Rank as Proxy for the Observation in Statistical Procedures

F.P.TARASENKO AND V.P. SHULENIN
National Research Tomsk State University, Russia

Abstract

The properties of rank tests are discussed and it is shown that besides computational convenience, in many cases they have advantages over their counterparts on observations.

Keywords: Statistical procedures, effectiveness and efficiency of procedures, rank tests, statistical properties of ranks.

Introduction

Ranks often are preferred to actual observation values in processing experimental data. There are a few good reasons for that:

- Ranks are pure whole numbers and, hence, are very convenient to calculate. In contrast to this, observations often are continuous values that need rounding (with unpredictable consequences), and registered in various measuring scales (with each scale having different set of allowed operations over its values).
- Ranks are related to observations and, hence, contain some of the same (sought by observer) information as well as observations themselves.
- Relation between the sample value and its rank becomes even stronger with growth of a sample size; this promises the good asymptotic properties to procedures based on ranks.
- Last but not least: some distribution-free properties of ranks insure robustness to the rank procedures, – much appreciated property in statistical practice.

Here follows a brief survey of old and a few new results on these issues.

1 Basic Distributions

Let $\vec{X} = (X_1, ..., X_n)$ be a sample from p.d.f. $F_X(x)$ with a density $f_X(x)$, $x \in R^1$. Let, then, $\vec{X}_{(\cdot)} = (X_{(1)}, ..., X_{(n)})$ be the ordered statistics, and $\vec{R} = (R_1, ..., R_n)$ be a vector of ranks for the sample$\vec{X} = (X_1, ..., X_n)$. Between the sample $\vec{X}$ and the pair $\left\{\vec{X}_{(\cdot)}, \vec{R}\right\}$ there exists mutual one-to-one correspondence, which means that the information contained in observations $\vec{X} = (X_1, ..., X_n)$ maybe split into two parts. One part belongs to order statistics $\vec{X}_{(\cdot)} = (X_{(1)}, ..., X_{(n)})$, the other – to ranks $\vec{R} = (R_1, ..., R_n)$. Therefore, a seeking the same aim statistical procedures

may be built either on raw observations $\vec{X} = (X_1, ..., X_n)$, or on order statistics $\vec{X}_{(\cdot)} = (X_{(1)}, ..., X_{(n)})$, or on ranks $\vec{R} = (R_1, ..., R_n)$.

The vector random variable of a "mixed" type (i.e. consisting of discrete and continuous components [1]), which our pair $\left\{\vec{X}_{(\cdot)}, \vec{R}\right\}$ belongs to, is characterized by corresponding probability distributions:

C.d.f. for i.i.d.r.v. $\vec{X} = (X_1, ..., X_n)$ is equal to

$$F_{X_1,...,X_n}(x_1, ..., x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n) = \prod_{i=1}^{n} F_X(x_i). \tag{1}$$

C.d.f. for r-th order statistics ($1 \leq r \leq n$, $x \in R^1$) is

$$F_{X_{(r)}}(x) = P\left\{X_{(r)} \leq x\right\} = \sum_{i=r}^{n} C_n^i F_X^i(x)(1 - F_X(x))^{n-i} = I_{F(x)}(r, n - r + 1), \tag{2}$$

where $I_p(n\ ,\ m)$ is the incomplete beta-function tabulated in [2]. Corresponding density is

$$f_{X_{(r)}}(x) = nC_{n-1}^{r-1} F_X^{r-1}(x)(1 - F_X(x))^{n-r} f_X(x). \tag{3}$$

The joined p.d.f. of random vector $\vec{X}_{(\cdot)} = (X_{(1)}, ..., X_{(n)})$ is

$$f_{\vec{X}_{(\cdot)}}(x_{(1)}, ..., x_{(n)}) = \begin{cases} n! f_{\vec{X}}(x_{(1)}, ..., x_{(n)}) = n! \prod\limits_{i=1}^{n} f_X(x_{(i)})\ , x_{(1)} < ... < x_{(n)} \\ 0\ ,\ \textit{otherwise} \end{cases} \tag{4}$$

In case of symmetrical (invariant to permutations of arguments) distribution of X, order statistics and rank vector are independent:

$$f_{\vec{X}_{(\cdot)}\vec{R}}(x_{(1)}, ..., x_{(n)}; r_1, ..., r_n) = f_{X_{(1)},...,X_{(n)}}(x_{(1)}, ..., x_{(n)}) \cdot P\left\{\vec{R} = \vec{r}\right\} \tag{5}$$

and their conditional and unconditional distributions coincide [3].

