The Philippine Statistician
Volume XVI, Nos. 3-4, 1967

RATIO ESTIMATORS IN FIELD RESEARCH

by
BURTON T. OÑATE[1]

1. **Introduction.** A sampling survey may be considered as an absolute experiment where the main objectives are to obtain estimators of parameters and to derive measures of precision of these estimators. In comparative experiments, one of the objectives is to test statistical hypotheses using the appropriate estimation procedures. Thus, it is apparent that simple and sound estimation techniques must be developed to provide precise statistics in the conduct of field research whether such endeavors are of the absolute or the comparative types, or both.

The purpose of this paper is to point out the theoretical framework of ratio estimators and their uses and applications to field research. The problem may involve the estimation of grain yield from experimental or paddy fields, the estimation of incidence of stem borer infestation and the use of these estimators in replicated field experiment, or the estimation from multi-stage samples of demographic and/or socio economic characteristics from finite but rather large universes.

2. **Theoretical Framework.** If from a universe of size N, a random sample of size n is drawn, and if the following pairs of characteristics are observed (X_1, Z_1), (X_2, Z_2), ..., and (X_n, Z_n) for the purpose of estimating the parameters of X such as the total T(X) and the mean X, then we can consider three general types of estimators for these parameters. The estimators are:

[1] Statistical Consultant, Asian Development Bank.

1. the X-only estimate,

2. the biased ratio estimate, and

3. the unbiased ratio estimate.

The variances of these estimators will be given and the corresponding estimators of these variances will be used to show the gain in efficiency in the use of ratio estimators as compared to the usual X-only estimate [1], [2], [3], [4], [7], [9]. These estimators of the population total $T(X)$ and their respective variances are shown in Table 1 where the population constants are:

N is the size of the universe,

Z is the population total of the concomitant variable which is easily obtained and is available at no cost or at a relatively low cost,

$S^2(Z_i) = \sum_{i=1}^{N} (Z_i - \bar{Z})^2/(N-1)$ is the population variance of the Z_i's (similar formula for the X_i's),

$S(X_i, Z_i) = \sum_{i=1}^{N} (X_i - \bar{X})(Z_i - \bar{Z})/(N-1)$ is the covariance of (X_i, Z_i),

$Q = \bar{X}/\bar{Z} = X/Y$ is the ratio of population means or population totals,

$R = \sum_{i=1}^{N} r_i /N$ is the population mean of ratios where $r_i = X_i /Z_i$,

with the sample statistics are

$\bar{q} = \bar{x}/\bar{z}$ is the ratio of sample means,

$\bar{r} = \sum_{i=1}^{n} r_i/n$ is the mean of ratios,

for the ratio of means to be more precise than the X-only estimate.

In developing economics, one of the fundamental problems is to bridge the gap between theory and application, between theory and empirical results which could increase the accuracy of research studies.

3. Empirical Results.

By and large, to establish the precision of estimate, we will compare the estimated variances from a given sample n. The choice of the concomitant variable will depend upon the correlation it has with the characteristics under study and on the ease and economy of observing this variable. Empirical results will be given for the estimation of grain yield in experimental and paddy fields and also the estimation of stem borer incidence in experimental fields. The use of ratio estimators in national sample surveys will be given in Section C.

The results for grain yield are shown in Tables 2, 3, and 4. The concomitant variable chosen to estimate grain yield (X_j) was the whole panicle weight (Z_i).

Table 2. RELATIVE EFFICIENCY, IN PERCENTAGE, OF RATIO ESTIMATORS FOR GRAIN YIELD OF SIX VARIETIES, IRRI, 1963.

Type of estimators	Variety					
	1	2	3	4	5	6
X-only	100	100	100	100	100	100
Ratio of means	21300	13650	32300	3421	856	3386
Mean of ratio	21300	13650	32300	3283	852	3386
Umbiased ratio	< 21300	< 13650	< 32300	< 3283	< 852	< 3386

Table 3. COEFFICIENT OF VARIATION AND CORRELATION BETWEEN GRAIN WEIGHT (X_i) AND WHOLE PANICLE WEIGHT (Z_i), IRRI, 1963.

