

SOME PRACTICAL SAMPLING PROCEDURES FOR DEVELOPMENT RESEARCH – Ian Wilson

*Statistical Services Centre, University of Reading, P.O. Box 240, Harry Pitt Building,
Whiteknights Rd., Reading RG6 6FN, UK. Phone 0118 931 8034, e-mail i.m.wilson@rdg.ac.uk*

INTRODUCTION

This paper is about how, despite appearances to the contrary, elements of the quantitative approach to sampling can be applied to certain sorts of mixed mode and qualitative research, and aid their generalisability.

The perspective

It is written from the point of view of statisticians familiar with teaching, and providing help to, non-statisticians. Our advisory input is often at stages of the research process when informal qualitative studies have already been done. Whether or not we were directly involved, we have, as long as we can remember, regarded mixing of methods, especially sequential mixing - as natural and inevitable. Early-phase work involving scoping studies, consultation processes, prioritisation and hypothesis-generating activities should be informal – what we now have to refer to as qualitative, and in the right circumstances as participatory. Our perception is that in these more exploratory studies, so-called qualitative sampling is quite reasonably seen as staking out the limits of little-known territory, contrasting with the later-phase detailed work to achieve proper coverage which is often associated with descriptive statistical work, e.g. hypothesis-testing. It is when the qualitative work is the main approach to this later-phase work that we see some potential for strengthening the sampling practice.

The sequential use of qualitative methods for in-depth follow-up work after broader, shallower quantitative work, we have usually seen as having an 'obvious' sampling set-up defined or informed by the latter.

LIMITATIONS OF TRADITIONAL STATISTICAL THEORY

Sampling Theory

What one can call 'statistical sampling theory' is a peculiarly – and regrettably – esoteric black art: most of the serious books are packed with formulae of surpassing boringness and very limited direct usefulness except sometimes to the initiate. The starting points it assumes are usually impossible hurdles for the development practitioner – both in terms of

the readers' mathematical hardiness (or foolhardiness) and in terms of the assumptions underlying the theory. A huge part of this theory is concerned with estimation of a single numerical quantity, not of course the most enticing objective to the qualitative researcher. Unsurprisingly, then, many researchers ignore the gobbledygook and jump straight in – with quite variable consequences: further on, this paper tries to suggest a few slabs to pave a middle way.

Random Sampling

Of course, for those who get exposed to sampling theory, the first big idea is that of random sampling. What people usually seem to recall later is that you must give every member of the population an equal chance of being picked – and this immediately loses many would-be researchers when they can't enumerate the entire population, or when the units are not all equal e.g. big households and small ones. You must indeed give every member of the population an equal chance of being picked if you want to work out the optimal properties of some mathematical estimators, but the important practical idea is both more basic and more interesting.

This is that *in order to produce conclusions that can reasonably be defended as generalising to a larger population, you must set up an objective (in principle a repeatable) procedure for sampling*, which prevents the implementers from choosing their informants in undisclosed and subjective ways – ways which invisibly bias the results, by unknowable amounts. A clearly stated procedure, with careful arguments in its support, is referred to below as a *protocol*; a good one allows a fair assessment of what the sample does represent.

Another kind of claim frequently made for random sampling, but all too often over-interpreted and misunderstood, is that the results can reflect very precisely the population mean. This is the claim best supported by statistical theory, but not the most important way in which qualitative conclusions need to be backed up.

Furthermore the claim to precision is true only if you meaningfully can, and successfully do, (a) elicit meaningful and accurate values, (b) sample at random, and (c) have a very large sample. A large sample allows us to select units disregarding the many complicated ways in which they may be characterised, e.g. human respondents' levels of education, social connectedness, access to loans etc. These will – as the sample becomes huge – be averaged out in the overall picture. There is no such claim at all if the sample size is very small or if we don't want to ignore the characterisations.

Small samples are all too common, because of lack of resources, but small sample sizes often do more harm than they need – and achieve less than they might – because of poor disposition of the resources available. This theme is developed further below as we look at some of the realities of sampling as perceived in our professional practice, especially in advising or reviewing DFID projects.

HIERARCHIES OF UNITS

Information in hierarchies

Very frequently, social research reflects a broad social setting where for example we might be concerned, say, with interviewing children as individuals, the primary caregiver or the head as representing the child's household, the head of the village development committee to convey formal information about the community. Here there is a hierarchy of units – child, household, community, and maybe more levels. If one is concerned with the impact of social policy on children, say, we also need to add some monitoring of district-level implementation and of the level where policy is defined and legislation enacted.