If the distribution $g_{X_1,...,X_n}(x_1, ..., x_n)$ is non-invariant to permutations, the famous Hoeffding's Theorem [4] holds:

$$P_g\{\vec{R} = \vec{r}\} = P_g\{R_1 = r_1, ..., R_n = r_n\} = \frac{1}{n!} M\left\{\frac{g_{\vec{X}}(X_{(r_1)}, ..., X_{(r_n)})}{f_{\vec{X}}(X_{(r_1)}, ..., X_{(r_n)})}\right\} \tag{6}$$

which in case of independence of variables takes the form of

$$P_g\{\vec{R} = \vec{r}\} = \frac{1}{n!} M\left\{\prod_{i=1}^{n} \frac{g_{X_i}(X_{(r_i)})}{f_X(X_{(r_i)})}\right\} \tag{7}$$

The joint d.f. $F_{R_iX_i}(x, y)$ of random variables R_i and X_i is [5]

$$F_{R_iX_i}(x, y) = n^{-1} \sum_{j=1}^{n} C(x - j) \cdot \int_{-\infty}^{y} f_{X_i|R_i=j}(x\,|j)dx =$$

$$n^{-1} \sum_{j=1}^{n} C(x - j) \cdot F_{X_{(j)}}(y), 1 \leq i, j \leq n \tag{8}$$

and a formal expression for joint density of the mixed type random variable (X, Y) is

$$f_{XY}(x,y) = \sum_{i-1}^{n} p_X(\tilde{x}_i) f_{Y|X}(y|\tilde{x}_i)\delta(x-\tilde{x}_i). \tag{9}$$

2 Some Characteristics of Independence between Observations and Their Ranks

Suitableness of ranks for coming out as a proxy of the sample measurements in statistical processing of experimental data, clearly depends on how tight is the connection between them. Most general presentation of interdependence between random variables is given by their joint distribution function (8). Its one-sided presentations are made by conditional distributions of each of them conditioned by value of the other one. Such conditional distributions may be obtained by corresponding integration of d.f. (8).

But there are several particular indicators characterizing different aspects of the statistical connectedness. Let us describe some of these quantitative indices for observations and their ranks.

2.1 Regression

The regression function determines the relationship between a random variable and corresponding values of dependent value. If both regression lines coincide, it means that the relationship between the two variables is strictly functional. The more they differ, the weaker is the relationship. In case of independency the lines are orthogonal to each other.

Let us denote a regression of the observation X_i of its rank R_i as $M(X_i|R_i = j)$, $1 \le i, j \le n$, and regression of the rank R_i of X_i as $M(R_i = j\,|X_i = x)$, $1 \le i, j \le n$, $x \in R^1$. It can be shown [5] that

$$M(R_i = j\,|X_i = x) = 1 + (n-1)F_X(x), x \in R^1, \tag{10}$$

$$M(X_i\,|R_i = j) = M(X_{(j)}), 1 \le i, j \le n. \tag{11}$$

Quantitative and qualitative analyses of these lines behavior for different distributions show [5] that the lines are crossing under a certain angle which is monotonously decreases with sample size increasing. It means that interdependence between rank and observation becomes only stronger under enlarging n.

2.2 Correlation

The correlation coefficient is a measure of connexion, which is very popular among data analysis practitioners. Its calculation for observations and ranks gives a re-

sult [5]:

$$\rho_{X_iR_i}(F) = \frac{\sqrt{3}}{2}\left(\frac{n-1}{n+1}\right)^{1/2}\frac{\Delta(F)}{S(F)}, \forall i \in (1,...,n), \tag{12}$$

where $\Delta(F)$ is the Geeny's average difference defined as

$$\Delta(F) = \int_{-\infty}^{+\infty}\int |x-y|dF(x)dF(y) \tag{13}$$

and $S(F)$ is standard deviation defined as

$$S(F) = \left(\frac{1}{2}\int_{-\infty}^{+\infty}\int (x-y)^2 dF(x)dF(y)\right)^{1/2} \tag{14}$$

It turns out that correlation between observation and its rank is always positive, equal for any observation in a sample, fast approaches, with growing n, to a value typical for the length of tails of the distribution. Here are values of $\rho_{XR}(F)$ for some distributions:

$F(x)$	Uniform	Gaussian	Logistic	Laplasian
$\rho_{XR}(F)$	1,00	0,98	0,95	0,92

The longer tails of a distribution are, the less correlated are ranks and observations. This explains, in a way, difference between effectiveness of the same rank procedure being applied to data from different distributions.

