Manual on Surveys of Informal Employment and Informal Sector

Draft Chapter 7:

Independent informal sector surveys using the mixed household and enterprise survey approach

January 2010

1. Design of an independent mixed survey of the informal sector: technical requirements

1.1. Characteristic features and consequences for survey design

At core, the independent (or stand-alone) informal sector survey using the mixed household and enterprise survey approach (ISS) consists of a survey of economic units comprised of certain types of production activities of private households. While no such activity may be present in many households, where it does exist there is often a one-to-one correspondence between these economic units and households. Consequently, many features of the ISS design are the same as those encountered in a typical household survey. For instance, as in the case of a household survey, a survey of the informal sector deals with a population of numerous and small units. For such units, no explicit lists are generally available from which a sample can be selected in a single stage. In any case, sampling in a single stage is hardly a practical option because the units are widely dispersed in most situations. Hence it is necessary to resort to multi-stage sampling which imposes a sufficient degree of clustering on the sample obtained. Typically, a multi-stage sample involves the selection of area units in one or more stages, and then the selection of households (or other appropriate ultimate units) at the last stage of sampling. Areas are usually selected after stratification by type of place, region, and other basic geographical or ecological variables, often with probabilities proportional to some measure of size. Areas are more stable units, and generally existing frames can be used for their selection, possibly with some updating. By contrast, the ultimate units involved are much less stable and it is therefore necessary to freshly prepare (or use very recent) lists for the last stage of selection. In most cases, at least in the circumstances of developing countries, data have to be collected through face-to-face personal interviewing.

Despite such basic similarities, however, the sampling requirements of a survey of the informal sector, aimed at measuring household economic activity of various types, differ from those of a typical household survey of the general population in some fundamental respects. These differences arise from the fact that we are dealing with a population of units which tend to be less stable than typical households, are often unevenly distributed, and tend to be very heterogeneous. Furthermore, precision requirements, and possibly even the measurement objectives, may differ from one category of units to another.

Small-scale economic activity of various types, though widely dispersed, is often far from being uniformly distributed across the population. To be reasonably efficient, the sample design must take into consideration the patterns of concentration of various type of economic activity. The patterns of concentration can be complex and differ from one type (branch) of activity to another. These also differ by size and other characteristics of the units.

The units involved are heterogeneous by virtue of differences in the type or branch of activity, size, location and type of premises, criteria of eligibility for inclusion in the survey, the manner in which the economic units relate to the household, and many other characteristics. Apart from differences in numbers and spatial distribution of different types of units, each type may have its own sample size requirements. To some extent, the content and form of the information required may also differ.

All these factors contribute to increased complexity of the survey design, including sample selection and estimation procedures. While such complexity in the design cannot be avoided, it remains highly desirable that it is kept to a minimum. Even more important is the requirement that the survey procedures remain as uncomplicated as possible as concerns their implementation in the field. By their very nature, many operations in a large-scale survey of wide scope and coverage have to be decentralised.

In the following sections, a number of special features of ISS design are discussed:

- Sampling frame. The frame of area units should provide information on the type and distribution of informal sector activity, apart from information on the population and households.

- Stratification. An important objective is to isolate areas of concentration for different types of activities. This is in addition to the usual stratification of areas by urban-rural and region etc.

- Selection of area units. The most suitable units may differ in type and size from normal household surveys, and there may be differences in the methods of selection to ensure that the required numbers of different types of economic activities are captured in the sample.

- Listing. Several factors make the listing of households or other ultimate units in the selected areas a critical and relatively expensive operation. A large number of units may have to be listed to secure sufficient samples; information is required for screening, stratification, and differential sampling of different types of units; visible as well as invisible units must be listed to ensure good coverage; different types of units such as households and establishments may be involved, requiring different arrangements, and so on.

- Last stage of sample selection. More elaborate stratification and sampling procedures and differing sampling rates may be required to deal with different types of units.

Several related issues concerning sample implementation and estimation will be considered in section 2. These include special problems resulting from factors such as instability and changing characteristics of the units over time, inaccuracies in the sampling frame, the possibility of finding unexpected numbers or concentrations of units etc. Such problems may require special, sometimes non-uniform, procedures or adjustments. It is important to ensure that these do not distort the probability nature of the sample achieved. Another issue concerns the procedures for weighting and estimation. Unlike many household surveys, self-weighting samples are often not appropriate or achievable in an ISS. Also, external control totals required for ratio-type adjustments are usually more varied than simple population totals; even more importantly, the available external information is often incomplete and unreliable. Consequently, more complex weighting procedures may be involved, requiring carefully kept records of sample selection and outcome.

1.2 Sampling frame

As in the case of a household survey based on a multi-stage area sample, the sampling frame sought of an ISS is that of area units. However, several special requirements may be noted in comparison with an ordinary household survey.

- For good stratification and selection of areas as discussed below, it is necessary to have information on the number and distribution of economic units of various types in the areas. This is in addition to the information on population size and characteristics.

- • Good coverage is a more critical and difficult requirement in the case of an ISS. This is because results of the survey are usually more seriously and directly affected by coverage error than in the case of a population-based survey for which more reliable external information is usually available to adjust ratios and distributions obtained from the survey to estimate population totals. Such information is often lacking for economic variables. Furthermore, many household surveys are primarily concerned with estimating ratios and distributions rather than aggregates, while an important objective of any ISS includes the estimation of the total size (number of units, employment, output etc.) of the informal sector. 'An important consequence of this requirement is that clarity of boundary of the area units (natural divisions, accurate maps, stability) is much more important than the uniformity in size of the units.

- Uniformity in population size is also less important because the number of units of interest, namely economic units of different types in various sectors, are not necessarily closely related to the population size of the areas.

1.2.1 Population census frame

Many population censuses contain information on the numbers of persons by status in employment. Those classified as working for own-account or as employers provide an approximate measure of the number of small, informal sector units since, at least in absolute numbers, small-scale enterprises predominate over large units. To be usable, such information needs to be tabulated at the level of individual area units. Much more useful is the cross-tabulation of status in employment against the sector or branch of activity at the level of individual areas, since the sampling plan for the ISS requires to cater separately for different types of activity. The usefulness of the population census frame is further enhanced if it contains additional information on the nature of economic units. Two desirable items may be mentioned in particular.

- Information on the number of workers (especially regular paid employees) can help to separate out micro-establishments from household-based own-account activity, and these from larger establishments of different sizes. Such information is useful not only for an ISS, but generally in the planning of economic censuses and surveys.

- Information on the legal status and type of enterprise can help in identifying the target population for the ISS more clearly and precisely. Of course, such information is not easily (or always even wisely) collected in a census for the whole population. However, it should be recommended in situations where the census is conducted in two components: a complete count to obtain information on demographic and other basic characteristics of the population, and a large sample attached to the census to collect more detailed information on other variables, including economic variables.

1.2.2 Area frame from the economic census

Typically, the source of the frame is the population census. A recent economic census may be even more, pertinent for the ISS in so far as it contains more detailed information on economic activity which can be used for stratification and selection of area units. However, unless based entirely on the population census areas, an economic census area frame is usually not as complete in coverage and as good in the demarcation and mapping of area units as a population census frame. This is because of differences in the resources typically available for the two types of censuses. Hence there are advantages of basing the economic census on population census areas. In some countries, economic censuses have been conducted independently of the population census in the sense that the two do not use the same set of area units. Consequently, the quality of the area frame resulting from the economic census is not as good as that from the population census. Another common limitation of economic censuses is that their coverage excludes household-based enterprises and often also other small establishments. Both these factors tend to limit the usefulness of economic censuses as a source of frame for sample surveys.

A good way to overcome these limitations is to link the economic census with the house-listing operation of the population census, and widen its scope to cover not only small establishments but also household and other establishments lacking recognisable features. This for example has been the system adopted in India since her second economic census in 1980. A somewhat similar approach has been followed in conjunction with the 1990 population census of Indonesia: along with the house listing operation, information was obtained on the number of establishments with fixed premises (whether located within or outside private households) by sector of activity, while the actual economic census was to be conducted later. One of the explicit objective of this undertaking in Indonesia was to create a frame for surveys of household and small establishments.

1.2.3 Other sources

For special purposes, it may be possible to use alternative sources for the frame, or at least to supplement the main census-based frames. Certain types of activities may be concentrated in a few known locations which can serve as part of the frame. In certain urban areas, usable lists or registers of street vendors or other informal businesses may be available which can be aggregated over areas. Sometimes use can be made of lists of electricity users, especially if domestic and business use can be distinguished.

1.3 Stratification of area units

Apart from the usual reasons for stratification, stratification has a very specific added objective in the case of an ISS: to identify and separate out different degrees of concentration of different types of units. It is necessary to discuss this aspect of the ISS design in some detail because of its importance, and also because it has not been discussed adequately in the literature.

