

Two-Sample Median Test for Vague Data

Przemysław Grzegorzewski

Systems Research Institute, Polish Academy of Sciences

Newelska 6, 01-447 Warsaw, Poland

e-mail: pgrzeg@ibspan.waw.pl

Abstract

Classical statistical tests may be sensitive to violations of the fundamental model assumptions inherent in the derivation and construction of these tests. It is obvious that such violations are much more probable in the presence of vague data. Thus nonparametric tests seem to be promising statistical tools. A generalization of the median test for the two-sample problem with vague data is suggested.

Keywords: Fuzzy sets, Median test, Necessity, Vague data.

1 Introduction

Most of statistical procedures are based on fairly specific assumptions regarding the underlying population distribution, like normality, exponentiality, etc. However, quite often such stringent assumptions on distributions are not satisfied. In such a case distribution-free methods – also called nonparametric – are very useful. If the data are vague the difficulties concerning the distribution even increase. In fact we still do not have satisfactory goodness-of-fit techniques for imprecise data. Therefore, it seems that there is a great need for nonparametric statistic for vague data.

The present paper is devoted to hypotheses testing with fuzzy data. This general problem has been considered by many authors (for the review papers on testing hypotheses in fuzzy environment we refer the reader to [10] and [11]). Some nonparametric test for vague data were also proposed (see [5], [9]). Below we suggest another very

useful distribution-free statistical test for comparing two samples. This test is a generalization of the well-known two-sample median test for fuzzy data. The paper is organized as follows: in Sec. 2 we recall the classical two-sample median test for crisp data. In Sec. 3 we introduce basic notation used for modelling vague data. Then we propose how to modify the classical median test for vague data (Sec. 4). In our approach we utilize the necessity index of strict dominance, suggested by Dubois and Prade [3]. Using this tool we obtain a fuzzy test showing the grade of necessity for rejecting the underlying hypothesis.

2 Two-sample median test

Suppose our data consist of two mutually independent random samples V_1, \dots, V_{n_1} and W_1, \dots, W_{n_2} from populations with continuous cumulative distribution functions F_V and F_W , respectively. Usually the hypothesis of interest in the two-sample problem is that the two populations have the same distribution, i.e.

$$H_0 : F_V(z) = F_W(z) \quad \text{for all } z. \quad (1)$$

against one-sided alternative stating that V is stochastically larger than W , i.e.

$$H_1 : \begin{aligned} F_V(z) &\leq F_W(z) \quad \text{for all } z, \\ F_V(z) &< F_W(z) \quad \text{for some } z. \end{aligned} \quad (2)$$

As a particular case of we may consider the difference of location alternative, i.e.

$$H_2 : F_V(z) = F_W(z - \theta) \quad \text{for all } z, \quad (3)$$

where $\theta \neq 0$. Under the location model V is stochastically larger than W if and only if $\theta > 0$.

The available statistical literature on the two-sample problem is quite extensive. For a general case when no assumptions on underlying distributions are made several test have been proposed, like the Wald-Wolfowitz runs test, the Kolmogorov-Smirnov two-sample test, the Mann-Whitney-Wilcoxon test, etc. Among them we also find the so-called median test, attributed to Brown and Mood. The idea of the median test is as follows: let $S_{1:n} \leq \dots \leq S_{n:n}$ denote order statistics from the combined samples of V_1, \dots, V_{n_1} and W_1, \dots, W_{n_2} , where $n = n_1 + n_2$. Moreover, let M_S denote a sample median obtained for these combined samples, i.e.

$$M_S = \begin{cases} S_{\frac{n+1}{2}:n} & \text{if } n \text{ is odd,} \\ \frac{1}{2} (S_{\frac{n}{2}:n} + S_{\frac{n}{2}+1:n}) & \text{if } n \text{ is even.} \end{cases} \quad (4)$$

