

# 1 Surveys

Extremely important for collection of information

Have already considered generally -focus in more detail on some issues -taken mainly from Angus Deaton's book

## 1.0.1 Survey design:

Obvious way to start is with a list of all households and take a random sample, but things are more complex in practise

Sampling frame: is list of households from which take sample

Accuracy of sample statistics increases less than proportionately to size ( in prop to its square root actually

Sampling fractions smaller in larger populations

Frames: often a census:

- Most common is 2 stage selection
  1. select from a cluster of households -e.g. villages
  2. select households
- Can have an equal probab of being chosen in frame as long as self weighting design.
- Outdated inaccurate frames are a source of error
- Even if accurate can have incomplete coverage -armed forces /rural etc.
- wont prevent us making inferences, but have to take care.

Two stage design: means not randomly distributed over space -geographically grouped

- Cost effective
- Allows repeat visits
- Makes worthwhile to collect more agg infor that may be imp for households eg at village level
- May have partic interest in partic target group -e.g. AIDS patents
- May need representation over whole range of types -need enough obs for each group

- Random sample could choose all people with same characteristics
- May also be statistical reasons -using prior knowledge can improve efficiency of statistical inference

When stratify: convert one population into many

- Explicit stratification: e.g. region
- Implicit stratification: order on the list e.g. % African in SA Living Standards Survey

While stratification will usually enhance the precision of sampling estimates, clustering will normally reduce it

- Clusters usually have more similar households so don't get as much info/variation as would if not clustered
- Consider

$$\text{Design effect} = \frac{\text{variance of an estimate}}{\text{variance under simple random sample}}$$

stratification will reduce this to below ones, but variables in stratified and clustered sample normally greater than one.

- In practice cost/convenience concerns predominate

Unequal selection probabilities: common for the probabs of inclusion to differ because of stratification and clustering

1. costs more to sample some households
2. differential probabs of inclusion can enhance precision
3. some types of households may refuse to participate

Cost variation: common between rural and urban

prior info may allow focus -best precision for cost

use probab proportional to size 'pps sampling' e.g. if looking at mean income mor precise if overrepresent households that contribute more to the mean -higher income.

Now though don't have income data at start may have prior info on variables that are correlated e.g. household size, land holdings

When selection probabilities differ across households , each household stands for a different number of households in population

need weight sample data to make estimates for popn

weights undo sample selection

weight according to reciprocal of sampling probabs -inflation factor

Difference in weights can come from differences in selection  
by design /construction = 1/sampling probab and are controlled and  
ex ante  
by accident -e.g. non-response- are not controlled often added to survey  
ex post -determined by judgement/modelling and so care needed in using them

## 1.1 Sample design: Theory and practice can differ

Are statistical arguments for stratification and differential sampling probabilities, but more likely to be more practical concerns esp in developing countries.

Can design optimal survey with single purpose, say to measure mean income, but unusual not to be for more than one purpose and each piece of info might imply a different survey structure

trade off coverage and precision  
selecting for efficiency can compromise usefulness: e.g. use of on board mode choice surveys

can use 'choice based sampling' and deal with sample selection statistically

complicated designs can be problematic: sampling errors are likely to be smaller than non sampling errors and may more than offset theoretical benefits: best to keep survey as simple as possible.

## 1.2 Types of survey data

1. Standard cross section: designed as a snapshot.  
May take time to collect, so each household interviewed at different time, or may be designed to cover areas at different times over say a year.
3. Supplemented cross section: merge administrative and survey data,  
may revisit some households,  
may create 'panel' by asking households to give info for period of time  
-e.g. diary  
problems of accuracy of recall
2. panel data: longitudinal survey -tracks household over time and gives multiple observations on them  
allows study of dynamics  
can be rotating: losing some observations and replacing

Problems:  
attrition -households move away, refuse get lost, families split by death  
etc.

trade off representative sample and tracking individuals

Benefits

can enhance precision of estimates especially changes in them as panel will look at same households, whereas cross section wouldn't

Has benefits when there is possibility of measurement (non sampling) error. Consider when household reports  $x_{it}^*$  rather than the true value  $x_{it}$

$$\begin{aligned}x_{it}^* &= x_{it} + \epsilon_{it} \\ E(\epsilon_{it}) &= 0 \\ E(\epsilon_{it}^2) &= \omega^2\end{aligned}$$

if reporting error uncorrelated with truth

$$\begin{aligned}\Delta x_{it}^* &= \Delta x_{it} + \Delta \epsilon_{it} \\ \text{var}(\Delta x_{it}^*) &= \text{var}(\Delta x_{it}) + 2\omega^2(1 - \rho)\end{aligned}$$

where  $\rho$  is the correlation between the errors over the two periods.

