CENSUS OF INDIA, 1951

SAMPLING STUDIES

by

U. SIVARAMAN NAIR, M. A., Ph. D., F. A. Sc., F. N. I.,

Superintendent of Census Operations,

Travancore-Cochin.

PRINTED BY THE SUPERINTENDENT OF GOVT. PRESSES
AT THE GOVT. CENTRAL PRESS, TRIVANDRUM.
1953.

PREFATORY NOTE

In organising the abstraction and compilation of data collected during the 1951 Census of India, a special arrangement was made to select one out of every ten slips on a strictly random basis. The main object of this arrangement was to secure a wider range of tabulations than was possible, within the prescribed limits of time and cost, on the total count. It was also the intention that material should be collected at this Census whereby one may settle the size and structure of the sample which may be regarded as adequate for purposes of Census tabulation in future.

Shri R. A. Gopalaswami, the Registrar-General, discussed with me the lines of organising this study and he entrusted me with the work. The discussions led him to define a suitable yard stick for measuring the degree of approximation of the Sample Value to Total Count Value. The Index of Approximation, thus introduced, is explained in the appendices to the introduction.

I have had very liberal help and unsparing criticism from Shri R. A. Gopalaswami. His letters to me have been most thought-provoking. I am grateful to him for his guidance in this study.

The large mass of data analysed in this work has come from the various tabulation offices in India. The Superintendents of Census have been certainly put to great inconvenience in the despatch of relevant material and that at a time when they were frightfully busy with the preparation of their final tables. For their co-operation in this all-India study, I tender to them my very sincere thanks.

Dr. K. Bhaskara Varma was my collegue in organising this study. The calculations have been heavy and I am happy to record that he took great pains to check at each stage the work done by the computing staff under him. He has placed me under a deep sense of obligation.

The printing of the elaborate tables in this volume has been a heavy responsibility on the Superintendent of Government Presses. The scrupulous care and speed with which the work was done at each stage of printing go to the great credit of the large band of earnest workers in the Press. To the Superintendent and his staff, I extend my sincere thanks.

In presenting this volume, I have strictly followed Shri R. A. Gopalaswami's advice—"analyse and tabulate observed results in such a manner that all may study them and make their own researches." I shall feel happy if this purpose is served.

U. S. NAIR

CONTENTS

KEY TO STATE TABLES

A - State B - Division C - District

Name of State	Table I			Table II			Table III			Table IV	
	A	B	C	A	B	C	A	B	C	A	B
1. Uttar Pradesh ..	1	1	2	63	64	69	221	222	227	413	414
2. Orissa ..	11	11	11	90	91	92	253	254	256	419	420
3. West Bengal ..	14	14	14	98	99	100	263	264	266	422	423
4. Assam ..	17	17	17	106	107	108	274	275	277	425	426
5. Manipur, Tripura, Sikkim...	20	..	..	114	..	..	285	..	..	428	..
6. Madras ..	21	21	21	117	118	121	288	289	293	431	432
7. Mysore ..	26	..	26	132	..	133	306	..	307	436	..
8. Travancore-Cochin ..	28	..	28	137	..	138	312	..	313	437	..
9. Coorg ..	29	..	..	140	..	..	315	..	..	438	..
10. Bombay ..	30	30	31	141	142	145	316	317	322	439	440
11. Saurashtra ..	36	..	36	156	..	157	336	..	337	..	...
12. Kutch ..	37	..	..	159	..	..	340	..	..	446	..
13. Madhya Pradesh ..	38	38	38	160	161	163	341	342	345	447	448
14. Madhya Bharat ..	43	43	43	178	179	181	356	357	360	451	452
15. Bhopal ..	47	..	47	188	..	189	368	..	369	455	..
16. Vindhya Pradesh ..	48	..	48	190	..	191	370	..	371	456	..
17. Rajasthan ..	50	50	50	194	195	197	375	376	380	457	458
18. Punjab ..	55	55	55	208	209	210	393	394	396	462	463
19. Pepsu ..	57	..	57	212	..	213	..	..	..	465	..
20. Ajmer ..	59	..	..	216	..	..	..	..	..	466	..
21. Delhi ..	60	..	..	217	..	..	..	..	..	467	..
22. Himachal Pradesh ..	61	..	..	218	..	219	..	..	..	468	..

INTRODUCTORY NOTE

I have no faith in anything short of actual Measurement and Rule of Three.

