On the stability of comparing histograms with help of probabilistic methods

Alexander Lepskiy

National Research University - Higher School of Economics,
Moscow, Russia

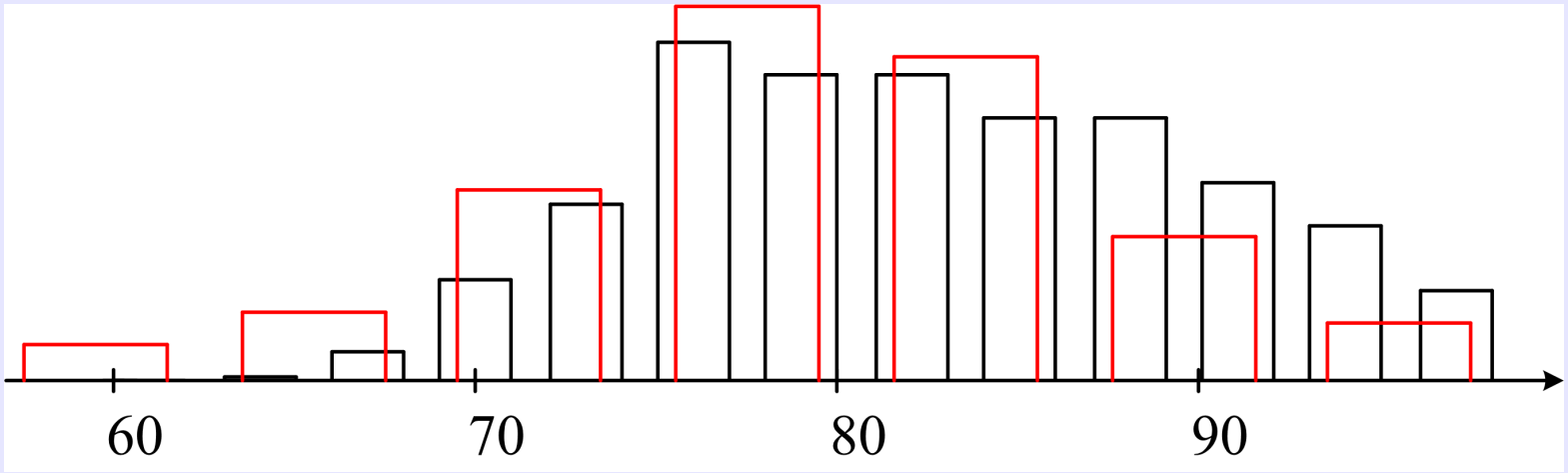
The 2^{st} International Conference on Information Technology
and Quantitative Management,
June 3 - 5, 2014, Moscow, Russia

Outline of Presentation

1. Comparison of histograms
   - Problem statement of comparison of histograms
   - Applied problems where comparison of histograms is used
   - Main approaches for comparison of histograms
   - Some Probabilistic Indices of Comparison
2. Distortions of Histograms
3. Conditions of Preservation for Comparison of Distorted Histograms
4. Comparison of the Sets of Admissible Distortions
5. Example. Histograms of Unified State Exam of Universities
6. Summary and conclusion

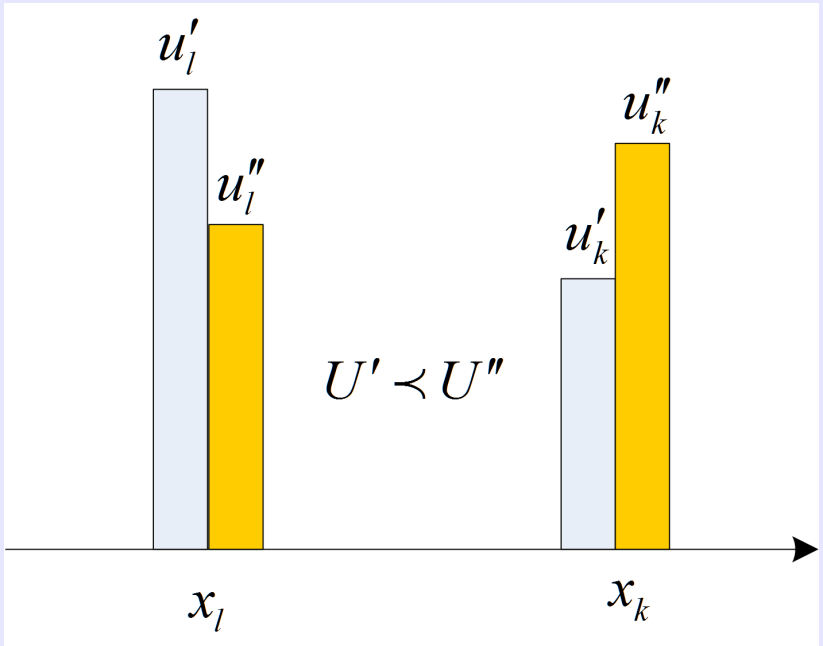
Problem statement of Comparison of Histograms