If the two populations have the same distribution we would expect the sample observations from each population to be similarly spread throughout the combined order statistics. Thus the median test uses as a test statistic τ the random variable that counts the number of observations from the sample V_1, \dots, V_{n_1} which exceed the sample median M_S of the combined samples. If the null hypothesis H_0 holds then τ has a hypergeometric distribution given by

$$P(\tau = t) = \frac{\binom{n_1}{t} \binom{n_2}{m-t}}{\binom{n}{m}}, \quad (5)$$

where

$$m = \begin{cases} \frac{n-1}{2} & \text{if } n \text{ is odd,} \\ \frac{n}{2} & \text{if } n \text{ is even,} \end{cases} \quad (6)$$

and $t = 0, 1, \dots, \min\{n_1, m\}$. Values of τ significantly bigger than m will make us reject H_0 in favor of H_1 . The critical values are obtained either directly from the distribution (5) or – if sample sizes n_1 and n_2 are large enough – a normal approximation can be used. Moreover, if sample sizes are large we may use the chi-square approximation for the two-sided alternative.

3 Vague data

It may happen that a sample used for making decision consists of observations that are not necessarily crisp but may be vague as well. In order to describe the vagueness of data we use the

notion of a fuzzy number, introduced by Dubois and Prade [2]. We say that a fuzzy subset A of the real line \mathbb{R} , with the membership function $\mu_A : \mathbb{R} \rightarrow [0, 1]$, is a fuzzy number if and only if

- (a) A is normal (i.e. there exists an element x_0 such that $\mu_A(x_0) = 1$),
- (b) A is fuzzy convex (i.e. $\mu_A(\lambda x_1 + (1-\lambda)x_2) \geq \mu_A(x_1) \wedge \mu_A(x_2)$, $\forall x_1, x_2 \in \mathbb{R}$, $\forall \lambda \in [0, 1]$),
- (c) μ_A is upper semicontinuous,
- (d) $\text{supp}A$ is bounded, where $\text{supp}A = \text{cl}(\{x \in \mathbb{R} : \mu_A(x) > 0\})$ and cl is the closure operator.

A useful notion for dealing with a fuzzy number is a set of its α -cuts. The α -cut of a fuzzy number A is a nonfuzzy set defined as $A_\alpha = \{x \in \mathbb{R} : \mu_A(x) \geq \alpha\}$. A family $\{A_\alpha : \alpha \in (0, 1]\}$ is a set representation of the fuzzy number A . According to the definition of a fuzzy number it is easily seen that every α -cut of a fuzzy number is a closed interval. Hence we have $A_\alpha = [A_\alpha^L, A_\alpha^U]$, where

$$A_\alpha^L = \inf\{x \in \mathbb{R} : \mu_A(x) \geq \alpha\}, \quad (7)$$

$$A_\alpha^U = \sup\{x \in \mathbb{R} : \mu_A(x) \geq \alpha\}. \quad (8)$$

A space of all fuzzy numbers will be denoted by $\mathbb{FN}(\mathbb{R})$.

A notion of fuzzy random variable was introduced by Kwakernaak [14], [15]. Other definitions of fuzzy random variables are due to Kruse [12] or to Puri and Ralescu [16]. Our definition is similar to those of Kwakernaak and Kruse. Suppose that a random experiment is described as usual by a probability space (Ω, \mathbb{A}, P) , where Ω is a set of all possible outcomes of the experiment, \mathbb{A} is a σ -algebra of subsets of Ω (the set of all possible events) and P is a probability measure. Then mapping $X : \Omega \rightarrow \mathbb{FN}(\mathbb{R})$ is called a fuzzy random variable (f.r.v.) if it satisfies the following properties:

- (a) $\{X(\alpha, \omega) : \alpha \in [0, 1]\}$ is a set representation of $X(\omega)$ for all $\omega \in \Omega$,
- (b) for each $\alpha \in [0, 1]$ both $X_\alpha^L = X_\alpha^L(\omega) = \inf X_\alpha(\omega)$ and $X_\alpha^U = X_\alpha^U(\omega) = \sup X_\alpha(\omega)$, are usual real-valued random variables on (Ω, \mathbb{A}, P) .

