распределения случайной величины $U$. Противоположное неравенство к сравнению объясняется необходимостью соответствия условию согласования сравнения с упорядоченностью аргументов гистограмм по возрастанию их важности. Это принцип стохастического доминирования 1-го порядка, который используется, например, в теории риска [13], [18]. Подход, основанный на применении принципа стохастического доминирования к широко используемым в микроэкономике кривым спроса, был реализован в работе [19] для сравнения образовательных программ по результатам ЕГЭ зачисленных студентов.

Соответствующий разностный индекс сравнения будем обозначать через $\Delta_F(U,V) = \inf_x (F_U(x) - F_V(x))$. Заметим, что если $U = (x_i, u_i)_{i=1}^n$ и $V = (x_j, v_j)_{j=1}^n$ – две случайные величины, то $F_U(x) - F_V(x) = 0$ для всех $x \leqslant x_1$ или $x > x_n$. Поскольку нас интересует свойство сохранения знака (точнее, неотрицательности) разности $F_U(x) - F_V(x)$, то вместо разностного индекса сравнения $\Delta_F(U,V) = \inf_x (F_U(x) - F_V(x))$ будем рассматривать индекс $\inf_{x\in(x_1,x_n]} (F_U(x) - F_V(x))$, который также будем обозначать через $\Delta_F(U,V)$. Заметим, что индекс $\Delta_F(U,V)$ определен не на всем множестве $\mathcal{U}^2$.

**3. Стохастическое предшествование (stochastic precedence)**: $U \succcurlyeq V$, если $P\{U \geqslant V\} \geqslant P\{U \leqslant V\}$ (распределение $V$ предшествует распределению $U$). Некоторые свойства стохастического предшествования можно найти в [15], [10]. Такой способ сравнения использовался, например, в [22]. Если считать, что случайные величины $U = (x_i, u_i)_{i\in I}$ и $V = (x_j, v_j)_{j\in I}$ независимы, то

$$P\{U \geqslant V\} = \sum_{(i,j):\, x_i \geqslant x_j} u_i v_j.$$

Соответствующий разностный индекс сравнения будем обозначать через $\Delta_P(U,V) = P\{U \geqslant V\} - P\{U \leqslant V\}$.

Заметим, что неравенство $\Delta_P(U,V) \geqslant 0$ не задает транзитивное отношение, как видно из следующего примера. Однако, как показывает численное моделирование, вероятность появления нетранзитивной тройки гистограмм при равномерной их генерации очень мала.

**Пример 1.** Пусть $U = (u_i)_{i=1}^4$, $u_1 = 0.23$, $u_2 = 0.72$, $u_3 = 0.04$, $u_4 = 0.01$; $V = (v_i)_{i=1}^4$, $v_1 = 0.62$, $v_2 = 0.01$, $v_3 = 0.12$, $v_4 = 0.25$; $W = (w_i)_{i=1}^4$, $w_1 = 0.45$, $w_2 = 0.2$, $w_3 = 0.3$, $w_4 = 0.05$. Тогда $\Delta_P(U,V) = 0.064 \geqslant 0$, $\Delta_P(V,W) = 0.051 \geqslant 0$, но $\Delta_P(U,W) = -0.028 < 0$.

В литературе рассматриваются и другие вероятностные индексы сравнения. Так, в [4] рассматривался следующий индекс включения $\psi_\beta$: $\psi_\beta\{U \subseteq V\} = P_U\left\{(V)_\beta | (U)_\beta\right\}$, $(U)_\beta = \{x : F_U(x) < \beta\}$. Тогда $U \succcurlyeq V$, если $\psi_\beta\{U \subseteq V\} \geqslant \psi_\beta\{V \subseteq U\}$. Если $F_U(x) \geqslant F_V(x)$ для

всех $x \in \mathbb{R}$, то $\psi_\beta\{U \subseteq V\} = 1$ для любого $\beta \in [0,1]$ и $U \succcurlyeq V$. Таким образом, с помощью индекса включения обобщалось понятие стохастического доминирования.

## Искажения гистограмм

Предположим, что вместо сравниваемых гистограмм $U = (x_i, u_i)_{i\in I}$ и $V = (x_j, v_j)_{j\in I}$ мы имеем две «близкие» к ним гистограммы $\tilde{U} = (x_i, \tilde{u}_i)_{i\in I}$ и $\tilde{V} = (x_j, \tilde{v}_j)_{j\in I}$. Причины искажения гистограмм могут быть различными. Это может быть умышленное манипулирование данными, из которых формируются гистограммы. Это может быть результат действия случайных факторов. Это может быть результат применения процедур обработки гистограмм (сглаживания, приведения к унимодальному виду и пр.). Поэтому и само описание неопределенности гистограммы может быть разным. Например, эта неопределенность может иметь интервальный характер, может быть стохастическим процессом, может иметь нечеткостный характер и т.д.

Ниже рассмотрим интервальное искажение гистограмм. Выбор такого типа искажений обусловлен, как относительной простотой анализа гистограмм, подвергнутых интервальному искажению, так и возможностью «огрубления» других типов искажений (вероятностных, нечеткостных и др.) до интервальных. Пусть $U = (x_i, u_i)_{i\in I}$ – «идеальная» гистограмма, а $\tilde{U} = (x_i, \tilde{u}_i)_{i\in I}$ – ее интервальное искажение: $\tilde{u}_i = u_i + h_i$, $i \in I$, где $\sum_{i\in I} h_i = 0$ и $|h_i| \leqslant \alpha u_i$, $i \in I$ где $\alpha \in [0,1]$. Величина $\alpha$ характеризует порог искажения в том смысле, что изменение $i$-го столбика гистограммы не может больше $100\cdot\alpha\%$. Будем называть такое искажение $\alpha$ -искажением. Обозначим через $N_\alpha(U)$ класс всех $\alpha$ -искажений гистограммы $U = (x_i, u_i)_{i\in I}$, т.е.

$$N_\alpha(U) = \left\{H = (h_i)_{i\in I} :\ \sum\nolimits_{i\in I} h_i = 0,\ |h_i| \leqslant \alpha u_i,\ i \in I\right\}. \tag{1}$$

Предположим, что $\Delta_r(U,V) > 0$. Основной вопрос, который исследуется в этой статье: в каком случае $\Delta_r(\tilde{U},\tilde{V}) \geqslant 0$ для любых $H \in N_\alpha(U)$ и $G \in N_\beta(V)$? Другими словами, в каком случае сравнение искаженных гистограмм не изменится после $\alpha$ -искажения гистограммы $U = (x_i, u_i)_{i\in I}$ и $\beta$ -искажения гистограммы $V = (x_j, v_j)_{j\in I}$?

Получим необходимые и достаточные условия сохранения сравнения в случае описанного интервального искажения для разных типов сравнений.

## Условия сохранения сравнения гистограмм

### Условия сохранения сравнения гистограмм относительно индекса $\Delta_E$

Для гистограммы $U = (x_i, u_i)_{i\in I}$ рассмотрим величину

$$\mathcal{E}_U = \sup\left\{\sum\nolimits_{i\in I} x_i^0 h_i :\ (h_i)_{i\in I} \in N_1(U)\right\}, \tag{2}$$

где $N_1(U)$ – множество вида (1) с $\alpha = 1$, $x_i^0 = \frac{1}{\Delta x}(x_i - x_{\min})$ для всех $i \in I$. Отметим следующие свойства величины $\mathcal{E}_U$.

