Entropy Estimation from Judgment Post Stratified Data

Ehsan Zamanzade
Department of Statistics, University of Isfahan
Isfahan, 81746-73441, Iran
ehsanzamanzadeh@yahoo

Mahdi Mahdizadeh
Department of Statistics, Hakim Sabzevari University
P.O. Box 397, Sabzevar, Iran
mahdizadeh.m@live.com

Abstract

This article concerns entropy estimation using judgment post stratification sampling scheme. Some nonparametric estimators are developed and shown to be consistent. Monte Carlo simulations are used to compare these estimators with their competitors in simple random sampling. The results indicate the preference of the new estimators.

Keywords: Judgment ranking; Judgment post stratification; Entropy estimation.

Mathematics Subject Classifications: 94A15; 94A17; 62D05.

1. Introduction

Judgement post stratification (JPS) sampling scheme, introduced by MacEachern et al. (2004), has wide applications in situations where auxiliary information is available to induce an additional ranking structure in simple random sampling (SRS) scheme. This structure is obtained via a ranking process to determine the position of the measured units in the simple random sample among additional independent $H-1$ sample units from the target population. These positions are then used to put homogeneous units in the SRS in the same strata, and therefore an increased efficiency of a judgement post stratified data is expected from theory of stratified sampling in survey sampling designs.

To draw a judgement post stratified sample of size N, using set size H, one first draws a simple random sample of size N, say $Y_1,...,Y_N$, from the interest population and measures them. Next, for each Y_i in the simple random sample, he draws $H-1$ additional independent sample units from the population of interest to create the set of $\{Y_{i1},...,Y_{iH}\}$, (for $i=1,...,N$). This set is then ranked from smallest to largest and the rank of Y_i is recorded. This ranking process in JPS sampling scheme is done by any inexpensive method which does not require actual quantifications of the units in the set (e.g. eye inspection, personal judgement or using a covariate). If the ranking process is done based on eye inspection, the researcher should be blinded to the actual value of Y_i to avoid possible biases in the ranking process. Here the term judgement rank indicates that the ranking process in the JPS sampling scheme is done without actual measurement of the additional units, and thereby is prone to errors. Let F be the cumulative

distribution function (CDF) of the population, then the ranking process is called consistent if

$$F(y) = \frac{1}{H}\sum_{h=1}^{H} F_{[h]}(y),$$

where $F_{[h]}$ is the CDF of the hth judgement order statistics of a sample of size H.

A judgement post stratified sample of size N consists of a simple random sample of size N, with their corresponding judgement ranks, can be represented as $(Y_1, R_1),...,(Y_N, R_N)$, where R_i is the judgement rank of Y_i among the $H-1$ additional units. Let n_h be the number of observations with judgement rank h (for $h = 1,...,H$), one then can simply show that the vector $\mathbf{n} = (n_1,...,n_H)$ follows a multinomial distribution with mass parameter N and probability vector $(H^{-1},...,H^{-1})$.

Conditionally on the vector of judgement ranks $\mathbf{R} = (R_1,...,R_N)$ JPS sampling scheme can be regarded as ranked set sampling (RSS) scheme which was firstly introduced by McIntyre (1952). To construct an unbalanced ranked set sample of size N using set size H, one first determines the vector $\mathbf{n} = (n_1,...,n_H)$, where $N = \sum_{h=1}^{H} n_h$ is the total sample size. For $h = 1,...,H$, he draws n_h simple random samples of size H from the target population, and ranks them in increasing magnitude. The ranking process is accomplished similar to that in the JPS scheme, i.e. without obtaining the precise values of the units. He then actually measures the precise value of the hth judgement ordered unit in the rth sample (for $r = 1,...,n_h$). A ranked set sample is called balanced if $n_1 = ... = n_H$.

The main difference between the JPS sampling scheme and the RSS is about the ranking process. In the JPS setting, the ranking process is performed after measurements of the sample units, so the judgement ranks are loosely related to the measured units, and can be ignored. Therefore, a judgement post stratified sample can still be analyzed with standard SRS procedures. This is very useful in situations in which the researcher believes that the ranking process is too poor or the required statistical method has not been developed yet for the JPS setting. However, in the RSS setting, the ranking process is preformed prior to measurements of the sample units, so the judgement ranks of the units are strongly attached to them and cannot be disregarded. So, a ranked set sample must be analyzed with an appropriate procedure specially developed for the involved situation. Up-to-date references for the RSS scheme can be found in Wolfe (2012).

Both RSS and JPS sampling schemes are useful in situations in which exact measurement of sample units is expensive or time-consuming but ranking them (without obtaining their precise values) is easy and cheap. These situations frequently happen in forestry (Halls and Dell, 1966), medicine (Chen et al., 2005), environmental monitoring (Kvam, 2003;

Nussbaum and Sinha, 1997; and Ozturk et al., 2005), reliability (Mahdizadeh and Zamanzade, 2016) and entomology (Howard et al., 1982).

