The Comparison of Sampling Precision under the Cost Fixed Condition

Pang Zhiqiang[1,3] , Wang Lixia[2]
[1]School of statistics Lanzhou University of Finance and Economics,Lanzhou ,Gansu, China
[2]School of statistics Lanzhou University of Finance and Economics,Lanzhou ,Gansu, China
[3] Corresponding author: Pang Zhiqiang, e-mail:pangzhq@lzcc.edu.cn

Abstracts

The basic criterion of determine optimum sampling scheme should be that achieving the highest precision in the approved cost range, or achieving the required accuracy by the lowest cost. However in practices, we expect for maximum accuracy through the sampling design under the costs fixed condition commonly. This paper researches the sampling accuracy of probability sampling when the costs fixed. At the same time, this paper gives the conditions which some complex sampling designs are more effective than simple random sampling.

Key words: cost function; covariance function; probability sampling; sampling precision

The cost and accuracy of sampling survey are involved in the whole sampling survey process. The level of cost and accuracy influence the effect and function of the sampling work. In fact, improve sampling accuracy need to increase the sampling cost. And the costs of survey will increase rapidly when the accuracy reached the critical level. Therefore, researchers are often forced to balance repeatedly between higher sampling accuracy and low sampling cost. They pursued that reach the highest precision in the approved cost range, or at the lowest cost satisfied the accuracy requirement. In the actual problem, one of the more common situations is the cost of a given condition, through the different sampling design in order to achieve the maximization of accuracy. This paper is based on the established cost; compare different probability sampling design precision size, looking forward to for the reality of choice to provide a new way to judge.

To convenient purpose, this paper regards simple estimation of means as premise. The sampling accuracy is described by variance of estimator. The foundation of comparison is simple random sampling. We denote the simple random sampling cost function as follow linear function: $C_{srs} = c_0 + cn$. Where, c_0 is the fixed cost unrelated with sample size, $c$is unit average cost of survey, n is the sample size.

1.Accuracy comparison between stratified sampling and simple random sampling

1.1 "unit" cost of proportion of stratified sampling

When the sampling cost mainly concentrated in each unit, the cost function of proportional stratified sampling is

$$C_{prop1} = c_0 + \sum_{h=1}^{L} c_h n_h = c_0 + \sum_{h=1}^{L} c_h n w_h = c_0 + n \sum_{h=1}^{L} c_h w_h$$

Where,$C_{prop1}$is the cost function of proportional stratified sampling, $c_0$is the fix cost unrelated with sample size (organization propaganda cost, prepare sampling frame cost, stratified expenses, etc.),$n_h$is the sample size of each strata, $c_h$is the unit average cost of survey in $h$strata, $n$is the number of the sample unit,w_h the weighted of each strata. To make the costs of proportional stratified sampling and simple random sampling equals,

$$n\sum_{h=1}^{L} c_h w_h = c' n'$$

thus

$$n = \frac{c' n'}{\sum_{h=1}^{L} c_h w_h}$$

Arrangestratified sampling variance function with this this formula:

$$V_{prop}(\overline{y_{st}}) = \frac{1-f}{n}\sum_{h=1}^{L} w_h {S_h}^2 = \frac{1-f}{n'} S^2 \frac{\sum_{h=1}^{L} c_h w_h}{c'} \frac{\sum_{h=1}^{L} w_h {S_h}^2}{S^2}$$

$$= V(\overline{y_{srs}}) \frac{\sum_{h=1}^{L} c_h w_h}{c'} deff_{prop1} = V(\overline{y_{srs}}) \frac{deff_{prop1} c}{c'}$$

Where, $c = \sum_{h=1}^{L} c_h w_h$is theunit of average cost of each strata.

Obviously,$V_{prop}(\overline{y_{st}}) = V(\overline{y_{srs}})$when$deff_{prop1} c = c'$, $V_{prop}(\overline{y_{st}}) < V(\overline{y_{srs}})$ when $deff_{prop1} c < c'$,;$V_{prop}(\overline{y_{st}}) > V(\overline{y_{srs}})$ when $deff_{prop1} c > c'$,.

