ACC DAR Adapting to Climate Change in Coastal Dar es Salaam

Technical Workshop on Data Analysis

SAMPLING PROCEDURES

Dar es Salaam, 27 September 2011

Centro Interuniversitario di Ricerca Per lo Sviluppo sostenibile - CIRPS

Sapienza Università di Roma

ARDHI UNIVERSITY TANZANIA UTAFITI ELIMU USHAURI

Pietro DEMURTAS

IRPPS – National Research Council

p.demurtas@irpps.cnr.it

Simple random sampling…

… is the simplest probability sampling procedure. In this method of selecting a sample of the population units, every sample of a fixed size is given an equal chance to be selected. Every population unit is given an equal chance of appearing in the sample. Similarly, every pair of units has en equal chance of appearing in the sample. In general, every collection of units of a fixed size has an equal chance of being selected.

- *i.e.: you want select a random sample of size n=2 from N =6 students (population)*

  One can select the sample by writing the names, numbers, or labels of the six students on pieces of paper or cards, thoroughly mixing them, selecting one randomly, setting it aside, and then selecting the second one randomly from the remaining five. This is one of procedures for selecting a sample randomly **without replacement.**

- Random samples from a population, especially large, can also be selected from random number tables or computer software packages.

- for selecting a simple random sample from a population, **you must have a list of the units of populations.**

Systematic sample

- In the common practice for selecting a systematic sample, the population is divided into *k* groups of size *n =N/K* in each.
- ((N is the size of population that we study))

  One unit is chosen randomly from the first *k* units and every *kth* unit following it is included in the sample.

Arrangement of 40 population units into 4 systematic samples

SAMPLES			
1°	2°	3°	4°
1	2	3	4
5	6	7	8
9	10	11	12
…	…	…	…
37	38	39	40

If the random number drawn from the first four numbers is 2, for example, the sample would consist of the units (2, 6, 10, ...). Selecting a systematic sample as above is equivalent to selecting randomly one of the k groups

- It is very convenient to draw systematic samples from telephone and city directories, automobile registries, and electoral rolls. A sample of households from a residential neighborhood can be easily obtained by this method.
- It also becomes convenient to draw a sample systematically when the population frame in not available. For example, selection of every 10th farm on the spot from the approximately 500 farms in a village will provide a sample of 50 farms. In industrial quality control, a sample of item produced, for example, every 30 minutes or every tenth person passing a certain location

The problem of NONRESPONSE

- In almost every survey, some of persons, households, and other types of unit selected into the sample are no contacted. Person away from home on business or vacation, wrong addresses and telephone numbers, inability of the interviewers to reach households in remote places, and similar reasons contribute to the non contact.

i.e.

- Public polls are also frequently affected by the non contacts and non response. In the Truman-Dewey presidential contest, although Truman emerged as the winner from the final count of the ballots, *The Literary Digest*, a newspaper in Chicago, initially declared Dewey the winner. A badly conducted poll and large amount of non response were blamed for this “fiasco”

Probability and Nonprobability sampling

- Sampling procedures described before are also frequently known as **random samples**. Randomness, however, does not imply that the sample units are selected haphazardly. On the contrary, in this method of sampling, the probabilities for selecting the different sample from a population are specified, and the probabilities for the units of population to appear in the sample are known.
- Sampling error is a measure of the departure of all the possible estimates of a probability sampling procedure from the population quantity being estimated. A very important feature of probability sampling is that in addition to providing an estimate of the unknown population quantity, it enables the assessment of the sampling error of the estimate, the **standard error**

- **Non probability sampling:** to estimate the total sales or average prices of computer in a metropolitan area, a handful of selected computer stores may thought to be adequate. To estimate the agricultural production in a village, a sample consisting of a certain number of small, medium and large farms may thought to be **representative**. In these illustrations, prior information on the population units is frequently utilized for selecting sample.
- QUOTA sampling: in this method, the survey is continued until a predetermined number of the people, households, hospitals, corpoations, and similar populations of units with specified charactheristics are contacted and interviewd. In a political survey, for example, the interviewers may be asked to obtein 200 responses from each of the male and female groups between the ages 25 and 65. In several surveys, people are contacted until specified percentages of responses are obtained to the different items in the questionnaires

Survey design

Generally speaking, all the surveys must have a design in which all the details are considered to reach a specific target. In our case the target is to improve the amount of information about population related to environmental habits. This objective involves the need **to enlarge the most is possible the space in whic the data are collected**. But, at the same time, **the data have to be representatives of the territory in whic they are detected**.

So, it's very important to avoid an excess of concentration of interviews in a territory and absence of them in other places. The best situation is to have a number of interviews proportional to the population living each place. If this is not easy to calculate, the number of intervews have to be calculated following a common sense rule.

