

11-1-2002

On Distribution Function Estimation Using Double Ranked Set Samples With Application

Walid A. Abu-Dayyeh

Yarmouk University, Irbid Jordan, abudayyehw@yahoo.com


Hani M. Samawi

Sultan Qaboos University, Sultanate of Oman

Lara A. Bani-Hani

Yarmouk University, Irbid Jordan

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Abu-Dayyeh, Walid A.; Samawi, Hani M.; and Bani-Hani, Lara A. (2002) "On Distribution Function Estimation Using Double Ranked Set Samples With Application," *Journal of Modern Applied Statistical Methods*: Vol. 1: Iss. 2, Article 53.

DOI: 10.22237/jmasm/1036110420

Available at: <http://digitalcommons.wayne.edu/jmasm/vol1/iss2/53>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized administrator of DigitalCommons@WayneState.

On Distribution Function Estimation Using Double Ranked Set Samples With Application

Walid A. Abu-Dayyeh
Department of Statistics
Yarmouk University, Irbid Jordan

Hani M. Samawi
Dept. of Mathematics & Statistics
Sultan Qaboos University
Al-Khod, Sultanate of Oman

Lara A. Bani-Hani
Department of Statistics
Yarmouk University, Irbid Jordan

As a variation of ranked set sampling (RSS); double ranked set sampling (DRSS) was introduced by Al-Saleh and Al-Kadiri (2000), and it has been used only for estimating the mean of the population. In this paper DRSS will be used for estimating the distribution function (cdf). The efficiency of the proposed estimators will be obtained when ranking is perfect. Some inference on the distribution function will be drawn based on Kolomgrov-Smirnov statistic. It will be shown that using DRSS will increase the efficiency in this case.

Key words: Double ranked set sample, distribution function estimation, Kolomgrov-Smirnov, ranked set.

Introduction

In some practical situations, collecting units from the population is not too costly comparing with quantification of the sampling units. A large number of those units may be identified to represent the population of interest and yet only a carefully selected subsample is to be quantified. This potential for observational economy was recognized for estimating the mean pasture and forge by McIntyre (1952). He proposed a method, later called ranked set sampling (RSS) by Halls and Dell (1966), currently under active investigation.

RSS procedure can be described as follows: Identify a group of sampling units randomly from the target population. Then, randomly partition the group into disjoint subsets each having a pre-assigned sizer r , in the most practical situations, the size r will be 2, 3 or 4. Then, rank each subset by a suitable method of ranking such as prior information, visual inspection or by the experimenter himself.

In terms of sampling notation,

where $X_{j(i)}$ denotes the i -th ordered statistic in the j -th set. Then the i -th ordered statistic from the i -th subset will be quantified, $i = 1, \dots, r$. Then $X_{1(1)}, X_{2(2)}, \dots, X_{r(r)}$ will be obtained. The whole process can be repeated k -times, to get a RSS of size $n = kr$. The resulting sample is called the balanced ranked set sample (RSS). Through all the paper, only balanced RSS will be used.

Al-Saleh and Al-Kadiri (2000) extended RSS to double rank set sample (DRSS). DRSS can be described as follows:

1. Identify r^3 elements from the target population and divide these elements randomly into r subsets each of size r^2 elements.
2. Use usual RSS procedure to obtain r RSS each of size r .
3. Apply again the RSS procedure in Step 2, on the r RSS's.

We may repeat steps 1, 2 and 3 k -times to obtain DRSS sample of size $n = rk$. In DRSS, ranking in the second stage is easier than ranking in the first stage, (see Al-Saleh and Al-Kadiri, 2000).

Moreover, an up-to-date annotated bibliography for RSS can be found in Kaur et al., (1995) and Patil et al. (1999). Stokes and Sager (1988) estimate the distribution functions, $F(x)$ say, for a random variable X by the empirical cdf (F^*) based on the RSS, which will be given in

The contact person for this article is Hani M. Samawi. Email him at hsamawi@squ.edu.om.

