

Coverage Probabilities for Confidence Intervals Based on Stratified Random Sampling

Jörgen Dalén



R & D Report
Statistics Sweden
Research - Methods - Development
1988:2

Från trycket Juni 1988 1988
Producent Statistiska centralbyrån, Utvecklingsavdelningen
Ansvarig utgivare Staffan Wahlström
Förfrågningar Jörgen Dalén, tel. 08-7834494

© 1988, Statistiska centralbyrån
ISSN 0283-8680
Printed in Sweden
Garnisonstryckeriet, Stockholm 1988

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Häri ingår Abstracts : sammanfattningar av metoderapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1988:2. Coverage probabilities for confidence intervals based on stratified random sampling / Jörgen Dalén.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

COVERAGE PROBABILITIES FOR CONFIDENCE INTERVALS BASED
ON STRATIFIED RANDOM SAMPLING

by

JÖRGEN DALÉN

Coverage probabilities for confidence intervals based on stratified random sampling.

1. Introduction.

In this paper some empirical studies are presented dealing with the quality of confidence intervals based on stratified random sampling from finite populations. The paper should be read as a complementary documentation to Dalén (1986), where some concepts and ideas in this paper are presented in greater detail.

Stratification by an auxiliary variable correlated with the target variable is intended to reduce the effective variance and skewness of the population, thereby improving not only the precision of the estimates but also the coverage properties of confidence intervals based on the normal approximation. However, in many surveys you could not find a good enough auxiliary variable to do that job well and you therefore end up with quite skewed stratum populations anyway. In these situations the question arises as to whether the confidence intervals based on the normal approximation are valid and if some easy-to-use "rule of thumb" could be formulated for this problem.

In Dalén (1986) the following statement is made:

"There are indications that a rule like

$$n > K_{\alpha} \sum w_i G_{2i}^2,$$

where summation is over strata, where G_{2i} is G_2 in stratum i , and w_i are weights such that $\sum w_i = 1$, would work satisfactorily. If a Neyman allocation is used, $w_i = N_i \sigma_i / \sum N_i \sigma_i$ seems to work in many cases". (N_i is the size and σ_i^2 the variance of stratum i . K_{α} is a constant such that a nominal 95% confidence interval based on a sample of n units could be counted on to have at least α % actual coverage probability.)

2. The empirical calculations

The empirical studies, presented in table 1-7, are done for populations with four different types of strata: dichotomous, lognormal, power function and uniform. In the lognormal, power function and uniform cases the population consists of fixed percentiles of the corresponding parametric distributions.

In the dichotomous case the ACPs (=actual coverage probability, defined in Dalén (1986) are calculated exactly but in the other cases the EACPs (=estimated actual coverage probability) are based on Monte Carlo experiments with 500 random samples.

The samples are in all cases distributed according to a Neyman allocation. Some definitions are:

N_i = stratum size,

X_{ij} = variable value of unit j in stratum i ,

$\sigma_i^2 = \Sigma (X_{ij} - \mu_i)^2 / N_i$, where $\mu_i = \Sigma X_{ij} / N_i$,

$G_{2i} = \Sigma |X_{ij} - \mu_i|^3 / N_i \sigma_i^3$,

n = total sample size,

Σ stands for summation over j from 1 to N_i ,

L is number of strata and

$K = n \frac{\Sigma_{i=1}^L N_i \sigma_i}{\Sigma_{i=1}^L N_i \sigma_i G_{2i}^2}$ is an indicator of the goodness of the normal approximation used for calculating confidence intervals.

Due to the many dimensions involved in this problem (number of strata, stratum sizes, sample sizes, variances, and degrees of skewness in each stratum), it is difficult and expensive to make extensive, systematic studies. Those presented here should only be seen as a beginning from which some clues could be got for further analysis.

