Cross-sectional Surveys

(Nutritional Anthropometry)

Survey design, sampling, data management, and data analysis groupwork sessions

Mark Myatt, Brixton Health / SCF(UK), January 2004

Section One - Survey Design

Background to the survey design exercise

Your task in this exercise is to develop a plan for a cross-sectional (descriptive) survey of nutritional status in children living in a single administrative district of a developing country.

This manual is intended to guide you through the exercise and contains background information required to complete the exercise, some practical advice to help you complete the exercise, and a series of practical tasks for you to complete.

The **`SampleXS`** sample size calculation program that is used in this exercise is to be found on the supplied floppy disk. Instructions for starting **`SampleXS`** under different operating systems are shown in the table below:

System	Instructions
Windows 95, 98, NT 4.xx, 2000, XP	Select the **`Run`**... item on the Start menu. In the **`Run`**... dialog box type: **`a:\samplexs`** and click the **`OK`** button.

`SampleXS` is a 32-bit Windows program and may also be run under 32-bit Windows emulation on systems running UNIX / LINUX.

The dataset (**`NUTSURV.REC`**) used in the data analysis practical is also on the supplied floppy disk.

The survey setting

Country 'X' (an imaginary country) is located in central Asia covering an area of 237,500 km^2. Approximately 65% of the country is mountainous with an average elevation of about 1300 m. The country is divided 39 administrative districts and there are ten major cities and towns.

The climate is semi-arid steppe but varies sharply between highlands and lowlands. Sub-polar in the mountainous areas with dry and cold winters (falling to -26ºC) and desert in the north with less than 70 mm of rainfall annually and summer temperatures exceeding 35ºC. Monsoons influence the climate of the mountains in the south-east with an annual rainfall in excess of 1100 mm. Elsewhere, summers are hot and dry with temperatures January -2.8ºC, July 24.5ºC. Annual rainfall 388 mm.

The fiscal year 1998 budget was for a revenue of US$437.2 million. Main items of expenditure were defence (29.6%), housing and welfare (8.7%), and education (3.4%). The balance of payments (current account fiscal year 1995) was a deficit of US$114.7 million. Inflation has recently been high but has now stabilised at around 6% per annum. Estimated GDP is US$4,167 million, per capita GDP is estimated at US$260. The total number of economically active persons in the country are estimated at around 5 million persons:

Economic sector	% of workforce
Industry	21%
Agriculture	68%
Other	11%

Total population (1988) was estimated at 16,120,000 *excluding* nomads of whom there were an estimated 2,674,000 in 1982. Approximately 55% are Pashtun with the remainder Uzbek (15.5%), Tadzhik (9.7%), Turkmen (16.2%), and Chahar Aimak (3.6%). The nomadic population are predominantly Hajara.

The proposed survey will take place in Administrative District 31 (AD31) which is located in the north-west of the country. The population of AD31 is estimated at 97,875 persons excluding nomads. The age distribution of the non-nomadic population is:

Age	Population	%
0 - 14 years	40491	41.37%
15 - 19 years	13869	14.17%
20 - 24 years	7938	8.11%
25 - 29 years	7370	7.53%
30 - 39 years	13957	14.26%
40 - 49 years	7360	7.52%
50 - 59 years	4209	4.30%
Over 60 years	2681	2.74%

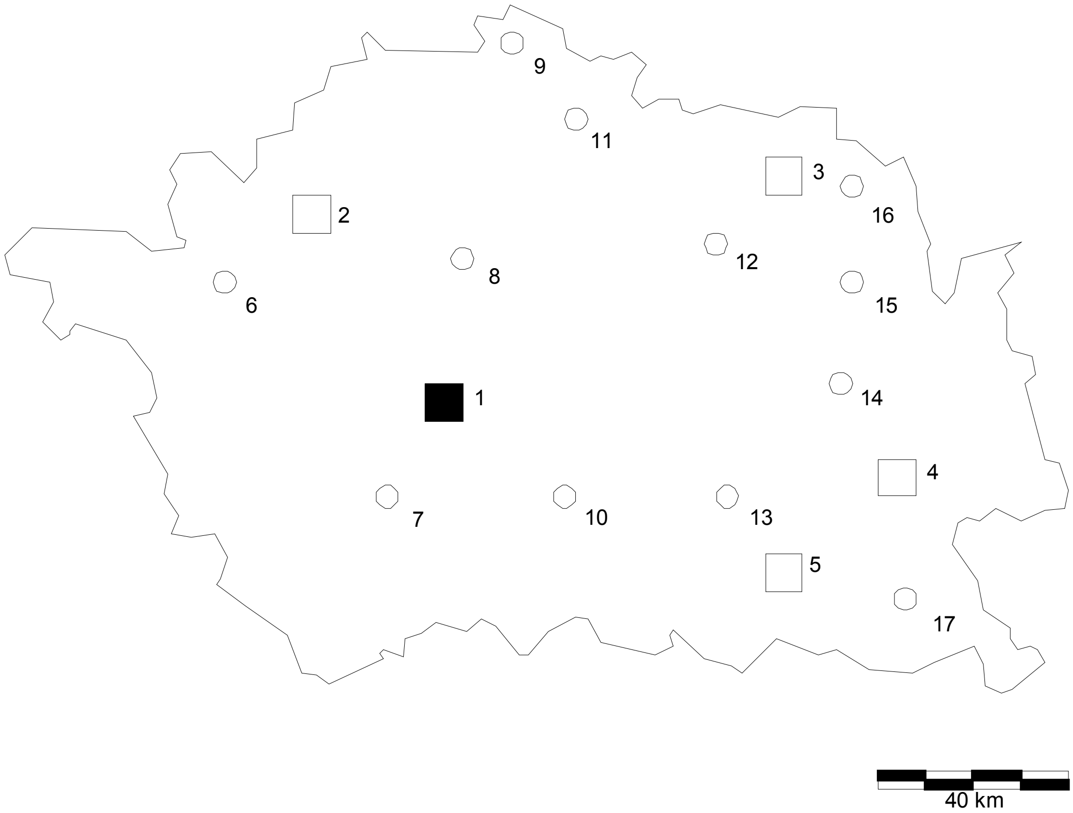
Population data is estimated from a census undertaken prior to the civil-war and may be unreliable.

The country has recently emerged from over a decade of civil war and existing preventative and curative health services are disorganised. The security situation is currently good.

The survey setting

There are 17 towns and villages in the district:

`Administrative District 31 showing towns and villages`



The populations of each town and village in the district are estimates as:

Centre	N	Centre	N
1	24515	10	370
2	18761	11	1678
3	14178	12	5432
4	11577	13	2091
5	12670	14	1133
6	1765	15	904
7	787	16	523
8	844	17	200
9	467		

Centres one and two are internally divided into twelve electoral districts of approximately equal population. Centres three, four, and five are internally divided into eight electoral districts of approximately equal population. The population estimates in these larger centres is likely to be upwardly biased due to internal displacement during the civil war. Some return of the population to the villages is expected.

Stating the aims of a survey

The first task in designing and planning a survey is to formalise the aims of the survey. These should be clearly and precisely defined and should be able to inform all other decisions about the design of the survey. Without well defined aims there is a risk that the survey will not be able to answer any useful questions about the survey's target population. There may be a tendency to collect far more data than is actually needed resulting in extra time and expense and poor data quality. It may be useful to define *primary* (or *essential*) aims and *secondary* (or *desirable*) aims for a survey as this helps in setting priorities and making decisions about the scope and size of a survey.

A set of aims for a nutrition survey might read:

1. To estimate, with reasonable precision, the prevalence of acute global, moderate, and severe wasting malnutrition in children aged between six (6) and fifty-nine (59) months living in Administrative District 31 of Country X.

2. To investigate possible *risk factors* and *risk markers* for acute global, moderate, and severe wasting malnutrition in children aged between six (6) and fifty-nine (59) months living in Administrative District 31 of Country X.

3. To estimate the size of programs (e.g. supplementary feeding programs, therapeutic feeding centre, community therapeutic care programs) required to address the problem of wasting malnutrition in children aged between six (6) and fifty-nine (59) months living in Administrative District 31 of Country X.

These aims have implications for the design of the survey:

The first aim contains the phrase 'with reasonable precision'. This indicates that the survey will have a design and a sample size that allows the results of the survey to be *generalised* from the survey sample to the target population.

All aims contain the phrase 'wasting malnutrition'. The minimum requirement for measuring wasting malnutrition is the measurement of mid-upper-arm-circumference (MUAC). MUAC measurements must be taken for the left arm while the arm is hanging down the side of the body and relaxed. An alternative approach is to collect weight and height data for comparison with a reference population. If height data is collected then children taller than 85 cm must be measured standing. Shorter children must be measured lying down (i.e. measure their crown-to-heel length). All children must be measured and weighed naked (or in underwear) and without shoes. Weight measurement is of little value if the respondent has kwashiorkor as oedema causes the weight to increase giving the impression of adequate nourishment. In surveys, the presence of *bilateral* pitting oedema (any grade) is often used indicate kwashiorkor. If you use the weight-for-height method of assessing nutritional status you will also need to collect data on the sex and age of the respondent as each sex and age-group has a different 'reference curve'. Whichever method you use to measure a respondent's nutritional status you will need to use a case-definition for malnutrition. The following case definitions are commonly used:

Level	**Definition (MUAC)**	**Definition (Wt/Ht)**
Global	MUAC of less than 125mm and / or bilateral pitting oedema.	< 2 weight-for-height z-scores below NCHS / WHO reference mean and / or bilateral pitting oedema.
Moderate	MUAC of between 110mm and 124mm (inclusive).	< 2 weight-for-height z-scores and >= 3 weight-for-height z-scores below NCHS / WHO reference mean.
Severe	MUAC of less than 110mm and / or bilateral pitting oedema.	< 3 weight-for-height z-scores below NCHS / WHO reference mean and / or bilateral pitting oedema.

The use of MUAC in surveys is considered by some agencies to be controversial. It is thought to overestimate malnutrition in very young children. Collection and analysis of MUAC data is usually, therefore, limited to children of one year or older.

Stating the aims of a survey (continued)

Weight for height data are considered by many to be the ‘gold-standard’ for assessing undernutrition. It may seem to be an unnecessary extra step to collect MUAC as well as weight and height data but doing so has some advantages:

MUAC data can be used as a quality check on weight for height data as we expect that the majority of children classified as undernourished by one measure will also be classified as malnourished by the other measure. Reviewing MUAC and weight for height data during the data collection phase of a survey can help identify problems with measurement and data collection.

Peripheral tissues, such as the tissues of the upper arm, are preferentially catabolised during infection. If analysis of MUAC and weight for height data yield broadly similar results then it is likely that the problem is one of undernutrition (either availability or uptake). If the analysis of MUAC data yields a considerably higher prevalence of undernutrition than analysis of weight for height data then there may also be a considerable infectious disease problem in addition to problems with food availability and uptake.

MUAC is a better predictor of short-term mortality than weight for height.

MUAC data is easy and quick to collect creates little extra work either in data collection, data-entry, or data-analysis. There is a revival of interest in using MUAC for height as a means of assessing undernutrition. Screening instruments are available (e.g. the QUAC stick) but software that allows the easy and rapid analysis of MUAC for height data in a survey setting is currently not available.

The stated aims have some additional implications for the design of the survey:

All aims describe the *target population*. It is essential that this population is defined at the outset of the planning process (e.g. ‘all children aged between six (6) and fifty-nine (59) months living in Administrative District 31 of Country X’). The specified age range was selected because this population is considered to be a *canary-in-the-coalmine* or *sentinel* population providing an early warning of community-level nutritional stress. Reliable age data is sometimes difficult to obtain so a proxy based on the height of the respondent is often used - children of between sixty-five (65) and one-hundred and ten (110) centimetres in height. This height range is an internationally accepted proxy for the selection of the required age range. This height range may cause your sample to include some older children if there is significant stunting (i.e. due to chronic undernutrition) in the population.

The second aim will probably require that each respondent be examined and / or a parent questioned. The data you collect to meet this aim will usually be based on the results of a *semi-quantitative / semi-qualitative* pre-survey. You may also want to collect data on variables closely associated with malnutrition (e.g. familial circumstances, night blindness / xerosis and common infections, particularly measles and gastro-intestinal infections / worm infestations).

