Proximity measure between samples with repetition factor greater than one

R.I.Andrushkiw

Department Mathematical Sciences and Center for Applied Mathematics and Statistics, New Jersey Institute of Technology, Newark, N.J., USA

D.A.Klyushin, Yu.I.Petunin

Department of Cybernetics, Kyiv National Taras Shevchenko University, Kyiv, Ukraine

Abstract. *A new proximity measure between empirical samples, having values x_k that may occur in a sample x more than once, is constructed. This proximity measure is based on confidence intervals containing the bulk of population constructed by means of order statistics.*

Key Words: proximity measure, atom, order statistics, confidence interval.

1. Introduction.

Let $x = (x_1, x_2, ..., x_n)$ be a sample drawn from general population G with distribution function $F(u)$ by simple random sampling. The *atom* of the sample x is a sample value x_k that occurs in the sample x more than once:

$$x_k = x_{k_1} = ... = x_{k_i},$$

$$k, k_1, ..., k_{i_k} \subset \{1, 2, ..., n\}.$$

The number of repetitions of the value x_k in the sample x shall be called a *repetition factor* $t(x_k)$. Thus, atoms are sample values with repetition factor greater than 1. If $F(u)$ is continuous and the values of x_k are exact, then the probability of atoms in x is zero and we shall refer to such a sample as *hypothetical*. However, as a rule, sample values are the results of measuring some random variable. Since every measurement is subject to some error, the measured sample $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$, $\tilde{x}_k \in \tilde{x}$, may contain atoms (such sample we shall call *empirical*). Unfortunately, the well-known proximity measures between two samples (Kolmogorov-Smirnov statistics, Wilkoxon statistics, p-statistics [1, 2]) cannot be applied to atomic samples. The purpose of this paper is to modify the *p*-statistics in a such way that it may be used to calculate the similarity of empirical samples and to construct a corresponding test.

2. Proximity measure between empirical samples. Let us introduce the following notation: P_α is the flooring operator up to the decimal number α,

$$|P(x) - x| \le \delta = 10^{-\alpha},$$

where δ is a rounding error, and

$$x_{(1)} < x_{(2)} < ... < x_{(n)},\ \tilde{x}_{(1)} < \tilde{x}_{(2)} < ... < \tilde{x}_{(m)}$$

are variational series constructed on samples x and $\tilde{x}$.

If x^* is a sample value drawn from the general population G and independent from x, then it is well-known that

$$p\left(x^* \in \left[x_{(k)}, x_{(k+1)}\right)\right) = \frac{1}{n+1}, \qquad (1)$$

$$k = 0, 1, ..., n,\ x_{(0)} = -\infty,\ x_{(n+1)} = \infty.$$

To extend this formula to empirical samples we need to prove the following lemma.

Lemma. *If the hypothetical distribution function $F(v)$ of the general population G is differentiable and satisfies Lipschitz condition with module of continuity K*

$$|F(v) - F(u)| \le K|v - u|, \qquad (2)$$

and the sample value x^* is independent from x, then for every $\delta > 0$ the following inequality is true,

$$p\left(x_{(k)} - \delta \le x^* < x_{(k)}\right) \le \le K\delta(n - k + 1) \qquad (3)$$

Proof. Let ξ and η be random variables with continuous distribution functions $F_\xi(u)$ and $F_\eta(u)$ respectively. It was proved in [3] that

$$p(\xi < \eta) = \int_{-\infty}^{\infty} F_{\xi}(v)\, dF_{\eta}(v).$$

Therefore,

$$p\left(x^* < x_{(k)} - \delta\right) = p\left(x^* + \delta < x_{(k)}\right) =$$

$$= \int_{-\infty}^{\infty} F_{x^*+\delta}(v)\, f_{x_{(k)}}(v)\, dv =$$

$$= nC_{n-1}^{k-1} \int_{-\infty}^{\infty} \left[F(v)\right]^{k-1} \left[1 - F(v)\right]^{n-k} \times$$

$$\times F(v - \delta)\, dF(v).$$

According to Lipschitz condition,

$$F(v - \delta) \geq F(v) - K\delta.$$

Substituting $u = F(v)$, we have

$$p\left(x^* < x_{(k)} - \delta\right) \geq nC_{n-1}^{k-1} \int_{-\infty}^{\infty} \left[F(v)\right]^{k-1} \times$$

