九州大学学術情報リポジトリ
Kyushu University Institutional Repository

INFERENCE BASED ON $ P(XY) $ IN TWO SAMPLE PROBLEMS

Fujii, Yoshinori
Faculty of Education and Culture, Miyazaki University

https://doi.org/10.5109/12584

出版情報：Bulletin of informatics and cybernetics. 36, pp.137-145, 2004-12. Research Association of Statistical Sciences
バージョン：
権利関係：

INFERENCE BASED ON $P(X < Y)/P(X > Y)$ IN TWO SAMPLE PROBLEMS

by

Yoshinori Fujii

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.36

FUKUOKA, JAPAN
2004

Bulletin of Informatics and Cybernetics, Vol. 36, 2004

INFERENCE BASED ON $P(X < Y)/P(X > Y)$ IN TWO SAMPLE PROBLEMS

By

Yoshinori Fujii*

Abstract

Let X and Y be two independent random variables. It is important in practical situations to evaluate the difference between the distributions of X and Y. We focus on the inference based on $\theta = P(X < Y)/P(X > Y)$. θ is an extension of the odds ratio in 2×2 tables. An unbiased estimating function for θ is proposed by using the pairwise estimating functions. The variance of the proposed estimating funciton can be estimated unbiasedly. We present methods for stratified data and examplify them in a practical example.

Key Words and Phrases: Wilcoxon test, Mantel-Haenszel procedure, Estimating functions, Stratified analysis, Ordered categorical data.

1. Introduction

Let X and Y be two independent random variables. It is important in practical situations to evaluate the difference between the distributions of X and Y. The difference of means is famous to be one of indices to measure the difference between two distributions. The confidence interval of the difference of means under the normal assumption and t test are basic statistical methods. For nonparametric situations the shift in location is well investigated, see Lehmann (1975) and Yanagawa (1982).

In this paper we are interested in the probability $P(X < Y)$. $P(X < Y)$ is intutively interpreted since $P(X < Y) = 1/2$ if X and Y have same continuous distribution. A nonparametric estimator for $P(X < Y)$ based on the sample is given by

$$R = \int F_n(x) dG_m(x)$$

where $F_n(x), G_m(x)$ are the empirical distributions of X and Y, respectively. R is called Wilcoxon-Mann-Whitney statistic. The confidence interval based on R and the asymptotic properties of R were developed (see Reiser and Guttman,1986). In parametic situations the inference of $P(X < Y)$ are also discussed. Especially when X and Y are normal,

$$P(X < Y) = \Phi\left(\frac{\mu_Y - \mu_X}{(\sigma_X^2 \sigma_Y^2)^{1/2}}\right),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and μ_X, σ_X^2, μ_Y and σ_Y^2 are the mean and variance of the X and Y variables, respectively. Downton(1973)

* Faculty of Education and Culture, Miyazaki University, Gakuen Kibanadai Nishi 1–1 Miyazaki 889-2192 Japan. tel 81–985–58–7493 yfujii@cc.miyazaki-u.ac.jp

discussed the estimation in the normal setting. However when X and Y are not continuous, $P(X = Y)$ may be nonzero. So $P(X < Y) \neq 1/2$ enen if X and Y have same distribution. Hochberg (1981) noted that $P(X < Y) - P(X > Y)$ is prefer to $P(X < Y)$ in categorical data. Also Wellek and Hampel (1999) used

$$\theta = P(X < Y)/P(X > Y)$$

to construct equivalent test. In this paper θ is an interesting parameter. θ is an extension of the odds ratio in 2×2 tables since θ is equivalent to the odds ratio if X and Y are binary random variables with same two values. The parameter θ can be interpreted as follows. Consider a situation that we can choose one of two treatments A and B. Then it may be important which treatment has better results. The parameter θ gives us such information. Suppose X and Y are the results corresponding to the treatments A and B, $\theta > 1$ means that the choice of the treatment B tends to have better results than the choice of A. Also θ is easy to assume that homogeneity for stratified data because it is less constrained than $P(X < Y) - P(X > Y)$.