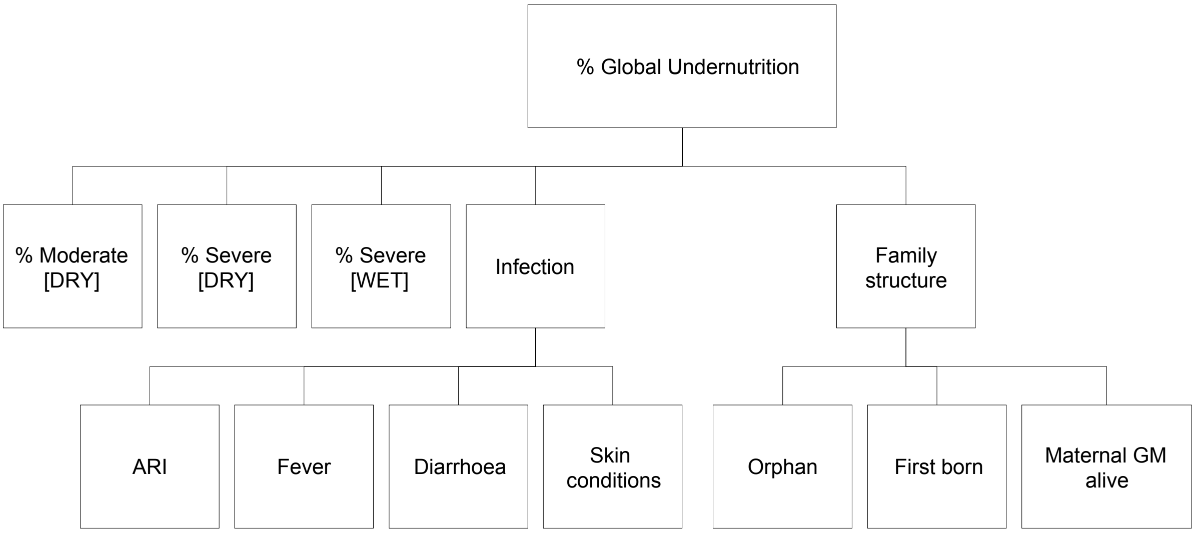
If one of your aims is to estimate program size then you will need to be able to apply program entry criteria to the collected data. Program entry criteria may be different from the case definitions used to estimate prevalence. For example, the weight for height z-score may be used to define wasting but the weight for height percentage of median (WHM) might be used for program entry.

Any survey of nutritional status will concentrate on anthropometric indices of nutritional status. You might consider, however, collecting data on specific nutritional deficiencies. If you do this you should concentrate on data that can be collected using readily ascertainable signs and symptoms. The prevalence of vitamin A deficiency, for example, can be estimated by asking about night blindness and direct examination of eyes for (e.g.) Bitot’s spots.

In general, the survey aims should inform the survey design, survey sample size, the survey dataset, the methods of measurement to be used, and the case-definitions used in the survey. Standard case definitions for most conditions are available.

Stating the aims of a survey (continued)

It may prove useful to use a framework to help organise your ideas about the primary and secondary aims of a survey. One framework that you might find useful is to organise your ideas in a hierarchy of aims using a *tree diagram* or *organisation diagram*:



In group discussion formulate the aims of a nutrition survey and complete the information requested on each of the following pages.

Stating the aims of a nutrition survey (continued)

The aims of the proposed nutrition survey are:

Stating the aims of a nutrition survey (continued)

The target population of the proposed nutrition survey is:

Stating the aims of a nutrition survey (continued)

The methods of measurement that will be used in the proposed nutrition survey are:

Defining a survey dataset

The aims of the survey will suggest the scope and nature of the data items that will need to be collected and a broad outline of the survey *dataset* (i.e. the data needed to meet the aims of the survey) should flow naturally from them. Much of the survey dataset is defined directly by the survey aims. It is, however, necessary to provide a detailed description of the required dataset.

You will always need to collect basic anthropometric data (i.e. age, sex, weight, height, MUAC, presence of oedema). It may prove useful to use an *epidemiological* framework to help you define a dataset. The two main epidemiological frameworks that you might find useful are the *Time-Place-Person* (TPP) and the *Host-Agent-Environment* (HAE) frameworks. You might also find frameworks that utilise existing epidemiological knowledge about the condition under survey useful when defining a dataset.

In a cross-sectional survey you are most interested in *describing* the pattern of disease within a population. Cross-sectional surveys are **not** particularly effective at answering questions about associations between *risk factors* and disease. Such questions are better handled by *cohort* and *case-control* studies and are best relegated to the secondary aims of a cross-sectional survey. Cross-sectional surveys provide an estimate of the *prevalence* of disease in a population. They cannot answer questions about disease *incidence*.

The *Time-Place-Person* framework

A common framework used for epidemiological surveys is *Time-Place-Person* (TPP). This is a convenient way of describing the occurrence and distribution of disease within populations in terms of time, place, and person. These characteristics provide clues that might help explain the causative process or agent of the disease and the varying susceptibility of subgroups within a population to disease when an exposure occurs:

> **TIME** : This usually refers to the time of occurrence of disease (e.g. year, month, week, day of week, date of onset, time of onset) but may also refer to the time or length of exposure to a particular risk factor, the latency or incubation period of a disease, or the duration of symptoms. Collection of time data is essential when you want to describe changes in the *trend* of disease over time. Time data is difficult to collect in cross-sectional surveys. In cross-sectional surveys 'time' usually means the time (i.e. the date) of the survey.
>
> **PLACE** : You may also be interested in where a disease occurs. The place of occurrence can refer to a geographical point or an area that may be defined geographically, politically, or by a common, often natural, attribute (e.g. altitude, vegetation, proximity to a river). *Natural* boundaries such as rivers, mountains, and roads may be homogeneous in terms of (e.g.) the presence of a disease vector but political boundaries may be more convenient in terms of local reporting and control. There may be great variation in the size and population compared (e.g. villages, towns, compounds, regions). Place may be also be categorised in terms of oppositions such as urban / rural or forest / savannah. Note that place may also refer to place of exposure (e.g. place of residence, place of work) or place of treatment.
>
> **PERSON** : These are the characteristics of persons that may influence their exposure or susceptibility to disease. People may be described in terms of both their genetic (e.g. sex) and acquired characteristics (e.g. age, tribe, social class) or by activities, occupations, or infection.

A single characteristic of a person or their environment may lead to that person having an increased risk of disease, more commonly, several characteristics of time, place, and person are involved in the development of a pattern of disease within a *population*.

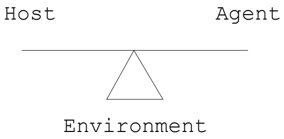
Selection of appropriate variables will depend on the results of a *semi-quantitative / semi-quantitative* pre-survey, your previous knowledge and experience, and an appreciation of the TPP factors known to be associated with undernutrition in similar contexts.

The *Host-Agent-Environment* framework

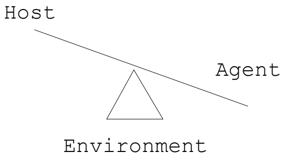
An alternative epidemiological framework is the *Host-Agent-Environment* (HAE):

> **HOST** : This is the same as *person* in the TPP framework.
>
> **AGENT** : This attribute is most commonly used for infectious diseases. Characteristics of the agent commonly used are its reservoir, vector, mode of transmission, incubation period, and virulence. Agents implicated in non-infectious diseases are usually considered to be *environmental* factors. In nutritional surveys, 'agent' might be the lack of a dietary factor or co-infection (e.g. worms).
>
> **ENVIRONMENT** : This is usually seen as having three components: the *biological* environment (e.g. animals and plants external to the host), the *social* environment (e.g. familial, social, cultural, economic, political factors *external* to the host), and the *physical* environment (e.g. temperature, humidity, water, air, radiation, chemicals) to which the host is exposed.

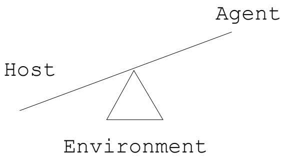
The HAE framework is of particular use when investigating changes in disease patterns. It proposes a *susceptible* host population in an environment that may affect the host population's susceptibility or exposure to a disease agent or the ability of the agent to cause disease. The pattern of disease results from an interaction between these three factors which are usually in a steady state. If any of the factors changes, sufficiently to affect the steady state, changes occur in one or both of the other factors. The HAE model can be seen as a balance:



Any change in the steady state that favours the agent or an increase in the pathogenicity of the agent will result in an increase in disease in the community:



Any changes that favour the host's ability to resist the disease agent or reduces the host's exposure to the disease agent will result in a decrease in disease in the community:

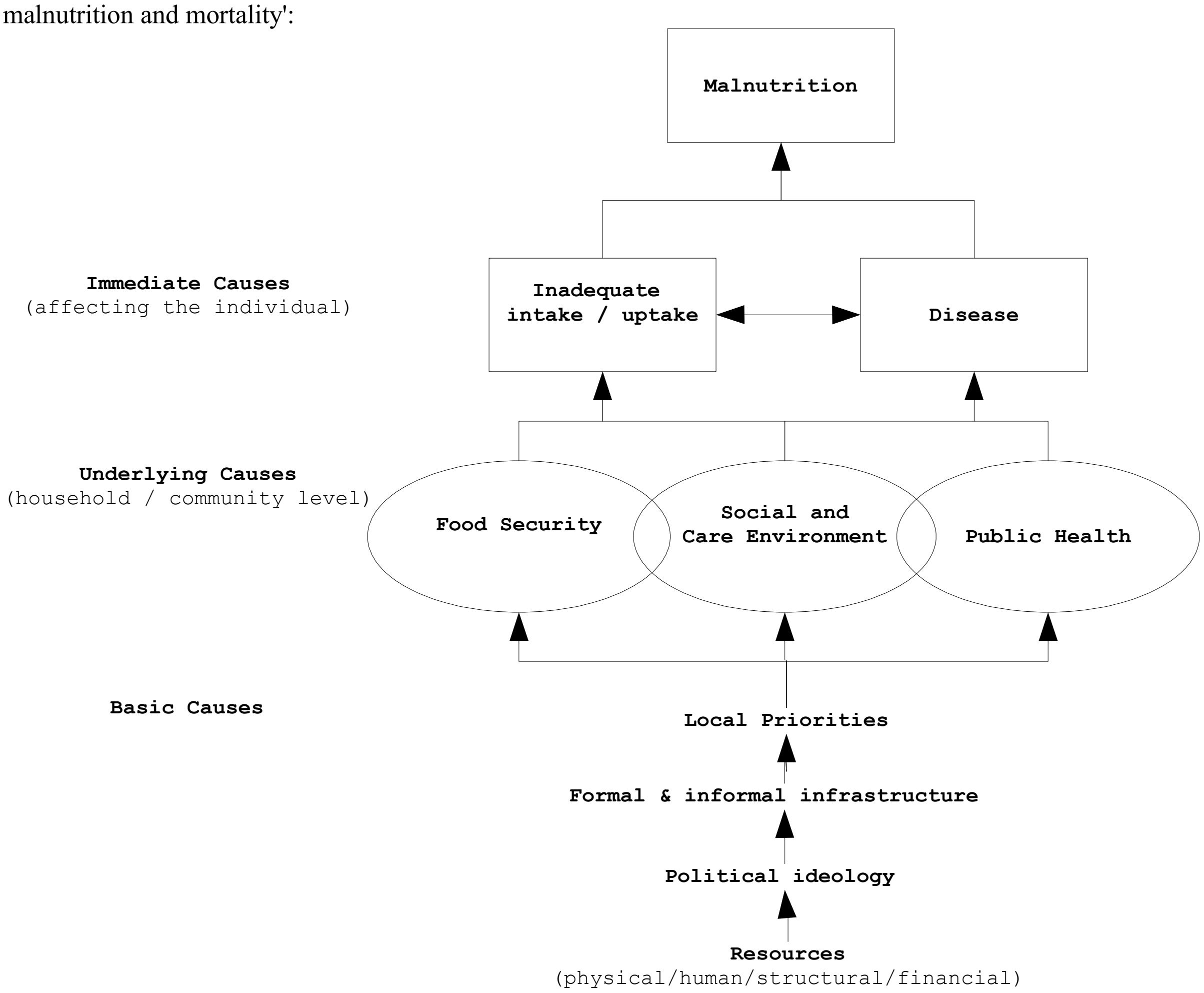


The HAE framework is most useful when considering communicable diseases.

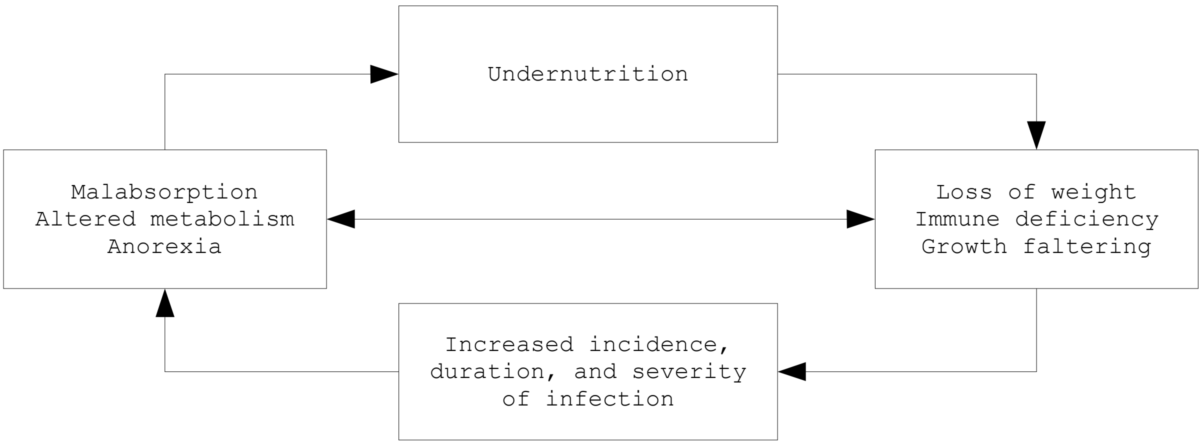
Selection of appropriate variables will depend on the results of a *semi-quantitative / semi-quantitative* pre-survey, your previous knowledge and experience, and an appreciation of the HAE factors known to be associated with undernutrition in similar contexts.