Let $\mathcal{U} = \{U\}$ be a set of all histograms of form $U = (x_i, u_i)_{i \in I}$, $x_i < x_{i+1}$, $i \in I$.
We want define the total preorder relation R (reflexive, complete and transitive relation) on $\mathcal{U}$: $(U, V) \in R \Leftrightarrow U \succeq V$.



Ordering Arguments of Histograms

The relation R should be in accord with the condition of the ordering of histogram arguments by ascending their importance:
if $U' = (x_i, u'_i)$, $U'' = (x_i, u''_i)$ be two histograms for which $u'_i = u''_i$ for all $i \neq k, l$ and $u'_l - u''_l = u''_k - u'_k \geq 0$ then $U'' \succeq U'$ for $k > l$ and $U' \succeq U''$ for $k < l$.



Application of Comparison of Histograms

- comparison of results of different experiences;
- comparison of indicators of functioning of the organizational, technical systems etc.;
- decision-making under fuzzy uncertainty;
- simulation of fuzzy preferences;
- comparisons of income distribution within the framework of socio-economic analysis;
- ranking of histogram data

  etc.

Main Approaches for Comparison of Histograms

- probabilistic approach;
- ranking methods of income distribution in the theory of social choice.
  Histograms income has the form $U = (i, u_i)_{i=1}^{n_U} = (u_i)_{i=1}^{n_U}$, where $u_1 \leq u_2 \leq ... \leq u_{n_U}$ in this case. These histograms are compared with help of welfare functions $W(U)$ that satisfy the conditions of symmetry, monotonicity, concavity, etc.
- using the tools of comparison of fuzzy numbers.
  The histogram $U = (x_i, u_i)_{i \in I}$ is associated with fuzzy set (or fuzzy number) by means of membership function $U = (u_i)_{i \in I}$ which is defined on the universal set $X = (x_i)_{i \in I}$.

Some Probabilistic Indices of Comparison

We consider a numerical **index** $r(U,V)$ **of pairwise comparison** of histograms U and V in $\mathcal{U}^2$.

Let index $r(U,V)$ is consistent with increasing of importance of arguments: if $U=(x_i,u_i)$, $V=(x_i,v_i)$ be two histograms for which $u_i=v_i$ for all $i\neq k,l$ and $u_l-v_l=v_k-u_k\geq 0$ then $r(U,V)\geq 0$ for $k>l$ and $r(U,V)\leq 0$ for $k<l$. In particular $r(U,U)=0$.

Let $\Delta_r(U,V)=r(U,V)-r(V,U)\geq 0$ be a **differential index of comparison**.

Let $U=(x_i,u_i)_{i\in I}$ and $V=(x_j,v_j)_{j\in I}$ are random variables taking values $\{x_i\}_{i\in I}$ with probabilities $\{u_i\}_{i\in I}$ and $(v_j)_{j\in I}$ respectively.

Examples of Indices Pairwise Comparison of Histograms

1. Comparison of mathematical expectations

Let $U \succeq V$ if $E[U] \geq E[V]$. In general $U \succeq V$ if $E[f(U)] \geq E[f(V)]$, where f is some utility function.

Let $E_0[U] = \frac{1}{\Delta x}(E[V] - x_{\min})$ be a normalized index, where $\Delta x = x_{\max} - x_{\min}$, $E_0[U] \in [0, 1]$.

Let $\Delta_E(U, V) = E_0[U] - E_0[V] = \frac{1}{\Delta x}(E[U] - E[V])$ be a corresponding differential comparison index.

2. Comparison of distribution functions

Let $U \succeq V$ if $F_U(x) \leq F_V(x)$ for all $x \in \mathbb{R}$, where $F_U(x) = \sum_{i:x_i<x} u_i$ is distribution function of random variable U.