During the data analisys, the geographical data have to be available, to be able to read the result taking into account the spatial dimention. In that way, even if the concentration of intervews is not really proportional to the resident population, a spatial statistical analysis (i.e. an analysis in whic the geographial variable is considered) can enable us to answer to the most important aims of the research.

Finally, if the respondents are selected in a random way inside a particular geographical location, special statistical technics can be used to **extend the result** to non-respondent people. But, to apply that, no reasons to suspect that the selection of the respondents is not random have to exist, including refuse or substitution, for any reason. Otherwise just a simple statistical analysis is permitted, and no probabilistic methods have to be applied.

Bias of sample

- The bias (distorsion) of sample depends from as you have retrieved the people for the interview
- For infinite populations, an estimator is defined to be consistent if it approaches the population quantity being estimated as the sample size becomes large ((If the sample is larger, we have a greater chance of having respondents with different profiles and then to respect differences in the reference population!!!))

To estimate the distortion among the sample and the population we must have information about the criteria that you used to administer the questionnaires:

- As you have selected the families?
- Did you have select randomly the first person to be interviewed? As you chosen the following persons to be interviewed?
- Have all the people that you have contacted answered to the questionnaire?

Lime Survey

Why to use a CAWI software to collect survey data?

- The CAWI (Computer Assisted Web Interviewing) It's just a questionnaire on a web page.
- We used Limesurvey software, an open source product (so, free of charge for all users!), to collect the data from the survey. It's an open source survey application. One of the most common free software to install on a web server. The result of this installation is a web site in wich an administrator and other authorized people can manage questionnaires (manage: create, set up, give authorizations to see results or questions, extract data, etc.).

Why do that and don't use only directly a statistical software?

The fact is that in this way we have a list of advantages:

1) we can use more people even if they are differently located for inserting the data;

In the present project the questionnaire is administered face to face using a paper version that is inserted in the database only in a different time. Using an internet based software presents the advantage to digit the data in different places without the mismatch problems that can occur using a stand alone database. In every time the database will be complete and available for statistical analysis of the collected data.

2) the data bank is always available everywhere at the only condition to have an internet connection

All the people, at the only condition to be autorized to access to the Limesurvey system, can participate to the check and analysis of the data also during the data collection. In that way more than the only people directly involved in the administration of the survey have the possibility to support the research.

3) the possibility follow the data collection using simple statistics and having the possibility to monitorate and correct situations and problems during the survey field

The Limesurvey system offers the possibility to generate in few click simple statistics about the questions. This fact is really useful as a check of the trend of interviews and it allows to correct some imperfection or problem on the data before the end of the survey field.

4) the last advantage, but not the list, is related to the possibility to assist the data digiting with a simple software check that limits the possibility to insert incorrect data in the database

- To digit a questionnaire can generate trivial errors that can be avoid simply using software barriers that prevent not required information (i.e. text instead number, number out of scale, etc). Some of these features are contained in the present questionnaire on Limesurvey software.

- This software also provides a system to export data in different formats: text, csv, SPSS, R.

The Limesurvey version installed on the IRPPS-CNR web site is reachable at the address http://www.irpps.cnr.it/limesurvey/admin/admin.php.

- First of all you need to provide a **username and password**. The system administrator will give you an account to enter and manage the software as your competence.
- Then you arrive in the home page of the software in which you have to select your survey in the little window at the top, right side. Then you can choose both **insert a new questionnaire** (selecting the appropriate link) or **view the inserted questionnaire** at the link http://www.irpps.cnr.it/limesurvey/admin/admin.php?action=browse&sid=16732.

In this page, just selecting the appropriate icon, you can **extract the dabase**. But, first of all, you have the possibility to **check the data** inserted until now selecting the related icon (that's similar to a paper). You enter in an environment in wich is possible to see in detail all the inserted data, but also correct, eliminate, or **manage the data.** Finally, the icon like a cake, enable you to see data **statistics**, both for all the data or just a selection of them.

- Selecting the icon to extract the database, you arrive in a web page in wich you have to choose the kind of responses, the SPSS version, the syntax exportation and finally the data exportation.

- After selecting the kind of responses and the SPSS version, if you select the syntax exportation button, the software produce an **SPSS syntax file** that create a complete version of SPSS data base. That feature is very usefull as you can personalize the data set, changing, if is the case, the label or variable caracteristics before collecting the data under SPSS software.

- Selecting the 'step 2' button, the software generates a .dat file useful to produce an **SPSS data set** by the previous SPSS syntax file.

- The SPSS syntax file is very usefull because preserves tracks of your SPSS work and enables you to repeat in the same way all the operartion you did in advance. It also allows to prepare a great part of work in advance or to test the analysis on partial database and repeat it on the whole database.