Variety	Sample size n	$cv(X_i)$	$cv(Z_i)$	$(1/2)cv(Z_i)/cv(X_i)$	$\hat{\rho}(X_i,Z_i)$
1	15	17.9	17.5	.489	0.999
2	14	13.0	12.7	.488	0.998
3	25	22.6	21.6	.478	0.999
4	34	39.9	37.1	.465	0.987
5	34	29.2	29.5	.505	0.942
6	36	25.6	26.2	.510	0.986

Table 4. UPPER BOUND OF RELATIVE BIAS OF RATIO ESTIMATORS FOR GRAIN YIELD OF SIX VARIETIES, IRRI, 1963.

Variety	Sample size n	Upper bound of relative bias (percentage)	
		Ratio of means ($\bar{q}$)	Means of ratios (r)
1	15	5	68
2	14	3	47
3	25	4	108
4	34	6	216
5	34	5	172
6	36	4	157

Not shown in Table 3 are the results for another variety where a covariance analysis was attempted to obtain the within-plot correlation $\hat{\rho}_w = 0.997$. This correlation brought about a gain in statistical efficiency of 16,000 percent for the ratio of means estimate over that of the X-only estimator. The analysis of covariance also was utilized in sampling studies in replicated rice experiments. The objective was to establish the level of the correlation at each stage of sampling for different types of cultural and management practices.

Results from an experiment involving the removal or incorporation of straw are shown in Table 5. There were four treatments in four replications arranged in randomized blocks.

From each plot, the structure of the sampling scheme is as follows:

six rows/plot
five hills/row
three panicles/hill.

Table 5. ANALYSIS OF COVARIANCE FOR GRAIN YIELD (X_i) AND PANICLE WEIGHT (Z_i), IRRI, 1963.

Source of variation	Degrees of freedom	Mean squares			Correlation coefficient $\hat{(\rho)}$
		x^2	xz	z^2	
Total	*1400*				
Replicate (R)	3	48.34	53.14	59.20	0.993
Treatment (T)	3	7.80	6.88	6.09	.999
R x T	9	8.22	8.77	10.64	.938
Rows w/in plot	80	6.16	6.52	7.27	.975
Hills w/in rows	384	3.12	3.13	3.21	.990
Panicles w/in hills	921	1.70	1.73	1.79	.994

The level of correlation is high at all stages of sampling and this result indicates considerable gain in statistical efficiency in the use of ratio estimators at any stage. Also, estimates from the analysis of variance will be used to obtain optimum combinations of number of panicles, hills, rows or sub-rows, etc. for a comparable but acceptable level of precision for estimate of mean or total. Comparable levels of correlation also were observed in five other replicated experiments.

It is assumed that whole panicle weight (Z_a, Z_b or Z_c, etc.) is more easily obtained than grain weight and there is little or no loss in harvesting. In fact, the actual grain yield of the paddy field T(X) may be considerably different than the observed grain yield because of losses incurred in harvesting, threshing, winnowing, etc., so that the ratio estiators will be more precise estimates of [T(X)] than the observed grain yield. In crop cutting experiments, it is important to consider

the losses incurred by the crop between the time of actual crop cuts and the time of harvest. Generally, there will be two different biological universes. Thus, the estimate of the standing crop from crop cutting must be adjusted because of losses in harvesting so that the estimate will refer to the biological universe at time of harvest[8]. Precise estimates of production (kgm./Ha.) is in itself not useful without precise estimate of hectarage. A multi-stage and multi-phase sampling design with an interview phase must be devised in order to provide other necessary information.

In the incidence of stem borer study, the problems was to estimate the ratio rather than the total. Thus, our estimate of variance will be for rates or ratios, i.e.,

$$\text{var } (\bar{q}) = [\ (N-n)/Nn(n-1)\bar{z}^2] \ [\ \Sigma' x_i^2 + \bar{q}^2\ \Sigma x_i z_i\ -2\bar{q}\ \Sigma x_i z_i]$$

while the var $(\bar{r})$ will be similar as that for $\bar{q}$ except that we replace $\bar{q}$ by $\bar{r}$ in the var (q). The comparison between the levels of $\bar{r}$ and $\bar{q}$ for 20 varieties out of a total of 245 varieties tested and their estimated variances indicates that $\bar{q}$ is lightly more efficient that $\bar{r}$. It is important to repeat at this point that is if we are interested in

$$R = \sum_{i}^{N} r_i/N$$

then our unbiased estimators is $\bar{r} = \Sigma'(X_i/Z_i)/n$ and the appropriate estimate of variance, Var $(\bar{r})$, is

$$s^2(\bar{r}) = \Sigma'(r_i-\bar{r})^2/(n-1)n\ .$$

However, if the population constant is $Q = X/Z$, then the better estimator is $\bar{q}$. In this comparative experiment, $\bar{q}$ was used as conceptually we are interested in $Q = X/Z$.

