IMPUTASI MENGGUNAKAN PENAKSIR REGRESI UNTUK MENAKSIR RATA-RATA POPULASI PADA SAMPLING GANDA

Bernad Fundika Marpaung[1*], Rustam Efendi [2], Haposan Sirait[2]

[1] Mahasiswa Program Studi S1 Matematika
[2] Dosen Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Riau
Kampus Binawidya Pekanbaru (28293), Indonesia

[*]bernadfundika73@gmail.com

ABSTRACT

Estimators discussed in this paper are estimators that estimate a population mean with imputation method under double sampling design using regression estimators. Imputation methodsare used to estimate the missing data. This is a review of an article written by Thakur et. al [Journal of Reliability and Statistical Studies, 5(2): 21-31]. There are three estimators that discussed and each of them is an unbiased estimator. The minimum variance of each estimator is compared in order to decide the most efficient estimator.

Keywords**:** *Missing data, regression estimators, double sampling, simple random sampling, large sample approximation.*

ABSTRAK

Penaksir yang dibahas pada kertas kerja ini adalah penaksir untuk menaksir rata-rata populasi dengan metode imputasi pada sampling ganda dengan menggunakan penaksir regresi. Metode imputasi dilakukan untuk menaksir data hilang. Kertas kerja ini merupakan review dari artikel Thakur dkk [Journal of Reliability and Statistical Studies, 5(2): 21-31]. Terdapat tiga penaksir yang dibahas dan masing-masing penaksir adalah tak bias. Variansi minimum dari masing-masing penaksir dibandingkan untuk menentukan penaksir yang paling efisien.

Kata kunci: *data hilang, penaksir regresi, sampling ganda, sampling acak sederhana, hampiran sampel besar.*

1. PENDAHULUAN

Dalam penelitian terhadap populasi, akan lebih efektif diambil sebagian data sebagai sampel untuk diteliti karena bila seluruh populasi diteliti maka pada sebagian kasusdapat merusak populasi. Alasan lain yaitu akan lebih ekonomis dalam hal biaya, waktu dan tenaga. Metode pengambilan sampel ada dua cara, yaitu sampling probabilitas dan sampling nonprobabilitas. Pada kertas kerja ini akan lebih dibahas

mengenai sampling probabilitas, dan sampling probabilitas yang lebih dibahas adalah sampling ganda. Sampling ganda merupakan bagian dari sampling *multiphase*. Pada sampling ganda dilakukan dua kali pengambilan sampel dengan sampel pertama berukuran n' dan sampel kedua berukuran n.

Pada saat survei telah dilakukan terkadang terdapat *missing data* (data hilang atau data kurang lengkap) pada sampel sehingga menggangu analisa terhadap data hasil observasi. Untuk permasalahan ini perlu dilakukan proses membangkitkan data yang hilang tersebut dengan cara menaksir nilai dari data hilang tersebut. Ahmed dkk [1] sebelumnya telah membahas beberapa metode imputasi pada sampling sederhana sedangkan Thakur dkk [7] menyarankan tiga metode imputasi dengan masing-masing penaksir untuk rata-rata populasi pada sampling ganda. Penaksir ini akan lebih dibahas pada lembar kerja ini. Penaksir ini memanfaatkan penaksir regresi sebagai alat untuk imputasi. Karena ketiga dari penaksir ini merupakan tak bias maka akan dihitung variansi minimum dari ketiga penaksir, selanjutnya akan dibandingkan ketiga penaksir dan penaksir yang memiliki variansi terkecil merupakan penaksir yang paling efisien [2].