1.2 "between unit" cost of proportion of stratified sampling

When sampling cost mainly concentrated in each unit of travel cost, the cost function of proportional stratified random sampling is

$$C_{prop2} = c_0 + \sum_{h=1}^{L} t_h \sqrt{n_h} = c_0 + \sum_{h=1}^{L} t_h \sqrt{n w_h}$$

where,$C_{prop2}$is the cost function of proportional stratified random sampling of, $c_0$is the fixed cost unrelated sample size (organization propaganda cost, prepare sampling frame cost, stratified expenses, etc.),$n_h$is thesample size of each strata,$t_h$is average travel expenses of unit per strata, $n$is number of sample unit,$w_h$is the weighted of strata. At this time, if we think the costs of proportional stratified sampling and simple random sampling costs same, that is,

$$\sqrt{n}\sum_{h=1}^{L} t_h \sqrt{w_h} = c' n'$$

then

$$n = \left(\frac{c' n'}{\sum_{h=1}^{L} t_h \sqrt{w_h}}\right)^2$$

Arrange stratified sampling variance function with this this formula:

$$V_{prop}(\overline{y_{st}}) = \frac{1-f}{n}\sum_{h=1}^{L} w_h {S_h}^2 = \frac{1-f}{n'} S^2 \frac{\left(\sum_{h=1}^{L} t_h \sqrt{w_h}\right)^2}{n' c'^2} \frac{\sum_{h=1}^{L} w_h {S_h}^2}{S^2}$$

$$= V(\overline{y_{srs}}) \frac{\left(\sum_{h=1}^{L} t_h \sqrt{w_h}\right)^2}{n' c'^2} deff_{prop2} = V(\overline{y_{srs}}) \frac{deff_{prop2} t^2}{n' c'^2}$$

where, $t = \sum_{h=1}^{L} t_h \sqrt{w_h}$is the unit of average cost of each stratum. (weighted by$\sqrt{w_h}$).

Obviously, $V_{prop}(\overline{y_{st}}) = V(\overline{y_{srs}})$ when $deff_{prop2} t^2 = n' c'^2$; $V_{prop}(\overline{y_{st}}) < V(\overline{y_{srs}})$when$deff_{prop2} t^2 < n' c'^2$;$V_{prop}(\overline{y_{st}}) > V(\overline{y_{srs}})$when $deff_{prop2} t^2 > n' c'^2$,.

2.Accuracy comparison between systematic sampling and simple random sampling

Systematic sampling and simple random sampling are very similar.Usually the cost function of systematic sampling is:

$$C_{sy} = c_0 + cn$$

Where,$C_{sy}$is total cost of systematic sampling,$c_0$is fixed cost (organization propaganda cost, prepare sampling frame cost etc.),nis sample size, cisthe unit of average cost of each strata. (investigators wages, survey test cost, travel cost , etc.). At this time,if we think the costs of the system sampling and simple stratified random sampling costs same, namely

$$cn = c'n'$$

then

$$n = \frac{c'n'}{c}$$

2.1 "Cluster" systematic sampling

We can regard cluster samplingas a special kind of systematic sampling. The estimator varianceof population mean can be calculated by correlation coefficient ρ_{wsy} of "system sample" (group) when group size is equal.Arrange stratified sampling variance function with $n = \frac{c'n'}{c}$:

$$V(\overline{y_{sy}}) = \frac{N-1}{N}\frac{S^2}{n}\left[1 + (n-1)\rho_{wsy}\right]$$

$$= \frac{N-1}{N}\frac{S^2}{\frac{c'n'}{c}}\left[1 + (n-1)\rho_{wsy}\right]$$

$$= \frac{N-n'}{N}\frac{S^2}{n'}\frac{N-1}{N-n'}\frac{c}{c'}\left[1 + (n-1)\rho_{wsy}\right]$$

$$= V(\overline{y_{srs}})\frac{N-1}{N-n'}\frac{c}{c'}\left[1 + (n-1)\rho_{wsy}\right]$$

Obviously,when

$c(N-1)\left[1 + (n-1)\rho_{wsy}\right] = c'(N-n')$, $V(\overline{y_{sy}}) = V(\overline{y_{srs}})$;when $c(N-1)\left[1 + (n-1)\rho_{wsy}\right] < c'(N-n')$, $V(\overline{y_{sy}}) < V(\overline{y_{srs}})$;when $c(N-1)\left[1 + (n-1)\rho_{wsy}\right] > c'(N-n')$,$V(\overline{y_{sy}}) > V(\overline{y_{srs}})$.