A lot of research has been done in the JPS sampling scheme in recent years. Wang et al. (2006) developed a class of estimators for population mean by using concomitant of multivariate order statistics. Wang et al. (2008) and Zamanzade (2016) used stochastic order constraint to develop estimators for population mean and variance, respectively. By conditioning on the observed vector of ordered in-stratum sample sizes, Frey and Feeman (2012) and Frey and Feeman (2013) developed conditionally unbiased estimators of population mean and variance, showed that their proposed estimators have better performance than their competitors in the literature. The problem of the estimation of the CDF has been considered by Wang et al. (2012). Frey and Feeman (2012) developed some mean estimators when the actual quantifications of the concomitant variable are available, which is followed by Zamanzade and Vock (2015) and Zamanzade and Mohammadi (2016). Dastbaravarde et al. (2013) provided a theoretical framework for JPS sampling scheme, and proved that the JPS moment estimators are less efficient than their counterpart in RSS. Ozturk (2014) developed an estimator for p th quantile of a distribution F. Two sample distribution free procedures in the JPS setting have been developed by Ozturk (2015). Zamanzade and Vock (2016) proposed some nonparametric tests for assessing assumption of perfect ranking in JPS.

In Section 2, we discuss the problem of CDF estimation in the JPS setting. In Section 3, we propose some nonparametric entropy estimators for judgment post stratified data. We then prove that the proposed estimators are consistent. In Section 4, we compare the proposed entropy estimators with their counterparts in the SRS setting. We end with a conclusion in Section 5.

2. Nonparametric estimation of CDF in JPS sampling scheme

Let $(Y_1, R_1),...,(Y_N, R_N)$ be a judgement post stratified sample of size N from a population with CDF F. Let $n_h = \sum_{i=1}^{N} I(R_i = h)$ be the number of the observations with judgement rank h, where $I(.)$ is an indicator function.

The standard CDF estimator of JPS sampling scheme is

$$\hat{F}_{st}(y) = \frac{1}{d_H} \sum_{h=1}^{H} \hat{F}_h(y) I(n_h > 0) \tag{1}$$

where $d_H = \sum_{h=1}^{H} I(n_h > 0)$, and $\hat{F}_h(y)$ is the empirical distribution function (EDF) based on the observations with judgement rank h.

Dastbaravarde et al. (2013) showed that this estimator is unbiased for the population CDF, and has less variance than its SRS rival, $F_n(t) = \frac{1}{N}\sum_{i=1}^{N} I(Y_i \le t)$, provided that the sample size N is not too small. They also proved that this estimator is strongly consistent and established its asymptotic normality as

$$\sqrt{N}\left(\hat{F}_{st}(y) - F(y)\right) \xrightarrow{d} N(0, \sigma_{st}^2)$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\sigma_{st}^2 = \frac{1}{H}\sum_{h=1}^{H} F_{[h]}(y)\left(1 - F_{[h]}(y)\right)$, and $F_{[h]}$ is the CDF of the hth judgement order statistics of a sample of size H. One can simply show that if the ranking process is consistent, then $\sigma_{st}^2 \le F(y)\left(1 - F(y)\right)$, meaning that $\hat{F}_{st}$ is asymptotically more efficient than F_n.

Wang et al. (2012) proposed some nonparametric CDF estimators which improve the performance of $\hat{F}_{st}$ when sample size is not too large. Their suggested estimators were constructed based on the idea that the CDF of judgement order statistics are often stochastically ordered, i.e.

$$F_{[1]}(y) \ge \ldots \ge F_{[H]}(y) \tag{2}$$

However, the constraint (2) may be violated by their sample estimates due to sampling noise. Therefore, one natural way to improve $\hat{F}_{st}$ is to impose the constraint (2) onto estimation process. Wang et al. (2012) proposed to estimate $\{F_h\}_{h=1}^{H}$ such that the weighted mean square $\sum_{h=1}^{H}\left(\hat{F}_h - F_h\right)^2$ is minimized under the constraint (2). There are two analytical forms for solving this optimization problem as

$$F_{[h]}^{+}(y) = \max_{s \ge h} \min_{r \le h} \sum_{g=r}^{s} \frac{n_g \, \hat{F}_{[g]}(y)}{n_{rs}} \tag{3}$$

and

$$F_{[h]}^{-}(y) = \min_{r \le h} \max_{s \ge h} \sum_{g=r}^{s} \frac{n_g \, \hat{F}_{[g]}(y)}{n_{rs}}, \tag{4}$$

where $n_{rs} = \sum_{g=r}^{s} n_g$.

If there is no empty stratum, then it turns out from properties of isotonic regression that the two above formulas are equivalent (Robertson and Waltman, 1968). However, if there exists at least one empty stratum, then $\left\{F_{[h]}^{+}(y)\right\}_{h=1}^{H}$ and $\left\{F_{[h]}^{-}(y)\right\}_{h=1}^{H}$ are not equivalent any more. Thus, two CDF estimators can be constructed as

$$\hat{F}_{iso^{+}}(y)=\frac{1}{H}\sum_{h=1}^{H}F_{[h]}^{+}(y) \tag{5}$$

and

$$\hat{F}_{iso^{-}}(y)=\frac{1}{H}\sum_{h=1}^{H}F_{[h]}^{-}(y). \tag{6}$$

Wang et al. (2012)'s simulation results indicate that $\hat{F}_{iso^{+}}$ works well on the left tail of $F(y)$ and $\hat{F}_{iso^{-}}(y)$ works well on the right tail of $F(y)$. To have an estimator with overall good performance, Wang et al. (2012) suggested to estimate the CDF with

$$\hat{F}_{iso}(y)=\frac{1}{2}\left[\hat{F}_{iso^{+}}(y)+\hat{F}_{iso^{-}}(y)\right]. \tag{7}$$

It is worth mentioning that the asymptotic behavior of the CDF estimators proposed by Wang et al. (2012) is the same as $\hat{F}_{st}$.