CHARLES DARWIN

1. Sampling has not played any prominent part in Indian Censuses prior to 1941. At that Census, as a measure of war economy tabulation was restricted in the former British Provinces and small States to the classification of the total population by 'communities'. However to secure information that might be required later, Mr. Yeatts, the Census Commissioner arranged for the extraction and preservation of a two per cent sample of the original census slips. These sample slips were subsequently used to construct the age tables for the 1941 census.

At the 1951 census a number of sampling experiments have been tried. A verification of the accuracy of census enumeration by a sample check was undertaken for the first time at this census. The size of the sample was one-in-one -thousand of the total number of households. The frame for this survey was the National Register of Citizens which is a list of all the households and the inmates in each, according to the 1951 census. The results of this enquiry are contained in the Census of India Paper No. 1, of 1953.

A second sample survey was conducted in most of the States in India to verify the accuracy of registration of births and deaths. This experimental census of births and deaths was intented to be the first step towards improvement of population data and was confined to a one per cent sample of households in the districts selected for the survey. The births and deaths in the households were recorded and verified with the entries made in the official birth and death registers.

Besides these surveys, a detailed sampling experiment on population data was conducted in all the states in India at the time of sorting the enumeration slips of the 1951 census in the various tabulation offices. Instructions were given by the Registrar-General to extract a ten per cent sample of slips for each village and municipal ward. The procedure was to break open enumeration pads, one by one, shuffle and cut as in a game of cards and deal into ten pigeon holes, the slips of displaced persons being placed in a eleventh hole*. After all the pads belonging to a village or ward were subject to the above process, the slips in the pigeon holes were collected in three lots—

i. the slips of displaced persons
ii. the slips in the fifth pigeon hole
iii. the slips in the remaining nine pigeon holes

The slips in (*ii*) formed the ten per cent sample.

The three lots of slips were kept separate at all stages of sorting. The tables prepared from the slips in (*ii*) give the results of the sampling experiment while those obtained by consolidating the tables from (*ii*) and (*iii*) contain the results of total count.

Ordinarily, in sampling studies, the results of complete enumeration are unknown. Statistical techniques have been developed to estimate the unknown values for the total population from the known values in the sample. The object of the present study is not to prepare estimates of unknowns—in fact, the results of complete enumeration are known—but to assemble the sample values and the total count values and indicate the extent of agreement of divergence between them.

2. The samples were drawn in the various Tabulation Offices and sorted for (*a*) sex and (*b*) eight livelihood classes for each village and ward. Fuller sorting was done after combining the slips of villages for larger geographical areas called Tracts, each tract having a population of nearly two lakhs. The results are contained in the various tables published as Part II of the Census Report for each State. It

* A displaced person was defined as a person who has entered India, having left or being compelled to leave his or her home in Western Pakistan on or after the 1st March 1947, or his/her home in Eastern Pakistan on or after the 15th October 1946 on account of civil disturbances or the fear of such disturbances or on account of the setting up of the two dominions of India and Pakistan.

Displaced persons have been excluded in this study.

was however decided to select the following population characteristics for sampling study.

1. Sex Composition
2. Livelihood classes
3. Livelihood classes and dependency
4. Employment in Industries and Services

The tables for each district for these four characteristics were made available for the present study.* This data correspond to an approximately ten per cent sample for the districts.

Samples with smaller sampling fraction were also studied. These samples were derived by pooling together the slips of :-

(*a*) Every tenth tract (selection being made with a random start) in each district

(*b*) Every tenth village (with a random start) in all the tracts

(*c*) Every tenth village of every tenth tract. The four sample types thus evolved are:—

(*i*) S.1— Approximately ten per cent sample

(*ii*) S.01— Approximately ten per cent 'thick' sample formed by pooling the ten per cent samples of every tenth tract

(*iii*) S. '01— Approximately ten per cent 'thin' sample formed by pooling the ten per cent samples of every tenth village in all the tracts.

(*iv*) S.001— Approximately one-in- thousand sample formed by pooling the samples of every tenth village of every tenth tract.

The district samples were combined to give corresponding samples for larger areas within the State—called Divisions—and the division samples were combined to give the state samples. The population characteristics, sample types and geographical areas included in this study are as follows:—

Population characteristic	Sample Type *State*	*Division*	*District*
Sex composition ..	All the four sample types		
Livelihood classes ..	All the four sample types		S.1 & S.01
Livelihood classes and dependency ..	S.1 and S.01		S.1
Employment in industries and services ..	S.1 and S.01		Nil

*The data for Bihar, Hyderabad and Andaman Nicobar Islands were not available.