The objectives of this added stratification are:

- to increase sampling efficiency;
- to permit different designs and sampling rates for different categories of units;
- and to minimise the complexity and differences in the design at the last stage of selection for different types of units.

The last mentioned point is of considerable practical importance because it helps to simplify survey implementation in the field, notwithstanding the complexity of the overall design. The idea is to accommodate at the higher stages of sampling any differences in the sampling procedures required for different types of units, so that the units can be sampled in a more uniform way at the last stage of selection which involves large and decentralised operations.

The type of stratification possible depends on the information available. Such information need not be very precise or up-to-date to be useful for the purpose of stratification, so long as it is reasonably correlated with current relevant characteristics of the area units of interest. Also, patterns of distribution are usually much more stable than the fate of individual establishments.

If nothing more than population figures and physical areas are available, at least something may be gained from stratification by population density. With information on numbers of establishments as well, better stratification can be done on the basis of "economic density" defined for each area as the ratio of the number of establishments to the number of households in the area. The potential is greatly increased when the numbers of establishments are available by sector of activity, as illustrated below.

Consider a situation in which information is available on numbers of own-account and other establishments classified by sector of activity, for each EA or such area unit. Generally, such numbers can be expected to be closely related to the size of the informal sector in each area. The objective is to divide up the areas into non-overlapping groups or strata, such that each stratum represents the concentration of establishments belonging to a particular sector. To the extent a stratum 'captures' a large proportion of the establishments in the corresponding sector, the sampling requirements (such as selection rates or sample sizes) for the latter can be applied uniformly to the stratum itself. This increases efficiency, and can greatly simplify the sampling procedure by reducing the need to treat units in different sectors differently within any given stratum.

Various measures may be used to assign area units to different 'strata of concentration'. Classification simply in terms of the sector having the largest number of establishments in the area is unlikely to be useful: often sectors differ greatly in size, and large ones (such as small trade) tend to predominate in most areas. Sometimes the difficulty can be reduced by excluding from the exercise a very large and widely distributed sector for which it is not necessary or meaningful to identify the stratum of concentration. In any case, *relative* measures are likely to be more useful. For example, one may compare economic sectors within an area on the basis of the index

$$Ri = \frac{\text{no. of sector i establishments in the area}}{\text{average number of sector i establishments per area}}$$

and identify the sector having the largest concentration in the area relative to overall size of the sector. In practice, we have found the index to work better if the average in the denominator is computed after excluding areas with no establishment in the sector concerned. This avoids giving undue weight to small sectors unevenly distributed across the areas. Table 1 shows an application of the method. Areas, in this example small segments of census EAs, have been classified into one of four strata - trade, services, industry, and other (mainly construction and water transport) - depending on which sector has the largest value of the index Ri. In addition there are a large number of 'empty' areas in which no establishments were recorded in the census, though at the time of the survey they may contain some informal sector units. Rows of the table show strata of concentration; the degree to which they have been able to capture establishments from the sector with the same name is shown by the diagonal cells in the lower panel of the table. For example for sector 'industry', 67% of the establishments fall in the industry stratum, while 20% are scattered in stratum trade and 12% in stratum services.

Tables 2 and 3 illustrate the application of the method with further refinements introduced with two objectives:

- to improve stratification for the smaller sectors (services, industry, and other), possibly at the expense of the dominant sector (trade) which is bound to be well represented in any case; and
- to isolate areas with high density of establishments from those with low density.

The first objective is achieved by looking for the sector which ranks second on the index in each area to identify sub-strata, which can then be shifted across the original main strata to adjust their boundaries so as to better capture the smaller sectors. Such sub-stratification is illustrated in Table 2, and the adjusted strata after shifting some sub-strata (marked *) are shown in Table 3. To separate out high and low density areas, each sub-stratum was further divided on the basis of whether the total number of establishments in each area

TABLE 1: IDENTIFYING THE MAIN STRATA OF CONCENTRATION							
Number of units							
			SECTOR				
STRATUM	AREAS	HHs	All ESTs	TRADE	SERVICES	INDUSTRY	OTHER
1 TRADE	1246	47920	4878	4315	401	158	4
2 SERVICES	381	13577	1329	498	724	95	12
31NDUSTRY	183	6618	862	264	54	529	15
4 OTHER	9	440	74	1	2	4	67
9 empty	1552	31536	0	0	0	0	0
TOTAL	3371	100091	7143	5078	1181	786	98
Percentage distribution according to stratum							
STRATUM			All ESTs	TRADE	SERVICES	INDUSTRY	OTHER
1 TRADE			68.5	85.0	33.9	20.1	4.0
2 SERVICES			18.5	9.8	61.3	12.1	12.2
31NDUSTRY			11.9	5.2	4.6	67.3	15.3
4 OTHER			1.0	0.0	0.0	0.5	68.5

The code in column SECTOR of Tables 2-3 is as follow:

1st digit: 1= areas with below average number of establishments; 4= average or above.

2nd digit: sector with the largest value of the relative index R.

3rd digit: sector with the second largest value of the index.

exceeded the average per area for the whole population. The above is meant simply as an illustration; details will always depend on the actual data at hand. Incidentally, it may also be mentioned that to improve capture of the trade sector, one substratum, 412, was further split on the bases of whether the number of trade establishments in the area exceeded a certain value, in which case it was shifted to the trade stratum. This is just an example of ad-hoc adjustments which we are free to make subjectively at the stratification stage, prior to actual selection of the .sample, without affecting probability nature of the sample. As a result of adjusting the stratum boundaries, 80% of trade establishments for instance lay in areas included in the trade stratum, so that a fairly good control on the sample for the sector could be achieved by applying the required sampling rates and procedures to the corresponding stratum itself.

1.4 Sample size and allocation

Before deciding on a particular design and method of selection, it is necessary to consider the required sample size and the implied overall selection rates (sampling probabilities) for the ultimate units of observation and analysis. In any survey the required sample size is determined by numerous theoretical and practical considerations, which need not be discussed here in so far as the issues involved are common to any sample survey and not

TABLE 2. DISTRIBUTION OF ESTABLISHMENTS BY SECTOR AND INITIAL STRATUM

		Number of units							Percentage distribution according to stratum				
				SECTOR					SECTOR				
STRATUM	subSTR.	AREAs	HHs	All ESTs	TRADE	SERVICES	INDUSTRY	OTHER	All ESTs	TRADE	SERVICES	NDUSTRY	OTHER
1 TRADE	112	47	1671	141	94	47	0	0	2.0	1.9	4.0	0.0	0.0
	113	21	764	63	42	0	21	0	1.0	0.8	0.0	2.7	0.0
	119	767	21850	1176	1176	0	0	0	16.5	23.0	0.0	0.0	0.0
	412	194	12265	1946	1562	337	45	2	27.2	30.8	28.5	5.7	2.0
	413	50	2588	433	323	17	92	1	6.2	6.4	1.4	11.7	1.0
	414	1	49	4	3	0	0	1	0.1	0.1	0.0	0.0	1.0
	419	166	8733	1115	1115	0	0	0	15.6	22.0	0.0	0.0	0.0
total		**1246**	**47920**	**4878**	**4315**	**401**	**158**	**4**	68.6	**85.0**	33.9	20.1	4.0
2 SERVICES	121	96	2638	214	96	118	0	0	3.0	1.9	10.0	0.0	0.0
	123	34	1313	88	13	39	34	2	1.2	0.3	3.3	4.3	2.0
	129	135	3694	175	0	175	0	0	2.4	0.0	14.8	0.0	0.0
	421	97	5096	744	373	336	33	2	10.4	7.3	28.5	4.2	2.0
	423	15	582	80	16	39	25	0	1.1	0.3	3.3	3.2	0.0
	424	1	174	15	0	4	3	8	0.2	0.0	0.3	0.4	8.2
	429	3	80	13	0	13	0	0	0.2	0.0	1.1	0.0	0.0
total		**381**	**13577**	**1329**	**498**	**724**	**95**	**12**	18.5	9.8	**61.3**	12.1	12.2
3 INDUSTRY	131	48	1310	103	48	0	55	0	1.4	0.9	0.0	7.0	0.0
	132	3	68	9	0	3	6	0	0.1	0.0	0.3	0.8	0.0
	139	70	1531	82	0	0	82	0	1.1	0.0	0.0	10.4	0.0
	431	49	3084	569	206	36	327	0	8.0	4.1	3.0	41.6	0.0
	432	10	369	59	10	13	34	2	0.8	0.2	1.1	4.3	2.0
	434	2	232	25	0	2	10	13	0.3	0.0	0.2	1.3	13.3
	439	1	24	15	0	0	15	0	0.2	0.0	0.0	1.9	0.0
total		**183**	**6618**	**862**	**264**	**54**	**529**	**15**	11.9	5.2	4.6	**67.3**	15.3
4 OTHER	149	3	125	3	0	0	0	3	0.0	0.0	0.0	0.0	3.1
	441	1	59	6	1	0	0	5	0.1	0.0	0.0	0.0	5.1
	442	1	28	7	0	1	1	5	0.1	0.0	0.1	0.1	5.1
	443	2	108	49	0	1	3	45	0.7	0.0	0.1	0.4	46.0
	449	2	120	9	0	0	0	9	0.1	0.0	0.0	0.0	9.2
total		**9**	**440**	**74**	**1**	**2**	**4**	**67**	1.0	0.0	0.2	0.5	**68.5**
EMPTY	199	**1552**	**31536**	**0**	**0**	**0**	**0**	**0**	0.0	0.0	0.0	0.0	0.0
TOTAL		**3371**	**100091**	**7143**	**5078**	**1181**	**786**	**98**	100.0	100.0	100.0	100.0	100.0