3. Nonparametric estimation of entropy

Let Y be the continuous random variable with the density function $f(y)$ and the CDF $F(y)$. The entropy $H(f)$ of this random variable, as a measure of uncertainty, is defined by Shannon (1948) as

$$H(f)=-\int_{-\infty}^{\infty}\log\left(f(y)\right)f(y)\,dy. \tag{8}$$

The problem of estimation of *H(f)* has been considered by many researchers in the literature. Vasicek (1976) was the first who proposed to estimate *H (f)* based on spacings. His estimate was obtained by using the fact that Equation (8) can be rewritten as

$$H(f)=-\int_{0}^{1}\log\left(\frac{d}{dp}F^{-1}(p)\right)dp. \tag{9}$$

The estimate was constructed by replacing distribution function *F* by the EDF, and the derivative $\frac{d}{dp}F^{-1}(p)$ is estimated by a function of order statistics.

Let $Y_{(1)} < ... < Y_{(N)}$ be an ordered random sample of size N from the population of the interest. Vasicek (1976)'s entropy estimator is given by

$$H_V = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{Y_{(i+m)} - Y_{(i-m)}}{2m/N}\right), \tag{10}$$

where $Y_{(i)} = Y_{(1)}$ for $i < 1$, $Y_{(i)} = Y_{(N)}$ for $i > N$, and $m \le N/2$ is an integer which is called window size.

Vasicek (1976) showed that $H_V \xrightarrow{p} H(f)$, as $m, N \to \infty$ and $m/N \to 0$, where $\xrightarrow{p}$ indicates convergence in probability.

Ebrahimi et al. (1994) modified Vasicek (1976)'s entropy estimator at the boundaries which are replaced by $Y_{(1)}$ and $Y_{(N)}$. Their estimator is given by

$$H_E = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{Y_{(i+m)} - Y_{(i-m)}}{F_N\left(Y_{(i+m)}\right) - F_N\left(Y_{(i-m)}\right)}\right), \tag{11}$$

where $F_N(.)$ is the EDF. Ebrahimi et al. (1994)'s simulation results indicate that this modification of the entropy estimator has less mean square error than the Vasicek (1976)'s entropy estimator.

The problem of estimation of *H (f)* in the RSS setting has been considered by Mahdizadeh and Arghami (2009), Al-Omari and Haq (2016).

Let $(Y_1, R_1), ..., (Y_N, R_N)$ be a judgement post stratified sample of size N from the population of interest. Motivated by Ebrahimi et al. (1994)'s entropy estimator, we propose to estimate the entropy from JPS samples by

$$H_Z = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{Y_{(i+m)} - Y_{(i-m)}}{\hat{F}_Z\left(Y_{(i+m)}\right) - \hat{F}_Z\left(Y_{(i-m)}\right)}\right), \tag{12}$$

where $Y_{(1)} < ... < Y_{(N)}$ are the ordered values of Y_i's, and $Z = st, iso^+, iso^-, iso$.

The next proposition establishes the consistency of the proposed estimators.

Proposition 1. *Let* $(Y_1, R_1), ..., (Y_N, R_N)$ *be a judgement post stratified sample from a continuous population with the CDF* F *and the entropy* $H(f)$. *If the ranking process is consistent, then* $H_Z \xrightarrow{p} H(f)$ *as* $m, N \to \infty$ *and* $m/N \to 0$, *where* $Z = st, iso^+, iso^-, iso$.

Proof. Note that

$$\begin{aligned}\left|H_Z - H(f)\right| &= \left|H_Z - H_E + H_E - H(f)\right| \\ &\le \left|H_Z - H_E\right| + \left|H_E - H(f)\right| \xrightarrow{p} 0, \qquad as\ \ N \to \infty, \frac{m}{N} \to 0\end{aligned}$$

where the last limit holds because under a consistent ranking process F_Z is a consistent estimator of F.

4. Monte Carlo comparison

We compare the performance of the proposed entropy estimators in the JPS sampling scheme with their competitors in SRS scheme via Monte Carlo simulation in terms of root of mean square error (RMSE). We have generated 10,000 judgement post stratified samples of size $N = 10, 20, 30, 50$, with set sizes $H = 3, 4, 10$ from standard normal, standard exponential and standard uniform distributions. So, we consider both effects of increasing sample size (N) and the set size (H) on the performance of the estimators. We control the quality of the ranking by using a concomitant variable in adaptive perceptual model proposed by Dell and Clutter (1972). Let (Y, X) be the interest variable and the concomitant variable, respectively, the judgement rank Y_i in the set $\{Y_i, Y_{i2}, ..., Y_{iH}\}$ is determined by using the concomitant vector $\mathbf{X}_i = (X_{i1}, ..., X_{iH})$, where

$$\mathbf{X}_i = (X_{i1}, ..., X_{iH}) = \rho\left(\frac{Y_i - \mu_y}{\sigma_y}, \frac{Y_{i2} - \mu_y}{\sigma_y}, ..., \frac{Y_{iH} - \mu_y}{\sigma_y}\right) + \sqrt{1-\rho^2}\,(Z_{i1}, ..., Z_{iH}),$$

and $(Z_{i1}, ..., Z_{iH})$ is an H-dimensional independent random vector from standard normal distribution.