In this paper we propose an unbiased estimating function for θ based on the pairwise estimating functions. An unbiased estimator for the variance of the proposed estimator are obtained in section 2 in order to construct the confidence interval and testing the hypothesis for θ. Especially we show methods in the ordered categorical data and compare them with the methods proposed before. In section 3 summary statistics and hypothesis tests are consider in stratified data. The proposed methods are considered to be extensions of Mantel-Haensel procedures and include an extension of Wilcoxon test in stratified data. We examplify the proposed methods to apply the practical data in section 4. Some discussions are given in the last section.

2. Estimation

2.1. General situations

Let $X_1, X_2, \cdots, X_n$ be an independent sample from distribution $F(x)$ and $Y_1, Y_2, \cdots, Y_m$ be from distribution $G(x)$. In this paper $\theta = P(X < Y)/P(X > Y)$ is of interest and other structures of the distributions are treated to be nuisance. Fujii and Yangimoto(2004) showed that pairwise conditional estimating functions are useful to remove the effect of nuisance paremeters in estimating the canonical parameter in exponential families. In this setting we can remove the effects of other structures of the distributions by using pairwise estimating functions. For any pair (X_i, Y_j), we define two variables

$$S_{ij} = \begin{cases} 1 & X_i < Y_j \\ 0 & otherwise, \end{cases} \qquad T_{ij} = \begin{cases} 1 & X_i = Y_j \\ 0 & otherwise \end{cases}$$

The joint distribution of (S_{ij}, T_{ij}) is given by

$$\begin{aligned} P(S_{ij} = s, T_{ij} = t) &= P(X_i < Y_j)^s P(X_i > Y_j)^{1-s-t} P(X_i = Y_j)^t \\ &= \exp\{s \log\theta + t \log\psi - \log(1 + \theta + \psi)\}. \end{aligned}$$

where $\psi = P(X_i = Y_j)/P(X_i > Y_j)$. It belongs to two parameter exponential families. Andersen (1970) recommended the inference based on the conditional distribution given by the sufficient statistic for the nuisance parameter. In this case the conditional

probability is given by

$$P(S_{ij} = s|T_{ij} = t) = \frac{\theta^s}{(1+\theta)^{1-t}},$$

and does not depend on the nuisance paremeter ψ. We construct a log likelihood by summing up log of above pairwise conditional probabilities. It is

$$\ell = \sum_{i=1}^{n}\sum_{j=1}^{m}\{S_{ij}\log\theta - (1-T_{ij})\log(1+\theta)\}.$$

The first derivative of ℓ is

$$\frac{\partial\ell}{\partial\theta} = \frac{1}{\theta(1\theta)}\sum_{i=1}^{n}\sum_{j=1}^{m}\{S_{ij} - \theta(1 - S_{ij} - T_{ij})\}$$

and we can get the estimator $\hat{\theta}$ to solve the equation $\frac{\partial\ell}{\partial\theta} = 0$.

Here we put

$$g(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} g_{ij}(\theta) \tag{1}$$

where $g_{ij}(\theta) = S_{ij} - \theta(1 - S_{ij} - T_{ij})$. Hochberg (1981) gave an unbiased estimator for the variance of $g(1)$. We can lead an unbiased estimator in the general case.

Lemma 2.1.

$$\begin{aligned} Var[g(\theta)] &= mn\big\{(n-1)(\pi_{XYY} - 2\theta\pi_{YXY} + \theta^2\pi_{YYX}) \\ &+ (m-1)(\pi_{XXY} - 2\theta\pi_{XYX} + \theta^2\pi_{YXX}) + (\pi_{XY} + \theta^2\pi_{YX})\big\} \end{aligned} \tag{2}$$

where

$$\pi_{XY} = P(X_1 < Y_1), \qquad \pi_{YX} = P(X_1 > Y_1)$$

$$\begin{aligned} \pi_{XYY} &= P(X_1 < Y_1, X_1 < Y_2) \\ \pi_{YXY} &= P(Y_1 < X_1 < Y_2) \\ \pi_{YYX} &= P(Y_1 < X_1, Y_2 < X_1) \\ \pi_{XXY} &= P(X_1 < Y_1, X_2 < Y_1) \\ \pi_{XYX} &= P(X_1 < Y_1 < X_2) \\ \pi_{YXX} &= P(Y_1 < X_1, Y_1 < X_2). \end{aligned}$$