3. As has been mentioned at the outset, the object of this study is to assemble the results of complete enumeration and of sampling and give an indication of the agreement between them. The number per ten thousand has been calculated from the data for all the population characteristics. The proportions based on complete enumeration are called total count values and those from the samples, sample values. To judge the agreement between the total count values and sample values, the percentage error and the coefficient of variation have also been calculated for each proportion. Theoretically the percentage error in large samples, will be less than thrice the coefficient of variation almost always. Thus the sample value is in agreement with the total count value if the corresponding percentage error is less than thrice the coefficient of variation. As an illustration the sample value and the total count value are in agreement in the following cases :—

i. Percentage error is 1.0 and the coefficient of variation is 0.5.

ii. Percentage error is 10 and the coefficient of variation is 5.

Clearly, the sample value in (*i*) is a much better estimate than that in (*ii*). To bring out this distinction, a new index called the Index of Approximation has been introduced.

'A' will be defined as the index of approximation of a sample value if the percentage error exceeds half A in A out of 100 cases utmost. Low values of A correspond to good approximation; the approximation becomes coarse as A increases. The theory of the index of approximation is given in the appendices.

4. To represent the large volume of results for the various population characteristics, sample types and geographical areas, without loss of any relevant information, the following details are given for each sample value:—

i. Sample size and sampling fraction
ii. The total count value and the sample value
iii. The percentage error
iv. The index of approximation
v. The coefficient of variation

There are four series of tables corresponding to the four population characteristics. These are

Table I. Sex Composition
Table II. Livelihood Classes
Table III. Livelihood Classes and Dependency
Table IV. Employment in Industries and Services

Each Table consists of three parts—A for the State, B for the Divisions and C for the Districts.

APPENDIX I

Degrees and Indices of Approximation

R. A. GOPALASWAMI, I.C.S.,

Registrar-General, India.

I

1. Let the following assumptions be made—

(*i*) (*a*) There are n different population groups comprising $P_1, P_2, \ldots P_n$ individuals respectively.

(*b*) From each population group, a sample is drawn on a strictly random basis using the same sampling fraction. Each of these samples comprises $S_1, S_2, \ldots S_n$ individuals respectively.

(*ii*) (*a*) Each population group and the corresponding sample are stratified according to different characteristics.

(*b*) For each of the population groups, the number of individuals having a specified characteristic is ascertained and the proportion which this number bears to the total number of individuals in the population group determined, as the value of the specified *characteristic proportion* for that population group. The values thus determined are found to be $\xi_1, \xi_2, \ldots \xi_n$.

(*c*) The values of the same *characteristic proportion* are also determined similarly for the population group samples. These are found to be $p_1, p_2, \ldots p_n$.

2. It is necessary to define in precise terms, the degree of approximation of the set of values p_i to ξ_i.

Three degrees of approximation are defined as follows—

(*a*) *Very Close Approximation*—if the number of cases in which p_i exceeds or falls short of ξ_i by more than one per cent, does not exceed .2 per cent of n.

(*b*) *Close Approximation*—where the approximation is not very close, if the number of cases in which p_i exceeds or falls short of ξ_i by more than 2.5 per cent does not exceed 5 per cent of n.

(*c*) *Fair Approximation*—where the approximation is not very close nor even close, if the number of cases in which p_i exceeds or falls short of ξ_i by more than 5 per cent does not exceed 10 per cent of n.

3. Generalising the foregoing, it is possible to define an *index of approximation* between the set of values p_i and the corresponding set ξ_i as follows—

The index of approximation is A if the number of cases in which p_i exceeds or falls short of ξ_i by more than $A/2$ per cent is as nearly as may be, equal to A per cent of n.

II

1. A somewhat different conception of the *degrees and indices of approximation* may be based on the following assumptions.

(*i*) (*a*) There is only one population group comprising P individuals.

(*b*) From this group n different samples are drawn on a strictly random basis, each sample consisting of the same number of individuals say S. (n is large).

(*ii*) (*a*) The value of a specified *characteristic proportion* is determined for the entire population group and found to be ξ.

(*b*) The values of the same *characteristic proportion* are determined for all the n samples and found to be $p_1, p_2, \ldots p_n$ respectively.