In the Dell and Clutter (1972)'s model, the quality of ranking is controlled by the parameter ρ.

In this simulation study, we take $\rho = 1$ for perfect ranking, $\rho = 0.8$ for the ranking which is good enough that one expect improvement in the JPS sampling scheme, and $\rho = 0$ for random ranking. The value of window size (m) is selected according to the Grzegorzewski and Wieczorkowski (1999)'s heuristic formula as

$$m = \left\lfloor \sqrt{N} + 0.5 \right\rfloor,$$

where $\lfloor x \rfloor$ is the largest integer part of x.

Simulation results for three distributions of standard normal, standard exponential and standard uniform are presented in Tables 1-3, respectively.

Table 1: RMSE and bias of different entropy estimators in SRS and JPS settings when parent distribution is standard normal

			SRS		JPS $(\rho=1)$					JPS $(\rho=0.8)$					JPS $(\rho=0)$				
N	H		H_V	H_E	H_{st}	H_{iso^+}	H_{iso^-}	H_{iso}	H_L	H_{st}	H_{iso^+}	H_{iso^-}	H_{iso}	H_L	H_{st}	H_{iso^+}	H_{iso^-}	H_{iso}	H_L
10	3	RMSE	0.618	0.403	0.380	0.379	0.379	0.379	0.350	0.388	0.388	0.388	0.388	0.370	0.396	0.401	0.400	0.400	0.376
		Bias	-0.556	-0.300	-0.259	-0.258	-0.257	-0.258	-0.269	-0.268	-0.269	-0.268	-0.269	-0.259	-0.282	-0.293	-0.293	-0.293	-0.212
	4	RMSE	0.618	0.403	0.377	0.373	0.373	0.373	0.360	0.388	0.386	0.385	0.386	0.362	0.395	0.401	0.399	0.400	0.367
		Bias	-0.559	-0.303	-0.257	-0.255	-0.250	-0.254	-0.263	-0.269	-0.270	-0.265	-0.269	-0.251	-0.283	-0.296	-0.294	-0.296	-0.189
	10	RMSE	0.618	0.403	0.382	0.358	0.353	0.355	0.340	0.392	0.382	0.378	0.380	0.343	0.398	0.403	0.396	0.400	0.350
		Bias	-0.559	-0.303	-0.270	-0.251	-0.214	-0.240	-0.239	-0.279	-0.273	-0.241	-0.262	-0.221	-0.290	-0.297	-0.285	-0.292	-0.118
20	3	RMSE	0.374	0.246	0.233	0.233	0.233	0.233	0.220	0.238	0.238	0.238	0.237	0.228	0.242	0.245	0.245	0.244	0.230
		Bias	-0.330	-0.172	-0.151	-0.152	-0.151	-0.152	-0.154	-0.156	-0.157	-0.157	-0.157	-0.148	-0.164	-0.169	-0.169	-0.169	-0.110
	4	RMSE	0.374	0.246	0.231	0.230	0.230	0.230	0.220	0.236	0.237	0.237	0.237	0.224	0.241	0.246	0.246	0.246	0.229
		Bias	-0.330	-0.172	-0.144	-0.145	-0.145	-0.145	-0.150	-0.151	-0.153	-0.153	-0.153	-0.141	-0.160	-0.169	-0.169	-0.169	-0.088
	10	RMSE	0.374	0.246	0.226	0.219	0.217	0.218	0.214	0.235	0.233	0.231	0.232	0.217	0.240	0.246	0.245	0.246	0.240
		Bias	-0.332	-0.174	-0.139	-0.132	-0.126	-0.131	-0.136	-0.149	-0.149	-0.144	-0.148	-0.121	-0.157	-0.169	-0.168	-0.169	-0.022
30	3	RMSE	0.281	0.184	0.177	0.177	0.177	0.177	0.165	0.180	0.180	0.180	0.180	0.173	0.183	0.184	0.184	0.184	0.176
		Bias	-0.242	-0.116	-0.104	-0.104	-0.104	-0.104	-0.105	-0.107	-0.108	-0.108	-0.108	-0.101	-0.113	-0.115	-0.115	-0.115	-0.064
	4	RMSE	0.281	0.184	0.175	0.175	0.175	0.175	0.164	0.179	0.179	0.179	0.179	0.170	0.183	0.185	0.185	0.185	0.175
		Bias	-0.243	-0.117	-0.100	-0.101	-0.101	-0.101	-0.102	-0.105	-0.106	-0.106	-0.106	-0.096	-0.112	-0.116	-0.116	-0.116	-0.046
	10	RMSE	0.281	0.184	0.167	0.164	0.164	0.165	0.160	0.174	0.174	0.174	0.175	0.166	0.178	0.184	0.184	0.184	0.155
		Bias	-0.245	-0.119	-0.089	-0.087	-0.086	-0.087	-0.092	-0.097	-0.101	-0.100	-0.101	-0.079	-0.105	-0.116	-0.116	-0.116	0.020
50	3	RMSE	0.199	0.128	0.124	0.124	0.124	0.124	0.120	0.126	0.126	0.126	0.126	0.121	0.127	0.128	0.128	0.128	0.123
		Bias	-0.166	-0.066	-0.059	-0.059	-0.059	-0.059	-0.059	-0.061	-0.061	-0.061	-0.061	-0.056	-0.065	-0.066	-0.066	-0.066	-0.022
	4	RMSE	0.199	0.128	0.122	0.122	0.122	0.122	0.117	0.124	0.124	0.124	0.124	0.118	0.126	0.127	0.127	0.127	0.120
		Bias	-0.167	-0.067	-0.057	-0.057	-0.057	-0.057	-0.057	-0.061	-0.061	-0.061	-0.061	-0.053	-0.064	-0.066	-0.066	-0.066	-0.004
	10	RMSE	0.199	0.128	0.117	0.117	0.117	0.117	0.116	0.121	0.122	0.122	0.122	0.117	0.125	0.128	0.128	0.128	0.131
		Bias	-0.166	-0.066	-0.046	-0.047	-0.046	-0.047	-0.049	-0.053	-0.055	-0.055	-0.055	-0.038	-0.058	-0.065	-0.065	-0.065	0.060