TABLE 3. PERCENTAGE DISTRIBUTION ACCODING TO ADJUSTED STRATA

				SECTOR				
STRATUM	subSTR	AREAs	HHs	All ESTs	TRADE	SERVICES	NDUSTRY	OTHER
1 Trade	119	767	21850	16.5	23.2	0.0	0.0	0.0
	412.A	181	10884	20.9	23.4	22.2	4.7	1.0
	414	1	49	0.1	0.1	0.0	0.0	1.0
	419	166	8733	15.6	22.0	0.0	0.0	0.0
total		1115	41516	53.1	68.7	22.2	4.7	2.0
2 Services	112*	47	1671	2.0	1.9	4.0	0.0	0.0
	121	96	2638	3.0	1.9	10.0	0.0	0.0
	123	34	1313	1.2	0.3	3.3	4.3	2.0
	129	135	3694	2.4	0.0	14.8	0.0	0.0
	412.B*	13	1381	6.4	7.3	6.4	1.0	1.0
	421	97	5096	10.4	7.3	28.5	4.2	2.0
	423	15	582	1.1	0.3	3.3	3.2	0.0
	429	3	80	0.2	0.0	1.1	0.0	0.0
total		440	16455	26.7	19.0	71.4	12.7	5.0
3 Industry	113*	21	764	0.9	0.8	0.0	2.7	0.0
	131	48	1310	1.4	0.9	0.0	7.0	0.0
	132	3	68	0.1	0.0	0.3	0.8	0.0
	139	70	1531	1.1	0.0	0.0	10.4	0.0
	413*	50	2588	6.1	6.4	1.4	11.7	1.0
	431	49	3084	8.0	4.1	3.0	41.6	0.0
	432	10	369	0.8	0.2	1.1	4.3	2.0
	439	1	24	0.2	0.0	0.0	1.9	0.0
total		252	9738	18.6	12.4	5.8	80.4	3.0
4 Other	149	3	125	0.0	0.0	0.0	0.0	3.1
	424*	1	174	0.2	0.0	0.3	0.4	8.2
	434*	2	232	0.3	0.0	0.2	1.3	13.3
	441	1	59	0.1	0.0	0.0	0.0	5.1
	442	1	28	0.1	0.0	0.1	0.1	5.1
	443	2	108	0.7	0.0	0.1	0.4	45.9
	449	2	120	0.1	0.0	0.0	0.0	9.2
total		12	846	1.5	0.0	0.7	2.2	89.9
empty	199	1552	31536	0.0	0.0	0.0	0.0	0.0
TOTAL		3371	100091	100.0	100.0	100.0	100.0	100.0

TABLE 4. DETERMINATION OF THE REQUIRED SAMPLING RATES

DISTRIBUTION OF ESTABLISHMENTS BY SECTOR AND STRATUM

				Number of establishments by sector :				
STRATUM	AREAs	HHs		TRADE	SERVICES	NDUSTRY	OTHER	All ESTs
1 (Trade)	1115	41516		3484	262	37	2	3785
2 (Services)	440	16455		964	842	100	5	1911
3 (Industry)	252	9738		629	69	632	3	1333
4 (Other)	12	846		1	8	17	88	114
9 (empty)	1552	31536		0	0	0	0	0
total	3371	100091		5078	1181	786	98	7143
Required sample size (as example) :				500	200	250	50	1000
Implied sampling fraction (%)				9.8	16.9	31.8	51.0	

SAMPLE OBTAINED BY APPLYING UNIFORM RATE WITHIN STRATA

			f (%)	Resulting sample size (establishments selected)				
The original sampling rates :				TRADE	SERVICES	NDUSTRY	OTHER	All ESTs
	STRATUM	1 (Trade)	9.8	343	26	4	0	383
		2 (Services)	16.9	163	143	17	1	341
		3 (Industry)	31.8	200	22	201	1	456
		4 (Other)	51.0	1	4	9	45	109
			total	707	194	230	47	1178
With slight adjustments in strata 2-4 :								
		1 (Trade)	9.8	343	26	4	0	383
		2 (Services)	17.3	167	146	17	1	348
		3 (Industry)	34.8	219	24	220	1	499
		4 (Other)	54.4	1	4	9	48	116
			total	729	200	250	50	1229
With "Trade" enumeratee in only a subsample of areas in strata 2 and 3:								
		1 (Trade)	9.8	343	26	4	0	383
		2 (Services)	17.3	(1/2) *83	146	17	1	264
		3 (Industry)	34.8	(1/3) *73	24	220	1	353
		4 (Other)	54.4	1	4	9	48	116
			total	500	200	250	50	1000

special to an ISS. In the context of an ISS, two issues need to be discussed in particular: (i) the diversity of sampling rates required for different types of units in the same survey, and (ii) the problem of controlling sample sizes to achieve what is planned.

- In most circumstances, it is necessary to determine the sample size and hence the sampling rates required separately for different sectors. The sectors can differ greatly in size and sample size requirements. Different sampling rates are often also desirable for establishments of different sizes; for instance the number of micro-establishments, as opposed to own-account household economic units, is usually small, and the former often have to be sampled at higher rates to obtain adequate numbers. How can different sampling rates be applied to different types of units without unduly complicating the last stage selection process, which has to deal with different types of units in the same area?

- Several factors contribute to the difficulty of controlling sample size in an ISS. In many household surveys, the 'PPS method' of sampling area units is used to obtain self-weighting samples of households with a reasonable control on the sample size (overall and at the area level) at the same time. The success of the method depends on how well the measures of size used for PPS selection correspond to the actual sizes of the areas. However, the kind of information available in the frame for an ISS is often rather approximate because the sources of the area frame such as population or economic censuses tend not to be specially geared for the purpose of a survey of the informal sector. Control over sample size often requires departures from the desired sampling rates, and certainly from self-weighting. How to minimise this problem? Even more importantly, how to ensure that the procedures do not result in departures from probability sampling? Unfortunately, many times in practice non-probability procedures have been adopted.

It may be noted that sometimes simple approximate solutions are possible to the first requirement above. For example a survey in Sri Lanka covered both agricultural and non-agricultural households, and it was required to sample non-agricultural households at a higher rate. This was achieved by over-sampling the strata (districts) which contained such households in greater concentration. This resulted in their automatic over-representation *on the average*, thus reducing the need for differential selection at subsequent stages.

A more satisfactory and precise procedure is as follows. The correspondence between the sectors and strata of concentration means that the sectoral samples can be controlled to an extent by applying appropriate but uniform rates within strata, thus reducing the need for differential sampling at subsequent stages. Table 4 above provides an illustration based on the same set of data as the preceding tables. Assume that the sample sizes required by sector are as shown at the bottom of the first panel of the table, and that the total numbers of establishments by sector and strata shown in the table correctly reflect the actual numbers. This gives the implied over-all sampling rates by sector (last row of the top panel, Table 4). In practice, these rates are applied to the corresponding *strata*. The results are shown in the second panel. The resulting sample sizes by *sector* are not exactly as desired because the correspondence between the sectors and strata is not perfect. The next step is to adjust the strata rates to more nearly achieve the required sectoral sample sizes. Theoretically, this simply requires the solution of a set of linear equations by inverting a matrix. The required constraints are:

$$\sum N_{ij} \cdot f_j = n_i$$

with

N_{ij} as the number of establishments in sector i stratum j

f_j the stratum sampling rates to be determined and

n_i the required sectoral sample sizes

In practice some simple adjustment may suffice to obtain approximate but adequate results. For example, in the illustration very small proportionate adjustments to the strata sampling rates give the required sample sizes for the three smaller sectors (though such precision may not be actually necessary in practice because of approximate nature of the information available in the frame). The sampling rate for trade requires a more substantial downward adjustment. Without a large reduction in the rate for the trade stratum, this has been achieved in the illustration by reducing the rate for trade establishments in the services and industry strata. By applying this sub-sampling at the *area* level (i.e., by covering trade establishments only in a subset of the sample areas in the strata concerned), differential sampling can be minimised at the last stage of selection.