2. The index of approximation of any of the values p to ξ is A if the number of cases in which p_i exceeds or falls short of ξ by more than $A/2$ per cent is, as nearly as may be, equal to A per cent of n.

3. The different degrees of approximation may be defined as below:—

(*a*) *Very Close Approximation*—where A does not exceed 2.

(*b*) *Close Approximation*—where A exceeds 2 but does not exceed 5.

(*c*) *Fair Approximation*—where A exceeds 5 but does not exceed 10.

APPENDIX II

The Index of Approximation

1. Consider a population group of size N and a random sample of size n from it. The population group and the sample are stratified into k strata each stratum corresponding to a particular characteristic of the population.

$X_1, X_2, \ldots X_k$ and $x_1, x_2, \ldots x_k$ are the numbers in the strata in the population and sample respectively.

$$\varrho_i = \frac{X_i}{N} \text{ and } p_i = \frac{x_i}{n}$$

are the proportions of the ith characteristic in the population and sample respectively. p_i is an estimate of ϱ_i. The suffix will be omitted if there is no need to specify the stratum.

The error of estimation may be measured in terms of

$$(a) \quad \varepsilon = p - \varrho \quad \text{(absolute error)}$$

$$\text{and } (b) \quad \delta = \frac{p-\varrho}{\varrho} 100 \quad \text{(percentage error)}$$

2. In practical situations, the acceptance of p as an estimate of ϱ depends on a high probablity for a small error. The acceptance specification reduces to,

$$\left.\begin{aligned} &\text{(i)} \quad |\delta| \leq \mu \\ &\text{(ii)} \quad P[|\delta| \leq \mu] \geq \beta \end{aligned}\right\} \qquad (1)$$

μ determins the probability α that the percentage error is numerically less than μ. If α is greater than β the conditions in (1) are satisfied and p is considered to be in agreement with ϱ. If α is smaller than β the conditions of acceptance (1) are violated and p is not in agreement with ϱ. The two parameters μ and β that specify the conditions of acceptability of p will vary with the requiremencs of the problem studied and therefore the specification itself is arbitrary.

A scale of measurement of agreement of p with ϱ can be defined by placing some restriction on μ and β in (1). One way of connecting μ and β is to assume that

$$\mu = \frac{A}{2} \text{ and } \beta = 1 - \frac{A}{100}.$$

Then percentage error exceeds $\frac{A}{2}$ numerically in A out of 100 cases. A will be called the index of approximation of p.

2. The index of approximation A is defined by the relation

$$P\left(-\frac{A}{2} \leq \delta \leq \frac{A}{2}\right) = 1 - \frac{A}{100} \qquad (2)$$

In the case when the sample size n is large, the normal approximation for binomial probability leads to elegant mathematical formula for A. Let

$$Z = \frac{p-\varrho}{\sqrt{\frac{\varrho(1-\varrho)}{n}}} = \frac{\delta}{100}\sqrt{\frac{n\varrho}{1-\varrho}} = \frac{\delta}{\gamma} \qquad (3)$$

In (3), $\gamma = 100\sqrt{\frac{1-\varrho}{n\varrho}}$ is the coefficient of variation of p. δ/γ follows the normal distribution and

$$P\left\{-\frac{A}{2} \leq \delta \leq \frac{A}{2}\right\} = \sqrt{\frac{2}{\pi}} \int_{0}^{\frac{A}{2\gamma}} e^{-\frac{x^2}{2}} dx$$

Thus A is defined by

$$\sqrt{\frac{2}{\pi}} \int_{\frac{A}{2\gamma}}^{\alpha} e^{\frac{x^2}{2}} dx = \frac{A}{100} \qquad (4)$$

The calculation of A from (4) is not possible unless the value of γ is known. In a sample, p alone is known and the formula for A has to be modified to give the index of approximation knowing the sample proportion p. Denoting this by $\mathbf{a}$

$$P\left\{\frac{\mathbf{a}}{2} \leq \delta \leq \frac{\mathbf{a}}{2}\text{; given } P\right\} = 1 - \frac{\mathbf{a}}{100} \qquad 4{\cdot}1$$