If we use a concept of average ACP similar to that in Dalén (1986) but now only referring to the calculated and tabled sample sizes we obtain the following spans for the two α -levels 90% and 94% in terms of K:

<u>Population</u>	<u>90 %</u>	<u>94 %</u>
Dichotomous	1.7 - 13.8	16.0 - 65.5
Lognormal	3.3 - 7.3	9.4 - 19.7
Power function, uniform	3.0 - 8.9	16.6 - 64.0

(The values are extracted from tables 1-7. There the 90% average ACP point is designated * and the 94% point **).

Roughly speaking, the spans of the constant K are similar in level with those in the simple random sampling case.

A general observation is that the ACPs for increasing sample sizes quickly and stably rise above 90% but thereafter in a quite long interval they oscillate between, say, 91% and 96%.

3. An example

In table 8 an example is given from the Survey on Exports and Imports of Services, a yearly financial survey carried out by Statistics Sweden. In the survey there are three principal parameters to estimate: total exports, total imports and total exports minus total imports. The population consists of around 60 000 enterprises among which a stratified random sample of 2 300 units is drawn.

We have done calculations on five consecutive surveys: from 1980 to 1984 for the three parameters. In the sample we have estimated G_{2i} in each stratum by

$$g_{2i} = \sum_S |X_{ij} - \bar{X}_i|^3 / n_i s_i^3, \text{ where}$$

$$\bar{X}_i = \sum_S X_{ij} / n_i \text{ and}$$

$$s_i^2 = \sum_S (X_{ij} - \bar{X}_i)^2 / n_i.$$

A weighted mean is calculated over all sampled strata (not take-all):

$$g_2^2 = \sum N_i s_i g_{2i}^2 / \sum N_i s_i$$

and so is a k -value based on it:

$$k = n/g_2^2.$$

These k -values now serve as instruments for a judgment on the coverage properties of the standard 95% confidence intervals calculated from these surveys.

We must carefully note the difference between g in this real-life example and G in the studies of parametric distributions above. g is an estimated property based on sample third moments and is very sensitive to the presence or absence of large population units in the sample.

But when it comes to practical applications data like those in table 8 are all you have. Having data from several years instead of just a single survey makes things easier. If, as in this case, most of the k -values are above 10, we should be able to feel quite confident that a nominal 95% confidence interval for any of the three parameters of interest has at least a 90% coverage probability.

4. Conclusion

Until more research in this area is done, the following "rule of thumb" may be applied for stratified random sampling, where approximate Neyman allocation is used:

For k -values below 5 the normal approximation is generally not reliable. Between 5 and 20 it is useful at least as a crude approximation. For k -values above 20 the normal approximation should be quite good. In applying this rule it is essential to have many observations on k and not just a single one. k should be based on an average g_2^2 calculated over all sampled strata (not the take-all strata) and in no sampled stratum should the sample size be close to the population size.

Reference:

Dalén, J. (1986): Sampling from Finite Populations: Actual Coverage Probabilities for Confidence Intervals on the Population Mean, Journal of Official Statistics, 2, 13-24.