Thus a f.r.v. X is considered as a perception of an unknown usual random variable $V : \Omega \rightarrow \mathbb{R}$,

called an *original* of X (if only vague data are available, it is of course impossible to show which of the possible originals is the true one). Similarly n -dimensional fuzzy random sample X_1, \dots, X_n may be treated as a fuzzy perception of the usual random sample V_1, \dots, V_n (where V_1, \dots, V_n are independent and identically distributed crisp random variables). For more information we refer the reader to [13].

Let \mathbb{V} denote a set of all possible originals of X . Then a fuzzy median of a f.r.v. X is defined as a fuzzy set M with a membership function

$$\mu_M(t) = \sup \left\{ \inf_{\omega \in \Omega} \mu_{X(\omega)}(V(\omega)) : V \in \mathbb{V}, F_V(t^-) \leq \frac{1}{2} \leq F_V(t) \right\}, \quad (9)$$

where $t \in \mathbb{R}$ and F_V denotes the c.d.f. of V .

Grzegorzewski [5] shown that a fuzzy sample median M from the fuzzy random sample X_1, \dots, X_n via a fuzzy number with α -cuts $[M_\alpha^L, M_\alpha^U]$ given by

$$M_\alpha^L = M_\alpha^L(X_1, \dots, X_n) \quad (10)$$

$$= \begin{cases} (X_\alpha^L)_{\frac{n+1}{2}:n} & \text{if } n \text{ is odd,} \\ \frac{1}{2} \left((X_\alpha^L)_{\frac{n}{2}:n} + (X_\alpha^L)_{\frac{n}{2}+1:n} \right) & \text{if } n \text{ is even,} \end{cases}$$

$$M_\alpha^U = M_\alpha^U(X_1, \dots, X_n) \quad (11)$$

$$= \begin{cases} (X_\alpha^U)_{\frac{n+1}{2}:n} & \text{if } n \text{ is odd,} \\ \frac{1}{2} \left((X_\alpha^U)_{\frac{n}{2}:n} + (X_\alpha^U)_{\frac{n}{2}+1:n} \right) & \text{if } n \text{ is even,} \end{cases}$$

where $(X_\alpha^L)_{k:n}$ denotes the k -th order statistic of a sample $(X_1)_\alpha^L, \dots, (X_n)_\alpha^L$ while $(X_\alpha^U)_{k:n}$ denotes the k -th order statistic of a sample $(X_1)_\alpha^U, \dots, (X_n)_\alpha^U$. It can be shown that the fuzzy sample median becomes a traditional crisp sample median if the observations are not vague but crisp. Moreover, the fuzzy sample median is a fuzzy-consistent estimator of the median, provided that we restrict ourselves to distributions with unique median (for more details see [5]).

4 Median test for vague data

Now suppose that we want to verify H_0 against one-sided alternative H_2 having only fuzzy ran-

dom samples X_1, \dots, X_{n_1} and Y_1, \dots, Y_{n_2} . However, when we try to apply the median test directly to fuzzy data we meet immediately a serious difficulty: the test statistic τ depends on the number of observations from the first sample bigger than the sample median M of the combined sample, which is also fuzzy. As it is known, fuzzy numbers are not linearly ordered, hence there is no such ordering system that could univocally determine which of any two fuzzy numbers is a bigger one. Therefore, we cannot say explicitly how many observations from the fuzzy sample X_1, \dots, X_{n_1} exceed fuzzy sample median M .