**Лемма 1.** *Справедлива оценка* $0 \leqslant \mathcal{E}_U \leqslant \min\{E_0[U], 0.5\}$, *причем неравенства являются точными.*

**Доказательство.** Имеем $\mathcal{E}_U = \sup\left\{\sum_{i\in I} x_i^0 h_i :\ (h_i)_{i\in I} \in N_1(U)\right\} \geqslant \sum_{i\in I} x_i^0 \cdot 0 = 0$, так как $(0)_{i\in I} \in N_1(U)$. Неравенство $\mathcal{E}_U \geqslant 0$ является точным, поскольку $\mathcal{E}_U = 0$, например, для гистограммы с $u_i = 0$ для всех $i \neq i_0$ и $u_{i_0} = 1$. С другой стороны, $\sum_{i\in I} x_i^0 h_i \leqslant \sum_{i\in I} x_i^0 u_i = E_0[U]$. Поэтому $\mathcal{E}_U \leqslant E_0[U]$ и это неравенство является точным, поскольку равенство

$\mathcal{E}_U = E_0[U]$ достигается, например, для гистограммы $U = (x_i, u_i)_{i=1}^n$ с $u_1 = u_n = 0.5$, $u_i = 0$ для всех $i \neq 1, n$. Покажем, что $\mathcal{E}_U \leqslant 0.5$. Имеем

$$\sum_i x_i^0 h_i = \sum_{i:\, h_i>0} x_i^0 h_i + \sum_{i:\, h_i\leqslant 0} x_i^0 h_i \leqslant \max_{i:\, h_i>0} x_i^0 \cdot \sum_{i:\, h_i>0} h_i + \min_{i:\, h_i\leqslant 0} x_i^0 \cdot \sum_{i:\, h_i\leqslant 0} h_i.$$

Так как $\sum_{i:\, h_i>0} h_i = -\sum_{i:\, h_i\leqslant 0} h_i$, то

$$\sum_i x_i^0 h_i \leqslant \left(\max_{i:\, h_i>0} x_i^0 - \min_{i:\, h_i\leqslant 0} x_i^0\right) \cdot \sum_{i:\, h_i>0} h_i \leqslant \sum_{i:\, h_i>0} h_i.$$

Оценим величину $A = \sum_{i:\, h_i>0} h_i$. Пусть $B = \sum_{i:\, h_i>0} u_i$. Тогда $B \in [0,1]$ и, с одной стороны, $A \leqslant B$. С другой стороны, $A = \sum_{i:\, h_i>0} h_i = -\sum_{i:\, h_i\leqslant 0} h_i \leqslant \sum_{i:\, h_i\leqslant 0} u_i = 1 - B$. Поэтому, $A \leqslant \min\{B, 1-B\} \leqslant 0.5$. Таким образом, $\sum_{i\in I} x_i^0 h_i \leqslant 0.5$ и, следовательно, $\mathcal{E}_U \leqslant 0.5$. Это неравенство также является точным и достигается, например, для гистограммы $U = (x_i, u_i)_{i=1}^n$ с $u_1 = u_n = 0.5$, $u_i = 0$ для всех $i \neq 1, n$. ■

**Лемма 2.** *Для гистограммы $U = (x_i, u_i)_{i=1}^n$ верно равенство*

$$\mathcal{E}_U = \sum_{s=s_0}^{n} x_s^0 u_s a_s - \sum_{s=1}^{s_0-1} x_s^0 u_s b_s,$$

*где $1 \geqslant a_n \geqslant ... \geqslant a_{s_0} \geqslant 0$, $1 \geqslant b_1 \geqslant ... \geqslant b_{s_0-1} \geqslant 0$, $\sum_{s=s_0}^n u_s a_s = \sum_{s=1}^{s_0-1} u_s b_s$, а индекс $s_0$ удовлетворяет неравенству $s_0 - 1 < m_U \leqslant s_0$, $m_U$ – медиана распределения $U$.*

Доказательство очевидно.

**Утверждение 1.** *Пусть $\tilde{U} = (x_i, u_i + h_i)_{i\in I}$, $\tilde{V} = (x_j, v_j + g_j)_{i\in I}$ – $\alpha$ - и $\beta$ -искажения гистограмм $U = (x_i, u_i)_{i=1}^n$ и $V = (x_j, v_j)_{j=1}^n$ соответственно. Тогда $\Delta_E(\tilde{U}, \tilde{V}) \geqslant 0$ для любых $(h_i)_{i\in I} \in N_\alpha(U)$ и $(g_i)_{i\in I} \in N_\beta(V)$, $\alpha, \beta \in [0,1]$ в том и только том случае, если*

$$\Delta_E(U, V) \geqslant \alpha\mathcal{E}_U + \beta\mathcal{E}_V.$$

**Доказательство.** Имеем

$$\Delta_E(\tilde{U}, \tilde{V}) = \sum_{i\in I} x_i^0 \,(u_i + h_i - v_i - g_i) = \Delta_E(U, V) - \sum_{i\in I} x_i^0 \,(g_i - h_i).$$

Поэтому условие $\Delta_E(\tilde{U}, \tilde{V}) \geqslant 0$ для любых $(h_i)_{i\in I} \in N_\alpha(U)$ и $(g_i)_{i\in I} \in N_\beta(V)$ равносильно выполнению неравенства

$$\Delta_E(U, V) \geqslant \sup_{(h_i)_{i\in I}\in N_\alpha(U)} \sum_{i\in I} x_i^0 \,(-h_i) + \sup_{(g_i)_{i\in I}\in N_\alpha(V)} \sum_{i\in I} x_i^0 g_i \Leftrightarrow$$

$$\Delta_E(U, V) \geqslant \alpha \sup_{(h_i)_{i\in I}\in N_1(U)} \sum_{i\in I} x_i^0 h_i + \beta \sup_{(g_i)_{i\in I}\in N_1(V)} \sum_{i\in I} x_i^0 g_i =$$

$$= \alpha\mathcal{E}_U + \beta\mathcal{E}_V$$

и утверждение доказано. ■

Пусть $\bar{\mathcal{E}}_U = \min\{E_0[U], 0.5\}$. Тогда из Леммы 1 вытекает справедливость следствия.

**Следствие 1.** *Если*

$$\Delta_E(U,V) \geqslant \alpha\bar{\mathcal{E}}_U + \beta\bar{\mathcal{E}}_V,$$

*то* $\Delta_E(\tilde{U},\tilde{V}) \geqslant 0$ *для любых* $(h_i)_{i\in I} \in N_\alpha(U)$ *и* $(g_i)_{i\in I} \in N_\beta(V)$.

**Условия сохранения сравнения гистограмм относительно индекса $\Delta_F$**

Аналогичные условия сохранения знака можно получить и для разностного индекса сравнения $\Delta_F(U,V)$. Введем в рассмотрение функцию

$$\mathcal{F}_U(x) = \sup\left\{\sum\nolimits_{i:x_i<x} h_i :\ (h_i)_{i\in I} \in N_1(U)\right\}, \tag{3}$$

где $N_1(U)$ – множество вида (1) с $\alpha = 1$.