Of course, hierarchies are not unique, e.g. there are different representations of a community other than the holder of an officially defined post. However, it is important to recognise that information exists at different levels, and to think systematically about linking them. The largest sampling entities are referred to in the quantitative literature as 'primary sampling units', the next level down as 'secondary' and so on till we reach the 'ultimate sampling units'. For example an attempt to trace the impact of social and economic policies on the welfare of children needs to link levels, and trace pathways, from the policymaking arena down to its effects on individual children, the ultimate units.

Example

This is being attempted in the DFID SSRU Young Lives¹ project. In that project, as well as policy monitoring we talk about four levels where both qualitative and quantitative information is collected: the site (sometimes 'sentinel site'), the community, the household and the individual. The 'site' is an administrative entity (a mandal in Andhra Pradesh, a commune or two in Vietnam) from within which we recruit 100 index children aged between 6 and 18 months of age at enrolment.

¹ That particular study poses other demands too, e.g. it needs longitudinal collection of information, but this must not obscure proper consideration of the levels.

Resource Allocation in Hierarchies

In our limited scrutiny of qualitative research design literature, there seem to be quite a number of sources that do not properly acknowledge how hierarchies should be considered in study design². There is very often a resource trade-off that has to be thought about rather carefully – more sites means less information per site, and sometimes – as with the policy-tracking objective, the selections at one level bear critically on those at another. One of the main messages is that sampling strategies are incomplete and incoherent unless they define reasoning and procedures at all necessary levels. In the following we assume there is some hierarchical structure, and that this implies collecting information and addressing objectives at two levels at least.

Example

In reviewing DFID research projects, SSC have strongly criticised the design of studies where a couple of sites (each comprising a handful of villages) have been selected, and very intensive studies conducted within them – to the tune of a third of a million pounds in some cases – to give very large (indeed unmanageable and excessive) amounts of site-specific information. In these cases the sample contained two primary sampling units and some hundreds of secondary units, i.e. households.

The salient elements of our objections to such designs were that the sampling at the level of primary units was poorly conceptualised and poorly justified, e.g. (a) no attempt was made to demonstrate where the two sites sat in any sort of relation to other places, and (b) the sample of size two was (beyond doubt) unable to represent the larger reality about which DFID or government might usefully have been informed. In one case at least, the two primary units were chosen within a large and important agro-eco-system – fine – but the main feature recorded in the final technical report was that several other sites were rejected because they did not have the ‘right’ combination of features, indeed that the chosen sites were of a type that was quite hard to find. So what did they represent³? There were no confirmatory studies where conclusions from the two main study sites were even briefly calibrated against the realities of other places. Note that

² In statistical parlance, the standard terminology refers to multi-stage sampling, but that term is avoided here as it carries a good deal of baggage irrelevant to our purpose as well as misleading overtones of multiple time points.

³ This question has remained with the author for 38 years. Working briefly as a statistician in a marine biology lab he was told, while helping a struggling scientist, “only one crab in 100 is typical.”

the work done within sites was argued to be of wider general relevance i.e. to be research, rather than being defended as locally relevant empowerment, for example.

The conclusion thus is that too much effort has been expended *within* and too little *between* the primary sampling units. We would have advised, or gone along with, (i) a not-very-different first phase with a few sites, and quite a lot of households, then (ii) an additional light-touch second phase where some key results and predictions from phase (i) were shown to apply to a range of other sites.

The argument in this paper is largely about research funded because of a claim or implication, maybe not an explicitly stated promise, of generalisability to a domain that a body like DFID regards as important. Note that there was no distinction made in the comments above between work done within sites that was quantitative, and that which was qualitative or indeed participatory. The argument in this paper is largely about issues about evidence, rather than about the qualitative or quantitative tools used to gather it.

Protocols for Hierarchies

The usual situation when sampling in a hierarchical setting is this. At the top level, the primary units are large entities about which a great deal is public knowledge e.g. a geographical region is characterised – with varying degrees of relevance and accuracy – as to its climate, transport, ethnic mix, employment types, social problems, local government personalities and so on, and on. In contrast, the bottom level of the hierarchy, the household or individual, is probably unknown to outsiders until after recruitment. Sampling at the top level involves selecting a very little of the vast amount of possible information and using this to provide a rationale for particular choices.