A problem arises when one uses the results on a hill basis, $(r_i = X_i/Z_j)$ in the analysis of variance. On a hill basis $r_i = q_i$. One can, however, use a collection of hills or clusters in order to derive a $\bar{q} = \bar{x}/\bar{z}$ for the cluster ratios of means. In the study of incidence of stem borer, the size of the cluster of vairance will be for rates or ratios, i.e.,

was n = 5. The optimum combinations of number of replications (r), number of cluster (c) and number of hills or plants per cluster (p) for desired levels of precision in the sampling for stem borer incidence are given by the author[10].

The analysis of variance technique used by author[10] may be used to estimate incidence of stem borer in a farmer's paddy field or in a bigger area such as a barrio, village, region, or country. Yield data from crop-cutting or other sources may be used with these data on incidence to develop a ratio or regresson estimate on the probable losses resulting from incidence of stem borer. A properly designed sampling survey is needed in order to evolve this relationship. Experimental results must be used to check the validity of the estimates[5]. In practical application, a considerable area is usually the subject of the estimate. A multi-stage sampling technique with complete replacement of the primaries may be attempted. This technique will require simple estimation procedures.

4. **Ratio Estimators in Statistical Surveys.** Several multistage sample designs utilize the ratio estimators. One disadvantage is the added complexity of the estimation procedures. are at a premium. Priority must be given to a change in the estimation procedure rather than in a change of design. One of the techniques which the author [7, 9] considered was the multi-stage estimator $\hat{X}_{hi}$ as the X-variable and the multi-stage estimator $\hat{Z}_{hi}$ as the Z-variable. Thus, instead of the usual observational pairs (X_i, Z_i), we have the multi-stage estimators $(\hat{X}_{hi}, \hat{Z}_{hi})$. The theoretical framework given in Section A was utilized and applied to the Philippine Statistical Survey of Households but with some changes in the concept of paired observations.

Briefly, the Philippine Statistical Survey of Households (PSSH) consists of a division of the country into four sectors, namely: (a) Metropolitan Manila, (b) chartered cities and provincial capitals, (c) poblacions, and (d) the barrios. Sector (a) and (b) comprised the urban area, (c) the urban-rural, and (d) the purely rural area. Attemps presently are being

made to make sharper distinctions between rural and urban areas[6]. The urban area was stratified using geographical criteria, while the rural area was dissected through the use of paper strata. In the urban areas, equal probability was used, while probability proportional to size (pps) was used in the rural areas. In both areas sampling was done with complete replacement of primaries. A description of the PSSH is shown in Table 6. Figure 1 illustrates the sampling technique used in the PSSH. The national significance of the use of ratio estimators in the design of the PSSH will be given for Metropolitan Manila. Population count is usually highly or moderately correlated with other important socio-economic variables in a given area. Thus, the population in listed households will be used as the sample variable and this variate will generate the X-component ($\hat{X}_i$) while the number of registered voters by precinct is the Z-component ($\hat{Z}_i$) . The number of registered voters is available every 2 years by precinct, by stratum and by the whole universe from the usual administrative channels with little or no cost to the survey. Empirical results for November, 1957, and November, 1959 are shown in Table 7 for the six types of estimators studied.

Table 7. STATISTICAL EFFICIENCY OF SIX TYPES OF ESTIMATORS FOR POPULATION IN LISTED HOUSEHOLDS: METROPOLITAN MANILA, 1957 1959.*

Type of estimator	Relative efficiency, percentage 1957[a]		1959[b]	
	separate	combined	separate	combined
Regular unbiased (X-only)				
Separate	100	—	100	—
Combined	—	100	—	100
Biased ratio of means				
Separate	127	—	105	—
Combined	—	135	—	148
Unbiased ratio				
Separate	125	—	105	—
Combined	—	136	—	148

[a]/ Seven strata were used.
[b]/ Ten strata were used.
* Source of basic data: Philippine Statistical Survey of Households and Philippine Electoral Commission.