Section 2. They pointed out that, $F^*(t)$ is an unbiased for $F(t)$ and is more efficient than the empirical distribution function of a SRS

$(\hat{F}(t))$ of size n with

$$\text{Var}(F^*(t)) = \frac{1}{kr^2} \sum_{i=1}^m F_i(t) [1 - F_i(t)], \quad \text{where}$$

$$F_i(t) = F_{(i)}(t) = I_{F(t)}(i, r - i + 1) \quad (1.1)$$

for perfect ranking, and $I_{F(t)}(i, r - i + 1)$ is the incomplete beta ratio function.

Basic Setting of DRSS

Let Y_1, \dots, Y_r be a DRSS, and assume that $Y_i \sim g_i(y)$ with df, mean and variance are: $G_i(y), \mu_i^*$ and σ_i^{*2} , respectively. Al-Saleh and Al-Kadiri (2000) showed that:

(i) $f(y) = \frac{1}{r} \sum_{i=1}^r g_i(y), \quad (1.2)$

(ii) $F(y) = \frac{1}{r} \sum_{i=1}^r G_i(y), \quad (1.3)$

(iii) $\mu = \frac{1}{r} \sum_{i=1}^r \mu_i^*, \quad (1.4)$

(iv)

$$\sigma^2 = \frac{1}{r} \left[\sum_{i=1}^r \sigma_i^{*2} + \sum_{i=1}^r (\mu_i^* - \mu)^2 \right], \quad (1.5)$$

where f, F, μ and σ^2 are the pdf, cdf, mean and variance of the population.

In this paper, we will consider the problem of estimating the distribution function F using DRSS. In Section 2, the empirical cdf estimator based on DRSS (\hat{FDR}) will be considered. The efficiency between the DRSS estimator and those estimators based on SRS and RSS will be obtained when ranking is perfect. In Section 3 the Kolmogrov-Smirnov statistic will be studied based on a DRSS. Also, a confidence interval of $F(t)$ will be constructed using the Kolmogrov-Smirnov statistic based on DRSS.

Estimating The Distribution Functions Using DRSS

In this Section the distribution function will be estimated using the DRSS, in the cases where ranking is perfect and when ranking is imperfect. The suggested estimator will be compared with the cdf estimators based on SRS and RSS via their variances.

Definition and Some Basic Results

For the l -th cycle, let $\{Y_{1l}, Y_{2l}, \dots, Y_{rl}\}$, $l = 1, \dots, k$, be a DRSS of size r , and assume that Y_i has the probability density function (pdf) $g_i(y)$ and the cdf $G_i(y)$. Note that $g_i(y)$ is the density of the i -th ordered statistic of a RSS with densities $f_{(1)}, f_{(2)}, \dots, f_{(r)}$ and distribution functions $F_{(1)}, F_{(2)}, \dots, F_{(r)}$ respectively. Then

$$G_i(y) = \sum_{j=i}^r \sum_{S_j} \prod_{L=1}^j F_{(L)}(t) \prod_{L=j+1}^r [1 - F_{(L)}(t)] \quad (2.1)$$

where the set S_i consists of all permutations (i_1, i_2, \dots, i_r) of $1, 2, \dots, r$ for which $i_1 < \dots < i_j$ and $i_{j+1} < \dots < i_r$ (see Al-Saleh and Al-Kadiri, 2000).

Let \hat{FDR}, \hat{F} and F^* be the edf's (empirical distribution functions) of DRSS, SRS and RSS from the population with cdf F , then:

$$\hat{FDR}(t) = \frac{1}{kr} \sum_{j=1}^k \sum_{i=1}^r I[Y_{ij} \leq t] \quad (2.2)$$

$$\hat{F}(t) = \frac{1}{kr} \sum_{i=1}^{rk} I[X_i \leq t] \quad (2.3)$$

$$F^*(t) = \frac{1}{kr} \sum_{j=1}^k \sum_{i=1}^r I[X_{i(i)} \leq t] \quad (2.4)$$

respectively, where $I(\cdot)$ is the indicator function. Then, we have the following results.