So-called 'random sampling' has little or nothing to say at this level, and attempting to use it usually entails discarding or ignoring readily available information which might quite logically guide one's choices. In the author's view, doing so in order to make a claim to 'statistical' generalisability is usually bogus. For a start the sample size of primary-stage units, e.g. the number of sub-districts, would have to be very large. A sample of two sub-districts is still a miserable little sample of size two regardless of how many households are interviewed in each one. Any claim to objectivity, generalisability or broad usefulness will usually depend on a well-argued case for the selection made. A good protocol, at the top level of a hierarchy, will discuss the information that could be used as the basis for selection, will

explain which of it is relevant, recent and reliable, and will illustrate clearly how the most appropriate information is used. It will admit that practicalities constrained the choices, and will assess the impact of restrictions so imposed.

At the bottom level, say the household, the external reader of a research report will never meet child Reddy, aged 8, and is weakly positioned to interpret a researcher's unexplained procedure for selecting him. The contents of the protocol at this level are a quite separate issue: the protocol needs to assure the reader about different things, e.g. that researcher Mrs. Rao recruited child Reddy following a purposeful, well-defined procedure, with adequate assurances of objectivity as one part of the requirement – alongside issues of ethics, qualifying characteristics and so on. The report of the implementation of this 'ultimate sampling unit protocol' should again acknowledge, and assess the effects of, practicalities, such as the hit rate, in judging what the achieved sample represents e.g. if 11 children had to be approached to get the sample of size 10 to which child Reddy belonged, that has a different evidential implication than if 10 could only be assembled by approaching 60. If the hit rate is a lot less than 100%, good practice demands some attempt at characterising the misses, and the impact on the conclusions of their omission or non-compliance.

TYPES OF OBJECTIVE

The 'Numericalness' of Objectives

Crusading zeal for sampling rigour is often misplaced. Its importance does depend on the topic, as well as the prevailing culture of the audience.

If the objective of a study is to provide an accurate overall picture, sampling rigour is of course important. One exercise in the evaluation of the 1999-2000 Starter Pack 2 programme in Malawi was a set of 54 village censuses. Interestingly, this involved careful 'statistical' sampling of villages combined with participatory mapping exercises to define village boundaries and membership. One objective was to generate an estimate of the rural population, for comparison with the census figure for about the same time. The villages were of differing sizes, and selected from different regions and districts, themselves of differing sizes, so the scaling-up process required very careful consideration of how to weight up the numbers found. The weights used were worryingly varied in size – the smaller the base of quantitative data, the greater the weight that had to be used to scale it up and the greater the potential for a small value to influence the total.

Many of the other components of SP2 evaluation involved research that, even though quantitative, was much less critically numerical e.g. questionnaires and semi-structured surveys where people across the country recorded hungry periods, food preferences and the like. Most of these were reported, largely in tables, with no form of weighting. Why? (i) Without full censuses, there were village size elements missing so weights would have had to be guessed; (ii) in many cases the data was much more contentious than just household sizes and we were unsure of its uniform reliability, so we wanted to suppress any heavy weights for fear of exaggerating the influence of poor data; (iii) in the main the results were designed to illustrate patterns of response, not numerical estimates; (iv) we needed the widest possible range of analysts and users to understand the data and feel some ownership of the outputs. To minimise the maximum weight given to any one response, we encouraged the research teams to eschew all weights and produce tables as if the surveys had been nationally-representative simple random samples. The key objective was to convey honest overall views, through carefully-constructed verbal descriptions.

Thinly anonymised, one research proposal to DFID, reviewed by SSC, proposed to go to one of DFID's favoured countries and conduct "about 20 PRAs" in "a selection of locales", with a view to telling DFID the most important problems facing farmers in areas where NR research might help them. We objected that implicit in, but crucial to, the words "most important" are numerical scores which could be scaled up to represent varying numbers of farms, or maybe different-sized agroecological zones.

COMPARATIVE SAMPLING

Comparison as opposed to Estimation

Many studies are concerned, not primarily with prevalences or population sizes, but with comparisons e.g. between before and after or between project areas and 'untreated control' areas. This usually imposes sample size requirements that sharply disagree with those for estimation objectives whether the information collection process is quantitative or more qualitative: in a simple instance, a population split roughly 80:20 between two important strata should usually be sampled proportionately to produce an overall estimate, but to compare the two groups one with the other would usually suggest 50:50 sampling so both groups are characterised equally well and the differences are as clear as possible.

