OCCASION

This publication has been made available to the public on the occasion of the 50th anniversary of the United Nations Industrial Development Organisation.

DISCLAIMER

FAIR USE POLICY

CONTACT

Please contact publications@unido.org for further information concerning UNIDO publications.

For more information about UNIDO, please visit us at www.unido.org

UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION
Vienna International Centre, P.O. Box 300, 1400 Vienna, Austria
Tel: (+43-1) 26026-0 · www.unido.org · unido@unido.org

Consultant Jack Gross
Backup Mr. Ballance
PSD/STAT

DRAFT

18827

AUG 81 1983

INDUSTRIA.

Sampling in The Framework of an Industrial Census.

1. Introduction: Chapter overview

   1.1 Primary goals of an industrial census

   1.2 Objections to a complete canvass

       Expensive, inefficient, and inaccurate operation

       Exhaustive detail unnecessary and poor allocation of resources

   1.3 Use of sampling in an industrial census

       For supplementary information

       Instead of a complete canvass

2. Fundamental concepts of sampling

   2.1 Definitions: A sample, sampling, sample design, sample estimate

   2.2 Probability sampling

   2.3 Cut-off sampling

3. Sampling frames

   3.1 The directory: source, content, applications, deficiencies

   3.2 Geographic area sampling frames

       Complete coverage

       Well defined, recognizable segments, EA maps

       Descriptive information available: number of establishments by size and industry

       Updating operations considered

4. Cut-off sampling for supplementary information

   4.1 Long forms versus short forms

   4.2 Factors favoring cut-off sampling

   4.3 Estimating universe totals

DRAFT

5. The canvass for basic statistics: complete canvass, cut-off sample or probability sample?

5.1 Publication specifications for basic statistics

Considerable detail (industry, geography, size)

Consistent detail and totals

Estimates of measurable quality

Quality that is related to importance

5.2 Objections to cut-off sampling for basic statistics

5.3 Objections to a complete canvass

5.4 Probability sampling: Satisfies all the publication specifications

Estimates by any desired classifications

Consistency within and between tables

Measurable quality

Quality related to importance

5.5 Administrative advantages of probability sampling

Reduces cost and workload

Less dependent on inexperienced (temporary) staff

More careful processing; better control of reporting errors and processing errors

Improves distribution of respondent burden and census budget

6. Formal mathematical properties of probability sampling

6.1 Definitions of universe values (parameters): N, X, $\bar{X}$, S^2, V^2

6.2 The foundations of probability sampling: Simple random sampling

6.3 Estimates of totals from a simple random sample

Simple unbiased estimates of totals, X'

Relation of X' to X

$E(X')$, $S^2(X')$, $V^2(X')$, $V(X')$. Normal distribution of $X'-X$.

6.4 Relationship of X' to X for other probability sample designs:

$$X'_p = \sum_{i=1}^{n} X_i/P_i,\ EX'_p = X,\ V^2(X'_p) = (\text{design effect})\ V^2(X'_r).$$

7. Frequently used probability sample designs

7.1 Stratified random sampling

Selection and estimation procedures

Purpose and effects

Optimum allocation

7.2 Random systematic sampling

Description of procedure, operational simplicity; control problems

Benefit of ordering

7.3 Probability cluster sampling

Definitions and examples of clusters, ultimate clusters

Effects of clustering: Homogeneity and cluster size; design effect

Cost advantage of cluster sampling

7.4 Sampling with probability proportional to measures of size

Variable p_i but retain chance

Advantages over constant p

Use in combination with stratified, systematic and cluster sampling, and in multi-stage sampling

7.5 Sampling with or without replacement

7.6 Other probability sampling designs

8. A uniform rule for selecting the quality of the estimates to their importance: Quality proportional to the square root of importance.

8.1 Specific definitions: Economic importance = size of labor force for category

Quality = inverse of relative sampling error, $1/V(X'_c)$

Formal relationship = $K/V(X'_c) = X_c$, or $V(X'_c) = K/X_c$

Assign desired quality level by choice of K.

8.2 Comparison of recommended rule with other rules for specifying quality.

Recommended rule: Quality changes systematically but slowly with changing importance

Proportional rule: Changes appear too abrupt

Constant quality for a given publication level (e.g. major industry group by region)

Can demand far more for economically negligible categories than for major ones

Ambiguous for some classes of estimates.

Exceptions to the general rule:

For special conditions adjust size measure to reflect importance

Cost differentials - similarly adjust sizes by factors

8.3 Mathematical properties that support the rule

Consistent for independent estimates and their sums

The rule conforms, approximately, with the conditions for maximizing the efficiency of the sample design

8.4 Approximations in seeking to optimize the sample design

Exactly optimum design involves unknown universe values

Necessary to accept some approximations

Central approximation: All independently sampled groups have the same relative standard deviation, $V_g=V$

Assumption is generally satisfactory

Efficiency curve is flat in region of optimum

9. The certainty class

Necessary to include all large establishments with certainty

Choice of a lower certainty cut-off for sampling from the directory

For protection against understatement of some sizes

To improve efficiency

10. Sampling jointly from the directory and the geographic frame

10.1 Different strategies adopted for urban and rural samples

Urban: Directory sample supplemented by geographic sample

Rural: Geographic sample only

Cost considerations influence on choice of strategies

10.2 Stratification of the urban samples

Directory: 5-9 and 1-4 persons engaged

Geographic: EAs with 5 or more establishments and EAs with less than 5 establishments

10.3 The rural stratum

10.4 Allocation of the sample to strata

Formula for approximately optimum allocation

Relative influence of MOS, design effect and unit costs

10.5 Estimates of number of establishments and persons engaged by strata

Available data

Reasonable speculations regarding unknown factors

10.6 Design effects and average unit costs

Design effect = 1 for stratified samples from directory

Design effect = $1+(n_h-1)\delta_h$ for geographic samples

Conservative speculative value of δ_h

Calculation of n_h for each of those geographic strata

Unit costs: factors that entered calculation for each stratum

10.7 Optimum number of sample reports per stratum

Effect of different quality specifications, K, on the allocation and cost of the sample.

Practical application of the theory, using the numerical values previously derived for each stratum

11. Selecting the samples

11.1 Features common to all strata: Information for checking and control purposes

11.2 Procedures for sampling from the directory frame

Arrangement of record file

Detailed steps for random systematic selection

11.3 Procedures for sampling from the geographic, urban-low stratum

11.4 Sampling from the geographic, urban-high and rural strata

Reason for sampling with PPMOS from these strata

Procedure for systematic sampling with PPMOS

12. Estimating totals, sampling variances and relative standard errors

12.1 Requirements prescribed for estimates of totals (unbiased and additive) and of sampling variances (reasonable, for few selected items)

12.2 Estimates of totals

Formula

Properties

12.3 Estimating sampling variances and relative standard errors

Sampling unit totals needed

Software packages for computing variance estimates

Alternative Simple Approximation

Computational advantages

Bias of formula

INDUSTRIA

Sampling

1. Introduction

This chapter discusses the use of sampling methods for data collection purposes in the 1993 Industrial Census. It explains why sampling is preferable to a complete canvass, the particular purposes sampling can serve, the basic concepts involved and the principal methods pertinent to an industrial census (or survey). The chapter also deals with the issue of defining quality standards, the specific steps involved in planning and selecting the sample and developing estimates for the 1993 Industrial Census, and some alternative approaches that also should be considered.

1.1 Goals of an Industrial Census

An industrial census has two primary goals: A comprehensive body of accurate statistics which describe the industrial activity of the country in considerable detail, and a good foundation for later, more limited industrial surveys. The first goal, traditionally, has been defined as calling for the compilation and publication of highly detailed tables which present basic measures of industrial activity cross-classified by industry, geography and establishment size, and less detailed breakdowns of secondary, more specialized

statistics. The second goal calls for developing a list of units and their characteristics (industry, location, size) from which selections can be made as needed for particular surveys following the census. Traditionally, in hope of satisfying the goals, efforts were made to collect a census report from every industrial establishment in the country.

1.2 Objections to a Complete Canvass

It is questionable whether efforts to collect reports from all the industrial establishments in the country are justifiable. They demand that an exhaustive canvass, similar in scope to that of a population census, be conducted. Such a canvass is extremely expensive. It requires hiring a large staff of inexperienced people to collect and to process the reports, which implies that the project will be inefficiently conducted. Completion is slow. Moreover, the hoped for accuracy is more theoretical than achievable. Errors in coverage, reporting and processing inevitably occur and diminish the accuracy of the results.

Valid questions can also be raised as to the need for the elaborate detail the complete census promises. Is it really important to know that eleven people were engaged in baking in one group of rural villages in Yendi province, and nine

in the next group of villages in the same province? By the time these figures become available they may well have changed by at least two or three persons in each case. Would it not be sufficient to know that the numbers were quite small?

One additional objection to a full industrial census must be noted. It is one of the most serious criticisms. Industrial activity is highly concentrated. For example, in Industria, the largest manufacturing establishments, less than 11 percent of the total number, employ 65 percent of all persons engaged in manufacturing. The bottom end of the size distribution, the establishments having fewer than ten persons engaged, include nearly 65 percent of all the manufacturing establishments, but account for less than 10 percent of the manufacturing labor force. A census program which treats all establishments alike would expend most of its funds on those small establishments. This would be so not only because of their disproportionate numbers, but also because it would cost more to collect and to process their individual reports. In some countries, most large establishments would respond well to a mail canvass. A high percentage of the small establishments would require personal visits to get reports from them. Additionally, more complete and more accurate reports (based on good business records) can be expected from the large

establishments than from the small ones, and correcting reporting deficiencies can prove expensive. The combination of relatively large numbers and high unit costs for the small establishments would result in an extremely unbalanced effort. A very large majority of the census budget would be devoted to the economically least important data.

1.3 Use of Sampling in an Industrial Census

Instead of treating all establishments alike, sampling can be used to improve the balance between expenditures and importance. It can be applied in two ways:

(i) Reports can be collected from only some rather than all establishments.

(ii) Information on supplementary topics - all except the basic topics - can be collected from only some of the establishments included in the canvass.

In an industrial statistics program both procedures can be applied to reduce the attention small establishments get, and thereby give the large ones proportionally more. That is what we decided to do in the current industrial census. How we went about it and the reasons behind the procedural choices we made are discussed in this chapter. We begin by defining some fundamental terms.

2. Fundamental Concepts of Sampling

Everyone has a general understanding of the terms sampling and a sample. Perhaps less familiar is the associated term sample design. All of these need to be explicitly defined. In our context they have the following meanings:

Sampling is the act of selecting a set of n units from a finite universe of N units.*

A sample is a particular set of n units selected from the N.*

The term sample design means the plan and method used to select a sample. Since estimates for the entire universe of N will be developed from the sample, the sample design also covers the method to be used to develop those estimates. Note that we have implicitly defined a fourth term:
The sample estimate - the estimate of a universe value, as derived from the sample values.

The concept of a sample design is crucial, for a given design may yield many different samples, and some of the

---

* These are not the most general definitions of sampling and a sample. They do not cover, for example, biological or other experimental studies which involve samples from infinitely large universes. The definitions are appropriate for our situation.

samples produced by different designs can be identical. Certain designs will yield a single unique sample.

We cannot determine from any given sample whether it is good or bad, i.e. how closely its estimates correspond to the universe values. If we know the sample design, however, we may be able to infer how its samples behave. We may be able to calculate, objectively, the likelihood that the particular sample estimate and its corresponding universe value differ by specified amounts.

The ability to evaluate the quality of its estimates is a great strength of probability sample designs. The essential characteristics of such designs are that chance determines which units are selected for any sample, that every unit in the universe has a positive probability of being selected and that the probabilities are known. Thus a particular sample obtained when using a probability design is one of a number of different samples that might occur.

Designs which can yield only one unique sample have very different characteristics. They limit the selection to units that have specified properties. Cut-off sampling designs are the most important example. Under that method all establishments larger than a specified size, and no others, are selected for the sample. How good its estimates

might be can only be judged subjectively, unlike the case for probability sampling.