The above example does not control separately for units of different types, such as own-account enterprises versus enterprises of employers. It is difficult to avoid differential sampling at later stages for this purpose because of the lack of information on this aspect in most area frames. Furthermore, to control the overall sample size, it is necessary to have a idea of the relationship of the numbers of establishments (or some other measure of size) in the frame to the actual number of informal sector units expected at the time of the survey. The two numbers may be more or less highly correlated, but in general are not equal. The numbers in the frame may fall short of the actual numbers because of growth, but also because the frame may not fully cover the less visible informal sector units. Pre-testing, involving relisting in a small sample or areas of different types may provide some information on the overall relationship between the two numbers for different categories of units.

1.5 Selection of area units

In a multi-stage design, the required overall sampling rate in any stratum is applied in two steps: (i) the selection of area units in one or more stages; and (ii) the final selection of ultimate units within the selected areas.

1.5.1 Type of area units

Often census enumeration areas (EAs) are the most suitable basic area units for the type of survey under discussion. As noted earlier, the source may be the population or the economic census, depending on the type and quality of the information available.

What is the optimal size of the area units to be used for sampling? One can discuss this question at least in relative terms. Often it has been argued that, in comparison with normal household surveys, it is, desirable to work with larger areas in an ISS. Larger areas may have certain advantages: it is easier to find the required number of establishments of various types in larger areas; for the same reason, larger areas are more suitable where the same set of units are to be retained in an ongoing survey; larger areas permit larger sample takes per area, thereby reducing the number of areas to be taken into the sample and thereby reducing travel costs and improving supervision; and boundary errors are likely to be less problematic with larger areas, improving coverage which is a particularly important consideration in an ISS.

From a given frame of basic area units such as EAs, 'larger' areas can be created in two ways: by grouping geographically contiguous areas; or by introducing higher stages of selection, such as towns or villages as PSUs within which EAs are selected at the next stage. The balance of advantages can be different in the two situations.

The use of larger areas also has a number of disadvantages. Using fewer larger areas tends to reduce efficiency of the sample (larger design effects); this can be serious to the extent economic activities of the same type tend to be concentrated. If the sampling rates are not increased in proportion to the size of the areas, then the listing cost would increase; this can be a serious matter in an ISS where listing may already be a major operation (see below). A particularly serious problem with grouping areas (or introducing higher stages of sampling) is the resulting reduction in effectiveness of the primary stratification on the basis of patterns of concentration as described above. The rather good results in the illustration above arise partly from the fact that it is based on small areas (an average of 30 households per area). A major part of the ISS sample is expected to lie in urban areas, where travel cost is generally a less critical consideration, usually better maps are available, and often small pockets of high concentration of similar activities exist. All this argues for the use of small, compact areas for the purpose of sampling.

1.5.2 Sampling stages

From the above considerations, it appears that a two-stage sample with areas such as EAs as the primary units, and households (and micro establishments) as the second stage units is often the most suitable design. In some cases with particularly high concentrations, even smaller areas such as segments or blocks may be better, provided of course that suitable frames exist for such units. Grouping of EAs or introducing higher stages may be worth considering in rural areas, but especially in areas of low concentration of informal sector activities.

The above are of course general guidelines, and the actual choice must depend on particular circumstances.

Sometimes samples confined to a rather small number of cities or towns have been used. This is done for reasons of cost and difficulty in managing the survey in more than a small number of centres. Usually such designs are very inefficient. Furthermore, when the number of units involved is very small, the process of 'random' selection can easily become a mere pretence; there can for instance be too many areas which appear too 'important' to be left out of the sample. It is best to recognise it explicitly if the sample is really a purposive sample. Indeed, with a very small sample of PSUs, a thoughtfully selected purposive sample may actually 'represent' the population better than a random selection of a few units. As a rule, however, such designs should be avoided.

1.5.3 Selection of area units

As in the case of household surveys, the selection of area units with probabilities proportional to some measure of size (PPS) is generally a suitable method for an ISS. The difference lies in what constitutes a suitable measure of size. It may be the total number of informal sector economic units, or some similar measure correlated with that number. When the survey covers a number of sectors and different types of units within sectors, a composite measure of size may be formed by giving them different weights. For example, a multi-sectoral survey with special focus on industry may give a higher weight to industrial activity in selecting the areas. This will help in over-representing areas with more industry. Subsequently, even if different types of activities are selected at a uniform rate within each area (so as to simplify the last stage sampling), the average overall rate for industry will still remain higher.

Another factor to be taken into account is the density of economic activity, i.e. the ratio of number of economic units to the number of households in the area. It is more efficient to select more dense areas at a higher rate (for example with probabilities proportional to the square-root of density, apart from the usual proportionality to some measure of size), so as to obtain more of the sample from areas of greater density. This will lower the relative cost of listing, which is an important consideration in the case of an ISS.

Usually it is efficient to select areas systematically from lists ordered by variables such as location or density.

1.5.4 Low density areas

Areas with low density of economic activity generally require special treatment. It is often simpler to select such areas with constant probabilities rather than with the usual PPS scheme. This removes the need for special treatment of areas with too small a number of ultimate units (i.e. with too small a measure of size), something which is necessary when areas are selected with PPS. The constant probability scheme also facilitates the use of 'take-all' sampling at the last stage, which is convenient when each area has only a small number of ultimate units of interest.

There can be good reasons in low density strata to group areas or 'cluster' them through the introduction of higher stages of sampling. This helps in concentrating the sample and reduces the chance of not finding the required number of units in the area.

A decision has to be taken as to whether or not areas with no reported activity should be included in the sample. At the initial stages of implementation of an ISS, it may be prudent to exclude such areas in view of the high cost of covering them and the probably small effect of their exclusion on the overall survey results. However, such assumption of small effect may be quite invalid; indeed, one cannot always trust the information of 'no activity' in the frame. Where possible, the likely effect of their exclusion should be estimated (e.g. from a special investigation on the basis of a small sample). When the survey has been more established, it is desirable to extend it to include such areas even if at a low sampling rate.

1.6 Listing

The quality of estimates of aggregates (total number of economic units, employment, output etc.) from the survey depends on completeness of coverage, which in turn depends on the quality of listing. The requirement of good coverage is particularly important in the case of an ISS.

1.6.1 Segmentation, listing and sample selection

In a multi-stage sample, the ultimate units need to be listed only within the last stage area units selected into the sample. However it is generally necessary to prepare fresh lists prior to each survey to ensure that the most up-to-date situation is reflected. The period between listing and main survey interviewing should be minimised to the extent possible. At the same time it is generally desirable that the listing operation be organised *separately* from the main survey, ideally using different enumerators for the two operations. Two other operations also have to be accommodated: (i) possible segmentation of the areas considered too-large to be' completely listed, followed by selection of a sample of segments where created; and (ii) selection of the final sample after listing.

Segmentation is a difficult operation, and in many surveys a tendency has been noticed for the selected segments to be systematically under-sized, especially if the staff involved are those responsible for subsequent listing. This makes it desirable to separate listing from segmentation as well. The cost of adding-a new operation to the survey can be considerable. Another complication with ad hoc segmentation is that its introduction requires a revision of the last stage sampling rates if the overall rate has to be kept unchanged. All these considerations imply that a special operation for segmentation should be introduced only if it results in a substantial saving in the listing work.

To summarise, after the selection of sample areas, there may be up to four field operations - segmentation (perhaps on a selective basis), listing, sample selection, and enumeration of the final sample. Each step needs to be controlled and checked in turn, making it highly desirable that the steps are operationally separated to the extent possible. It is clearly desirable to design any survey such that the number of steps involved is minimised.

In an ISS, the listing stage has several objectives:

- to identify and produce a complete list of survey units;
- to obtain information on characteristics of the units listed so that those within the scope of the survey as defined by the eligibility criteria can be identified precisely (eligibility may be determined by a combination of several criteria);
- to obtain information for secondary stratification of the ultimate units;
- and to obtain other information required for sample selection.

In principle, a representative sample of informal sector economic units can be obtained by listing the activity of all members in a sample of households. However, a better alternative is to organise it in the form of a dual system combining the household and establishment approaches. The alternatives are described below in turn.

1.6.2 Listing through households

In an ISS, households form the core or basic units because a large proportion of the economic units of interest consist of own-account activity with one-to-one correspondence with households. Even when a number of activities exist within the same household, it may be unavoidable (sometimes even advantageous) to treat the household as a single integral economic unit, though in general it is desirable to list each activity separately. Completeness in the identification of economic units depends on the type and detail of questions asked at the listing stage. Ideally, this should involve the listing of all household members individually, and asking information on variables such as current activity status, type of activity, status in employment, secondary activity and location of work of each member, using a definite short reference period.