a) $E[\hat{FDR}(t)] = F(t)$

$$b) \text{ var}(\hat{FDR}(t)) = \frac{1}{kr^2} \sum_{i=1}^r G_i(t)[1 - G_i(t)], \tag{2.5}$$

(see the Appendix for the prove of these results.) Also, we show in the Appendix that

$[\hat{FDR}(t) - E(\hat{FDR}(t))]/[\text{var}(\hat{FDR}(t))]^{1/2}$ converges in distribution to a standard normal random variable as $k \rightarrow \infty$ when r and t are held fixed. Moreover, it can be shown that an unbiased estimator of

$\text{var}[\hat{FDR}(t)]$ is given by

$$\hat{\text{var}}[\hat{FDR}(t)] = \frac{1}{(k-1)r^2} \sum_{i=1}^r \hat{G}_i(t)[1 - \hat{G}_i(t)], \tag{2.6}$$

where $\hat{G}_i(t) = \frac{1}{k} \sum_{j=1}^k I[Y_{ij} \leq t]$ is the edf based

on all k of the i -th judgment order statistic and hence it can be shown also that

$[\hat{FDR}(t) - E(\hat{FDR}(t))]/[\text{var}(\hat{FDR}(t))]^{1/2}$ converges in distribution to a standard normal random variable as $k \rightarrow \infty$ when r and t are held fixed. (See the Appendix for the prove of the above results.) Therefore, when k is large for a specified value t , an approximate $100(1-\alpha)\%$ confidence interval for $F(t)$ is

$$\hat{FDR}(t) \pm Z_{\alpha/2} \sqrt{\hat{\text{var}}[\hat{FDR}(t)]} \tag{2.7}$$

Finally, as a special case when $r = 2$, it can be shown that $\text{var}[\hat{FDR}(t)] \leq \text{var}[\hat{F}(t)]$

and $\text{var}[\hat{FDR}(t)] \leq \text{var}[F^*(t)]$. (See the Appendix Lemma 2 for the prove of this results.)

Efficiency of \hat{FDR}

The edf is used for making pointwise estimates of $F(t)$, as well as for making inference concerning the overall population distribution. In this section, we will examine the magnitude of the

improvement in precision that results when estimating $F(t)$ by $\hat{FDR}(t)$ rather than by $\hat{F}(t)$ or $F^*(t)$.

Now, the relative precision (RP) of the double ranked set to the simple random sampling estimator and to ranked set sample estimator, are defined by:

$$RP_1(t) = \frac{\text{var}[\hat{F}(t)]}{\text{var}[\hat{FDR}(t)]} = \frac{F(t)[1 - F(t)]}{F(t) - \left[\frac{\sum_{i=1}^r G_i^2(t)}{r} \right]} \tag{2.8}$$

$$RP_2(t) = \frac{\text{var}[F^*(t)]}{\text{var}[\hat{FDR}(t)]} = \frac{rF(t) - \sum_{i=1}^r F_{(i)}^2(t)}{rF(t) - \sum_{i=1}^r G_{(i)}^2(t)} \tag{2.9}$$

Table 1 and Table 2 show the value of $RP_1(F^{-1}(p))$ and $RP_2(F^{-1}(p))$ respectively, for some values of p and $r = 2, 3, 4, 5$. It can be noticed that both of RP_1 and RP_2 are monotone increasing from $p = 0$ to $p = 0.5$, to achieve their maximum at $p = 0.5$. Also, they are symmetric about $p = 0.5$. Table 1 and Table 2 show that the gain in efficiency from DRSS for estimation of $F(t)$ is substantial when the ranking can be done perfectly.

Table 1. $RP_1(F^{-1}(p))$ when ranking of X is perfect.

		P							
r	0.0	0.0	0.1	0.1	0.2	0.3	0.4	0.5	
	1	5	0	5	0	0	0	0	
2	1.0	1.0	1.1	1.1	1.2	1.4	1.5	1.6	
	1	5	2	9	7	4	8	4	
3	1.0	1.1	1.2	1.4	1.6	1.9	2.0	2.1	
	2	1	6	2	0	1	8	2	
4	1.0	1.1	1.4	1.6	1.9	2.3	2.5	2.6	
	3	8	1	8	4	2	2	0	
5	1.0	1.2	1.5	1.9	2.2	2.8	3.4	4.2	
	4	5	8	5	9	8	3	7	

Table 2. $RP_2(F^{-1}(p))$ when ranking of X is perfect.