2.2 "Stratification" systematic sampling

We can regard systematic sampling as a special kind of stratified sampling. The estimator variance of population mean can be calculated by correlation coefficientρ_{wst} of population (strata) when cluster size is equal. Arrange stratified sampling variance function withn $= \frac{c'n'}{c}$:

$$V(\overline{y_{sy}}) = \frac{1-f}{n} S_{wst}{}^2[1+(n-1)\rho_{wst}] = \frac{1-f}{n'} S^2 \frac{c}{c'} \frac{S_{wst}{}^2[1+(n-1)\rho_{wst}]}{S^2}$$

$$= V(\overline{y_{srs}}) \frac{c}{c'} \mathrm{deff}_{sy}$$

Obviously, when $\mathrm{deff}_{sy} c = c'$, $V(\overline{y_{sy}}) = V(\overline{y_{srs}})$;when $\mathrm{deff}_{sy} c < c'$, $V(\overline{y_{sy}}) < V(\overline{y_{srs}})$;when $\mathrm{deff}_{sy} c > c'$,$V(\overline{y_{sy}}) > V(\overline{y_{srs}})$.

3.Accuracy comparison between cluster sampling and simple random sampling

For more convenient, generally the cost functionof cluster samplingis $C_{cs1} = c_0 + nc_1 + nMc_2$, where, $C_{cs1}$is the total cost of cluster sampling, $c_0$is fixed fee (organization propaganda cost, prepare sampling frame cost, etc.), $c_1$is the average group survey fees (investigators wages, survey please, travel, etc.), $c_2$is he average unit cost of the investigation (investigators wages, survey please, travel, etc.), $n$isunit number of PSU, M is the size of SSU. At this time, to make the whole group sampling and simple stratified random sampling costs the same, i.e:

$$nc_1 + nMc_2 = c'n'$$

thus

$$n = \frac{c'n'}{c_1 + Mc_2}$$

Will this generation into the whole group sampling variance function arrangement:

$$V(\overline{\overline{y_{cs}}}) = \frac{1-f}{n} \frac{(NM-1)}{M^2(N-1)} S^2[1+(M-1)\rho_c]$$

$$\approx \frac{1-f}{nM} S^2[1+(M-1)\rho_c]$$

$$= \frac{1-f}{\frac{c'n'}{c_1+Mc_2} M} S^2[1+(M-1)\rho_c]$$

$$= \frac{1-f}{n'} S^2 \frac{c_1 + Mc_2}{c'M} [1+(M-1)\rho_c]$$

$$= V(\overline{y_{srs}}) \frac{c_1 + Mc_2}{c'M} [1+(M-1)\rho_c]$$

Obviously,when $(c_1 + Mc_2)[1+(M-1)\rho_c] = c'M$, $V(\overline{\overline{y_{cs}}}) = V(\overline{y_{srs}})$;when $(c_1 + Mc_2)[1+(M-1)\rho_c] < c'M$, $V(\overline{\overline{y_{cs}}}) < V(\overline{y_{srs}})$;when $(c_1 + Mc_2)[1+(M-1)\rho_c] > c'M$,$V(\overline{\overline{y_{cs}}}) > V(\overline{y_{srs}})$.

4.Accuracy comparison between multistage sampling and simple random sampling

In order to facilitate comparison, assuming overall is very big, the sampling ratio is small, the calculated variance of finite population correction coefficient right now but no. Second order (section) sampling in the first stage of the overall variance can be seen as between group variance, the second phase of the variance can be seen as within group variance, so the total variance for combined, namely$S^2 \approx S_1{}^2 + S_2{}^2$. So, when the second order (section) sampling in the first phase extraction n a primary unit and the second stage in the pumping unit extraction m a secondary unit, the total sample size for nm. Visible multistage sampling is the whole sampling process is divided into several stages, a phase one stage on sampling to complete the sampling process, this kind of sampling that which is order sampling, thus it can be seen, multistage sampling and cluster sampling, the situation is similar, we refer to the definition of smoke cluster sampling, the multistage sampling cost function is$C_{tss1} = c_0 + nc_1 + nmc_2$, among them, C_{tss1} for two stage sampling of total cost, c_0 for fixed cost, c_1 every survey a first-order unit costs (organization propaganda cost, prepare sampling frame and other fees), c_2 every survey a second order unit costs (investigators wage, travel, etc.).At this time, to make the multistage sampling and simple stratified random sampling costs the same, namely

$$nc_1 + nmc_2 = c'n'$$

thus

$$n = \frac{c'n'}{c_1 + mc_2}$$

As the research is convenient for us in this study only when a order unit equal and two phase of the sampling are simple random sampling, and because in the sample size phase and design effect for $deff_{tss} = \frac{mS_1{}^2 + S_2{}^2}{S^2}$, it is:

$$V(\overline{\overline{y_{tss1}}}) - V(\overline{y_{srs}}) = \frac{1}{n}S_1{}^2 + \frac{1}{mn}S_2{}^2 - \frac{1}{n'}S^2$$
$$= \frac{mS_1{}^2 + S_2{}^2}{mn} - \frac{mS_1{}^2 + S_2{}^2}{n'deff_{tss}}$$
$$= \frac{(n'deff_{tss} - mn)(mS_1{}^2 + S_2{}^2)}{mnn'deff_{tss}}$$
$$= \frac{[deff_{tss}(c_1 + mc_2) - mc']n'(mS_1{}^2 + S_2{}^2)}{mnn'deff_{tss}(c_1 + mc_2)}$$