Given that the number of households to be listed can be several folds larger than the final sample to be enumerated, the cost of listing involving detailed questioning can be high. Consequently many surveys in practice use a much simpler (hence cruder) approach. For example, only a minimal set of questions (e.g. "Does any persons in this household have their own business or other activity?", followed by a listing of individual activities by type may be asked in an attempt to identify self-employed economic activity. However, with such a simplified approach, usually a heavy price has to be paid in the form of poor quality of coverage. Some surveys have used a more elaborate form involving explicit probing using a specified list of activities, though still at the level. of the whole household rather than that of individual members. A useful improvement would be to precede the above with a question on the main source of income of the household. Such a question, requiring a specific answer from each household rather than a simple 'yes-no' response, can be more effective in identifying economic activity of household members.

1.6.3 Listing of establishments

The identification of economic activity exclusively through households is obviously the appropriate approach for listing household-based informal sector activity. The approach is often less satisfactory in dealing with informal sector economic units which are located separately from the household. Such units are often fewer in number and may require special sampling and data collection procedures. Examples are micro-economic units employing one or more hired workers, or enterprises operated jointly by more than one household which are frequently located outside the premises of any one household. Concentration of establishments such as in the market place also falls in this category. The quality and coverage of such units can be improved by identifying and listing them using the establishment approach. Separate identification and enumeration of establishments located outside private households, in non-residential buildings in the sample area, also has a number of additional advantages: they can be sampled separately, possibly using different (usually higher) rates and procedures; subsequent interviewing at the location of work is facilitated; some estimation complexities such as those arising from an establishment being operated by more than one household can be avoided, and so on.

A purely establishment approach is of course not suited to identifying the smaller own-account activity units which have no fixed location or are otherwise not visible from the outside: hence the need for a combined (dual) household-establishment approach.

1.6.4 The dual approach

The idea of the dual approach is to divide the population of units into two categories, in principle non-overlapping and 'exhaustive: the bulk of smaller units which are best covered through a household listing operation; and units which require special treatment and are appropriately listed using the establishment approach. To define these categories clearly, the following three types of situations may be identified.

Type A	Establishments located, within the sample area, in a building or structure other than an occupied residential dwelling. The owner(s) of the establishment may or may not reside within the sample area.
Type B	One or more informal sector activities carried out within the household premises, owned and operated by persons resident in the household. Hired workers may or may not be employed. In principle, a variation is possible (Type B'): as above, but of a person not resident in the household, i.e. self-employment activity carried out in someone else's household. This is likely to be a rare case, and may be treated as if identical to Type B.
Type C	All other informal sector activities of persons residing in the sample area, carried out without a fixed or definite location, irrespective of whether the activity is conducted within or outside the sample area.
To keep the categories distinct, it is important to define in operational terms what is meant by a 'building', 'structure', 'occupied dwelling unit' etc. In particular, it should be made clear whether kiosks, stalls and other make-shift structures are to fall under type A (having a separate fixed premises) or type C.	

Listing involves the coverage of all structures in the area, whether residential or non-residential. All residential or mixed-use buildings are covered using the *household component* of the listing operation. It identifies all households, and all informal sector activity carried out by household members, obtaining information on the sector or type of activity, and its location and type of premises if any. Information on the size (the number of regular hired workers for example) and other characteristics for the identification of economic units and determining whether they are in scope of the survey is also obtained. From the list of informal activities so obtained, those of type A, i.e. conducted at a fixed location outside the household (or more strictly, outside the dwelling unit), are separated out and eliminated from the list. Instead, they are added to the second list described below *if* they are located within the sample area, after eliminating any duplicates in that list of course. In this way the household component of the list covers types B and C activities. (If needed, a question can be added to include here the residual category B' as well.)

Through coverage of all structures in the sample area the second, *establishment component* of listing identifies all establishments located in buildings other than occupied residential dwellings (type A). To these are added for completeness any missed establishments of type A located within the area which have been discovered through the household listing operation as described above.

The two components of the list can be kept apart and sampled and enumerated separately as discussed in the next section. As mentioned in Chapter 3, the dual approach can be extended to the use of different samples of areas for activities of (i) type A and (ii) types B and C.

1.6.5 Cost and quality

The main drawback of the dual approach of course is its higher cost. Also special care is needed to ensure that there is no double counting or gaps in the coverage. Despite these, there is much to recommend this approach in a good ISS.

A number of steps can be taken to reduce the size of the listing operation, for example: reducing the number of area units selected and increasing the last stage sampling rate, even to 'take-all' sampling; over-representing more dense areas (i.e. areas with more economic units for a given number of households) in the sample; and perhaps in some cases dropping areas with no or little economic activity from the study population. Generally, these features make the sample less efficient, i.e. increase design effects. In most cases however, some reduction in sampling efficiency is a less worrying factor than the non-sampling error resulting from poor quality of coverage. Coverage error has a direct and proportionate impact on the estimation of aggregates from the survey. The improvement of the quality (coverage) of listing should therefore be a primary concern.

1.7 Last stage of sampling

1.7.1 Basic practical requirements

Before discussing the specifics of sampling from the dual household-establishment lists, it is useful to note several desirable features of the design in relation to the last stage of sampling of households or economic units within the sample areas:

From a practical point of view, it is highly desirable that the process is uncomplicated. Selecting units in, different sectors of activity with different rates should be avoided if at all possible. It is better and often possible to absorb any required differences at preceding stages of sampling, which involve much smaller and better controlled operations. In any case, too much reliance cannot be placed on the classification of units on the basis of often approximate information obtained during listing. This by no means precludes the use of such information for stratification of the units by sector of activity prior to sample selection. (Indeed, experience shows that it is quite simple to design forms which divide the list into columns by sector, from which a stratified systematic sample is selected easily.) However, special procedures and sampling rates may be necessary for certain special categories of units, such as enterprises of employers, or more generally, units covered through the establishment approach as described above. When the procedures must be varied within the same area, such variation should be minimised. When an entirely uniform procedure is not possible, one should explore before introducing any further complications whether the objectives can be reasonably met by dividing the units into only two categories for the purpose of sampling: those which are sampled using the normal or uniform procedure; and others which can all be included in the survey without the need for sampling.

It is essential that the procedures adopted do not result in departures from probability sampling. The requirement of probability sampling is that each unit has a known and non-zero chance of being selected into the sample. There are examples in country surveys (and in even some international documents) where the desire to achieve certain specified sample sizes for various categories of units resulted in the adoption of procedures (such as the assignment of zero probability of selection for some units) which do not yield a probability sample. Given that the procedures and rates of sampling may vary by type of unit, a particularly important requirement is to ensure that records are kept of the number of units of various types listed and the number selected in each sample area, so that the sampling rates (and sampling weights to be applied at the estimation stage) can be computed.

A desirable feature of the design is good control over the sample sizes and workloads. In household survey practice sample areas are usually selected with PPS, and there is often a debate as to whether at the last stage it is preferable (i) to aim at a self-weighting sample, or (ii) a sample with fixed workloads. For theoretical as well as practical reasons the first option is generally favoured in normal household surveys, except in 'heavy' or repeated surveys where strict control of interview workloads is considered a critical requirement. The balance of argument is likely to be more in favour of the second option in the case of an ISS. As noted earlier, variations in sample sizes and workloads are likely to be a more serious problem in such surveys; consequently, the need to control these variations is typically greater. Samples tend to depart from self-weighting in any case because of the need to cater for different sectors and types of units.

1.7.2 Selecting the sample from a dual establishment-household list

The two components of the list can be kept apart and. sampled and enumerated separately. In the household component, the ultimate sampling units are households with one or more informal sector economic activities. If a household is selected, all its activities can be included in the sample, even though each activity is identified and enumerated separately at the subsequent stage. In the interest of operational simplicity, it is desirable that a uniform sampling procedure is applied to all households with one or more relevant activities, without distinguishing by sector or type of activity. Generally, the convenient arrangement would be to conduct the main interview at the location of the household, irrespective of where the activity is carried out.

In the establishment component, different sampling and enumeration procedures can be used. The appropriate unit of sampling will be the establishment, rather than the household. And generally, the most convenient arrangement may be to conduct the main interview at the location of the establishment. Several types of establishments are likely to fall within this component:

- Micro-units with one or more hired workers. These are usually few in number, and may be selected at a high rate or even taken all into the sample. If sampled, stratification by type and size can be useful.

- Concentrated establishments, such as in a market place. These can be stratified by type, and if too many in number, sampled at a low rate.

- Establishments operated by more than one household in partnership are most likely to fall in this category. Using the establishment rather than the household as the sampling unit avoids some of the estimation complexities which would otherwise be present in such cases.

- Other establishments. For simplicity, it is generally desirable that they are sampled in the same way as the household component, the difference only being that here the sampling unit will be the establishment.

2. Issues in measurement and estimation

This section considers selected problems relating to data collection and estimation. Most of these problems are common to the design and implementation of the various types of informal sector mixed household and enterprise surveys discussed in this and the preceding chapter (Chapter 6).