The distinction between A and $\mathbf{a}$ has to be noted. A is the index of approximation in a sample from a population with an unknown proportion ϱ while $\mathbf{a}$ is the index in a sample giving a proportion p. To evaluate (4.1) given p,

$$\varrho = \frac{100\,p}{100+\delta}$$

and $$\gamma = 100\sqrt{\frac{100\,q+\delta}{100\,np}}$$

$$Z = \frac{\delta}{\gamma} = \frac{\delta}{100}\sqrt{\frac{100\,np}{100\,q+\delta}}$$

where $q = 1-p$

Since Z monotonically increases with δ, $-\frac{\mathbf{a}}{2} \leq \delta \leq \frac{\mathbf{a}}{2}$ is equivalent to

$$-\frac{\mathbf{a}}{200}\sqrt{\frac{np}{q-\frac{\mathbf{a}}{200}}} \leq Z \leq \frac{\mathbf{a}}{200}\sqrt{\frac{np}{q+\frac{\mathbf{a}}{200}}}$$

Hence $\mathbf{a}$ is defined by

$$\frac{1}{\sqrt{2\pi}} \int_{-\frac{\mathbf{a}}{200}\sqrt{\frac{np}{q-\frac{\mathbf{a}}{200}}}}^{\frac{\mathbf{a}}{200}\sqrt{\frac{np}{q+\frac{\mathbf{a}}{200}}}} e^{-\frac{x^2}{2}} dx = 1 - \frac{\mathbf{a}}{100} \qquad 4{\cdot}2$$

Denoting the sample coefficient of variation by v,

$$v = 100\sqrt{\frac{q}{np}}$$

$$\text{or} \quad P = \frac{10000}{10000 + nv^2} \text{ and}$$

$$q = \frac{nv^2}{10000 + nv^2}$$

Substituting these in (4.2),

$$\frac{1}{\sqrt{2\pi}} \int_{-\frac{a}{2\sqrt{v^2\left(1+\frac{a}{200}\right)+\frac{50a}{n}}}}^{\frac{a}{2\sqrt{v^2\left(1+\frac{a}{200}\right)-\frac{50a}{n}}}} e^{-\frac{x^2}{2}}\, dx = 1 - \frac{a}{100} \qquad 4.3$$

When n is large, $\frac{50a}{n}$ may be neglected and (4.3) reduces to

$$\sqrt{\frac{2}{\pi}} \int_{-\frac{a}{2v\sqrt{1+\frac{a}{200}}}}^{\alpha} e^{-\frac{x^2}{2}}\, dx = \frac{a}{100} \qquad 4.4$$

(4.4) is independent of n and involves only the sample coefficient of variation. Thus in large samples, though n and p may vary, the index of approximation remains constant if the sample co- efficient of variation does not change.

3. In large sample theory, the acceptance limits of a sample value are defined by the t—sigma rule where t depends on the value of the probability for acceptance. This rule leads to limits for ϱ when p is known or for p when ϱ is known. The limits based on the t—sigma rule, are

$$\frac{2np + t^2 - \sqrt{t^2 + 4npq}}{2(n+t^2)} \leq \varrho \leq \frac{2np + t^2 + \sqrt{t^2 + 4npq}}{2(n+t^2)} \quad \text{(for } \varrho\text{)} \qquad (5)$$

$$\text{and} \quad \varrho - t\sqrt{\frac{\varrho(1-\varrho)}{n}} \leq p \leq \varrho + t\sqrt{\frac{\varrho(1-\varrho)}{n}} \quad \text{(for } p\text{)}, \qquad (6)$$

Keeping only terms of the order of $\frac{1}{n}$, (5) becomes

$$p\left(1-\frac{vt}{100}\right)+\frac{t^2}{2n}(q-p)\leq\varrho\leq q\left(1+\frac{vt}{100}\right)+\frac{t^2}{2n}(q-p)$$

If **a** is the index of approximation of p, the limits of p for probability $1-\frac{\mathbf{a}}{100}$ are

$$\frac{p}{1+\frac{\mathbf{a}}{200}}\leq\varrho\leq\frac{p}{1-\frac{\mathbf{a}}{200}} \qquad (7)$$

In (5) if t corresponds to probability $1-\frac{\mathbf{a}}{100}$, even though the limits for ϱ are determined, the per- centage error is not directly known, whereas in (7), the limits and percentage are explicitly indicated.

In statistical investigations it is usual to take 0.99, 0.95 or 0.90 as the probability levels in the t—sigma rule; the approximations defined as 'very close', 'close', and 'fair', in appendix (1) relate to probabilities greater than 0.98, 0.95 to 0.98, and 0.90 to 0.95. Thus the probability levels 0.95 and 0.90 in the t—sigma rule correspond to 'at least close' and 'at least fair' approximations.