2. SAMPLING GANDA

Penarikan sampel secara sampling acak sederhana merupakan suatu metode untuk mengambil n unit sampel dari N unit populasi dimana setiap elemen memiliki kesempatan yang sama untuk diambil sebagai unit sampel. Beberapa teorema pada sampling acak sederhana adalah

Teorema 1 [3:h.27] Apabila sampel berukuran n diambil dari populasi berukuran N yang berkarakter Y pada sampling acak sederhana maka variansi rata-rata sampel $\bar{y}$ dinotasikan dengan $V(\bar{y})$ dan dirumuskan sebagai

$$V(\bar{y}) = \frac{S^2}{n}(1-f)$$

dengan $f = n/N$ adalah fraksi penarikan sampel dan $S^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2 \Big/ (N-1)$ adalah variansi y_i pada populasi berkarakter Y.
Bukti dari Teorema ini dapat dilihat pada [3: h.27].

Teorema 2 [3:h.29] Jika y_i, x_i adalah sebuah pasangan yang bervariasi ditetapkan pada unit dalam populasi dengan $\bar{y}, \bar{x}$ masing-masing adalah rata-rata dari sampel acak sederhana berukuran n maka kovariansi dari $\bar{y}, \bar{x}$ atau dinotasikan $\text{cov}(\bar{y}, \bar{x})$ adalah

$$\text{cov}(\bar{y}, \bar{x}) = \frac{(1-f)}{n}\rho S_y S_x$$

dengan $\rho = \sum(y_i - \bar{Y})(x_i - \bar{X}) \Big/ \sqrt{\sum(y_i - \bar{Y})^2 \sum(x_i - \bar{X})^2}$ adalah koefisien korelasi antara y_i dan x_i.

Bukti dari Teorema ini dapat dilihat pada [3:h.29].

Misalkan terdapat populasi berukuran N dan akan ditaksir rata-rata populasinya dengan variabel bantu X, tetapi $\bar{X}$ tidak diketahui sehingga pada kasus seperti ini akan lebih efektif jika penarikan sampel dilakukan menggunakan sampling ganda. Pada sampling ganda dilakukan dua kali pengambilan sampel, pertama diambil sampel berukuran n' secara sampling acak sederhana dari populasi berukuran N untuk menaksir $\bar{X}$ selanjutnya diambil sub-sampel berukuran n secara sampling acak sederhana dari sampel pertama tadi dan kali ini untuk memperkirakan variabel utama. Penaksir yang digunakan dalam penarikan sampling ganda dilambangkan dengan $\bar{y}_{ds_j}$ (ds_j merupakan singkatan dari double sampling). Akan dibuktikan jika dalam rata-rata sampel pertama $\bar{y}_{ss_i}$ adalah tak bias maka $\bar{y}_{ds_j}$ adalah penaksir yang tak bias dari $\bar{Y}$.

$$E\left(\bar{y}_{ds_j}\right)=\sum_{j=1}^{C_n^{n'}}\bar{y}_{ds_j}P\left(\bar{y}_{ds_j}\mid i\right).$$

$$E\left(\bar{y}_{ds_j}\right)=\frac{1}{C_n^{n'}}\sum_{j=1}^{C_n^{n'}}E\left(\bar{y}_{ds_j}\mid i\right).$$

$$E\left(\bar{y}_{ds_j}\right)=\frac{1}{C_n^{n'}}\sum_{j=1}^{C_n^{n'}}E\left(\bar{y}_{ss_i}\right)$$

$$E\left(\bar{y}_{ds_j}\right)=\bar{Y}.$$

Dengan demikian penaksir $\bar{y}_{ds_j}$ adalah penaksir yang tak bias dari $\bar{Y}$.

Teorema 3 [6 :h.287] Jika pengambilan sampel pertama secara acak yang berukuran n' dengan rata-rata sampel adalah $\bar{y}_{ss_i}$, serta sampel kedua adalah subsampel yang diambil secara acak berukuran n dari sampel pertama dengan rata-rata sampelnya adalah $\bar{y}_{ds_j}$, maka variansi rata-rata sampel pada sampling ganda adalah

$$\mathrm{Var}\left(\bar{y}_{ds_j}\right)=\left(\frac{1}{n'}+\frac{1}{N}\right)S_{ss_i}^2+\left(\frac{1}{n}-\frac{1}{n'}\right)S_{ds_j}^2.$$

Bukti dari Teorema ini dapat dilihat pada [6:h.287].