Despite the disadvantage of non-measurable quality, cut-off sampling has a role in the industrial census program, as has probability sampling. Both methods require a sampling frame: lists of units that (analogously to a complete canvass) cover all the industrial establishments in the universe. We consider next, therefore, the subject of sampling frames.

3. Sampling Frames

The National Statistical Office (NSO) has a directory of industrial establishments. Its primary source is the 1983 Industrial Census. The census list has been updated to some extent by later surveys conducted by the NSO and by registration records of the Ministry of Trade and Business.

The Ministry of Trade and Business requires all new businesses and all purchasers of existing businesses to register with it. The ministry provides, annually, copies of the registration documents to the NSO.

The NSO does not maintain lists of establishments that supply electric power or water. National and local governmental agencies operate all the establishments engaged in those activities and can provide complete lists as needed.

A directory provides several potential benefits in conducting an industrial census. The names and addresses that tell where to go for reports can be imprinted on the report forms in advance, along with NSO assigned control numbers. They also offer the possibility for collecting reports by mail. Industry codes and size codes in the directory enable the NSO to select the appropriate report form to imprint for each establishment. Check-in control files can be established before the canvass begins.

All the above mentioned benefits of a directory apply whether a complete canvass or sample coverage is contemplated. As a frame for sampling, the directory has the additional advantageous feature that its records can be sorted into groups according to distinguishing characteristics, notably, size, industry and geographic location.

The existing directory, however, has a number of deficiencies. It omits a considerable number of industrial establishments. Conversely it includes some out-of-business establishments, and it carries the wrong owner's name, industry code or size code in some other cases.

Of these deficiencies, the undercoverage is the most serious. It is believed that records have been obtained for virtually all of the large new businesses, but that the coverage of small new businesses is weak. There are also doubts concerning the completeness of the small size class in the last industrial census. Additional omissions occur because some businesses have changed their operations from non-industrial to industrial. A census which relies on the directory's coverage alone would not satisfy the objectives of providing a detailed description of the country's industrial activities, and good lists that can be drawn on for future surveys.

The directory's coding errors and its inclusion of out-of-business establishments lessen it efficiency as a frame for conducting the census. Both types of errors, however, can be corrected in the course of conducting the canvass at some extra cost. Unlike them, the coverage gap of the directory can be overcome only by use of a geographic area frame.

The concept of a geographic area frame is quite simple. It calls for dividing the entire country into a finite number of distinct areas or geographic segments. Then, since every industrial establishment must be located in one segment or another, the complete list of segments will also cover all the industrial establishments in the country. Thus the geographic frame - the list of segments - will provide complete coverage of the industrial universe.

To be operationally satisfactory the segments of a geographic frame must be uniquely defined. That means they must have distinct boundaries that are clearly shown on available maps. A good example is the set of enumeration areas (EAs) used for the latest population census of Industria 1990. Maps showing their boundaries are readily available.

The availability of the EA maps, which can readily be copied as needed, gives the EAs great cost and time advantages over other methods for segmenting all of Industria to develop an area sample

frame. The cost and time advantages alone might be considered sufficient reasons to choose the EAs as the geographic sampling frame. The case for choosing the EAs, however, is further strengthened by the fact that for each establishment the directory records show the EA in which the establishment is located, as well as its industry code and a total persons engaged figure. From that information, rough, comparative measures of size, can be developed for all the EAs. As a practical matter, no other set of segments is competitive with the EAs, so we shall use them as our geographic sampling frame.

Inaccurate information in the sampling frames reduces their efficiency. However, efforts to improve the information would be surely worthwhile only if they yielded substantial improvements at low cost. Possibilities considered included an advance mail canvass of the directory establishments - one limited to inquiries about ownership, nature of activity and total persons engaged - a similarly limited advance field canvass, and request to municipal authorities and other local sources (e.g. village leaders) for industrial establishment lists or related information.

None of these procedures were considered fully satisfactory. A mail canvass would not get an adequate response. A full field canvass would be costly. Matching lists obtained from cities, etc. to the directory would present serious difficulties. However, two limited steps were taken to improve the frames. For the geographic

frame, cities and towns were requested to identify their areas in which unusual industrial growth has occurred during the last 10 years. The NSO then canvassed the EAs involved, and compiled lists of the industrial establishments in each of them. For the directory, selected companies were mailed, in advance, pre-canvass questionnaires for reporting changes to their lists of industrial establishments: newly built, purchased, sold or discontinued. Those pre-canvass questionnaires were sent to all companies that owned two or more industrial establishments, or that owned a single industrial establishment with 500 or more persons engaged. Both of these two programs for updating the frames provided significant improvements at modest costs.

4. Cut-off Sampling for Supplementary Information

The industrial census will use two types of report forms: long forms and short forms. Establishments with 10 or more persons engaged will receive the long forms. Smaller establishments will receive the short forms. The distinction will apply to establishments selected from the directory according to the size information therein. It may also apply, selectively, to establishments included in the area sample canvass.

The short forms will ask only for the most basic information: a description of activity (for classification purposes), number of persons engaged (detail and total), payroll (operatives, others and total) and total cost of materials and total value of shipments and receipts (from which value added can be derived). The long forms will include all those inquiries, plus many more including employees' fringe benefits, detailed material costs and products shipped, stocks, capital expenditures and others.

This program illustrates the use of sampling - in this case of cut-off sampling - to collect supplementary information, i.e. all the topics included exclusively in the long form. Cut-off sampling is singularly appropriate in this case for

several reasons.

The distinction by size realistically acknowledges the futility of trying to collect more than the simplest, most basic information from small establishments. As a class, their records regarding the supplementary topics of the long forms range from none to poor. Few of the small establishments would respond to the supplementary inquiries if asked. Attempting to force them to report the supplementary detail would have little merit. Many of the limited number of responses received would involve respondents' guesses of dubious accuracy. The combination of non-response and inaccurate reporting from most small establishments argues strongly against attempting to collect the supplementary information from them.

The long form reports will account for a high proportion of the total value for most items. For example, Exhibit I-6-3 of Chapter 1 indicates that the long forms will account for approximately 90 percent of the total number of persons engaged. It is likely that they will account for roughly the same percentage of the totals for other items. The cut-off sample totals by themselves, therefore, would provide useful lower limits for the supplementary items.

In general, though, users would prefer to have projections

to universe levels, for they are easier to interpret and to use than incomplete totals, especially in relation to other statistics. Accordingly, the NSO will develop estimated totals for all the supplementary items of the long forms. The estimates will be derived by dividing the cut-off sample totals for the respective supplementary items by coverage ratios for associated items which appear on both the short and the long forms. To illustrate: The amount paid to operatives appears on both the short and the long forms. Its coverage ratios, therefore, will be calculable by dividing the cut-off sample totals for that item by the corresponding universe totals derived from all reports. The associated item, number of days worked by paid operatives, appears only on the long form. To derive its universe level estimates its cut-off sample totals will be divided by the corresponding coverage ratios for the amount paid to operatives. Exhibit 4-1 shows which items have been paired for purposes of these projections.

Coverage factors will be calculated and applied to produce estimates both by industry groups and by provinces. The estimates by province then will be conformed to those by industry group. This will be done for each item by calculating the ratio of the national total of the industry group estimates to the total of the province estimates, and then multiplying each province estimate by that ratio.

Major industry group totals will be obtained by summation. No industrial-geographic cross classification estimates will be derived for the supplementary items.

One last issue concerning the estimates for supplementary items needs to be addressed. What should be done when an establishment reports on the wrong form for its size? The answer generally is to accept the report and process it according to its form type. There are two exceptions. (1) When an unusually sizable establishment - one with 50 or more persons engaged - reports on a short form we shall request it to submit a long form report. (2) When an establishment with fewer than 10 persons engaged submits a grossly incomplete long form report, we shall recode it and process it as a short form.

5. The Canvass for the Primary Statistics: Complete, Cut-off Sample or Probability Sample?

Statistics for the primary items which the short forms cover--total persons engaged, payroll total shipments and receipts, and total cost of materials - can adequately describe the level of industrial activity in Industria.

5.1 Publication Objectives for the Primary Statistics

To satisfy the census objectives, statistics for the primary items must be published in considerable detail by industrial, geographic and size classifications, including major cross-classifications. The tables must all be consistent. (All sums of detail must conform to the totals of each table and corresponding totals of different tables must match.) The estimates must have objectively calculable sampling errors, and the quality of the estimates must be related to their importance.

5.2 Objections to Cut-off Sampling

Cut-off sampling cannot satisfy the publication objectives. If it were to be used, no current measures of the undercoverage would be available and therefore satisfactory projections to universe totals could not be made. Historic coverage measures might be calculated, but they would be unreliable for major

totals and more so for the many detailed figures desired from the census.

5.3 Objections to a Complete Canvass

A complete canvass theoretically could satisfy most of the publication objectives. As discussed in Section 1, however, there are serious objections to such a program. Here, it is sufficient to recall that a complete canvass would require an expensive field effort, that the program would indiscriminately treat economically trivial and economically paramount data alike, and that the heavy cost would be incurred mostly in collecting information about a minor fraction of the country's industrial activity.

5.4 Probability Sampling - A Tool Which Satisfies all the Publication Objectives

Probability sampling, which entails selecting units from the entire universe at the rates given by their assigned probabilities, is a highly flexible tool. As such, it can satisfy all the publication objectives.

Estimates can be developed from probability samples for any desired category. Fully consistent tables can be produced. Measures of quality (in terms of the error associated with

sampling) can be calculated for as many estimates as desired. Most important, on the average, the quality of the estimates will vary directly with their importance.

5.5 Administrative and Other Advantages of Probability Sampling

Quite aside from its technical advantages regarding the estimates, probability sampling offers administratively significant advantages. Obviously it can slash the total cost by drastically cutting the workload, particularly the field canvass. Lower costs from lower workload, however, are not the sole gains. A lower workload also means a lessened need for temporary, inexperienced field and office staff. Those reductions in turn imply quicker results, and pay-offs in better quality as well as money. The more experienced staff can exercise more care in handling the smaller volume of reports, thus achieving better control over reporting and processing errors. Lastly, probability sampling distributes the census budget more satisfactorily between large and small establishments, and simultaneously lightens the reporting burden for the smaller size classes.

6. Formal Mathematical Properties of Probability Sampling

To this point we have discussed probability sampling mainly in qualitative terms. Its quantitative behavior depends on certain descriptive universe values and how their corresponding sample estimates relate to them. This section defines the formal mathematical relations involved.

6.1 Universe Values (parameters)

The number of universe values of concern is small. They are:

(i) The total number of units in the universe, N.

(ii) The <u>total</u> value of an item (such as receipts), or the sum of the individual values. It is written as X, or $\sum_{i=1}^{N} X_i$. (If the total is for a particular category - an industry group, a province, etc. - $X_i=0$ for every establishment which is not in that category.)

(iii) The <u>mean</u> value per unit. It equals X/N and is written as $\bar{X}$.

(iv) An average of the squares of the differences of the individual values from the mean, called the <u>variance</u>. Its symbol and formula are $S^2 = \sum_{i=1}^{N} (X_i-\bar{X})^2/(N-1)$.

(v) The relative variance, or for brevity, the rel-variance, the ratio of the variance to the square of

the mean, $S^2/\bar{X}^2$. Its symbol is V^2.

Corresponding to these five universe values (called <u>parameters</u>) are the number of units in the sample, n, and estimates of the item total and mean, written as X' and $\bar{X}'$, respectively, and of the variance and rel-variance, written as s^2 and v^2.

6.2 The Foundation of Probability Sampling: Simple Random Sampling

As was remarked briefly in Section 2, any particular probability sample is one of many possible samples of the same design and size. Thus, whereas the universe parameters are constant, but unknown values, the sample estimates vary from one sample to another. When the principles of probability sampling are strictly observed, the behavior of the estimates in relation to the universe values follows known mathematical laws. These laws involve the unknown parameters. Nevertheless, with minor practical reservations, we can use them to predict how good our estimates will be. Later, from the sample data themselves, we can evaluate the actual quality achieved. These ideas are well illustrated by simple random sampling, the foundation of all probability sampling methods.