		P							
R	0.0	0.0	0.1	0.1	0.2	0.3	0.4	0.5	
	1	5	0	5	0	0	0	0	
2	1.0	1.0	1.0	1.0	1.0	1.1	1.2	1.2	
	0	0	2	4	7	4	0	3	
3	1.0	1.0	1.0	1.1	1.1	1.2	1.3	1.3	
	0	1	5	1	7	8	2	3	
4	1.0	1.0	1.1	1.1	1.2	1.3	1.4	1.4	
	0	3	0	8	6	6	0	2	
5	1.0	1.0	1.1	1.2	1.3	1.5	1.7	2.1	
	0	4	4	6	6	3	2	0	

Inference on the distribution function

Because the distribution function F can be estimated more efficiently from a double ranked set sample than from a SRS and a RSS, it suffices to note that the statistics based on an estimate of F(t), such as the Kolmogorov-Smirnov statistic, would be improved in some sense as well.

In particular, we observe that the null distribution of the statistic

$D^{**} = \sup_t [\hat{FDR}(t) - F_0(t)]$ is stochastically smaller than $D^* = \sup_t [F^*(t) - F_0(t)]$ and smaller

than $D = \sup_t [\hat{F}(t) - F_0(t)]$ when D^{**} , D^* and D are all based on the same number of measured observations. We mean that

$$H_{(r)k}^{**}(d) \geq H_{(r)k}^*(d) \text{ and } H_{(r)k}^{**}(d) \geq H_{(r)k}(d)$$

with strict inequality for some d, where

$$H_{(r)k}^{**}(d) = p(D^* \leq d]$$

$$H_{(r)k}^*(d) = p(D^* \leq d) \text{ and } H_{rk}(d) = p(D \leq d].$$

Where D, D*, and D** are calculated from a SRS, a RSS and a DRSS of size rk respectively.

This implies that 100(1-α)% of D**,

which be denoted by C_{α}^{**} , will always be less than or equal to corresponding percentile of the statistics D and D*, denoted by C_{α} and C_{α}^* respectively. A confidence band for F based on

D** is

$$\hat{FDR} \pm C_{\alpha}^{**}, \tag{3.1}$$

is narrower than the corresponding band based on D and D*.

In this section, the simulations which we done, is true for some finite values of r and k in the case of perfect judgment ranking. To find the table of critical values of D** (C_{α}^{**}) we draw a double ranked set sampling (Y_i 's) of size n from uniform distribution with parameters 0, 1. Then all elements in the sample will be rank ($X_{(i)}$'s).

Now for k=1,

$$D^{**} = \max \left\{ \max_{1 \leq i \leq n} \left| \frac{1}{n} - F_0(Y_{(i)}) \right|, \max_{1 \leq i \leq n} \left| F_0(Y_{(i)}) - \frac{i-1}{n} \right|, 0 \right\}$$

where

$$F_0(X_{(i)}) = X_{(i)}.$$

The previous procedure will be repeated until we get, $D_1^{**}, D_2^{**}, \dots, D_{10000}^{**}$. Also, D_i^{**} 's will be ranked to find C_α^{**} such that, $P(D^{**} \leq C_\alpha^{**}) = 1 - \alpha$, i.e., the $C_\alpha^{**} = D_{(i)}^{**}$ where $i = [(1 - \alpha)10000]$, where $[d]$ is the greatest interge of d .