2.3 Information

Various "quantities of information" are used for estimating degree of connexion tightness. In our case of considering ties inside a pair (X_i, R_i), the Shannon's quantity of information

$$I(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{XY}(x,y)\ln\left[\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right]dxdy \tag{15}$$

after cumbersome calculations, appeared to be

$$I(X,Y) = \ln n - \left(\sum_{k=1}^{n-1}\ln k + \frac{n-1}{2} - \frac{2}{n}\sum_{k=1}^{n-1} k\ln k\right), \tag{16}$$

or, asymptotically, with accuracy of $ARE_F(U:t) = 3$, is

$$I(X,R) = \ln\sqrt{ne/2\pi}. \tag{17}$$

So, quantity of information in ranks about observations does not depend neither on index i of the observation, nor on its d.f. $W_1, ..., W_k$, and increases, together with n, with velocity $\ln\sqrt{n}$. This ensures, that qualities of the rank statistical procedures will asymptotically approximate merits of procedures based on observations themselves.

3 On Some Advantages of Ranks over Observations

It was already mentioned that ranks have attracted interest from statisticians and data analysts due to their content (they share the information with observations) and to their form (they are integers, which are very convenient to work with). But it does not mean that the straightforward replacement of observations by their ranks in a statistical procedure will bring a desired effect. First, observations and their ranks usually belong to different measuring scales, with different permissible operations for their processing. This restricts usage of direct similarity of procedures to the case of their containing equivalent permissible operations only. Second, ranks of sample values contain the same kind of information as the values themselves if only this information is connected with own size of each value (when large-sized value receives higher rank). But if the information of interest is about other relations between observations, then another, the appropriate way of ordering values is required to map the information onto ranks. And the third, last but not least: the algorithms (sequences of operations) of statistical processing of data depend on a'priori knowledge of stochastic nature of the data. This is why the same sample must be treated much differently under conditions of parametric, non-parametric, and robust statistics. And here again an important role belongs to proper way of put observations in order to preserve useful information on ranks. But the most surprising and admiring feature of ranks manifests itself in complicated circumstances of robust statistics: rank test could be more effective than its counterpart based on observations.

Let us discuss briefly the abovementioned peculiarities of ranks and give some illustrative examples.

3.1 Ordering that transfers target information from observations onto ranks

Usefulness of ranks as substitutes to observations is primarily based on their attachment to the values of observations. But sometimes a statistical procedure is designed to extract from the sample such information that is indirectly defined by the values of observations but directly by their relevancy to other random events. In such a case, neither the sample alone, nor its rank vector are valid for achieving the purpose of data processing.

Typical example is homogeneity tests. The purpose is to reveal the identity or distinction between two distributions, judging by a comparison of the samples taken from them. The test is made by combining the two samples into one, and detecting a degree of their overlapping. If distributions are different then observations from one sample will dominate in number over another one in those regions where their probability is higher. For instance, if distributions are shifted (differ in location parameter) then observations from one of them will overwhelm the other in number at one side of the whole range of values; if distributions differ in scale parameter, then the observations from the wider one will outnumber those from narrower at both far ends of the range. The same will happen to the ranks of observations, if ordering

was made on the whole joined sample but with retained information of belonging observations to their distributions.

3.2 Comparison of rank tests with their counterparts based on observations

The general theory of rank tests is presented in books by Lehman [6], Hayek and Shidak [3], Pury and Sen [4], Hettsmanspreger [7]. Here we give only a few examples revealing merits of rank tests in comparison with analogous tests based on observations.

The notion of the Pitman asymptotic relative efficiency (ARE) is widely used for comparison of two tests, T_n and S_n. $ARE_F(T_n : S_n)$ characterizes the ratio of sample sizes n_1 and n_2 under which T_n and S_n with equal levels of significance ensure equal ARE against the same sequence of contigual alternatives converging to zero hypothesis.

For the Wilcoxon sign rank test S^+ and Student's t-test

$$ARE_F\,(S^+ : T(\vec{X})) = 12\sigma^2 \left[\int_{-\infty}^{\infty} f^2(x)dx\right]^2. \tag{18}$$

In Table 1 the numerical values of $ARE_F\,(S^+ : T(\vec{X}))$ are presented for some symmetric distributions.

Table 1

Distribution $F(x)$	$ARE_F(S^+ : T(\vec{X}))$
Uniform	1
Gaussian	$3/\pi = 0,955$
Logistic	$\pi^2/9 = 1,097$
Double exponential	1,5

For the Wilcoxon sign rank test S^+ and the sign test S

$$ARE_F\,(S : S^+) =_F^2 (S)/_F^2(S^+)) = 4\sigma^2 f^2(0)/3 \cdot \left[\int f^2(x)dx\right]^2. \tag{19}$$

Its numerical values are given in Table 2.