A *disease-specific* framework

In most situations much will already be known about the *global* epidemiology (as opposed to the *local* epidemiology) of the condition being surveyed. This information may also be used to help you define the survey dataset. A common framework for undernutrition is the 'UNICEF framework for underlying causes of malnutrition and mortality':


Malnutrition
Immediate Causes
(affecting the individual)
Inadequate intake / uptake
Disease
Underlying Causes
(household / community level)
Food Security
Social and Care Environment
Public Health
Basic Causes
Local Priorities
Formal & informal infrastructure
Political ideology
Resources
(physical/human/structural/financial)

Another useful framework is the *undernutrition-infection cycle*:



Surveys are likely to concentrate on the *immediate* and *underlying* causes in the UNICEF conceptual framework and the morbidity aspects of the undernutrition-infection cycle framework.

Defining a survey dataset

The *Time-Place-Person* (TPP) variables you will collect in the proposed survey are:

Defining a survey dataset

The *Host-Agent-Environment* (HAE) variables you will collect in the proposed survey are:

Defining a survey dataset

The *disease-specific* variables you will collect in the proposed survey are:

Defining a survey dataset

Combine the TPP, HAE, and disease-specific variables eliminating any duplicates:

Defining a survey dataset

Check to see if the variables you have listed are sufficient to meet the stated aims of the survey. Remove all non-essential variables and list the remaining variables:

Defining a survey dataset

Another approach to defining a survey dataset is to design *dummy tables*. These are the tables and figures that you will need to include in the final survey report (which should be implied by the survey aims). You can go about this by writing down a description of the table:

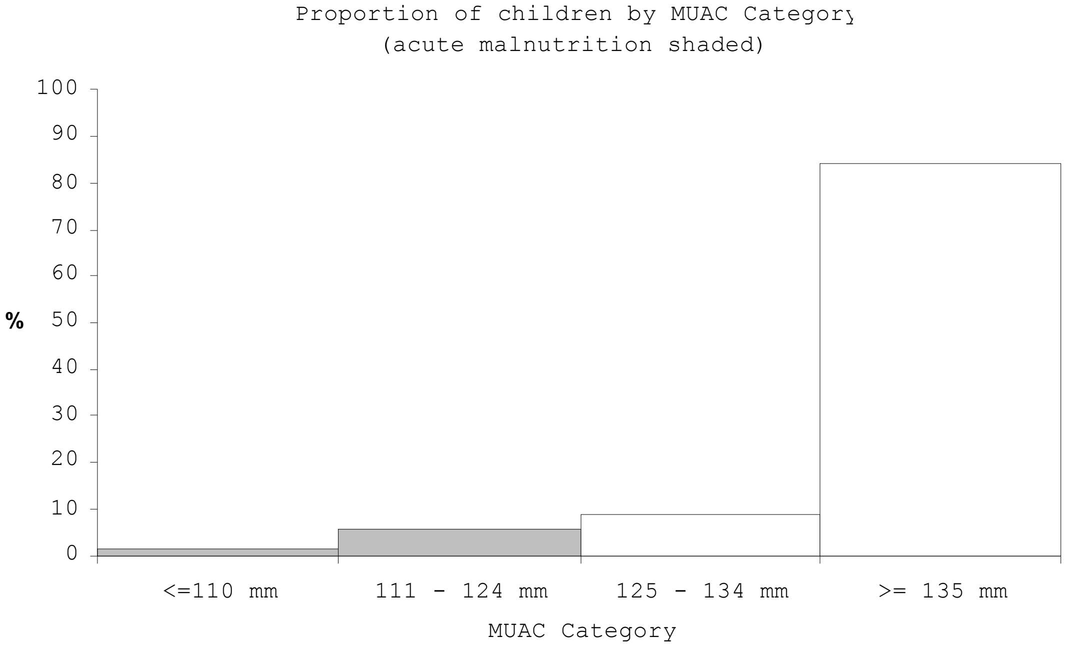
> *A table of sex (male, female) by nutritional status (adequate, moderate, severe) showing numbers and relative proportions.*

Or by actually sketching out the required table:

```
                        NUTRITIONAL STATUS

             ¦   Adequate¦   Moderate¦     Severe¦      Total¦
 +-----------+-----------+-----------+-----------+-----------+
 ¦Male       ¦           ¦           ¦           ¦           ¦
 ¦ Count     ¦       ####¦       ####¦       ####¦      #####¦
 ¦ Percent   ¦       ##.#¦       ##.#¦       ##.#¦           ¦
 +-----------+-----------+-----------+-----------+-----------+
 ¦Female     ¦           ¦           ¦           ¦           ¦
S¦ Count     ¦       ####¦       ####¦       ####¦      #####¦
E¦ Percent   ¦       ##.#¦       ##.#¦       ##.#¦           ¦
X+-----------+-----------+-----------+-----------+-----------+
 ¦Total      ¦           ¦           ¦           ¦           ¦
 ¦ Count     ¦       ####¦       ####¦       ####¦      #####¦
 ¦ Percent   ¦       ##.#¦       ##.#¦       ##.#¦           ¦
 +-----------+-----------+-----------+-----------+-----------+
```

Or graph:



By doing this you can work backwards from the desired tables and graphs to make a list of variables required to produce the desired output from the survey. The example table requires data for sex, anthropometry (MUAC or weight and height), and the presence or absence of oedema. The example graph requires data for MUAC.

Defining a survey dataset

You might use a table to present a large quantity of data (e.g. nutritional status by age-group and sex) but such large tables can be confusing for readers of survey reports. You should also consider using separate tables and charts for (e.g.) each sex. Large and complex tables may be best included in the appendices of survey reports.

When designing dummy tables and charts you should be thinking about the presentation of data in the survey report as well as using the dummy tables and charts to inform the definition of the survey dataset. You should always bear the 'target audience' in mind when designing dummy tables and reports.

Defining a survey dataset - dummy tables

Make a list of the tables and figures (in words or as sketches) needed for the final survey report. List the variables you need to collect and check them against your previous list:

Defining a survey dataset - dummy tables (continued)

Make a list of the tables and figures (in words or as sketches) needed for the final survey report. List the variables you need to collect and check them against your previous list:

Defining a survey dataset

Now you have a list of the variables you will need to collect to meet the aims of your survey you should start thinking about how data will be collected, coded, and stored. You can use a framework that defines the main *attributes* of each variable to help you do this. The main attributes you need to consider are:

> **NAME** : What is the name of the variable?
>
> **TYPE** : What is the type of variable (i.e. the type of data that will be collected and recorded)? It is easiest to think in terms of variables being either numbers or text.
>
> **LENGTH** : How many numbers or letters will the variable have to hold?
>
> **CODES** : Which codes will be used to represent different categories of a variable?

For example, you may have a variable that holds the sex of each respondent using the number **1** to indicate that the respondent is male and the number **2** to indicate that the respondent is female. In this case you might decide to use a NUMBER type variable called SEX of sufficient LENGTH to hold a single digit.

Here is an example of some variable definitions:

Name	Type	Length	Codes
CL	Number	2	None (cluster ID)
SEX	Number	1	1 = Male 2 = Female
HEIGHT	Number	3+1 e.g. 104.3	None (cm)
WEIGHT	Number	2+1 e.g. 15.6	None (cm)
MUAC	Number	3	None (mm)
AGE	Number	2	None (months)
ARI	Number	1	1 = Yes 2 = No
ORPHAN	Number	1	1 = Mother 2 = Father 3 = Both 4 = Neither

Sometimes the choice of codes may be obvious (e.g. sex) but may be more difficult to define for other variables (e.g. occupations). It is often helpful to draw up a set of possible and useful coding schemes before deciding which one to use. Do **<u>not</u>** finalise coding schemes until you have piloted the survey questionnaire (data-collection form) and survey data-systems.

Defining your dataset in this way will make it easier to design data collection forms and computerise data-entry, data-management, and data-analysis.

In a *cluster sampled survey* (described later in this document) you will need to record the cluster number of each subject. As you will know the location of each individual cluster this variable can be grouped during analysis to create derived variables of place (e.g. urban vs. rural, lowland vs. highland, &c.).

Defining a survey dataset

Define the variables you have decided to collect in the previous exercises in terms of their names, types, lengths, and codes:

Name	Type	Length	Codes

Defining a survey dataset

Define the variables you have decided to collect in the previous exercises in terms of their names, types, lengths, and codes:

Name	**Type**	**Length**	**Codes**

Prepare the survey data collection forms

Data collection forms are used for recording of the survey dataset during interview and/or examination. The data collection form will accurately reflect the dataset and should be clear and easy to use. It should also show coding and routing instructions. Here is part of a data-collection form that might be used to record data from a nutritional survey:

AD31 Childhood Nutrition Survey **Data Collection Form**

Cluster : ¦__¦__¦   Village : ¦__¦__¦__¦__¦__¦__¦__¦__¦__¦__¦   Date DDMM: ¦__¦__¦__¦__¦

Ch	**Age (months)**	**Sex (M/F)**	**Wt (Kg)**	**Ht (cm)**	**MUAC (mm)**	**Oedema (Y/N)**	**FB (Y/N)**	**GM (Y/N)**	**Orphan (CODE)**
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

Each row of the form holds data for an individual respondent. Each column corresponds to a variable in the survey dataset. This type of tabular form is suited to the collection of simple datasets (i.e. few variables, simple coding) for which each row should contain no missing values (i.e. there are no skips or filters and all variables are collected for all respondents). Complex datasets are best collected using separate forms for each respondent.

Data-collection forms should be *piloted* (i.e. tested) before being used in the field. This is particularly true of forms used to collect complex datasets.

Prepare the survey data collection forms

Design and print (using a word-processor) a data-collection form to collect your survey dataset.

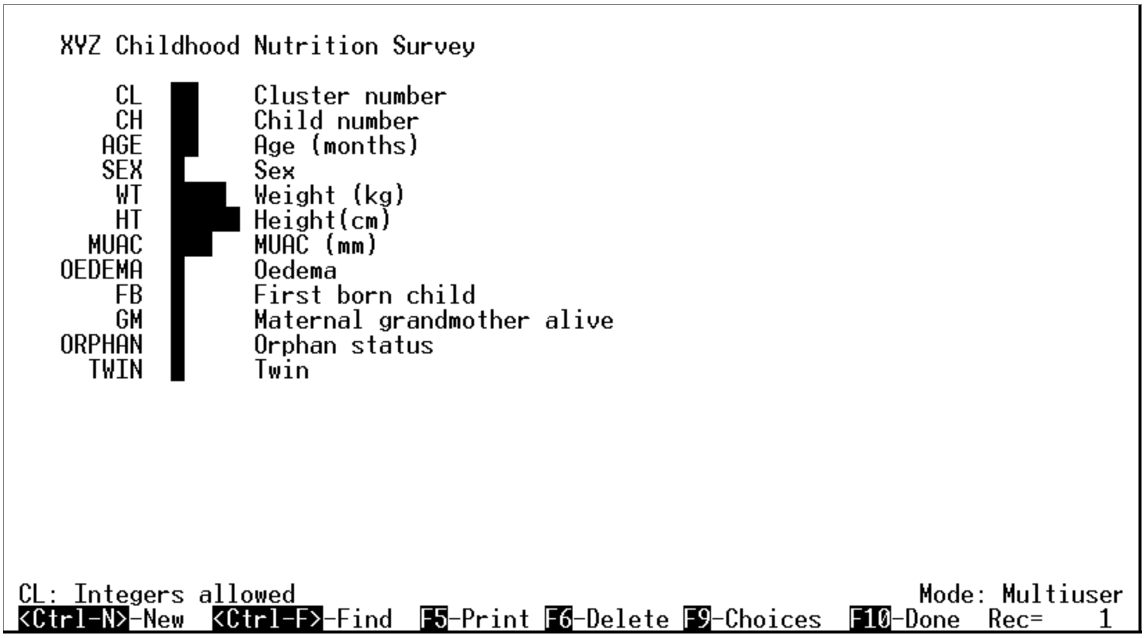
Creating a data-entry system

Once you have decided on a survey dataset and designed the data-collection form you can proceed to create a data-entry system. Here is how the example form might be translated into an **`EpiInfo/EpiData .QES`** file:

```
XYZ Childhood Nutrition Survey

    CL  ##    Cluster number
    CH  ##    Child number
   AGE  ##    Age (months)
   SEX  <A>   Sex
    WT  ##.#  Weight (kg)
    HT  ###.# Height(cm)
  MUAC  ###   MUAC (mm)
OEDEMA  <Y>   Oedema
    FB  <Y>   First born child
    GM  <Y>   Maternal grandmother alive
ORPHAN  #     Orphan status
  TWIN  <Y>   Twin
```

Which, when used with **`ENTER`** module of **`EpiInfo/EpiData`** to create a **`.REC`** file will result in a data-entry screen:



```
XYZ Childhood Nutrition Survey

    CL        Cluster number
    CH        Child number
   AGE        Age (months)
   SEX        Sex
    WT        Weight (kg)
    HT        Height(cm)
  MUAC        MUAC (mm)
OEDEMA        Oedema
    FB        First born child
    GM        Maternal grandmother alive
ORPHAN        Orphan status
  TWIN        Twin

CL: Integers allowed                                        Mode: Multiuser
<Ctrl-N>-New  <Ctrl-F>-Find  F5-Print F6-Delete F9-Choices  F10-Done  Rec=    1
```

The layout of the data-entry screen should follow the layout of the data-collection form. This is not possible with tabular forms. If you use tabular forms make sure that you follow the order of the columns on the form. You can then use the **`UPDATE`** command in **`EpiInfo ANALYSIS`** to enter data in tabular format. Complex coding information can be discarded as this tends to clutter up the data-entry screen. Remember to give your variables short and meaningful names as this will make it easier to perform data-management and data-analysis tasks.