Obviously, when $deff_{tss}(c_1 + mc_2) = mc'$, $V(\overline{\overline{y_{tss1}}}) = V(\overline{y_{srs}})$;when $deff_{tss}(c_1 + mc_2) < mc'$,$V(\overline{\overline{y_{tss1}}}) < V(\overline{y_{srs}})$;when$deff_{tss}(c_1 + mc_2) > mc'$,$V(\overline{\overline{y_{tss1}}}) > V(\overline{y_{srs}})$.

5.Conclusion

We can draw the following conclusion from the above analysis.

（1）The relationship of survey accuracy between proportional stratified random sampling and simple random sampling survey depends on relationship, which between the ration of unit average cost and design effect of stratify sampling, when cost fixed and cost of the investigation mainly concentrated in each unit. Only when both the ratio of not less than the design effect, proportional stratified random sampling to more accurate than simple random sampling; When the cost of the investigation mainly concentrated in each unit of travel between use, proportional stratified random sampling and simple random sampling survey accuracy, mainly depends on the simple random sampling of $n'c'^2$ and stratified random sampling t^2（$t = \sum_{h=1}^{L} t_h\sqrt{w_h}$ for each layer by each unit share of the average travel expenses, here to quality $\sqrt{w_h}$ for weight) than with proportional stratified random sampling design effect relationship, only when both the ratio of not less than the design effect, proportional stratified random sampling to more accurate than simple random sampling.

（2）The relationship of survey accuracy between system sampling and simple random samplingdepends on the relations between the $c(N-1)\left[1+(n-1)\rho_{wsy}\right]$and $c'(N-n')$ when the cost fixed and variance of population mean can be calculated by correlation coefficient ρ_{wsy}. Only the current is not greater than the latter, the system sampling to than simple random sampling to more accurate; When the system sampling ensemble average estimator of the variance in general (layer) in correlation coefficient ρ_{wst} said, the system sampling and simple random sampling survey precision mainly depends on the relationship between simple random sampling, each survey a unit of the average cost and system sampling in each survey a unit of the ratio of the average cost with systematic sampling design effect relationship, only when the ratio of not less than systematic sampling design effect, the system sampling to more accurate than simple random sampling.

（3）The relationship of survey accuracy between cluster sampling and simple random sampling survey depends on relationship, which between the ration of unit average cost of simple random sampling and to deleverage cost of cluster sampling (the average cost of the investigation and each unit the contribution of each group of the sum of the average cost (i.e.$c_1 + Mc_2$) with the ratio of $\frac{1+(M-1)\rho_c}{M}$ relationship), when cost fixed. Only when the ratio is not less than$\frac{1+(M-1)\rho_c}{M}$, the cluster sampling is more accurate.

（4）The relationship of survey accuracy between tow stage sampling with equal probabilityand simple random sampling survey depends on relationship, which between the ratioof unit average cost of simple random sampling and total average cost of SSU in tow stage sampling with equal probability(each survey a second order unit average cost and each second order unit have allocated first-order unit is the sum of the average cost of the investigation $c_1 + mc_2$ and m ratio) and $deff_{tss}$, when cost fixed. Only when the $\frac{c'}{\frac{c_1+mc_2}{m}}$not less than$deff_{tss}$, the cluster sampling is more accurate.

References:

[1] the W.G.C ochran, ZhangYaoTing, WuHuiYi. Sampling technology [M]. Beijing: China statistics press, 1985:359-363

[2], KishL the NiJiaXun translation. Sampling technology [M]. Beijing: China statistics press, 1997:342-348

[3] theDuZiFang. Sampling technique and its application [M]. Beijing: tsinghua university press, 2005.8:259-263

[4] ZhiJiang, MrZhangPengCheng. Accuracy when certain various sampling design investigation cost comparison [J].journal of statistics and information BBS, 2008.7:5-8

[5] ZhangPengCheng. Sampling design investigation cost comparative study [D]. Lanzhou, lanzhou business school, 2009.4