**Лемма 3.** $\mathcal{F}_U(x) = \min\{F_U(x), 1 - F_U(x)\}$ *для всех* $x \in \mathbb{R}$.

**Доказательство.** Пусть $m_U$ – медиана распределения $U = (x_i,u_i)_{i=1}^n$ и $m_U \in (x_{i_0}, x_{i_0+1}]$. Если $x < m_U$, то $\mathcal{F}_U(x) = \sup\left\{\sum_{i:x_i<x} h_i :\ (h_i)_{i\in I} \in N_1(U)\right\} = \sum_{i:x_i<x} u_i = F_U(x) < 0.5$, причем супремум в последнем равенстве достигается при $h_i = u_i$ для всех таких $i$, что $x_i < x$ и $h_i = -b_iu_i$ для всех таких $i$, что $x_i \geqslant x$, где $b_i \in [0,1]$ и $\sum_{i:x_i<x} u_i = \sum_{i:x_i\geqslant x} b_iu_i$.

Если же $x \geqslant m_U$, то, в силу равенства $\sum_{i:x_i<x} h_i = -\sum_{i:x_i\geqslant x} h_i$, имеем $\mathcal{F}_U(x) = \sup\left\{-\sum_{i:x_i\geqslant x} h_i :\ (h_i)_{i\in I} \in N_1(U)\right\} = \sum_{i:x_i\geqslant x} u_i = 1 - F_U(x) \leqslant 0.5$, причем супремум в последнем равенстве достигается при $h_i = -u_i$ для всех таких $i$, что $x_i \geqslant x$ и $h_i = b_iu_i$ для всех таких $i$, что $x_i < x$, где $b_i \in [0,1]$ и $\sum_{i:x_i<x} b_iu_i = \sum_{i:x_i\geqslant x} u_i$. ■

**Утверждение 2.** *Пусть* $\tilde{U} = (x_i, u_i + h_i)_{i\in I}$, $\tilde{V} = (x_j, v_j + g_j)_{i\in I}$ – $\alpha$ - *и* $\beta$ - *искажения гистограмм* $U = (x_i,u_i)_{i\in I}$ *и* $V = (x_j,v_j)_{i\in I}$ *соответственно. Тогда* $\Delta_F(\tilde{U},\tilde{V}) \geqslant 0$ *для любых* $(h_i)_{i\in I} \in N_\alpha(U)$ *и* $(g_i)_{i\in I} \in N_\beta(V)$, $\alpha,\beta \in [0,1]$ *в том и только том случае, если*

$$F_U(x) - F_V(x) \geqslant \alpha\mathcal{F}_U(x) + \beta\mathcal{F}_V(x) \tag{4}$$

*для всех* $x \in \mathbb{R}$.

**Доказательство.** Действительно, условие $\Delta_F(\tilde{U},\tilde{V}) \geqslant 0$ для любых $(h_i)_{i\in I} \in N_\alpha(U)$ и $(g_i)_{i\in I} \in N_\beta(V)$ равносильно выполнению для всех $x \in \mathbb{R}$ неравенства

$$F_{\tilde{U}}(x) - F_{\tilde{V}}(x) \geqslant 0 \Leftrightarrow \sum_{i:x_i<x} (u_i + h_i - v_i - g_i) \geqslant 0 \Leftrightarrow$$

$$F_U(x) - F_V(x) \geqslant \sum_{i:x_i<x} (-h_i) + \sum_{i:x_i<x} g_i.$$

Но выполнение последних неравенств для любых $(h_i)_{i\in I} \in N_\alpha(U)$ и $(g_i)_{i\in I} \in N_\beta(V)$ и $x \in \mathbb{R}$ равносильно условию

$$F_U(x) - F_V(x) \geqslant \sup_{(h_i)_i\in N_\alpha(U)} \sum_{i:x_i<x} (-h_i) + \sup_{(g_i)_i\in N_\beta(V)} \sum_{i:x_i<x} g_i$$

для всех $x \in \mathbb{R}$ или

$$F_U(x) - F_V(x) \geqslant \alpha \sup_{(h_i)_i\in N_1(U)} \sum_{i:x_i<x} h_i + \beta \sup_{(g_i)_i\in N_1(V)} \sum_{i:x_i<x} g_i,$$

что равносильно (4). ■

**Следствие 2.** *Неравенство* $\Delta_F(\tilde{U},\tilde{V}) \geqslant 0$ *верно для любых* $(h_i)_{i\in I} \in N_\alpha(U)$ *и* $(g_i)_{i\in I} \in N_\beta(V)$ *в том и только том случае, если*

$$0 \leqslant \sup_x \frac{\alpha\mathcal{F}_U(x) + \beta\mathcal{F}_V(x)}{F_U(x) - F_V(x)} \leqslant 1$$

*(считаем, что дробь равна нулю, если ее числитель и знаменатель равны нулю).*

**Следствие 3.** *Если*

$$\Delta_F(U,V) \geqslant \sup_x \left\{\alpha\mathcal{F}_U(x) + \beta\mathcal{F}_V(x)\right\},$$

*то* $\Delta_F(\tilde{U},\tilde{V}) \geqslant 0$ *для любых* $(h_i)_{i\in I} \in N_\alpha(U)$ *и* $(g_i)_{i\in I} \in N_\beta(V)$.

**Условия сохранения сравнения гистограмм относительно индекса $\Delta_P$**

Для разностного индекса сравнения $\Delta_P(U,V)$ справедливы следующие условия сохранения знака.

**Утверждение 3.** *Пусть* $\tilde{U} = (x_i, u_i + h_i)_{i\in I}$, $\tilde{V} = (x_j, v_j + g_j)_{j\in I}$ – $\alpha$ - *и* $\beta$ -*искажения гистограмм* $U = (x_i, u_i)_{i\in I}$ *и* $V = (x_j, v_j)_{j\in I}$ *соответственно. Тогда* $\Delta_P(\tilde{U},\tilde{V}) \geqslant 0$ *для любых* $(h_i)_{i\in I} \in N_\alpha(U)$ *и* $(g_i)_{i\in I} \in N_\beta(V)$, $\alpha,\beta \in [0,1]$ *в том и только том случае, если*

$$\Delta_P(U,V) \geqslant \Delta\eta_{\alpha,\beta}(U,V),$$

*где*

$$\Delta\eta_{\alpha,\beta}(U,V) = \sup_{\substack{(h_i)_i\in N_\alpha(U),\\ (g_i)_i\in N_\beta(V)}} \sum_{(i,j):\, x_i<x_j} (u_i g_j + h_i v_j + h_i g_j - u_j g_i - h_j v_i - h_j g_i). \tag{5}$$

**Доказательство.** Так как

$$P\left\{\tilde{U} \geqslant \tilde{V}\right\} = \sum_{(i,j):\, x_i\geqslant x_j} \tilde{u}_i\tilde{v}_j = \sum_{(i,j):\, x_i\geqslant x_j} (u_i + h_i)(v_j + g_j) =$$

$$\sum_{(i,j):\, x_i\geqslant x_j} u_i v_j + u_i g_j + h_i v_j + h_i g_j = P\left\{U \geqslant V\right\} + \eta(\tilde{U},\tilde{V}),$$

где $\eta(\tilde{U},\tilde{V}) = \sum\limits_{(i,j):\, x_i\geqslant x_j} u_i g_j + h_i v_j + h_i g_j$, то

$$\Delta_P(\tilde{U},\tilde{V}) = P\left\{\tilde{U} \geqslant \tilde{V}\right\} - P\left\{\tilde{V} \geqslant \tilde{U}\right\} = \Delta_P(U,V) - \Delta\eta(\tilde{U},\tilde{V}),$$

где $\Delta\eta(\tilde{U},\tilde{V}) = \eta(\tilde{V},\tilde{U}) - \eta(\tilde{U},\tilde{V}) = \sum\limits_{(i,j):\, x_i<x_j} u_i g_j + h_i v_j + h_i g_j - u_j g_i - h_j v_i - h_j g_i$. Из последнего равенства и вытекает справедливость утверждения. ■

**Следствие 4.** *Если*

$$\Delta_P(U,V) \geqslant \frac{\alpha+\beta}{1+\alpha\beta}\left(1 + P\left\{V = U\right\}\right), \tag{6}$$

*то* $\Delta_P(\tilde{U},\tilde{V}) \geqslant 0$ *для любых* $H = (h_i)_{i\in I} \in N_\alpha(U)$ *и* $G = (g_i)_{i\in I} \in N_\beta(V)$.