It is also advisable to add some form of data checking to a data-entry system to enforce the coding scheme and apply other (e.g. range) checks.

Creating a data-entry system

Build and test a data-entry and checking system that accurately reflects the survey dataset and your data-collection form.

Sampling considerations

If all individuals in the target population are sampled then you would get a precise picture of the health or nutritional status of the target population. Such a survey would be expensive and time-consuming so a sub-group of the target population (called the *sample*) is surveyed. This sample **must** be *representative* of the target population if the results of an analysis of the collected data are to be representative of the target population

A representative sample

The sample taken **must** be representative of the target population if the results of the survey are to be generalised to the target population. For a sample to be representative:

1. Each individual in the target population must have an **equal** chance of being included in the sample.

2. The selection of one individual should be **independent** of the selection of another individual.

Precision

The results of a sample survey give an *estimate* of the results that would arise from a survey of the entire target population. If a second sample were drawn from the same target population, slightly different results would arise because of the variation of the individuals selected for the sample. This variation is known as *random variation*.

The *sample* prevalence is considered to be the *best estimate* of the *true* or *population* prevalence. It is called the *point estimate* of the prevalence. The effect of random variation can be accounted for statistically by calculating a range of values around the point estimate of the prevalence that has a specified *probability* of including the true value of the prevalence . The specified probability is called the *confidence level* and is usually 95%. The range of values that the true prevalence could take is called the *confidence interval*. The endpoints (minimum and maximum) of the confidence interval are called the *confidence limits*.

With a 95% confidence interval we are 95% sure that the confidence interval will include the true prevalence. We do not know for certain that the confidence interval includes the true prevalence. The chances are 95% that the confidence interval does include the true prevalence. The chances are 5% that the confidence interval does **not** include the true prevalence.

The size of the confidence interval is influenced by size of the sample taken and is usually specified, in sample size calculations, as a *margin of error* (i.e. the prevalence **plus** or **minus** a given margin of error). The width of the confidence interval is twice the acceptable margin of error.

Probability of error

When a sample is taken there is a probability that the sample will not be representative of the target population. In most surveys a risk of 5% is acceptable. This means that there is a 5% chance that the survey results will not reflect the health status of the target population.

Sampling methods

The sampling method used for a particular survey will often be decided by the nature of the target population and the context of the survey. There are three main sampling methods:

> **RANDOM SAMPLING** : This is the best method but requires the existence of a sampling frame that lists every individual in the target population together with information on how to locate them. The list **must** be accurate and up-to-date. Individuals are randomly drawn from the list using a random number table. In most situations, such a list will not be available.
>
> **SYSTEMATIC SAMPLING** : This method takes into account some existing ordering of the target population. In a household survey you would usually rely upon the geographical organisation of the area to be surveyed. Each household must have the same chance of being sampled by a team moving systematically across the target area. One household out of every *N* households will be visited where *N* (the *sampling interval*) is defined by the size of the target population, the number of households in the target population area, and the required sample size. Systematic sampling is most useful in well organised refugee camps or grid-layout towns. Alternatively all dwellings can be enumerated (huts are easier to count than heads) and a systematic sample of dwellings taken from the resulting list. It is common practice to sample all eligible individuals in a selected household.
>
> **CLUSTER SAMPLING** : This method is used when random or systematic sampling is not possible (e.g. when no reliable population register is available, when the geographical organisation does not permit every household to be visited, or resources are not available to visit every household). With cluster sampling the population is grouped into smaller units for which relative population sizes can be estimated. The smallest unit for which populations can be estimated is chosen as the basis for sampling. A number of these units (*clusters*) are sampled. In each cluster a certain number of individuals will be sampled. The chance that a unit will be sampled is proportional to its population size. This sampling technique does **not** meet the second criteria for representativeness of a sample. The fact that individuals are selected within a cluster by proximity means that the selection of one individual is **not** independent of the selection of another individual. Within each cluster it is likely that individuals will be similar to each other. This is called the *design effect* and it is taken into account when calculating the required sample size for a survey and, later, when analysing data from a cluster sampled survey. When cluster sampling is used, the sample size will be larger than if random or systematic sampling is used. It is **not** possible to accurately predict the design effect in advance of calculating the required sample size. This means that you will have to rely on informed guesswork when specifying a design effect in sample size calculations. Experiences from previous surveys suggest that for estimating the prevalence of conditions that may have a uniform distribution in the target population (e.g. low vision or blindness due to cataract), it is safe to specify a design effect of 1.5 or 2.0. For estimating the prevalence of conditions that might be highly clustered (e.g. measles, trachoma), larger design effects (5.0 or higher) should be specified. In such situations, a survey that provides a *wide-area* estimate might not be the best tool for estimating prevalence since a single estimate for a wide area might not reflect the prevalence in any single community. For other situations you will have to review the results of similar surveys or seek the advice of an experienced survey statistician. For surveys of infectious diseases or diseases related to environmental exposure the design effect is likely to be large as individuals within a cluster will tend to have a similar status with regard to the condition under study. The number of clusters and the number of individuals within each cluster has been the subject of much study and as a general rule it is **not** advisable to sample fewer than about thirty clusters. It is better to have many small clusters than a few large clusters.

The cluster sampling method described here is also known as *two-stage* cluster sampling (the sampled clusters are stage one and the individuals sampled within each cluster are stage two). Variations on this sampling method are frequently used. These usually involve an additional sampling stage based on population units or a characteristic of the target population (*stratified* sampling). All *complex* or *multi-stage* sampling methods have a design effect that must be accounted for in sample size calculations and data-analysis. More complex sampling methods tend to have larger design effects than simpler methods. When considering sampling methods you should always try to choose the least complex method capable of meeting the aims of the survey.

Choice of sampling method

When an accurate population register is available then you should use random sampling. If no accurate population register is available but the population is living in a small well-defined area then systematic sampling should be used. In all other situations, cluster sampling must be used.

Surveys that use random or systematic sampling are relatively easy to analyse using readily available computer programs such as the **`EpiInfo ANALYSIS`** module but surveys that use cluster sampling need special analytical techniques that can control for the design effect of the sampling strategy used. The **`CSAMPLE`** program in **`EpiInfo`** is a simple tool for analysing data from complex sample surveys.

Practical sampling - random sampling

Random sampling requires the existence of a reliable population register. Sample size calculation requires an estimate of the size of the target population, an expected prevalence, a probability of error, and a desired precision. A serial number is assigned to each individual on the population register and a list of random numbers of the required sample size is generated. Individuals with serial numbers corresponding to the generated random numbers are included in the sample. Statistical tables will generally include a table of random numbers and computer packages such as **`EpiTable`** (in **`EpiInfo`**) and **`EpiCalc`** can generate random number lists. You should make some provision for refusals after calculating the required sample size (e.g. increase the sample size by 20%).

Practical sampling - systematic sampling

Systematic sampling is best used for small well-defined areas. Sampling is based on a register of households or the geographic arrangement of dwellings. The steps required for systematic sampling are:

1. Estimate the size of the target population and the number of households in the target population.

2. Estimate the number of eligible individuals in the target population.

3. Calculate the sample size as for random sampling.

4. Estimate the required number of households by dividing the calculated sample size by the average number of eligible individuals in each household (i.e. the number of eligible individuals in the population divided by the number of households in the survey area). You may need to estimate the size of the eligible population using the information such as age structure or sex ratio from a census.

5. Determine the sampling interval by dividing the number of households in the sample area by the number of households required to meet the sample size.

6. Select the households to be sampled by selecting the first household using a random number between one and the sampling interval and then systematically sample every N^{th} house (N is the sampling interval). If more than one eligible individual is found in a selected household then all eligible individuals should be sampled. If no eligible individuals are found in a selected household then the closest household is selected. If an eligible individual is not present at the time of the visit then the household will have to be revisited at a later time or date.

It is important that that you do not overestimate the number of eligible individuals in the target population as this will cause the sampling interval to be too large and you will fail to meet the required sample size. Also, it is important that you select a sampling interval that ensures that all households could have been selected. It is possible that you may meet your sample size after covering (e.g.) only 75% of a camp. In this situation you should continue sampling until the entire sample area has been covered.

Practical sampling - cluster sampling

The steps required for taking a two-stage cluster sample are:

1. Determine the geographic units and their relative populations. Cluster sampling assumes the target population is grouped into smaller geographical units. The smallest geographic unit for which the population can be estimated is chosen. For each of these units the population of eligible individuals is estimated.

2. Calculate the cumulative eligible population by compiling a table of population units and their eligible populations (you may need to estimate the eligible population in each unit using the information such as the age structure or sex ratio from a census). The cumulative population is calculated by adding the population of each unit to the sum of the populations of all preceding population units.

3. Calculate the sample size using a sensible expected design effect of **at least** 2.0 unless you have good reasons to suspect that the design effect will be lower than 2.0.

4. Calculate the sampling interval as the total eligible population divided by the number of clusters. A minimum of about 30 clusters is required. In each of these clusters the number of individuals to be selected is the sample size divided by the number of clusters. You may vary the number of clusters to more than thirty clusters so as to make each cluster a manageable size (e.g. the number of individuals it is viable for a data-collection team to sample in a single day). It is better to have many small clusters than a few large clusters.

5. Determine the location of the first cluster. The location of the first unit to be used in the sample is randomly selected within the first sampling interval using a random number between 1 and the sampling interval

6. Select the clusters. The sampling interval is added to this random number and the unit which includes this sum in its cumulative eligible population is selected. The location of further clusters is determined by adding the sampling interval to the previous sum and selecting the unit containing the resulting sum in the cumulative eligible population. A large unit may be sampled more than once. A small unit (i.e. one smaller than the sampling interval) may be skipped over and not sampled at all.

7. Select individuals within clusters. A data-collection team visits the unit where the cluster is located and goes to the geographical centre of the cluster. A random direction is picked by spinning a bottle. The bottleneck indicates the sampling direction. A member of the team (called the surveyor) walks in that direction to the edge of the unit counting the number of households encountered on the way. The first household to be selected is chosen at random using a random number (it may prove useful to have made a sketch map of the households on the route). Subsequent households are chosen by their proximity to the first household. This may be determined as the first household to the left or right but this should be the same for all clusters and decided in advance. All eligible individuals in a household are sampled. If an eligible individual is not present at the time of the visit then the household will have to be revisited at a later time or date. It is possible to use other *within-cluster* sampling strategies. If a village level population register is available then it may be possible to use that to take a random sample. It may also be possible to use a map or co-ordinate scheme to take a random or systematic sample of households.

Warning

In a cluster sampled survey it does **not** make sense to analyse data and produce results for individual clusters. What is observed in one cluster is **not** representative of the whole population or the sub-population from which it is drawn. To avoid misinterpretation of results by people not acquainted with survey techniques, results should **not** be presented by cluster.

Cluster sampling - a worked example

The figures used in this worked example are **different** from those for AD31. The area to be sampled consists of ten villages and towns:

Unit	Population
1	1500
2	2100
3	750
4	1150
5	3600
6	200
7	1800
8	4500
9	17000
10	3400

The age structure of the target population taken from a recent census is:

Age	%
0 - 14 years	42.85%
15 - 19 years	6.82%
20 - 24 years	6.92%
25 - 29 years	7.57%
30 - 39 years	13.46%
40 - 49 years	9.78%
50 - 59 years	6.53%
Over 60 years	6.07%

We are interested in children of five years of age and younger. In the absence of detailed data, it is reasonable to assume that this age-group makes up about 20% of the population of a developing country. In severe famine, however, there may have been considerable mortality in this age-group and population estimates should be revised downwards. The population estimates for the 0-14 years group in the table above means that is probably sensible to use the 20% estimate in this situation. We can use this proportion to estimate the eligible population within each cluster. We also keep a running total of the eligible population:

Unit	**Population**	**Eligible population** †	**Cumulative eligible population**
1	1500	300	300
2	2100	420	720
3	750	150	870
4	1150	230	1100
5	3600	720	1820
6	200	40	1860
7	1800	360	2220
8	4500	900	3120
9	17000	3400	6520
10	3400	680	7200
ALL	36000	7200	7200

† The eligible population is estimated as 20% of total population.