3. DATA HILANG

Salah satu permasalahan yang sering dialami dalam survei adalah terdapatnya data hilang. Data hilang tersebut terjadi karena terjadi kegagalan saat mengukur beberapa unit dalam sampel. Adanya data hilang pada sampel yang diteliti, mengakibatkan data hasil observasi tidak dapat dianalisa dengan baik. Cara mengatasi data hilang tersebut yaitu peneliti tetap melakukan penelitian menggunakan data yang ada atau

peneliti dapat melakukan survei ulang tetapi cara ini akan kurang efektif bila dilakukan. Cara lain yang dapat digunakan yaitu dengan membangkitkan data hilang tersebut dengan cara menaksir nilai dari data tersebut. Cara ini disebut dengan metode imputasi, cara ini lebih efektif digunakan dari pada cara yang sebelumnya.

Pada sampel kedua pada sampling ganda, kemungkinan dapat terjadi data hilang. Saat terjadi data hilang, sampel kedua dapat dibagi menjadi dua kelas yaitu kelas respon (R_1) dan kelas tidak respon (R_2). Kelas respon merupakan kelas yang memuat data yang lengkap sedangkan kelas tidak respon adalah kelas yang memuat data hilang.

4. VARIANSI DARI PENAKSIR REGRESI UNTUK RATA-RATA POPULASI DENGAN METODE IMPUTASI PADA SAMPLING GANDA

Diberikan y_{Vji} adalah notasi dari observasi ke-i untuk imputasi ke-j, dengan a,b dan c adalah konstanta sehingga variansi dari penaksir minimum. Misalkan $\bar{x}_1$, $\bar{x}_n$, dan $\bar{x}'$ masing-masing adalah rata-rata dari X untuk kelas respon, rata-rata dari X untuk sampel kedua, rata-rata dari X untuk sampel pertama. Metode imputasi yang digunakan adalah

1. Metode imputasi yang pertama

$$y_{Vji} = \begin{cases} y_i & jika \quad i \in R_1 \\ \left[\bar{y}_1 + a(x_i - \bar{x}_1)\right] & jika \quad i \in R_2 \end{cases} \tag{1}$$

untuk estimator

$$\bar{Y}_{V4} = \bar{y}_1 + a(\bar{x}_n - \bar{x}_1) \tag{2}$$

2. Metode imputasi yang kedua

$$y_{Vji} = \begin{cases} y_i & jika \quad i \in R_1 \\ \bar{y}_1 + \dfrac{b}{(1-W_1)}\left(\bar{x}' - \bar{x}_n\right) & jika \quad i \in R_2 \end{cases} \tag{3}$$

untuk estimator

$$\bar{Y}_{V5} = \bar{y}_1 + b(\bar{x}' - \bar{x}_n) \tag{4}$$

3. Metode imputasi yang ketiga

$$y_{Vji} = \begin{cases} y_i & jika \quad i \in R_1 \\ \bar{y}_1 + \dfrac{c}{(1-W_1)}\left(\bar{x}' - \bar{x}_1\right) & jika \quad i \in R_2 \end{cases} \tag{5}$$

untuk estimator

$$\bar{Y}_{V6} = \bar{y}_1 + c(\bar{x}' - \bar{x}_1) \tag{6}$$

dengan $W_1 = \dfrac{N_1}{N}$.