Table 2: RMSE and bias of different entropy estimators in SRS and JPS settings when parent distribution is standard exponential

			SRS		JPS $(\rho=1)$					JPS $(\rho=0.8)$					JPS $(\rho=0)$				
N	H		H_V	H_E	H_{st}	H_{iso^+}	H_{iso^-}	H_{iso}	H_L	H_{st}	H_{iso^+}	H_{iso^-}	H_{iso}	H_L	H_{st}	H_{iso^+}	H_{iso^-}	H_{iso}	H_L
10	3	*RMSE*	0.561	0.399	0.401	0.401	0.403	0.402	0.390	0.405	0.405	0.406	0.405	0.400	0.407	0.408	0.408	0.408	0.397
		Bias	-0.432	-0.175	-0.134	-0.133	-0.132	-0.133	-0.144	-0.149	-0.151	-0.150	-0.151	-0.122	-0.158	-0.169	-0.169	-0.169	-0.087
	4	*RMSE*	0.561	0.399	0.406	0.405	0.412	0.408	0.390	0.405	0.408	0.411	0.409	0.393	0.399	0.401	0.401	0.401	0.398
		Bias	-0.432	-0.176	-0.130	-0.128	-0.123	-0.127	-0.135	-0.147	-0.150	-0.145	-0.149	-0.109	-0.157	-0.170	-0.167	-0.169	-0.063
	10	*RMSE*	0.561	0.399	0.398	0.389	0.418	0.401	0.385	0.402	0.403	0.416	0.409	0.386	0.401	0.400	0.400	0.410	0.398
		Bias	-0.441	-0.185	-0.152	-0.134	-0.098	-0.123	-0.121	-0.165	-0.156	-0.135	-0.150	-0.077	-0.172	-0.179	-0.167	-0.174	-0.001
20	3	*RMSE*	0.352	0.260	0.262	0.262	0.262	0.262	0.250	0.261	0.262	0.262	0.262	0.256	0.259	0.260	0.260	0.260	0.258
		Bias	-0.256	-0.097	-0.077	-0.077	-0.077	-0.077	-0.08	-0.085	-0.086	-0.086	-0.086	-0.064	-0.09	-0.095	-0.095	-0.095	-0.036
	4	*RMSE*	0.352	0.260	0.260	0.260	0.260	0.260	0.250	0.259	0.260	0.261	0.261	0.252	0.256	0.258	0.258	0.258	0.258
		Bias	-0.255	-0.096	-0.069	-0.069	-0.069	-0.069	-0.074	-0.079	-0.082	-0.082	-0.082	-0.052	-0.084	-0.093	-0.093	-0.093	-0.013
	10	*RMSE*	0.352	0.260	0.263	0.258	0.267	0.262	0.251	0.265	0.266	0.270	0.268	0.255	0.263	0.265	0.265	0.265	0.265
		Bias	-0.254	-0.095	-0.061	-0.054	-0.047	-0.053	-0.057	-0.074	-0.075	-0.072	-0.074	-0.016	-0.08	-0.091	-0.090	-0.091	-0.058
30	3	*RMSE*	0.272	0.205	0.206	0.206	0.206	0.206	0.200	0.206	0.206	0.206	0.206	0.202	0.205	0.205	0.205	0.205	0.205
		Bias	-0.189	-0.063	-0.051	-0.051	-0.051	-0.051	-0.052	-0.056	-0.056	-0.056	-0.056	-0.038	-0.060	-0.062	-0.062	-0.062	-0.011
	4	*RMSE*	0.272	0.205	0.206	0.206	0.206	0.206	0.200	0.206	0.207	0.207	0.207	0.202	0.205	0.206	0.206	0.206	0.205
		Bias	-0.190	-0.064	-0.047	-0.047	-0.047	-0.047	-0.049	-0.054	-0.055	-0.055	-0.055	-0.030	-0.058	-0.063	-0.063	-0.063	-0.008
	10	*RMSE*	0.272	0.205	0.204	0.202	0.205	0.204	0.200	0.204	0.205	0.207	0.206	0.202	0.203	0.205	0.205	0.205	0.210
		Bias	-0.189	-0.063	-0.033	-0.032	-0.031	-0.032	-0.036	-0.044	-0.049	-0.048	-0.049	-0.001	-0.049	-0.061	-0.060	-0.061	-0.076
50	3	*RMSE*	0.196	0.152	0.153	0.153	0.153	0.153	0.150	0.152	0.153	0.153	0.153	0.151	0.152	0.152	0.152	0.152	0.151
		Bias	-0.127	-0.027	-0.020	-0.020	-0.020	-0.020	-0.020	-0.023	-0.023	-0.023	-0.023	-0.009	-0.026	-0.027	-0.027	-0.027	-0.017
	4	*RMSE*	0.196	0.152	0.151	0.151	0.151	0.151	0.151	0.151	0.151	0.151	0.151	0.151	0.150	0.151	0.151	0.151	0.154
		Bias	-0.126	-0.026	-0.016	-0.016	-0.016	-0.016	-0.017	-0.021	-0.021	-0.021	-0.021	-0.002	-0.024	-0.025	-0.025	-0.025	-0.036
	10	*RMSE*	0.196	0.152	0.151	0.151	0.151	0.151	0.151	0.151	0.151	0.152	0.151	0.151	0.152	0.151	0.151	0.152	0.151
		Bias	-0.127	-0.027	-0.007	-0.007	-0.007	-0.007	-0.009	-0.015	-0.018	-0.018	-0.018	-0.022	-0.019	-0.026	-0.026	-0.026	-0.100