The essential properties of random sampling are that:

(i) Chance (a random process) strictly governs the selection of the sample units.

(ii) All the different possible samples of n have the same chance of being selected.

(iii) As a corollary (consequence) of (ii) all N units have the same chance of being selected, $P_i=p=n/N$.

6.3 Estimates of Totals From Simple Random Samples

Estimates of totals (persons engaged, payrolls, receipts, cost of materials and value added) are our primary interest. To compute them from a simple random sample, weight each observed sample value by the reciprocal of the probability of selecting the unit, and sum over all the sample units, i.e.

$$X'=\sum_{i=1}^{n} X_i/p_i=\sum_{i=1}^{n} X_i/(n/N)=(N/n)\sum_{i=1}^{n} X_i \qquad (6.3.1)$$

(Note that here too, if the estimate is for a particular category, $X_i=0$ for establishments which are not in that category.)
The set of all possible estimates, X', also constitutes a universe with a limited number of descriptive measures. The most important of these are its mean value, technically called the expected value and its variance. Their symbolic expressions are E(X') and $S^2(X')$. They are closely related to the analogous parameters of the original universe. In fact,

$$E(X')=X, \qquad (6.3.2)$$

which shows that the average of all the sample estimates of the universe total exactly equals that universe total. The relationship between the variances is a little more complex, namely

$$S^2 (X') = (1-f) N^2 S^2/n, \quad (6.3.3)$$

where for simplicity we substituted f, the sampling fraction, for n/N.

The corresponding rel-variance, $V^2(X')$, looks simpler, for

$$V^2(X')=(1-f)V_2/n. \quad (6.3.4)$$

We shall also refer to the relative standard error of the estimate, V(X'), the square root of $V^2(X')$.

Under fairly broad conditions, particularly when the sample number for the category is at least 30, the normal probability curve will describe the distribution of the X' quite well. For example, suppose that the relative standard error, V (X') is 3 percent. Then close to two-thirds of the sample estimates will differ from X by no more than 3 percent in either direction; about 95 percent of the estimates will differ from X by no more than 6 percent, and almost none of the estimates will differ from X by more than 9 percent. These relations for multiples of one, two or three times V(X') are general. If V(X') is 2 percent nearly two-thirds of the estimates will differ from X by 2 percent or less. At the planning stage, speculations regarding the value of V^2, together with selected values of n, would provide useful, advance approximations of the precision of the estimates. After the sample has been

selected we can estimate V^2 and $V^2(X')$ from the values observed.*

6.4 The Relationship of X' to X for Other Probability Sample Designs.

Universe totals can be estimated from a probability sample of any design by the general formula

$$X'_p = \sum_{i=1}^{n} X_i/P_i \qquad (6.4.1)$$

where the p_i is the probability that unit i will be selected in a sample of size n. The p_i can differ among the different units. When they are properly applied as weights to their corresponding X_i values, the expected value of X'_p, $EX'_p=X$, just as in the case of simple random sampling.

The rel-variance of X'_p, $V^2(X'_p)$, likewise is related to its corresponding value, the universe rel-variance V^2. This relationship can conveniently be expressed in terms of the rel-variance of a simple random sample, $V^2(X'_R)$. Symbolically

*The standard or "textbook" formula for the estimate of V^2 is

$$v^2 = \sum_{i=1}^{n} (X_i-\bar{X}')^2/[(n-1)(\bar{X}')^2].$$

It can appropriately be substituted for V^2 in (6.3.4.) when the sample number for the category is not small. Samples from industrial universes often fail to satisfy that condition.

$$V^2(X'_p) = (\text{design effect})\ V^2(X'_R), \qquad (6.4.2)$$

where the factor <u>design effect</u> indicates by how much the particular design changes the rel-variance.

That factor may be less than or greater than one, depending on the characteristics of each particular design. The next section briefly describes some commonly used designs, including those most pertinent to our purpose.

7. Frequently Used Probability Sample Designs

Simple random sampling is seldom used in practice. Many other probability sample designs have been developed. Their aims are to improve efficiency, to simplify the selection process, to deal with administrative restraints or to serve a combination of those purposes. All probability sample designs preserve the key features that every unit has a known chance of being selected, and that chance determines the choices for any given sample.

7.1 Stratified Random Sampling

A useful variation from simple random sampling is <u>stratified random sampling</u>. This design divides the universe into a number of distinct sets - the strata - from each of which a random sample is selected. Estimated totals are calculated for each stratum by the standard formula,

$$X'_h = (N_h/M_h) \sum_{i=1}^{M_h} X_{hi}, \qquad (7.1.1)$$

where the subscript h has been added to indicate the individual stratum. Then the stratum estimates are summed to derive the estimated universe total.

The strata may be defined in any manner whatsoever. For example, they could be individual file drawers, which are defined as the strata purely for convenience. More significantly the strata might

be regions with an urban - rural split within regions, or they might be size classes by broad industrial group within size class. Stratification can benefit the sampling process when it groups together units with similar values and assigns to different strata units with dissimilar values. In the ideal case perfect stratification - all units within each stratum would be alike, only the between stratum values would differ. In this extreme situation a sample of one unit from each stratum would be enough for an error free estimate of the universe total.

Such extreme results, of course, are never achieved. Commonly the increased homogeneity achieved by stratifying produces moderate gains; the design effect factor is not much smaller than one.

There is another way, however, in which stratified sampling can improve the quality of the results. With simple random sampling, when calculating the estimated total, each of the n sample values gets the same weight, N/n. With stratified sampling the weights N_i/n_i are of the same form, but can vary from one stratum to another depending on how the total sample size of n is allocated to the strata. A judicious allocation can further reduce the design effect factor, and for any particular set of strata a theoretically optimum allocation exists.

7.2 Random Systematic Sampling

Random systematic sampling is often used in place of simple random

sampling. Under this method the sample units are selected in order and at equal intervals, beginning with a random start. To select a 10 percent sample, for example, a random number between 1 and 10 inclusive would be chosen. The corresponding unit and every tenth thereafter then would be selected until the entire sample had been chosen.

The procedure is operationally exceedingly simple. It has been used often to select samples in the field from lists compiled there. Unfortunately the simplicity of the method sometimes has proved disadvantageous. Enumerators have found it tempting to manipulate the selection in ways that exclude difficult units from the sample. One technique involved manipulating the order in which they listed units. Since they could identify the sample lines in advance, they filled them with easy cases as quickly as possible instead of listing units in the order in which they came to them. Careful controls are needed to prevent such distortions of the sampling process.

When properly controlled, random systematic sampling has proved to be a satisfactory method. Experience shows that its results usually are comparable to or somewhat better than simple random sampling.

When the sample is to be selected from a file that either is or easily can be arranged in some desirable order, random systematic

sampling can capture some of the benefits of stratification. It would be cumbersome and undesirable to define a large number of distinct strata - say by size crossed by industry crossed by province - each of which is to be sampled at a different rate. However, one of those classifications say size, might serve well for defining the strata, and within each stratum the records might be arranged according to the secondary classifications. The systematic selection then would distribute the sample from each stratum fairly evenly by industry and geography. Simple random sampling from the strata would not have the same effect. Some of its samples would be heavily concentrated in certain industries and in certain regions, with correspondingly thin representation of other industries and regions. The secondary level of stratification, characteristic of random systematic sampling, gives the method its edge over simple random sampling.

7.3 Probability Cluster Sampling

It is not always feasible or desirable to sample establishments singly, that is, one at a time. Instead, in some circumstances, it is better to work with sampling units that cover clusters of establishments. Typically such units are definable geographic areas that together cover the country. They might be as large as governates, as small as city blocks or of any other favorable size provided they meet the criteria for geographic area sampling frames. (Section 3)

Probability samples of such units can be selected in the same manner as samples of any other units. The establishments in the selected geographic units would constitute a perfectly valid probability sample. Alternatively, a sub-sample might be chosen from the cluster of establishments in each initially selected geographic unit. Those sub-samples, too, would constitute a valid probability sample of establishments, provided that probability sampling principles were fully observed at the sub-sampling stage as well. The set of establishments from each initial unit - all or the sub-sample as the case may be - is called the ultimate sample cluster, or more briefly the ultimate cluster, and the design naturally enough is called probability cluster sampling.

Neighboring establishments tend to be more like one another than establishments are in general. Examples abound: the concentrations of meat packing plants in the north, sawmills and planing mills in the east, etc. This tendency for clusters to be more homogeneous than the universe as a whole reduces the efficiency of cluster designs relative to simple random sampling. The design effect factor for cluster sampling typically is greater than one. The effect is most pronounced when the clusters are large, i.e. when they include many establishments, on the average. Specifically the expression for the design effect factor of cluster sampling is

$$\text{design effect} = 1 + (\bar{n}-1)\bar{\delta}, \qquad (7.3.1)$$

Where $\bar{n}$ is the average number of establishments per cluster, and $\bar{\delta}$ is a measure of the average homogeneity of the clusters.

Since the design effect is greater than one[1], we need to include more establishments in a cluster sample than in a simple random sample to get the equivalent precision. Nevertheless, for a given precision, cluster sampling may be more economical than simple random sampling.

The comparative total costs rather than simply the numbers of sample establishments determines which design is preferable. For example: To be satisfactory, the sampling frames for both designs must closely approach the ideal of complete coverage. For the random sampling design, an expensive canvass, aimed at listing all the establishments in the country, would be needed (making realistic assumption that an adequate directory is not available). For the cluster sampling design it would be relatively easy to compile a complete list of geographic segments, and it would be sufficient to canvass only two selected samples of those segments for listing purposes. The reduction in listing costs could more than compensate for the cost of including more establishments in the sample.

---

[1] Unless $\bar{\delta}$ is zero or negative, a theoretical possibility we can disregard.

Cluster sampling sometimes can be economical even when complete lists are on hand. When enumerators must make personal visits to collect the reports, travel costs can mount rapidly for simple random sampling. This can be especially important in rural areas. Clustering the sample can drastically reduce the amount of travel. In this case too, it may produce substantial net savings over the total cost of the smaller, simple random sample needed for comparable precision.

7.4 Sampling with Probabilities Proportional to Measures of Size (PPMOS)

Probability sampling does not require that all units have the same chance of being selected. The units can, alternatively, be sampled with probabilities proportional to measures of size (PPMOS). Under this method chance still determines the units that are selected at any trial, but units with large measures are selected more often than those with small measures. The procedure is intuitively appealing, and it is better than sampling with equal probabilities, provided that the measures of size are well correlated with the values to be estimated from the sample.

Sampling with PPMOS can replace sampling with constant probabilities in stratified sampling designs, in systematic selection and in cluster sampling. It is widely applied in multi-stage cluster sampling where moderate numbers of broadly defined

units are selected at the first stage, and sub-sampling probabilities are assigned such that every unit (e.g. household) in the universe has the same chance of being selected.

7.5 Sampling With and Without Replacement

In thinking of sampling from a finite universe, we ordinarily think of sampling without replacement, that is, of removing each selected unit before selecting another one. Without violating any basic principles, we could replace each unit as selected, so that it could be selected again. Sampling with replacement, as the latter method is called, is somewhat less efficient than sampling without replacement. This is shown by the factor (1-f) in the formula for the rel-variance of a total estimated from a simple random sample drawn without replacement (6.3.4). The comparable factor is simply one when sampling with replacement. The difference often is quite unimportant.

7.6 Other Probability Sampling Designs

Other probability sampling designs have been devised, but are not important for our purpose. The fundamental methods described above offer a sufficient range of choices and a framework for deciding the specific details of our sample design.

8. A Uniform Rule for Relating the Quality of Estimates to Their Importance.

A key issue in deciding the specific details of the sample design is: How good must the estimates be? Thousands of estimates are desired. It would be hopeless to consider each one individually. As a step toward solving the problem, and as suggested in Section 2, we might adopt the uniform rule for all estimates that their quality should vary with their economic importance.