This is a principle of **stochastic dominance of the 1st order**.

Let $\Delta_F(U, V) = \inf\limits_{x \in (x_{min}, x_{max}]} (F_U(x) - F_V(x))$ be a corresponding differential comparison index.

3. Comparison of probabilities

Let $U \succeq V$ if $P\{U \geq V\} \geq P\{U \leq V\}$. This approach to comparison called by **stochastic precedence** (V precedes U).

If we assume that the random variables $U = (x_i, u_i)_{i \in I}$ and $V = (x_j, v_j)_{j \in I}$ are independent then $P\{U \geq V\} = \sum\limits_{(i,j):\, x_i \geq x_j} u_i v_j$.

The corresponding differential comparison index is denoted by $\Delta_P(U, V) = P\{U \geq V\} - P\{U \leq V\}$.

Notice that the inequality $\Delta_P(U, V) \geq 0$ does not specify a transitive relation.

Distortions of Histograms

The compared histograms may be distorted.

The reasons of distortions:

- random noise;
- deliberate distortion of data;
- filling gap in incomplete data;
- etc.

The α-distortion of histogram.

Let $U = (x_i, u_i)_{i \in I}$ is a "ideal" histogram and $\tilde{U} = (x_i, \tilde{u}_i)_{i \in I}$ is an interval distortion of U: $\tilde{u}_i = u_i + h_i$, $i \in I$, where $\sum_{i \in I} h_i = 0$ and $|h_i| \leq \alpha u_i$, $i \in I$, where $\alpha \in [0, 1]$. The value α characterize the threshold of distortion.

Let

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

be a class of all α-distortion of histogram $U = (x_i, u_i)_{i \in I}$.

Main problem

Suppose that $\Delta_r(U, V) > 0$. In what case do we have $\Delta_r(\tilde{U}, \tilde{V}) \ge 0$ for all $H \in N_\alpha(U)$ and $G \in N_\beta(V)$?
By other words, when the comparison of histograms will not changed after α-distortion of histogram $U = (x_i, u_i)_{i \in I}$ and β-distortion of histogram $V = (x_j, v_j)_{j \in I}$?

Conservation Conditions of Comparison w.r.t. $\Delta_E(U,V) = \frac{1}{\Delta x}(E[U] - E[V])$ Index

We consider the value

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

for $U = (x_i, u_i)_{i\in I}$, where $x_i^0 = \frac{1}{\Delta x}(x_i - x_{\min}) \in [0,1]\ \forall i \in I$.

Lemma

The estimation $0 \le \mathcal{E}_U \le \min\{E_0[U], 0.5\}$ *is true.*

Proposition

Let $\tilde{U} = (x_i, u_i + h_i)_{i\in I}$, $\tilde{V} = (x_j, v_j + g_j)_{i\in I}$ *be a* α- *and* β-*distortion of histograms* $U = (x_i, u_i)_{i=1}^n$ *and* $V = (x_j, v_j)_{j=1}^n$ *respectively. Then we have* $\Delta_E(\tilde{U}, \tilde{V}) \ge 0$ *for all* $(h_i)_{i\in I} \in N_\alpha(U)$ *and* $(g_i)_{i\in I} \in N_\beta(V)$, $\alpha, \beta \in [0,1]$ *iff* $\Delta_E(U,V) \ge \alpha\mathcal{E}_U + \beta\mathcal{E}_V$.

Let $\bar{\mathcal{E}}_U = \min\{E_0[U], 0.5\}$.

Corollary

If we have $\Delta_E(U,V) \geq \alpha\bar{\mathcal{E}}_U + \beta\bar{\mathcal{E}}_V$, *then inequality* $\Delta_E(\tilde{U},\tilde{V}) \geq 0$ *is true for all* $(h_i)_{i\in I} \in N_\alpha(U)$ *and* $(g_i)_{i\in I} \in N_\beta(V)$.

Conservation Conditions of Comparison w.r.t $\Delta_F(U,V) = \inf_{x\in(x_{min},x_{max}]} (F_U(x) - F_V(x))$ Index

Let $\mathcal{F}_U(x) = \sup\left\{\sum_{i:x_i<x} h_i : \ (h_i)_{i\in I} \in N_1(U)\right\}$.

Lemma

$\mathcal{F}_U(x) = \min\{F_U(x), 1 - F_U(x)\}$ *for all* $x \in \mathbb{R}$.