The issues discussed briefly include:

Scope of the survey. Is it preferable to cover all branches of economic activity uniformly in a single survey, or to organise separate branch-specific surveys?

Seasonality. How best to capture the seasonal nature of much of the informal sector activity?

Estimation. How to weight sample results, especially for the estimation of population aggregates such as numbers of units, employment and total output?

Special types of units. How to deal with situations where the units of observation and analysis differ from the type of units used for sampling.

Sample implementation. How to deal with common departures from the sample as designed, such as due to non-response, the failure to find sufficient numbers of sample cases, or finding unexpectedly large concentrations? Conditions and problems are diverse, and only a few salient points are noted on this topic.

2.1 Coverage of different branches of economic activity

The informal sector encompasses activities of diverse types. An important practical issue therefore is whether it is preferable (even necessary) to (i) cover these activities through a series of surveys each focussed on a relatively homogeneous subset; or (ii) to aim at the coverage of diverse types of informal sector activities in a single comprehensive enquiry.

Country experiences and practices differ in this respect. In a number of developing countries, surveys have been designed to cover all or most sectors of activities together (for example Tanzania, South Africa, Brazil, Mexico). In some cases, a distinction has been made between industry and related activities on the one hand, and trade and services on the other (Indonesia). In India, where the experience is more extensive and on a larger scale, a series of surveys have been conducted covering different sectors (see Chapter 5).

Reasons for considering separate branch-specific surveys may include the following:

1. A survey covering all branches may be considered too large and complex. A series of separate surveys may be operationally simpler and more manageable. Design simplicity of separate surveys can be a related advantage.

2. Different branches may require different types of information; modes of data collection may also not be identical. It may be easier for the purpose of data collection to cover the branches in groups of relatively homogeneous subsets.

3. Focussing on a subset of branches at a time also facilitates fine-tuning the design and operations to the specific requirements of each subset. This may be particularly useful for the purpose of stratification, and for adjusting the survey design to take into account

differing patterns of seasonality.

4. Different branches may be concentrated in different areas, reducing the advantage of covering them together.

5. Finally, the survey objectives may be limited, and may not require uniform coverage of all the branches.

However, there are a number of disadvantages of branch-specific surveys, in so far as the ultimate objective is complete coverage of all informal sector activity.

1. Firstly, separate operations are likely to be more expensive than a single combined enquiry. There are overhead costs in launching each separate operation. Apart from the increased cost of separate training, travel, supervision and data collection, a major disadvantage can be the increased cost of listing. Listing is in any case a major operation because of the fairly detailed and accurate information required for the identification of units in scope of the survey, and for the stratification and selection of the units.

2. The separate listing operation can be avoided only if the same set of lists can be used for the selection of samples for different branches. This already links not only the design and selection of the separate samples, but also the operations for their implementation. Common lists can be used only if the different surveys are spaced closely in time, which reduces any operational advantages which may be seen in their separate implementation. The need to use common lists also removes the possibility of using fine-tuned branch-specific sample designs.

3. In any case, more complex information is required at the listing stage for the clear and separate identification of different types of activities. In a combined design, such information is required only for distinguishing *groups* of branches which need to be differentiated for the purpose of sample selection and data collection. Furthermore, it is desirable (and often also possible) in a good combined design to minimise the need for such differentiation at the stage of final sample selection and enumeration - as detailed in section 1 of this chapter.

4. There are a number of serious *substantive* problems in conducting separate branch-specific surveys. Firstly, industrial classification categories may be difficult to identify precisely. It may be difficult in practice to assign complex and mixed informal sector activities to the standard list of categories. There can be problems in separating out different informal sector units within the same household. Also, a single informal sector unit may be involved in activities belonging to different branches. These problems become more serious when the enumeration of different types of activities has to be operationally separated.

5. There are also analytical complexities. For example, units included in a branch-specific survey may turn out at the time of enumeration to have activities outside the branches

covered in that survey. Similarly, units not included on the basis of characteristics recorded at the time of listing may in fact have activities in the branches which are meant to be covered in the survey. In an integrated survey covering all branches together, all types of activities can be enumerated and tabulated appropriately, irrespective of their classification at the time of listing and sample selection. It is much more difficult, if not impossible, to do so in separate surveys.

6. Perhaps the most serious drawback of separate surveys is the chance of *increased coverage error*. As a result of difficulties of the type noted above, it is hard to ensure that separate surveys cover the total population of informal sector units without duplication or omission.

7. Finally, it should also be noted that by design or due to practical and cost constraints, the separate surveys actually undertaken often do not add up to cover all the branches of interest in a uniform way. The picture of the informal sector obtained is therefore incomplete.

In conclusion, an approach covering all branches of economic activity in an integrated enquiry is clearly the desirable arrangement. Separate branch-by-branch surveys should be considered only when the interest is confined to certain branches, or when the scale of the combined survey is considered too large to be manageable on practical grounds.

2.2 Seasonal and other aspects of variation in time

A programme of informal sector surveys may involve one or both of the following two basic types of arrangements:

1. A survey on a continuing basis, aimed primarily at obtaining current estimates of levels and trends. The size and scope of such surveys may be limited to more important topics on which frequent and up-to-date information is required. Typically, the arrangement would be to attach such a survey to an ongoing labour force survey.

2. One-time or occasional surveys, aimed at obtaining information of longer-term interest, pertaining to the structure of the informal sector, and average conditions and patterns prevailing over a period of time. The survey size may be large so as to be able to produce disaggregated estimates for various domains or subpopulations. The survey content may also be more detailed than that of a more frequent or continuing survey.

Each type of survey has its specific requirements in relation to survey timing and its structure over the period covered.

For a continuing type of survey, representativeness of the sample over time is achieved by adopting a 'rotation pattern' in accordance with the type of estimates to be produced. The survey is typically organised in the form of an ongoing series of rounds, each round being designed to produce separate estimates covering a period defined by the required frequency of reporting. Detailed discussions of this aspect of survey design of labour force and other regular surveys are available elsewhere, and need not be elaborated here. It may be noted, however, that for an informal sector module attached to a labour force survey, it is not necessary that the two have the same frequency or duration of the survey round. The module may be attached selectively, for example during one particular quarter each year in the case of a quarterly labour force survey. It is also possible to apply the module to a subsample of each round, and then cumulate data over several rounds (e.g. months in a monthly labour force survey) to construct less frequent and extended (e.g. annual) rounds for the informal sector component.

The basic consideration in the case of an occasional survey is that, if data collected during a particular time are to be applied more generally to a longer period of interest, then the former should in some sense be representative of the longer period. By the same token, the survey period should be long enough to capture seasonal and other variations in time; for this purpose it is preferable to spread out the survey period to the extent possible. Furthermore, to estimate as well as properly average out seasonal and other variations, the total survey period should be divided into *sub-rounds*, over each of which a spatially representative sample is enumerated.

This basic requirement - of ensuring representativeness of the sample not only in space but also over time - is common to any type of survey arrangement, whether a one-time survey or individual round of a continuing survey. This issue merits some further discussion.

There are a number of advantages of dividing the survey period into sub-rounds. With a spatially representative sample for each sub-round, the total sample is representative in both space and time. Data from various sub-rounds can be cumulated to produce overall estimates; at the same time comparison between sub-rounds provides information on seasonal and other variations. Some selected variables can be estimated more frequently, even if with less disaggregation. The survey workload can be distributed more uniformly. The main disadvantage can be the increased cost of enumerating a representative sample for each sub-round separately.

The samples for the sub-rounds can be related in various ways: independent or entirely different samples at all stages; samples with overlapping areas, but different ultimate units; partial overlap in the ultimate units covered; and re-enumeration of the same ultimate units in all sub-rounds.

For an informal sector survey, the first arrangement, i.e. an independent sample for each subround, may generally be the most suitable. Typically the survey period may be spread out over a whole year, and divided into quarterly or shorter sub-rounds. Covering a different sample of areas in each sub-round is convenient for several reasons. The number of areas to be covered in each round is minimised. This also means that for a given sample size, the sampling rates within the areas are maximised. Consequently, an efficient use is made of the listings. The listing operation can be phased in the same way as the main survey, thus minimising the time lag between listing and enumeration. It is important to note that to measure seasonal and other changes at the aggregate level, it is not necessary to have overlapping samples from one period to another, though overlaps generally increase precision of the estimates. Non-overlapping samples between sub-rounds are preferred because they maximise the efficiency for aggregating the sub-round results over the whole survey period.

In so far as the pattern of seasonality differs geographically, it is necessary to ensure that each major geographic domain is covered in the sample for each season or sub-round. The same applies to urban-rural differences. This can be done through appropriate stratification and selecting the sample for each sub-round to include areas from each stratum.