Variansi dari masing-masing penaksir adalah

1. Variansi dari penaksir $\bar{Y}_{V4}$

$$V(\bar{Y}_{V4}) = \left(L - \frac{1}{n'} \right) S_y{}^2 + \left(L - \frac{1}{n} \right) \left(a^2 S_x{}^2 - 2a\rho S_y S_x \right)$$

sehingga variansi minimum dari $\bar{Y}_{V4}$ adalah

$$\left[V(\bar{Y}_{V4}) \right]_{Min} = \left[\left(L - \frac{1}{n'} \right) - \left(L - \frac{1}{n} \right) \rho^2 \right] S_y{}^2 . \qquad (7)$$

2. Variansi dari penaksir $\bar{Y}_{V5}$

$$V(\bar{Y}_{V5}) = \left(L - \frac{1}{n'} \right) S_y{}^2 + \left(\frac{1}{N} - \frac{2}{n'} + \frac{1}{n} \right) \left(b^2 S_x{}^2 - 2b\rho S_y S_x \right)$$

sehingga variansi minimum dari $\bar{Y}_{V5}$ adalah

$$\left[V(T_{V5}) \right]_{Min} = \left[\left(L - \frac{1}{n'} \right) - \left(\frac{1}{N} - \frac{2}{n'} + \frac{1}{n} \right) \rho^2 \right] S_y{}^2 . \qquad (8)$$

3. Variansi dari penaksir $\bar{Y}_{V6}$

$$V(\bar{Y}_{V6}) = \left(L - \frac{1}{n'} \right) S_y{}^2 + \left(\frac{1}{N} - \frac{2}{n'} + L \right) \left(c^2 S_x{}^2 - 2c\rho S_y S_x \right)$$

sehingga variansi minimum dari $\bar{Y}_{V6}$ adalah

$$\left[V(\bar{Y}_{V6}) \right]_{Min} = \left[\left(L - \frac{1}{n'} \right) - \left(\frac{1}{N} - \frac{2}{n'} + L \right) \rho^2 \right] S_y{}^2 . \qquad (9)$$

Dengan notasi $L = E\left(\frac{1}{n'} \right) = \left[\frac{1}{nW_1} + \frac{(N-n)(1-W_1)}{(N-1)n^2 W_1{}^2} \right]$; $a = b = c = \rho \frac{S_y}{S_x}$.

5. PENAKSIR YANG EFISIEN

Selanjutnya ditentukan penaksir yang efisien diantara ketiga penaksir yang diajukan, yaitu dengan membandingkan variansi minimum dari masing-masing penaksir tersebut dengan menggunakan efisiensi relatif. Misalkan digunakan notasi

$$A = \frac{(1-W_1)}{(N-1)W_1{}^2}, \; B = \frac{2}{n'} - \frac{1}{N} \text{ dan } C = \left(2 - \frac{1}{W_1} + A \right)$$

sehingga

$$n_1{}^* = \frac{C + \sqrt{C^2 - 4BNA}}{2B}$$

$$n_2{}^* = \frac{C - \sqrt{C^2 - 4BNA}}{2B}.$$

Dilakukan perbandingan ketiga estimator berikut

1. Perbandingan penaksir $\bar{Y}_{V4}$ dan $\bar{Y}_{V5}$

   $Var(\bar{Y}_{V5}) < Var(\bar{Y}_{V4})$ jika

   $$n_2^* < n < n_1^* . \tag{9}$$

2. Perbandingan penaksir $\bar{Y}_{V4}$ dan $\bar{Y}_{V6}$

   $Var(\bar{Y}_{V6}) < Var(\bar{Y}_{V4})$ jika

   $$n'n + N(n' - 2n) > 0 \tag{10}$$

   atau

   $$n' - 2n > 0 . \tag{11}$$

3. Perbandingan penaksir $\bar{Y}_{V5}$ dan $\bar{Y}_{V6}$

   $Var(\bar{Y}_{V6}) < Var(\bar{Y}_{V5})$ dalam kondisi apapun.