Structured comparisons

In our experience it is quite common to find that there are many possible situational factors and distinguishing features that might be taken into account in a process of comparison. At the simplest, each of these might again distinguish population members into two groups. For example at the top level in a hierarchy we might have districts with high or low proportions of scheduled caste populations, with good or bad transport infrastructure, with early or late implementation of the District Poverty Initiative Programme and so on. Similar multiple characterisations can appear at each of several levels in a hierarchy.

There is a great deal of theory in the statistical literature on experimental design, little mentioned in sampling theory, but relevant here. For the simple example above, we might propose a study⁴ in four sites at the level of choice of districts, to get a bit of information on the above three forms of groupings:-

Transport	High % Scheduled Castes		Low % Scheduled Castes	
	DPIP +	DPIP –	DPIP +	DPIP –
Good	—	Include	Include	—
Poor	Include	—	—	Include

Note that this element of design does not sit alone. It does not imply that the study at the next level of the hierarchy, within each chosen district, is qualitative or quantitative, cross-sectional or longitudinal, nor does it say how the study is structured within districts. We do not explore this theme further below, but note that the above is the merest mention of what can become quite a complicated balancing act, when the levels are combined. It may be evident even here that some of the body cells of the table will be more common in the population than others, and we want to include relatively significant ones.

COMBINING AND CONTEXTUALISING STUDIES

Multiple studies

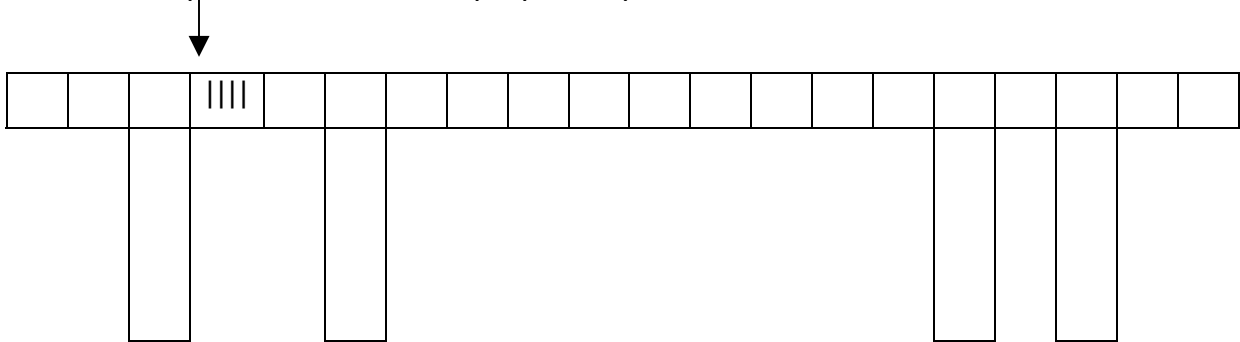
It is of course often the case that individual studies are part of a larger whole, where different themes are to be developed and explored over a period, as in many DFID development projects. Exercises needing more involvement and time commitment from

⁴ A half-replicate of a factorial design

local people (and the researchers) will be restricted to smaller numbers than the briefer, shallower sections of the work e.g. responding to a short survey.

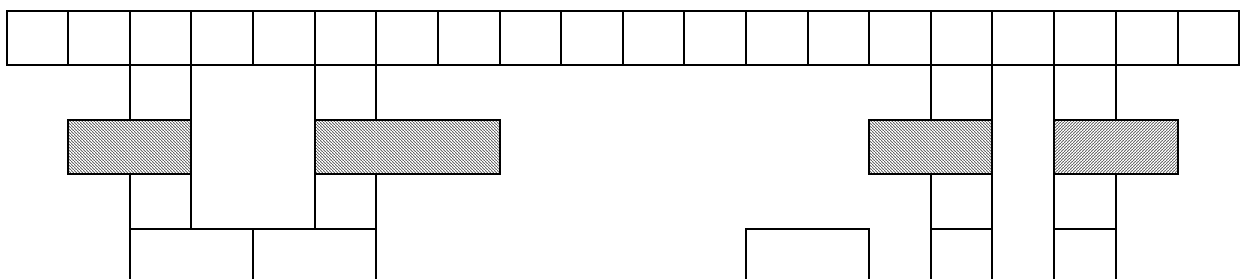
A simple version of this, which we refer to as the “table-top”, typically involves a relatively large number of people across a good number of places as respondents in a short questionnaire-based study at a relatively early phase, then following this up with more intensive, deeper – probably qualitative – work in a few communities selected partly on the basis of the first study.