Proposition

Let $\tilde{U} = (x_i, u_i + h_i)_{i\in I}$, $\tilde{V} = (x_j, v_j + g_j)_{i\in I}$ *be a* α- *and* β-*distortion of histograms* $U = (x_i, u_i)_{i\in I}$ *and* $V = (x_j, v_j)_{i\in I}$ *respectively. Then we have* $\Delta_F(\tilde{U}, \tilde{V}) \geq 0$ *for all* $(h_i)_{i\in I} \in N_\alpha(U)$ *and* $(g_i)_{i\in I} \in N_\beta(V)$ *iff*

$$F_U(x) - F_V(x) \geq \alpha\mathcal{F}_U(x) + \beta\mathcal{F}_V(x) \ \text{ for all } \ x \in \mathbb{R}.$$

Corollary

The inequality $\Delta_F(\tilde{U},\tilde{V}) \geq 0$ *is true for all* $(h_i)_{i\in I} \in N_\alpha(U)$ *and* $(g_i)_{i\in I} \in N_\beta(V)$ *iff* $0 \leq \sup\limits_x \frac{\alpha\mathcal{F}_U(x)+\beta\mathcal{F}_V(x)}{F_U(x)-F_V(x)} \leq 1$ *(the fraction is equal to zero if its numerator and denominator are equal to zero).*

Corollary

If $\Delta_F(U,V) \geq \sup\limits_x \{\alpha\mathcal{F}_U(x) + \beta\mathcal{F}_V(x)\}$ *then inequality* $\Delta_F(\tilde{U},\tilde{V}) \geq 0$ *is true for all* $(h_i)_{i\in I} \in N_\alpha(U)$ *and* $(g_i)_{i\in I} \in N_\beta(V)$.

Conservation Conditions of Comparison w.r.t. $\Delta_P(U,V) = P\{U \geq V\} - P\{U \leq V\}$ Index

Proposition

Let $\tilde{U} = (x_i, u_i + h_i)_{i \in I}$, $\tilde{V} = (x_j, v_j + g_j)_{j \in I}$ *be a* α- *and* β-*distortion of histograms* $U = (x_i, u_i)_{i \in I}$ *and* $V = (x_j, v_j)_{j \in I}$ *respectively. Then we have* $\Delta_P(\tilde{U}, \tilde{V}) \geq 0$ *for all* $(h_i)_{i \in I} \in N_\alpha(U)$ *and* $(g_i)_{i \in I} \in N_\beta(V)$, $\alpha, \beta \in [0,1]$ *iff* $\Delta_P(U,V) \geq \Delta\eta_{\alpha,\beta}(U,V)$, *where*

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

Corollary

If

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

then inequality $\Delta_P(\tilde{U},\tilde{V}) \geq 0$ *is true* $\forall (h_i)_{i\in I} \in N_\alpha(U)$, $(g_i)_{i\in I} \in N_\beta(V)$.

Corollary

If

$$\Delta_P(U,V) \geq \frac{\alpha+\beta+\alpha\beta}{1+\alpha+\beta+\alpha\beta}, \tag{2}$$

then inequality $\Delta_P(\tilde{U},\tilde{V}) \geq 0$ $\forall (h_i)_{i\in I} \in N_\alpha(U)$, $(g_i)_{i\in I} \in N_\beta(V)$.

Remark. The condition (2) gives weaker restrictions on distortions of histograms which preserve their comparison relative differential index $\Delta_P(U,V)$ than condition (1).

Comparison of the Sets of Admissible Distortions

Let

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

be a set of admissible distortions of histograms U and V for given comparison $\Delta_r(U,V)=c$.
The set $\Omega_r^c(U,V)$ has a form

$$\Omega_r^c(U,V)=\left\{(\alpha,\beta):\ \alpha\geq 0,\beta\geq 0,\Phi_r^c(\alpha,\beta)\leq 1\right\},$$

where $\Phi_r^c(\alpha,\beta)$ is a ray function (i.e. continuous, non-negative and homogeneous).

Stability to Distortion

We have

- $\Phi_E^c(\alpha,\beta) = \frac{1}{c}\left(\alpha\mathcal{E}_U + \beta\mathcal{E}_V\right)$ for index $\Delta_E(U,V)$;
- $\Phi_F^c(\alpha,\beta) = \sup\limits_x \left\{\frac{\alpha\mathcal{F}_U(x)+\beta\mathcal{F}_V(x)}{F_U(x)-F_V(x)}\right\}$ for index $\Delta_F(U,V)$;
- $\Phi_P^c(\alpha,\beta) = \frac{1}{c}\Delta\eta_{\alpha,\beta}(U,V)$ for index $\Delta_P(U,V)$.