Table 3: RMSE and bias of different entropy estimators in SRS and JPS settings when parent distribution is standard uniform

			SRS		JPS $(\rho=1)$					JPS $(\rho=0.8)$					JPS $(\rho=0)$				
N	H		H_V	H_E	H_{st}	H_{iso^+}	H_{iso^-}	H_{iso}	H_L	H_{st}	H_{iso^+}	H_{iso^-}	H_{iso}	H_L	H_{st}	H_{iso^+}	H_{iso^-}	H_{iso}	H_L
10	3	*RMSE*	0.454	0.235	0.219	0.217	0.217	0.217	0.200	0.224	0.223	0.223	0.223	0.208	0.231	0.234	0.233	0.234	0.207
		Bias	-0.423	-0.167	-0.127	-0.126	-0.125	-0.126	-0.135	-0.135	-0.136	-0.135	-0.136	-0.124	-0.151	-0.162	-0.161	-0.161	-0.079
	4	*RMSE*	0.454	0.235	0.216	0.212	0.213	0.212	0.201	0.224	0.221	0.222	0.221	0.203	0.231	0.235	0.234	0.235	0.204
		Bias	-0.422	-0.166	-0.121	-0.119	-0.113	-0.118	-0.127	-0.133	-0.134	-0.129	-0.132	-0.113	-0.147	-0.16	-0.158	-0.159	-0.053
	10	*RMSE*	0.454	0.235	0.215	0.197	0.197	0.197	0.173	0.223	0.214	0.213	0.211	0.183	0.223	0.234	0.229	0.234	0.203
		Bias	-0.421	-0.165	-0.132	-0.115	-0.078	-0.104	-0.102	-0.142	-0.132	-0.101	-0.122	-0.081	-0.152	-0.160	-0.148	-0.155	-0.020
20	3	*RMSE*	0.274	0.134	0.120	0.120	0.120	0.120	0.114	0.124	0.125	0.125	0.125	0.116	0.130	0.132	0.132	0.132	0.120
		Bias	-0.260	-0.102	-0.082	-0.082	-0.082	-0.082	-0.085	-0.087	-0.088	-0.088	-0.088	-0.076	-0.095	-0.099	-0.099	-0.099	-0.040
	4	*RMSE*	0.274	0.134	0.117	0.117	0.117	0.117	0.110	0.123	0.123	0.123	0.123	0.112	0.129	0.132	0.132	0.132	0.116
		Bias	-0.261	-0.103	-0.075	-0.076	-0.076	-0.076	-0.081	-0.083	-0.085	-0.085	-0.085	-0.068	-0.091	-0.100	-0.100	-0.100	-0.019
	10	*RMSE*	0.274	0.134	0.114	0.105	0.106	0.105	0.100	0.122	0.119	0.119	0.119	0.101	0.128	0.134	0.133	0.134	0.124
		Bias	-0.262	-0.104	-0.069	-0.062	-0.056	-0.061	-0.065	-0.079	-0.079	-0.074	-0.078	-0.043	-0.087	-0.099	-0.098	-0.099	-0.049
30	3	*RMSE*	0.210	0.096	0.087	0.087	0.087	0.087	0.080	0.090	0.090	0.090	0.090	0.083	0.094	0.095	0.095	0.095	0.085
		Bias	-0.201	-0.076	-0.063	-0.063	-0.063	-0.063	-0.064	-0.067	-0.067	-0.067	-0.067	-0.056	-0.072	-0.074	-0.074	-0.074	-0.023
	4	*RMSE*	0.210	0.096	0.083	0.083	0.083	0.083	0.073	0.088	0.088	0.088	0.088	0.078	0.093	0.095	0.095	0.095	0.085
		Bias	-0.202	-0.076	-0.058	-0.059	-0.059	-0.059	-0.061	-0.064	-0.065	-0.065	-0.065	-0.050	-0.070	-0.074	-0.074	-0.074	-0.003
	10	*RMSE*	0.210	0.096	0.075	0.072	0.072	0.072	0.070	0.084	0.084	0.083	0.084	0.071	0.088	0.094	0.094	0.094	0.103
		Bias	-0.201	-0.076	-0.045	-0.043	-0.042	-0.043	-0.048	-0.054	-0.058	-0.057	-0.057	-0.027	-0.061	-0.073	-0.072	-0.073	-0.064
50	3	*RMSE*	0.155	0.063	0.057	0.057	0.057	0.057	0.050	0.059	0.059	0.059	0.059	0.053	0.062	0.062	0.062	0.062	0.060
		Bias	-0.151	-0.051	-0.044	-0.044	-0.044	-0.044	-0.044	-0.067	-0.066	-0.066	-0.046	-0.037	-0.049	-0.050	-0.050	-0.050	-0.007
	4	*RMSE*	0.155	0.063	0.055	0.055	0.055	0.055	0.049	0.058	0.058	0.058	0.058	0.050	0.061	0.062	0.062	0.062	0.060
		Bias	-0.151	-0.051	-0.042	-0.042	-0.042	-0.042	-0.042	-0.045	-0.045	-0.045	-0.045	-0.033	-0.049	-0.050	-0.050	-0.050	-0.011
	10	*RMSE*	0.155	0.063	0.047	0.047	0.047	0.047	0.041	0.054	0.055	0.055	0.055	0.043	0.058	0.062	0.062	0.062	0.069
		Bias	-0.151	-0.051	-0.030	-0.031	-0.031	-0.031	-0.033	-0.038	-0.041	-0.041	-0.041	-0.014	-0.043	-0.050	-0.050	-0.050	-0.076