**Доказательство.** Оценим величину $\eta(\tilde{U},\tilde{V})$ (см. доказательство Утверждения 3). Так как

$$u_i v_j(\alpha\beta-\alpha-\beta)\leqslant u_i g_j + h_i v_j + h_i g_j \leqslant u_i v_j(\alpha+\beta+\alpha\beta)$$

для $(h_i)_i \in N_\alpha(U)$ и $(g_i)_i \in N_\beta(V)$,то

$$(\alpha\beta-\alpha-\beta)P\{U\geqslant V\}\leqslant \eta(\tilde{U},\tilde{V})\leqslant(\alpha+\beta+\alpha\beta)P\{U\geqslant V\}.$$

Из последней оценки получим

$$\Delta_P(\tilde{U},\tilde{V}) = P\{\tilde{U}\geqslant\tilde{V}\}-P\{\tilde{U}\leqslant\tilde{V}\} = \Delta_P(U,V)+\eta(\tilde{U},\tilde{V})-\eta(\tilde{V},\tilde{U})\geqslant$$

$$\Delta_P(U,V)+(\alpha\beta-\alpha-\beta)P\{U\geqslant V\}-(\alpha+\beta+\alpha\beta)P\{V\geqslant U\}=$$

$$(1+\alpha\beta)\Delta_P(U,V)-(\alpha+\beta)\left(P\{U\geqslant V\}+P\{V\geqslant U\}\right).$$

Поэтому если

$$(1+\alpha\beta)\Delta_P(U,V)-(\alpha+\beta)\left(P\{U\geqslant V\}+P\{V\geqslant U\}\right)\geqslant 0$$

или, что то же самое, верно (6), то $\Delta_P(\tilde{U},\tilde{V})\geqslant 0$. ■

**Следствие 5.** *Если*

$$\Delta_P(U,V)\geqslant\frac{\alpha+\beta+\alpha\beta}{1+\alpha+\beta+\alpha\beta}, \tag{7}$$

*то* $\Delta_P(\tilde{U},\tilde{V})\geqslant 0$ *для любых* $H=(h_i)_{i\in I}\in N_\alpha(U)$ *и* $G=(g_i)_{i\in I}\in N_\beta(V)$.

**Доказательство.** Оценим величину $\Delta\eta(\tilde{U},\tilde{V})$ (см. доказательство Утверждения 3). Так как

$$\sum_{(i,j):\,x_i<x_j} u_i g_j - u_j g_i = \sum_i u_i g_i + 2\sum_{(i,j):\,x_i<x_j} u_i g_j,$$

$$\sum_{(i,j):\,x_i<x_j} h_i v_j - h_j v_i = -\sum_i v_i h_i - 2\sum_{(i,j):\,x_i<x_j} v_i h_j,$$

$$\sum_{(i,j):\,x_i<x_j} h_i g_j - h_j g_i = \sum_i h_i g_i + 2\sum_{(i,j):\,x_i<x_j} h_i g_j,$$

то

$$\Delta\eta(\tilde{U},\tilde{V}) = \sum_i (u_i+h_i)\,g_i + 2\sum_{(i,j):\,x_i<x_j}(u_i+h_i)\,g_j - \sum_i v_i h_i - 2\sum_{(i,j):\,x_i<x_j} v_i h_j.$$

Тогда

$$\Delta\eta_{\alpha,\beta}(U,V) = \sup_{\substack{(h_i)_i\in N_\alpha(U),\\(g_i)_i\in N_\beta(V)}}\Delta\eta(\tilde{U},\tilde{V}) = \sup_{(h_i)_i\in N_\alpha(U)}\left\{-\sum_i v_i h_i - 2\sum_{(i,j):\,x_i<x_j} v_i h_j + \right.$$

$$\left. + \sup_{(g_i)_i\in N_\beta(V)}\left\{\sum_i (u_i+h_i)\,g_i + 2\sum_{(i,j):\,x_i<x_j}(u_i+h_i)\,g_j\right\}\right\}.$$

Для внутреннего супремума, учитывая, что $u_i + h_i \geqslant 0$ для всех $i \in I$, имеем

$$\sup_{(g_i)_i \in N_\beta(V)} \left\{ \sum_i (u_i + h_i)\, g_i + 2 \sum_{(i,j):\, x_i < x_j} (u_i + h_i)\, g_j \right\} \leqslant$$

$$\beta \left\{ \sum_i (u_i + h_i)\, v_i + 2 \sum_{(i,j):\, x_i < x_j} (u_i + h_i)\, v_j \right\} =$$

$$\beta \left\{ P\{U = V\} + 2P\{U < V\} + \sum_i h_i v_i + 2 \sum_{(i,j):\, x_i < x_j} h_i v_j \right\}.$$

Теперь получим

$$\Delta\eta_{\alpha,\beta}(U,V) \leqslant \beta \left(P\{U = V\} + 2P\{U < V\}\right) +$$

$$\sup_{(h_i)_i \in N_\alpha(U)} \left\{ (\beta - 1) \sum_i h_i v_i + 2\beta \sum_{(i,j):\, x_i < x_j} h_i v_j - 2 \sum_{(i,j):\, x_i < x_j} v_i h_j \right\}.$$

Поскольку

$$\sum_{(i,j):\, x_i < x_j} v_i h_j = -\sum_i h_i v_i - \sum_{(i,j):\, x_i < x_j} h_i v_j,$$

то

$$\Delta\eta_{\alpha,\beta}(U,V) \leqslant \beta \left(P\{U = V\} + 2P\{U < V\}\right) +$$

$$(\beta + 1) \sup_{(h_i)_i \in N_\alpha(U)} \left\{ \sum_i h_i v_i + 2 \sum_{(i,j):\, x_i < x_j} h_i v_j \right\} \leqslant (\alpha + \beta + \alpha\beta) \left(P\{U = V\} + 2P\{U < V\}\right).$$

Таким образом, $\Delta_P(\tilde{U}, \tilde{V}) \geqslant 0$ для любых $H = (h_i)_{i \in I} \in N_\alpha(U)$ и $G = (g_i)_{i \in I} \in N_\beta(V)$, если

$$\Delta_P(U,V) \geqslant (\alpha + \beta + \alpha\beta) \left(P\{U = V\} + 2P\{U < V\}\right). \tag{8}$$

Но

$$P\{U = V\} + 2P\{U < V\} = 1 - \Delta_P(U,V).$$

Поэтому неравенство (8) равносильно неравенству

$$\Delta_P(U,V) \geqslant (\alpha + \beta + \alpha\beta) \left(1 - \Delta_P(U,V)\right).$$

Откуда и следует справедливость условия (7). ■

**Замечание 1.** Нетрудно показать, что $(\alpha+\beta+\alpha\beta)/(1+\alpha+\beta+\alpha\beta) \cdot (1+\alpha\beta)/(\alpha+\beta) < 1$ для всех $\alpha, \beta \in (0, 1]$. Поэтому правая часть в (6) не меньше, чем правая часть в (7). Следовательно, условие (7) задает более слабые ограничения на искажения гистограмм, при которых сохраняется их сравнение относительно разностного индекса $\Delta_P(U,V)$.