It is equally important to ensure that different types of activities are properly represented in each sub-round. The pattern of seasonality can be markedly different for different types or branches of activity. Sometimes it is argued, quite incorrectly, that the survey for a particular type of activity should be timed to coincide with the period at which the activity peaks or is concentrated. However, the results will be biased unless in estimating annual aggregates account is taken of the slack periods as well. In the overall estimates, all periods must be represented equally. This by no means implies that the *sample sizes* need to be the same: in fact it is more efficient to select larger samples for periods of more intense activity, but then the results must be appropriately weighted before aggregation. If a survey is confined to a particular time of the year, then information on seasonal variation has to be obtained though retrospective questioning.

To measure gross changes at the level of individual units, it is necessary either

(a) to use a long reference period, such as covering the whole year; or

(b) to enumerate the same units repeatedly, such as in every sub-round, using a shorter reference period.

The same applies to measuring at the individual level variables (such as annual income or output) which by definition require data pertaining to an extended period.

In the case of an informal sector survey, neither of these options may be particularly suitable. Given the type of units and the nature of the information sought, retrospective questioning with long reference periods is often inappropriate. But enumerating the same sample repeatedly drastically reduces the sample size available for the production of annual estimates. Often it is already difficult to produce such estimates with the required precision and level of disaggregation.

On the other hand, the follow-up of a small subsample over an extended period can be fruitful. The objective of such a follow-up would not be the production of overall estimates of gross changes (for which the sample may be quite insufficient), but a study of the dynamics of change in the informal sector.

2.3 Weighting and estimation

2.3.1 Basic requirements

A number of special requirements concerning weighting and estimation may be noted in the context of informal sector surveys.

1. In general, informal sector surveys require the selection of different types of units at different rates, often the differences in the rates being quite large.

2. Complete coverage of the informal sector units is difficult to achieve, and coverage errors may differ markedly among different types and sizes of units and different branches of economic activity. This can distort the distribution of the resulting sample according to important characteristics of the population.

3. Response rates may also differ by type of unit.

4. Apart from estimates of proportions, means, ratios and distributions etc., the estimation of various aggregates (such as the total number of units, employment and output) for various categories is usually a basic objective of the survey. These estimates are often directly affected in proportion to the magnitude of the coverage error.

5. Often external information on the size of the population surveyed for different types of units is lacking. This makes the effect of coverage errors more serious and adversely affects in particular the reliability of the estimates of aggregates.

6. Certain types of units of observation and analysis lack one-to-one correspondence with the units of sampling, and thereby require special treatment at the stage of estimation. This last mentioned issue will be considered in the next section.

The weighting and estimation procedures have to be devised to reduce the impact of such problems on the survey results.

2.3.2 Systematic approach to weighting

Sample weighting is introduced for several reasons, such as to take into account selection probabilities, under-coverage, non-response, and other factors resulting in departures between sample results and the corresponding information about the population available more reliably from other sources. When sample data are to be weighted, it is highly desirable to follow a systematic, step-by-step procedure which separates out the different aspects of weighting. The weights to be applied may be calculated at each step in a series, the final weight to be used being the product of the weights at individual steps. This allows one to examine the correctness and the effect of each step in the weighting procedure. The basic steps include the following.

1. A set of **design weights** to compensate for differences in selection probabilities. These may also include adjustments for known large errors in coverage.

Basically, the design weights are inversely proportional to the selection probabilities. They produce what may be termed as 'simple unbiased estimates', meaning that the estimates are produced directly from the survey results on the basis of units' selection probabilities, without recourse to data external to the survey. Even though these initial estimates may be adjusted or refined at subsequent stages, they provide a test of the quality of the sample design and implementation.

2. A set of **non-response weights** to reduce the effect of different response rates in different parts of the sample and among units of different types.

The basic procedure is to divide the sample into a number of 'domains' and apply a uniform correctional weight to all units in a domain. The weights are inversely proportional to the domain response rate. Such weights can be defined only on the basis of characteristics which are available for both the responding and non-responding units, which generally means on the basis of the information available in the sample lists, though other sources of information such as results from a previous enumeration may also be available in special circumstances. Since any informal sector survey requires fairly elaborate information for the identification of units to be included in the survey, it is usually possible to use a variety of variables to adjust for non-response. One should choose variables the categories of which capture large differences in response rates.

3. One or more sets of **external weights**, introduced with the objective of making the distribution in the sample on important characteristics agree with the same distribution in the population as available from some more reliable external source.

Such adjustments based on external information are particularly important when the sample size is small, rates of non-response are high, there are serious departures from probability sampling in the selection or implementation of the sample, or the sample as designed cannot control the distribution of the outcome on certain important characteristics. The common lack of reliable external information by variables of interest in an informal sector survey limits the use of external weighting.

4. **Inflation factor(s)**, used for blowing up the sample results to estimates of population aggregates.

The previous steps involve weighting in relative terms, such as on the basis of the relative selection probabilities, relative response rates, or relative distribution by certain characteristics. The weights can be scaled arbitrarily. Generally for convenience and clarity, it is best to scale the weights at each stage such that the average per unit weight is

1.0. The inflation to population aggregates requires absolute values, and it is good practice to keep it as a separate step in the weighting process.

There are two basic ways in which a population aggregate may be estimated:

- As a simple unbiased estimate as in (1) above, except that actual rather than relative values of selection probabilities are used in the denominator.

- In the form of ratio-type estimates, where a ratio from the sample is inflated on the basis of an estimate of the denominator obtained from some more reliable external source.

In multistage samples, particularly when the sample size is not large and good coverage is difficult to achieve, the first type of estimates can be subject to large variance and bias. This unfortunately may be the case in informal sector surveys because of the difficulty in enumerating small, unstable and not always easily identified units. However, the use of ratio-type estimates requires the availability of relevant and more reliable external information than that available directly from the sample survey. In the case of informal sector units such information is usually limited. One should try to make the maximum use of whatever relevant information is available. For instance, total size and characteristics of the population estimated from the households listing may be compared with more reliable external estimates when available, and the survey data adjusted as appropriate. It may be possible to make similar checks and adjustments to economic characteristics of the population obtained during the listing operation. However, it is important to emphasise that sample data should not be adjusted automatically and indiscriminately: external information should be used only when it is clearly more reliable than what can be expected directly from the survey.

2.3.3 Trimming of extreme weights

The problem of large weights can sometimes turn out to be important in an informal sector survey. It is possible in informal sector surveys, perhaps more so than in usual household surveys, to encounter unexpectedly large concentrations or numbers of units at the enumeration stage. This often requires taking a reasonable subsample and subsequently weighting the results upwards to compensate for this. However, it is desirable to avoid assigning extremely large weights to any units in the sample. The use of large and variable weights, even if affecting only a small part of the sample, can result in a substantial increase in variance. It is a common practice therefore to trim extreme weights to some maximum value to limit the associated increase in variance. The justification for this procedure is that the effect of any bias introduced due to arbitrary trimming of extreme weights is smaller than the benefit arising from reduced variance.

2.4 Units requiring special treatment

This section discusses several cases requiring special treatment in situations when units of observation and analysis lack one-to-one correspondence with the ultimate units of sampling.

2.4.1 Several informal sector units in the same household

As a sampling issue, this presents no special problem. If a household is selected all informal sector units in it can be taken into the sample, and each unit receives the same selection probability (and hence the same design weight) as the household to which it belongs. The problem can be that of data collection. Ideally in an informal sector survey each separate economic unit should be the unit of observation and analysis, but in practice it may not be possible to separate out multiple economic units in the same household. By definition, informal sector units are unincorporated (not legally separated from the household as an economic entity), and maintain no separate-accounts. Indeed, it has been argued that all economic activity of a household should be treated as a single integrated whole for the purpose of data collection and analysis (cf. paragraph 12 of the 15th ICLS resolution). This may be unavoidable in practice in many situations. Nevertheless, for proper analysis of the structure and functioning of the informal sector, it is desirable to try and obtain, separate information on each type of activity in the household.

2.4.2 Unit with several types of activities

Again, this is not a problem in terms of sampling. If a unit has been selected into the sample, details of all its activities can be enumerated as recommended in paragraph 13 of the 15th ICLS resolution. All data on the unit are given the same weight - determined according to the procedure for selecting the unit. At the stage of tabulation and analysis, however, different types of activities of the same unit may appear under different classifications such as different branches of economic activity, depending on their characteristics. It is not necessary for these branches to correspond with the sector or stratum from which the unit was selected, though the sampling weight is always determined in accordance with the latter.

2.4.3 Changes in the type of activity and other characteristics of the units

The above argument applies to this situation as well. The sampling weight of a unit is always determined by how it was selected on the basis of its characteristics as determined at the time of listing. Information pertaining to the unit is always tabulated according to its characteristics as determined at the time of enumeration.