Cluster sampling - a worked example (continued)

The required sample was calculated as 694 using the following parameters:

Parameter	Value
Population size	7200
Estimated prevalence	10%
Maximum error	3%
Probability of error	5%
Design effect	2.0

It is estimated that the a single data collection team can collect data from 50 individuals per day. It seems sensible that each team collect two clusters of 25 individuals per day so this is set as the cluster size. Using a cluster size of 25 individuals for a sample size of 694 requires that 28 clusters be sampled (694 / 25 = 27.76). twenty-eight is close enough to 30 for this to be a reasonable number of clusters to sample. The sampling interval is thus 257 (7200 / 28 = 257.14). The final sample size is rounded up to 700 (28 clusters of 25 individuals).

The location of the first cluster is determined using a random number between one and the sampling interval (we will use 33 which was generated using the random number function on a scientific calculator) and other clusters are selected by successive additions of the sampling interval.

The first cluster is at position 33 and is in unit one, the second cluster is at position 290 (33 + 257 = 290) and is also in unit one, the third cluster is at position 547 (290 + 257 = 547) and is in unit two, the fourth cluster is at position 804 (547 + 257 = 804) and is in unit three. We continue to allocate clusters in this manner until we have allocated all 28 clusters:

Unit	Cumulative eligible population	Positions	Clusters in this unit
1	300	1-300	2
2	720	301-720	1
3	870	721-870	1
4	1100	871-1100	1
5	1820	1101-1820	2
6	1860	1821-1860	1
7	2220	1861-2220	1
8	3120	2221-3120	4
9	6520	3121-6520	13
10	7200	6521-7200	2

In practice, large population centres (e.g. unit 9 in the example above) are broken down into smaller population units based upon geographical (e.g. roads, rivers, railways) or political boundaries and clusters allocated accordingly.

Cluster sampling - Advantages and problems

The choice of cluster sampling as a sampling method is often presented as a negative choice (i.e. only to be used when simple random sampling or systematic sampling are not feasible) but there are some positive reasons for choosing cluster sampling. Principal amongst these are lower cost, easier supervision of data-collection (leading to better quality data), and the ability to 'nest' the collection of cluster-level (e.g. village / neighbourhood) *qualitative* data (e.g. from focus groups) within a *quantitative* survey. Unlike simple random samples, refusals do not prevent you from reaching the required sample size surveys as within-cluster sampling continues until the required sample size is met. You should, however, record and report refusals.

One problem that occasionally occurs with cluster sample surveys is the inability to visit a selected village (e.g. due to severe weather or poor security). If you think that this might be a problem in your survey then you should select *contingency clusters* that will only be sampled if another cluster cannot be sampled. It is usual to make a 10% provision. A thirty cluster survey would have thirty clusters plus three contingency clusters that would be sampled only if one of the thirty main clusters could not be visited. The contingency clusters would usually be clusters 11, 22, and 33. Specifying contingency clusters in this way ensures that they do not always come from the end of the list of population units: If the list of population units is sorted by location then selecting contingency clusters from the end of the list might cause one area to be oversampled resulting in a non-representative sample of the survey area.

It is important to note that it is **not** valid to analyse the *quantitative* survey data by cluster. The spin-the-bottle technique selects households by proximity to each other and also selects individuals within the same household. This violates the rule that the selection of one individual should be independent of the selection of another individual. In addition, the cluster size is usually much too small to produce reliable estimates of within-cluster prevalence. What is observed in one cluster is **not** representative of the whole population or the sub-population from which it is drawn

Cluster sampling - an alternative approach

An alternative approach to using the successive addition of a sampling interval is to generate a list of unique random numbers (e.g. 30 random numbers for 30 clusters) ranging between 1 and the total eligible population. This list can then be used to locate clusters within the cumulative eligible population.

Statistical tables will generally include a table of random numbers and computer packages such as **`EpiTable`** (in **`EpiInfo`**) and **`EpiCalc`** can generate random number lists.

Continuing with the previous example, we aim for a sample size of 700 from an eligible population of 7200 collected as 28 clusters of 25 persons. We, therefore, need to generate a list of 28 unique random numbers between 1 and 7200. The list below was created with **`EpiTable`**:

```
  97       128       381       553      1040      1628
1784      2135      2326      2849      2882      2998
3102      3425      3434      4046      4274      4439
4754      5122      5360      5442      5486      5581
6642      6863      7137      7154
```

These random numbers are then used to select the cluster positions from the cumulative eligible population:

Unit	Positions	Random numbers (from list)	Clusters in this unit
1	1-300	97, 128	2
2	301-720	381, 553	2
3	721-870		0
4	871-1100	1040	1
5	1101-1820	1628, 1784	2
6	1821-1860		0
7	1861-2220	2135	1
8	2221-3120	2326, 2849, 2882, 2998, 3102	5
9	3121-6520	3425, 3434, 4046, 4274, 4439 4754, 5122, 5360, 5442, 5486 5581	11
10	6521-7200	6642, 6863, 7137, 7154	4

The clusters selected using this method may differ slightly from those selected using successive addition of a sampling interval but this method still ensures that clusters are selected proportional to their population size.

If you choose to use this method of selecting clusters then you should ensure that the computer package you use is configured to produce a *uniform* distribution of random numbers.

Determining the required sample size

To calculate a sample size for a survey we need the following items of information:

POPULATION : An estimate size of the target population. Larger populations require a larger sample size in order to achieve a reliable result.

EXPECTED PREVALENCE : This must be a guess. If you knew the true prevalence of a health condition in the target population you would not need to undertake a survey. The nearer the expected prevalence is to 50%, the more individuals are needed in the sample for the same degree of precision. It is a good idea to conduct a small pre-survey in order to get a rough idea of the expected prevalence. For nutrition surveys, a sample of one hundred or so children using MUAC may prove useful. For most developing countries, a global prevalence of about seven percent is considered to be ‘normal’.

REQUIRED PRECISION : How precise should the estimate be? The greater the precision required, the more individuals needed in the sample. The required precision specifies the width of the confidence interval. It is the margin of error. The precision needed also depends upon the expected prevalence. For example, if the expected prevalence of a condition is 1% then an estimate of 1% ± 0.25% may be acceptable whereas an estimate of 1% ± 5% is useless.

PROBABILITY OF ERROR : What chance of accidentally drawing a non-representative sample is acceptable? The smaller the probability of error, the more individuals are needed in the sample.

DESIGN EFFECT : This is 1.0 for random or systematic samples but will be greater than 1.0 for cluster sample surveys. Complex (e.g. cluster) sample surveys need larger samples than random or systematic sample surveys. It is **not** possible to accurately predict the design effect in advance of calculating the required sample size. This means that you will have to rely on informed guesswork when specifying a design effect in sample size calculations. Surveys of immunisation coverage and nutritional status often use a design effect of 2.0 but for other situations you will have to review the results of similar surveys or seek the advice of an experienced survey statistician. For surveys of infectious diseases or diseases related to environmental exposure the design effect is likely to be large as individuals within a cluster will tend to have a similar status with regard to the condition under study. In such situations, a survey that provides a *wide-area* estimate might not be the best tool for estimating prevalence since a single estimate for a wide area might not reflect the prevalence in any single community.

Another factor needs to be accounted for:

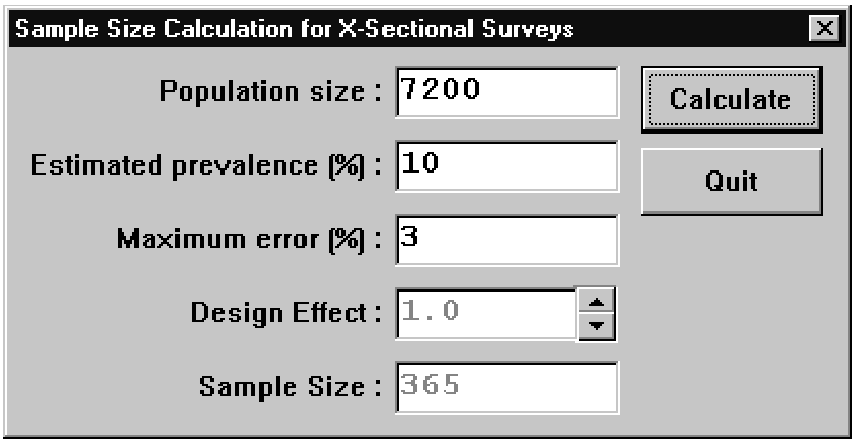
PRACTICALITY : The best sample size yields an estimate of high precision with a small probability of error but, often, the resources or time available to a survey will dictate that a smaller than ideal sample be taken. It is important that a sample size that is attainable given the available resources is selected otherwise there is a risk that the survey will not achieve its sample and produce misleading results. You may need to compromise on the level of precision or the probability of error to make a survey practicable. You may also need to rethink the aims of the survey or reduce the width of the dataset (i.e. the number of variables) to be collected in order to achieve a reasonable sample size collectable within a reasonable period of time. For extremely rare conditions (e.g. childhood blindness) it may be necessary to adopt a *purposive* sampling method where an attempt is made to find all cases in the target population through exhaustive *case-finding* (detective) work rather than attempting to find cases using a (very large) probability sample.

Using `SampleXS` to determine the required sample size

`SampleXS` is a program that you can use to help you calculate the required sample size for a cross-sectional health survey. In this exercise you will use **`SampleXS`** to calculate a sample size for a prevalence survey of a health condition (e.g. undernutrition) using the following assumptions:

Parameter	Value
Population size	7200
Estimated prevalence	10%
Maximum error	3%

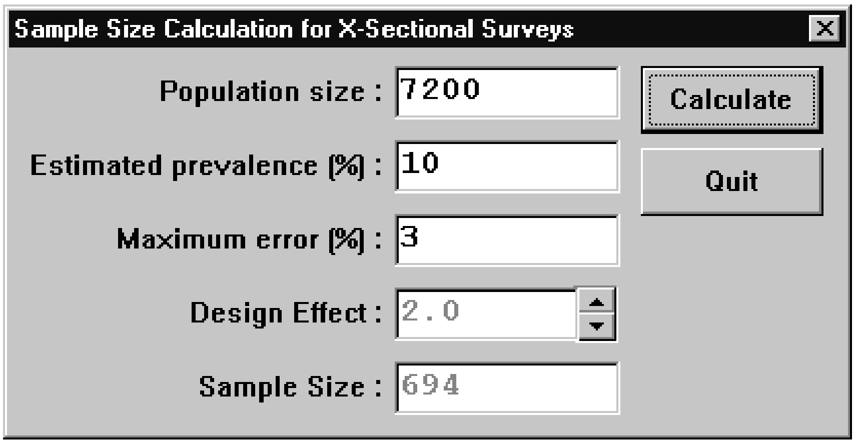
Start **`SampleXS`** and enter these values into the relevant data-entry boxes (**`SampleXS`** assumes that the maximum error (i.e. the margin of error or required precision) specified is for a 95% confidence interval and that the probability of error is 5%). Specify a design effect of 1.0 and click the **`Calculate`** button:



Population size : 7200
Estimated prevalence (%) : 10
Maximum error (%) : 3
Design Effect : 1.0
Sample Size : 365

The resulting sample size (365) is the sample size required for a survey using random or systematic sampling.

To calculate a sample size for a cluster sampled survey you should specify a design effect. A design effect of 2.0 is usually specified for nutrition surveys. To specify a design effect click on the up and own arrow buttons next to the design effect box and click the **`Calculate`** button:



Population size : 7200
Estimated prevalence (%) : 10
Maximum error (%) : 3
Design Effect : 2.0
Sample Size : 694

The required sample size increases as the design effect increases. Experiment with different design effects and see what effect this has on the required sample size for a survey. Experiment with the different parameters and note the effect that they have on the required sample size.

Sample size and sampling method

Calculate the sample size and cluster locations for a two-stage cluster sample survey of AD31 using the background information provided in the first few pages of this manual.