Calculations of ARE for many other pairs of tests were made (e.g. in [7 – 12]). Some general conclusions follow from their consideration:

– In most cases ARE does not depend on scale parameter and is connected to the distributions' family type only.

Table 2

Distribution $F(x)$	$ARE_F(S : S^+)$
Uniform	$1/3$
Gaussian	$2/\pi = 0,637$
Logistic	$\pi^2/12 = 0,822$
Double exponential	2

– AREs may take various values not limited from above, but have non-zero lower limits. For instance, $ARE_F(U : t) \geq 0,864$, which means that in two-sampled problem of shift we may loose in efficiency not more than 13,6% using Wilcoxon's test instead of Student's one. Under Gaussian distribution the loss is 5% only. The most favorable distribution ($ARE_F(U : t) = 3$) is gamma-distribution with $p = 1$. So, under these circumstances the Wilcoxon test is always preferable among other tests.

Robust statistics is an approach to designing statistical procedures at an intermediate (between parametric and non-parametric) level of a priori knowledge about stochastic nature of observations. Underlining them distribution is considered as known approximately: it belongs to a "supermodel", a certain vicinity of some parametric function. The procedures are designed that remain effective ("robust") until actual distribution lies inside the vicinity; there are among those the rank procedures, too. And they demonstrate certain advantages.

For example, efficiency of H-test of Kruskall-Wallis against its Gaussian competitor, Fisher's F-test is [7]

$$ARE_F(H : F) = 12\sigma_f^2 \left[\int_0^1 f(F^{-1}(u)du) \right]^2, \tag{20}$$

and this formula is valid for several other counterparts of tests [8]. Numerical values of it for Gaussian model with a scale obstruction

$$F \in \Im_{\varepsilon,\tau}(\Phi) = \{F : F_{\varepsilon,\tau}(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/\tau)\}, 0 \leq \varepsilon < 1/2, \tau \geq 1$$

are given in Table 3.

Table 3

	ε	0.00	0.01	0.03	0.05	0.08	0.10	0.15	0.20
$ARE_{F_{\varepsilon,\tau}}(H : F)$	$\tau = 3$	0.955	1.009	1.108	1.196	1.309	1.373	1.497	1.575
	$\tau = 5$	0.955	1.150	1.505	1.814	2.201	2.412	2.795	3.006
	$\tau = 7$	0.955	1.369	2.115	2.759	3.553	3.977	4.724	5.099

It is seen that H-test looses in efficiency only 5% to the optimal F-test of Fisher in Gaussian case, but much overwhelms it under deviations from normality.

References

[1] Wilks S. Mathematical statistics. – M.: Nauka, 1967. – 632 p.

[2] K. Pearson. Tables of the Incomplete Beta-Function. Cambridge, England: The University Press, 1934.

[3] Gaek Ya., Shidak Z. Theory of rank tests (in Russian). – M.: Nauka, 1971. – 376 p.

[4] Puri M.L., Sen P.K. Nonparametric methods in Multivariate Analysis. John Wiley, N.Y., 1970, 440 p.

[5] Shulenin V.P., Tarasenko F.P. Regression functions for observations and their ranks (In Russian) // Tomsk State University Journal of Control and Computer Science 2003, N. 280, pp. 213 – 216.

[6] Lehmann E. L. Nonparametric: Statistical Methods Based on Ranks. Holden – Day. San Francisco, 1975. - 326 p.

[7] Hettmansperger T.P. Statistical inference based on ranks. John Wiley and Sons, New York. 1984. 323 p.

[8] Kendall M, Stuart A. Statistical inference and communication (in Russian). – M.: Nauka, 1973. – 899 p.

[9] Gibbons J.D. Nonparametric Statistical Inference. New York, McGraw-Hill, 1971.

[10] Randles R.H. and Wolfe D.A. Introduction to the theory of nonparametric statistics. New York: Wiley. 1979.

[11] Tarasenko F.P. Nonparametric statistics (in Russian). Tomsk. TSU Publisher, 1976. – 292p.

[12] Tarasenko F.P., Shulenin V.P. On statistical relation between observations and its rank (in Russian) // Trudy SFTI pri Tomskom universitete. 1971, Vol. 60 , p. 220 – 228.

[13] Hollander M., Wolfe D. A. Nonparametric Statistical Methods: John Wiley and Sons, New York. 1973. 503 p.