Each box represents a number of people sampled within each of a number of communities



This provides some ‘breadth’ which can justify the selection of the in-depth study sites as being somewhat representative with respect to survey findings, rather than being ‘just case studies’. There is an element of ‘read-through’ of data from the first phase quick, shallow study to the deeper second phase, which may give some time-related information.

If a project, e.g. a DFID development project, does carry out a succession of studies, the above table legs may cease to be unitary and become more like piles of stones, each ‘layer’ of stones representing a component study, participatory discussion, on-farm-trial etc.

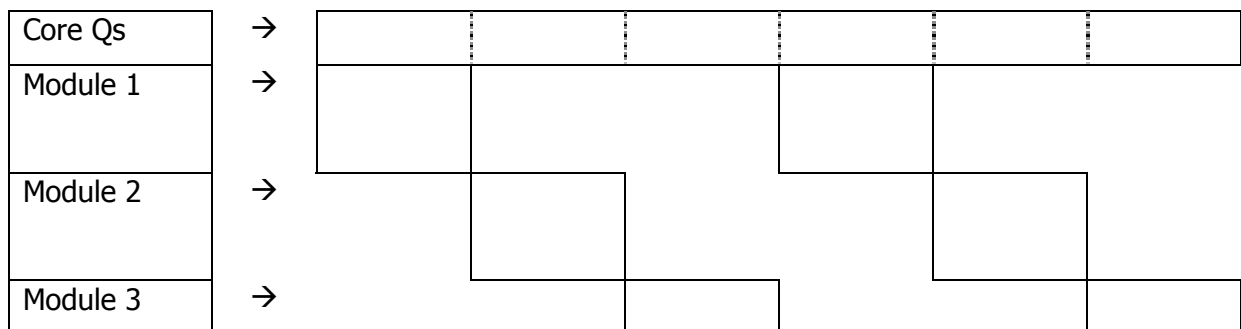


The greatest interest of the entomologist may be in the ‘horizontal view’ of the shaded layer representing data from a study of pest management strategies, for example.

If there is the 'read-through' suggested by the 3rd, 6th, 16th and 18th columns, then it is worth noting that the combination data from these compliant (and maybe much-burdened) communities, provides a 'vertical view', which with very careful thought about the studies, might be the 'livelihoods' view of the population.

Segmenting a study or articulating several?

Very often a single survey, whether formal and quantitative, or less formal and more open, is effectively a combination of several segments, each exploring a particular theme. There is a tendency to treat the survey instrument as a unitary entity and to assume that every respondent must 'answer every question'. Especially where there has been a multi-disciplinary team involved in the design of the instrument we have seen survey instruments that have become far too cumbersome and burdensome. It appears there is often scope to prioritise within such surveys, saying that a core of questions must be answered by all, whereas some other themes can be adequately covered by asking the questions of only a subset of the respondents:-



Of course this assumes an effective study design so that that analysis themes in module 1 do not require joint analysis with those in module 2.

Another way of thinking about the above is to say that each of modules 1 – 3 in the example above is a separate study, maybe done by different researchers or at different times, but that the three studies have by design agreed on a set of common-core questions that all will agree to use, with adequate commonality of training, interpretation, and approach that the set of responses to this core can be amalgamated to strengthen its evidential power.

The above corresponds to a simple generalisation of the commonplace method in censuses where a percentage of the households receive a 'long form', the rest just a 'short form' subset. This creates an L-shaped dataset. The diagram above has multiple Ls.

RANKED SET SAMPLING

All the above notions are concerned with having both a relatively large group as the grounding for a study, and smaller groups which are more intensively studied when selected from that context. A different concept from sampling theory also operates at a conceptual level where it is equally applicable to qualitative, including participatory work, and brings with it some claims to objectivity of selection, lack of systematic bias, and generalisability. The ranked set sampling approach is illustrated here in a simple case, where there is a single key measure or characterisation used to determine that the sample chosen is reasonable.