2.4.4 Units with partners in different households

Informal sector enterprises may be owned or operated in partnership. Even if the overall proportion of such enterprises is small, the proportion may be significant in certain branches of economic activity or in certain areas. Consider a situation in which a partnership enterprise is selected through a sample of households, but its partners reside in different households. The problem is to compute the probability with which the enterprise appears in the sample, given the probabilities of selection of the associated households.

There is no problem if all owners of a partnership reside in the same household, since in that case the unit appears in the sample with the same probability as the household. There is also no problem if such units are selected into the sample directly using the establishment approach - as for example may be done in an informal sector survey for units located outside of private households (see section 1.6). The problem arises when the owners of a partnership unit reside in different households, and the units concerned are taken into the sample on the basis of the selection of associated households.

As a sampling issue, the last mentioned situation is in fact similar to that in an economic survey, where the units of sampling and observation are generally individual establishments, but in a proportion of the cases information can be collected only at the level of enterprises each with multiple establishments. Hence the following technical discussion has a wider relevance than the context of an informal sector survey alone.

Basically, two approaches are possible to the problem. One approach is to establish a rule on the basis of which each observation unit is associated with one and only one sampling unit, in which case the former's selection probability (and the corresponding sampling weight) is the same as that of the associated sampling unit. For instance, for each partnership unit the 'main' owner may be identified and the unit taken into the sample only if the household of the main owner is selected, irrespective of whether or not the households of any other partners are selected. (One may note for general interest that similarly in an economic survey, an enterprise may be included only if its principal establishment is selected.) Technically this approach is the simplest, but has a number of practical problems. The unit may be run on equal footing, so that there is no 'main' owner. Or there may be differences as to the relative position of the partners, or even as to whether someone is a partner, a helper, or an employee. An additional major difficulty is that the fairly elaborate information required to identify partnerships, the partners, and the main one among the partners has to be obtained at the *listing stage*, prior to the main interview.

The second, preferred approach is to take the partnership unit into the sample if *any* of the households of its owners is selected, and appropriately determine the former's selection probability (and sampling weight) on the basis of the known selection probabilities of *all* the owners' households associated with it. This approach is more practical because information for the identification of whether the unit is operated as a partnership and on the number and location of the associated households is not required for sample selection, and can be collected more easily and accurately during the main interview for the selected units. Another possible advantage follows from the fact that the approach automatically gives an increased chance of selection to partnership units, which is often desirable because of the small number of and special interest in such units.

The selection probability of a partnership unit is computed as follows.

(i) Consider first a simple situation in which a simple random sample of households is selected at a small constant rate **f**. For a unit with partners in **n** households, the probability of appearing in the sample is

$$x = n.f.$$

The expression involves an approximation, which is of little consequence if **(n.f)** is small.

(ii) The approximation in (i) results from the fact that it double-counts the joint selection probabilities of the partner households. A more precise expression, again assuming a simple random sample of households, is

$$x = 1 - (1-f)^n$$

which differs from (is smaller than) the above only in terms of order (n.f) squared. This effect is generally negligible in an informal sector survey of small and numerous units. (Incidently, this may not be the case in a survey of large establishments.)

The above expression is obtained as follows. The same logic will apply to other relationships derived below. The probability of a household *not* being selected in a simple random sample is **(1-f)**; therefore, that of *none* of the **n** households being selected is **(1-f)^n**; and that of at least one of the households (and hence the associated partnership unit) being selected is its complement, **1-(1-f)^n**.

(iii) The difference between (i) and (ii) can be important in the more realistic situation of a *multi-stage* design, because the joint selection probabilities of the households in the same area unit can be quite large. Consider a two-stage design, with random sampling at uniform rates at both stages: areas selected at the rate **a**, and households within an area at the rate **p**, giving the overall rate as **f = a.p**. With an argument similar to the above it follows that for a unit with **n** partners *all residing in the same area*, the selection probability is **x = n'.f** with **n'**, which replaces **n** in (i), given by

$$n' = [1 - (1-p)^n] / p \; ; \; x = n'.f = a.[1 - (1-p)^n].$$

In practice the last stage sampling rate, **p**, can be quite large; indeed, in some sectors all relevant units in an area may be taken into the sample **(p = 1)**. Table 5 below compares **n'** with **n**, for different values of **n** (the number of partner households in the same area) and **I=1/p** (the last stage selection interval or inverse of the within area sampling rate). Clearly, with all households taken (**I=1**), the probability of selection of the unit is not affected by the number of its owners residing in the area (i.e., **n'** is always =1). For small **I** (large **p**), the difference between **n** and **n'** is marked, i.e. the probability of selection of the partnership unit is grossly over-estimated by (i).

Table 5: Factor by which the probability of selection is increased for a unit with n partners from different households in the same area

	I = 1/P					
n	1	2	3	5	10	20
2	1.00	1.50	1.67	1.80	1.90	1.95
3	1.00	1.75	2.11	2.44	2.71	2.85
4	1.00	1.88	2.41	2.95	3.44	3.71

With varying probabilities of selection for households (j) in an area, the expression can be written more generally in terms of the product of individual(**1- $\mathbf{p_j}$**) values:

$$\mathbf{x = [1 - \pi_j(1-p_j)].}$$

(iv) Another generalisation required is for the situation when the partners do not all come from households in the same area. First considering only the $\mathbf{n_i}$, partners coming from a particular area **i**, we can compute $\mathbf{x_i}$ as above. Considering all areas in this way, the final selection probability of the unit is

$$\mathbf{x = 1 - \pi_i(1-x_i)\ .}$$

(v) The above expressions have been developed for the case of a two stage design with random selection of units at each stage. A similar logic can be used to extend them to more complex designs, such as designs with more than two stages or stratification. Often the added complexity may not be worthwhile. In applying (iv) to a design with more than two stages, the term 'area' should be taken to mean units selected at the last area stage of sampling.

2.5 Sample implementation

Unfortunately, it is not uncommon to find examples of more or less serious departure from the standards of probability sampling in the design and implementation of informal sector surveys. Several factors may contribute to this state of affairs.

Often samples are not designed to ensure good control over the sample size, particularly on the sample takes in individual areas. There are also difficulties in ensuring such control due to the lack of relevant and accurate information on the numbers and characteristics of units in the sampling frame. Another contributing factor is need to ensure that minimum sample targets are achieved for each of a number of different types of units as required: the problem is not only that of overall sample size but also of the sizes required for different types or branches of activity. Large variation in the sample sizes from individual areas is a related problem. Finding large and unexpected concentrations of units contributes to this loss of control, as do large and variable rates of non-response. Often decisions to adjust survey procedures to deal with these problems have to be taken ad hoc in the field, at lower levels of the survey organisation. These adjustments can result in departures from the standards of probability sampling if the procedures are not correctly formulated and controlled.

The identification of some common errors should be helpful in reducing their incidence. Here are a few examples.

Sometimes it is argued that the survey timing should coincide with peak periods of the activity or activities to be covered. This can be a biased procedure. As noted earlier, it is better to spread out the enumeration over time so as to capture seasonal and other variations.

The same argument applies to substitution for activities found to be dormant at the time of the survey. The substitution of inactive units by active units obviously results in over-estimation of the extent of activity at any particular time.

Sometimes quotas are fixed in order to achieve a pre-specified, fixed sample size for different categories of units. For instance, a survey is carried on until a certain sample size is achieved and discontinued thereafter. Since the sample areas are not covered in a random order, such a procedure would normally result in a non-probability sample.

Another example with similar results is provided by certain procedures for the selection of units at the last stage. For instance, in a survey a specified number of units of a certain type is taken from each sample area, if the area contains a sufficient number of units of the required type. Only otherwise are units from other categories taken to achieve the required quota. Clearly this does not yield a probability sample for the last mentioned categories of units.

Taking fixed sample sizes rather than fixed sampling rates increases the problem, though this is sometimes unavoidable if adequate control cannot be achieved otherwise.

Finally, uncontrolled substitution for non-responding units is a common source of bias in informal sector surveys, as in many other types of surveys.

Many of the problems during sample implementation arise from the undue importance given to fix sample sizes and sample takes from individual areas in absolute terms, when steps have not been taken to ensure such control in the design itself. In this connection, a few important points may be noted.

1. It is neither necessary nor useful to aim at absolutely fixed sample sizes. A considerable amount of variation from the 'target' sample size can usually be accommodated without much difficulty and effect on the survey results.

2. Beyond a certain point of course, it becomes necessary to control such variation. Variation in sample size is often a more serious problem in informal sector surveys than in certain other types of surveys, because of the diversity of the branches of economic activity and the types of units to be covered and the specific requirements for each type. A balance is required between accepting these variations and making adjustments to the sampling process at the ultimate stage.

3. Often it is possible to reduce the unusual problems of variation in sample sizes by adopting an appropriate survey design. In an informal sector survey, it is desirable to introduce the necessary variations and controls at higher stages of sampling to the extent possible, so that the need to introduce special measures at the stage of final sample selection and enumeration can be minimised.