Table 1: Populations with dichotomous strata of equal size and G_2

Stratum size = 200									Stratum size = 400											
Both $\sigma_1=0.5$ $G_{2i}=1.00$ $\sigma_2=1.0$			Both $\sigma_1=0.49$ $G_{2i}=1.06$ $\sigma_2=0.98$			Both $\sigma_1=0.46$ $G_{2i}=1.27$ $\sigma_2=0.92$			Both $\sigma_1=0.43$ $G_{2i}=1.44$ $\sigma_2=0.87$			Both $\sigma_1=0.40$ $G_{2i}=1.70$ $\sigma_2=0.80$			Both $\sigma_1=0.36$ $G_{2i}=2.09$ $\sigma_2=0.71$			Both $\sigma_1=0.30$ $G_{2i}=2.73$ $\sigma_2=0.60$		
n	K	ACP	n	K	ACP	n	K	ACP	n	K	ACP	n	K	ACP	n	K	ACP	n	K	ACP
6	6	75.4	6	5.3	88.6	6	3.7	84.9	6	2.9	78.5	6	2.1	70.8	6	1.4	59.1	6	0.8	45.6
9	9	89.1	9	8.0*	90.4	9	5.6*	93.1	9	4.3	90.9	9	3.1	85.3	9	2.1	75.9	9	1.2	61.2
12	12*	91.3	12	10.7	89.6	12	7.5	90.4	12	5.8	83.0	12	4.2	92.2	12	2.8	85.4	12	1.6	71.5
15	15	92.4	15	13.3	92.6	15	9.4	94.5	15	7.2*	90.8	15	5.2	82.2	15	3.4*	91.0	15	2.0	79.4
18	18	91.1	18	16.0**	95.1	18	11.2	92.6	18	8.6	93.8	30	10.4*	95.0	30	6.9	94.6	30	4.0	81.6
21	21	93.0	21	18.6	93.4	21	13.1	89.7	21	10.1	91.3	45	15.6	90.6	45	10.3	91.7	45	6.0*	94.3
24	24**	94.3	24	21.3	94.4	24	15.0	94.1	24	11.5	94.5	60	20.8	93.6	60	13.8	94.5	60	8.0	94.3
27	27	94.9	27	24.0	93.5	27	16.9	93.4	27	13.0	92.2	75	26.0**	94.6	75	17.2	94.5	75	10.0	94.7
30	30	94.5	30	26.6	94.3	30	18.7	91.8	30	14.4	93.6	90	31.1	94.3	90	20.7	93.2	90	12.0	92.8
45	45	94.5	45	39.9	94.7	45	28.1	94.2	45	21.6**	95.1	105	36.3	93.8	105	24.1	93.1	105	14.1	91.7
60	60	93.4	60	53.3	94.9	60	37.5	93.4	60	28.8	94.8	120	41.5	94.7	120	27.6	93.1	120	16.1	93.0
75	75	93.8	75	66.6	94.7	75	46.8	93.5	75	36.0	94.7	135	46.7	94.1	135	31.0**	94.5	135	18.1	93.7
90	90	94.7	90	79.9	95.1	90	56.2	93.5	90	43.2	94.2	150	51.9	94.6	150	34.5	95.4	150	20.1**	94.5
105	105	93.6	105	93.2	94.0	105	65.5**	94.1	105	50.4	95.2				165	37.9	95.0	165	22.1	94.7
120	120	94.3	120	106.5	94.8	120	74.9	94.8	120	57.6	94.1				180	41.3	95.5	180	24.1	93.6
						135	84.3	95.5	135	64.8	95.2				195	44.8	95.1	195	26.1	94.2
									150	72.0	94.3				210	48.2	95.2	210	28.1	95.2
															225	51.7	95.1	225	30.1	93.3
															240	55.1	94.1	240	32.1	94.5
															255	58.6	94.2	255	34.1	95.1
															270	62.0	94.3	270	36.1	93.9
															285	65.5	94.3	285	38.1	95.2
															300	68.9	95.0	300	40.2	93.6