One frequent problem where ranked set sampling can help is the following. If there is no baseline study or existing sample frame available, we are denied the option that all the site selections could be made from a reasonable if not comprehensive list of communities. If there is no such list, how might we proceed, using more localised knowledge to help choose a few communities?

Example

A participatory problem diagnosis study is to be carried out in **four** food-insecure village communities in one of the eight Agricultural Development Divisions (ADDs) of Malawi. How do we find these four communities? Four* Extension Planning Areas (EPAs; sub-units of ADDs) are selected at random from those which have featured in the Famine Early Warning System (FEWS) as having food-insecure communities. Within these, a set of 'qualifying' criteria are set up which exclude unusual or untypical communities e.g. trading centres adjacent to metalled roads. Four* village communities per EPA are selected and it is verified that they 'qualify'. Knowledgeable extension staff from each EPA are asked to think about the last five years and to rank the set of four villages, say in terms of the proportion of their population who suffered three or more months of hunger in the year with the worst rains out of the last five.

* The number of communities and the number of EPAs should be the same - as the method description indicates - but of course four is just an example.

The 1, 2, 3, 4 rankings from the four EPAs are brought together. Taking the sets of ranks in an arbitrary order the community ranked 1 in the first EPA is selected, that ranked 2 in the next EPA is selected, that ranked 3 is taken from the EPA that happens to be third in the review, and that ranked 4 from the fourth.

This set of four selected villages now has one per EPA, but also some claim to span the range of levels of food insecurity in the target area, and not to represent unconscious selection biases of the researchers, insofar as it has some elements of objectivity in its selection. The four villages selected are a set chosen in a 'random' way to be representative of a larger sample of 16. This sampling process has in no way affected the research methodology decisions which can now be made by the qualitative researchers working in each of the four villages.

Of course the status in the entire population of the four villages ranked 1, say, will not be identical, and *a fortiori*, the differences between those ranked 1 and those ranked 4 will not be the same. This does not matter to the argument that we have an "objective" subset of a first sample of 16, and an enhanced claim to representativeness. If the four constituent rankings are done at random, the resulting sample is neither better nor worse than a random sample of size four, while if it is better than random the resulting ranked set is better than that. Once again the argument described here can be applied at more than one of the levels in a hierarchical sample. There is no bar to choosing four communities in this way then having n households within each one selected on a similar basis.

SAMPLE STRATIFICATION

Set in the context of communities, households and individuals, the above discusses the sampling issues that arise even if the units of study are all treated as interchangeable e.g. one community is the same as any other when sampling at primary unit level, within villages households are treated as being undifferentiated, and so on. Of course this is usually not the case, and as well as worrying about hierarchies, we need to think about male vs. female headed households, occupation or livelihood characterisations, caste, religion and the like.

Effective stratification

The statistical concept of stratification is widely cited, but not always relevant. Its essential meaning is not technical, and can be expressed clearly by considering a wildly extreme case:

suppose a population comprises subsets of individuals where every member is identical within each subset – in terms of the response we observe – even though the subsets differ from each other. We then need only a very small sample (of one) from each subset to typify it. In combination with information about how big the subsets are we can typify the whole.

In reality stratification can be very effective if the members who form a subgroup are a *relatively* homogeneous subset of the population i.e. have a greater degree of similarity to one another in response terms than would a completely random subset. The education level of head of household, the land tenure status, or other such factor used for stratification, brings together subsets of people who have something in common. Relatively small numbers of people can be supposed to typify each group, so the method is to a degree economical in fieldwork terms, though we probably need to search out all strata. Also it is common that a report of a study will produce some results at stratum level, so it is sensible to control how many representatives of each stratum are sampled, so the information base is fit for this purpose, as well as to represent the whole population.

Ineffective stratification

Populations are often divided into subgroups for administrative reasons, and results may be needed for separate subdivisions e.g. provinces. Unless the administrative grouping happens to coincide with categories of participants who are homogeneous in response, it is not an effective stratification in the above sense. As an over-simplified example, if every village contains farmers, traders and artisans in vaguely similar proportions, villages will be of little relevance as a stratification factor if the main differences in livelihood situation are between farmers, traders and artisans.