Estimate of correlation, $\hat{\rho}(\hat{X}_i, \hat{Z}_i)$ from covariance analysis ranged from 0.4 to 0.5 and these correlations established gains in precision from 5 to 27 percent for the separate ratio estimators and from 36 to 48 percent for the combined ratio estimators as compared to the regular unbiased (X-only) estimates.

Three forms of ratio estimators may be obtained, one of which is applicable to any given stratum. These forms are as follows:

(a) Complete ratio estimation. The precincts are divided into segments which are known, recognizable, and identical for both electoral list and survey list. Street names or city blocks may be used to identify segments in precincts with electoral list for the precinct. Consider as estimate of X (the stratum total)

$$\approx\!\!\!\!\!\!\sim\ \bar{X}^{**} = \approx\!\!\!\!\!\!\sim\ \bar{q}\, Z$$

where

$$\approx\!\!\!\!\!\!\sim\ \bar{q} = [(M/m)\,\Sigma'_i (S_i/s)\ \Sigma'_j X_{ij}] \,/\, [(M/m)\ \Sigma'_i (S_i/s)\ \Sigma'_j Z_{ij}]$$

X_{ij} is the listed segment total,

Z_{ij} is the corresponding electoral list total for the segment,

Z is the electoral list total for the stratum, and

s is the number of segments (segmentation of precinct is done for possible rotation of sample).

(b) Ratio estimation in the primary stage. In this case the segment totals X_{ij} $(j = 1, 2, 3)$ are available for the survey data only, while for the electoral counts Z only precinct (psu) totals, Z_i $(i = 1, 2, \ldots, M)$, are available. The ratio estimator has the form

$$\hat{\bar{X}}^{**} = \hat{\bar{q}}\, Z$$

where

$$\hat{\hat{q}} = [(M/m) \sum_i' (S_i/s) \sum_j 'X_{ij}] \; / \; [(M/m) \sum_i 'Z_i] \; .$$

(c) No segmentation. The precinct is completely listed by the survey. We have available at the precint or psu level, the psu totals X_i $(i = 1, 2, 3, 4, 5)$ and the psu electoral counts, Z_i $(i = 1, 2, \ldots, M)$.
Let the estimator be

$$\hat{\hat{X}}^{**} = q \; Z$$

where

$$q = [\,(M/m) \sum_i' X_i\,] \; / \; [\,(M/m) \sum_i' Z_i\,] = \hat{\hat{X}}^{*}/\hat{Z}^{*}.$$

This is the theoretical framework which was used to obtain the statistical efficiency of the six types of estimator given in Table 7.

One may attempt to estimate from the sample household stage. It should be pointed out that this technique is too difficult to implement in the field.

5. **Summary and Conclusions.** This paper discussed the theoretical results for ratio estimators and how the results may be extended to multi-stage sampling on a national basis. Empirical results show the versatility and usefulness of ratio estimators in comparative and absolute experiments. Generally, ratio estimators will have lower variance or will provide more precise estimates than the X-only estimate. To be of practical use, these ratio estimators must be derived with little or no cost to the experiment or survey.

Results strongly indicate that whole panicle weight (Z_i) can be used to estimate grain (X_i) at a higher level of precision than be attained by using grain yield (X_i) alone at various levels or stages of sampling.

Estimates of stem borer incidence from the analyses of variance can be used to increase precision of estimates for larger

areas than experiments on paddy fields. Estimates of both incidence and yield from crop-cuts can be used in a ratio or regression estimate in order to have a yield loss equation.

On a national or regional level, the significance of the use of simple ratio estimators as applied to the Philippine Statistical Survey of Households becomes more apparent in view of the following:

a) sampling was with complete replacement of PSU, and the reduction in variance of the X-only estimate due to application of the finite population correction (1-f) is lost where f is the sampling fraction,

b) the PSU's were drawn with equal probability and not with probability proportional to size (PPS).

While conditions (a) and (b) will result in rather simplified estimating procedures which are of advantage when applied in developing economics, considerable losses in precision may exist if the X-only estimate is used. The use of ratio estimators generally will recover these losses in statistical efficiency with little or not cost to the survey proper. On the average, there is a gain of about 28 per cent precision in the use of ratio estimators. This implies that we can reduce the number of PSU's by 28 percent and still maintain the level of precision of the X-only estimate. Peso-wise, this reduction in the number of PSU, will also result in the reduction of cost of the survey by about 28 percent. Four forms of ratio estimators may be employed.