Table 3: Populations with 2 lognormal strata of 100 units each

$G_{21}=1.94$ $\sigma_1=0.45$ $G_{22}=3.02$ $\sigma_2=1.25$			$G_{21}=1.94$ $\sigma_1=0.45$ $G_{22}=3.02$ $\sigma_2=1.25$			$G_{21}=1.94$ $\sigma_1=0.45$ $G_{22}=2.68$ $\sigma_2=0.99$			$G_{21}=1.94$ $\sigma_1=0.45$ $G_{22}=2.68$ $\sigma_2=0.99$			$G_{21}=1.67$ $\sigma_1=0.21$ $G_{22}=1.94$ $\sigma_2=0.45$			$G_{21}=1.67$ $\sigma_1=0.21$ $G_{22}=2.39$ $\sigma_2=0.77$		
n	K	EACP	n	K	EACP	n	K	EACP	n	K	EACP	n	K	EACP	n	K	EACP
8	1.0	81.2	103	13.4	93.2	6	1.0	82.8	87	14.2	93.2	6	1.7	86.4	10	2.0	87.0
11	1.4	87.4	106	13.8**	94.2	10	1.6	87.6	90	14.7	92.6	10	2.9	89.8	14	2.8	89.4
15	1.9	86.8	110	14.3	94.2	13	2.1	89.8	93	15.2	94.0	13	3.8*	93.2	19	3.7*	92.0
19	2.5	87.6	114	14.8	95.6	16	2.6	87.6	96	15.7	93.6	16	4.6	91.2	24	4.7	92.0
23	3.0	90.2	118	15.3	94.6	19	3.1	90.6	99	16.2	95.6	19	5.5	93.2	29	5.7	91.2
27	3.5	88.0	122	15.8	94.2	22	3.6	89.2	103	16.8	93.0	22	6.4	90.6	33	6.5	93.4
30	3.9*	93.2	125	16.2	95.4	26	4.3*	92.4	106	17.3	93.8	25	7.2	90.4	38	7.5	92.6
34	4.4	91.4	129	16.7	95.0	29	4.7	90.0	109	17.8	92.4	29	8.4	92.4	43	8.5	92.6
38	4.9	93.8	133	17.3	90.4	32	5.2	93.6	112	18.3**	94.8	32	9.3	92.0	48	9.4**	94.8
42	5.5	93.0	137	17.8	94.8	35	5.7	93.8	115	18.8	94.6	35	10.1	94.0	52	10.2	93.6
46	6.0	92.6				38	6.2	93.2	119	19.5	93.4	38	11.0	93.2	57	11.2	95.2
49	6.4	92.2				42	6.9	92.0	122	20.0	94.8	41	11.9	93.6	62	12.2	94.4
53	6.9	92.0				45	7.4	91.2	125	20.4	93.8	44	12.7	95.0	67	13.2	93.8
57	7.4	93.0				48	7.9	93.8	128	20.9	95.4	48	13.9	92.2	71	14.0	94.2
61	7.9	91.4				51	8.3	92.4	131	21.4	96.0	51	14.8**	94.6	76	14.9	95.6
65	8.4	91.8				55	9.0	92.2	135	22.1	95.2	54	15.6	93.8	81	15.9	93.2
68	8.8	95.0				58	9.5	96.0	138	22.6	93.0	57	16.5	93.6	86	16.9	95.6
72	9.3	91.8				61	10.0	91.8	141	23.1	94.2	60	17.4	96.4	90	17.7	95.0
76	9.9	93.0				64	10.5	94.6	144	23.6	94.8	64	18.5	93.4	95	18.7	93.6
80	10.4	92.4				67	11.0	93.0				67	19.4	93.4	100	19.7	91.2
84	10.9	91.2				71	11.6	94.8				70	20.3	95.0	105	20.6	94.8
87	11.3	91.8				74	12.1	91.0				73	21.1	94.8	110	21.6	93.6
91	11.8	92.4				77	12.6	94.8				76	22.0	95.2	114	22.4	94.0
95	12.3	92.6				80	13.1	95.2				79	22.9	95.4	119	23.4	95.2
99	12.9	94.6				83	13.8	93.8				83	24.0	94.4			
												86	24.9	95.2			
												89	25.8	94.8			