Now, Table 3 reports the critical values C_α^{**} for the test statistic D^{**} for $\alpha = 0.01, 0.05$ and 0.10 for $r = 2, 3, 4, 5$ and $k = 2, 3, \dots, 20$. The table shows that DRSS can result in a substantial decrease in width of the simultaneous confidence band. The amount of the improvement can be described by the quantities,

$$R_{\alpha 1} = \left(\frac{C_\alpha}{C_\alpha^{**}} \right)^2 \quad (3.2)$$

$$R_{\alpha 2} = \left(\frac{C_\alpha^*}{C_\alpha^{**}} \right)^2 \quad (3.3)$$

Because $R_{\alpha 1}$ and $R_{\alpha 2}$ are the square of the ratio of confidence-band widths, then they can be interpreted as a measure of relative precision. The ratios $R_{\alpha 1}$ and $R_{\alpha 2}$ are computed from the entries of Table 3 (C_α^{**}), Table 2 (C_α^*) (from Stokes and Sager; 1988) and the Table of critical values for the Kolmogrove-Smirnov statistic D (from Gibbons and Chakraborti (1992)).

Table 4 gives the values of $R_{\alpha 1}$ and $R_{\alpha 2}$ at $r = 2, \dots, 5$ and $k = 2, \dots, 10$. These values are comparable with those of Table 1 and Table 2. So, $R_{\alpha 1}$ and $R_{\alpha 2}$ indicate the same thing which given by $R_{p1}(t)$ and $R_{p2}(t)$, when ranking of X is perfect.

Table 3. Critical values of $D^{**}(C_{\alpha}^{**})$

	r=2			r=3			r=4			r=5		
	α :											
k	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
2	0.43	0.47	0.01	0.36	0.40	0.47	0.13	0.35	0.42	0.28	0.32	0.38
3	0.33	0.36	0.57	0.27	0.30	0.36	0.24	0.26	0.31	0.21	0.24	0.28
4	0.27	0.29	0.44	0.22	0.24	0.28	0.19	0.21	0.26	0.17	0.19	0.23
5	0.23	0.25	0.34	0.19	0.21	0.24	0.17	0.18	0.21	0.15	0.16	0.19
6	0.20	0.22	0.29	0.16	0.18	0.21	0.14	0.16	0.18	0.13	0.14	0.17
7	0.18	0.19	0.25	0.15	0.16	0.19	0.13	0.14	0.16	0.12	0.13	0.15
8	0.16	0.17	0.22	0.13	0.15	0.17	0.11	0.13	0.15	0.10	0.11	0.13
9	0.15	0.16	0.20	0.12	0.13	0.15	0.11	0.12	0.13	0.10	0.10	0.12
10	0.14	0.15	0.19	0.11	0.12	0.14	0.10	0.11	0.12	0.09	0.10	0.11
11	0.13	0.14	0.17	0.10	0.11	0.13	0.09	0.10	0.12	0.08	0.09	0.10
12	0.12	0.13	0.16	0.10	0.11	0.12	0.08	0.09	0.11	0.08	0.08	0.10
13	0.11	0.12	0.15	0.09	0.10	0.12	0.08	0.09	0.10	0.07	0.08	0.09
14	0.10	0.11	0.14	0.09	0.09	0.11	0.08	0.08	0.09	0.07	0.07	0.08
15	0.09	0.11	0.13	0.08	0.09	0.10	0.07	0.08	0.09	0.06	0.07	0.08
16	0.09	0.10	0.12	0.08	0.08	0.10	0.07	0.07	0.09	0.06	0.07	0.08
17	0.09	0.10	0.12	0.07	0.08	0.09	0.07	0.07	0.08	0.06	0.06	0.07
18	0.09	0.09	0.11	0.07	0.08	0.09	0.06	0.07	0.08	0.06	0.06	0.07
19	0.08	0.09	0.11	0.07	0.07	0.08	0.06	0.06	0.07	0.05	0.06	0.07
20	0.08	0.09	0.10	0.07	0.07	0.08	0.06	0.06	0.07	0.05	0.06	0.06