8.1 The Specific Recommended Relationship between Quality and Importance.

The recommended rule is sensible, but to be operational it must be defined more specifically. The specific relationship recommended is that, on average, quality change proportionally with the square root of importance. The importance of any category - industry group, major industry group by region, all manufacturing by province, etc. - is well measured by the category's labor force, or total persons engaged. The quality of an estimate is well defined by the inverse of its relative standard error, $1/V(X'_c)$. [It should be recalled that $V(X'_c)$ indicates how widely an estimate X'_c may differ from its corresponding universe value, X_c.] With persons engaged taken as the measure of importance and $1/V(X'_c)$ taken as the measure of quality, the

recommended rule is

$$K/V(X') = \sqrt{X_c} \quad \text{or} \quad V(X'_c) = K/\sqrt{X_c} \qquad (8.1.1)$$

Raising or lowering the value of the proportionality constant K will raise or lower the general level of the relationship between quality and importance. A reasonable level can be determined by trial. Insert in formula (8.1.1) for a few categories the persons employed totals, X_c, and acceptable corresponding values of $V(X'_c)$. Then K can be calculated readily. While the results may differ for each category tested, choosing a reasonable compromise value should not prove difficult.

8.2 Comparison of Recommended Rule with Other Rules for Specifying Quality.

An important feature of the recommended rule is that it calls for quality to change systematically with importance, but to do so slowly. The rate of change is much slower, for example, than would be the case of a simply proportional relationship. In the latter case the changes appear too abrupt. Such judgments necessarily are subjective. However, to consider an example, wouldn't it be objectionable to have one category with 400 persons engaged, a second with 3,600 persons engaged and relative standard

errors respectively of 16 percent and 2 percent for them?

The recommended rule also has more appeal than one which specifies equal quality for all estimates at a given publication level, such as major industry group by region. The latter rule makes little sense for any publication cells at the reference level which are of negligible economic importance, or worse, are empty. Moreover what guidance that rule provides for other categories is unclear. Does it imply that better quality is desired for higher order categories - national, major industry group totals, regional, all industry totals, etc. - and the reverse for lower order categories? In cases of inconsistencies, which takes precedence, the industrial classification or the geographic classification? Lastly, what about size tables which present the data in a completely independent dimension? A rule based on any arbitrarily chosen publication level cannot well define the quality objectives of the sample.

While size generally defines importance well, exceptions can occur. Special circumstances, perhaps, for example, an experimental development program initiated a few years ago in one or two provinces, may justify giving extra attention to particular categories. In such instances we merely need to increase the category's original size by an appropriate

factor before sampling.

Conversely, per unit costs for some categories may differ widely from the average. If exceptionally high costs are ignored those categories will absorb too much of the total budget at the expense of poorer results for all other estimates. Reducing the quality specified for the high cost categories would be justified. Conversely, categories which involve lower than average costs merit better than average quality. We can obtain the desired effects by suitably adjusting the measures of size of the units in the respective categories.

8.3 Mathematical Properties That Support the Rule

The attractive concept that quality should change slowly with importance can be satisfied by many formulas. Two important mathematical properties of formula (8.1.1) support its choice.

First, it applies consistently to the sums of independent estimates that individually satisfy the rule. This implies that estimates for industries, for industry groups and major industry groups, and estimates for provinces and for regions, etc. will all show the same relationship. For example, an estimate for a table total of four,

approximately equal size sub-categories would have about half as large a relative standard error as each of the subordinate estimates.

The second mathematical property that supports this particular rule is even more compelling. The rule conforms approximately with the conditions for maximizing the efficiency of the sample design. Maximizing efficiency is a fundamental objective of every sample design. It is a powerful force in planning and if efficiency had dictated some other rule it would certainly have been considered.

8.4 Approximations in Seeking to Optimize the Sample Design

A sample design can involve many numerical variables, and the values assigned to them will influence the efficiency of the design. The exactly optimizing values involve unknown universe parameters. Necessarily, therefore, approximations must be accepted.

The first, and most central approximation we shall introduce is that all sample groups of units, g, have the same unit

relative standard deviation, V_g.[1] In practice V_g is quite stable and differs only moderately among different groups. With rare exceptions the assumption works well. (If exceptional circumstances lead to the belief that for some group V_g greatly exceeds the average V, adjust the measure of size for the group as discussed in Section 8.2.) This and other necessary approximations generally will cause little harm, because the mathematical curve that describes the efficiency achieved is quite flat over a wide range around the exact optimum. Reasonable deviations from the exactly optimizing values will cause only small losses in efficiency compared with the theoretically optimum results.

---

[1] $V_g = V_g^2$, where $V_g^2 = S_g^2/\bar{X}_g^2$, and is the rel-variance for a group of N_g units. (See Section 6.1)

9. The Certainty Class.

When sampling for industrial statistics all of the most important establishments must be included with certainty. Unless such a lid is imposed the sampling errors can increase uncontrollably. The exact size chosen for this absolute limit is not crucial, but experience suggests that it should be low enough to account for at least 50 percent of the total industrial activity of the country.

This criterion can be met for Industria by setting the certainty cutoff size at 50 persons engaged. It is believed that the directory includes all such establishments, although it may show a smaller size for some of them. As a safeguard against such mis-classifications, it would be prudent to set the certainty cutoff at a lower size.

Moreover, a lower certainty cutoff might be more efficient for sampling from the directory frame, even though its coverage is incomplete for establishments with fewer than 50 persons engaged.

To test that hypothesis the total numbers of establishments - certainty plus sample non-certainty - needed to achieve various precision levels were calculated for certainty cutoffs of 50, 25, and 10 persons engaged.

These calculations utilized the specified relationship between the relative standard error and size, $V(X')=K/\sqrt{X}$, and the following simplified approximation for $V(X')$:

$$V(X')=(X_0/X)/\sqrt{M_0} \qquad (9.1)$$

where

$V(X')$ is the relative standard error of the estimate X',

X_0 is the total value of the non-certainty class,

X is the total value (of the certainty and non-certainty classes together),

M_0 is the sample number of non-certainty establishments.

These calculations are based on the data of Exhibit I-6-4, as modified to allow for incomplete directory coverage of establishments with less than 50 persons engaged; they assume incomplete certainty coverage for all of the lower size classes. The results are summarized in Table 9-1.

TABLE 9-1

TOTAL NUMBER OF ESTABLISHMENTS REQUIRED FOR SPECIFIED QUALITY LEVELS BY SELECTED DIRECTORY CERTAINTY SIZE CUT-OFFS

Specified Quality Level, K	Directory Certainty Size Cut-off (Number of Persons Engaged)		
	10	25	50
1.0	9,318	16,633*	36,354*
1.5	6,886	8,893	16,952*
2.0	6,035	6,183	10,261
2.5	5,641	4,929	7,018
3.0	5,427	4,248	5,310

*Exceeds 100 percent

The numbers of Table 9-1 do not reflect differences in costs and some other factors that influence the efficiency of the sample design. The numbers are satisfactory, nevertheless, as a guide. They show that the optimum directory certainty cut-off is in the region of 10 to 25 persons engaged. The cut-off at 10 persons engaged gives better results for the higher quality levels, and more stable results over the whole range of K values, than the cut-off at 25 persons engaged. These features, together with the fact that 10 persons engaged has been designated as the cut-off for the long forms, led to choosing 10 persons engaged as the certainty cut-off for sampling from the directory.

The analysis did not include data for the government operated public utility facilities or for mining establishments. Data for all the public utilities will be obtained by special arrangements with the government offices concerned. They therefore are of no concern when developing the sample design for the industrial census. Data for mining establishments were not included simply because detailed figures by size classes are not readily available and the totals are so small they would not affect the results. The 10 persons engaged cut-off for sampling from the directory will be applied to the mining establishments as well as to the manufacturing establishments, and no distinction will be made between the two classifications in the following phases of designing the sample.

10. Sampling Jointly from the Directory and the Geographic Frames

In accordance with the decisions of Section 9, all directory establishments classified as having 10 or more persons engaged - both urban and rural - will be included with certainty in the sample. They will be canvassed by mail and followed up by personal visits as necessary. The remaining establishments, all those in the directory which are classified as having less than 10 persons engaged and all establishments omitted from the directory, will be represented by samples selected with less than certainty probabilities.

10.1 Different Strategies for Urban and Rural Samples

Different sampling strategies will be employed for urban places than for rural areas. For urban places a sample of the establishments with less than 10 persons engaged will be selected from the directory and canvassed by mail. Then, in order to cover the urban establishments (large and small) which the directory omits, a supplementary sample of urban geographic areas will be selected. No rural establishments from the less than 10 persons engaged class will be selected from the directory. Instead, the rural, small establishments will be covered solely by a geographic sample. Cost considerations led to adopting this twofold approach.

A moderately good response to a mail canvass is anticipated for urban establishments. In contrast, a mail canvass of small rural establishments would be wasteful. Interviewers would have to make follow-up visits to nearly all the small rural establishments canvassed in order to get their reports.

A sample selected randomly from the directory would be widely scattered and would involve high travel costs, whereas in urban places travel costs would be comparatively low. Clustering the rural sample is the natural way to deal with the travel problem. Then, since a rural area sample is needed in any case, it is economical to use the same sample to cover all the rural establishments with less than 10 persons engaged and the larger rural establishments which are missing from the directory.

10.2 Stratification of the Urban Sample

For sampling purposes we divided the directory file of urban establishments into two size strata, 1-4 persons engaged and 5-9 persons engaged. Within each stratum the records were sorted by major industry group by region, thereby providing a secondary level of stratification within the size strata.*

The cost of stratifying was nominal and the operation considered worthwhile, therefore, even though it is expected to yield only modest benefits. The difference in the average establishment sizes for the two strata accounts directly for some of the gain. The larger size class averages more than five persons engaged; the smaller size class averages less than three persons engaged. Stratification eliminated that difference of 2+ persons, which would contribute to the variance of a single class, 1-9.

---

* Optionally we might have stratified by industry and geography, then sampled from those strata with probabilities proportional to size. There is no assurance that the resulting more complex design would give better rather than poorer results.

Additional gains occur in less obvious ways. One source is the greater stability of the larger establishments. A lower proportion of them than of the smaller establishments will have gone out of business and consequently have current values of zero. The presence of such "zeros" inflates the variances, an effect which the stratification reduces. Lastly, stratification enables us to apportion funds to the two classes in a way that takes better account of their different response rates to the mail canvass than is possible for a single combined class.

We similarly divided the urban EAs into two size strata, once again in order to improve the efficiency of the sample design. In this instance we attempted to separate EAs that contain many industrial establishments from EAs that contain relatively few. Industrial establishments tend to congregate, although some may be found in scattered locations throughout cities and towns. In the belief that the non-directory establishments - the targets of the area sample - tend to concentrate in the same EAs as the directory establishments, urban EAs that contained more than four directory establishments were assigned to an urban-high stratum. Based on the industrial growth reports from local sources, the NSO had conducted field listing operations in about two dozen EAs. The new information rather than the directory counts was used to assign those EAs to strata. As a result seven EAs were transferred to the urban high stratum, for a total of 198 EAs. The remaining 253 urban EAs of the geographic frame were assigned to an urban-low stratum.

10.3 The Rural Stratum

As implied earlier, but perhaps needing explicit statement, the rural EAs constitute a separate stratum of the geographic sampling frame. They present quite a different sample design problem than the urban EAs.

Even though the stratum includes all rural establishments, except directory establishments in the 10 or more persons engaged class, the rural EAs have an average of less than two establishments each. That number is too small to be economical. Those with less than three small directory establishments, therefore, were joined with adjacent EAs in the same province to form enlarged geographic frame sampling units. No more than four EAs, however, were linked together into a single unit. This process produced 541 rural sampling units, with an estimated average number of establishments of about 4.5 each.