The above suggests that the subsets by occupation correspond to clearly distinguished, identifiable groups, internally similar to each other but very different from group to group. In this clear situation, stratification - by occupation - is an obvious sampling tactic. In many cases, however, the groups are by no means so distinct, and the subdivisions may be as arbitrary as the colonial borders of some African states. Usually this makes for ineffectual and delusory stratification.

Pre- and post-stratification

Where stratification is meaningful, it is sensible to pre-stratify where the groupings can be detected before study work commences. In some cases the information for stratifying only becomes apparent during the study, and the cases are sorted into strata after the event – post-stratification. That does not allow the same control over the number of representatives included from each stratum, so it is usually a bit weaker, except of course where the fieldwork is necessary to decide on the stratifier to be used.

Participatory stratification

It is sometimes suggested that useful subdivisions of community members within communities can be achieved by getting them to divide into their own groups using their own criteria. This provides useful functional subdivisions for participatory work at local level. If results are to be integrated across communities, it is important that the subgroups in different villages correspond to one another from village to village.

Thus a more formal stratification may require (i) a preliminary phase where stratification criteria are evolved with farmer participation, (ii) a reconciliation process between villages, and then (iii) the use of compromise "one size fits all" stratification procedures in the stratified study. If so the set of strata should probably be the set of all subsets needed anywhere, including strata that may be null in many cases, e.g. fisher folk, who may only be found in coastal villages.

Quantile subdivision

Stratification is not natural where there is a continuous range rather than an effective classificatory factor. If there is just one clear-cut observable piece of information which is selected as the best basis to be used, a pseudo-stratification can be imposed. For example a wealth ranking exercise may put households into a clear ordering, and this can be divided into quantiles, e.g. the bottom, middle and top thirds, or four quartiles, or five quintiles. This permits comparisons between groups derived from the same ranking e.g. the top and bottom thirds of the same village. Since the rankings are relative, they may be rather difficult to use across a set of widely differing communities, some of which are overall more prosperous than others.

Sub-sampling

The last paragraph hints at one way of choosing sub-samples for later phase, more detailed work, the "table legs" of the metaphor above. One result of the broad, shallow "table top" study - maybe a baseline study - could be a ranking or ordering of primary study units such as communities, and it would then be plausible to select a purposive sample to represent quantiles along the range of variation found.

Stratification for Comparing Groups

If as above we were intent on comparison, villages might be classified as Near/Remote from a metalled road, their land as mostly Flat/Steeply Sloping, their access to irrigation water as Good/Poor - three stratification factors each at two crudely defined levels – arguably more. The 8 possible combination characterisations such as [Near, Flat, Good] suggest we might have 8 sub-samples if possible. If each factor gave 4 levels, we would have $4 \times 4 \times 4 = 64$. This illustrates a common difficulty that possible strata are often all too numerous, and care is needed to select stratifiers of critical importance: if too many are used the number of sub-studies to be conducted and reported can get out of hand.

TARGETED SAMPLING

The processes described above are mainly concerned with ensuring that the sample selected can be justified on the basis of being representative. In some cases the aim is to target, exclusively or mainly, special segments of the general population e.g. members of a geographically dispersed socio-economic or livelihood subgroup. The problem is that there is not a sampling frame for the target group and we are never going to enumerate them all, so methods are based on *finding* target population members. There are several approaches to doing this. Mostly, the theoretical basis is less than perfect.

General population screening

If the target population is a reasonably big fraction of the overall population, and if it is not contentious or difficult to ascertain membership, it may be possible to run a relatively quick screening check that respondents qualify as target population members e.g. "Are there any children under 16 living in the household now?" As well as finding a sample from the target population, this 'hit rate' of this method will provide an estimate of the proportion which the target population comprises of the general population, so long as careful records are kept of the numbers screened. If the target population is a small proportion of the whole, this method is likely to be uneconomical.

Snowball sampling

The least formal method of those we discuss is "snowball" sampling. The basis of this is that certain hard-to-reach subgroups of the population will be aware of others who belong to their own subgroup. An initial contact may then introduce the researcher to a network of further informants. The method is asserted to be suitable in tracking down drug addicts, active political dissidents and the like. The procedure used is serendipitous, and it is seldom possible to organise replicate sampling sweeps. Thus the results are usually somewhat anecdotal and convey little sense of how completely the subgroup was covered or how large it really is.