An unbiased estimator $\hat{V}(\theta)$ for the variance is obtained by replacing each probability in the formula (2) with its unbiased estimator. For example, an unbiased estimator for π_{XY} is given by

$$\hat{\pi}_{XY} = \frac{1}{nm}\#\{(i,j)|X_i < Y_j\}.$$

When n amd m tend to infinite, the asymptotic distribution of $g(\theta)/\sqrt{\hat{V}(\theta)}$ is the standard normal distribution. So we can construct an asymptotic confidence interval and hypothesis test for θ.

2.2. Ordered categorical data

We consider ordered categorical data for a special case. Assume that X_i and Y_j take only K values $c_1 < c_2 < \cdots < c_K$. f_k and g_k are frequencies of observations with value c_k in X_i's and Y_j's, respectively. In this case we have

$$g(\theta) = \sum_{k=1}^{K-1} \sum_{l=k}^{K} (f_k g_l - \theta f_l g_k).$$

The estimating function is rewritten by

$$g(\theta) = \sum_{k=1}^{K-1} \left(f_k \sum_{l=k}^{K} g_l - \theta g_k \sum_{l=k}^{K} f_l \right)$$

and this expression shows that $g(\theta)$ is also an unbiased estimating function for common odds ratio under the assumption of homogeneity in continuation-ratio logits model (see Agresti, 1990). Note that we can interpret θ as $P(X < Y)/P(X > Y)$ without the homogeneity assumption. The unbiased estimator $\hat{V}(\theta)$ for the variance of $g(\theta)$ is given by

$$\hat{V}(\theta) = V_1 - 2\theta V_2 \theta^2 V_3$$

where

$$\begin{aligned}
V_1 &= \sum_{k=1}^{K-1} f_k (\sum_{l=k}^{K} g_l)^2 \sum_{l=2}^{K} g_l (\sum_{k=1}^{l-1} f_k)^2 - \sum_{k=1}^{K-1} \sum_{l=k}^{K} f_k g_l \\
V_2 &= \sum_{k=1}^{K-2} \sum_{l=k}^{K-1} \sum_{h=l}^{K} (f_k g_l f_h g_k f_l g_h) \\
V_3 &= \sum_{k=1}^{K-1} g_k (\sum_{l=k}^{K} f_l)^2 \sum_{l=2}^{K} f_l (\sum_{k=1}^{l-1} g_k)^2 - \sum_{k=1}^{K-1} \sum_{l=k}^{K} g_k f_l
\end{aligned}$$

When $K = 2$, we have

$$g(\theta) = f_1 g_2 - \theta g_1 f_2.$$

It leads the estimator $\hat{\theta} = f_1 g_2 / f_2 g_1$ which is equivalent to the ordinary odds ratio. The estimation problem of the variance of this estimating functions are discussed in the situations of estimate the variance of the Mantel-Haenszel estimator. Robins et al. (1986) emphasized that the unbiasedness of estimators is important to keep the consistency for sparse data. Phillips and Holland (1987) and Sato (1991) proposed other estimators with unbiasedness and invariance when replacing the rows and coloumns in 2×2 tables. The estimator for the variance given in this section,

$$\hat{V} = f_1 g_2 (f_1 + g_2 - 1) \theta^2 f_2 g_1 (f_2 + g_1 - 1),$$

is different from those estimators but it has the unbiasedness and invariance.