10.4 Allocation of the Sample to Strata

How we allocate sample numbers of reports to collect from the various strata will affect the efficiency of our design. For an optimum allocation, the numbers should take account of three factors: the size of the stratum (total persons engaged), the applicable design effect and the average cost per report. Specifically, the formula for the optimum number of reports from each stratum is

$$R_h = t X_h \sqrt{D_h / C_h} \quad ^{**} \qquad (10.4.1)$$

** The formula is an approximation. It incorporates the simplifying assumption made earlier that the relative standard deviation is constant over all classes, i.e. $V_h = V$.

where

M_h is the optimum number for stratum h,

t is a proportionality constant,

X_h is the total number of persons engaged,

D_h is the design effect factor,

and

C_h is the average cost per report.

The formula shows that size affects the sample number far more strongly than the design effect or the average cost. Quite reasonably, we should allocate more sample reports to large strata than to small ones, more to strata with large design effects than to strata with small design effects, and more to inexpensive strata than to the expensive ones. The value of the proportionality constant, t, will depend on the quality specified for our estimates (or the amount budgeted to collect and to process the sample reports). It also involves the sum of the N_h, the total sample number. Before we can evaluate t, therefore, we must assign to each stratum values for X_h, D_h, and C_h. We incidentally shall need the N_h, the total number of establishments per stratum.

10.5 Estimates of the Number of Establishments and Persons Engaged, by Strata

For our purposes we need to estimate how many establishments are missing from the directory, separately by persons engaged size classes and within size classes by urban and rural classifications. Additionally, we need to estimate the number of rural establishments with fewer then 10 persons engaged. For

each category we also need to estimate the corresponding total number of persons engaged.

Exhibits I-6-4, I-6-5 and I-6-6 give helpful figures. They present estimates of the numbers of establishments in some detail by number of persons engaged size classes and by urban and rural classifications. They also present very useful numbers of persons engaged totals and averages for various categories. The exhibits use broader size classes, in general, than those that define our strata, and they provide only some of the urban-rural breakdowns we would like to have. They do not, of course, give any direct figures on the completeness of the directory.

Comparisons of the numbers of establishments in the directory and the current estimates cannot help with this problem, for unknown numbers of the directory establishments have gone out of business, and the proportionate distribution by size may have changed as well.

Fortunately, we can assume with great confidence that the directory is complete for establishments with 50 or more persons engaged, that its coverage then declines very slowly down to establishments with 25 persons engaged, and more rapidly below that size to a minimum of 60 percent for the 1-9 persons engaged size class.

Guided by these assumed conditions and the data in the exhibits cited, we developed the needed stratum by stratum estimates of the number of establishments and of total persons engaged. They are shown in Table 10.5-1.

The estimates are rough, but they are plausible and should be satisfactory for purposes of allocating the sample budget to the strata.

TABLE 10.5-1

ESTIMATED NUMBER OF ESTABLISHMENTS AND OF PERSONS ENGAGED TO BE COVERED BY SAMPLING, BY STRATA

Stratum description	Stratum code	No. of establishments	No. of persons engaged
Urban, directory, 5-9 persons engaged	D9	950	5,200
Urban, directory 1-4 persons engaged	D4	2,900	6,600
Urban, nondirectory, high	UH	2,630	13,700
Urban, nondirectory, low	UL	400	1,950
Rural, all 1-9 persons engaged plus all larger non-directory	R	2,490	9,550
TOTAL	ALL	9,370	37,000

10.6 Design Effects and Average Unit Costs

In order to determine the number of sample establishments to select from each stratum we must weight the persons engaged figures of Table 10.5-1 by the design effect factors, D_h and unit costs, C_h, as indicated by formula (10.4.1).

We determined values for them as follows.

For the two urban directory strata, D9 and D4, we plan to use random systematic sampling. The results are apt to be a little better than simple random sampling, which suggests that a design effect factor below one would be appropriate. We conservatively set the factors for both of these strata equal to one.

All three of the other strata involve cluster sampling. Their design effect factors are harder to assign. Their values depend on $\overline{n_h}$, the average number of sample establishments included in each cluster, and $\overline{\delta_h}$ the average measure of the homogeneity of the clusters. We can calculate the values of $\overline{n_h}$ by dividing the number of establishments of Table 10.5-1 by the number of sampling units for the corresponding stratum, as shown in Table 10.6-1.

TABLE 10.6-1
AVERAGE NUMBER OF SAMPLE REPORTS PER CLUSTER n_h
BY STRATA

Stratum code	Number of sampling units	Number of establishments	
		Total in stratum	Average per sampling unit, $\overline{n_h}$
UH	198	2,630	13.3
UL	253	400	1.6
R	541	2,490	4.6

It is harder to decide on an appropriate value for $\overline{\delta_h}$. Its largest possible value is one, which it reaches only when, in every cluster, all the establishments are exactly alike. Values as large as one-third for $\overline{\delta_h}$ are rare. Acting conservatively again, we set $\overline{\delta_h}$ equal to 0.4 for all three strata. The design effect factors as then calculated by the formula $D_h=1+(\overline{a_h}-1)\overline{\delta_h}$, appear in Table 10.7-1.

In computing the other factor needed for allocation purposes, the average cost per sample report, C_h, we considered solely, costs which vary with the sample sizes. Substantially fixed costs, such as those for preparing the frames and sampling from them were excluded. Processing costs were taken to be the same for every sample report, without regard to its stratum. Collection costs, however, were compiled separately for every stratum, because the operations differed both in type and degree.

As remarked above, we plan to canvass the samples from both of the directory strata by mail, and to follow-up non-respondents with personal visits. Anticipated differences in the mail response rates resulted in different unit costs for the two strata.

Mail collection is not contemplated for the establishments of the urban geographic samples. For these, the plan is to have enumerators compile lists

of the industrial establishments in each selected geographic sampling unit, to have those lists matched (independently of the enumerators) against the directory, and then to have interviewers return to those establishments from which reports are required.*** Substantial differences in the listing cost for the high density stratum versus the low density stratum are expected, and are reflected in the unit costs derived.

Mail will not be used to canvass sample rural establishments either, although it will be used for the directory list of rural establishments with 10 or more persons engaged. In the rural sampling units, interviewers will not p e-list. They, instead, will collect reports from all the industrial establishments they find in their assigned geographic sampling units. (Excepted will be any establishment that has already reported by mail and shows the interviewer the file copy of its report.) The high cost of travel to rural areas compared to travel costs to urban places influence this variation in methodology. It, of course, also influences the unit costs.

In developing estimated costs per sample report for the geographic strata, it was necessary to distinguish between operations that involve entire sampling units and operations that involve individual reports. Costs associated with entire geographic sampling units included copying maps, travel to and from the unit and travel within the unit when it is to be canvassed completely, either

*** We shall not provide the enumerators with lists of the establishments in the directory, which they then could skip when compiling their listings in urban places, or collecting reports in rural areas. Use of such skip lists weakens control too much. Similarly, because we want to maintain strong control we chose full coverage of small clusters over sub-sampling from listings for large ones.

for listing purposes as in the case of the urban geographic sampling units or to collect reports, as in the case of the rural geographic sampling units. Travel costs included listers' and interviewers' time as well as allowances for fares, automobile mileage, etc. The geographic unit costs were reduced to a per establishment basis by dividing their sum by the average number of establishments per sampling unit. The result, the estimated cost per interview to collect a report, and the unit processing cost then were combined to arrive at a total per unit cost. Table 10.7-1 shows those total unit cost figures, together with the other values needed to calculate the relative numbers of sample reports for the approximately optimum design.[****]

10.7 Optimum Number of Sample Reports per Stratum

To convert the proportional numbers of sample reports of Table 10.7-1 to absolute values, we must specify either how much we can spend, or how much error we can tolerate in our estimates. We had decided that we would like our relative standard errors, $V(X_c')$ to vary inversely as the square roots of the importance of their related totals, X_c, i.e. that the relationship $V(X_c') = K/\sqrt{X_c}$ should hold on the average. Moreover, Section 8.2 suggested a method for choosing a provisional value for the proportionality constant K, which defines the general quality level of the estimates. We now have more

[****] The considerations that entered the derivation of the unit costs are important. The specific numbers are not. They are needed to illustrate methodological issues, and should not be considered real, or typical for any country. For a comprehensive discussion of costs, see Reference (1) pp. 270-284

information we can use to determine an appropriate value of K. With the data of Table 10.7-1 we can see how different quality levels affect the census budget, and can more fully judge the implications of different choices.

TABLE 10.7-1

CALCULATION OF NUMBERS PROPORTIONAL TO THE APPROXIMATELY OPTIMUM NUMBER OF SAMPLE REPORTS FOR EACH STRATUM

Stratum	Design Effect D_h	Cost Per report C_h	$\sqrt{\frac{D_h}{C_h}}$	Total persons engaged	Optimum $m_h = f_{h/c}$ (Approx.)	Total Establishments N_h
D9	1.00	6.00	0.41	5,200	2,123	950
D4	1.00	7.25	0.37	6,600	2,451	2,900
UH	5.92	10.25	0.76	13,700	10,412	2,630
UL	1.24	21.80	0.24	1,950	465	400
R	2.44	22.70	0.33	9,550	3,131	2,490

We can assign a few values of K and then calculate the corresponding number of sample establishments needed from each stratum to satisfy each such trial specification. Alternatively, we might make some provisional decisions regarding the sample, then examine their budgetary and quality implications. For both approaches the solution involves equating our specified quality

importance relationship to the mathematical formula for the relative standard error, namely:

$$K/\bar{X} = (1/\bar{X}) \sqrt{\sum_{h=1}^{s} (1-f_h) D_h \bar{X}_h^2 V_h^2 / n_h}$$

or squaring, to get an easier form to work with,

$$K^2 = (1/\bar{X}) \sum_{h=1}^{s} (1-f_h) D_h \bar{X}_h^2 V_h^2 / n_h \qquad (10.7.1)$$

We assume, as previously discussed, that the approximation $V_h^2 = V^2$, a constant, is satisfactory, and set $V^2=1$, a generally conservative value. As a further simplifying approximation the factors $1-f_h=1-(n_h/N_h)$ often are rounded off to one. However, our universe is small, and we can anticipate that the sampling fractions for at least some of the strata will be appreciable. In order to capture their effects, therefore, we express the term $(1-f_h)n_h$ in the alternative form $1/n_h-1/N_h$. Lastly, since, $\hat{n}_h$, the optimum number of sample reports for each stratum is approximately proportional to the corresponding value m_h of Table 10.7-1, we substitute tm_h for n_h. With those simplifications and substitutions we re-write equation (10.7.1) as

$$K^2 + (1/\bar{X}) \sum_{h=1}^{s} D_h \bar{X}_h^2 / N_h = (1/t) \ (1/\bar{X}) \sum_{h=1}^{s} D_h \bar{X}_h^2 / m_h$$

We can derive all the terms of equation (10.7.2) except K and t. Exhibits I-6-4 and I-6-6 provide the total number of persons engaged in manufacturing and mining, 292,400. Each term of the summations, and thus the sums, can be calculated from the figures of Table 10.7-1. After making these calculations, and substituting the numerical results in equation (10.7.2) we get

$$K^2 + 1.960 = 0.755/t. \qquad (10.7.3)$$

Now, if we specify a value for K we can easily find the necessary value of t, and in turn the implied values of the $\hat{n}_h$.

To illustrate: We considered an estimate for a category that covers about 400 persons engaged. We deemed that size to be marginally economically important, and felt that a relative standard error of about 10 percent would be tolerable for such estimates. In terms of our basic quality-size relationship that specification called for 0.10=K/20, or K=2, which in turn, when substituted in equation 10.7.3 gave t=0.208. The required $\hat{n}_h$ values then were calculated using the relation $\hat{n}_h = tm_h$ and the values of m_h in Table 10.7-1. The $\hat{n}_h$, their corresponding sampling fractions, f_h, and costs, TC_h, appear in Table 10.7-2.