Table 1 gives the results when the parent distribution is standard normal. It is seen that the entropy estimators in the JPS setting beats their competitors in SRS setting regardless of the quality of the ranking. The performance of the entropy estimators based on isotonic regression in the JPS setting ($H_{iso^+}, H_{iso^-}, H_{iso}$) are almost the same and slightly better than H_{st} for $\rho \geq 0.8$. It is also interesting to note that even in the case of random ranking ($\rho = 0$), the proposed entropy estimators have competitive performances with H_E. It is also evident from this table that H_E has always less RMSE and absolute Bias than H_V. However, the RMSE and Bias of all estimators improve as the value of sample size (N) decreases, and all entropy estimators have acceptable performances for $N \geq 30$. The performance of entropy estimators in the JPS setting improve as the value of set size (H) increases provided that the quality of ranking is fairly good ($\rho \geq 0.8$). It is worth mentioning that all entropy estimators underestimate the true value of the population entropy.

The analogous results for standard exponential distribution are presented in Table 2. We observe that for exponential distribution, although the JPS entropy estimators do not provide any improvement, their performances are slightly worse than H_E. Furthermore, H_E still has less RMSE and absolute Bias than H_V in the SRS setting. It is also evident that in the case of perfect ranking ($\rho = 1$), has less RMSE than H_{iso^-} and H_{iso}. The value of Bias of entropy estimators is negative, so they underestimate the true value of the population entropy. The RMSE and Bias of all estimators improve as the value of sample size (N) decreases, and all entropy estimators have acceptable performances for $N \geq 30$, however, their performances do not necessary improve with the value of the set size (H). Table 3 presents the simulation results when the parent distribution is standard uniform. The performance of the entropy estimators in this case is very similar to that of Table 1. The only clear difference is that the values of RMSE and Bias of entropy estimators in the standard uniform distribution are less than those in the standard normal distribution.

5. Conclusion

In this paper, we developed some nonparametric entropy estimators for judgement post stratification sampling scheme. The estimators were obtained by using different cumulative distribution function estimators in the JPS setting. We proved that the proposed entropy estimators are consistent. Our simulation results show that the entropy estimators in the JPS setting typically have better performance than their competitors in SRS setting, especially when the quality of ranking is fairly good.

In this paper, we confined our attention to estimation of entropy. However, it would be interesting to evaluate the performance of different entropy estimators for goodness of fit tests in the JPS setting, as well. This will be studied in the subsequent work.

Acknowledgments

The authors are thankful to an anonymous referee for his/her valuable comments and suggestions.