3. Stratified Analysis

We next consider the stratified data case. Let $X_1^k, X_2^k, \cdots, X_{n_k}^k$ be a sample of size n_k from distribution F_k and $Y_1^k, Y_2^k, \cdots, Y_{m_k}^k$ be a sample of size m_k from distribution G_k for $k = 1, 2, \cdots, K$. We assume that $P(X_i^k < Y_j^k)/P(X_i^k > Y_j^k)$ does not depend on k and set the common value as θ. An estimating function $g^k(\theta)$ for k-th stratum is defined by the similar way to getting the function (1). The weighted sum of these estimating functions, $\sum w_k g^k(\theta)$, is an unbiased estimating function for θ. We apply the Godambe's criteria (Godambe, 1991),

$$M(\sum_{k=1}^{K} w_k g^k(\theta)) = \frac{Var[\sum w_k g^k(\theta)]}{E[\frac{\partial}{\partial\theta}\sum w_k g^k(\theta)]^2},$$

to get optimum weights. One of the optimum weights , which minimise the Godambe's criteria, is given by

$$w_k = \frac{E[\frac{\partial}{\partial\theta} g^k(\theta)]}{Var[g^k(\theta)]}. \tag{3}$$

However w_k depends on the distributions F_k and G_k. If we used the weight which depended on samples, the unbiasedness of the estimating function could not be guranteed. Yanagimoto (1990) noted that the unbiasedness is kept if we use the optimum weights in a special situation. The optimum weights in the case that $F_k = G_k$ are given by the following lemma. We call them the locally optimum weights.

Lemma 3.1. *Let X, Y and Z be independent random variables with same distribution F_k. The optimum weight w_k in the case that $F_k = G_k$ is given by*

$$w_k = \frac{-1}{(n_k + m_k) - (n_k + m_k - 2)\alpha}$$

where $\alpha = P(X < Y < Z)/P(X < Y)$.

Proof. Note that

$$P(X < Y) = 3P(X < Y < Z)(X < Y = Z)(X = Z < Y).$$

Using the above formula, we can easily lead the following equations.

$$\begin{aligned} Var[g^k(\theta)] &= n_k m_k \left[(n_k + m_k - 2)\left\{P(X < Y, X < Z) - 2P(X < Y < Z)\right.\right. \\ &\quad \left.\left.(X < Z, Y < Z))P(X < Y)\right\}\right] \\ &= n_k m_k \left\{(n_k + m_k)P(X < Y) - (n_k + m_k - 2)P(X < Y < Z)\right\} \\ E[\frac{\partial g^k(\theta)}{\partial\theta}] &= -n_k m_k P(X > Y) = -n_k m_k P(X < Y) \end{aligned}$$

The lemma is shown from these equations.

If we assume that F_k and G_k are continuous distributions, the optimum weights are $w_k = \frac{-3}{2(n_k+m_k)}$. If we assume that F_k and G_k are binary distiributions with same values, $w_k = \frac{-1}{n_k+m_k}$. Unfortunately the locally optimum weights depend on the structure of

the distributions F_k and G_k in general. One of the choice of the weights is that we use $1/(n_k + m_k + 1)$ if the number of ties in the samples is small and $w_k = 1/(n_k + m_k)$ for other cases. We set the weights as $\tilde{w}_k$. An estimator $\hat{\theta}$ is defined as the solution of $\sum \tilde{w}_k g^k(\theta) = 0$. The variance of $\hat{\theta}$ can be estimated by

$$\frac{\sum_{k=1}^{K} \tilde{w}_k^2 \hat{V}^k(\hat{\theta})}{\left\{\sum \tilde{w}_k \frac{\partial g^k(\hat{\theta})}{\partial \theta}\right\}^2}$$

where $V^k(\theta)$ is defined as similar way to getting formula (2). When $F_k = G_k$ for all k,

$$\frac{\sum_{k=1}^{k} \tilde{w}_k g^k(1)}{\sqrt{\sum_{k=1}^{K} \tilde{w}_k^2 \hat{V}^k(1)}}$$

follows the standard normal distribution asymptotically. We can test the hypethesis $F_k = G_k$ from this result.