TABLE 10.7-2
NUMBERS OF SAMPLE REPORTS, SAMPLING FRACTIONS AND TOTAL COSTS FOR SPECIFIED QUALITY LEVELS, K, BY STRATA

Stratum		K = 2.0			K = 3.0			K = 1.0			K = 1.0		
	N_h	$\hat{n}_h$	f_h	TC_h	$\hat{n}_h$	f	TC_h	$\hat{n}_h$	f_h	TC_h	$\hat{n}_h$	f	TC_h
D9	950	269	0.25	1,614	146	0.15	876	541	0.57	3,246	546	0.57	3,276
D4	2,900	311	0.11	2,255	169	0.06	1,225	625	0.22	4,531	630	0.22	4,568
UH	2,630	1,319	0.50	13,520	717	0.27	7,349	2,655	1.01	27,214	2,630	1.00	26,958
UL	400	59	0.15	1,266	32	0.08	688	119	0.30	2,564	120	0.30	2,616
R	2,490	397	0.16	9,012	216	0.09	4,903	798	0.32	18,115	805	0.32	18,274
TOTAL	9,370	2,355	0.25	27,667	1,280	0.14	15,051	4,736	0.51	55,700	4,731	0.50	55,692

For comparison with the sampling pattern for the quality level K = 2.0, similar calculations were made for K = 3.0 and K = 1.0. Those results also are shown in Table 10.7-2. Relaxing the sampling error specification by 50 percent (from K = 2.0 to K = 3.0) would substantially reduce the size and cost of the sample - both about 45 percent - while tightening the quality specification from K = 2.0 to K = 1.0 would nearly double the sample size required and its cost.[*****]

The result for K = 1.0 presents an interesting situation. For the stratum UH the theoretical optimum number of sample establishments exceeds the total number in the stratum. Accordingly the sampling fraction for that stratum was set equal to 1.00 and the allocation to the other strata recalculated. The results are shown in the column K'= 1.0. The changes are trivial.

[*****]The relative changes in the size and cost of the sample for different K values do not apply universally. A different distribution of the N_h, of the mos_h, or different values for any of the other factors that influenced the sample design would also have resulted in different sample size and cost patterns.

11. Selecting the Samples

Similar, although not identical procedures will be used to select the samples from the various strata. None are difficult. The availability of readily manipulable computer records for both the directory and geographic frames makes it easy to control the selection operations.

11.1 Features Common to All Strata

For checking purposes after the sample is selected for a stratum the following details will be printed out: The random number, the sampling fraction (or equivalent), the number of units selected, the total number of units and where applicable, the total measure of size. Also, for potential reference during later operations, during the process of selecting the samples, the stratum code, probability of selection and a selection indicator - yes or no - will be entered in the record for every sampling unit in each frame. A separate file will also be produced of the complete records for all the selected units.

11.2 Procedure for Sampling from the Directory Frame

Essentially the same sampling procedure will apply to the two strata of the directory frame. After the urban records have been extracted from the complete directory file and separated by size

class they are to be sorted by major industry group and city or town. Then a random systematic sample will be selected from each stratum.

The desired sampling fractions, given in Table 10.7-2 are 28/100 for stratum D-9 and 11/100 for stratum D-4. Computers can deal with such fractions perfectly well.** The procedure is as follows:

(i) The reciprocal of the sampling fraction will be computed to find the sampling interval, I.

(ii) A random number, r, will be chosen such that $0<r\leq I$, and the sequence, r, r+I, r+2I, etc. will be constructed.

(iii) Sequential numbers will be assigned to the units in the frame and the first unit for which the sequential number equals or exceeds r will be designated as selected; the first unit for which the sequential number equals or exceeds r+I will be the second unit selected, etc.

For example: In selecting the sample for stratum D9, I=100/28=3.571. The random number might be 2.218. The sequence of selection numbers would be 2.218, 5.789, 9.360,..., and the third, sixth, tenth, etc. units would be selected.

**Some people feel more comfortable with integral sampling fractions. Rounding to 1/4 for stratum D-9 and to 1/9 or 1/10 for stratum D-4 would be tolerable. (Reference: United Nations Recommendations for the 1983 World Programme of Industrial Statistics. Part Two, Annex I. Practical Sampling Techniques in Industrial Censuses. Paragraph 25.)

11.3 Procedures for Sampling From the Urban-low Stratum (UL)

Random systematic sampling at a constant probability for all units will also be used for the urban-low stratum, UL. The sampling unit records for the EAs show how many establishments, by size class, each contains. However, except for its use in defining the records assigned to this stratum that information will be disregarded because the range for the total number of establishments is small. Before the sample is selected the records will be arranged by province, and by city or town within province.

11.4 Sampling from the Urban-high Stratum (UH) and the Rural Stratum (R).

The cluster sampling units of the urban-high stratum (UH) and of the rural stratum (R) may include highly variable numbers of establishments. Such large variation in cluster size can seriously inflate the sampling variance of estimated totals. In order to offset that effect, sampling units will be selected from the urban-high stratum and from the rural stratum with probabilities proportional to measures of size (ppmos), where the measures of size (mos) are the estimated numbers of non-directory

establishments included in each unit.**

Before the units of the urban-high stratum are selected they will be arranged by province and by city or town within province. The units of the rural stratum will be ordered by EA number within province before sampling. (For rural sampling units that consist of two or more EAs, the lowest EA number of the combination will be used.) The following procedure will be applied separately to each stratum to select its sample systematically with ppmos:

(i) The units in the file will be numbered from 1 through N.

(ii) Cumulative mos totals, (mos_1), ($mos_1 + mos_2$), ($mos_1 + mos_2 + mos_3$),... will be computed and recorded.

(iii) The mean mos, $\bar{mos}$ (the total mos divided by the total number of units in the stratum) will be computed.

(iv) The sampling interval, I, will be computed by dividing $\bar{mos}$ by f.

(v) Every record in the file will be examined to see if its mos equals or exceeds I.

**[2]The design effect formula for cluster sampling given in Section 10 is a simplification which omits a term that reflects the variation in the size (number of establishments) per cluster. The simplified formula applies strictly only when the number per cluster is constant. It is a satisfactory approximation when the sampling is with probabilities proportional to size, or when the variation in size is small. The exact current sizes of the sampling units are unknown. Hopefully they are well enough correlated with the assigned measures of size for satisfactory results. The method used to derive the mos is described in the Appendix.

a. If any such records are found:

   1. They will be removed.
   2. The corresponding unit will be designated as a certainty unit.
   3. The number of certainty units so designated and their total mos will be recorded.
   4. The previous total number of units and total mos will each be reduced by their corresponding values for the certainty units.
   5. The operation will be repeated, starting at step (i).

b. If no records are found which have a mos equal to or greater than I:

   1. A random number r will be chosen, such that $0<r\leq I$.
   2. The sequence, r, r+I, r+2I,..., will be computed and recorded.
   3. The successive, individual mos will be divided by I to find the probability of selection of each unit. The probabilities will be recorded.
   4. The successive cumulative mos figures will be compared with the sequence recorded in b2. The first unit for which the corresponding cumulative mos equals or exceeds r will be the first unit selected. The first unit for which the corresponding cumulative mos equals or exceeds r+I will be the second unit selected, etc.

12. Estimating Totals, Sampling Variances and Relative Standard Errors

The formulas to be used to calculate the estimates are an integral feature of all sample designs. Formulas are needed for both estimates of totals and of their sampling variances and corresponding relative standard errors.

In choosing the formula for the estimates of totals two requirements were prescribed: First, that they should be mathematically unbiased (the average value of the estimates from all possible samples should equal the universe totals); second, that the estimates should be strictly additive (that all of the estimates - for all classifications - should sum to the same totals). Unbiasedness was prescribed because biased estimates derived from small samples may behave very erratically, and for many of the detailed estimates the number of contributing sample units will be small. Additivity was prescribed because totals which are inconsistent with detail or from table to table are confusing and irritating to data users even when the differences are statistically negligible.

Much less stringent requirements were imposed on the estimates of sampling variances. It was considered sufficient that they provide reasonable measures of the quality of the estimated totals. Additivity was not applicable, and unbiasedness was unnecessary. If biased, however, the variance estimates

preferably should, on average, overstate rather than understate the exact value. It was also decided that variance estimates should be computed for all planned publication levels, but for only the two measures, total persons engaged and value added. (Experience has shown that the relative standard errors for these two items cover the range of the standard errors for all the short form items quite well.)

12.1 Estimates of Totals

Section 6.4 gave a general formula for unbiased estimates of totals. The same formula applies for the estimated total of any particular category, c, such as an industry or a province, namely

$$X'_c = \sum_{i=1}^{n} X_i/p_i = \sum_{i=1}^{n} w_i X_i \qquad (12.1)$$

where

X'_c is the estimated total for category c,

n is the total number of establishments in the sample,

X_i is the value with respect to category c of the i-th sample establishment (and equals zero if establishment i is not a member of category c),

p_i is the total probability of selecting establishment i.

$w_i=1/p_i$ is the weight for establishment i.

This simple formula is to be used throughout to estimate the totals. As noted, it is unbiased, and it is completely additive and general. It can be applied to any group of sample establishments - an industry, a province, a size class, cross-classifications of those categories or any non-standard groupings - that might be desired. For purposes of deriving totals the weighted establishment values, W_iX_i afford the same flexibility in tabulations as a complete coverage data file. The stratum codes and even the sampling unit identification in the establishment records can be disregarded without affecting the results.

12.2 Estimating Sampling Variances and Relative Standard Errors

Estimates of the sampling variances cannot be computed quite as simply as the estimated totals. In this case the sampling unit must be taken into account. The sample establishments selected from the directory present no problems. Each such establishment is a sampling unit. For the geographic sample the entire cluster of establishments in each EA constitutes the sampling unit, and cluster totals must be developed as the first step in the computation of the sampling variances. Separate totals are needed for each computation category. This preliminary step is

necessary because the variances involve the squares of the sample unit values rather than the squares of the values of the separate establishments included in a cluster. This is a minor complication. No processing difficulties are expected, because the sampling unit identification, which will appear in every establishment record, will be used simultaneously with the category code as a control to develop the needed sample unit sub-totals.

After the needed sampling unit totals have been obtained, a software package such as PC-CARP (for micro-computers) might be used to compute the estimates of the sampling variances. Such programs are highly convenient and have only moderate requirements. PC CARP, for example, requires an IBM compatible micro-computer with a math coprocessor. It also requires the input data to be formatted in a specified way, and use of a supplementary program (such as Lotus 1-2-3) to convert the output into a readily readable and publishable form. None of these requirements are apt to be serious limitations.

Alternately, the estimates of the sampling variances can be computed by means of a custom program. A sample, approximate formula which can be used for the purpose is

$$\hat{s}^2(\bar{x}'_c) = \sum_{h=1}^{n} w_h (w_h - 1) \chi_h^2 \qquad (12.2)$$

where

$\hat{s}^2(X'_c)$ is the estimated sampling variance,

X'_c is the estimated total for category c,

m is the total number of sampling units (clusters and establishments as applicable) in the sample,

w_h is the weight for sampling unit h,

$X_h = \sum_{n=1}^{m_n} X_{hi}$

n_h is the total number of establishments in sampling unit n,

x_{hi} is the value for the i-th establishment of sampling unit h with respect to category c (and is zero if i is not in k),

x_h is the total value of sampling unit h for category k.

Notable features of formula (12.2) are, that like the formula for estimating totals, it does not require separate computations for each stratum, and that in particular it does not involve the stratum totals or means. These features offer considerable computational advantages over more standard formulas for estimating variances.

As might be guessed from the absence of a term for the total or

mean, Formula (12.2) is biased. The bias almost always is positive. Also it is small when the total is small and the sample size is large. These relations can be seen by comparing $\hat{s}^2(x'_c)$ with $s^2(x'_c)$, the standard formula for an unbiased estimate of the sampling variance for an estimated total as given in section 6. The standard formula would apply individually to each of the three strata, D9, D4 and UL which are to be sampled with constant probabilities. For each of these stratum g. c4 can be written in the alternative form

$$s_2(X'_g) = \hat{s}(X'_g) - w_g(w_g-1)[(\sum_{h=1}^{m_g} X_{gh})^2 - \sum_{h=1}^{m_g} X_{gh}^2]/(m_g-1). \quad (12.3)$$

At its maximum $s^2(X'_{gc})=[m_g/(m_g-1)]\hat{s}^2(X'_{gc})$, a value it attains only when $\sum_{n=1}^{m_g} X_{gh}=0$, which shows why the bias of $\hat{s}^2(x'_{gc})$ almost always is positive. The bias is small for many estimates because while m_g is the total number of units selected from the stratum, the X_{gh} equal zero for many of them.