Other sampling strategies

We have considered the three main sampling methods commonly used in surveys (simple random, systematic, and two-stage cluster sampling). There are, however, other sampling strategies that could be used. These are:

> **STRATIFIED SAMPLING :** This is used when the population consists of distinct subgroups (strata) which may differ with respect to the variable under study or are themselves of interest. For example you may have a population consisting of residents living in their own homes and displaced persons living in camps. A sample (simple random, systematic, or cluster) is taken from each stratum. The calculation of a confidence interval is based on a combination of the standard errors of the estimated prevalence in each stratum and, like cluster sampling, requires specialist software. The advantage of using a stratified sample is that a sufficient sample size is taken in each stratum so as to ensure that we can obtain a reasonably precise prevalence estimate for each stratum. You can think of stratified sampling as undertaking several surveys that may be combined.
>
> **COMPREHENSIVE / EXHAUSTIVE SAMPLING :** This involves sampling the entire population. This results in a *true* prevalence rather than an *estimate* of the true prevalence and there is no need to calculate confidence intervals. This form of sampling is rarely used in nutrition surveys except in the case of a screening exercise to identify households (e.g.) to receive a dry supplementary ration.
>
> **PURPOSIVE SAMPLING :** This is a non-probabilistic sampling method. Purposive sampling is a search for cases. It is sometimes called *active case-finding* or *optimally-biased sampling*. It is often used in preliminary surveys where it is important to find undernourished children and identify potential risk-factors that could be topics of investigation in a full survey or a case-control study. This sampling method is also used with very low prevalence conditions (e.g. childhood blindness) which would require a very large sample for a conventional survey. In some cases it may be possible for a case-finding method to find all, or nearly all, cases in surveyed communities. In such situations the sample may be treated as an *exhaustive* sample.
>
> **CONVENIENCE / OPPORTUNISTIC SAMPLE :** This method relies on a sample being readily available. Children attending an outpatient clinic might, for example, be sampled. Such a sample is unlikely to be representative of the overall population and results should be interpreted with caution.

Practical planning tasks

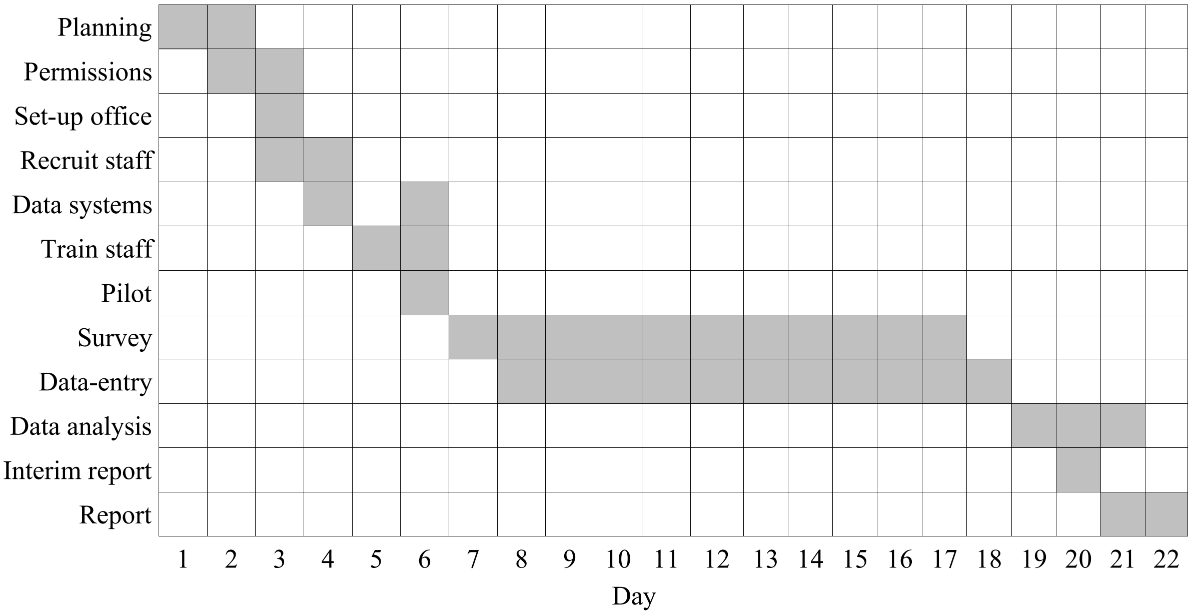
There are practical management tasks that need to be completed before a survey can be started. These are:

STAFFING : You should make a list of the personnel required to undertake the survey. It is probably best to do this by first listing the main tasks involved in executing the survey and drawing up outline job descriptions. Make sure that you allocate adequate staff and leave a margin for possible loss of staff due to (e.g.) ill health. For nutritional survey, a survey team will consist of three people if weight for height is used. If MUAC is used then teams can be as small as a single worker.

RECRUITMENT and **TRAINING** : In most situations you will not have an experienced survey team available. This means you will have to recruit and train for data-collection and data-entry work. Ensure that you allow adequate time for recruitment and training. Training should not be an ad-hoc affair. Make sure that you have a well-thought out training program prior to recruitment.

EQUIPMENT : Make a list of the equipment required and arrange for its purchase and transportation. Take time to ensure that this list is exhaustive as equipment can often be impossible or prohibitively expensive to purchase in the field.

TIMETABLE : Make a detailed timetable of activities. You will need to allow time for recruitment and training as well as for a pilot study of the data-collection form and methods. A special type of graph called a *Gantt* chart is useful for project planning as it shows which activities depend upon other activities having been completed and which activities can run concurrently. For example:



Once a Gantt chart has been created it can be transferred in greater detail to a year-planner or separate and detailed Gantt charts can be created for each activity.

BUDGET : All of the above tasks should inform the size of the budget you will need to complete the survey. Make sure you build some ‘slack’ into the budget to cope for unforeseen circumstances. You should produce a detailed and categorised cash-flow projection using a spreadsheet program if your survey will last for a period of longer than a few weeks.

Practical tasks

Complete the following tasks:

1. Make a list of the members of the survey team giving a brief description of the main tasks and responsibilities allocated to each member.
2. Prepare a syllabus for staff training.
3. Make a list of equipment required.
4. Prepare an outline project timetable.
5. Prepare an estimated budget

Prepare a short (c. 15 minutes) presentation of your proposed survey design. This Should include:

1. Study aims.
2. Description of the survey dataset.
3. Data collection form.
4. Sampling strategy.
5. Organisational issues (staff, training, subject flow, budget, timetable, &c.).

Section Two - Survey Analysis

Background to the survey analysis exercise

This is a hands-on exercise that you should follow seated at your computer. It concentrates on the practical skills needed to manage and analyse data from a two-stage cluster sampled health survey.

The exercise assumes a familiarity with the **`ANALYSIS`** module of **`EpiInfo`** and with basic computing and statistical concepts.

The dataset used in the exercise (**`NUTSURV.REC`**) is to be found on the supplied floppy disk.

The structure of the **NUTSURV.REC** dataset

Data from a nutrition survey is stored in an EpiInfo file called **NUTSURV.REC**. The following table lists the names, types, and codes of the variables in the **NUTSURV.REC** dataset. The variable types and the EpiInfo / EpiData *pictures* that would be used to define variables in EpiInfo **.QES** files or with the **DEFINE** command in **EpiInfo ANALYSIS** are also listed:

Variable	Contents	Type	Picture	Values / Codes
CL	Cluster number	Number	##	N/A
CH	Child number	Number	##	N/A
AGE	Age in months	Number	##	N/A

SEX	Sex	Character	<A>	M = Male
				F = Female

WT	Weight in Kg	Number	##.#	N/A
HT	Height in cm	Number	###.#	N/A
MUAC	MUAC in mm	Number	###	N/A
OD	Oedema	Logical	<Y>	Y/N
FB	First born	Logical	<Y>	Y/N
GM	Maternal grandmother alive	Logical	<Y>	Y/N

ORP	Orphan	Character	<A>	M = Mother
				F = Father
				B = Both
				N = Neither

EpiInfo tools for survey analysis

`EpiInfo` provides two modules of use in handling and analysing data from surveys:

`ANALYSIS`	Provides data-management facilities for all types of surveys and data-analysis facilities for random and systematic sampled surveys. **`ANALYSIS`** does **not** produce reliable results (e.g. when calculating confidence intervals) when used to analyse data from a cluster sampled survey.
`CSAMPLE`	Provides simple data-analysis facilities for data from a cluster sampled survey but has few data-management facilities.

If your data comes from a random or systematic sampled survey then **`ANALYSIS`** will probably provide all the analytical tools you will need. If your data comes from a cluster sampled survey then you must use **`ANALYSIS`** for data-management tasks (e.g. combining and recoding variables) and **`CSAMPLE`** for data analysis.

The sample survey data comes from a cluster sampled survey. This means you must first use **`ANALYSIS`** for data-management and then use **`CSAMPLE`** for data analysis.

If you are using weight-for-height data then you will also need to use the **`EPINUT`** module of **`EpiInfo`**. This module calculates various nutritional indices, produces graphs, and performs basic data analysis. In this exercise we only will use **`EPINUT`** to calculate nutritional indices.

Data management

The data-management tasks you need to perform on a survey dataset will depend upon the data you collected and how you have coded it. With the sample survey dataset we will do some typical recoding and combining of variables.

Start **EpiInfo**. Start the **ANALYSIS** module. With the supplied floppy disk in the A: drive of your computer enter the command:

```
read a:\nutsurv.rec
```

to retrieve the sample survey dataset. Type the command:

```
browse
```

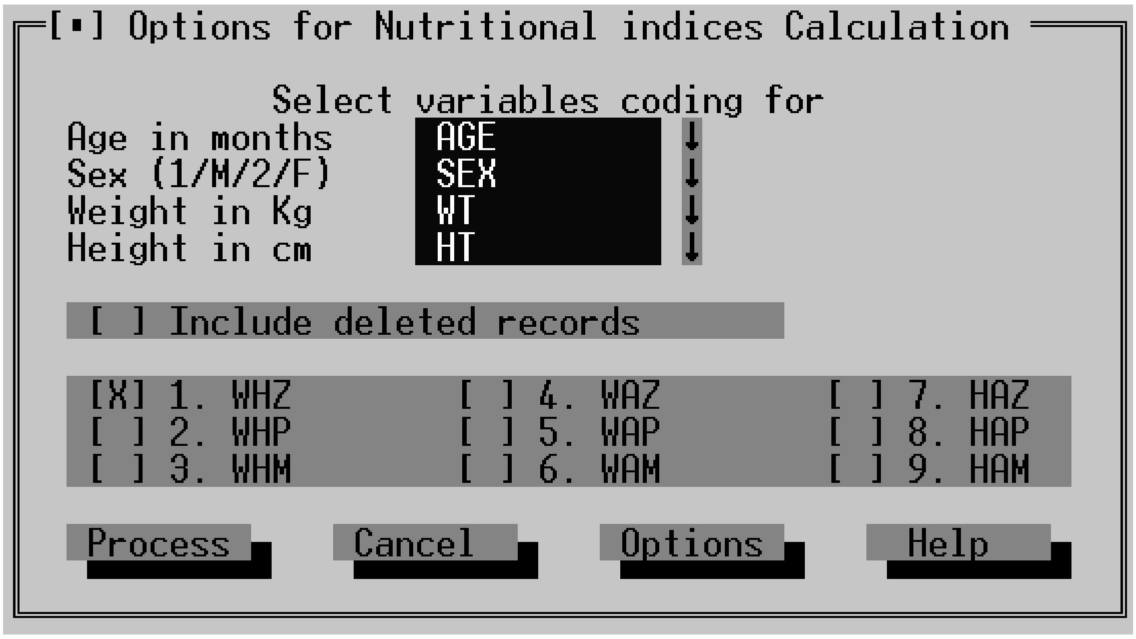
and examine the dataset. Make yourself familiar with the variables in the dataset.

When you have finished press Esc. Press F10 to quit **ANALYSIS**.

We will now use the **EPINUT** module to calculate nutritional indices and add them to our data. Start the **EPINUT** module. Select the option **Indices -> Add to a File** and specify the file **NUTSURV.REC**. You will then be asked for some information about the data that **EPINUT** needs to calculate nutritional indices. Specify the variable names:

Prompt	Variable
Age in months	AGE
Sex (1/M/2/F)	SEX
Weight in Kg	WT
Height in cm	HT

We do not want to include deleted records and we are only interested in the weight-for-height z-score (**WHZ**):



[•] Options for Nutritional indices Calculation

Select variables coding for

Age in months AGE

Sex (1/M/2/F) SEX

Weight in Kg WT

Height in cm HT

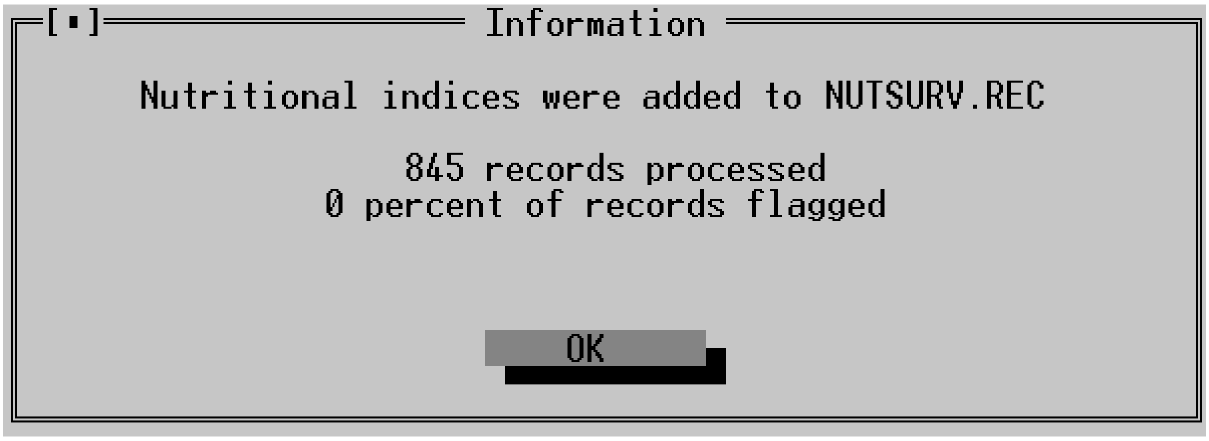
☐ Include deleted records

☒ 1. WHZ
☐ 2. WHP
☐ 3. WHM
☐ 4. WAZ
☐ 5. WAP
☐ 6. WAM
☐ 7. HAZ
☐ 8. HAP
☐ 9. HAM

Once you have specified the variables required to calculate the nutritional indices and specified the nutritional that indices you require, click the **Process** button.