In this analysis we assume that θ is homogeneous for all stratum. We need to verify the assumption in practice. When the number of strata is fixed and n_k, m_k tend to be large, we can apply the homogeneity test by Fujii (1994). The test statistic is given by

$$\sum_{k=1}^{K} \frac{g^k(\hat{\theta})^2}{\hat{V}^k(\hat{\theta})} - \frac{\left[\sum_{k=1}^{K} g^k(\hat{\theta}) E\left\{\frac{\partial}{\partial\theta} g^k(\hat{\theta})\right\} / \hat{V}^k(\hat{\theta})\right]^2}{\sum_{k=1}^{K} E\left\{\frac{\partial}{\partial\theta} g^k(\hat{\theta})\right\}^2 / \hat{V}^k(\hat{\theta})}$$

and under the homogeneity assumption it follows chi-square distribution with $(K-1)$ degree of freedom asymptotically.

4. Example

This section presents an application of our methods to the data from Liu and Agresti (1996). Table 1 shows data from a duble-blind, parallel-group clinical study conducted at a large number of centers. The purpose of the study was to compare an active drug with placebo in the treatment of patients suffering from asthma. Patients were randomly assigned to the treatments. Investigators described their perception of the patients' change in condition using the ordinal scale. Let's use the notation + for "better", = for "unchanged" and − for "worse". Liu and Agresti (1996) analysed the data using the cumulative logit model(see Agresti,1990). The log cumulative odds ratio is -1.153 and the standard error of the estimator is 0.571. In our methods $\log\theta = -1.306$ with standard error 0.501 when $\tilde{w}_k = 1/(n_k+m_k)$. The proposed methods shows smaller estimate and smaller standard error. In the view of pairwise comparison the difference comes from that Liu and Agresti (1996) twice used the pairwise estimating functions which compared column 1 with column 3. In general Liu and Agresti (1996) tends to lay emphasis on pairs with large difference. Liu and Agresti (1996) also assume the homogeneity with respect to the choice of base column. It needs to check the homoneity and the meaning of the parameter is difficult to understand even if we can assume the homogeneity. On the other hand the parameter θ is based on $P(X < Y)$, so it is easy to understand and we don't have to need the homogeneity asumption. Of course our methods assume homogeneity of θ over all strata. In this case it is difficult to check it

Table 1: Evaluations of patients suffering from asthma from Liu and Agresti (1996)

		Response					Response		
Center	Drug	=	–	Center	Drug	=	–		
1	Placebo	0	2	1	2	Placebo	0	1	0
	Active	1	1	0		Active	1	1	0
3	Placebo	1	1	0	4	Placebo	1	0	0
	Active	0	1	0		Active	1	1	0
5	Placebo	1	0	0	6	Placebo	1	0	0
	Active	1	0	0		Active	2	1	0
7	Placebo	0	1	0	8	Placebo	0	0	1
	Active	2	1	0		Active	0	1	0
9	Placebo	1	1	0	10	Placebo	0	2	0
	Active	1	1	0		Active	1	0	0
11	Placebo	2	0	0	12	Placebo	0	1	0
	Active	1	0	1		Active	1	0	0
13	Placebo	1	0	0	14	Placebo	0	1	0
	Active	1	0	0		Active	2	0	0
15	Placebo	1	0	0	16	Placebo	0	1	0
	Active	1	0	0		Active	1	0	0
17	Placebo	0	2	0	18	Placebo	0	1	0
	Active	1	1	0		Active	1	0	0
19	Placebo	1	0	0	20	Placebo	1	0	0
	Active	1	0	0		Active	1	0	0
21	Placebo	0	3	0	22	Placebo	0	2	0
	Active	0	1	0		Active	1	0	0
23	Placebo	1	0	0	24	Placebo	1	1	0
	Active	1	0	0		Active	1	0	0
25	Placebo	1	0	0	26	Placebo	0	1	1
	Active	1	0	0		Active	1	0	0
26	Placebo	0	1	0	28	Placebo	1	0	0
	Active	0	2	0		Active	1	1	0

because the sample sizes of each stratum is small. But one might still use $\hat{\theta}$ to summarise the association if the degree of heterogeneity is mild.