Contoh :

Sebagai contoh dari pembahasan, diberikan data dari Kadilar dan Cingi [5] yakni tentang produksi apel Tahun 1999 di daerah Turki yang tersebar di 106 desa di daerah Aegean. Sebagai informasi tambahan (variabel bantu) digunakan banyak pohon apel yang ada di setiap desa yang telah diteliti sebelumnya. Pada pembahasan ini dimisalkan terjadi data hilang pada sampel kedua. Tabel 1 menunjukkan jumlah pohon apel dan banyaknya produksi apel di 106 desa pada daerah Aegean, dengan nama desa ditandai dengan nomor 1 sampai 106.

Tabel 1: Jumlah pohon apel dan produksi apel pada desa Aegean

NO	JUMLAH POHON APEL (X)	PRODUKSI APEL (Y)	NO	JUMLAH POHON APEL (X)	PRODUKSI APEL (Y)
1	86500	8650	54	10000	500
2	10800	432	55	30300	3030
3	300	30	56	2250	90
4	19000	1330	57	350	8
5	104900	3671	58	2715	163
6	1000	70	59	19000	475
7	11920	834	60	8400	210
8	33950	2886	61	200	4
9	12500	1000	62	200	5
10	280	17	63	140	3
11	350	21	64	3000	60
12	6000	156	65	9800	196

13	3500	280	66	58000	2030
14	5000	450	67	1500	38
15	1000	60	68	2500	88
16	56500	5650	69	24500	490
17	220000	15400	70	200	6
18	15000	390	71	21500	538
19	8000	320	72	0	0
20	53000	1325	73	54000	1350
21	27100	1084	74	12700	445
22	27670	1107	75	450	45
23	100	2	76	1750	36
24	100	2	77	7900	474
25	150	3	78	6225	1245
26	350	12	79	151050	5287
27	111000	2220	80	138000	11730
28	4750	153	81	5200	260
29	3000	90	82	18700	1496
30	6700	201	83	36500	1460
31	3500	70	84	8800	352
32	69500	2085	85	27400	137
33	113000	3164	86	4170	71
34	5050	131	87	315	6
35	53365	2668	88	4200	126
36	3750	94	89	88500	2655
37	19841	595	90	75000	2250
38	25500	637	91	30000	390
39	2350	108	92	7895	188
40	1600	48	93	350	7
41	5065	203	94	13000	325
42	700	5	95	18400	828
43	4950	248	96	3500	98
44	3950	126	97	700	11
45	13500	338	98	6200	248
46	44249	1947	99	5700	68
47	129750	6488	100	3640	73
48	1445	65	101	21250	1914

49	470800	117700	102	45350	1587
50	11200	67	103	6470	388
51	15050	677	104	1450	36
52	12000	960	105	57647	4450
53	2600	130	106	6600	165

Selanjutnya diperoleh informasi dari data pada Tabel 1 sebagai berikut

$N = 106$	$S_x = 57460{,}61$	$S_y^2 = 133437845{,}3409$
$n' = 45$	$S_y = 11551{,}53$	$L = 0{,}0617$
$n = 20$	$S_{xy} = 568176176{,}10$	$A = 0{,}002976$
$n_1 = 16$	$C_y = 5{,}22$	$B = 0{,}03501$
$\overline{X} = 27421{,}70$	$C_x = 2{,}10$	$C = 0{,}752976$
$S_x^2 = 3301721701{,}5721$	$\rho = 0.86$	

Dengan menggunakan informasi dari data tersebut, diperoleh bahwa

1. $n_1^* = 21{,}0797$
2. $n_2^* = 0{,}427469$

sehingga

$$n_2^* < n < n_1^*.$$

Dari informasi tersebut dapat dilihat bahwa pertidaksamaan (9) terpenuhi sehingga penaksir $\overline{Y}_{V5}$ lebih efisien daripada $\overline{Y}_{V4}$. Selanjutnya dari informasi tersebut diperoleh