In general, biological data vary less than socio-economic data. Also, the level of correlation observed in the former is higher ($\hat{\rho} \doteq 0.9$) than the latter ($\rho \doteq 0.5$). The estimates for ρ given in this paper illustrate this level of difference.

We also can state that the use of ratio estimators is an excellent example of an increase in the precision of estimates through the introduction of new estimation procedures rather than by a change in the design.

LITERATURE CITED

[1] Cochran, W. G. *Sampling Techniques.* New York: John Wiley and Sons, Inc., 1953. pp. 111-159.

[2] Goodman, L. A. and Hartley, H. O. "The precision of unbiased ratio-type estimators." Jour. Amer. Stat. Assoc. 53. (1958) pp. 491-508.

[3] Hartley, H. O. *Theory of Advanced Design of Survey.* (Mimeographed Lecture Notes). Dept. of Statistics. Iowa State University. 1959. pp. 20-30.

[4] ————————, and Ross, A. "Unbiased ratio estimators"; *Nature* 174 (1954). pp. 270-271.

[5] Ishikura, H. "Short review on the recent works on the estimation of loss of rice crops by insect pests in Japan". IRC. 1957.

[6] Nazaret, F. V. and F. R. Barreto. "Concepts and definitions of urban-rural areas in the Philippines." Read at the 11th annual conference: Philippine Statistical Assoc. 1963.

[7] Añate, B. T. *Development of multi-stage designs for statistical* surveys in the Philippines. Multi-11th Series No. 3. Stat. Lab., Iowa State University. 1960. pp. 57-62.

[8] ————————. *Crop-cutting experiments.* Proceedings of the Seminar on Sampling and Sample Surveys. Statistical Center, University of the Philippines. March 1962. pp. 170-182.

[9] ————————. "Ratio estimators in multi-stage designs." The *Philippine Statistician* 11 (2). (June, 1962.) pp. 58-67.

[10] ————————. *Statistics in Rice Research* (Mimeographed Lecture Notes). International Rice Research Institute. 1963. Chapter V.

Table 1.

THREE TYPES OF ESTIMATORS AND THEIR VARIANCES.