Table 4. The values of $R_{\alpha 1}$ and $R_{\alpha 2}$

$R_{\alpha 1}$									
	r= 2			r=3			r=4		
	$\alpha:$								
k	0.01	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
2	1.17	1.76	1.79	1.70	1.72	1.74	1.75	1.65	1.65
3	2.03	2.09	2.15	2.09	2.05	20.1	2.01	2.14	2.11
4	2.31	2.41	2.52	2.39	2.51	2.58	2.49	2.47	2.25
5	0.59	2.69	2.85	2.49	2.62	2.78	2.52	2.60	2.78
6	2.89	2.98	3.24	3.06	2.97	3.10	2.94	2.85	3.16
7	2.97	3.39	3.64	.300	3.29	3.20	3.13	3.19	3.52
8	3.52	3.77	3.80	3.41	3.24	3.45	3.64	3.13	3.48
9	3.48	3.75	3.79	3.67	3.70	4.27	3.30	3.36	4.31
10	3.72	3.74	4.24	4.00	4.00	4.29	4.00	3.64	4.34
$R_{\alpha 2}$									
2	1.41	1.42	1.34	1.23	1.16	1.18	1.15	1.27	1.22
3	1.70	1.70	1.62	1.49	1.44	1.43	1.36	1.33	1.27
4	1.88	2.00	2.08	1.74	1.78	1.74	1.60	1.54	1.42
5	2.19	2.19	2.30	1.87	1.78	20.1	1.67	1.78	1.78
6	2.40	2.39	2.56	2.25	2.09	2.18	2.04	1.72	2.09
7	2.60	2.66	2.83	2.15	2.25	2.33	1.92	2.04	2.07
8	2.85	3.11	3.06	2.61	2.35	2.52	2.39	2.14	2.15
9	2.78	3.06	3.02	2.78	2.86	2.78	2.12	2.25	2.61
10	2.94	3.00	.354	2.98	2.51	2.94	2.56	2.39	2.78

References

Al-Saleh, & Al-Kadiri (2000). Double ranked set sampling. *Statistics and Probability Letters*, 48(2), 205-212.

Gibbons, J. D., & Chakraborti, S. (1992). *Nonparametric statistical inference*. Marcel Dekker, Inc. New York, Basel, Hong Kong.

Halls, L. K., & Dell, T. R. (1966). Trial of ranked set sampling for forage yields. *Forest Science*, 12, 22-26.

Kaur, A. Patil, G. P., Sinha, A. K. and Tailie, C. (1995). Ranked set sampling. An annotated bibliography. *Environmental and Ecological Statistics*, 2, 25-45.

McIntyre, G. A. (1952). A method for unbiased selective sampling using ranked set. *Australian Journal of Agricultural Research*, 3, 385-390.

Patil G. P., A. K. Sinha and Tillie C. (1999). Ranked set sampling: A Bibliography. *Environmental Ecological Statistics*, 6, 91-98.

Stokes, S. L., & Sager, T. (1988). Characterization of a ranked set sample with application to estimating distribution functions. *Journal of American Statistical Association*, 83, 374-381.

Appendix

Proposition 1. \hat{FDR} is an unbiased estimator of F.

- a) $E[\hat{FDR}(t)] = F(t)$
- b)

$$\text{var}(\hat{FDR}(t)) = \frac{1}{kr^2} \sum_{i=1}^r G_i(t)[1 - G_i(t)].$$

Proof:

From the definition of a DRSS the proof will follow simply by using (1.3) and (2.1).

Proposition 2.

$[\hat{FDR}(t) - E(\hat{FDR}(t))]/[\text{var}(\hat{FDR}(t))]^{1/2}$ converges in distribution to a standard normal random variable as $k \rightarrow \infty$ when r and t are held fixed.

Proof: This follows from rewriting \hat{FDR} as

$$\hat{FDR} = \frac{1}{k} \sum_{j=1}^k U_j, \quad \text{where}$$

$$U_j = \sum_{i=1}^r \frac{I[Y_{ij} \leq t]}{r}, \quad \text{then } U_j \text{'s are iid,}$$

therefore the proof follows directly from the Central Limit Theorem.

Lemma 1.

- (a) $\text{var}[\hat{FDR}(t)]$ is an unbiased estimator of

$$\text{var}[\hat{FDR}(t)].$$

where:

$$\text{var}[\hat{FDR}(t)] = \frac{1}{(k-1)r^2} \sum_{i=1}^r \hat{G}_i(t)[1 - \hat{G}_i(t)]$$

$$\text{and } \hat{G}_i(t) = \frac{1}{k} \sum_{j=1}^k I[Y_{ij} \leq t] \text{ is the edf}$$

based on all k of the i-th judgment order statistic.