## Сравнение множеств допустимых искажений

Рассмотрим множество всех тех $\alpha$ - и $\beta$ -зашумлений гистограмм $U$ и $V$ соответственно, при которых сохраняется сравнение гистограмм относительно определенного индекса $\Delta_r(U,V)$ при условии, что он равен $c > 0$:

$$\Omega_r^c(U,V) = \left\{(\alpha,\beta):\ \Delta_r(U,V) = c,\ \Delta_r(\tilde{U},\tilde{V}) \geqslant 0\ \forall H \in N_\alpha(U),\ G \in N_\beta(V)\right\}.$$

Такое множество назовем множеством допустимых искажений гистограмм $U$ и $V$ для данного сравнения $\Delta_r(U,V) = c$. Нетрудно видеть, что множество $\Omega_r^c(U,V)$ является звездным [16] с центром в начале координат, т.е. если $(\alpha_0,\beta_0) \in \Omega_r^c(U,V)$, то и $(t\alpha_0, t\beta_0) \in \Omega_r^c(U,V)$ для всех $t \in [0,1]$. Известно [16], что звездному множеству с центром в начале координат можно взаимно однозначно сопоставить такую лучевую функции $\Phi_r^c(\alpha,\beta)$ (т.е. непрерывную, неотрицательную и однородную: $\Phi_r^c(t\alpha,t\beta) = t\Phi_r^c(\alpha,\beta)$ для всех $t \geqslant 0$), что

$$\Omega_r^c(U,V) = \{(\alpha,\beta):\ \alpha \geqslant 0, \beta \geqslant 0, \Phi_r^c(\alpha,\beta) \leqslant 1\}.$$

Функции $\Phi_E^c(\alpha,\beta)$, $\Phi_F^c(\alpha,\beta)$ и $\Phi_P^c(\alpha,\beta)$ множеств допустимых искажений для индексов $\Delta_E(U,V)$, $\Delta_F(U,V)$ и $\Delta_P(U,V)$ соответственно, как следует из Утверждений 1-3, будут равны

$$\Phi_E^c(\alpha,\beta) = \tfrac{1}{c}\left(\alpha\mathcal{E}_U + \beta\mathcal{E}_V\right), \Phi_F^c(\alpha,\beta) = \sup_x\left\{\frac{\alpha\mathcal{F}_U(x) + \beta\mathcal{F}_V(x)}{F_U(x) - F_V(x)}\right\}, \Phi_P^c(\alpha,\beta) = \tfrac{1}{c}\Delta\eta_{\alpha,\beta}(U,V).$$

Вообще говоря, функция $\Phi_F^c(\alpha,\beta)$ в случае дискретных распределений является кусочно-линейной. Однако можно выделить широкий класс пар распределений, для которых эта функция будет линейной. Опишем этот класс.

Пусть $U$ и $V$ две случайные величины с функциями распределений $F_U$ и $F_V$ соответственно, а $m_U$ и $m_V$ – медианы соответствующих распределений. Заметим, что если $F_U(x) \geqslant F_V(x)$ для всех $x \in \mathbb{R}$, то $m_U \leqslant m_V$. Обозначим через $\varphi_{\alpha,\beta}(x) = \frac{\alpha\mathcal{F}_U(x)+\beta\mathcal{F}_V(x)}{F_U(x)-F_V(x)}$.

**Лемма 4.** *Пусть $F_U(x) \geqslant F_V(x)$ для всех $x \in \mathbb{R}$ и функции распределения $F_U$ и $F_V$ удовлетворяют условиям:*

*а) $F_V(x) \leqslant 2F_U(x)F_V(m_U)$ для всех $x \leqslant m_U$;*

*б) $1 - F_U(x) \leqslant 2\left(1 - F_V(x)\right)\left(1 - F_U(m_V)\right)$ для всех $x \geqslant m_V$;*

*в) на промежутке $(m_U, m_V)$ выполняется либо неравенство а), либо б).*

*Тогда $\Phi_F^c(\alpha,\beta) = \max\{\varphi_{\alpha,\beta}(m_U), \varphi_{\alpha,\beta}(m_V)\}$. Причем*

$$\Phi_F^c(\alpha,\beta) = \begin{cases} \varphi_{\alpha,\beta}(m_U) & \text{при } F_U(m_V)F_V(m_U) \geqslant \frac{1}{4}, \\ \varphi_{\alpha,\beta}(m_V) & \text{при } \left(1 - F_U(m_V)\right)\left(1 - F_V(m_U)\right) \geqslant \frac{1}{4}. \end{cases} \tag{9}$$

*Если же*

$$\begin{cases} F_U(m_V)F_V(m_U) < \frac{1}{4}, \\ \left(1 - F_U(m_V)\right)\left(1 - F_V(m_U)\right) < \frac{1}{4}, \end{cases} \tag{10}$$

*то*

$$\Phi_F^c(\alpha,\beta) = \begin{cases} \varphi_{\alpha,\beta}(m_U) & \text{при } \alpha < \beta\frac{\frac{1}{4} - F_U(m_V)F_V(m_U)}{\frac{1}{4} - (1-F_U(m_V))(1-F_V(m_U))}, \\ \varphi_{\alpha,\beta}(m_V) & \text{в противном случае.} \end{cases} \tag{11}$$

**Доказательство.** Пусть $x \leqslant m_U$. Тогда $\mathcal{F}_U(x) = F_U(x)$. Кроме того, так как $F_U(x) \geqslant F_V(x)$, то $\mathcal{F}_V(x) = F_V(x)$. Поэтому

$$\varphi_{\alpha,\beta}(x) \leqslant \varphi_{\alpha,\beta}(m_U) \Leftrightarrow (\alpha+\beta)F_V(x) \leqslant 2(\alpha+\beta)F_U(x)F_V(m_U) \Leftrightarrow \text{а).}$$

Если $x \in (m_U, m_V)$, то $\mathcal{F}_U(x) = 1 - F_U(x)$, $\mathcal{F}_V(x) = F_V(x)$ и

$$\varphi_{\alpha,\beta}(x) \leqslant \varphi_{\alpha,\beta}(m_U) \Leftrightarrow$$

$$\alpha\left(1 - 2F_V(m_U) + F_V(x) - 2F_U(x)\left(1 - F_V(m_U)\right)\right) \leqslant \beta\left(F_U(x)F_V(m_U) - F_V(x)\right). \qquad (12)$$

Правая часть в последнем неравенстве неотрицательна, если верно неравенство а) для $x \in (m_U, m_V)$. Если $F_V(m_U) = 0$, то из а) следует, что $F_V(x) = 0$, $x \in (m_U, m_V)$ и тогда выражение в скобках в левой части неравенства (12) будет равно $1 - 2F_U(x) \leqslant 0$. Следовательно, неравенство (12) будет верным для $x \in (m_U, m_V)$. Если же $F_V(m_U) > 0$, то из а) следует, что $2F_U(x) \geqslant F_V(x)/F_V(m_U)$, $x \in (m_U, m_V)$ и выражение в скобках в левой части неравенства (12) можно оценить так:

$$1 - 2F_V(m_U) + F_V(x) - 2F_U(x)\left(1 - F_V(m_U)\right) \leqslant$$

$$1 - 2F_V(m_U) + F_V(x) - \frac{F_V(x)}{F_V(m_U)}\left(1 - F_V(m_U)\right) = \left(1 - 2F_V(m_U)\right)\left(1 - \frac{F_V(x)}{F_V(m_U)}\right) \leqslant 0.$$

Поэтому неравенство (12) будет верным и в этом случае.