Many researchers have considered the problem of ranking fuzzy numbers (see, e.g. [1] or [17]). Below we apply a method for comparing fuzzy numbers via coefficient called the necessity index of strict dominance (NSD), suggested by Dubois and Prade [3]. Let us recall that for any fuzzy numbers A and B with membership functions μ_A and μ_B , respectively, we can evaluate the degree of necessity to which the relation $A > B$ is fulfilled

$$Ness(A > B) = 1 - \sup_{x,y:x \leq y} \min\{\mu_A(x), \mu_B(y)\}. \quad (12)$$

Dubois and Prade proposed also the possibility of strict dominance index and other indices. However, we decided to use NSD index because of its natural interpretation and effectiveness in solving real-life problems (including statistics – see, e.g., [8]). NSD index takes values from the interval $[0, 1]$ and provides the grade of necessity for $A > B$. Let $[A_\alpha^L, A_\alpha^U]$ and $[B_\alpha^L, B_\alpha^U]$ denote α -cuts of fuzzy numbers A and B , respectively. By (12) $Ness(A > B) = 1$ if and only if $A_\alpha^L > B_\alpha^U$ for all $\alpha \in (0, 1]$ which corresponds to situation where A could be univocally classified as "greater" than B . However, if $A_\alpha^L < B_\alpha^U$ for some $\alpha \in (0, 1]$ then the relationship between A and B is not so evident and $Ness(A > B) < 1$. Thus it corresponds to soft concept of inequality rather than the crisp one and hence it seems more adequate for handling with vague data. This is why we have also found it suitable for our statistical testing problem.

Let Z_1, \dots, Z_n denote a set of observations created by combined samples of V_1, \dots, V_{n_1} and W_1, \dots, W_{n_2} and let M denote a fuzzy sample

median computed from Z_1, \dots, Z_n according to (10)–(11). Following lemmas will be useful:

Lemma 4.1

For any fuzzy random sample Z_1, \dots, Z_n we have

$$\# \{Z_i : (Z_i)_\alpha^L > M_\alpha^U, \text{ for all } \alpha \in (0, 1]\} \leq m, \tag{13}$$

where m is given by (6) and $\#\{\cdot\}$ stands for the cardinality of given set $\{\cdot\}$.

As a conclusion we get immediately

Lemma 4.2

For any fuzzy random sample Z_1, \dots, Z_n we have

$$\# \{Z_i : Ness(Z_i > M) > 0\} \leq m. \tag{14}$$

We also get

Lemma 4.3

If $\alpha > \beta$, where $\alpha, \beta \in (0, 1]$, then for any fuzzy random sample Z_1, \dots, Z_n we have

$$\begin{aligned} \# \{Z_i : Ness(Z_i > M) \geq \alpha\} &\leq & (15) \\ &\leq \# \{Z_i : Ness(Z_i > M) \geq \beta\}. \end{aligned}$$

Now we can define a test statistic T as follows

$$T = \bigcup_{\alpha \in (0, 1]} \# \{X_i : Ness(X_i > M) \geq \alpha\} / \alpha. \tag{16}$$

By lemmas given above T is a fuzzy subset of a set $\{0, 1, \dots, \min\{n_1, m\}\}$ with a nonincreasing membership function μ_T given by

$$\begin{aligned} \mu_T(t) &= \sup \{ \alpha \in [0, 1] : \\ &\# \{X_i : Ness(X_i > M) \geq \alpha\} = t \}. \end{aligned} \tag{17}$$

For simplicity of notation let $l = \min\{n_1, m\}$. As it is seen μ_T takes at most $l + 1$ distinct values. They correspond to no more than $l + 1$ distinct numbers of observations which exceed fuzzy sample median M with different grades of necessity. Thus we may decompose our testing problem for fuzzy data into no more than $l + 1$ crisp testing problems and then appropriately aggregate the results.

More formally, we may denote T in a following way:

$$T = 0/\alpha_0 + 1/\alpha_1 + \dots + l/\alpha_l, \tag{18}$$

where $\alpha_0 \geq \alpha_1 \geq \dots \geq \alpha_l$ (by Lemma 3). Moreover, assume that there are only k distinct values among $\alpha_0, \dots, \alpha_l$ and denote then by β_1, \dots, β_k (of course, in general, $1 \leq k \leq l$). Let us also adopt a following notation:

$$t_j = \max\{i : \alpha_i = \beta_j, i = 0, 1, \dots, l\}. \tag{19}$$

Hence our test statistic (16) reduces to

$$T = t_1/\beta_1 + \dots + t_k/\beta_k, \tag{20}$$

where $t_1 < \dots < t_k$ and $\beta_1 > \dots > \beta_k$. Test statistic (20) has a nice interpretation: the grade of necessity that t_j observations from the first sample X_1, \dots, X_{n_1} exceed fuzzy sample median M is not smaller than β_j .