Adaptive Sampling

This relatively new method allows the sampling intensity to be increased when one happens upon a relatively high local concentration of the target group during a geographical sweep such as a transect sample. It provides some estimation procedures which take account of the differing levels of sampling effort invested, and is efficient in targeting the effort. Until now this method has been developed primarily for estimating the abundance of sessile species and it is not yet in form suitable for general use with human populations. It does not carry any suggestion of networking through a succession of connected informants and is not a straightforward route to formalising snowball sampling.

Protocol-derived Replicated Sampling

In conclusion we offer a possible solution to the targeted sampling problem. The combination of ideas, and the suggestion to use it in the development setting make this solution novel in the sense of being untried. It clearly needs further development through practical application. The notion of replicated sampling is highly adaptable as a basis of valid statistical inference about a wider population⁵.

⁵ The original idea concerned a standard quantitative survey, probably with a complication such as multi-stage structure. If this could be organised as a set of replicates - miniature surveys, each with identical structure - then an estimate of some key measure could be derived from each one and that set of estimates treated just as basic statistics treats a simple random sample of data. The replicate-to-replicate standard error would incorporate the whole set of complexities within the stages of each miniature survey and we would get an easy measure of precision of the final answer.

We need to combine that idea with two other notions introduced here before using that of replication. The first is the idea of developing a prescriptive sampling protocol to be used in the field as a means of systematic targeting, say of particular households.

The protocol prescribes in detailed terms how to reach qualifying households in practice. As an example, suppose our target comprises "*vulnerable, female-headed rural households*" in a particular region. This involves sorting out all necessary procedural details. One element thereof might concern interviewing key informants at primary unit level, e.g. NGO regional officers - maybe presenting them with a list of twelve areas within the region and getting them to agree on two areas where they are sure there is a high number of target households. There are numerous procedural steps at several hierarchical levels. In the preceding example, the use of key informants is just an example; it is not an intrinsic part of every such protocol.

Samples are often derived in some such manner: they get at qualifying respondents cost-effectively, but the method usually carries overtones of subjectivity, and of inexplicit individual preference on the part of the selector. The protocol is supposed to address these difficulties. Naturally its development is a substantial process involving consultation, some triangulation, and pilot-testing of its practicability. It is thus a specially developed field guide which fits regional circumstances and study objectives, incorporating e.g. anthropological findings, local knowledge, and safeguards against fraud and other dangers. The protocol is a fully-defined set of procedures such that any one of a class of competent, trained fieldworkers could deliver a targeted sample with essentially interchangeable characteristics.

The second added notion is that if the protocol development involves appropriate consultation, brainstorming and consensus building, then the protocol can be used to define the *de facto* target population being reached. Developers of the protocol can effectively sign up to (i) accepting a term such as "*vulnerable, female-headed rural households*" as the title of the population who are likely to be sampled during repeated, conscientious application of the protocol, and to (ii) accepting that the population sampled is a valid object of study, & a valid target for the development innovation(s) under consideration in the locale for which the protocol is valid.

Repeated application of the procedure would produce equivalent "replicate" samples. These carry some "statistical" properties, provided that (i) the sampling is regulated as described above, and (ii) the information collection exercise within any given replicate is standardised. When the procedure is replicated, it is necessary that at least a common core of results should be collected in the same form, and recorded using the same conventions, in each replicate and it is for these results that we can make statistical claims.

For example, suppose we record the proportion (x) of respondents within a replicate sample who felt their households were excluded from benefits generated by a Farmers' Research Committee in their community. The set of x -values from a set of replicate samples from different places now have the properties of a statistical sample from the protocol-defined population. Even though the protocol itself encompassed various possibly complicated selection processes, we can, for example, produce a simple confidence interval for the general proportion who felt excluded.

The important general principle which follows from this is that if we can summarise more complicated conclusions (qualitative or quantitative) instead of a single number x , from each replicate, then we can treat the set as representing, or generalising to, the protocol-defined population. There are interesting ways forward, but the practical development and uptake of such a notion poses "adaptive research" challenges if the concept is put to use in the more complex settings of qualitative work in developing countries.

ACKNOWLEDGMENT

Substantial parts of the work reported in this presentation were carried out by the author as part of DFID NRSP (SEM) Project R7033, under a grant held jointly by the Natural Resources Institute, University of Greenwich, and the Statistical Services Centre, University of Reading.

Word count of entire text including title = 6411.