$$\times \left[1 - F(v)\right]^{n-1} \left[F(v) - K\delta\right] dF(v) =$$

$$= nC_{n-1}^{k-1} \int_{-\infty}^{\infty} \left[F(v)\right]^{k} \left[1 - F(v)\right]^{n-1} \times$$

$$\times dF(v) - K\delta n C_{n-1}^{k-1} \int_{-\infty}^{\infty} \left[F(v)\right]^{k-1} \times$$

$$\times \left[1 - F(v)\right]^{n-1} dF(v) =$$

$$= \frac{k}{n+1} - K\delta n C_{n-1}^{k-1} \times$$

$$\times \int_{0}^{1} u^{k-1} (1-u)^{n-1}\, du = J. \qquad (4)$$

Intergrating by parts, we obtain

$$\int_{0}^{1} u^{k-1} (1-u)^{n-k}\, du$$

$$= \frac{k-1}{n-(k+1)+1} \int_{0}^{1} u^{k-2} (1-u)^{n-k+1}\, du,$$

$$\int_{0}^{1} u^{k-1} (1-u)^{n-k}\, du =$$

$$= \frac{1 \cdot 2 \cdot \ldots \cdot (k-2)(k-1)}{(n-k+2)(n-k+3)\ldots(n-1)n}.$$

Therefore,

$$J = \frac{k}{n+1} - K\delta n \frac{(n-1)!}{(k-1)!(n-k)!} \times$$

$$\times \frac{(k-1)!}{n!} == \frac{k}{n+1} - K\delta(n-k+1) \quad (5)$$

Using relations (4) and (5), we have

$$p\left(x_{(k)} - \delta \leq x^* < x_{(k)}\right) =$$

$$p\left(x^* < x_{(k)}\right) - p\left(x^* + \delta < x^*\right) \leq$$

$$\leq \frac{k}{n+1} - \frac{k}{n+1} + K(n-k+1)\delta =$$

$$= K(n-k+1)\delta.$$

The lemma is proved.

Remark 1. If the distribution function $F(u)$ satisfies the Hölder condition $\left|F(u) - F(v)\right| \leq K\left|u - v\right|^{\theta} \quad \forall u, v \in R^1$ with index $\theta \in (0,1]$, then in (3) the value δ must be replaced with δ^n,

$$\begin{aligned} &p\left(x_{(k)} - \delta \leq x^* < x_{(k)}\right) \leq \\ &\leq K\delta^{\theta}(n-k+1). \end{aligned} \qquad (6)$$

Remark 2. Under the conditions of the above lemma it is not difficult to show that the following inequality holds,

$$\begin{aligned} &p\left(x_{(k)} - \delta \leq x^* < x_{(k)}\right) \leq \\ &\leq 2K(n-k+1)\delta. \end{aligned} \qquad \textbf{(7)}$$

Theorem. *If conditions of the lemma are satisfied and the order statistics* $\tilde{x}_k = P_{\alpha}\left(x_{(i)}\right)$, $k \leq i$, *of the empirical sample* $\tilde{x} = P_{\alpha}(x) = \left(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\right)$ *is an atom with repetition factor* $t\left(\tilde{x}_k\right)$, *then the following inequality holds*

$$\gamma\left(\tilde{x}_{(k)}\right) + \frac{1}{n+1} - 2K(n-1+\lambda)\delta \leq$$

$$\leq p\left(\tilde{x}^* \in \left[\tilde{x}_{(k)}, \tilde{x}_{(k+1)}\right)\right) \leq$$

$$\leq \gamma\left(\tilde{x}_{(k)}\right) + \frac{1}{n+1} + K\delta(n-i+1),$$

where $\tilde{x}^* = P_{\alpha}\left(x^*\right)$, $\tilde{x}_{(k)} = P_{\alpha}\left(x_{(i)}\right)$,

$$\gamma\left(\tilde{x}_{(k)}\right) = \frac{t\left(\tilde{x}_{(k)}\right) - 1}{n+1}, 1 \leq k \leq m,$$

$\lambda = t\left(\tilde{x}_k\right) - 1$, *and* $\delta = 10^{-\alpha}$ is the *rounding error.*