We can also test the hypothesis $\theta = 1$. The test statistic is -2.470 with p-value 0.014. The Mantel's generalized test with equally spaced score has 4.84 with p-value 0.028. These results are very similar. However Mantel's generalized test need to choose the column scores and the choice is sometimes plausible. On the other hand the proposed method does not need to choose them. It is one of the merits of the proposed method. But we have to note that there is another opinion about this issue if we are just interested in testing independent. Graubard and Korn (1987) recommended to assign reasonable column scores whenever possible and to consider equally spaced score when the choice is not apparent.

5. Discussions

Recently there are many clinical trials conducted in multiple centers. However some of studies did not use the information for the strata appropriately. The stratified analysis for 2×2 tables is well-known and the importance of stratification for sparse data was studied in many articles. But methods for other situations are not enough to be constructed. For example, Wilcoxon test is one of the basic analysis in two sample problems, but it is difficult to apply it for stratified data. In this paper we tried to construct a version of Wilcoxon test for stratified data based on probability $P(X < Y)$. The method for binary data is equivalent to the Mantel-Haenszel procedures. We can apply it for the ordered categorical data. For ordered cagtegorical data Wilcoxon test is known to be equivalent to the trend test if we choose appropriate scores. Another extension of Wilcoxon test may be considered to be the extended Mantel's test with the scores. Unfortunately it is not easy since the scores may be changed between strata.

The proposed methods can be extended for more general problems. Let X_i be an outcome which depends on a covariate vector, Z_i. If we assume that

$$logit \frac{P(X_i < X_j | Z_i, Z_j)}{P(X_i > X_j | Z_i, Z_j)} = \beta^T (Z_i - Z_j) \tag{4}$$

where T denotes the transposed operator and β is a vector of parameters with same dimension to Z_i, we can construct the methods based on pairwise estimating functions. Two sample problems are equivalent to the cases that Z_i is binary data. In other cases if we can assume

$$P(X_i < x | Z_i) = F(x)^{e^{\beta^T Z_i}},$$

we have the model (4). But the model is plausible in general. The analysis may need to check the model assumption. The methodology needs more study.

Acknowledgement

The idea of this paper comes from the Mantel-Haenszel procedures. The author would like to thank Prof. Takashi Yanagawa for introducing very atractive Mantel-Haenszel procedures and giving continuous encouragements. The author is also grateful to the editor and the referee for helpful comments.

References

Agresti, A. (1990). *Categorical data analysis.* John Wiley & Sons, New York.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators, *J. R. Statist. Soc.* **B32**, 283–301

Fujii, Y. (1994). On homogeneity test using estimating function, *Bulletin of Informatics and Cybernetics* **26**, 101–107

Fujii, Y. and Yanagimoto, T. (2004). The Mantel-Haenszel type estimator in the exponential family, to appear in *Journal of Statistical Planning and inference.*

Godambe, V. P. (1991). *Estimating Functions*, Clarendon Press, Oxford.

Graubard, B.I. and Korn, E.L. (1987). Choice of column scores for testing independence in ordered 2 K contingency tables, *Biometrics* **43**, 471–476.

Hochberg, Y. (1981). On the variance estimate of a Wilcoxon-Mann Whitney statistic for group ordered data, *Communications in Statistics: Theory and Methods* **10**, 1719–1732

Lehmann, E. L. (1975). *Nonparametrics, Statistical methods based on ranks* , Holden-Day, Inc., San Francisco.

Liu, I.-M. and Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response, *Biometrics* **52**, 1223–1234

Phillips, A. and Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate, *Biometrics*, **43**, 425–431.

Robins, J. M., Breslow, N. E. and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consisitent in both sparse data and large-strata limiting models, *Biometrics* **42**, 311–323.

Sato, T. (1991). An estimating equation approach for the analysis of case-control studies with exposure measured at several levels, *Statistics in Medicine* **10**, 1037–1042.

Wellek, S. and Hampel, B. (1999). A distribution-free two-sample equivalence test allowing for tied observations, *Biometrical Journal.* **41**, 171–186

Yanagawa, T. (1982). *Nonparametorikku Hou.* Baifukan, Japan.

Yanagimoto, T. (1990). Combining moment estimates of a parameter common through strata, *Journal of Statistical Planning and inference* **25**, 187–198

Received October 16, 2003
Revised June 16, 2004