Data management (continued)

EPINUT will display a progress bar followed by information about the data file listing the number of records and the proportion of records that have been 'flagged' because of unlikely combinations weight, height, age, and sex:



Nutritional indices were added to NUTSURV.REC

845 records processed
0 percent of records flagged

The sample dataset is 'clean' in the sense that **EPINUT** did not find any out-of-normal-range combinations of **WT**, **HT**, **AGE** and **SEX**.

Click the **OK** button and press F10 to return to the **EpiInfo** main menu.

Start the **ANALYSIS** module and enter the command:

```
read a:\nutsurv.rec
```

to retrieve the sample survey dataset. Type the command:

```
browse
```

to examine the dataset. Note that two variables (**WHZ** and **FLAG**) have been added to the dataset by **EPINUT**. The **FLAG** variable marks records with problems in the data used to calculate the nutritional indices. Before proceeding further you would examine records with a non-zero **FLAG** variable and, if possible, correct any errors. You do not need to do this with the **NUTSURV.REC** dataset. For your information, here is a list of the errors that can be detected by examining the **FLAG** variable:

	Problem calculating ...			
FLAG contains ...	**HAZ**	**WHZ**	**WAZ**	**Examine these variables first ...**
0				N/A
1	Y			Height / Age
2		Y		Weight / Height
3	Y	Y		Height
4			Y	Weight / Age
5	Y		Y	Age
6		Y	Y	Weight
7	Y	Y	Y	Weight / Height / Age

When you have finished examining the dataset press Esc.

Data management (continued)

The **AGE** variable contains the age of the respondent in months. If you want to produce age-specific prevalence rates you will need to group ages. First check the distribution of the **AGE** variable:

```
freq age
histogram age
```

We will groups the ages into bands centred on whole years. To group a variable you must create a new variable to hold the grouped data using the **DEFINE** command:

```
define agegrp #
```

then instruct **ANALYSIS** how to group the **AGE** variable using the **RECODE** command:

```
recode age to agegrp 6-17=1 18-29=2 30-41=3 42-53=4 54-65=5 66-hi=6
```

You should always check the results of any recoding operations:

```
freq agegrp
browse age agegrp
tables age agegrp
```

The sample survey dataset contains mid-upper-arm-circumference data (**MUAC**) and data indicating the presence of oedema (**OD**). You might want to combine these variables into a variables indicating undernutrition and grade of undernutrition. Remember the MUAC case-definitions:

Level	Definition (MUAC)
Global	MUAC of less than 125mm and / or bilateral pitting oedema.
Moderate	MUAC of between 110mm and 124mm (inclusive).
Severe	MUAC of less than 110mm and / or bilateral pitting oedema.

Check the distribution of the **MUAC** variable:

```
freq muac
histogram muac
```

We will create a variable to indicate undernutrition ascertained using MUAC (coded as 1 = Yes, 2 = No):

```
define globalmuac #
globalmuac = 2
if muac < 125 or od = "Y" then globalmuac = 1
```

We must remember not to include cases with missing values in our analysis:

```
if muac = . then globalmuac = .
```

The full-stop in the above command means 'missing'. We must also remember to exclude younger children from the analysis of **MUAC** data:

```
if age < 13 then globalmuac = .
```

You could use `age < 12` or `height < 75` (if you distrust the age data) to reject MUAC measurements taken from children aged less than one year.

Remember to check the effect of these commands using the **BROWSE** command:

```
browse age muac od globalmuac
```

The **GLOBALMUAC** variable is coded `(1)` to indicate undernourishment and `(2)` to indicate adequate nourishment.

Data management (continued)

Now we can create a variable to indicate grade of undernutrition:

```
define typemuac #
typemuac = 3
if muac < 125 then typemuac = 1
if muac < 110 then typemuac = 2
if muac = . then typemuac = .
if od = "Y" then typemuac = 2
if age < 13 then typemuac = .
```

Remember to check the effect of these commands using the **`BROWSE`** command:

```
browse age muac od typemuac
```

You can examine the prevalence of undernutrition using the **`FREQ`** command:

```
freq globalmuac typemuac
```

The codes used for **`TYPEMUAC`** are **`(1)`** moderately undernourished, **`(2)`** severely undernourished, and **`(3)`** adequately nourished.

The prevalence reported by **`ANALYSIS`** are correct. If we were to produces confidence intervals with **`ANALYSIS`** they would **<u>not</u>** be correct. This is because **`ANALYSIS`** assumes that the data was collected using simple random sampling.

We will now do the same for the weight-for-height z-score (**`WHZ`**) variable. Recall the case definitions:

Level	**Definition (Wt/Ht)**
Global	< 2 weight-for-height z-scores below NCHS / WHO reference mean and / or bilateral pitting oedema.
Moderate	< 2 weight-for-height z-scores and >= 3 weight-for-height z-scores below NCHS / WHO reference mean.
Severe	< 3 weight-for-height z-scores below NCHS / WHO reference mean and / or bilateral pitting oedema.

To apply these case definitions type the following commands:

```
define globalwhz #
globalwhz = 2
if whz < -2.00 or od = "Y" then globalwhz = 1
if whz = . then globalwhz = .
define typewhz #
typewhz = 3
if whz < -2.00 then typewhz = 1
if whz < -3.00 then typewhz = 2
if whz = . then typewhz = .
if od = "Y" then typewhz = 2
```

Remember to check the effect of these commands:

```
browse whz od globalwhz typewhz
```

If you have any ‘flagged’ records that could not be corrected, you might want to exclude them from analysis:

```
if flag <> 0 then globalwhz = .
if flag <> 0 then typewhz = .
```

Data management (continued)

We have completed the data management tasks needed to get the sample dataset ready for further analysis:

Task	**How done ...**
Add indices to file	**`EPINUT (Indices -> Add to a File).`**
Examine data for flagged records	**`BROWSE`** command in **`ANALYSIS`**.
Edit 'bad' (i.e. flagged) records	**`UPDATE`** command in **`ANALYSIS`** or **`ENTER`** module (we did not need to do this with the sample dataset).
Create age-groups	**`DEFINE`** and **`RECODE`** commands in **`ANALYSIS`**.
Create indicators for global malnutrition	**`DEFINE`** and **`IF .. THEN`** commands in **`ANALYSIS.`**
Create indicators for type of malnutrition	**`DEFINE`** and **`IF .. THEN`** commands in **`ANALYSIS.`**

The final data-management task is to produce a file that can be used by the **`CSAMPLE`** module. **`CSAMPLE`**, the program we will be using to analyse this data, requires that data is sorted by cluster. Issue the command:

```
sort cl
```

Once you have recoded and combined the variables and sorted the data by the variable(s) indicating the sampling units (e.g. the variable used to indicate cluster, **`CL`** in this case) you should write the recoded data to a working file on disk:

```
route nutwork.rec
write recfile
```

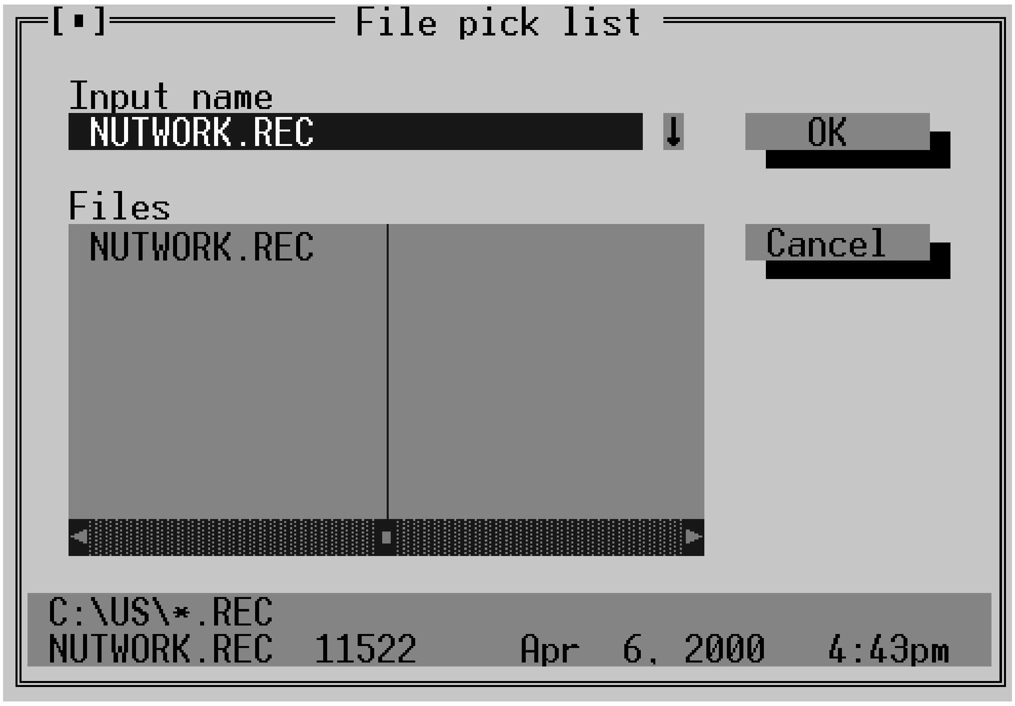
and leave **`ANALYSIS`**:

```
quit
```

Data analysis

When dealing with data from a cluster sampled survey you should remember that traditional analysis tools assume a random or systematic sample. You should use specially designed data-analysis tools such as **`CSAMPLE`** to analyse data from complex sample surveys.

Start **`CSAMPLE`** and select the file **`NUTWORK.REC`** when prompted:



Input name: NUTWORK.REC

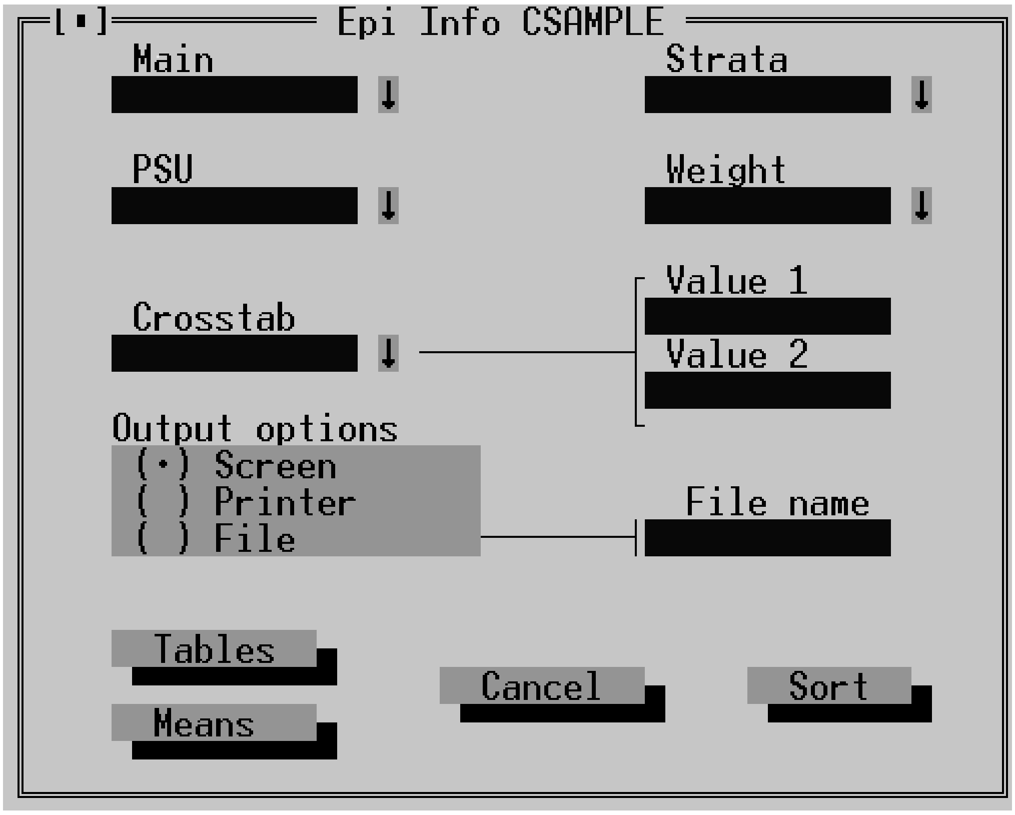
Files: NUTWORK.REC

```
C:\US\*.REC
NUTWORK.REC  11522      Apr  6, 2000   4:43pm
```

And press ↵ or click the **`OK`** button. The **`CSAMPLE`** analysis definition screen appears.