Proof. Let $t(\tilde{x}_k) = \lambda + 1$, $\lambda = 0,1,\ldots$, and

$$P_\alpha(x_{(i)}) = \tilde{x}_{(k)},\ P_\alpha(x_{(i+1)}) = \tilde{x}_{(k)},\ \ldots,$$

$$P_\alpha(x_{(i+\lambda)}) = \tilde{x}_{(k)},\ P_\alpha(x_{(i+\lambda+1)}) = \tilde{x}_{(k+1)}.$$

Consider the following random events:

$$A = \{x^* \in [x_{(i)}, x_{(i+\lambda+1)})\},$$

$$\bar{A} = \{x^* \in [x_{(i)}, x_{(i+\lambda+1)}]\},$$

$$B = \{\tilde{x}^* \in [\tilde{x}_{(k)}, \tilde{x}_{(k+1)})\},$$

$$\bar{B} = \{\tilde{x}^* \in [\tilde{x}_{(k)}, \tilde{x}_{(k+1)}]\},$$

$$\mathfrak{A} = \{x \in [\tilde{x}_{(k)}, x_{(i)})\},$$

$$\tilde{\mathfrak{B}} = \{\tilde{x}^* = \tilde{x}_{(k+1)}\},$$

$$\mathfrak{B} = \{x \in (x_{(i+\lambda+1)} - \delta, x_{(i+\lambda+1)} + \delta)\}.$$

If $x^* \in \bar{A}$, then $x_{(i)} \le x^* \le x_{(i+\lambda+1)}$. Therefore, $\tilde{x}_{(k)} \le \tilde{x}^* \le \tilde{x}_{(k+1)}$. Thus, $\tilde{x}^* \in \bar{B}$. It means that the event $\bar{A}$ implies the event $\bar{B}$. Therefore,

$$p(\bar{A}) = p(A) \le p(\bar{B}),$$

$$\gamma(\tilde{x}_{(k)}) + \frac{1}{n+1} \le p(\bar{B}), \tag{8}$$

as far as

$$A = \{x^* \in [x_{(i)}, x_{(i+1)})\} \cup \ldots$$

$$\cup \{x^* \in [x_{(i+\lambda-1)}, x_{(i+\lambda)})\} \cup \text{ i}$$

$$\cup \{x^* \in [x_{(i+\lambda)}, x_{(i+\lambda+1)})\}$$

$$p(A) = \frac{\lambda}{n+1} + \frac{1}{n+1}.$$

On the other hand, if $\tilde{x}^* \in \bar{B}$, then $\tilde{x}_{(k)} \le \tilde{x}^* \le \tilde{x}_{(k+1)}$. This implies that $x^* \in \mathfrak{A} \cup [x_{(i)}, \tilde{x}_{(k+1}]$. So, $x^* \in \mathfrak{A} \cup \bar{A}$. Thus,

$$\begin{aligned} p(\bar{B}) &\le p(\mathfrak{A}) + p(\bar{A}) = \\ &= p(\mathfrak{A}) + p(A). \end{aligned} \tag{9}$$

It is easy to see that from the condition $x \in \mathfrak{A}$ it follows that $x \in [x_{(i)} - \delta, x_{(i)})$, where $\delta = 10^{-\alpha}$ is a rounding error. By the above lemma,

$$p(\mathfrak{A}) \le p(x_{(i)} - \delta \le x^* \le x_{(i)}) \le$$

$$\le K\delta(n - i + 1),$$

and it follows that

$$p(\bar{B}) \le \gamma(\tilde{x}_k) + \frac{1}{n+1} + K\delta(n - i + 1).$$

Clearly,

$$p(\mathfrak{B}) = p(\bar{B})\, p\{\tilde{x}^* = \tilde{x}_{(k+1)}\} \text{ and}$$

$$p\{\tilde{x}^* = \tilde{x}_{(k+1)}\} \le p(\mathfrak{B}) \le 2K(n - i - \lambda)\delta.$$

Thus,

$$\gamma(\tilde{x}_{(k)}) + \frac{1}{n+1} - 2K(n - 1 + \lambda)\delta \le$$

$$\le p(B) \le \gamma(\tilde{x}_{(k)}) + \frac{1}{n+1} + K\delta(n - i + 1),$$

since $p(B) \le p(\bar{B})$. This completes the prove of the theorem.