- (b)

$$[\hat{FDR}(t) - E(\hat{FDR}(t))]/[\text{var}[\hat{FDR}(t)]]^{1/2}$$

converges in distribution to a standard normal random variable as $k \rightarrow \infty$ when r and t are held fixed.

Proof:

- (a)

$$E[\text{var}[\hat{FDR}(t)]] = \frac{1}{(k-1)r^2} \sum_{i=1}^r [E(\hat{G}_i(t)) - E(\hat{G}_i^2(t))]$$

$$\text{because } E(\hat{G}_i(t)) = G_i(t)$$

and

$$E(\hat{G}_i^2(t)) = \frac{1}{k^2} \sum_{j=1}^k \text{var}(I[Y_{ij} \leq t]) + [G_i(t)]^2$$

$$\begin{aligned}
 &= \frac{k}{k^2} \text{var}(\mathbb{I}[Y_i \leq t]) + G_i^2(t) \\
 &= \frac{G_i(t)[1 - G_i(t)]}{k} + \frac{kG_i^2(t)}{k} \\
 &= \frac{G_i(t) + (k - 1)G_i^2(t)}{k} .
 \end{aligned}$$

Then $E[\text{var}(\hat{FDR}(t))] =$

$$\begin{aligned}
 &\frac{1}{(k-1)^2} \sum_{i=1}^r \left[\frac{kG_i(t)}{k} - \frac{G_i(t) + (k-1)G_i^2(t)}{k} \right] \\
 &= \frac{1}{kr^2} \sum_{i=1}^r G_i(t)[1 - G_i(t)] \\
 &= \text{var}[\hat{FDR}(t)]
 \end{aligned}$$

Part (b) can be shown by noting that:

$$\frac{\text{var}(\hat{FDR}(t))}{\text{var}(F^*(t))} \xrightarrow{p \rightarrow 1} 1 \text{ as } k \rightarrow \infty, \text{ and}$$

because $\hat{G}_i(t) \xrightarrow{p} G_i(t)$.

Furthermore, by Lemma 1 when k is large for a specified value t , an approximate $100(1-\alpha)\%$ confidence interval for $F(t)$ is:

$$\hat{FDR}(t) \pm Z_{\alpha/2} \sqrt{\widehat{\text{var}}[\hat{FDR}(t)]}$$

Lemma 2. : For the special case when $r = 2$,

(a) $\text{var}[\hat{FDR}(t)] \leq \text{var}[\hat{F}(t)]$

(b) $\text{var}[\hat{FDR}(t)] \leq \text{var}[F^*(t)]$.

Proof: Let $k = 1$ and $F(t) = F$

$$F_1(t) = 2F - F^2,$$

$$F_2(t) = F^2, \quad G_1(t) = F^4 - 2F^3 + 2F, \text{ and}$$

$$G_2(t) = 2F^3 - F^4.$$

Then $\text{var}[\hat{F}(t)] = \frac{2F - 2F^2}{4}$

$$\text{var}[\hat{F}^*(t)] = \frac{-2F^4 + 4F^3 - 4F^2 + 2F}{4},$$

and $\text{var}[\hat{FDR}(t)] = \frac{1}{4}[-2F^8 + 8F^7 - 8F^6 - 4F^5 + 8F^4 - 4F^2 + 2F]$.

Then $\text{var}[\hat{FDR}(t)] = \text{var}[\hat{F}(t)] - \frac{2F^2(1-F)^2(F^4 - 2F^3 - F^2 + 2F + 1)}{4} \leq \text{var}[\hat{F}(t)]$

Also, $\text{var}[\hat{FDR}(t)] = \text{var}[F^*(t)] - \frac{2F^3[1-F]^3[2-F][F+1]}{4} \leq \text{var}[F^*(t)]$,

$0 \leq F \leq 1$.