References

1. Al-Omari, A.I., Haq, A., 2016. Entropy estimation and goodness-of-fit tests for the inverse Gaussian and Laplace distributions using paired ranked set sampling. Journal of Statistical Computation and Simulation, 86, 2262-2272.
2. Chen, H., Stasny, E. A., Wolfe, D. A., 2005. Ranked set sampling for efficient estimation of a population proportion. Statistics in Medicine, 24, 3319-3329.
3. Correa J. C. 1995. A new estimator of entropy. Communications in Statistics-Theory Methods. 24, 2439-2449.
4. Dastbaravarde, A., Arghami, N.R., and Sarmad, M. 2013. Some theoretical results concerning non-parametric estimation by using a judgment post-stratification sample. arXiv:1211.4040v2 [stat.ME].
5. Dell, T. R., and Clutter, J. L., 1972. Ranked set sampling theory with order statistics background. Biometrics, 28, 545-555.
6. Ebrahimi, N., Habibullah, M. and Soofi, E,. 1994. Two measures of sample entropy, Statistics and Probability Letters, 20, 225-234.
7. Frey, J., and Feeman, T. G., 2012. An improved mean estimator for judgment post-stratification. Computational Statistics and Data Analysis, 56, 418-426.
8. Frey, J., and Feeman, T. G., 2013. Variance estimation using judgment post-stratification. Annals of the Institute of Statistical Mathematics, 65, 551-569.
9. Grzegorzewski, P., Wieczorkowski, R. 1999. Entropy-based goodness-of-fit test for exponentiality, Communications in Statistics-Theory Methods 28, 1183-1202.
10. Halls, L. K., Dell, T. R., 1966. Trial of ranked-set sampling for forage yields. Forest Science, 12, 22-26.
11. Howard, R. W., Jones, S. C., Mauldin, J. K., Beal, R. H., 1982. Abundance, distribution, and colony size estimates for Reticulitermes spp. (Isopter: Rhinotermitidae) in Southern Mississippi. Environmental Entomology, 11, 1290-1293.
12. Kvam, P. H., 2003. Ranked set sampling based on binary water quality data with covariates. Journal of Agricultural, Biological, and Environmental Statistics, 8, 271-279.
13. Kvam, P. H., and Samaniego, F. J., 1993. On the inadmissibility of empirical averages as estimators in ranked set sampling. Journal of Statistical Planning and Inference, 36, 39-55.
14. McIntyre, G. A., 1952. A method for unbiased selective sampling using ranked set sampling. Austral. J. Agricultural Res. 3, 385-390.
15. MacEachern, S.N., Ozturk, O., Wolfe, D.A. and Stark, G.V., 2002. A new ranked set sample estimator of variance. Journal of the Royal Statistical Society: Series B. 64, 177-188.
16. MacEachern, S.N., Stasny, E.A., and Wolfe, D.A., 2004. Judgment Post-Stratification with imprecise rankings. Biometrics, 60, 207-215.
17. Mahdizadeh, M., Arghami, N.R. 2009. Efficiency of ranked set sampling in entropy estimation and goodness-of-fit testing for the inverse Gaussian law, Journal of Statistical Computation and Simulation, 80, 761-774.
18. Mahdizadeh, M. and Zamanzade, E. 2016, A new reliability measure in ranked set sampling, To appear in Statistical Papers, doi:10.1007/s00362-016-0794-3.

19. Nussbaum, B.D., and Sinha, B.K., 1997. Cost effective gasoline sampling using ranked set sampling, Proceedings of the Section on Statistics and the Environment, 83-87, American Statistical Association.
20. Ozturk, O., Bilgin, O., and Wolfe, D. A. 2005. Estimation of population mean and variance in flock management: a ranked set sampling approach in a finite population setting. Journal of Statistical Computation and Simulation, 75, 905-919.
21. Ozturk, O., 2014. Statistical inference for population quantiles and variance in judgment poststratified samples. Computational Statistics and Data Analysis, 77, 188-205.
22. Ozturk, O., 2015. Distribution free two-sample methods for judgement post-stratified data. Statistica Sinica. To appear.
23. Perron, F., and Sinha, B.K., 2004. Estimation of variance based on a ranked set sample. Journal of Statistical Planning and Inference, 120, 21-28.
24. Robertson, T., Waltman, P., 1968. On estimating monotone parameters. The Annals of Mathematical Statistics. 39, 1030-1039.
25. Shannon C. E. 1948. A mathematical theory of communications. Bell System Technical Journal. 27, 623-656.
26. Stokes, S. L., 1980. Estimation of Variance Using Judgment Ordered Ranked Set Samples. Biometrics, 36, 35-42.
27. Stokes, S. L., and Sager, T. W., 1988. Characterization of a Ranked-Set Sample with Application to Estimating Distribution Functions, Journal of the American Statistical Association, 83, 374-381.
28. Vasicek, O. 1976. A test for normality based on sample entropy, Journal of Royal Statistical Society, Ser B, 38, 54-59.
29. Wang, X., Stokes, S. L., Lim, J. and Chen, M. 2006. Concomitants of multivariate order statistics with application to judgement post stratification. Journal of the American Statistical Association, 101, 1693-1704.
30. Wang, X., Lim, J., Stokes, L., 2008. A Nonparametric mean estimator for judgment post stratified data. Biometrics, 64, 355-363.
31. Wang, X., Wang, K., and Lim, J., 2012. Isotonized CDF estimation from judgment post stratification data with empty strata, Biometrics, 68, 194-202.
32. Wolfe, D. A. 2012. Ranked set sampling: Its relevance and impact on statistical inference. ISRN Probability and Statistics, doi: 10.5402/2012/568385.
33. Zamanzade, E. and Vock, M., 2015. Variance estimation in ranked set sampling using a concomitant variable, Statistics and Probability Letters, 105, 1-5.
34. Zamanzade, E. 2016. An isotonized variance estimator for judgment post stratified data, Journal of the Korean Statistical Society, 45, 25-32.
35. Zamanzade, E. and Mohammadi, M. 2016. Some modified mean estimator in ranked set sampling using a covariate. Journal of Statistical Theory and Applications, 15, 142-152.
36. Zamanzade, E. and Vock, M. 2016. Some nonparametric tests of perfect judgment ranking for judgment post stratification, To appear in Statistical Papers, doi: 10.1007/s00362-016-0805-4.