We call the comparison $r(U,V) = c > 0$ by δ-**stable** to distortion if

$$\delta = \delta_r^{(i)}(U,V) = \max\left\{k(\alpha,\beta):\ \Phi_r^c(\alpha,\beta) \le 1\right\},$$

where $k(\alpha,\beta)$ is a some **criterial function**, as which the may be, for example: $k_1(\alpha,\beta) = \frac{1}{2}(\alpha+\beta)$, $k_2(\alpha,\beta) = \min\{\alpha,\beta\}$.
The δ-stability characterizes the **maximal level of distortions** of histograms for which the sign of comparison histograms will not change. In particular, $\delta_E^{(1)}(U,V) = \frac{c}{2\min\{\mathcal{E}_U,\mathcal{E}_V\}}$, $\delta_E^{(2)}(U,V) = \frac{c}{\mathcal{E}_U+\mathcal{E}_V}$.

Example. Histograms of Unified State Exam of Universities

We consider the comparison of the two histograms of USE (Unified State Exam) applicants admitted in 2012 on a speciality "Economy" in Moscow State Institute of the International Relations (MGIMO, the histogram U) and Moscow State University (MSU, the histogram V).

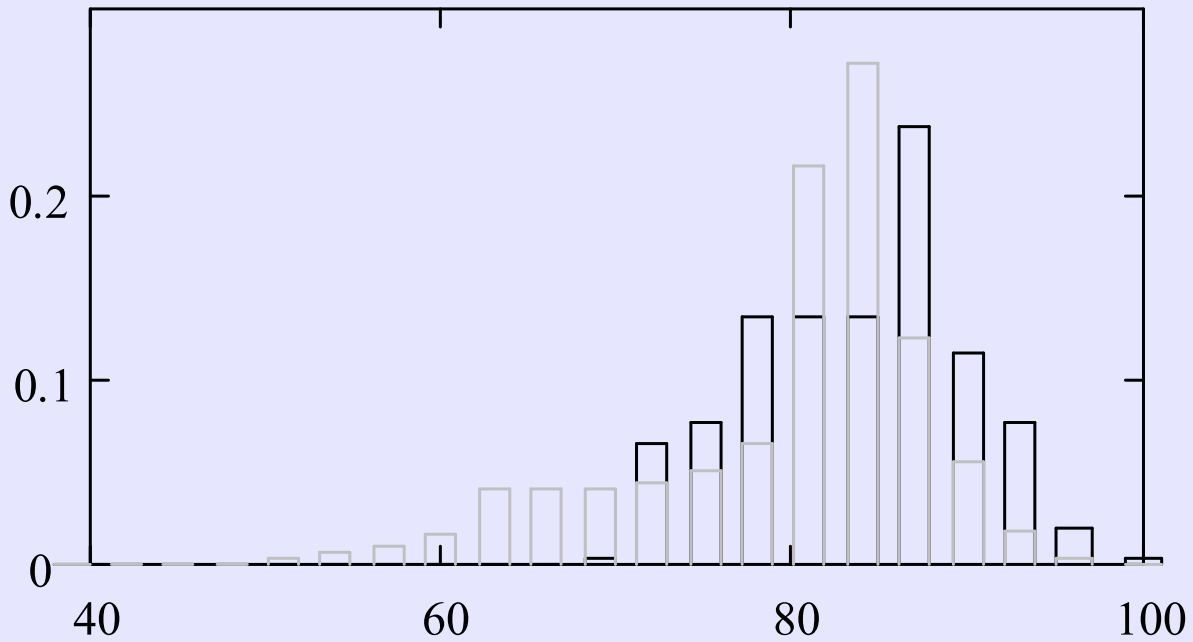


Figure: Histograms of MGIMO (dark color) and MSU (light color).