So 1. presence of measurement error enhances advantage of panel data over indep cross sections for measuring changes in means

2. Signal to noise ratio differs between levels and changes: difference depends on  $\omega^2$  and  $\rho$

### 1.3 Descriptive Statistics

When analysing surveys need to be sure that the statistics we use describe the particular population rather than the particular sample that is available for analysis

- need to know how sample designed
- need to use any differential weights (inflation factors)
- need to allow for stratification and clustering

Two approaches

1. population is finite and sample of households random selected. Survey data is random as replication of surveys would generate diff samples

-typically focus of description

2. Superpopulation: Actual population is regarded as a sample from all possible such populations the infinite superpopn

-focus on statistical laws and economic processes that generated the observed data in the superpopulation, so mean is less of interest for itself

-typically focus of modelling

Note recent trends in econometrics have emphasised description -so gap between these approaches is narrowing

### 1.3.1 Sample Mean and Variance

For finite simple random sample the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

is a random variable differing from one sample to another, or one popn to another in superpopulation model. Suppose want to know expectation, calculate mean for all possible samples and probabilities of them happening. Or take a shortcut:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^N a_i x_i$$

where  $a_i$  is a random variable taking the value 1 if in the sample zero otherwise and  $x$  is now fixed in the finite population and as its a simple random sample the probability of inclusion is  $n/N$  and the expectation of  $a_i$  is  $n/N$  so:

$$E(\bar{x}) = \frac{1}{n} \sum_{i=1}^N \frac{n}{N} x_i = \frac{1}{N} \sum_{i=1}^N x_i = \bar{X}$$

the population average

For sample variance.

$$var(\bar{x}) = \frac{1}{n^2} \left[ \sum_{i=1}^N x_i^2 var(a_i) + 2 \sum_{i=1}^N \sum_{j < i}^N x_i x_j cov(a_i, a_j) \right]$$

because  $a_i$  is binomial its variance is  $(n/N)(1 - n/N)$  and the probability that  $a_i, a_j$  are in the sample (equal to one) is  $(n/N).(n - 1)/(N - 1)$  as the sample is drawn without replacement. So

$$cov(a_i, a_j) = E(a_i, a_j) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = -\frac{n}{N(N-1)} \left[1 - \frac{n}{N}\right]$$

Substituting:

$$var(\bar{x}) = \frac{1-f}{n} S^2$$

where

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2$$

which can be thought of as the population variance and  $1 - f$  is the "finite population correction"  $(N - n)/N$ , which except in the unusual case where the sample is a large fraction of the population is close enough to one to be ignored. generally assume this is the case

Superpopulation approach makes more assumptions but can be more straightforward. For example postulate that  $x$  is distributed with mean  $\mu$  which is same for a all populations

$$E(\bar{x}) = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

The finite sample approach is more general as it doesn't need to assume homogeneity, but is more limited as its concerned with one population only.

For superpopulation each  $x_i$  is assumed IID mean  $\mu$  and variance  $\sigma^2$  so the sample variance is

$$var(\bar{x}) = E(\bar{x} - \mu)^2 = \frac{1}{n} \sigma^2$$

Since both  $\sigma^2$  and  $S^2$  are estimated from the sample variance

$$\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Can estimate variance as

$$var(\bar{x}) = \frac{\hat{S}^2}{n} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n(n-1)}$$

If we ignore  $1 - f$  there is no operational difference between this and the finite sample approach.

## 1.4 Using Weights

In most surveys households will have different probabilities of being selected, either deliberately (sample design) or by accident, because of response problems.