Из симметричности условий а) и б) вытекает, что если выполняется условие б) на промежутке $x \geqslant m_V$, то $\varphi_{\alpha,\beta}(x) \leqslant \varphi_{\alpha,\beta}(m_V)$ для всех $x \geqslant m_V$.

Таким образом, $\Phi_F^c(\alpha, \beta) = \max\{\varphi_{\alpha,\beta}(m_U), \varphi_{\alpha,\beta}(m_V)\}$, если функции распределения $F_U$ и $F_V$ удовлетворяют условиям а)-в). Непосредственно вычисляя, получим, что

$$\varphi_{\alpha,\beta}(m_U) \leqslant \varphi_{\alpha,\beta}(m_V) \Leftrightarrow \left(\tfrac{1}{4} - \left(1 - F_U(m_V)\right)\left(1 - F_V(m_U)\right)\right)\alpha \leqslant \beta\left(\tfrac{1}{4} - F_U(m_V)F_V(m_U)\right).$$

Поэтому будут верными формулы (9) и (11) (последняя – при выполнении условий (10)).
■

Заметим, что если $F_U(m_V)F_V(m_U) \geqslant \frac{1}{4}$, то $\left(1 - F_U(m_V)\right)\left(1 - F_V(m_U)\right) \leqslant \frac{1}{4}$ и наоборот. Поэтому другие случаи знаков, кроме описанных в Лемме неравенствами (9) и (10) исключены.

Условие (10) выполняется, если значения $F_U(m_V)$ и $F_V(m_U)$ расположены приблизительно симметрично относительно $\frac{1}{2}$. В этом случае функция $\Phi_F^c(\alpha, \beta)$ будет составлена из двух линейных функций. Если же значения $F_U(m_V)$ и $F_V(m_U)$ расположены «сильно несимметрично» относительно $\frac{1}{2}$, то функция $\Phi_F^c(\alpha, \beta)$ будет линейной.

Для численного измерения степени устойчивости сравнения к искажениям введем следующее понятие. Назовем сравнение $r(U, V)$ для пары гистограмм $U$ и $V$ с $r(U, V) = c > 0$ $\delta$ -устойчивым к искажениям, если

$$\delta = \max\left\{k(\alpha, \beta) : (\alpha, \beta) \in \Omega_r^c(U, V)\right\},$$

где $k(\alpha, \beta)$ – некоторая критериальная функция, в качестве которой могут выступать, например: $k_1(\alpha, \beta) = \frac{1}{2}(\alpha + \beta)$, $k_2(\alpha, \beta) = \min\{\alpha, \beta\}$.

Другими словами, $\delta$ -устойчивость характеризует максимальный уровень искажений гистограмм, при котором знак сравнения этих гистограмм не изменится. Через $\delta_r^{(i)}(U, V)$ будем обозначать значение $\delta$ -устойчивости сравнения гистограмм $r(U, V)$ относительно критериальной функции $k_i$. В частности, нетрудно видеть, что

$$\delta_E^{(1)}(U, V) = \frac{c}{2\min\{\mathcal{E}_U, \mathcal{E}_V\}}, \quad \delta_E^{(2)}(U, V) = \frac{c}{\mathcal{E}_U + \mathcal{E}_V}.$$

**Пример 2.** Рассмотрим сравнение двух гистограмм средних баллов ЕГЭ абитуриентов, поступивших в 2012 г. на специальность «Экономика» и только по конкурсному набору в Московский государственный институт международных отношений (МГИМО, гистограмма $U$) и Московский государственный университет им. М.В. Ломоносова (МГУ, гистограмма $V$). На рис. 1 приведены гистограммы средних баллов ЕГЭ абитуриентов, поступивших в эти вузы.

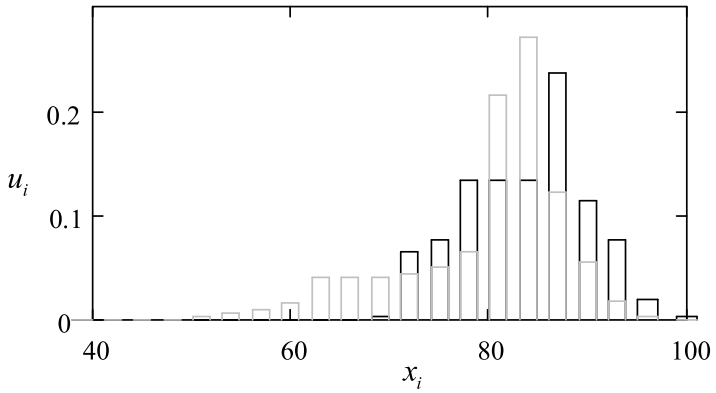


**Рис. 1.** Гистограмма средних баллов ЕГЭ абитуриентов, поступивших в 2012 г. на специальность «Экономика» в МГИМО (темный цвет) и МГУ (светлый цвет)

Для этих гистограмм нормированные математические ожидания равны $E_0[U] = 0.732$ и $E_0[V] = 0.669$, разностный индекс сравнения относительно математических ожиданий равен $\Delta_E(U,V) = E_0[U] - E_0[V] = 0.063$, разностный индекс сравнения относительно функций распределения равен $\Delta_F(V,U) = \inf\limits_{x\in(x_1,x_n]}(F_V(x) - F_U(x)) = 0.0031$, вероятности $P\{U \geqslant V\} = 0.684$, $P\{U \leqslant V\} = 0.434$, разностный индекс сравнения относительно вероятностей равен $\Delta_P(U,V) = P\{U \geqslant V\} - P\{U \leqslant V\} = 0.25$. Графики функций $\Phi_E^c(\alpha,\beta) = 1 \Leftrightarrow \alpha\mathcal{E}_U + \beta\mathcal{E}_V = c$ для $c = \Delta_E(U,V) = 0.063$, $\Phi_F^c(\alpha,\beta) = 1 \Leftrightarrow \sup\limits_x \left\{\frac{\alpha\mathcal{F}_U(x)+\beta\mathcal{F}_V(x)}{F_U(x)-F_V(x)}\right\} = c$ для $c = \Delta_F(V,U) = 0.0031$ и $\Phi_P^c(\alpha,\beta) = 1 \Leftrightarrow \Delta\eta_{\alpha,\beta}(U,V) = c$ для $c = \Delta_P(U,V) = 0.25$ приведены на рис. 2.

Заметим, что $F_U(m_V)F_V(m_U) = 0.415 \cdot 0.801 = 0.332 \geqslant \frac{1}{4}$. Поэтому функция $\Phi_F^c(\alpha,\beta) = 1$ является линейной и, согласно (9), $\Phi_F^c(\alpha,\beta) = \varphi_{\alpha,\beta}(m_U)$.