Suppose now that we consider a statistical test $\varphi : \Xi \rightarrow \{0, 1\}$, where Ξ is a sample space. As it is known, a classical test on significance level $\delta \in (0, 1)$ is given by

$$\varphi(s_1, \dots, s_n) = \begin{cases} 1 & \text{if } \tau(s_1, \dots, s_n) \in K_\delta, \\ 0 & \text{if } \tau(s_1, \dots, s_n) \notin K_\delta, \end{cases} \tag{21}$$

where K_δ is a critical region. In a fuzzy domain we have a fuzzy test $\phi : \mathbb{F}(\Xi) \rightarrow \mathbb{F}(\{0, 1\})$, where sample space $\mathbb{F}(\Xi)$ is now a set of all possible fuzzy outcomes of the experiment. We may decompose such a fuzzy test into a family $\{\phi_\alpha : \alpha \in (0, 1]\}$ of crisp tests corresponding to successive α -cuts, where $\phi_\alpha(Z_1, \dots, Z_n) = \varphi(\xi((Z_1)_\alpha), \dots, \xi((Z_n)_\alpha))$ and ξ is a function (compare, e.g., [6] [7]). Since in our case test statistic T assumes k values only, we may decompose the fuzzy median test into a finite family of crisp tests $\{\phi_{\beta_j} : j = 1, \dots, k\}$. In other words, we consider k crisp testing problems of verifying H_0 against H_2 assuming each time that the test statistic takes different values, namely $\tau = t_j$, where $j = 1, \dots, k$. Since $t_1 < \dots < t_k$, the following lemma holds:

Lemma 4.4

Let K_δ denote a critical region for the testing problem H_0 against H_2 on significance level $\delta \in (0, 1)$. If there exist such t^* , $t_1 \leq t^* \leq t_k$, that $\tau = t^* \in K_\delta$ then $\tau = t \in K_\delta$ for each $t > t^*$. Moreover, if t^* is the smallest value in a set $\{t_1, \dots, t_k\}$ such that $\tau = t^* \in K_\delta$ then $\tau = t \notin K_\delta$ for each $t < t^*$.

As a natural consequence of that lemma we get a following corollary

Corollary 4.1

If there exist such β^* , $\beta_1 \leq \beta^* \leq \beta_k$, that $\phi_{\beta^*} = 1$ then $\phi_\beta = 1$ for each $\beta < \beta^*$. Moreover, if β^* is the largest value in a set $\{\beta_1, \dots, \beta_k\}$ such that $\phi_{\beta^*} = 1$ then $\phi_\beta = 0$ for each $\beta > \beta^*$.

Thus finally, our fuzzy median test has a following form

$$\phi(Z_1, \dots, Z_n) = 1/\mu_\phi(1) + 0/\mu_\phi(0) \quad (22)$$

where

$$\mu_\phi(1) = \sup_{\beta \in \{\beta_1, \dots, \beta_k\}} \phi_\beta(Z_1, \dots, Z_n), \quad (23)$$

$$\mu_\phi(0) = 1 - \mu_\phi(1). \quad (24)$$

It is easily seen that our fuzzy median test, contrary to the classical crisp test, does not lead to a binary decision (acceptation or rejection of H) but to a fuzzy decision. We may get $\varphi = 1/1 + 0/0$ which indicates that we should reject H , or $\varphi = 1/0 + 0/1$ which means that there is no reason for rejecting H , but we may also get $\varphi = 1/\eta + 0/(1 - \eta)$, where $\eta \in (0, 1)$, which can be interpreted as a degree of necessity of rejection (η) or acceptation ($1 - \eta$) the null hypothesis H . Thus, in situation when η is neither 0 nor 1, a user must decide whether to reject or to accept given hypothesis actually (however value η would support his decision). The problem of defuzzification of a fuzzy test is considered in [7].