Corollary.

$$\begin{aligned} &p(\tilde{x}^* \in [\tilde{x}_{(k)}, \tilde{x}_{(k+1)})) \approx \\ &\approx \gamma(\tilde{x}_{(k)}) + \frac{1}{n+1} \end{aligned} \tag{10}$$

with precision up to the rounding error.

The estimate (10) implies that the probability $p(\tilde{x}^* \in [\tilde{x}_{(i)}, \tilde{x}_{(j)}))$, $i < j, 1 \le i, j \le m$ may be calculated by the formula

$$\begin{aligned} &p_{ij} = p(A_{ij}) = p(\tilde{x}^* \in [\tilde{x}_{(i)}, \tilde{x}_{(j)})) \approx \\ &\approx \gamma_i + \gamma_{i+1} + \ldots + \gamma_{j-1} + \frac{j-i}{n+1}, \end{aligned} \tag{11}$$

where $\gamma_l = \gamma(\tilde{x}_{(l)})$, $A_{ij} = \{\tilde{x}^* \in [\tilde{x}_{(i)}, \tilde{x}_{(j)}]\}$.

Note that when the sample value $\tilde{x}_{(l)}, i \le l \le j-1$ is not an atom , then $\gamma_l = 0$. Consequently, if the sample does not contain any atoms, then the formula (10) transforms into the well-known formula

$$p_{ij} = p(A_{ij}) = p(\tilde{x}^* \in [\tilde{x}_{(i)}, \tilde{x}_{(j)})) =$$

$$= \frac{j-i}{n+1}.$$

Denote by H the hypothesis that the continuous distribution functions $F_G(u)$ and $F_{G'}(u)$, of the respective general populations G and G', are equivalent. Suppose that $x = (x_1, \ldots, x_n) \in G$, $x' = (x'_1, \ldots, x'_m) \in G'$ and let $x_{(1)} \le \ldots \le x_{(n)}$, $x'_{(1)} \le \ldots \le x'_{(m)}$ be their variational series. Assume that

$F_G(u) = F_{G'}(u)$ and denote by $A_{ij}^{(k)}$, $k = 1, 2, ..., m$, the random event that x_k' belongs to the interval $\left(x_{(i)}, x_{(j)}\right)$, i.e. $A_{ij}^{(k)} = \left\{x_k' \in \left(x_{(i)}, x_{(j)}\right)\right\}$. If $F_G \equiv F_{G'}$ (i.e. $G = G'$) the probability of this event is calculated by formula (11). Let

$$p_{ij}^{(1)} = \frac{h_{ij}^{(n)} m + g^2/2 - g\sqrt{h_{ij}^{(n)}(1-h)m + g^2/4}}{m + g^2},$$
$$p_{ij}^{(2)} = \frac{h_{ij}^{(n)} m + g^2/2 + g\sqrt{h_{ij}^{(n)}(1-h)m + g^2/4}}{m + g^2}, \quad (12)$$

where $h_{ij}^{(n)}$ is the frequency of the event $A_{ij}^{(n)}$ in m trials. The value g determines the significance level of the confidence interval $I_{ij}^{(n,m)} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$. By the 3σ-rule when g =3, the significance level of this interval does not exceed 0.05.

Denote by N the number of all confidence intervals $I_{ij}^{(n,m)} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$, $N = n(n-1)/2$ and by L the number of the intervals $I_{ij}^{(n,m)}$ that contains probabilities $p_{ij}^{(n)}$. Let $h^{(n,m)} = \rho\left(F^*, F^{*\prime}\right) = \rho\left(x, x'\right) = \frac{L}{N}$.

Since $h^{(n,m)}$ is the frequency of the random event $B = \left\{p_{ij}^{(n)} \in I_{ij}^{(n,m)}\right\}$ with probability $p(B) = 1 - \beta$, substituting in equation (11) $h_{ij}^{(n,m)} = h^{(n)}$, $m = N$ and $g = 3$ we obtain the confidence interval $I^{(n,m)} = \left(p^{(1)}, p^{(2)}\right)$ for the probability $p(B)$. We shall call the statistics $h^{(n)}$ *modified p-statistics*. It is the proximity measure $\rho\left(x, x'\right)$ between x and x'.