- This will mean that the simple mean of a variable will be a biased estimate of the population mean
- It is possible get an unbiased estimator using weights

Consider  $N$  households assigned sampling probabilities  $\pi_i$  and assume selection with replacement

-though this doesnt happen in practice, as long as the sample is a small fraction of the population it doesnt really matter and can ignore finite sample corrections

- Low  $\pi_i$  means low ex ante probability of selection into sample so weights need to be inversely proportional to  $\pi_i$

$$w_i = \frac{1}{n\pi_i}$$

- Simple random sample probability of selection is  $1/N$  and so weights same for all observations and is  $N/n$ . This is the "inflation factor" that expands sample to popn
- When probabilities differ  $n\pi_i$  is the expected number of times household  $i$  shows up
- when sample small relative to population,  $n\pi_i$  is the probability of  $i$  being in the sample.
- So  $w_i$  is the approximate number of population households represented by the sample household  $i$  = specific household inflation factor

Can show:

1. Sum of weights  $\hat{N} = \sum_{i=1}^n w_i$  is an unbiased estimator of population size
2. That an unbiased estimator of the population total value of a variable  $X$  is  $\hat{X} = \sum_{i=1}^n w_i x_i$
3. An unbiased estimate of the sampling variance is

$$var(\hat{X}) = \frac{n}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2$$

where  $z_i = w_i x_i$

Sampling variance shows why different probabilities can enhance efficiency and tell us what the optimal probabilities are

- can show sampling variance is minimised when the  $\pi$ s are proportional to the  $x$ s
- this is sampling with "probability proportional to size" (PPS)
- in practice dont know  $x$  so use approximate PPS where set probabilities proportional to some other variable that is thought to be correlated with  $x$  and is known prior to sampling

Probability weighted mean of  $x$ ,  $\bar{X}$  is

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} = \sum_{i=1}^n v_i x_i$$

with

$$v_i = \frac{w_i}{\sum_{k=1}^n w_k}$$

because  $\bar{x}_w$  is a ratio of 2 random variables it is not unbiased, but it will converge on the population mean as  $n$  tends to infinity

- Can estimate the sampling variance and use this to find the probabilities that maximise the precision of the probability weighted mean
- Can show that optimal selection probabilities should be proportional to the absolute value of the deviation from the mean  $|x_i - \bar{X}|$ .
- This is why it is information on exceptional cases that adds most to the precision of the estimated mean

Whenever there is an association between sampling probabilities and quantities being measured unweighted estimates are biased.

So important to keep an eye out for the weights when using survey data.

Note that the probability weighted mean is not the only possible estimate. if we know  $N$  could use  $X/N$  where  $X$  is population total.

But:

- Population size often not known -particularly in developing countries- and is estimated from the survey itself eg randomly choose villages, survey all households, estimate number of households in population
- Some data unusable missing, implausible, transcription error. little option but to average 'good' observations and renormalise weights
- Often interested not in household mean but person mean. Weight not by number households data represents, but by number of people
- Often want to know mean for subgroups and to estimate in a way that is representative of subpopulations, then  $\bar{x}_w$  is relevant estimator unless know size of each subpopulation

Weights can be used to estimate other population statistics, such as population variance.

Can consider in similar way:

Stratification:

- break up one survey into a multitude of independent ones
- strata fixed households vary from sample to sample

- without stratification fraction sample in each strata depends on chance
- can reduce sampling variability whenever means differ across strata
- in practice estimates for stratified samples use simple adaptations of formulas for weighted estimates
- but population shares need not be same as sample shares: stratification is not about weighting

#### Two stage sampling and clusters

- With strata most household surveys collect data in two stages: sample clusters or primary sampling units (PSUs) and then select household from within each cluster
- Cluster sampling raises different statistical issues to stratification -if replicating a survey strata would be constant, clusters would change from sample to sample
- formulas for weighted and unweighted means are not affected by two stage design, but sampling variability is. Households in cluster similar so reduce variation and so precision.
- Can be serious mistake to treat 2 stage as if it were a simple random sample. Standard formulae seriously overstate precision of the estimates.

i