For the two strata, UH and R, which are to be sampled with ppmos, we do not have an exact, compact formula for an unbiased estimate of the variance. To a first order approximation the variance is

$$s^2(X'_c) = \sum^{m} \left(\frac{1-p_n}{p_n} + \frac{1}{m} \right)\left(X_h - \frac{p_h X_c}{m}\right)^2 \quad (12.4)$$

The values of the p_h would vary around $p_h = mX_h/X$, and the term $1/m$ will become neglible for large m, so that roughly, on the average

$$s^2(X'_c) = \sum_{h=1}^{m} \left(\frac{1-p_n}{p_n}\right) X_h^2 (1-X_c/X)^2 \quad (12.5)$$

This is somewhat smaller than the expected value of $\hat{s}^2(X'_c)$, which is

$$E\, s^2(X' = \sum_{h=1}^{m} \left(\frac{1-p_n}{p_n} \right) X_h^2, \qquad (12.6)$$

Hence, $\hat{s}(X'_c)$ is apt to be an acceptable estimate on these cases as well.

For publication purposes estimating the sampling variances is merely an intermediate step. The measures of real interest are the relative standard errors. These will be calculated simply by calculating the square roots of the estimated variances and dividing those results by the corresponding estimated totals.

EXHIBIT 4-1. SUPPLEMENTARY ITEMS AND ASSOCIATED ITEMS FOR CALCULATING THEIR COVERAGE RATIOS

Supplementary Long Form Item		Associated Items		
Number	Description	Long Form Number	Short Form Number	Description
6c	Number of homeworkers paid [m]	6a(3)	6a(3)	Number of paid operatives, November
7b(1),(2),(3)	Fringe benefits	7a(3)	7a(3)	Annual payroll, total
8	Number of operatives by quarterly pay period	6a(3)	6a(3)	Number of paid operatives, November
9	Days worked by paid operatives	7a(1)	7a(1)	Amount paid to operatives
10(1)	Stocks, finished goods [m]	16(4)	9	Total value of shipments and and receipts
10(2),(3)	Stocks, work in process, materials, supplies, etc. [m]	12(7)	8	Total cost of materials, supplies, etc.
10(4)	Stocks, total [m]	n	n	s
11	Fixed assets acquired during 1993	16(4)	9	Total value of shipments and receipts
12(1) -(6)	Cost of materials, detail	12(7)	8	Total cost of materials
13(1) -(7)	Fuels, detail	12(7)	8	Total cost of materials
14(1)	Electricity purchased	12(7)	8	Total cost of materials
14(2),(3)	Electricity, generated, sold	n	n	z

Notes: m: manufacturers only, s: sum of detail estimates, n: name used,
z: zero is assumed value for short

ADDENDUM

A1. Introduction

The United Nations Recommendations for the 1983 Programme of Industrial Statistics includes this caution: "Unless the sample is designed to fit the prevailing operating conditions and is satisfactorily controlled, losses rather than gains may result from introducing it. A highly elaborate and theoretically desirable sample design that required more skilled personnel, more resources and more equipment than were available would be worthless or worse".*

This caution was taken very seriously in planning the sample design for the industrial census. The individual operations called for are neither complex nor difficult. The total number of distinct steps, however, comes close to overburdening the small, trained and experienced staff available to manage the work. Several compromises were adopted to lessen the control burden. Given more time for preparatory operations and more staff qualified to direct execution of the sampling plans a technically more efficient design would have been adopted. This addendum discusses some refinements and alternatives to the present design which might be feasible for a future industrial census.

* United Nations: Statistical Papers. Series M No. 71 (part 11). P 98, par. 5.

A2. Use of Supplementary Data to Improve Estimates

The estimating formulas presented in Section 12 involve only the current sample data that are to be collected in the census. Under favorable conditions modifying those estimates by data from other sources will be advantageous, i.e. will produce modified estimates which have smaller sampling errors than those of the initial estimates.

The supplementary data available for this purpose are limited. They require a fairly considerable effort to convert them to a potentially useful form, and the degree of improvement that might be achieved is uncertain. For these reasons the sample design for the 1993 industrial census does not provide for using supplementary data to develop the estimates. The importance of the methodology demands that it be discussed, with possible future applications in mind.

A2.1 Basic Concepts and Procedures

Estimates from a given probability sample may be greater than or less than their corresponding universe values, and the deviations may be large or small. The relationship for one item will be similar for other items that are reasonably well correlated with the first. Therefore if both the universe values and sample

estimates are available for a supplementary data set, the relationship between them can be used to adjust the estimates derived from the sample data currently collected in the census.

Commonly, in practice, the supplementary figures have been taken from a recent complete census, but good guesses can be satisfactory as well. The number of persons engaged figures, by strata, of Table 10.7-1 are believed to be fairly accurate. They can serve as the foundation for the supplementary data file. Estimates will be required in fine detail, so that it will be necessary to distribute the totals to individual sampling units within each stratum and by industry within each of the sample clusters of the geographic frame. Adapting the distribution given by the directory is the best that can be done. The specific details are given in the Appendix. It is not certain that the results will be good enough.

After the persons engaged figures by industry have been assigned to each sampling unit, universe totals and corresponding sample estimates can be derived for any desired industrial and geographic categories. They then can be used to adjust the basic estimates developed from the census sample reports.

A2.2 Ratio Estimates

Widely used to make such adjustments is the ratio estimate

formula:

$$Y''_r = Y'(X/X') \qquad (A2.2.1)$$

where

Y''_r is the adjusted estimate of the total Y,

$Y' = \sum_{i=1}^{n} Y_i/p_i = \sum_{i=1}^{n} w_i Y_i$ is the unbiased estimate of Y,

$X = \sum_{i=1}^{N} X_i$ is the total of item X (persons engaged),

$X' = \sum_{i=1}^{n} w_i X_i$ is the unbiased estimate of X, derived from the identical sample as Y'.

The ratio X/X' describes how much larger or smaller, relatively, the known total is than the sample estimate of that total. That ratio is a pure (dimensionless) number. It can be used successively as an adjustment factor for all the short form items. Whatever their definitions, if Y' and X' are well correlated and the sample is at least moderately large, Y''_r will have a smaller sampling error than Y'.

Its flexibility is a highly convenient feature of Y''_r, but ratio ertimates have limitations. First, they do not yield consistent tables. The sums of detailed estimates do not equal separately derived totals, the sum of estimates by industry will not equal

the sum of estimates by geographic areas, etc. Statistically the inconsistencies are insignificant. Nevertheless they are troublesome. Sophisticated data users find them annoying. Unsophisticated users do not understand them.

The second limitation of ratio estimates is more fundamental and more serious. They are mathematically biased. If totals are developed by summing detailed ratio estimates, the effect of the biases on the sampling errors cumulate, and becomes substantial. Also, when the samples are small, the bias is a significant component of the sampling err · Such biased estimates may behave erratically in relation to their universe values, and the standard descriptions of their behavior in terms of the normal distribution may be invalid. Many of the detailed estimates sought from the industrial census would be so flawed, because they will depend on samples that fall far short of being moderately large.

A2.3 Difference Estimates

The limitations of ratio estimates can be overcome by using an alternative adjustment formula, the difference estimate formula:

$$Y''_d = Y' - bX' + bX \qquad (A2.3.1)$$

where

Y''_d = the adjusted (difference) estimate of Y,

Y' and X' are the unbiased estimates of Y and X, as in (A2.21),

and

b is a predetermined factor for converting the values of item X to the level of item Y.

Difference estimates require more work than ratio estimates, but are additive and are unbiased for all definitions of Y, X and b. Their efficiency, like that of ratio estimates, depends heavily on the correlation between Y' and X'. When the correlation is high the difference adjustment can result in a substantially lower sampling variance for Y''_d than the variance of Y'. Small gains result when the correlation is low. To a lesser degree a good or poor choice of the factor b also affects the efficiency of Y''_d.

Separate factors, b, will be needed for every pertinent short form item. They reasonably can be set equal to the anticipated (modal) ratios of Y_i to X_i, for example, value of shipments and receipts per person engaged, cost of materials per person engaged, etc. The factors assigned should be suitable for small establishments, and should, as appropriate, vary from industry to industry and within industry by geographic location where

important differences are known to occur. (Urban-rural differentials may be the most important.)

After all the factors have been assigned, the analytic estimates, bX_i, should be computed for every item and for every industry of each sampling unit, and recorded. The file so created would be comparable to a complete coverage file. For convenience, we shall represent the values bX_i by the symbol Y_i, and shall refer to the full set of records as the complete, analytic estimates file. Tabulations to all levels of detail derived for the census then could be computed, with the results being stored for later use.

A parallel file of the records for the selected sample units also should be compiled. Estimated totals could be derived, and the differences from the complete analytic file totals, Y'-Y, could then be computed and stored. With that procedure the differences could be reviewed at an early date, thus allowing ample time to correct any important errors that might have been made.

When the current reports are received they can be processed independently of the analytic estimate records; estimates Y' can be developed; then the Y' can be combined with the previously derived differences, Y'-Y, to produce the difference estimate Y''_d.

It is worthwhile to produce at least some of the estimates by a second procedure as well. Formula (A2.3.1), expanded, is

$$Y''_d = \sum_{i=1}^{n} w_i Y_i - \sum_{i=1}^{n} w_i Y_i + Y = \sum_{i=1}^{n} w_i (Y_i - Y_i) + Y = \sum_{i=1}^{n} w_i D_i + Y \qquad \text{(A2.3.2)}$$

where $D_i = Y_i - Y_i$, is the difference between the current reported value, Y_i, and the corresponding analytic estimate Y_i.

Matching the current and analytic sample records (which would be necessary to apply formula A2.3.2) adds an important element of control. If any records in either file remain unmatched, something has gone wrong. The individual weighted differences, $w_i D_i$, can also be useful in checking for data errors. Large errors, due, for example, to transcription mistakes, are apt to result in extremely large (positive or negative) values of $w_i D_i$, which would stand out in an array of such values.

Lastly, it is important to note that in estimating the variance of Y''_d, the differences D_{hc} replace X_{hc} in formula (12.2), so that the estimate of the variance becomes

$$\hat{s}^2(y'_c) = \sum_{h=1}^{m} w_h (w_h - 1) D^2_{hc}. \qquad \text{(A2.3.3)}$$

The principal bias term of this variance estimate equals $(Y-Y)^2/m$, instead of the term Y^2/m applicable tc (12.2). Practically always $(Y-Y)^2/_m$ will be smaller than Y^2/m, so that the

bias will be correspondingly smaller, and the estimate of the variance correspondingly more accurate.

A3. Contribution of the Non-certainty Class to Different Totals

The proportion the non-certainty class contributes to any given total strongly influences the relative standard error of the estimate of that total. A single, all industries average proportion was used for the standard error and sample allocation calculations of Section 10. Higher than average proportions for particular totals by industry or province will raise the relative standard error from the desired level given by the quality-importance relation, $V(X'_c)=K/\sqrt{X_c}$. Lower than average proportions will have the opposite effect. To reduce those effects offsetting increases or decreases can be made to the non-certainty sample sizes for individual categories. For the purpose the measures of size of the sampling units involved can be raised or lowered before the sample allocation is determined.

It might seem desirable to introduce such adjustments at the finest classification level that affects the tabulation program, namely industry group by city or town and by industry group for rural areas of provinces. To do so, however, would require an elaborate series of calculations, and would involve a large number of finely detailed categories. Many of the results would

be quite erratic. Appreciable numbers of small sampling units would be elevated to certainty size solely because their categories do not include any large certainty establishments. For other categories, in which large, certainty establishments do appear, the adjusted non-certainty measures of size would be reduced almost to the vanishing point.