3. Proximity measure between discrete samples. Let samples $x = \left(x_1, x_2, ..., x_n\right)$ and $y = \left(y_1, y_2, ..., y_m\right)$ be obtained by simple random sampling from discrete general populations $G_x = \left\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_N\right\}$ and $G_y = \left\{\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_M\right\}$ with distribution functions F_x and F_y.

Now, let us assume that all values of x and y are exact and let the null hypothesis be

$$H : F_x = F_y$$

and the alternative hypothesis be

$$H' : F_x \neq F_y.$$

Denote by A_i^x the random event that some sample value of x is equal to $\tilde{x}_i$, and by h_i^x the frequency of this event in x. Similarly, let B_i^y be the random event that some sample value is equal to $\tilde{x}_i$ and denote by h_i^y the frequency of this event in y. Consider these events as results of a series of independent random experiments that create the Bernoulli schemes $\left\{E_i^{(1)}\right\}_{i=1}^n$ and $\left\{E_i^{(2)}\right\}_{i=1}^m$, respectively. Using some of the results from [2-5], let us construct the confidence interval for h_i^y

$$I_i^x = \left(h_i^x - 2\tilde{\delta}_i, h_i^x + 2\tilde{\delta}_i\right),$$

where

$$\tilde{\delta}_i = \sqrt{\frac{h_i^x\left(1 - h_i^x\right)}{n}} + \sqrt{\frac{h_i^y\left(1 - h_i^y\right)}{m}}.$$

This interval contains the frequency h_i^y with probability that exceeds 0.95. Thus, the significance level of this interval, i.e. the value $p\left(h_i^y \notin I_i^x\right)$, does not exceed 0.05.

Let us introduce the random variable

$$\chi_i = \begin{cases} 1, if\ h_i^y \in I_i^x, \\ 0, if\ h_i^y \notin I_i^x, i = 1, 2, ..., k. \end{cases}$$

We shall define the proximity measure between x and y by the following statistics

$$\rho\left(x, y\right) = \frac{1}{k}\sum_{i=1}^{k} \chi_i.$$

If the general populations G_x and G_y are equivalent, then the proximity measure

$\rho(x, y)$ with probability greater than 0.95 is greater than 0.95, i.e.

$$p\left(\rho(x, y) \geq 0.95\right) \geq 0.95 .$$

Thus, the significance test for the test of hypothesis H about the equivalence of G_x and G_y may be formulated as follows:

1. If $\rho(x, y) \geq 0.95$ then x and y do not contradict H.
2. If $\rho(x, y) < 0.95$ then H is rejected.

By the central limit theorem the statistics

$$\rho(x, y) = \frac{1}{k} \sum_{i=1}^{k} \chi_i$$

has asymptotic normal distribution and the 2s-rule is valid in that case. Hence, the significance level of this test for the Bernoulli schemes $\left\{E_i^{(1)}\right\}_{i=1}^{n}$ and $\left\{E_i^{(2)}\right\}_{i=1}^{m}$ does not exceed 0.05.

Remark. The constructed proximity measure is not symmetric. To obtain a symmetric proximity measure we must swup x and y, calculate $\rho(y, x)$ and compute the value

$$\rho_{xy} = \frac{\rho(x, y) + \rho(y, x)}{2} .$$

4. Reference

1. Van der Waerdeb, B. L. Mathematische Statistik. Springer-Verlag, 1957.
2. Klyushin D.A. and Petunin Yu.I. Nonparametric population equivalence test based on proximity measure between samples //Ukr. Math. J. 2003. 55, № 2. p.147–163.
3. Andrushkiw R.I, Klyushin D.A., Petunin Yu.I., Savkiana M.Yu. The “exact” confidence limits for unknown probability in Bernoulli models // Proceedings 27th Int. Conf. Information Technology Interfaces ITI 2005, Cavtat – 2005. – P.175-179.
4. Madreimov I. and Petunin Yu.I. Characterization of the uniform distribution with the help of the order statistics // Theory Probab. and Math. Statist. - 1982. - 27. - P.96-102.
5. Klyushin D.A. and Petunin Yu.I. Statistical test for comparing two probabilities //Bull. of Kiev Univ. Cybernetics, 2005. № 6. p.35-40.