Table 4: Populations with 3 lognormal strata

$N_1=20 \quad N_2=N_3=50$			$N_1=N_2=N_3=50$						$N_1=N_2=N_3=100$																							
G_{21}	σ_1		G_{21}	σ_1		G_{21}	σ_1		G_{21}	σ_1		G_{21}	σ_1		G_{21}	σ_1		G_{21}	σ_1		G_{21}	σ_1										
G_{22}	σ_2		G_{22}	σ_2		G_{22}	σ_2		G_{22}	σ_2		G_{22}	σ_2		G_{22}	σ_2		G_{22}	σ_2		G_{22}	σ_2		G_{22}	σ_2							
G_{23}	σ_3		G_{23}	σ_3		G_{23}	σ_3		G_{23}	σ_3		G_{23}	σ_3		G_{23}	σ_3		G_{23}	σ_3		G_{23}	σ_3		G_{23}	σ_3							
n	K	EACP	n	K	EACP	n	K	EACP	n	K	EACP	n	K	EACP	n	K	EACP	n	K	EACP	n	K	EACP	n	K	EACP						
8	1.4	81.0	13	3.4	89.8	8	1.2	82.4	7	2.3	86.0	8	2.8	89.2	8	1.9	84.6	8	1.2	83.0	7	2.1	84.0									
16	2.9	87.4	28	7.3*	90.4	15	2.3	87.6	15	4.8*	91.0	12	4.1*	92.4	12	2.8	88.0	12	1.8	87.4	11	3.3*	91.2									
24	4.3*	91.6	41	10.8	95.8	23	3.5*	90.6	21	6.8	90.4	15	5.2	90.2	15	3.5	87.4	15	2.3	87.8	15	4.5	92.2									
31	5.6	91.2	54	14.2	94.6	30	4.6	92.2	29	9.4	93.8	19	6.6	92.2	20	4.7*	92.0	19	2.9	89.2	18	5.4	91.8									
39	7.0	92.2	69	18.1	93.8	39	6.0	91.2	36	11.6**	95.0	23	7.9	91.0	23	5.4	91.2	24	3.6*	90.8	21	6.3	93.0									
48	8.6	94.2	82	21.5	92.6	46	7.1	93.4	43	13.9	95.6	27	9.3	93.0	27	6.4	92.6	27	4.1	89.2	25	7.6	93.2									
55	9.9	93.0				54	8.3	90.6	51	16.5	96.0	30	10.4	93.0	31	7.3	94.2	31	4.7	91.0	29	8.8	92.0									
						57	8.8	91.8				34	11.8	94.2	35	8.3	92.2	34	5.1	91.4	33	10.0	93.0									
												39	13.5	93.4	38	9.0	90.2	39	5.9	91.4	36	10.9	94.0									
												42	14.5	92.2	43	10.2	93.2	43	6.5	93.6	39	11.8	94.4									
												46	15.9	94.6	46	10.9	92.8	46	7.0	92.2	43	13.0	94.0									
												50	17.3	94.0	50	11.8	93.8	51	7.7	93.0	47	14.2	93.6									
												54	18.7	92.2	54	12.8	96.8	54	8.2	94.2	51	15.4	93.6									
												57	19.7**	94.4	58	13.7	93.6	58	8.8	94.2	54	16.3**	95.2									
												61	21.1	94.4	61	14.4	91.2	62	9.4	92.6	57	17.2	94.6									
												65	22.5	94.6	66	15.6**	95.6	66	10.0	94.2	61	18.4	96.6									
												69	23.8	93.4	69	16.3	94.6	70	10.6	94.0	64	19.3	95.4									
												72	24.5	94.4	73	17.3	94.8	73	11.1	94.0	69	20.8	95.4									
												76	26.3	95.8	77	18.2	93.0	77	11.7	91.4	72	21.8	93.8									
												81	28.0	95.2	81	19.2	94.8	82	12.4**	95.0	75	22.7	93.8									
												84	29.0	94.2	84	19.9	93.0	85	12.9	93.8	79	23.9	93.6									
												88	30.4	95.8	89	21.1	94.8	89	13.5	94.8	82	24.8	93.4									
												91	31.5	93.6	92	21.8	94.0	93	14.1	95.2	87	26.3	93.8									
												96	33.2	94.0	96	22.7	94.0	97	14.7	94.6	90	27.2	94.4									
												99	34.2	95.0	100	23.7	94.0	101	15.3	93.4	93	28.1	96.6									
												103	35.6	94.4	104	24.6	94.0	104	15.8	92.6	97	29.3	96.4									
												107	37.0	96.0	107	25.3	92.8	109	16.5	94.8	100	30.2	93.2									
												111	38.4	95.4	112	26.5	92.0	113	17.1	92.6	105	31.7	94.6									
												115	39.7	95.2	115	27.2	92.6	116	17.6	93.8	108	32.6	94.6									
																		120	18.2													