A simpler, and more satisfactory approach is to consider only the primary tabulation levels of detail by industry group and by province, as given in Exhibits I-6-2, I-6-3 and I-6-6. Adjustment factors can be calculated for each province and each industry group, as detailed in the Appendix. Then since every sampling unit simultaneously has geographic and industrial classifications, all possible pairs from the two sets of factors can be simply averaged to determine the applicable factors, A_{pi}.

Small adjustments are not worth making. Only those A_{pi} that lie below 0.84 or above 1.2 should be used as multipliers of their corresponding mos_{pi}.

A4. An Alternative Strategy for Sampling Urban Establishments

The plans for the 1993 Industrial Census assign the existing directory the primary role in the canvass of urban industrial establishments, and assign field listing the secondary role of supplementing the directory's coverage. The sample design that

was developed in accordance with those plans proved to be expensive. It calls for field listing 50 percent of the urban EAs that include substantial numbers of industrial establishments.

An alternative strategy, which might be better, is to reverse the roles of the field listing and of the existing directory. According to this plan, all urban EAs would be canvassed for purposes of compiling a current list of the urban industrial establishments. That list then would be defined as the urban frame. To supplement it, a sample would be selected from the existing directory, matched against the new frame, and reports sought from the unmatched directory sample as well as from the sample selected from the new frame.

Timing might be a problem under this "reverse" plan. When the canvass is based on the directory, much of the preparatory work can be spread over a long period of time. That advantage would be lost, if work on the canvass must wait for the field listing to be completed. Simply starting the field listing operation very early won't solve the problem, because new establishments then would be omitted from the field lists.

A second objection to depending on a field listing operation rather than the existing directory is that the quality of field operations cannot be controlled as well as the quality of office

operations. This is a far more serious matter when the field listing aims at covering all establishments than when it aims at covering establishments which in total account for less than 10 percent of all industrial activity. (For example, omitting 5 percent of 100 percent is far more damaging than omitting 10 percent of 10 percent.)

If the timing problems can be resolved and a reasonably thorough field listing can be obtained, there is much to commend using the newly compiled field lists as the sampling frame for urban establishments. Extending the listing operation to all urban EAs would a little more than double the cost of listing. Otherwise all the advantages are in favor of that change. Briefly, the following points may be noted.

No sample of urban geographic clusters would be needed. Hence, the sample design would be simpler and much more efficient. The new lists would cover the urban places completely, so that the cluster design effect would be eliminated. Size and industrial classification information would be more up to date and thus more accurate, so that stratification of the frame would be more effective. Of particular importance, the establishments of certainty size could be identified with greater precision, while the substantial percentage of establishments in the present directory that are out of business would not appear in the new frame. The allocation of the sample to the strata (including the

rural stratum) would be improved for its dependence on the estimated omissions from the directory would be very sharply reduced. It would be easier to develop the analytic estimates needed for difference estimates, and they would be more precise. Similarly, it would be easier to make adjustments for variations in the proportionate contributions of the non-certainty class. Without any doubt the gains in sampling efficiency would more than offset the additional listing cost.

A5. Matching

In addition to the costs and sampling variances of the two strategies, their comparative coverage biases should be considered. On the average the combination of directory and field lists will cover the industrial establishment universe equally well regardless of which list is primary and which is secondary. But this is not necessarily true of coverage bias due to matching errors, and the matching procedures that can be used and their likelihood of error, are not identical for both plans.

Under the plan developed for this census, the sample lists from the field canvass would be matched against the complete directory. Rigorous, conservative rules would be applied and strictly observed. After the records for the definitely matching pairs have been removed from the files, the remaining records would be rematched. Liberal rules would be applied at this

stage, and all potentially matching pairs identified. Those pairs, and the completely unmatched sample records then would be returned to the field for clarification and verification. (Experience indicates that critical identifying information often is incomplete in these groups of records.) Reports would then be collected from those sample establishments that remain unmatched.

Under the second plan another feature would be added. When the field listers record an establishment they would give it a uniquely numbered census registration card. They would enter the number together with the establishment's name on the card. The card would contain instructions that it should be kept, as well as some general information regarding the census. The first matching stage, in the office, would be the same as for the current plan. The registration card device has proved quite useful for the second stage, when unmatched directory sample records would be checked in the field to determine whether or not their establishments had been listed in the canvass. The card device, which greatly tightens control over matching errors, is not applicable when the field listings is for a sample that supplements the directory.

APPENDIX

Computational Procedures

This Appendix describes, in detail, the computational procedures for deriving:

(i) Analytic estimates of number of persons engaged per sampling unit.

(ii) Adjustment factor for variation in proportions of totals accounted for by the non-certainty class.

(iii) Measures needed for sampling with probabilities proportional to size.

(i) Analytic estimates of the number of persons engaged per sampling unit

The estimates of the number of persons engaged for each sampling unit should be consistent with the total estimates of Table 10 - 7.1, by stratum.

For the sampling units which are individual establishments of the directory frame, calculate the per sampling unit values simply by dividing the respective stratum total persons engaged estimates (5,200 for D9 and 6,600 for D4) by the numbers of directory establishments in the stratum. (Round to 1 decimal place.) The resulting low "average value" will reflect the fact that a number of the establishments in the frame will no longer be in business; that their current zero values will depress the overall average.

For the cluster sampling units of the geographic frame, per sampling unit estimates of the number of persons engaged are needed by industry. Values can be derived from the distribution by industry and size of the directory establishments in each sampling unit, and the expected proportions of non-certainty establishments, and the average number of persons engaged for each size class. The specific steps, which should be carried out separately for each stratum, are as follows:

(1) for each sampling unit, u, obtain the counts, N_{uiz}, the number of establishments in each industry, i, and each size class, z.

(2) Multiply each N_{uiz} by its appropriate size class weight, from the table below, to derive the weighted values, $(W_z)(N_{uiz})=X'_{uiz}$.

(3) Sum the X'_{uiz} over all size classes: $\sum_z X'_{uiz} = X'_{ui}$.

store the X'_{ui}.

(4) Sum the X'_{ui} over all industries and all sampling units of the stratum: $\sum_u \sum_i X'_{ui} = X'$.

(5) Compute the adjustment factor, A, by dividing the estimated total numbers of persons engaged for the stratum (13,700 for UH, 1,950 for UL and 9,550 for R) by X'. Round to 2 decimal places.

(6) Multiply the stored values, X'_{ui}, by A: $(A)(X'_{ui}) = X_{ui}$. Round to one decimal place. These X_{ui} are the final, per sampling unit estimates by industry.

APPENDIX TABLE 1

WEIGHTING FACTORS BY STRATUM AND PERSONS ENGAGED SIZE CLASS

Persons engaged Size Class	Stratum UH and UL	Stratum R
1-9	1.2	3.0
10-24	3.0	3.0
25-49	1.5	1.5
50 or more	0.0	0.0

(ii) Adjustment factors for variation in proportions of totals accounted for by the non-certainty class

For all manufacturing and mining industries combined (excluding petroleum refining) the number of persons engaged figures used to compute the size and allocation of the sample were, for the total, X, 289,400, and for the

non-certainty class, X_o, 37,000.*[1] Corresponding non-certainty figures by industry and by province, X_{iu}, can be obtained by summing the per sampling unit analytic estimates defined under (i) of this Appendix.

The corresponding total figures for manufacturing and mining industries, X_i, can be taken directly from Exhibits I-6-2 and I-6-6. For provinces get total figures by combining the specific province figures of Exhibit I-6-6 with the figures of I-6-5. Reduce the total for Lioala province by 3,000. Also, for purposes of computing the adjustment factors by province reduce the total, X, from 289,400 to 287,200. Otherwise disregard the 2,160 persons engaged in mining and not distributed by province. The resulting slight distortion is negligible.

Given the values X, X_o, X_i and X_{io}, the respective factors for each industry and each province, then can be found by calculating the ratios,

$$R_i=(X/X_i)/(X_o/X) \qquad (AX\text{-}1).$$

(iii) Measures needed for sampling with probabilities proportional to measures of size.

Total measures of size per sampling unit can be found simply by summing the analytic estimates, X_{ui}, over all industries within each sampling unit. The resulting X_u values are the measures of size needed to select samples with probabilities proportional to measures of size (ppmos) from the urban high and the rural strata. However, any necessary adjustments for the contribution of the non-certainty class should be made to those X_{uc}, which need them before the samples are selected.

---

[1] The 3,000 persons engaged in the petroleum refining industry were excluded because all four petroleum refineries are so large they will be included with certainty.

GLOSSARY

biased estimate an estimate which has an expected value, different from the universe value

certainty class the set of units which are always selected

cluster a set of units which are linked together into one larger sampling unit

cluster sample a sample of clusters

correlation the closeness of the relationship between two variables

cut-off the size which separates units into two classes for different treatment

cutoff sample design a design that calls for including all units which have a measure of size equal to or larger than a specified size, and no others

design effect how much the given sample design changes the variance of the estimate from the variance given by a simple random sample of the same size.

directory a list of establishment names and addresses (with industry and size codes)

difference estimate an estimate of the form $y''=y'-\dot{y}'+\ddot{y}$

EA abbreviation for the term enumeration area.

enumeration area a well defined and mapped geographic area which was used in the latest population census

expected value the average value of the estimates from all possible samples of a given design.

geographic sampling frame a list of geographic segments which cover the entire area in scope of the project

homogeneity the extent to which the individual units of a cluster are like one another

optimum sample design a sample design in which the probabilities of selection are assigned as efficiently as possible; the cost is a minimum for a specified quality (or the quality is maximized if the cost is specified)

probability proportional to size probabilities of selection that change in direct proportion to the measures of size assigned to the sampling units

probability sample design a sample design in which every unit in the universe has a known, positive probability of being selected

ratio estimate an estimate of the form $y" = (y')(X/X')$

relative standard error the size of the standard error of the estimate relative to the value being estimated

rel-variance an abbreviation for the term relative variance

relative variance the value of the variance relative to the square of the universe mean value

rel-variance of the estimate the variance of the estimate relative to the square of the universe value being estimated

sample any particular set of units selected from a frame

sample design the plan and method for selecting units and deriving estimates

sampling the process of selecting a set of units from a frame

sampling, fraction the number of units included in the sample divided by the total number of units in the frame, $f=n/N$

sampling, frame a list of units from which a sample is selected

simple random sample design a sample design in which all possible combinations of a given number of units have the same chance of being selected

standard error of estimate the square root of the variance of the estimate; an absolute measure of the dispersion of the sample estimates about their expected value

stratification the process of dividing the units of a frame into distinct groups or strata

stratified random sampling the process of selecting random samples independently from each stratum

systematic sampling selecting units at constant intervals from an array

ultimate cluster the subset of units, from a sample cluster, included is the final sample (the subset can be all the units of the cluster, or a sub-sample, depending on the design)

unbiased estimate an estimate which has an expected value equal to the universe value

universe all the units in the specified class; the universe of industrial establishments consists of all industrial establishments defined to be within scope of the industrial census

universe mean the universe total divided by the total number of units in the universe

universe total the sum of the values for all units in the universe

variance a measure of the variation among the values in the universe; specifically an average of the squared deviations of the individual values from the universe mean

variance of the estimate a measure of the variation among all possible sample estimates for a given design

BIBLIOGRAPHY

Brewer, K.r.W. and Hanif, Muhammad (983) Sampling with Unequal Probabilities, Lecture Notes in Statistics, 15.

Cochran, William A. (1963) Sampling Techniques. 2nd Edition

Hansen, Morris H., Hurwitz, William M., and Madow, William G. (1953) Sample Survey Methods and Theory

Stuart, Alan (1976) Basic Ideas of Scientific Sampling. 2nd Edition, Griffin's Statistical Monographs and Course No. 4

United Nations (1981) Recommendations for the 1983 World Programme of Industrial Statistics, Part Two. Organizations and Conduct of Industrial Censuses. Statistical Papers Series M No. 71 (Part II) Sales No. E81.XVII.12.