$$n' - 2n = 5$$

dari informasi tersebut dapat dilihat bahwa pertidaksamaan (11) terpenuhi sehingga penaksir $\overline{Y}_{V6}$ lebih efisien daripada $\overline{Y}_{V4}$. Sedangkan penaksir $\overline{Y}_{V6}$ akan selalu lebih efisien daripada penaksir $\overline{Y}_{V5}$ dalam kondisi apapun. Dari keterangan-keterangan tersebut dapat disimpulkan bahwa penaksir $\overline{Y}_{V6}$ adalah penaksir yang paling efisien diantara ketiga penaksir untuk kasus data hilang pada contoh di atas. Selanjutnya nilai variansi minimum dari masing-masing penaksir disajikan pada Tabel 2.

Tabel 2: Nilai variansi minimum untuk ketiga penaksir

Penaksir	V_{Min}
$\bar{Y}_{V4}$	4163260.8
$\bar{Y}_{V5}$	3980451
$\bar{Y}_{V6}$	2684769

Berdasarkan Tabel 2 di atas, dapat dilihat bahwa penaksir $\bar{Y}_{V6}$ memiliki nilai variansi minimum terkecil dengan syarat bahwa kondisi lebih efisien dapat terpenuhi.

Untuk data yang sama variansi ini dibandingkan dengan MSE [4], seperti yang disajikan pada Tabel 3.

Tabel 3:*MSE* Penaksir Rasio dan Kombinasi Penaksir Rasio

Estimator	MSE
$\bar{y}_{KC4}$	2318718.59
$\bar{y}_{KC5}$	2317674.08
$\bar{y}_{pr3}$	1446719.34
$\bar{y}_{pr4}$	1446719.34

Dari Tabel 2 dan Tabel 3 dapat dilihat bahwa, jika terjadi data hilang pada sampel maka penaksir yang diberikan oleh Thakur dapat menaksir nilai dari rata-rata populasi dari data, dengan nilai variansi penaksir tidak jauh berbeda dari variansi sebelum terjadi data hilang.

6. KESIMPULAN

Variansi minimum dari masing-masing penaksir untuk rata-rata populasi yang diajukan telah diperoleh kemudian dengan membandingkan variansi minimum dari masing-masing penaksir sehingga dapat disimpulkan bahwa jika terjadi data hilang pada sampling ganda maka penaksir $\bar{Y}_{V6}$ akan lebih efisien daripada penaksir $\bar{Y}_{V4}$ dan penaksir $\bar{Y}_{V5}$ jika syarat efisien terpenuhi.

DAFTAR PUSTAKA

[1] Ahmed, M.S., O. Al-Titi., Z. Al-Rawi. and W. Abu-Dayyeh. (2006). *Estimation of a population mean using different imputation methods*, Statistics in Transition, 7 (6), p. 1247-1264.

[2] Bain, L.J. & M. Engelhard. 1991. *Introduction to Probability Mathematical Statistic, Second Edition*. Duxbury Press, California.

[3] Cochran, W. G. 1977. *Teknik Penarikan Sampel*, *Edisi ketiga*. Terj. Dari *Sampling Techniques,* oleh Rudiansyah & E. R. Osman. Universitas Indonesia, Jakarta.

[4] Ardhi, D. (2013). *Penaksir Rasio yang Lebih Efisien untuk Rata-rata Populasi pada Sampling Acak Sederhana*, Skripsi Sarjana Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Riau, Pekanbaru.

[5] Kadilar, C.& H. Cingi. 2006. *Improvement in estimating the population mean in simple random sampling*, Applied Mathematics and Computation. 19:75-79.

[6] Sukhatme, P. V. 1957. *Sampling Theory of Surveys with Applications*. The Indian Council of Agricultural Research, New Delhi.

[7] Thakur, N. S., K. Yadav.&S. Pathak.2012. *Imputation Using Regresion Estimators For Estimating Population Mean In Two-Phase Sampling*, Journal of Reliability and Statistical Studies, Vol. 5 issue 2, p. 21-31.