Table 5: Populations with 4 lognormal strata of 100 units each

$G_{21}=1.65$ $\sigma_1=0.20$ $G_{22}=1.69$ $\sigma_2=0.26$ $G_{23}=1.74$ $\sigma_3=0.32$ $G_{24}=1.81$ $\sigma_4=0.38$			$G_{21}=1.61$ $\sigma_1=0.15$ $G_{22}=1.64$ $\sigma_2=0.19$ $G_{23}=1.67$ $\sigma_3=0.24$ $G_{24}=1.71$ $\sigma_4=0.28$			$G_{21}=1.59$ $\sigma_1=0.10$ $G_{22}=1.60$ $\sigma_2=0.13$ $G_{23}=1.62$ $\sigma_3=0.17$ $G_{24}=1.65$ $\sigma_4=0.20$		
n	K	EACP	n	K	EACP	n	K	EACP
12	4.0	88.2	12	4.3	89.6	12	4.6	89.0
18	6.0*	93.0	18	6.5*	93.0	18	6.9*	91.4
22	7.3	92.4	22	7.9	93.0	24	9.1	92.4
28	9.3	93.4	28	10.1	94.2	30	11.4	92.6
34	11.3	94.0	34	12.2	91.8	36	13.7**	96.0
40	13.3	92.2	40	14.4	92.6	42	16.0	95.4
45	14.9**	95.6	46	16.6	92.2	49	18.7	93.8
51	16.9	94.4	52	18.7**	96.0	54	20.6	96.0
58	19.2	94.8	58	20.9	92.4	60	22.9	93.4
62	20.6	93.2	62	22.3	93.8	67	25.5	93.8
68	22.6	92.4	68	24.5	94.6	72	27.4	95.4
74	24.5	96.2	74	26.6	96.6	79	30.1	95.6
80	26.5	94.6	80	28.8	94.2	86	32.8	94.0
85	28.2	94.4	86	30.6	92.4	91	34.7	95.6
91	30.2	95.2	91	32.8	94.8	97	37.0	93.6
96	31.8	95.4	98	35.3	94.2	104	39.6	96.0
102	33.8	92.2	103	37.1	94.8	109	41.6	94.4
107	35.5	93.6	108	38.9	96.2	115	43.8	94.2
113	37.5	96.8	114	41.0	93.6	122	46.5	93.8
120	39.8	93.6	120	43.2	96.6	127	48.4	97.8
125	41.5	95.2	125	45.0	96.0	133	50.7	93.6
131	43.5	94.8	131	47.2	92.4	140	53.4	95.2
135	44.8	93.4	138	49.7	93.8	146	55.7	93.2
142	47.1	94.4	143	51.5	96.4			
147	48.8	95.0	148	53.3	97.0			

Table 8: Survey on Exports and Imports of Services, 1980-84.

		Exports	Imports	Exports./. Imports	Sample size (excl take- all strata)
1980	g_2^2	89	276	202	1 676
	k	18.8	6.1	8.3	
1981	g_2^2	376	129	351	1 516
	k	4.0	11.8	4.3	
1982	g_2^2	61	93	72	1 771
	k	29.0	19.0	24.6	
1983	g_2^2	61	135	116	1 641
	k	26.9	12.2	14.1	
1984	g_2^2	169	91	132	1 803
	k	10.7	19.8	13.7	

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie, som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

Nummer	Titel (författare)
1988:1 (beige)	Abstracts I - Sammanfattningar av metodrapporter från SCB

Kvarvarande BEIGE och GRÖNA exemplar av ovanstående promemorior kan rekvireras från Elisabet Klingberg, U/STM, SCB, 115 81 Stockholm, eller per telefon 08-7834178.

Dito GULA exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 Stockholm, eller per telefon 08-7834147.