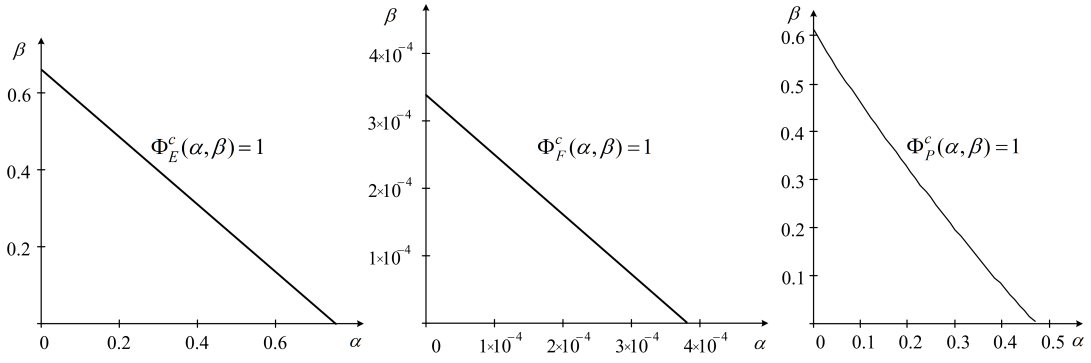


**Рис. 2.** Графики функций: а) $\Phi_E^c(\alpha,\beta) = 1$; б) $\Phi_F^c(\alpha,\beta) = 1$; в) $\Phi_P^c(\alpha,\beta) = 1$

Тогда для сравнений относительно:

— математических ожиданий: $\delta_E^{(1)}(U,V) = 0.375$, $\delta_E^{(2)}(U,V) = 0.351$;
— функций распределения: $\delta_F^{(1)}(U,V) = 0.001989$; $\delta_F^{(2)}(U,V) = 0.001788$;
— вероятностей: $\delta_P^{(1)}(U,V) = 0.306$, $\delta_P^{(2)}(U,V) = 0.254$.

Таким образом, наибольшую устойчивость (на уровне 35%-40%) демонстрирует сравнение с помощью математического ожидания. Немного уступает ему (25%-30%) сравнение с помощью вероятностей (стохастическое предшествование). Самая низкая устойчивость (0.15%-0.20%) у сравнения с помощью функции распределения (стохастическое доминирование).

## Анализ устойчивости сравнений гистограмм результатов ЕГЭ

Исследуем устойчивость к $\alpha$ –зашумлению сравнений гистограмм с помощью математических ожиданий средних баллов ЕГЭ абитуриентов, поступивших в 2012 г. на специальность «Экономика» и только по конкурсному набору. Результаты исследований для 10-ти вузов из верхней части ранжирования по среднему баллу приведены на рис. 3. Вузы расположены слева направо в порядке убывания среднего балла. Для каждой пары «соседних» гистограмм $U$ и $V$ двух вузов в верхней половине ячейки указано значение индекса $\Delta_E(U,V)$, а в нижней – значение $\delta$ -устойчивости $\delta_E^{(2)}(U,V) = \frac{c}{\mathcal{E}_U + \mathcal{E}_V}$ относительно максиминного критерия $k_2(\alpha,\beta)$.

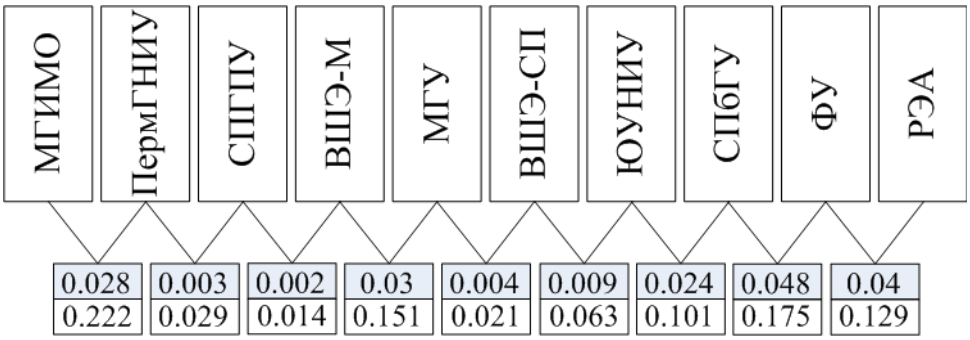


**Рис. 3.** Значения индекса $\Delta_E(U,V)$ (вверху) и $\delta$ -устойчивости $\delta_E^{(2)}(U,V)$ (внизу)

**Примечание.** Использованы следующие аббревиатуры вузов: 1) МГИМО — Московский государственный институт международных отношений; 2) ПермГНИУ — Пермский государственный национальный исследовательский университет; 3) СПГПУ — Санкт-Петербургский государственный политехнический университет; 4) ВШЭ-М — Национальный исследовательский университет «Высшая школа экономики», г. Москва; 5) МГУ — Московский государственный университет им. М. В. Ломоносова; 6) ВШЭ-СП — Национальный исследовательский университет «Высшая школа экономики», г. Санкт-Петербург; 7) ЮУ НИУ — Национальный исследовательский Южно-Уральский государственный университет, г. Челябинск; 8) СПбГУ — Санкт-Петербургский государственный университет; 9) ФУ — Финансовый университет при Правительстве Российской Федерации, г. Москва 10) РЭА — Российская экономическая академия им. Г. В. Плеханова, г. Москва.

Таким образом, на рис. 3 представлены максимальные значения зашумлений, при которых сравнения в парах не нарушатся. Эти значения колеблются от 1.4% до 22.2% . Нетрудно видеть, что хотя значения $\delta$ -устойчивости прямо пропорциональны значению

разностного индекса сравнения, но далеко не всегда большему значению разностного индекса сравнения соответствует и большее значение $\delta$ -устойчивости (сравните, например, пары «МГИМО–ПермГНИУ» и «ВШЭ-М – МГУ»).

## Заключение

В работе найдены необходимые и достаточные условия на уровень искажений гистограмм, при выполнении которых результат сравнения гистограмм вероятностными методами не изменится.

Из умозрительных соображений априори было понятно, что «интегральные» методы сравнение, такие как метод сравнения математических ожиданий, метод стохастического предшествования, будут более предпочтительными, чем методы поточечного сравнения, такие как стохастическое доминирование. Однако в результате проведенных исследований были не только подтверждены эти предположения, но и получены точные теоретические оценки возможных значений искажений гистограмм, при которых результат сравнения не будет меняться.

Найденные условия неизменности сравнения гистограмм могут быть использованы для оценивания надежности результатов различных ранжирований, обработки данных и пр. в условиях различных типов неопределенности: стохастической неопределенности; неопределенности, связанной с умышленным искажением данных, с заполнением пробелов в данных и т.д.

Дальнейшие исследования в направлении оценивания устойчивости сравнений могут быть связаны как с исследованием других методов сравнения (в том числе наиболее популярных и важных в приложениях нечеткостных методов сравнения), так и с другим описанием неопределенности.

## Литература


[1] *Rothschild M., Stiglitz J. E.* Some further results on the measurement of inequality // J. Economic Theory, 1973. Vol. 6. P. 188–204.

[2] *Алескеров Ф. Т., Белоусова В.Ю., Солодков В. М., Сердюк М. Ю.* Динамический анализ стереотипов поведения крупнейших российских коммерческих банков // Модернизация экономики и глобализация: В 3 кн. / Отв. ред. Е. Г. Ясин. — М.: Издательский дом ГУ-ВШЭ, 2009. Кн. 3. С. 371–381.

[3] *Sen A. K.* On economic inequality. Oxford: University Press, 1973.