5 Conclusions

In the present paper we have proposed a two-sample fuzzy median test for vague data. This test is a natural and proper generalization of the classical test since if all the data are crisp it reduces to this classical two-sample median test. Our two-sample fuzzy median test for vague data is based on the necessity index of strict dominance. This index seems to be a very useful tool for comparing fuzzy numbers because of its clear interpretation. However, it seems that one can also construct a similar test based on the possibility index. It is also worth noting that our

approach could be applied for the generalization of other nonparametric statistical test for vague data.

References

- [1] G. Bortolan, R. Degani, "A review of some methods for ranking fuzzy subsets", Fuzzy Sets and Systems, vol. 15 (1985), 1–19.
- [2] D. Dubois, H. Prade, "Operations on fuzzy numbers", Int. J. Syst. Sci., vol. 9 (1978), 613–626.
- [3] D. Dubois, H. Prade, "Ranking fuzzy numbers in the setting of possibility theory", Inform. Sci., vol. 30 (1983), 183–224.
- [4] J.D. Gibbons, S. Chakraborti, "Nonparametric Statistical Inference", Marcel Dekker Inc., 2003.
- [5] P. Grzegorzewski, "Statistical inference about the median from vague data", Control and Cybernetics, vol. 27 (1998), 447–464.
- [6] P. Grzegorzewski, "Testing statistical hypotheses with vague data", Fuzzy Sets and Systems, vol. 112(2000) 501–510.
- [7] P. Grzegorzewski, "Fuzzy tests – defuzzification and randomization", Fuzzy Sets and Systems, vol. 118 (2001), 437–446.
- [8] P. Grzegorzewski, "Testing fuzzy hypotheses with vague data", In: Statistical Modeling, Analysis and Management of Fuzzy Data, Bertoluzza C., Gil M.A., Ralescu D. (Eds.), Springer - Physica Verlag, Heidelberg, 2002, pp. 213–225.
- [9] P. Grzegorzewski, "Distribution-free tests for vague data", In: Soft Methodology and Random Information Systems, Lopez-Diaz M., Gil M.A., Grzegorzewski P., Hryniewicz O., Lawry J. (Eds.), Springer, Heidelberg, 2004, pp. 495–502.
- [10] P. Grzegorzewski, O. Hryniewicz, "Testing hypotheses in fuzzy environment", Mathware and Soft Computing, vol. 4 (1997), 203–217.

- [11] P. Grzegorzewski, O. Hryniewicz, "Soft methods in hypotheses testing", In: *Soft Computing for Risk Evaluation and Management*, Ruan D., Kacprzyk J., Fedrizzi M. (Eds.), Springer – Physica Verlag, Heidelberg, 2001, pp. 55–72.
- [12] R. Kruse, "The strong law of large numbers for fuzzy random variables", *Inform. Sci.*, vol. 28 (1982), 233–241.
- [13] R. Kruse, K.D. Meyer, "Statistics with Vague Data", D. Riedel Publishing Company, 1987.
- [14] H. Kwakernaak, "Fuzzy random variables, part I: Definitions and theorems", *Inform. Sci.*, vol. 15 (1978), 1–15;
- [15] H. Kwakernaak, "Fuzzy random variables, part II: Algorithms and examples for the discrete case", *Inform. Sci.*, vol. 17 (1979), 253–278.
- [16] M.L. Puri, D.A. Ralescu, "Fuzzy random variables", *J. Math. Anal. Appl.*, vol. 114 (1986), 409–422.
- [17] Q. Zhu, E.S. Lee, "Comparison and ranking of fuzzy numbers", In: *Fuzzy Regression Analysis*, Kacprzyk J., Fedrizzi M. (Eds.), Omnitech Press, Warsaw and Physica-Verlag, Heidelberg, 1992, pp. 21–44.