The Results of Analysis of Stability

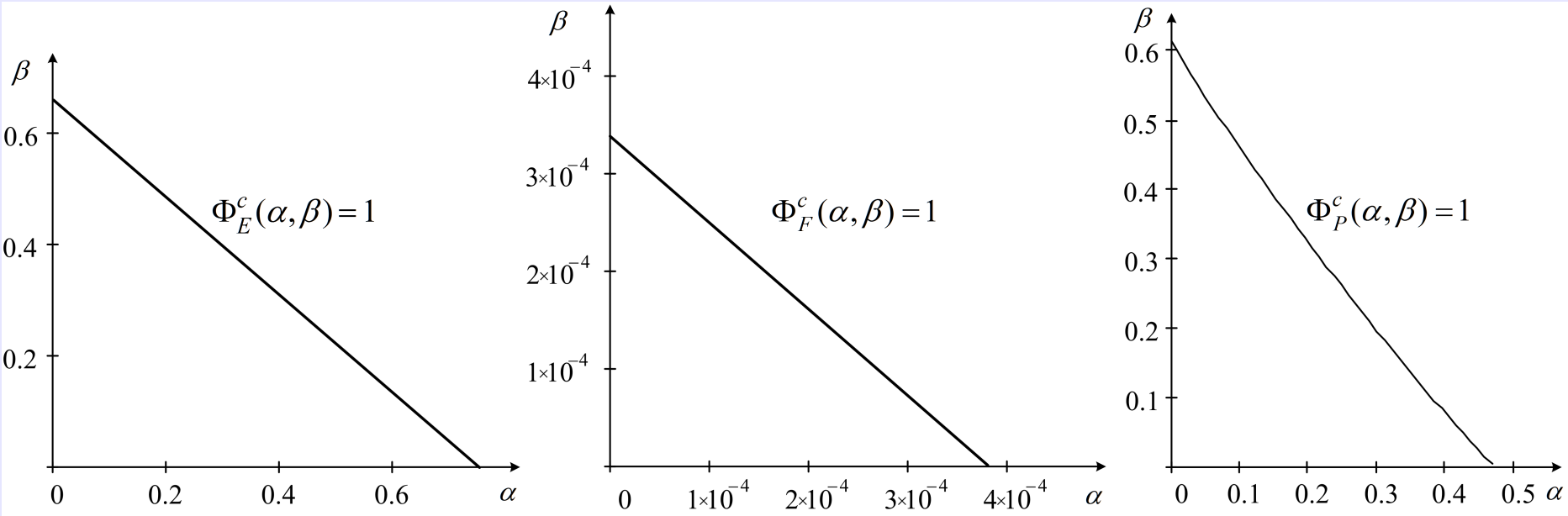
- differential index of comparison w.r.t. **expectations**:
  $\Delta_E(U,V) = E_0[U] - E_0[V] = 0.063$
- differential index of comparisons w.r.t. **distribution functions**:
  $\Delta_F(V,U) = \inf\limits_{x\in(x_1,x_n]} (F_V(x) - F_U(x)) = 0.0031;$
- differential index of comparisons w.r.t. **probabilities**:
  $\Delta_P(U,V) = P\{U \geq V\} - P\{U \leq V\} = 0.25.$

The values of δ-stability of comparisons of histograms w.r.t.:

- expectations: $\delta_E^{(1)}(U,V) = 0.375$, $\delta_E^{(2)}(U,V) = 0.351$;
- distribution functions: $\delta_F^{(1)}(U,V) = 0.00199$; $\delta_F^{(2)}(U,V) = 0.00179$;
- probabilities: $\delta_P^{(1)}(U,V) = 0.306$, $\delta_P^{(2)}(U,V) = 0.254$.

Thus the comparisons w.r.t. expectation shows the greatest stability (at the level of 35-40%). The comparisons w.r.t. probability slightly worse than the first comparison (25-30%). The comparison w.r.t. distribution function has the lowest stability (0.15-0.20%).

Graphs of Boundaries of Admissible Distortions Sets



Summary and Conclusion

- The necessary and sufficient conditions on the distortion level of histograms, under which the result of the comparison of histograms by probabilistic methods will not change, were found
- It was confirmed that "integral" methods of comparison (for example, method of comparing expectations, method comparisons of probability) are more stable than pointwise comparison methods, such as stochastic dominance.
- The conditions of invariability of comparing histograms can be used to estimate the reliability of results of different rankings, data processing, etc.
- The different types of uncertainty of data may be associated with considered model of distortion of histograms. For example, it may be stochastic uncertainty, the uncertainty associated with the distortion of the data in filling data gaps, etc.

Thanks for you attention

alex.lepskiy@gmail.com
http://lepskiy.ucoz.com