[4] *Bronevich A. G., Rozenberg I. N.* Ranking probability measures by inclusion indices in the case of unknown utility function // Fuzzy Optimization and Decision Making, 2014. Vol. 13. No. 1. P. 49–71.

[5] *Fodor J., Roubens M.* Fuzzy preference modelling and multicriteria decision support. — Dordrecht: Kluwer Academic Publs., 1994.

[6] *Shnoll S. E., Zenchenko K. I., Udaltsova N. V.* Cosmophysical effects in the structure of daily and yearly periods of changes in the shape of histograms constructed from the measurements of 239P u alpha-activity // Biophysics, 2004. Vol. 49. Suppl. 1. P. 155.

[7] *Shorrocks A. F.* Ranking income distributions // Economica, 1983. Vol. 50. P. 3–17.

[8] *Dubois D., Prade H.* Ranking fuzzy numbers in the setting of possibility theory // Information Sci., 1983. Vol. 30. P. 183–224.

[9] *Подиновский В. В.* Введение в теорию важности критериев в многокритериальных задачах принятия решений. — М.: Физматлит, 2007.

[10] *De Santis E., Fantozzi F., Spizzichino F.* Relations between stochastic orderings and generalized stochastic precedence. 2014. http://arxiv.org/pdf/1307.7546.pdf.

[11] *Бобров Р. А., Лепский А. Е.* Ранжирование вузов по баллам ЕГЭ методами сравнения нечетких чисел. WP7/2014/01. 2014. 24 с.

[12] *Wang X., Ruan D., Kerre E. E.* Mathematics of fuzziness — basic issues. Berlin–Heidelberg: Springer-Verlag, 2009.

[13] *Ватник П. А.* Теория риска: учеб. пособие. СПб.: С.-Петерб. гос. инж.-экон. ун-т., 2009.

[14] *Yager R. R.* A procedure for ordering fuzzy sets of the unit interval // Information Sci., 1981. Vol. 24. P. 143–161.

[15] *Boland Ph. J., Singh H., Cukic B.* The stochastic precedence ordering with applications in sampling and testing // J. Appl. Probability, 2004. Vol. 41, No. 1. P. 73–82.

[16] *Cassels J. W. S.* An introduction to the geometry of numbers. — Springer-Verlag, 1959.

[17] *Vanegas L. V., Labib A W.* Application of new fuzzy-weighted average (NFWA) method to engineering design evaluation // Int. J. Production Research, 2001. Vol. 39. P. 1147–1163.

[18] *Wolfstetter E.* Topics in microeconomics: Industrial organization, auctions, and incentives. — Cambridge: Cambridge University Press, 1999.

[19] *Польдин О. В., Силаев А. М.* Сравнение образовательных программ по результатам ЕГЭ зачисленных студентов // Вопросы образования, 2011. Т. 3. С. 192–209.

[20] *Baas S. M., Kwakernaak H.* Rating and ranking of multiple-aspect alternatives using fuzzy sets // Automatic, 1977. No. 13. P. 47–58.

[21] *Lepskiy A.* On the stability of comparing histograms with help of probabilistic methods // Procedia Computer Sci., 2014. Vol. 31. P. 597–605.

[22] *Шахнов И. Ф.* Задачи ранжирования интервальных величин при многокритериальном анализе сложных систем // Известия РАН. Теория и системы управления, 2008. Т. 1. С. 37–44.

[23] *Aleskerov F. T., Chistyakov V. V., Kaliaguine V. A.* Social threshold aggregations // Social Choice Welfare, 2010. Vol. 35. No. 4. P. 627–646.

## References

[1] *Rothschild M., Stiglitz J. E.* 1973. Some further results on the measurement of inequality. *J. Economic Theory* 6:188–204.

[2] *Aleskerov F. T., Belousova V. Y., Solodkov V. M., Serdyuk M. Y.* 2009. Dynamic analysis of the behavioural patterns of the largest commercial banks in the Russian Federation. *Economic Modernization and Globalization.* Ed. E. G. Yasin. Moscow: Publishing House of the HSE. 3:371–381.

[3] *Sen A. K.* 1973. *On economic inequality*. Oxford: University Press.

[4] *Bronevich A. G., Rozenberg I. N.* 2014. Ranking probability measures by inclusion indices in the case of unknown utility function. *Fuzzy Optimization Decision Making* 13(1):49–71.

[5] *Fodor J., Roubens M.* 1994. *Fuzzy preference modelling and multicriteria decision support.* Dordrecht: Kluwer Academic Publs.

[6] *Shnoll S. E., Zenchenko K. I., Udaltsova N. V.* 2004. Cosmophysical effects in the structure of daily and yearly periods of changes in the shape of histograms constructed from the measurements of 239P u alpha-Activity. *Biophysics* 49(Suppl. 1):155.

[7] *Shorrocks A. F.* 1983. Ranking income distributions // *Economica* 50:3–17.

[8] *Dubois D., Prade H.* 1983. Ranking fuzzy numbers in the setting of possibility theory // *Information Sci.* 30:183–224.

[9] *Podinovski V. V.* 2007. *Introduction to the theory of importance of criteria in multicriteria decision-making problems.* Moscow.: Fizmatlit.

[10] *De Santis E., Fantozzi F., Spizzichino F.* 2014. Relations between stochastic orderings and generalized stochastic precedence. `http://arxiv.org/pdf/1307.7546.pdf`.

[11] *Bobrov R. A., Lepskiy A. E.* 2014. Ranking universities according to the results of USE by means of fuzzy numbers comparison methods. Working Paper WP7/2014/01. Moscow: Publishing House of the HSE. 24 p.

[12] *Wang X., Ruan D., Kerre E. E.* 2009. *Mathematics of fuzziness — basic issues.* Berlin–Heidelberg: Springer-Verlag.

[13] *Vatnik P. A.* 2009. *Risk theory: a tutorial.* St. Petersburg: St. Petersburg State University of Engineering and Economics.

[14] *Yager R. R.* 1981. A procedure for ordering fuzzy sets of the unit interval. *Information Sci.* 24:143–161.

[15] *Boland Ph. J., Singh H., Cukic B.* 2004. The stochastic precedence ordering with applications in sampling and testing. *J. Appl. Probability* 41(1):73–82.

[16] *Cassels J. W. S.* 1959. *An introduction to the geometry of numbers.* Springer-Verlag.

[17] *Vanegas L. V., Labib A W.* 2001. Application of new fuzzy-weighted average (NFWA) method to engineering design evaluation. *Int. J. Production Research* 39:1147–1163.

[18] *Wolfstetter E.* 1999. *Topics in microeconomics: Industrial organization, auctions, and incentives.* Cambridge: Cambridge University Press.

[19] *Poldin O. V., Silaev A. M.* 2011. Comparison of educational programs on the USE results of enrolled students. *Educational Studies* 3:192–209.

[20] *Baas S. M., Kwakernaak H.* 1977. Rating and ranking of multiple-aspect alternatives using fuzzy sets. *Automatic* 13:47–58.

[21] *Lepskiy A.* 2014. On the stability of comparing histograms with help of probabilistic methods. *Procedia Computer Sci.* 31:597–605.

[22] *Shahnov I. F.* 2008. A problem of ranking interval objects in a multicriteria analysis of complex systems. *J. Computer Systems Sciences International* 47:33–39.

[23] *Aleskerov F. T., Chistyakov V. V., Kaliaguine V. A.* 2010. Social threshold aggregations. *Social Choice Welfare* 35(4):627–646.