The CSAMPLE analysis definition screen

The **CSAMPLE** analysis definition screen looks like this:



(•) Screen
( ) Printer
( ) File

The information that **CSAMPLE** requires to perform an analysis is:

Variable	Holds ...
Main	The variable of interest (e.g. **GLOBALMUAC**). You must specify this variable.
Strata	The variable that specifies the stratum that each record belongs to. This variable is optional.
PSU	The variable that identifies the cluster that each record belongs to (e.g. **CL** in the example dataset). Clusters are the first thing you select in a two-stage cluster sample so this variable is sometimes called the ***primary*** *sampling unit* or *PSU*. This variable is optional.
Weight	This variable is used if you want to give some records more weight than others in an analysis. This variable is optional.
Crosstab	The variable that identifies the subgroups that you wish to analyse the main variable by (e.g. **AGEGROUP** or **SEX**). You may limit the analysis to two subgroups if required. This variable is optional.

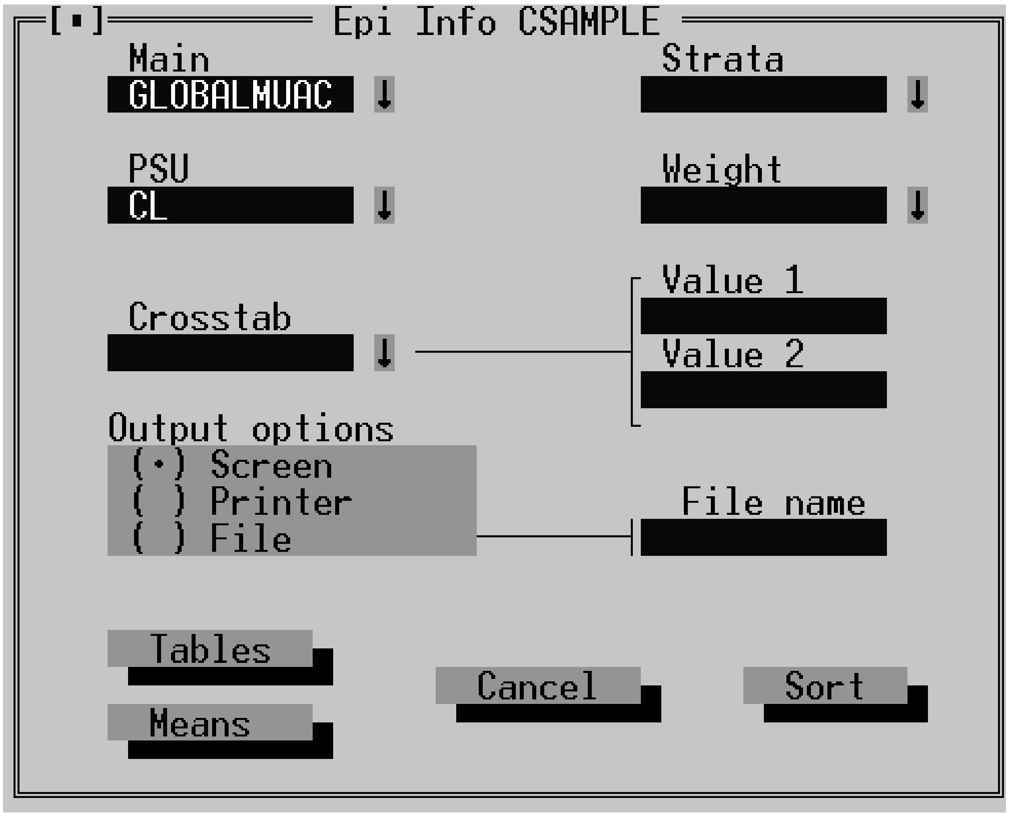
CSAMPLE provides two basic types of analysis:

Analysis	Used when . . .
Tables	Both the **Main** variable and the **Crosstab** variable (if specified) are categorical.
Means	The **Main** variable is continuous and the **Crosstab** variable is categorical.

The analyses available in **CSAMPLE** are limited to two variables. It is possible to produce (e.g.) a table showing age-specific prevalence rates but **<u>not</u>** a table of age- and sex-specific prevalence rates. To perform this analysis you would need to use **ANALYSIS** to produce separate files for males and females (using the **SELECT**, **ROUTE**, and **WRITE RECFILE** commands) or a single file with a combined age / sex variable and then use **CSAMPLE** to perform two separate analyses.

A simple **CSAMPLE** analysis

To perform a simple **CSAMPLE** analysis you need to specify the data (.REC) file, the Main and PSU variables and the type of analysis you wish to perform. Here is how you would specify a simple analysis of the prevalence of global undernutrition using the **GLOBALMUAC** variable:



Main: GLOBALMUAC

PSU: CL

Output options: (•) Screen

Specify the appropriate **MAIN** (**GLOBALMUAC**) and **PSU** (**CL**) variables and Click the **Tables** button. **CSAMPLE** produces the following output:

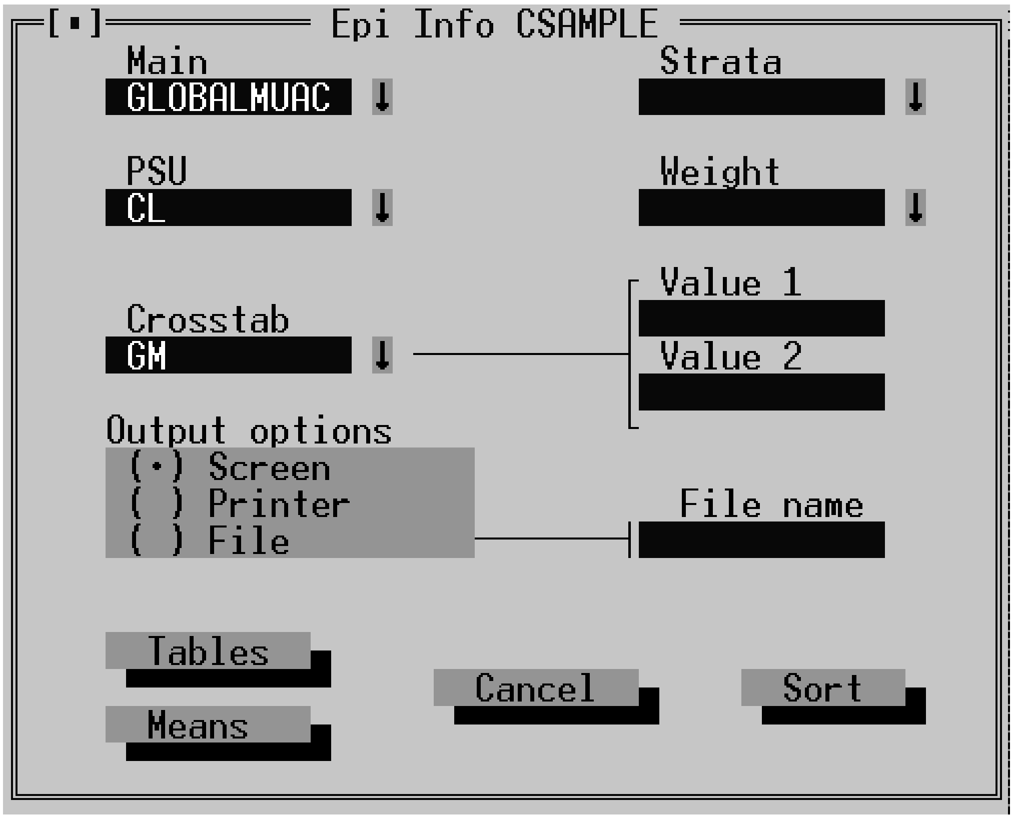
```
GLOBALMUAC
¦           ¦Total      ¦
+-----------+-----------¦
¦1          ¦           ¦
¦ Obs       ¦         62¦  <-- Count of records in each cell
¦ Percent   V      7.939¦  <-- Proportion of records in this cell
¦ SE%       ¦      1.164¦  <-- Standard error of proportion
¦ LCL%      ¦      5.658¦  <-- Lower 95% CI of proportion
¦ UCL%      ¦     10.219¦  <-- Upper 95% CI of proportion
+-----------+-----------¦
¦2          ¦           ¦
¦ Obs       ¦        719¦
¦ Percent   V     92.061¦
¦ SE%       ¦      1.164¦
¦ LCL%      ¦     89.781¦
¦ UCL%      ¦     94.342¦
+-----------+-----------¦
¦Total Obs  ¦        781¦  <-- Total number of records analysed
+-----------+-----------¦
¦Design eff.¦      1.447¦  <-- Calculated design effect
+-----------+-----------+
```

Identify the prevalence figures and 95% confidence interval for global undernutrition. Use **CSAMPLE** to produce prevalence estimates for moderate and severe undernutrition (i.e. **TYPEMUAC**).

Press Esc to close the output screen and repeat the analysis using the WHZ-based indicator.

More complex **`CSAMPLE`** analysis

More complex **`CSAMPLE`** analyses require you to specify a Crosstab variable. This analysis produces prevalence figures for children whose maternal grandmother is alive and for those whose maternal grandmother is dead:



Main: GLOBALMUAC
Strata:
PSU: CL
Weight:
Crosstab: GM
Output options
(•) Screen
( ) Printer
( ) File

Click the **`Tables`** button. **`CSAMPLE`** produces the following output:

```
¦GM         ¦GLOBALMUAC
¦           ¦1          ¦2          ¦Total      ¦
+-----------+-----------+-----------+-----------+
¦+          ¦           ¦           ¦           ¦  <-- Value of Crosstab variable
¦ Obs       ¦         31¦        448¦        479¦  <-- Count of records in cell
¦ Percent   V     50.000¦     62.483¦     61.489¦  <-- Column (V = vertical) %age
¦ Percent   H      6.472¦     93.528¦    100.000¦  <-- Row (H = horizontal) %age
¦ SE%       ¦      1.654¦      1.654¦           ¦  <-- SE of row percentage
¦ LCL%      ¦      3.231¦     90.287¦           ¦  <-- Lower 95% CI of row %age
¦ UCL%      ¦      9.713¦     96.769¦           ¦  <-- Upper 95% CI of row %age
¦ Deff.     ¦      2.164¦      2.164¦           ¦  <-- design effect
+-----------+-----------+-----------+-----------+
¦-          ¦           ¦           ¦           ¦
¦ Obs       ¦         31¦        269¦        300¦
¦ Percent   V     50.000¦     37.517¦     38.511¦
¦ Percent   H     10.333¦     89.667¦    100.000¦
¦ SE%       ¦      1.856¦      1.856¦           ¦
¦ LCL%      ¦      6.695¦     86.028¦           ¦
¦ UCL%      ¦     13.972¦     93.305¦           ¦
¦ Deff.     ¦      1.116¦      1.116¦           ¦
+-----------+-----------+-----------+-----------+
```

Identify the prevalence figures and 95% confidence interval for global undernutrition in children with and without a living maternal grandmother. Note that **`CSAMPLE`** displays confidence intervals for row (**`H`**) percentages only. Use **`CSAMPLE`** to produce age-specific prevalence rates of global undernutrition. Repeat the analysis using the WHZ-based indicator.

Further analysis of the sample survey

Working in your group, use **`ANALYSIS`** and **`CSAMPLE`** to produce a more complete analysis of the sample survey dataset.

Produce a **short** report using the data provided in the **`NUTSURV.REC`** dataset. This report should include:

A description of the sampling method used.

A list of the case-definitions used.

The prevalence of undernutrition in a compact format. For example:

		Severe (< -3.00 Z)		Moderate (< -2.00 Z)		Normal (>= -2.00 Z)	
Age	N	N	%	N	%	N	%
06 - 17							
18 - 29							
30 - 41							
42 - 53							
54 - Hi							
Total							

And:

	< -2.00 Z	>= -2.00Z
Oedema PRESENT	Marasmic Kwashiorkor N = % =	Pure Kwashiorkor N = % =
Oedema ABSENT	Marasmic N = % =	Normal N = % =

The distribution of the anthropometric indicators used (e.g. WHZ, MUAC).

An analysis of the risk factors and risk markers associated with undernutrition.