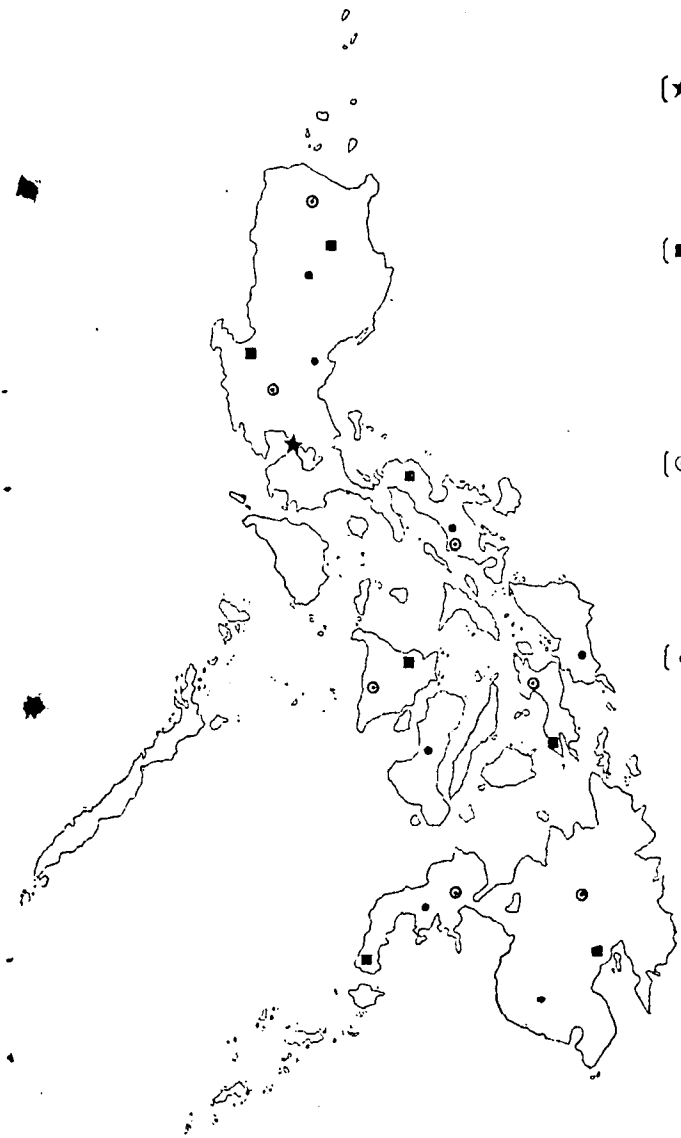
Type of estimator	Estimator of population total, $T(X)$	Variance of estimate	Estimate of variance
X-only	$\bar{T}_X = N\bar{x}$	$\sigma^2(\bar{T}_X) = (N\text{-}n)N\ S^2(X_i)/n$	$s^2(\bar{T}_X) = (N\text{-}n)N\ s^2(X_i)/n$
Biased ratio Ratio of means	$\tilde{T}_Q = \bar{q}Z$	$\sigma^2(\tilde{T}_Q) = (N\text{-}n)N$ $(S^2(X_i) + Q^2S^2(Z_i)$ $-2Q\ S(X_i, Z_i))/n$	$s^2\ (\tilde{T}_Q) = (N\text{-}n)N\ (s^2(X_i)$ $+\bar{q}^2s^2(Z_i)$ $-2\bar{q}\ s(X_i, Z_i))/n$
Mean of ratios	$T_R^* = \bar{r}Z$	$\sigma^2(T_R^*) = (N\text{-}n)N$ $(S^2(X_i)+R^2S^2(Z_i)$ $-2RS(X_i, Z_i))/n$	$s^2(T_R^*) = (N\text{-}n)N(s^2(X_i)$ $+\bar{r}^2s^2(Z_i)\ -2\bar{r}s(X_i, Z_i))/n$
Unbiased ratio	$T_U' = \bar{r}Z+(N\text{-}l)$ $(n/n\text{-}l)\ (\bar{x}\text{-}\bar{r}\bar{z})$	$\sigma^2(T_U') = (\sigma^2(T_R^*)$ $+$ negligible correction term)	$s^2(T_U') = (s^2(T_R^*)$ $+$ correction)
		or $= (\sigma^2(\tilde{T}_Q) +$ correction)	or $= (s^2(\tilde{T}_Q) +$ correction)

Table 6. DESCRIPTION OF THE PHILIPPINE STATISTICAL SURVEY OF HOUSEHOLDS

Sectors	Type of sampling	Primary sampling unit (PSU)	Number of strata	Number of precincts, municipalities or barrios	Number of sample households
Urban[a]					
Metropolitan Manila	Equal probability with complete replacement of PSU	precinct	32	160 (precincts)	800
Provincial capital and chartered cities	— ditto —	— ditto —	30	150 (precincts)	1050
Urban-rural[b]					
Poblacions	Unequal probability with complete replacement of PSU	municipality	30	150 (municipalities)	1500
Rural					
Barrios	— ditto —	— ditto —	30	300	3000

[a]/ two stages of sampling.
[b]/ three stages of sampling.

SAMPLING AND ESTIMATION PROCEDURES IN THE BUREAU OF THE CENSUS AND STATISTICS STATISTICAL SURVEY OF HOUSEHOLDS



ORDERED PSU's
(ORDERED ESTIMATORS)

(★) METROPOLITAN MANILA — P_{1M} P_{2M} P_{3M} P_{4M} P_{5M}

(■) PROVINCIAL CAPITALS *and/or* CHARTERED CITIES — P_{1P} P_{2P} P_{3P} P_{4P} P_{5P}

(⊙) POBLACIONS — M_{1P} M_{2P} M_{3P} M_{4P} M_{5P}

(•) BARRIOS — M_{1B} M_{2B} M_{3B} M_{4B} M_{5B}

PHILIPPINES (SAMPLE ESTIMATORS) — $\hat{X}_1$ $\hat{X}_2$ $\hat{X}_3$ $\hat{X}_4$ $\hat{X}_5$

(CONCOMITANT ESTIMATORS) — $\hat{Z}_1$ $\hat{Z}_2$ $\hat{Z}_3$ $\hat{Z}_4$ $\hat{Z}_5$